Source Separation of Single-Channel Mixed Speech in Computational Auditory Scene Analysis

ECE 6255 Final Project Report

Ha, Soo Kwon

Shin, Donghoon

Submitted 2019 April 28

Executive Summary

The audio source separation, or “cocktail-party effect,” attempts to separate the designated target signal from a mixed signal. In this project, the specific Computational Auditory Scene Analysis method was chosen to separate the mixed speech in *single-channel*. The single-channel source separation is particularly challenging in that the separation process is done with less amount of reference information compare to the multi-channel source separation method. To separate and recover each source from the mixed signal in one microphone, we decided to use the *refiltering* method rather than a traditional *unmixing* method to generate masking function and reweight each sub-band of the multiband signal with the processed masking coefficients. Sparse non-negative matrix factorization (SNMF) method was used as a mean to construct the statistical machine learning frame for training and processing the given data. Perceptual Evaluation of Speech Quality (PESQ), Short-Time Objective Intelligibility (STOI), Extended STOI (ESTOI) and Signal-to-Noise And Distortion (SINAD) were used as evaluation metrics to compare between the clean reference signal and the separated signal.

1. Introduction

1.1 Background

The source separation has become a research topic since Colin Cherry discovered in 1953, and it has been a very popular research area in various applications such as medical imaging and speech separation. From then to the 1990s, this problem was interpreted as BSS (Blind Source Separation) and solved with ICA (Independent Component Analysis). As shown in [1, Fig. 1], the premise was that the source signals were observed through the sensors at step A, and then the observed mixed signal was used to estimate each unknown source signal and unmix through the separation process at step B.

The speech separation algorithm used in BSS and ICA depend on the multi-channel mixed source observation, so it was somewhat challenging to construct an accurate separation algorithm for a mixed signal from a single microphone. However, as machine learning develops, this problem could also be solved with machine learning technology, even with a single channel.

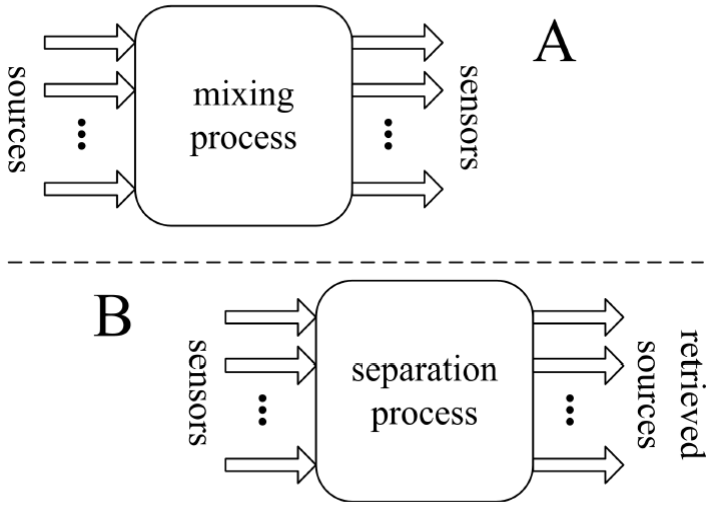

Figure 1. Visualization of the past sensor-dependent source separation problem

1.2 CASA vs. BSS

The two prominent speech source separation methods used nowadays include Computational Auditory Scene Analysis and Blind Source Separation. CASA involves machine learning technique to construct "machine listening" algorithm that resembles human auditory system separating a mixture of sound sources. This psycho-auditory method observes and analyzes the primitive feature of the speech data such as pitch, periodicity, and continuity. For example, CASA observes the behavior of each sub-band of multiband signal, and the similarly characterized frequencies are grouped together depending on the onset, offset upward or downward sweep. Then according to the grouped signals, the masking coefficients are created and applied to the corresponding sub-bands to reconstruct the separated signals. This is called *refiltering* as illustrated in [2, Fig 2].

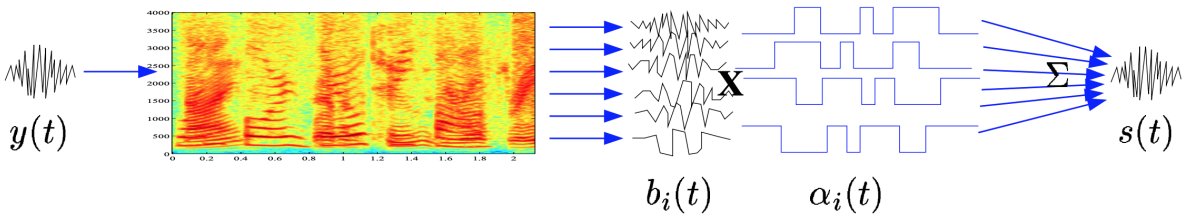

Figure 2. Refiltering for single-channel source separation

On the other hand, BSS is more of a pure machine-based separation algorithm which unmix and recover with multiple observation sequences. BSS and ICA decompose the mixed signals into independent non-Gaussian signals.

1.3 NMF and SNMF

Non-Negative Matrix Factorization (NMF) and Sparse Non-Negative Matrix Factorization (SNMF) are examples of CASA with high accuracy for separating sources in single channel recordings. The basic idea of NMF is with the assumption that each superpositioned signals in the mixed signal is positively added on top of each other and therefore can be decomposed into the separated signals. The basic NMF principle follows the ground spectrogram representation of Y=DH, where D is the dictionary matrix where the distinctive features are stored and H is a sparse code matrix. If the dictionary is diverse enough, Y can be decomposed or separated with respect to the sparsity represented in H.

However, NMF "does not provide a well-defined solution in the case of overcomplete dictionaries," [3] in other words, NMF is not very effective for obtaining sparse solution due to lack of ability to sufficiently detect and characterize sparse non-negative data. The solution to this problem is SNMF as name shows. The energy function for SNMF is shown below, with the

$$E = ||\mathbf{Y} - \bar{\mathbf{D}}\mathbf{H}||_F^2 + \lambda \sum_{ij} \mathbf{H}_{ij} \quad \text{s.t.} \quad \mathbf{D}, \mathbf{H} \geq \mathbf{0}$$

Equation 1. The cost function that the sparse non-negative matrix factorization (SNMF) optimizes from Schmidt and Olsson[3]

second portion of the RHS of the equation is added to compare to the original cost function, showing that there is a sparsity control parameter λ. The training and learning dictionaries follow the below equation, where Y contains all the training data and D is a fixed dictionary. Only the code matrix H needs to be updated to perform the source separation. The speech then goes through a separation with the recognition of sparse decomposition dictionaries.

$$\mathbf{H}_{ij} \leftarrow \mathbf{H}_{ij} \bullet \frac{\mathbf{Y}_i^\top \bar{\mathbf{D}}_j}{\mathbf{R}_i^\top \bar{\mathbf{D}}_j + \lambda}$$

$$\mathbf{D}_j \leftarrow \mathbf{D}_j \bullet \frac{\sum_i \mathbf{H}_{ij} \left[\mathbf{Y}_i + (\mathbf{R}_i^\top \bar{\mathbf{D}}_j)\bar{\mathbf{D}}_j\right]}{\sum_i \mathbf{H}_{ij} \left[\mathbf{R}_i + (\mathbf{V}_i^\top \bar{\mathbf{D}}_j)\bar{\mathbf{D}}_j\right]}$$

Equation 2. The recursive process to find optimum D and H from Schmidt and Olsson[3]

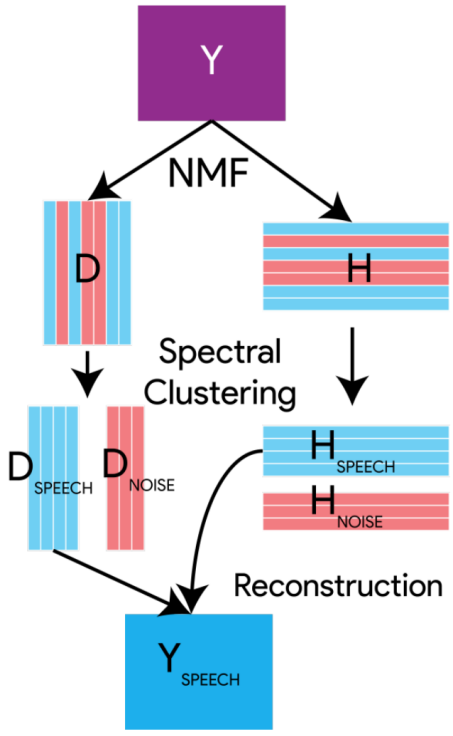

Figure 3. The general process that the sparse non-negative matrix factorization separates the speech from Nicholas Shu's qualifying exam[4].

2. Problem Statement

2.1 Objective

The main objective of the whole experiment is to extract the target voice from the mixed signal of two speech sources. The separated signals are compared with the original reference speech recordings to see if the separation was successful. We were aiming for constructing separation method that much more resembles the human auditory speech separation process rather than computing aspect, therefore although the phoneme-based separation is known as a more accurate way of executing the process, we used the direct unsupervised approach.

2.2 Design Approach

CASA is a more appropriate method for one microphone source separation, so we used it instead of BSS. The target signal to be extracted was Soo Kwon's voice signal. Dana Eun Bin Lee's voice was added on top of Soo Kwon's voice to produce a mixed signal for the male vs. female situation, and Donghoon's voice was mixed with Soo Kwon's voice for male vs. male situation. The primary experiment condition is that the target signal gain is greater than or equal to 1, in other words, Soo Kwon's voice is louder than Dana or Donghoon's voice. The STFT was taken for each speaker's data, and the obtained frequency matrices were trained through SNMF, and the obtained dictionaries were used to separate the signals. Each separated signals was reweighted with chosen masking coefficients.

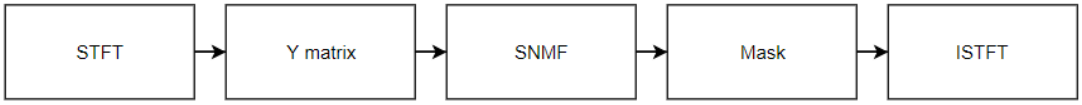

Figure 3. The chosen design of generalized whole separation process

3. Experiment

3.1 Data Collection

22 minutes of speech samples were collected from two male(Soo Kwon Ha and Donghoon Shin) and one female speaker(Eunbin Lee). From 0 to 20 minute of the speech samples were used for the training data to train dictionary array D. 20 to 22 minute of the speech samples were used to evaluate the speech separation algorithm. The speech data were from reading English article, and the speech data were recorded at the sampling rate of 44kHz but downsampled through Wavesurfer to 16kHz.

3.2 Implementation

3.2.1 Pre-processing and STFT

The speech data were passed through the lowpass filter that only accepts frequency lower than 4kHz. Then the data were normalized to make the gain of the signal to be 1. After that STFT was computed with the Hamming window of 64ms (= 1024 samples) at a sampling rate of 16kHz. An overlap of 50 percent (= 512 samples) was used between frames. FFT was done with the length of 2048 samples (= with 1024 padded zeros), that yielded 1025 frequency bins. Also, the absolute operator was used to make all the data index non-negative.

3.2.2 SNMF for training data

TIMIT data of each speaker was used to learn each speaker dictionaries using direct, unsupervised sparse non-negative matrix factorization. The method to calculate the distance between Y and DH was done by Kullback-Leibler divergence. The design parameter of SNMF was determined according to Schmidt and Olsson's literature [3] to maximize SNR. Sparsity parameter λ was chosen to be 0.1 among [0.0001 0.001 0.01 0.1] which means the algorithm values sparsity the most among choices. The dictionary size is 560, the largest among Schmidt and Olsson's suggestion, [70 140 280 560]. The maximum number of iteration of the algorithm for the training data was 300 which ends up with small energy function lower than 1.

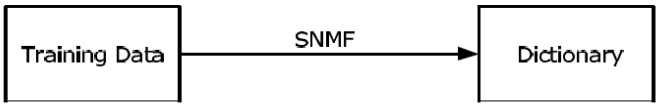

Figure 4. Conceptual visualization of SNMF

3.2.3 SNMF for test data

The TIMIT data of each speaker's test samples were added to have one mixed speech data. The dictionaries created in the training session was used to separate mixed speech data. In the SNMF algorithm to separate speech, the two dictionaries were concatenated. The size of 1120, concatenated dictionary was an input for the algorithm which was used to find concatenated code matrix H. The maximum number of iteration of the algorithm for the training data was 500 which ends up with small energy function, lower than 10.

$$\underbrace{s(t)}_{\text{estimated source}} = \overbrace{\alpha_1(t)}^{\text{mask 1}}\ \underbrace{b_1(t)}_{\text{sub-band 1}} + \overbrace{\alpha_2(t)}^{\text{mask 2}}\ \underbrace{b_2(t)}_{\text{sub-band 2}} + \ldots + \overbrace{\alpha_K(t)}^{\text{mask K}}\ \underbrace{b_K(t)}_{\text{sub-band K}}$$

Figure 5. Masking and reweighting

By applying the SNMF algorithm on the test data using the pre-trained dictionary, TIMIT data for each speaker is achieved. However, it is not possible to apply inverse STFT in this data, because this TIMIT data is non-negative, i.e only have the magnitude of the complex data. Therefore, it is needed to multiply with the phase of TIMIT data of the mixed signal in advance to apply inverse STFT. After applying inverse STFT, the real part of the time domain data would be the separated speech data.

3.3 Result

3.3.1 Male vs. Female Speaker

The result of speech separation between the male speaker and the female speaker was done enough for the listener to hear the speech of a single speaker. In other words, it was still possible to hear other speaker's mumbling sound, but the focused speaker's speech was dominating the other speaker's mumbling sound. The spectrogram in Figure 6 shows how the SNMF approach accurately separated the signal.

3.3.1 Male vs. Male Speaker

The result of speech separation between the male speaker and another male speaker was done poorly that the listener couldn't focus on the single speaker. It was more like amplifying one speaker's speech that made the listener understand which speaker's speech was separated. The spectrogram in Figure 7 shows how the SNMF approach separated the signal.

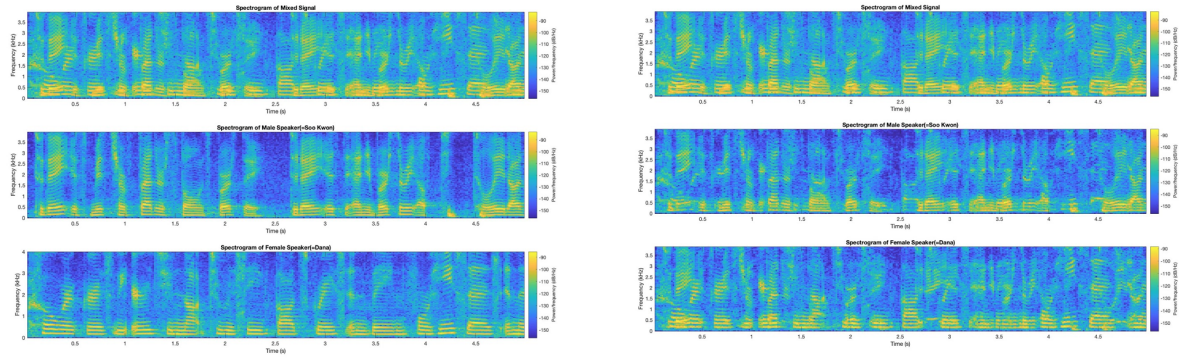

Figure 6. Speech separation between male speaker vs female speaker. Spectrogram of the mixed speech signal (top), male speaker (middle) and the female speaker (bottom). Compared with the ideally separated signals (left) and the signals separated with SNMF.

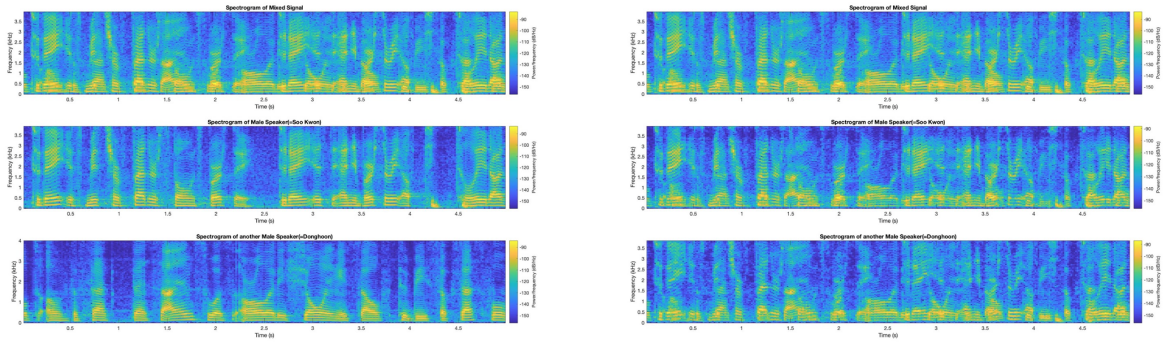

Figure 7. Speech separation between male speaker vs male speaker. Spectrogram of the mixed speech signal(top), male speaker(middle) and another male speaker (bottom). Compared with the ideally separated signals (left) and the signals separated with SNMF.

3.4 Evaluation

Table 1. Separation performance respect to the target signal

Case	PESQ	STOI	ESTOI	SINAD
Male vs. Female (Unseparated)	1.6360	0.6307	0.5892	-0.0842
Male vs. Female	1.9450	0.7404	0.5899	5.7488
Male vs. Male (Unseparated)	1.8720	0.7050	0.4592	0.4096
Male vs. Male	1.8890	0.7075	0.5313	3.9026

The separation performance and evaluation are shown in the above table. According to table 1, it is observable that the separation between mixed gender speech signal has better effect compare to that of the same gender situation. This is shown from the higher scores of each evaluation criterion for Male vs. Female situation compare to the Male vs. Male situation.

PESQ detect the average disturbance value and average asymmetrical disturbance value, and subtract the obtained values from 4.5 to calculate the overall speech quality. The evaluated speech segment was confirmed to be longer than 5 seconds because if the analysis frame is too short, PESQ underestimates linear frequency response distortions. [5] The quality score range was taken and adjusted to 0 ~ 5 scale, 5 as being the best quality and 0 being poor quality.

[6] STOI and ESTOI predict the intelligibility of the separated signal, but ESTOI observes the spectral correlation between noisy speech and clean speech and therefore more accurately analyzes the signal overlapped by highly modulated noise sources compare to the standard STOI. SINAD, or in other words SNDR (Signal-to-Noise & Distortion Ratio), measure the signal quality in respect to the noise and distortion level.

4. Conclusion

4.1 Discussion

As a conclusion, we have figured out that the separating target signal from the mixed gender speech signal using SNMF is more effective compare to the mixed signal created from identical gender.

As shown in the result and the evaluated scores, the separation was not very complete compare to our expectation. Also when we calculated each evaluation scores respect to Dana's voice or Donghoon's voice, we have observed that the scores were much lower than the values shown in table 1. We speculated that this could have been associated with the gain problem; Soo Kwon's voice was much louder than other two people's voice, so the separation of target signal might have been easier but vice versa is not quite as decent. The discrepancy between separation evaluation scores of each person might have been caused by the difference between silence length, pace of reading and dynamic level.

4.2 Future Application

As stated in the previous section, we were aiming for separating signals using human-method module rather than fully computational direction, and our findings and approach could be later developed with more advanced deep learning to create more accurate way of separating speech signals as human does. Other data training methods such as Hidden Markov Model (HMM) and Factorial HMM could also be used for pre-training the original speech signals from each person. The potential future research area would be: speech separation between moving vs. stationary and voiced vs. unvoiced sources in single-channel.

5. References

[1] Rasmus Kongsgaard Olsson: “Algorithms for Source Separation - with Cocktail Party Applications,” *Kongens Lyngby*, IMM-PHD-2006-181

[2] Sam T. Roweis: “One Microphone Source Separation,” *University College London*

[3] Mikkel N. Schmidt and Rasmus K. Olsson: “Single-Channel Speech Separation using Sparse Non-Negative Matrix Factorization,” *Informatics and Mathematical Modelling, Technical University of Denmark*

[4] Nicholas Shu: “Speech Recognition in a Medical Environment” *Georgia Institute of Technology*

[5] Scott Pennock: “Accuracy of the Perceptual Evaluation of Speech Quality (PESQ) algorithm,” *Lucent Technologies*

[6] Cees Taal, Richard Hendriks and Richard Heusdens: “A Short-Time Objective Intelligibility Measure for Time-Frequency Weighted Noisy Speech,” *Delft University of Technology*, Signal Information & Processing Lab

[7] Aapo Hyvarinen and Erkki Oja: “Independent Component Analysis: Algorithms and Applications,” *Helsinki University of Technology, Neural Networks, 13(4-5):411-430,* 2000

[8] "Computational Auditory Scene Analysis." Wikipedia. January 23, 2018. Accessed March 03, 2019. https://en.wikipedia.org/wiki/Computational_auditory_scene_analysis.

[9] Y. Salaun, E. Vincent, N. Bertin, N. Souviraa-Labastie, X. Jaureguiberry, D. T. Tran, and F. Bimbot: “The Flexible Audio Source Separation Toolbox Version 2.0,” *IEEE International Conference on Acoustics,* Speech and Signal Processing, 2014

[10] Kedar Patki: “Review of Single Channel Source Separation Techniques,” *University of Rochester*