ECE 6255
Final Project

Monaural Speech Separation with Sparse Non-Negative Matrix Factorization

Hua Chen
Kevin Cleary

April 28th, 2019

Introduction

Our project aims to study monaural speech separation problem from signal processing point of view and compare the result with equivalent Deep Neural Network approach.

In a crowded room, bar, or party setting, human perception of audio is not very good. All the conversations and sounds add and interfere, bounce off the walls and create "noise". Even though this noise is present, people have developed a biological mechanism to hear specific signals over the rest. Humans can focus on a particular conversation and "tune out" the noise. Aptly named the "Cocktail Party Effect", this ability is the result of processing that occurs within the human brain with the assistance of temporal, spatial, and image signals. For our final project, we aim to find a model through literature review that addresses this issue on a limited scope

Single-source speech separation is a subproblem based off the Cocktail Party problem. Two speakers are talking over one another and the goal is to separate the speech into individual sources. Since there is only one audio channel without any spatial data, the recording is denoted "single source". The problem is also called Blind Source Separation (BSS) due to ignorance of the original signals or how they were mixed. BSS could also apply to images or other types of signals. Theoretically, the mix can be composed of an arbitrary number of speakers. To simplify things, we are looking at two speakers.

In a real world design scenario, it may make more sense to integrate multiple microphones into a device to get better data. However, this problem is more focused on recordings that already have mixed signals. There are some instances where there is only one microphone is available. When two people are close, or two people share a microphone it would be valuable to have the technology to separate them. Software is almost always cheaper to implement than hardware. If this problem were to be solved it could cheapen recording technology. There are a whole host of single channel audio recordings of past films and media that could be separated retroactively. Another reason why many researchers still are interested in this area is because it is a clustering problem and there are many algorithms that tackle clustering for so many different applications. If an algorithm gets to the level of mimicking the human ability in a crowded room, surveillance agencies and others would be greatly interested.

Single-channel speech separation problem has been studied extensively in both signal processing and machine learning community, and a variety of models have been proposed and proved their efficiency. As part of the research and selection for this implementation, we want through numerous papers. Here are a few of the notable themes of papers we considered replicating:

- **Source-Filter** (VQ, HMM, Matrix factorization)
  Seeks to model the vocal tract filter and do pitch estimation using a different technique for each including vector quantization, HMM and NMF. This approach relies on and exploits the physical differences between the geometry of two speakers. [7]

- **MLE of vocal filter**
  Decompose the speech signal into the excitation signal and the vocal-tract-related filter and then estimate the components from the mixed speech using a hybrid model. This paper is a similar decomposition to the first but uses maximum likelihood as the technique to do the actual separation once the model is built. [8]

- **Latent variable decomposition of spectrograms**
  Attempts to construct the entire spectra for each speaker by identifying typical spectral structures for speakers through latent-variable decomposition of their magnitude spectra. [9]

- **HMM**
  Uses the ability of hidden Markov models to model dynamically varying signals and extends the approach to recognition in order to accommodate concurrent processes. There are many HMM approaches in this area.[10]

- **Neural Nets**
  Long Short-Term Memory recurrent neural networks are used for speech enhancement. Networks are trained to predict clean speech as well as noise features from noisy speech features, and a magnitude domain soft mask is constructed from these features. Obviously neural nets are a tool well suited to speech separation and many approaches are based on tweaking network parameters. [11]

Problem formulation

We wanted to use an approach that does not include black box techniques so we could concentrate on the digital signal processing. Our chosen approach is Sparse Nonnegative Matrix Factorization (SNMF).

Non-negative Matrix Factorization (NMF) has shown some significant improvement in speech separation. In [1], Schmidt and Olsson propose a Sparse NMF approach. Two sets of dictionaries are estimated for different speakers: one through computing SNMF over concatenated spectrograms for each speaker; and one through concatenating parts of training data corresponding to each phoneme for each speaker. For each mixed speech input, we concatenate the dictionaries of two speakers and compute code matrix using SNMF updates. We then reconstruct individual magnitude spectra. Spectral masking and spectrogram inversion are performed on separated waveforms using the original phase of mixed signal, yielding separated speech.

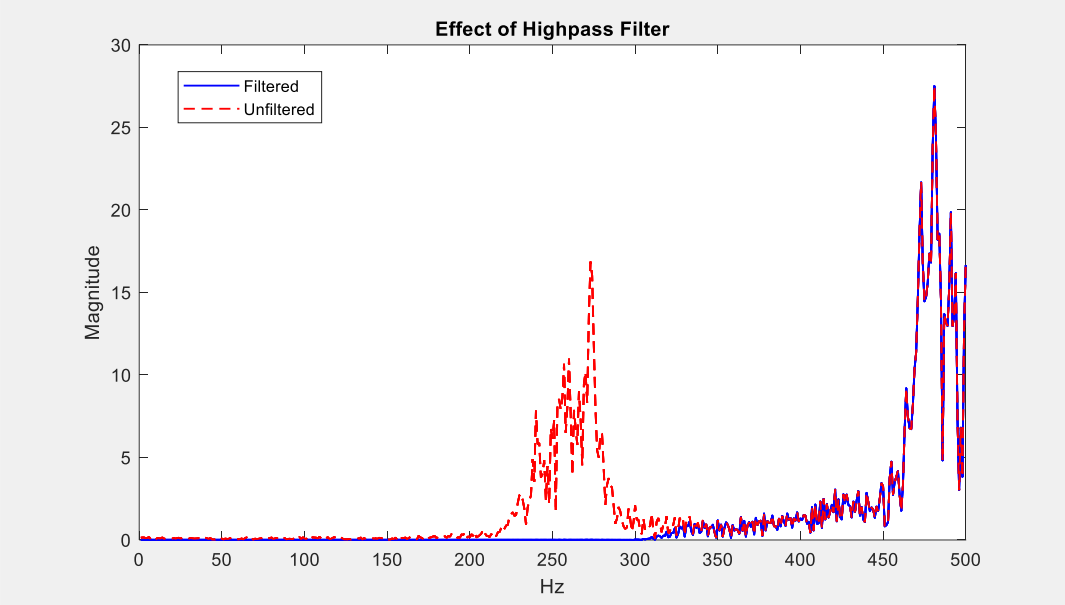
We also looked at a Deep Neural Net (DNN) implementation as a comparison and benchmark for a cutting-edge approach. In contrast to signal processing approach, DNN approaches rely on premixed datasets to build models of speech and/or noise. In [2], a joint optimization of a soft masking function and deep learning models (DNN or RNN) with discriminative training criterion is proposed. The idea is to learn the optimal hidden representations in order to reconstruct the target spectra. Soft-time-frequency masking function is incorporated into DNN directly and discriminative training technique that minimizes squared error between prediction and target ensures a better result. The DNN we used is a pre-trained model from the paper [6].

Preliminary findings from both of our chosen approaches yielded a noticeably improved qualitative sound. Seeing this, we proceeded with the rest of the process.

Experimental Configurations and Results

Our experimental approach closely tracked the Schmidt and Olsen paper [1] and Grais and Erdogan's paper [3]. For both the SNMF method and DNN method we used a MATLAB implementation with the TIMIT database as our corpus. We were able to find and use the MATLAB Audio Database Toolbox created by the Signal and Image Processing Lab at the Israeli Institute of Technology. This database tool helps query and filter results from the TIMIT database that would ordinarily have to be done manually. Using the tool, we selected 8 speakers from TIMIT. They are all from the "Army Brat" dialect group, meaning they do not have a strong regional accent. There are 4 female and 4 male speakers. They each have a corpus of 10 sentences. For each speaker, we took 8 sentences and played them back-to-back to form one long sentence. This was the training data and the remaining 2 are test data.

The training data was put through a high pass filter with a cutoff frequency of 350 Hz. This was done to clean up any low frequency noise that would occur prior to the first spectral peak.



A STFT was done on the training data with a Hamming window size of 800 samples (50ms) and an overlap of 50%. The SNMF was calculated on the STFT matrix. The purpose of the SNMF algorithm is to divide a matrix (Y) into two matrices, where the first is a matrix of basis vectors (D) and the second (H) is weight vectors for each basis. The SNMF algorithm optimizes the loss function:

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

The training of SNMF is the process of factorizing the optimal basis and corresponding weight with penalty on sparsity. If we set λ to be zero, the algorithm becomes Non-Negative Matrix Factorization. The number of basis vectors is a design choice. The "dictionary" for each speaker is the matrix of basis vectors.

Now the training is complete and the mix must be created, and then separated by the algorithm. The speech mixture was created by normalizing the two signals and then adding them together.

$$\mathbf{Y} = \sum_{i}^{R} \mathbf{Y}_i.$$

The STFT was performed on the mix to the same specifications as detailed for the dictionary. The two speakers' dictionaries are concatenated to make a super dictionary of basis vectors. We then run the SNMF to update only the weight matrix, so that we could optimally compose the new speech from the learnt features. Ideally, the speech that is produced by each speaker is captured by that speaker's dictionary. Imposing the L1 sparsity penalty on the weight matrix helps force the spectrogram component into one speaker or another's basis vector. Once the algorithm is complete, multiplying the speaker 1 dictionary with the speaker 1 weight matrix should reproduce the spectrogram of that speaker.

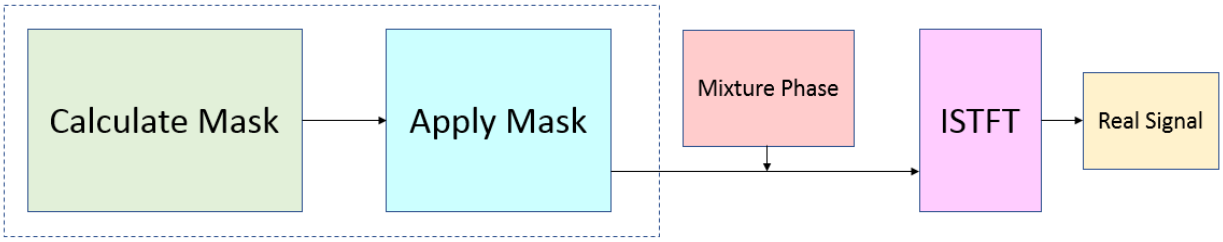
In reality, some of speaker 1 ends up in the spectrogram of speaker 2 and vice-versa. If we were to multiply speaker 1 weight by speaker 1 dictionary exclusively, we would introduce a lot of errors. Hence, we introduce a spectral masking technique to solve this issue, indicated below:

$$M = \frac{\tilde{Y}_1^{\,p}}{\tilde{Y}_1^{\,p} + \tilde{Y}_2^{\,p}}$$

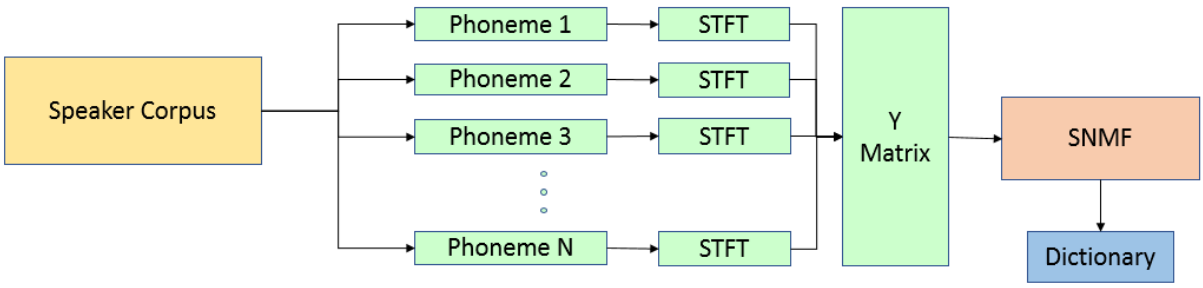
where $\tilde{Y}$ represents the SNMF estimate for each speaker. The mask is applied onto the mixture spectrum, extracting the magnitude weight proportionally. A p value of

2 is the well-known Weiner filter, and a *p* value of infinity is a hard matrix. It rounds the mask value to either 0 or 1. Choosing the p value is an exercise in trial and error, the best p value for our application was 5.

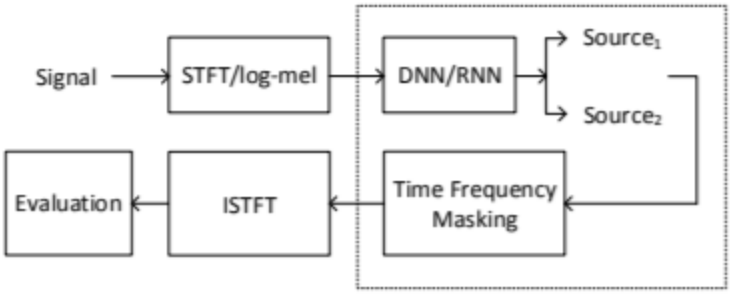
A feature of the SNMF algorithm is that is deals only in non-negative real values, meaning we cannot generate the phase information of each individual speaker. We assume that the phase of the mixture is the same as the phase of each speaker. With this phase and magnitude information, we can compute the inverse STFT and recover a real signal.



This real signal can be compared to the original and have metrics calculated off of its performance. A variant of this approach that Schmidt also employs is changing the original dictionaries. Instead of constructing the dictionary from every sentence back to back, every phoneme was placed back to back. The STFT was taken on each phoneme library instead of the entire corpus. This approach may create stronger basis vectors where transitions between phonemes are less pervasive.



The DNN process is extremely similar to the SNMF approach. The neural network essentially replaces the NMF algorithm.



The NMF algorithm can be thought of as a one-layer linear neural net with the basis vectors representing nodes and the weight matrix acting as the linear function. The DNN uses multiple layers and nonlinear functions. Therefore, the separation is not happening on the spectral domain or basis vectors of each speaker, but a nonlinear representation of the spectral domain, where hidden patterns are learned. The soft mask parameters are still deterministic but back propagation is used to feed the post-mask error back into the network. This was the nodes and functions are optimizing the final result. Since we were only interested in the DNN as a comparison, we did not experiment with the parameters but kept the pre-trained network as-is.

Now that the main process has been described we will go into the experimentation and variable manipulation. There are some features that can be tweaked to give different results.

Window size

J.R. Simpson's [4] indicates that a one size fits all approach for window size is not appropriate for speech separation via masks. He also proves that optimal SIR values occur at different window sizes for male speakers. We have found in our experiments that the female voice is separated more cleanly. Although we did not adjust this parameter, it may prove that a smaller window size may give better performance for male speech separation.

Mask type

The choice of mask as described previously will have an effect on the quality of the separation. Considering we tuned this value to give the best subjective performance, it is unlikely that investigating the parameter more will yield interesting results.

Sparsity

This is a critical element of the Schmidt paper. The database he trained on had hundreds of sentences from each speaker. It's feasible that the NMF algorithm would be able to split the weight between each speaker because there is so much data and

an erroneous match could be made. Enforcing the sparsity constraint forces the weight matrix to prefer one speaker over the other. Our dataset had so few sentences that sparsity was not really a concern in the weight matrix. However, since we don't have access to any other dataset, we had to set $\lambda = 0$. In future works, we will experiment on a bigger dataset to see how sparsity influence our result.

Number of basis vectors

There is no limit to the number of basis vectors that the NMF algorithm permits. There can be more basis vectors than vectors in the dictionary. Too few basis vectors will result in muddled noise that is not distinguishable as human speech. Generally, the higher the number of vectors the better since the algorithm can assign weights to more unique spectral characteristics. It is possible to tune the number of features for each mix, but a more general approach led us to stay at 560 for speaker dictionary and 280 for phoneme dictionary.

Sentence

The TIMIT dataset includes every speaker uttering the same two sentences: SA1 and SA2. In a mix of two female speakers with the same cadence speaking SA1, distinguishing the two speakers with human perception is almost impossible. We were curious whether the algorithm would have an equally difficult time. Comparing the results of the same sentence with that of two sentences will help address that problem.

dB level

In the real world there it's not possible to normalize the magnitude of each speaker prior to mixing. Inherently one speaker will be louder than the other. The louder party could change over the course of any recording. It's expected that the quality of separation will increase for the louder party and decrease for the quieter.

Gender

The difference between the speech of genders is well studied. Mean F0 would be around 120 Hz for men and 200 Hz for women. Also, vowel formants of female speakers tend to be located at higher frequencies. The spectral differences are due to physical differences in the vocal tract and glottis. Naturally if we are not tuning the algorithm to a particular gender, performance to be analyzed. Ideally both genders are separated equally.

The primary tool we used to do the analysis is MATLAB. We used both built-in functions and created new functions to simplify the main code. All code we used is available separately in the zip file.

Quantitative and Qualitative Evaluation:

In this section we will carefully examine our result with variant parameters. For each experiment, subject and objective evaluations are both carefully examined. Sparse Non-negative Matrix Factorization on both speaker dictionary and phoneme dictionary will be tested. We inspect four parameters, SAR, SDR [14] SNR and STOI[13]. The first three is standard separation evaluation criteria indicating the respectively Signal to Artifact Ratio, Signal to Distortion Ratio and Signal to Noise Ratio. There are two types of noise in separation, one is due to miss separation, known as interference or "cross talk", and the other is reconstruction algorithm or "artifacts". We also included Short-time Objective Intelligibility measure as an indication of our speech quality.

I. **Female** and **Male** Mixture ***(Different** Sentence, Same dB level)*

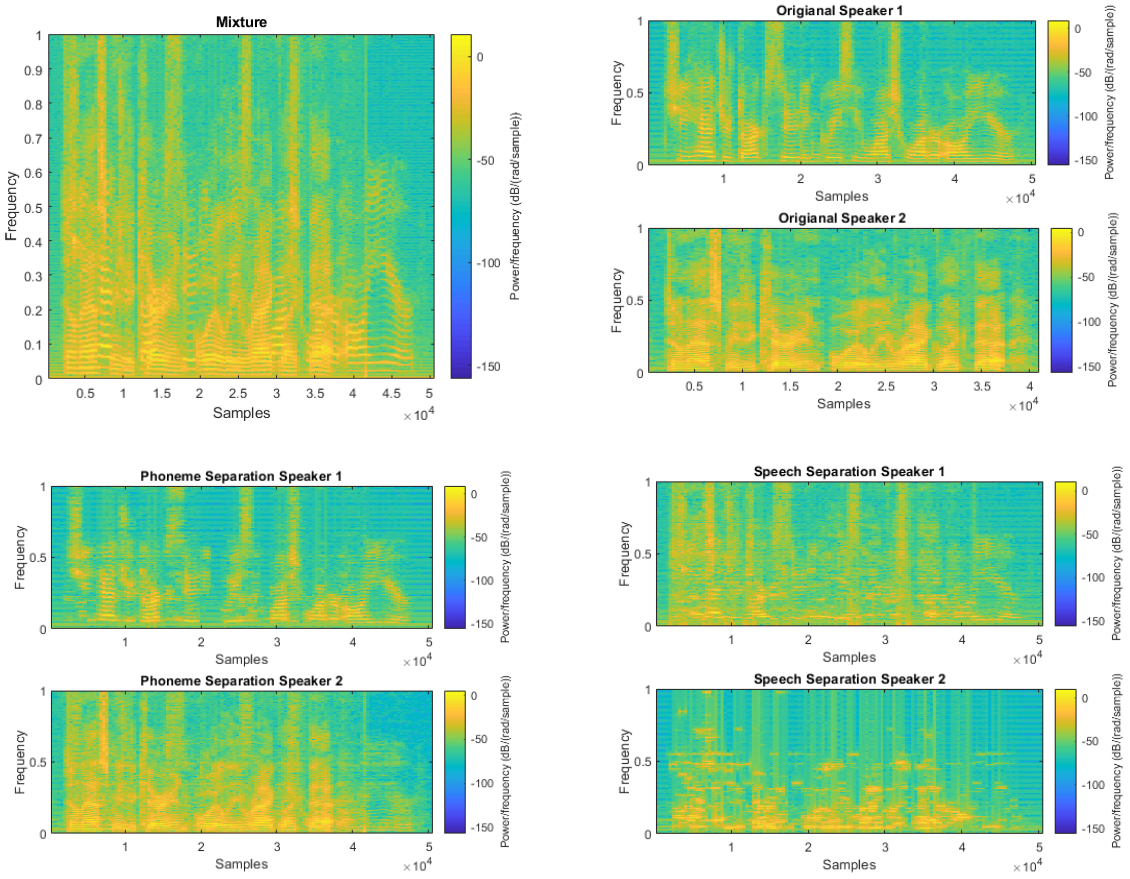
	SAR	SDR	SNR	STOI
Female Pho	10.13	10.12	6.26	0.686
Male Pho	9.93	9.81	8.66	0.817
Female Speaker	1.90	1.81	-3.18	0.531
Male Speaker	0.098	0.097	-2.625	0.522

(Pho indicates phoneme dictionary and Speaker indicates speaker dictionary)

In this experiment, we aim to separate female and male mixture utterances of different sentences. We see that phoneme dictionary separation gives all around better results than the speaker dictionary. Since TIMIT database has limited training dataset for each speaker, it is expected that the concatenation of speaker spectrogram yields unsatisfactory results. Subjectively, we can still interpret the separated speech, but there's a lot of musical noise, distortion and interference in the result. On the other hand, phoneme separation has very little crosstalk and artifacts when different sentences are spoken.

Phoneme separation methods give comparable results reported in the reference paper. Female and male separation results have similar SAR and SDR values, but male speech has higher SNR and STOI results. It is possible that female separation suffer more from low frequency range interference of the mixture, therefore has lower SNR and STOI. Subjectively, we believe that female speech is clearer, owing to the fact that human hearing system tend to favor high frequency range.

Degraded heavily by the lack of training sample, speaker dictionary reconstruction underperforms in all categories. Since the problem is due to dataset deficiency, we will not evaluate this method in the following sections. Nevertheless, quantitative results would still be provided. The speaker dictionary results are more in line with our expectation: male results are more distorted, and have much more artifacts.



(Speaker 1 is Female and Speaker 2 is Male)

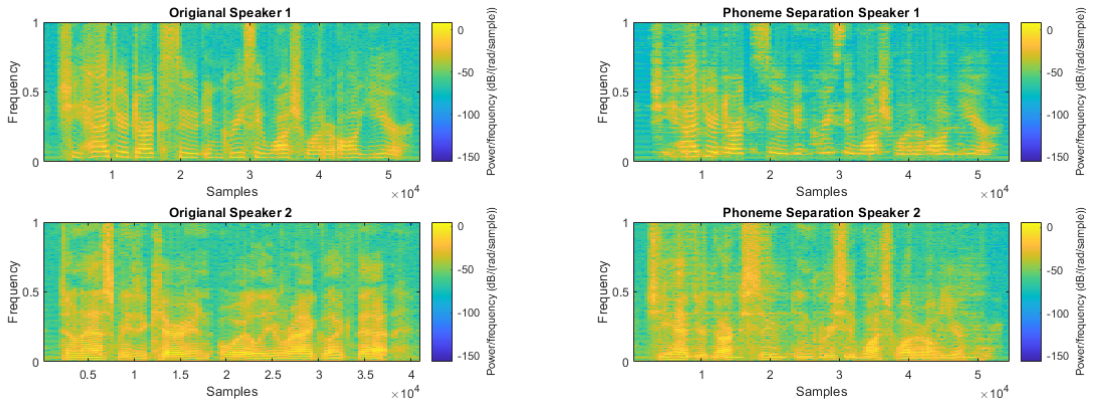
From the spectrograms, we see that phoneme separations almost perfectly reconstruct the individual spectrograms, preserving relatively the high and low frequency details without visible artifacts. Speech separation spectrograms show a lot more interference in low frequency range, and numerous ‘blocks’ are observed. These blocks contribute to the so called ‘musical noises’. The interference is also more severe for male cases for both methods, as most of the energy is registered in lower frequency range. In the following sections, we will only examine phoneme separated spectrogram and their original spectrograms.

II. **Female** and **Male** Mixture *(**Same** Sentence, Same dB level)*

	SAR	SDR	SNR	STOI
Female Pho	11.410	11.120	6.011	0.704
Male Pho	6.756	6.570	4.462	0.878
Female Speaker	1.274	1.132	-2.083	0.542
Male Speaker	0.183	0.153	-1.052	0.572

(Pho indicates phoneme dictionary and Speaker indicates speaker dictionary)

With both speakers' utterance of same sentence, we observe a drop in quality for male separation with increased artifact distortion and noise. The SNR for female speech is slightly degraded. Subjectively, we hear more cross-talking in both male and female results. Since we are using phoneme codebook, it is reasonable to expect interference in same sentence mixture. The quantitative results look promising for speaker dictionary, and intuitively, the speaker dictionary may top phoneme method in this scenario.



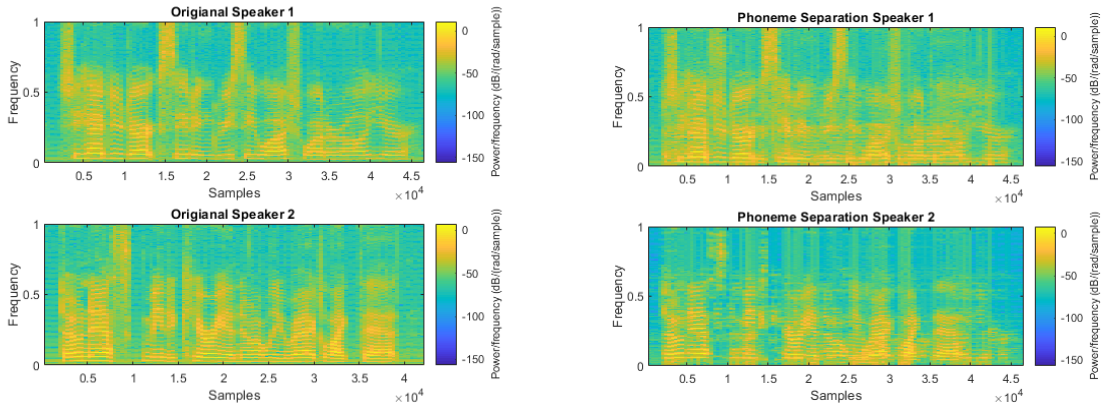
(Speaker 1 is Female and Speaker 2 is Male)

Spectrogram agrees that there is more cross talk. Note that female registers more low frequency power while male register more high frequency power, as compared to different sentence scenario. The separated speech hence will have a murmuring voice of the opposite gender in the background.

III. **Female** and **Female** Mixture *(**Different** Sentence, Same dB level)*

	SAR	SDR	SNR	STOI
Female Pho	4.097	3.868	1.655	0.685
Female Speaker	-1.075	-1.510	-1.572	0.469

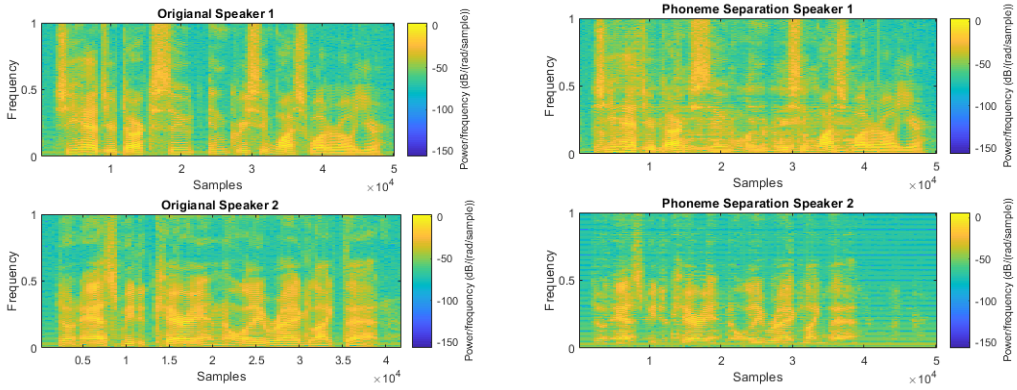
(Pho indicates phoneme dictionary and Speaker indicates speaker dictionary)



IV. **Male** and **Male** Mixture *(**Different** Sentence, Same dB level)*

	SAR	SDR	SNR	STOI
Male Pho	10.021	9.564	6.589	0.754
Male Speaker	0.041	-0.351	-1.539	0.489

(Pho indicates phoneme dictionary and Speaker indicates speaker dictionary)



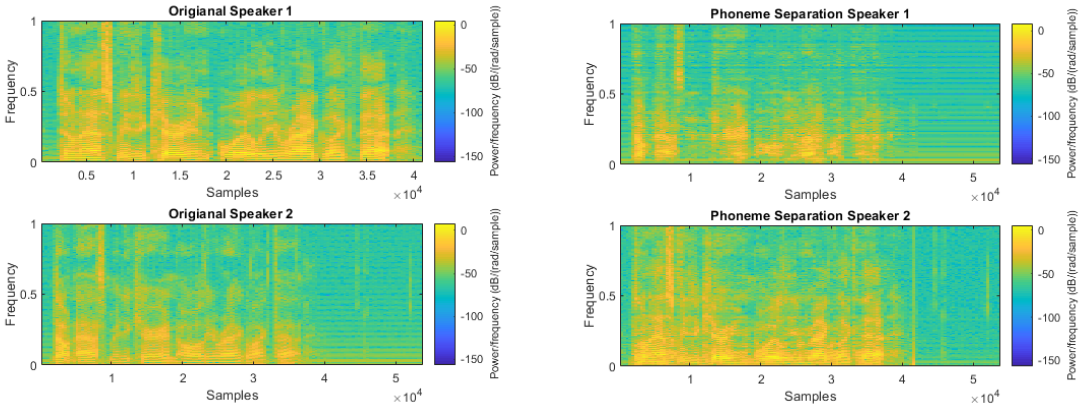
In the same gender different utterance scenario, male separations are far better than female separations. SAR, SDR, SNR and STOI parameters uniformly favors male results. Subjectively, female separations on average do appear to have more distortion and artifacts. However, we must note that the result is strongly relevant to the nature of two speakers. The female spectrograms denote a case where two speakers sound very much alike. Notice the reconstruction of female speaker 2 has much more interference from speaker 1. The male spectrograms also show that one speaker suffers more interference than the other. Note that male speaker one absorbs a lot of interference from speaker 2.

The speaker dictionary method is not promising either for both subject and objective assessments are unsatisfactory.

V. **Male** and **Male** Mixture *(**Same** Sentence, Same dB level)*

	SAR	SDR	SNR	STOI
Male Pho	8.5480	8.017	4.9532	0.7103
Male Speaker	-0.9665	-1.569	-2.6304	0.4537

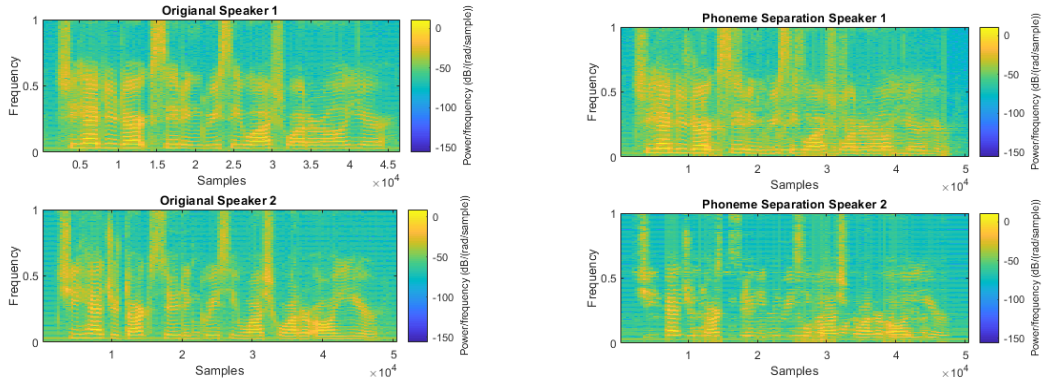
(Pho indicates phoneme dictionary and Speaker indicates speaker dictionary)



VI. **Female** and **Female** Mixture *(**Same** Sentence, Same dB level)*

	SAR	SDR	SNR	STOI
Female Pho	5.403	4.985	4.235	0.713
Female Speaker	-2.991	-3.514	-2.136	0.490

(Pho indicates phoneme dictionary and Speaker indicates speaker dictionary)



The male separation suffers small degradation but female results slightly better. In both gender cases, the STOI is higher than same gender

different sentence scenario. The reason is quite simple, since all the interference would actually help intelligibility as the utterance is the same. Examine the spectrograms, we will find that during the separation process, one speaker tends to have more features attributed to than the other. This finding is true to the last same gender experiment. Again, in this experiment, each speaker's own voice feature play a big part. If two speakers a very alike, only one separation will be of great quality. The two spectrograms above all show one separation suffering much from missing features.

VII. Varying dB level experiment (varying gender, different sentence, phoneme method)

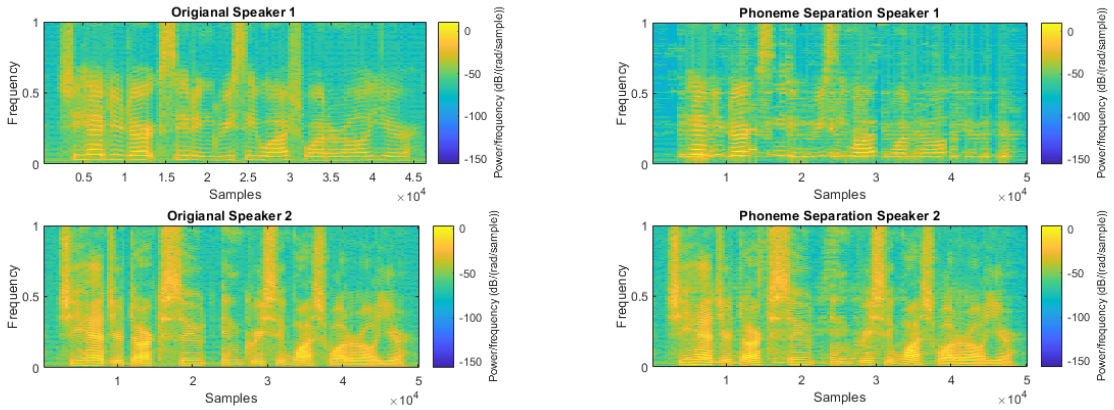
In this experiment, we test if SNMF algorithm could still perform with mixture of different sound level.

SAR	**0dB**	**3dB**	**6dB**	**9dB**
SG1	5.105	3.287	1.967	0.926
SG2	8.679	10.267	11.454	12.556
DG1	10.101	8.230	6.706	5.478
DG2	9.866	11.655	13.363	14.975

SNR	**0dB**	**3dB**	**6dB**	**9dB**
SG1	3.566	2.584	1.913	1.544
SG2	4.174	5.621	7.284	7.986
DG1	6.138	3.948	2.510	1.221
DG2	8.581	10.109	11.786	13.121

STOI	0dB	3dB	6dB	9dB
SG1	0.718	0.665	0.609	0.554
SG2	0.722	0.750	0.776	0.797
DG1	0.686	0.649	0.607	0.560
DG2	0.817	0.849	0.875	0.896

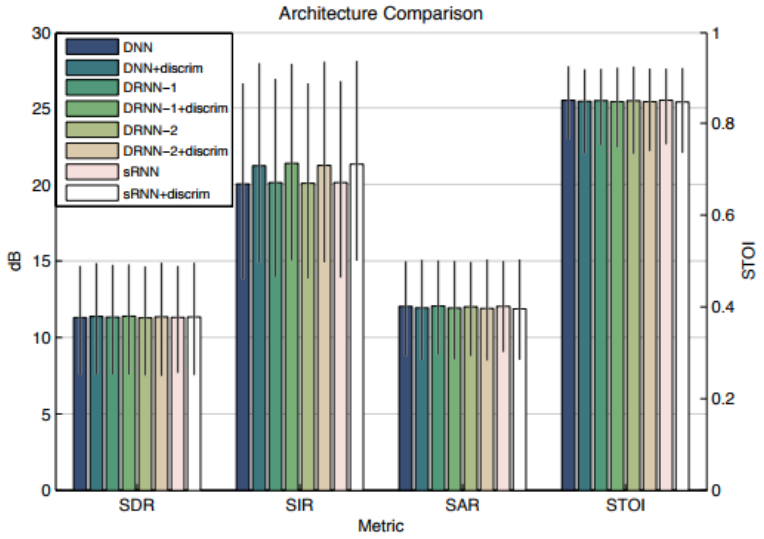
(SG: same gender, DG: different gender, dB indicates relative sound difference, the number indicates speaker identity)



9 dB levels apart

Above we compared original speaker spectrogram with separated spectrogram side by side. When varying the dB level, we suppress the first speaker and magnify the second speaker. The quantitative results align with intuition: the higher the volume in the mixture, the better the results; vice versa. A careful inspection of the spectrograms show that the magnified speaker spectrogram is nearly perfectly reconstructed, whereas the downgraded spectrogram suffers severe losses as well as distortion. Subjectively, speaker one reconstruction is overwhelmed by noise as indicated by SAR and SNR, but intelligibility is still quite good. Speaker two reconstruction is not perfect with small interference in the background. But the cross talk is then again overwhelmed by volume.

VIII. A brief look at DNN evaluations:



We included a graph from paper [2] as a comparison with SNMF approach. Note that our phoneme dictionary approach has comparable or even better values than the report above. However, weight update would take longer than a network excitation. Training time however would favor SNMF approach. Nevertheless, combining the knowledge in traditional signal processing with DNN, we are able to obtain strong results.

Conclusion

Considering the size of our training data compared to the training data of the source paper, it is fair to classify our results as “good”, especially the phoneme dictionary approach.

One difficulty in the process of this work was trying to interpret Schmidt’s paper where it claims, “Then, we reconstructed the individual magnitude spectra of the two speakers and mapped them from the Mel frequency domain into the linear frequency STFT domain.” The Mel frequency spectrum is a transformation of the linear frequency domain where vectors are passed through a bank of unevenly spaced triangular filters. The Mel frequency domain is relevant because it amplifies frequencies that have been shown to have better human perception. Analysis done with Mel frequency elements such as MFCCs often outperform methods not in the Mel scale. Unfortunately, we were not able to find a robust way to invert the spectrogram of the Mel frequency back into the linear frequency scale. It would effectively be undoing the filter bank operation, where since elements are summed information is lost. Schmidt has told us, “The transformation between linear and Mel-frequency can be implemented by a matrix multiplication, which can easily be "inverted" in a number of ways. The particular method is not of much importance in my experience.” Still, we resigned to do the analysis using only the STFT. It is expected our results would improve if we were to try this in the future.

Through our experiments, we find that phoneme dictionary approach is slower than speaker dictionary, but with much better results. We believe that the phoneme concatenation is already a good basis for the SNMF, or at least a good initialization to the algorithm. Speaker dictionary however, relies on the amount of data and number of basis to extrapolate meaningful features.

Surprisingly, the SNMF approach is comparable in quantitative and qualitative evaluations with the DNN counterparts. DNN still has the slight advantage, and we believe that a good understanding from the digital signal processing aspect would greatly benefit when designing a neural network for the same task.

References

[1] Schmidt, Mikkel N. and Rasmus Kongsgaard Olsson. “Single-channel speech separation using sparse non-negative matrix factorization.” *INTERSPEECH* (2006). http://www2.imm.dtu.dk/pubdb/views/edoc_download.php/4511/pdf/imm4511.pdf

[2] Huang, Po-Sen, et al. "Deep learning for monaural speech separation." *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014. https://posenhuang.github.io/papers/DNN_Separation_ICASSP2014.pdf

[3] E. M. Grais and H. Erdogan, "Single channel speech music separation using nonnegative matrix factorization and spectral masks," *2011 17th International Conference on Digital Signal Processing (DSP)*, Corfu, 2011, pp. 1-6. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6004924&isnumber=6004864

[4] Simpson, Andrew JR. "Time-Frequency Trade-offs for Audio Source Separation with Binary Masks." *arXiv preprint arXiv:1504.07372* (2015). https://arxiv.org/ftp/arxiv/papers/1504/1504.07372.pdf

[5] Pépiot, Erwan. "Voice, speech and gender: male-female acoustic differences and cross-language variation in english and french speakers." *Corela. Cognition, représentation, langage* HS-16 (2015).

https://halshs.archives-ouvertes.fr/halshs-00764811/document

[6] https://sites.google.com/site/deeplearningsourceseparation/

[7] M. Stark, M. Wohlmayr and F. Pernkopf, "Source–Filter-Based Single-Channel Speech Separation Using Pitch Information," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 2, pp. 242-255, Feb. 2011. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5443705&isnumber=5609232

[8] Radfar, M.H., Dansereau, R.M. & Sayadiyan, A. J AUDIO SPEECH MUSIC PROC. (2006) 2007: 084186. https://doi.org/10.1155/2007/84186

[9] B. Raj and P. Smaragdis, "Latent variable decomposition of spectrograms for single channel speaker separation," IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005., New Paltz, NY, 2005, pp. 17-20.

http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1540157&isnumber=32894

[10] A. P. Varga and R. K. Moore, "Hidden Markov model decomposition of speech and noise," International Conference on Acoustics, Speech, and Signal Processing, Albuquerque, NM, USA, 1990, pp. 845-848 vol.2.
http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=115970&isnumber=3385

[11] M. Kolbœk, Z. Tan and J. Jensen, "Speech enhancement using Long Short-Term Memory based recurrent Neural Networks for noise robust Speaker Verification," 2016 IEEE Spoken Language Technology Workshop (SLT), San Diego, CA, 2016, pp. 305-311.
http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7846281&isnumber=7846230

[12] C. H. Taal, R. C. Hendriks, R. Heusdens and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, TX, 2010, pp. 4214-4217.
http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5495701&isnumber=5494886

[13] Emmanuel Vincent, Shoko Araki, Pau Bofill. The 2008 Signal Separation Evaluation Campaign: A community-based approach to large-scale evaluation. 8th Int. Conf. on Independent Component Analysis and Signal Separation (ICA), Mar 2009, Paraty, Brazil. Pp.734--741.
https://hal.inria.fr/inria-00544168/en