Singing Voice and Music Separation

Haoyu Zhang, Songbo Wu, Vugar Javadov
Department of Computer Science
Boston University
[haoyuz,wsb97o,jvugar].bu.edu

Abstract

In the last decade, compressed sensing has been widely used for signal recovery. In this paper we utilized compressed sensing for the separation of the voice and background noise in music. The problem is considered under the assumption that the voice in the music is inherently sparse and that the music signal is low rank. Using those features in different sound sources, the separation problems can be posted as a RPCA problem. So in this project, we utilized robust principal component analysis for performing voice-music separation. And it gives a good result.

Indexed-terms: Robust Principal Component Analysis, Music/Voice Separation

1 Introduction

Compressed sensing has be wildly applied on a lot of signal process algorithms in the last decade. And one of its typical application is signal denoising. Given the fact that noise is usually dense, we can recover the original signal utilizing its low sparsity.

In this project we stepped forward and tried to tackle the voice/music separation problem. This problem can be also considered as a denoising problem, and recovery of any of the voice/music signal is enough to produce a good result. But in practice, due to the fact that neither music nor voice is actually dense, we have to make a different assumptions. In most cases, a voice signal is sparse and a music signal has low rank, and thus we can utilize the robust PCA to do the source separation.

2 Previous Work

In some previous papers, people have tried many methods to do the source separation, supervised or unsupervised. For supervised algorithms, sparse dictionary encoding and recurrent neural network are used to solve the problem. Y.-H. Yang,[2], P.S-. Huang, M. Kim, M. H-. Johnson, P.Smaragdis. For unsupervised learning algorithms, the most popular methods are robust PCA, K-SVD .e.t. In P.S-. Huang, M. Kim, M. H-. Johnson, P.Smaragdis[6], they proposed the method using robust PCA to separate human voice from background music.

3 Method & Experiments

As most of other signal processes based on compressed sensing, the input signal will be transformed into a sparse domain. In this project, we transformed the signal using short time Fourier transformation into time-frequency domain. Unlike original Fourier transformation, STFT can also reflect the changes of signal in the time domain. By comparing the voice signal and music signal, we can find that human

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

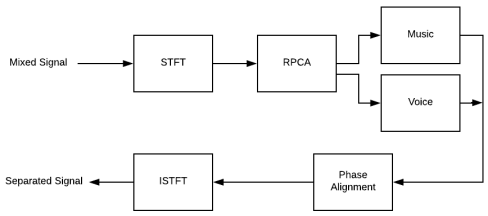

Figure 1: Framework

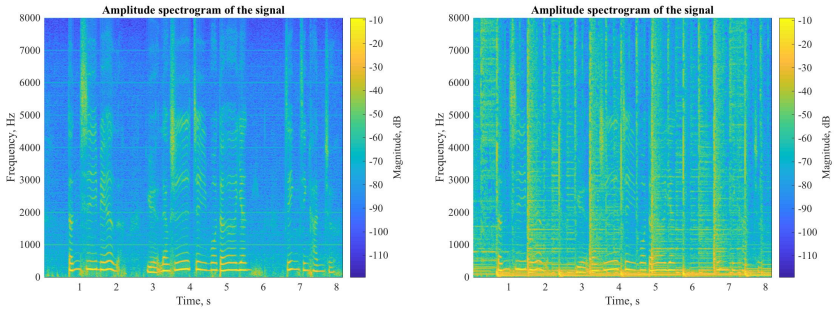

Figure 2: a) spectrogram of human voice. b) spectrogram of background music

voice is sparse in both time and frequency domain. While background music contains some repeating pattern and has low rank.

Under this assumption the source separation problem can be expressed in the following form,

$$minimize\ \|M\|_* + \lambda\|V\|_1$$
$$subject\ to\ S = M + V$$

Here M and V represents background music and singing voice respectively. And S is the input mixed signal. λ is used to control the weights on sparsity of V. Here $\|\cdot\|_*$ denotes the nuclear norm and $\|\cdot\|_1$ denotes the $L1 - norm$. This minimization problem can be carry out using RPCA. In our experiments, we used inexact ALM algorithm to find the solution. Notice that the STFT of the input signal is a complex matrix, to fit in the RPCA, we substitute S with $|S|$ and assume both music and singing voice have the same phase with S. And so the recovered singing voice and music is computed as follows,

$$M_{rec} = M * Phase(S), V_{rec} = V * Phase(S)$$

In practice, the choose of λ will influence the recovered quality of sing voice and music. We test the algorithm on the MIR-1K data set, and get the following observations. When λ is larger the recovered singing voice get better result, while when λ is smaller the music voice gets better. Because we are using inexact ALM, the sum of V_{rec} and M_{rec} may not equal to the S.

As you can see, to get a compromised result, we choose a λ approximately 0.2. And get the following result.

In our experiments, we found it very hard to determine, which λ produces the best answer. Typically, high frequency signal and low frequency signal get their best performance with different λ. As the curves in Figure 5 show, it seems the best λ is bigger for high frequency part. To solve this problem, intuitively we tried to divide the signal into two parts, high-frequency and low-frequency respectively. And used the same method to recover the signal.

And in our experiment, we found that it can somehow improved the current result.

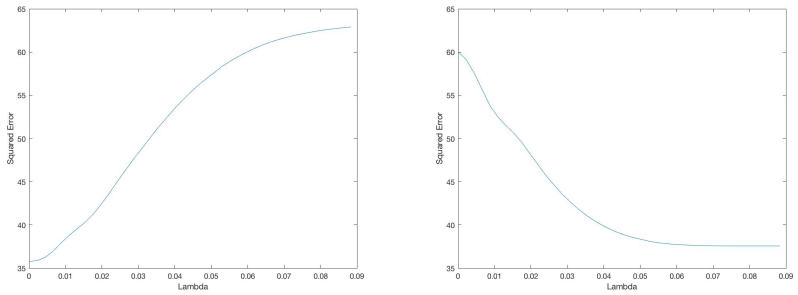

Figure 3: a) The squared error of recovered voice with different λ b) The squared error of recovered background music with different λ

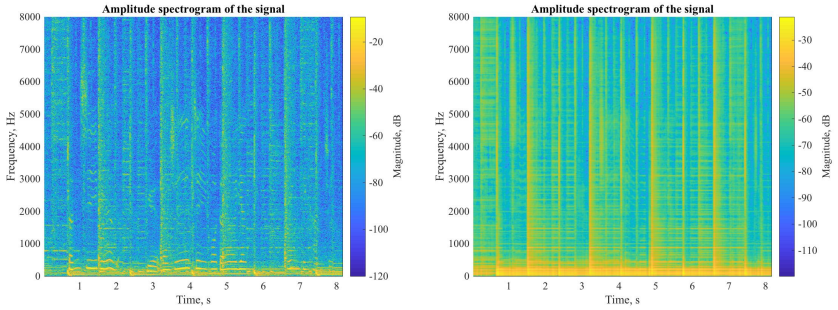

Figure 4: a) The recovered singing voice signal b) The recovered music signal

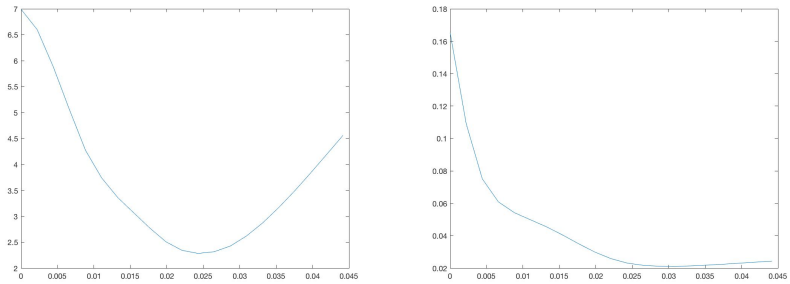

Figure 5: a) The squared error of recovered singing voice signal in low frequency b) The squared error of recovered singing voice signal in high frequency

4 Conclusion and Extension

In this project, we used a RPCA to solve the problem of singing voice and music separation. Afterwards, it was argued that under two main assumptions, namely, that music can be represented in low-rank and voice has sparse signals, a better algorithm can be implemented to solve the problem of music/voice separation. Specifically it was proposed to apply Robust Principal Analysis for voice, background noise separation in the music.

In our experiments, we find a very interesting phenomenon. When we checked the separated singing voice, we find that the background music is not removed when the singer is singing. But when the singing voice get smaller, the background music is separated much better. In other words, this method somehow performs in a way that it can pick those time segments when the singer is singing. But what it can't do is to separate those really mixed sounds. This indicates that the method fails in getting frequency features of two different sounds.

For future work, one promising solution is to utilize convolutional dictionary learning. Since the windows size of STFT is very hard to choose in practice. Using convolution operation can tackle the problem. In practice a lot of neural network has been built for image decoding. And for sound, it might also be useful. When we assume the human throat and instruments are linear systems. The the output sound will simply be the convolution of a sparse signal and a dictionary. And the music voice separation problems can be post as follows,

$$minimize_{V,M} \ \|V\|_1 + \lambda_1 \|M\|_1 + \lambda_2 \|S - D_M \star M - D_V \star V\|_2$$

Since the Fourier transformation of the convolution of two signals is equivalent to the element-vise multiplication of Fourier transformation of two signals. It can be converted to a classic $L1$ minimization problem.

References

[1] Minh Dao, Sang Chin, Yuanming Suo, Trac D.Tran, "Structured Sparse Representation with Low-rank Interference", Draper Laboratory, 2015.

[2] Y.-H. Yang, "Low-rank Representation of Both Singing Voice and Music Accompaniment via Learned Dictionaries", Research center for IT Innovation, 2013.

[3] M. Balouchestani, S. Krishnan, "Advanced K-means Clustering algorithm for Large ECG Data sets based on a collaboration of compressed sensing theory and K-SVD approach", Springer-Verlag London, 2014.

[4] C.G-.Cardona, B.Wohlberg, "Convolutional dictionary learning", IMT-SLM, IMT-LBM, CIF-SBR, 2017.

[5] E. Candes, J. Romberg, T. Tao, "Robust Uncertainty Principles: Exact Signal Reconstruction from Highly Incomplete Frequency Information", Caltech, University of California, 2005.

[6] P.S-. Huang, M. Kim, M. H-. Johnson, P.Smaragdis, "Joint Optimization of Masks and Deep Recurrent Neural Networks of Monaural Source Separation" IEEE/ACM Transactions on Audio, Speech and Language Processing, 2015.

[7] Y.-G. Zhang, C.-S. Zhang, "Separation of Music Signals by Harmonic Structure Modeling", Tsinghua University, 2009.

[8] P.-S. Huang, S.D. Chen, P. Smaragdis, M. H-. Johnson, "Singing-Voice Separation from Monaural Recordings Using Robust Principal Component Analysis", UIUC, 2012.