Towards Unsupervised Auditory Scene Analysis: Blind Source Separation using Non-Negative Matrix Factorization Methods

Christina Mai

Sharath Chandra Ramakrishnan

Thouseef Syed

Applied Cognition and Neuroscience Program

University of Texas at Dallas

Abstract

The world is full of noise, and the auditory environment is rarely simple, clean, and from clearly distinct sound sources. The cocktail party problem highlights this as a blind source separation problem where a scene can be separate concurrent speech signals drowned in a word of many sounds. While the brain can handle such a task fairly well, finding the computational equivalent requires testing many algorithms. This paper works specifically as an exploration of the GCC-NMF algorithm which combines negative matrix factorization (NMF) for unsupervised dictionary learning and generalized cross correlation (GCC) for spatial localization. Our work extends the algorithm by applying more source data, especially ones that are not necessarily speech signals. Results were similar to past work and shows high efficacy and accuracy across various sources. Eventually the goal is to apply classification after the separation with the goal to find patterns. Being able to compare the algorithm via factors such as target fidelity and lack of sound artifacts and ultimately pinpointing its strengths should lead to a better understanding of blind source separation. Higher efficacy for separation will benefit basic audio processing to work on assistive hearing devices and speech recognition systems.

Keywords: blind speech separation, auditory scene analysis, GCC, NMF, unsupervised learning

Towards Unsupervised Auditory Scene Analysis: Blind Source Separation using Non-Negative Matrix Factorization Methods

The human brain has the power to separate sounds and then categorize the various sound inputs accordingly. In other words, humans perform 'auditory scene analysis' by separating and organizing streams of sounds. The sound separation process has been considerably difficult to replicate with computers due to the complexity of sound and how sound waves interfere and mix with each other. It has usually been the goal of scientists and engineers to be able to separate different speech signals, as a human can do well and most notably through the 'cocktail party effect'. This effect takes into account environments where there are multiple speech components. Additionally, typical speech signals are also degraded by noisy environments or background sounds. The combination of noisy, degraded signals and multiple speech sources cause auditory scenes to be difficult to parse. The goal of the work here is to not only separate sound streams, but to be able to categorize it for a full understanding of the auditory scene. Understanding the gender of the speaker and noise categorizations such as music instruments is informative, yet this is just one step towards this goal. Initial stages with separation algorithms such as this can eventually, for example, lead to insights on characteristics of gendered speech. Proper separation also supports assistive hearing devices such as cochlear implants or hearing aids. Another area with many applications is with speech recognition systems.

The task at hand is blind speech separation which ideally assumes as little as possible about the mixed sources in the signal. The chosen algorithm to study is generalized cross-correlation (GCC) for spatial localization combined with non-negative matrix factorization (NMF) to learn events. Both will be expanded on in the methods section. Of course, there have been other supervised and semi-supervised methods to achieve this task. Studies from Schmidt

and Olsson (2006) uses a sparse NMF algorithm to separate multiple speech sources recording from a single microphone. While they used this learning on a personalized speech corpus dictionary, they improved performance by pre-processing the training data by segmenting it on a phoneme level. Although this process works well, it becomes too cumbersome to manage with an increased number of sources and variabilities. Other common supervised approaches have used isolated source recordings to adapt into source-specific dictionaries. These source specific dictionaries are then concatenated to encode mixture signals, and in the process of encoding, source separation is achieved as each source is encoded by its corresponding dictionary. Other methods involved specifying the nature of the sources in the mixture signal, thereby constraining the NMF dictionary to correspond to the sources of interest (Ozerov, Vincent, & Bimbot ,2012). Various other algorithms that also use NMF such as FASST (Salaün et al., 2014) were compared in previous literature, though overall did not perform as well.

The purpose of the project is to extend the applications of the GCC-NMF and further understand the accuracy and applicability of the separated data. We apply analysis to a musical mixture and compare to a base speech mixture for comparison. Once properly separated and understood, the sound components and their corresponding properties can contribute to the understanding of auditory scene analysis.

Method

The method being used is an unsupervised method that needs no prior information about the mixture signal. A common way to use unsupervised methods for source segregations has been to use spatially distributed microphones, and combing NMF along with spatial separation to perform source separation. This is achieved by learning a set of source specific dictionaries, to develop a set of mixing models that are represented by spatial covariance matrices; However, this process of using a spatial covariance matrix depends on creating a constrained dictionary and is also sensitive to initialization parameters, and in fact makes the process semi-supervised (Adiloglu & Vincent, 2012).

The method that we are using overcomes the previous problems encountered by combining spatial covariance with NMF and group dictionary components based on their spatial origin in an unsupervised manner (Wood & Rouat, 2016). We follow the methods used by Wood and Rouat, by using a generalized cross correlation source localization method, along with an NMF dictionary learned on a mixture signal, resulting in individual atoms being localized over time, and grouped according to their spatial origin.

Procedures

Blind Speech Separation.

The speech separation is the main task, and the code for the GCC-NMF algorithm is available on GitHub. Open source demonstrations are available through iPython Notebooks and this has allowed for us to practice with the data and observe how the data is obtained by code. The continuation of this method with be used on the data described in the following ‘Data Set’ section.

We start by inputting the mixture signal.

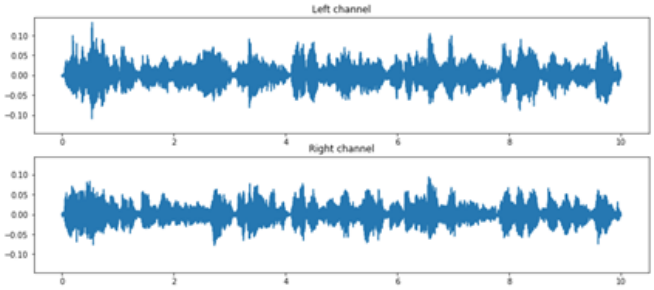


Figure 1. Live recording of 3 female voices was taken at reverberation of 130ms (Wood & Rouat, 2016).

A spectrogram of the source is plotted.

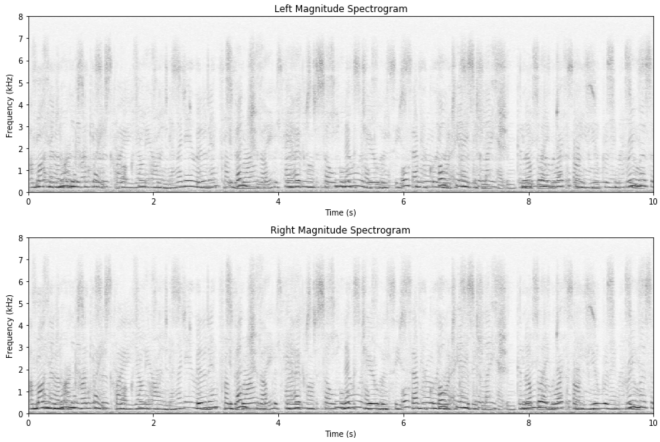


Figure 2. Left and right magnitude of the input mixture signal.

Then source localization is performed with GCC.

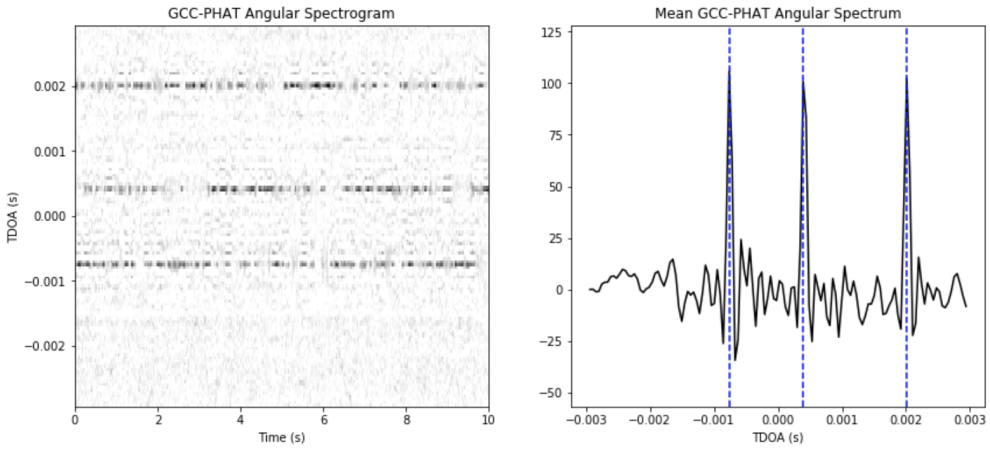


Figure 3. GCC-PHAT Angular Spectrogram and the mean spectrum is taken for localization.

Using non-negative matrix factorization, unsupervised learning is performed to obtain the dictionary atoms.

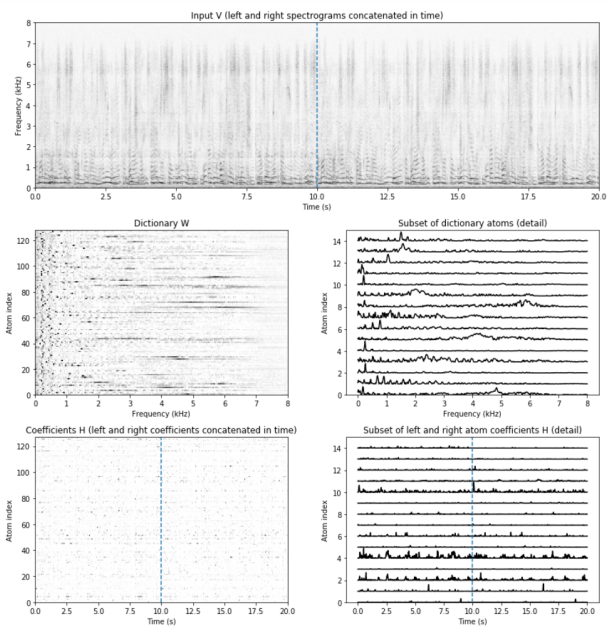


Figure 4. Female voice spectrogram factorized.

Dictionary atoms are masked with respect to coefficients for the individual sources.

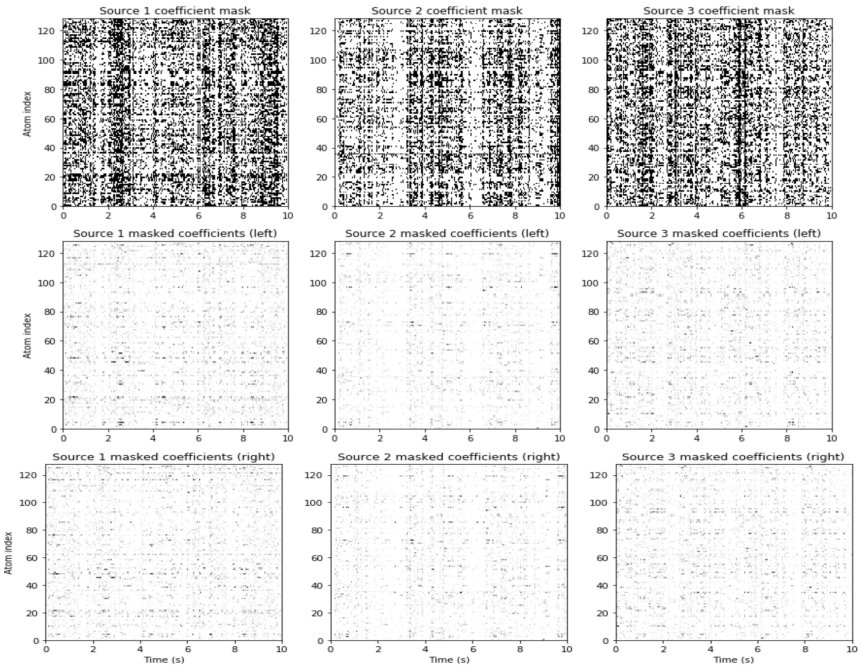
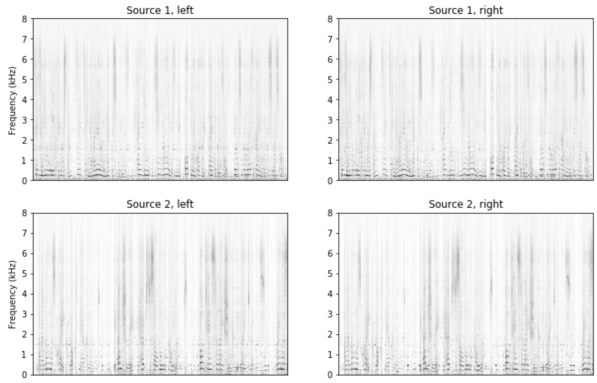


Figure 5. GCC-NMF target coefficient masks

From there the source spectrogram estimates can be constructed.



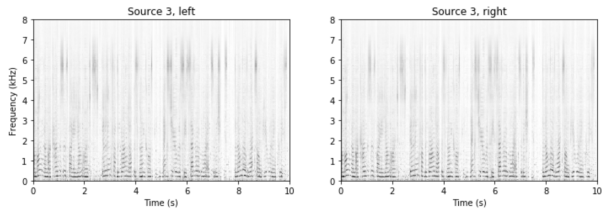


Figure 6. Estimated spectrogram with left and right channels of all sources.

Time domain target signal estimates can also be reconstructed.

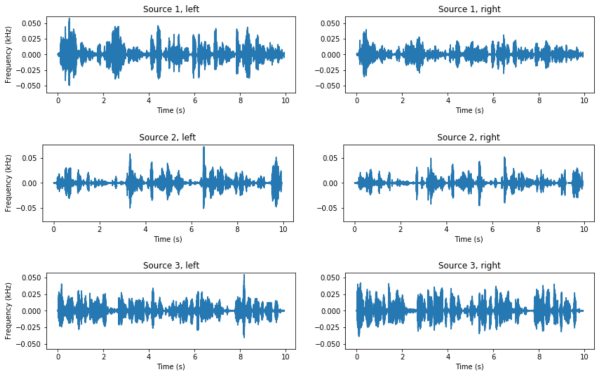


Figure 7. Estimated time domain signals left and right channels.

Classification.

One of our original goals was to be able to classify the data as a measure of the accuracy of the separation; however, the data set we used only created mixtures of sounds from the same type of source or were limited. This means we were unable to classify the data as planned. Some possibilities we considered initially were linear regression or support vector machines (SVM). This could be used as a future direction.

Data Set

The implementation of blind speech separation technique by GCC-NMF is applied to 2015 SiSEC dataset of underdetermined speech and music mixture files. It consists of live recordings of female speech, male speech and musical instruments. We decided to implement the

algorithm on two mixed signals. One was a mixed signal consisting of three simultaneous speaking voices, and the other was a mixture of a music track containing three musical sources where one was vocal, one drums, and one guitar. The recordings were obtained by two omnidirectional microphones strategically placed in a room of dimension 4.45 x 3.55 x 2.5 m where the distance between microphones ranged from five centimeters to one meter. The individual sources were placed in the arrival range from -60 degree to 60 degree with at least 15 degrees of spacing between them. Also, another parameter that plays a vital role in the recording of the source is reverberation. It can be defined as the persistence of sound after the sound has stopped playing. Typically, reverberation would vary from a control room to a open hall concert. The reverberation would be as low as 130ms in a control room and high as 1.44s in a concert hall. Hence, it is important to place the sources in the respective locations for obtaining optimized results.

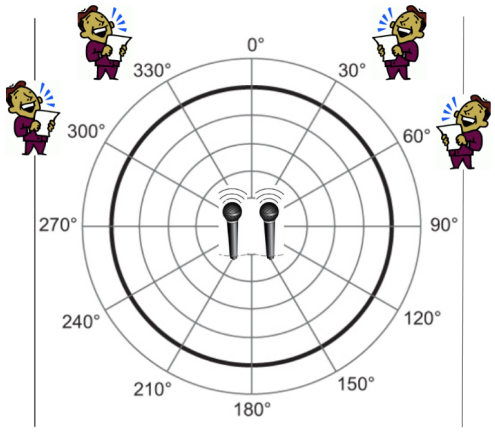


Figure 8. Two omnidirectional microphones receiving input from 4 sources.

The algorithm has been applied to the dev1 sub collection within 2016 SiSEC data, and it would be optimal for us to reuse this data for further investigation. The usage of other datasets

was considered but ultimately not implemented due to time constraints.

Algorithms

Nonnegative Matrix Factorization.

Non negative matrix factorization, also known as NMF, is a method by which a large dataset is decomposed to two relatively smaller matrices that are easier to work with. A matrix V of dimension ***f x t*** where each element is ***Vft*** ≥ □ (where f is the indexing frequency and t is the time) . NMF decomposes it into two matrices ***W*** and ***H*** of dimensions ***f x d*** and ***d x t*** respectively, where each element ***wft***≥ □and ***hft***≥ □and ***d<min(f,t)*** such that :

$$W*H=V$$

We assume ***V*** is a large dataset where each column is an observation and each row is a feature. In machine learning, it is often necessary to reduce the feature space of dataset for ease of computation. In our case, NMF behaves as an unsupervised algorithm that is essentially used for the source separation. The d columns represent the dictionary atoms which are the non-negative functions of frequency that are combined linearly with the corresponding coefficients at each point in time, in order to reconstruct the input spectrogram.

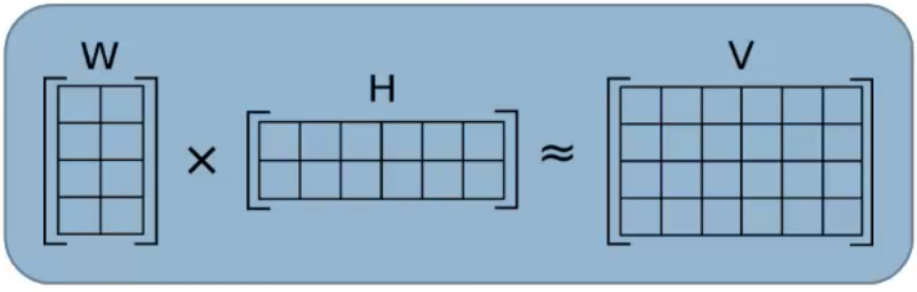


Figure 9. Matrix factorization. Templates (W) multiplied with activation (H).

Generalized Cross Correlation.

The generalized cross-correlation also known as GCC is used for source localization. The time difference of arrival (TDoA) is progressively computed during the recording of the channel in question with respect to the reference channel. In order to compute the TDOA between the reference channel and any other channel for any given segment it is usual to estimate it as the delay that causes the cross-correlation between the two signals segments to be a maximum. This method is definitely used for estimating TDoAs for a range of frequencies. It represents an angular spectrogram which is a function of time-delay τ and time t, as follows:

$$G_{\tau t} = \sum_{f} \psi_{ft} \mathrm{V}_{lft} \mathrm{V}^{*}_{rft} e^{j2\pi f\tau}$$

Where **Vlft** and **Vrft** are the left and right complex spectrograms , and **ψft** is a time-varying frequency-weighting function. In order to improve robustness against reverberation it is a normal practice to use the Generalized Cross Correlation with Phase Transform (GCC-PHAT). Hence, it can be depicted as:

$$\mathrm{G}^{\mathrm{PHAT}}_{\tau t} = \sum_{f} \frac{\mathrm{V}_{lft} \mathrm{V}^{*}_{rft}}{|\mathrm{V}_{lft}|\,|\mathrm{V}_{rft}|} e^{j2\pi f\tau}$$

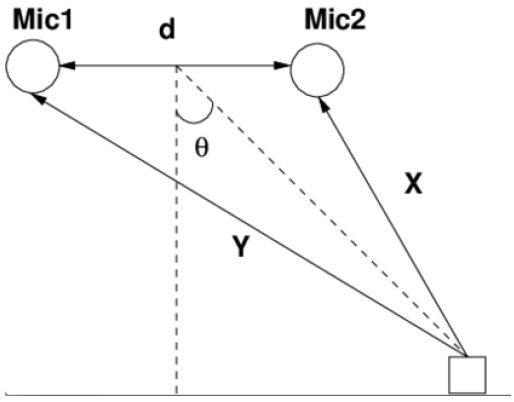


Figure 10. Source localisation using Time Difference of Arrivals (Umbarkar, Subramanian, & Doboli, 2010, p. 2).

Measures

An open source toolkits is used to help in benchmarking the performance of blind signal separation. We used the PEASS toolbox (Perceptual Evaluation for Audio Source Separation) which relies on subjective perceptual measures obtained from listening tests on the evaluation of audio source separation.

The measures of our system are based on the how each estimated source signal is decomposed into a number of features that contribute to the overall perceptual evaluation namely target source and interference from unwanted sources and artifacts (for instance, noise from continuous signal sources like music). With the PEASS toolbox, the scores better correlate with human assessments, as it is correlated with subjective ratings. Wood et al. found that varying some parameters of NMF showed no significant performance improvement while observing the S/N ratios but correlating to a model that captured subjective assessments did reveal effects to parametric variation of NMF.

Similar to BBSEval, another toolbox that was used by Wood et al., the distortion signal is decomposed into three components, namely target distortion, interference and artifacts. From these components, PEASS toolbox computes three quality scores, namely Overall Perceptual Score (OPS), Target-related Perceptual Score (TPS), Interference-related Perceptual Score (IPS). Of all these the IPS is the most important guiding factor to determine the efficacy of our source segregation algorithm.

We measured the effects of the performance by varying three parameters of the non-negative matrix factorization.

Dictionary Size.

This is the number of atoms that the factorization takes into account. Our expected result was that increasing the size of the dictionary should lead to to an increase in perceptual score in blind speech separation across target fidelity, and loss of artifacts. We had expected this improvement to saturate for larger dictionary sizes.

Number of Iterations.

We had expected the number of iterations to have a similar trend as the effect of increase in dictionary size has on the overall performance indicators, and also expected a saturation at higher values.

Sparsity Coefficient (Alpha).

We considered if explicitly including the notion of sparseness improve the performance of the NMF. Wood et al. found that increasing sparseness decreased the overall performance while at the same time increased interference suppression. Since the non-negativity constraints of NMF make the representation purely additive (no subtractive elements) compared to other linear transformations methods like PCA (Principal Component Analysis) or ICA (Independent Component Analysis), sparseness can be a very important variable.

Iter	Iterations
Dic	Dictionary Atoms
Alpha	Sparsity
OPS	Overall Perceptual Score
TPS	Targetted Perceptual Score
IPS	Interference Perceptual Score

Dic	Iter	Alpha	OPS	TPS	IPS
128	100	0.05	20	41	85
512	100	0.05	22	37	83
1024	100	0.05	20	39	82
Iter	**Dic**	**Alpha**	**OPS**	**TPS**	**IPS**
50	128	0.05	21	42	84
100	128	0.05	20	41	85
200	128	0.05	19	41	85
Alpha	**Dic**	**Iter**	**OPS**	**TPS**	**IPS**
0.1	128	100	21	43	83
0.3	128	100	21	42	83
0.5	128	100	20	41	83
0.9	128	100	23	45	84

Figure 11. Table of values of quality scores.

Dictionary Size

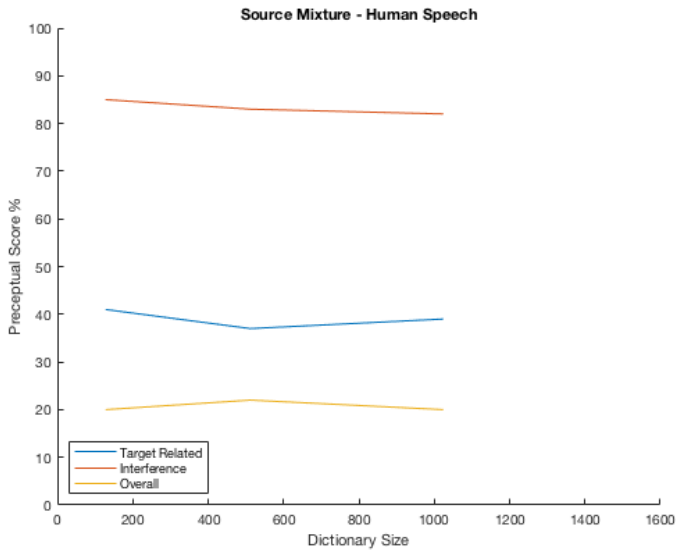


Figure 12. Varying dictionary size for speech.

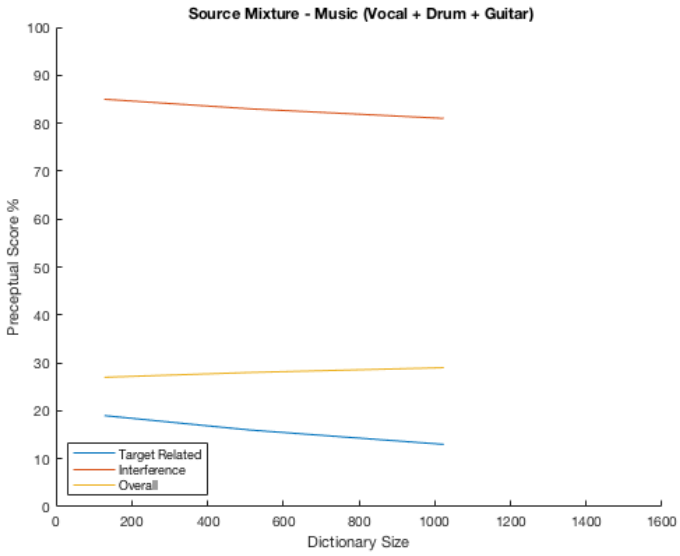


Figure 13. Varying dictionary size for music.

Number of Iterations

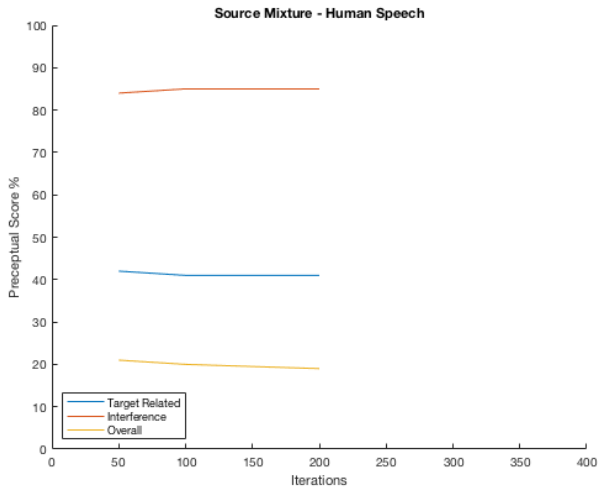


Figure 14. Varying iteration size for speech.

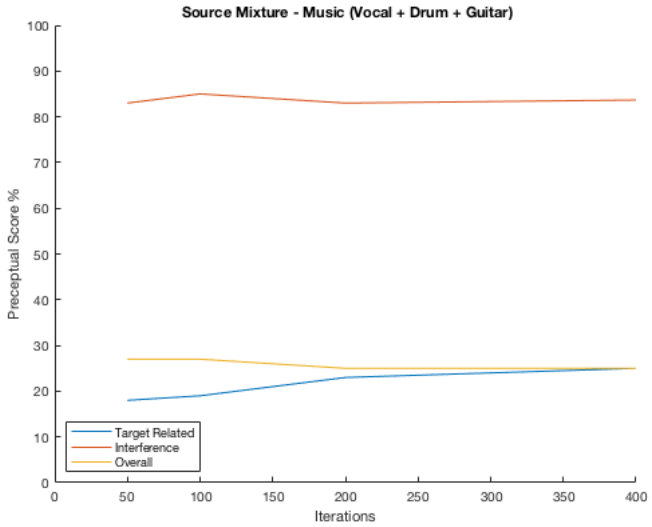


Figure 15. Varying iteration size for music.

Sparsity Coefficient (Alpha)

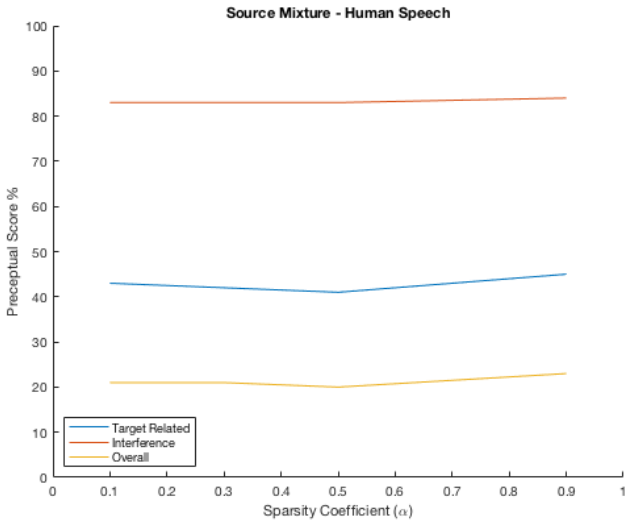


Figure 16. Varying sparsity coefficient for speech.

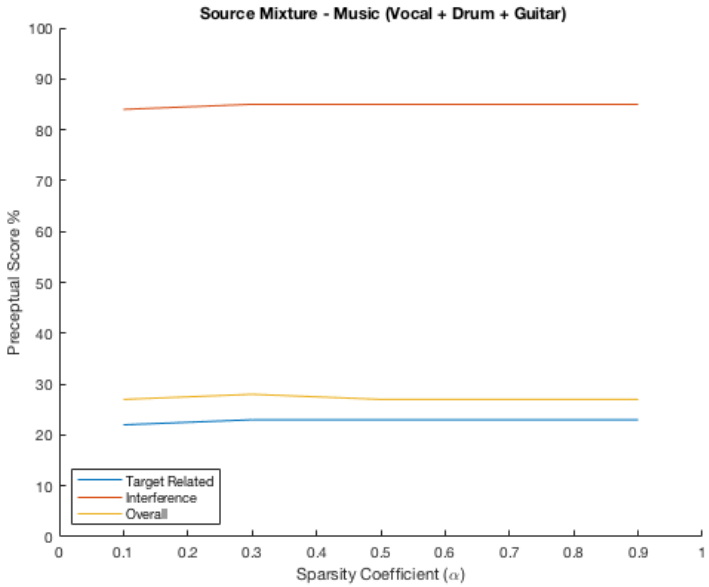


Figure 17. Varying sparsity coefficient for music.

Discussions

We see that in our results (Fig 12 to Fig 17), the Overall Perceptual Score of 20 to 40% maybe considered a normalized baseline (compared to human subjects) who rate the perfect

distortion less sounds as an 100% clear source. Noisy artifacts are often a feature of source segregation in noisy environments, and while they still maybe intelligible, might have a got a low perceptual score according to the quality constraints imposed by the PEASS Toolbox.

We can see increase in Dictionary size resulted in an increase in overall perceptual score until about 450 units in the case of the Speech Mixture. The general trend of the performance on Interference decreased gradually as number of dictionary units in increased in both the Speech mixture as well as the Music Mixture. The number of iterations in both cases was set to the default of 100 iterations with a sparsity alpha value of 0.05.

While changing the number of iterations and keeping dictionary size fixed at 128 and sparsity coefficient alpha at 0.05, we see an gradual increase the interference performance until about 100 iterations after which the benefit of increasing the number of iterations leads to saturation in the interference performance. In general, the gradual increase the performance measure of interference leads to corresponding decrease in perceptual measure for the target related perceptual score. Perhaps at some point more iterations at this fixed dictionary size of 128 resulted in some cross-source Target related artifacts that reduced the perceptual score although interference was being minimized.

The effect of varying sparsity coefficient while keeping the number of Dictionary units fixed at 128 and the number of Iterations at 100, resulted in an overall increase in interference suppression that was more prominent in the source separation of music. While Wood et al reported significant decrease in Overall Perceptual and Target Related Perceptual score, in our results we saw an overall saturation in those scores after the alpha value exceeded 0.5. Ideally keeping the alpha values low (0.2) with the number of iterations at about 150 and the dictionary

size in between 300 to 450 is an optimum setting for computational as well as higher perceptual performance of blind separation of mixture signals consisting of speech and music sources.

One major limitation with the separation algorithm is that is fairly dependent on the sound sources being locally and spatially separated by a distinguishable factor. There were no tests on sources that are near each other or have overlaps between speakers and the recording microphones. And as various types of sound sources are mixed, though it is able to do a decent level of separation, the algorithm simply does not perform as well.

By the nature of human perception of music, PEASS also was not able to rate mixtures with musical instruments the same way as it would for mixtures of human speech. We experimented on various sources including general popular music found via YouTube, but did not perform analysis on them since we did not have the original source files or their localizations for comparison.

Our mixtures also did not incorporate or consider general background noise. It might be beneficial to consider additional algorithms or certain levels of preprocessing in order to remove extra noises that are not part of the sources needing to be separated. Unwanted background noises are typical of most environments and for full application must be considered. Future work would need to take that into account poor environments with various forms of audio degradation. Therefore, we can conclude that separating speech in a noisy environment is quite challenging and cumbersome when compared to a control room which is relatively pleasant.

References

Adiloglu, K., & Vincent, E. (2012) Variational bayesian inference for source separation and robust feature extraction. Ph.D. dissertation, INRIA.

Ozerov, A.,Vincent, E., & Bimbot F. (2012). A general flexible framework for the handling of prior information in audio source separation. *Audio, Speech, and Language Processing*, IEEE , 20(4), 1118–1133.

Salaün, Y., Vincent, E., Bertin, N., Souviraà-Labastie, N., & Jaureguiberry, X. (2014). The Flexible Audio Source Separation Toolbox Version 2.0. *ICASSP*.

Schmidt, M. N., & Olsson, R. K. (2006). Single-channel speech separation using sparse non-negative matrix factorization. *Spoken Language Processing*. ISCA International Conference on INTERSPEECH.

Umbarkar, A., Subramanian, V., & Doboli, A. (2010). Improved sound-based localization through a network of reconfigurable mixed-signal nodes. In *Robotic and Sensors Environments (ROSE), 2010 IEEE International Workshop on* (pp. 1–6). IEEE Publishing. https://doi.org/10.1109/ROSE.2010.5675318

Vincent, E., Araki, S., Theis, F. J., Nolte, G., Bofill, P., Sawada, H., Ozerov, A., Gowreesunker, B. V., Lutter, D., and Duong, N. Q. K. (2012). The Signal Separation Evaluation Campaign (2007-2010): Achievements and remaining challenges. *Signal Processing*, 92, 1928-1936.

Vincent, E., Gribonval, R., & Févotte, C. (2006). Performance measurement in blind audio source separation. *IEEE Trans. Audio, Speech and Language Processing*, 14(4), 1462-1469.

Wood, S. (2018). *seanwood/gcc-nmf*. [online] GitHub. Available at: https://github.com/seanwood/gcc-nmf [Accessed 5 Oct. 2018].

Wood, S., Rouat, J., Dupont, S., & Pironkov, G. (2016). Blind Speech Separation and Enhancement With GCC-NMF. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), 25(4), 745–755. https://doi.org/10.1109/TASLP.2017.2656805