Non-negative Tensor Factorization of Modulation Spectrograms for Monaural Sound Source Separation

Team members:

Avijit Jaiswal, Mohit Kumar, Kumari Rashmi Bala, Shahid K I and Chaitanya Ahuja

Indian Institute of Technology, Kanpur

Email: avijit, krmohit, rbala, shahidee, chahuja@iitk.ac.in

Introduction

- Sound Separation is used in Speech enhancement, recognition and manipulation.
- For sound separation we are using Non Negative matrix factorization(NMF).
- NMF techniques decompose a mixture signal into a sum of components having having fixed spectrum and time varying gain.

ALGORITHM

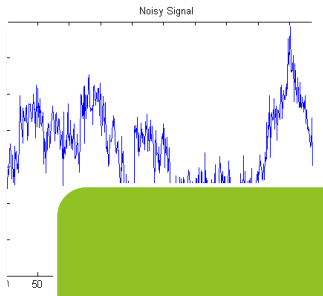
Modulation spectrogram feature representation.

Tensor factorization model.

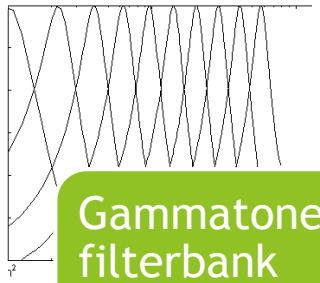
Synthesis of components from factorized tensors.

Simulation experiments.

Modulation Spectrogram Feature Representation

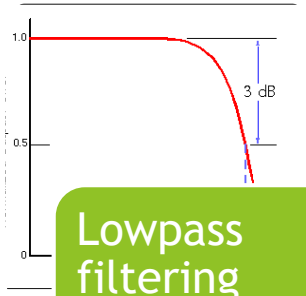


Mixture signal



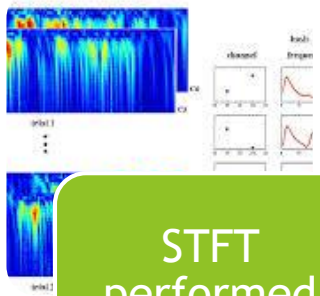
Gammatone filterbank

- Multiple channel output.



Lowpass filtering

- Filtering corresponding to every channel output



STFT performed

- Tensor with Size R*N*M

Tensor factorization model

- We model χ as a sum of K components.
- Each component is modelled as a product of three factors G, A and S, each of which characterizes one of the tensor dimensions.

$$\chi_{r,n,m} \approx \hat{\chi}_{r,n,m} = \sum_{k=1}^{K} G_{r,k} A_{n,k} S_{m,k} \quad (1)$$

- The model parameters are estimated by minimizing the generalized Kullback-Leibler (KL) divergence D

$$D(\chi \| \hat{\chi}) = \Sigma_{r,n,m} \chi_{r,n,m} \log \frac{\chi_{r,n,m}}{\hat{\chi}_{r,n,m}} - \chi_{r,n,m} + \hat{\chi}_{r,n,m} \quad (2)$$

- Iterative update equations for minimizing the KL divergence and initializing G,A and S to nonnegative values ensures non-negativity throughout updates.
- The update equations use the definition of C = χ/χ^.

$$G_{r,k} \leftarrow G_{r,k} \frac{\sum_{n,m} C_{r,n,m} A_{n,k} S_{m,k}}{\sum_{n,m} A_{n,k} S_{m,k}}$$

$$A_{n,k} \leftarrow A_{n,k} \frac{\sum_{r,m} C_{r,n,m} G_{r,k} S_{m,k}}{\sum_{r,m} G_{r,k} S_{m,k}}$$

$$S_{m,k} \leftarrow S_{m,k} \frac{\sum_{r,n} C_{r,n,m} G_{r,k} A_{n,k}}{\sum_{r,n} G_{r,k} A_{n,k}}$$

- C is reevaluated between each update of G, A and S.
- The total number of entries in the factor matrices G,A and S is K(M+R+N).

Synthesis of components from factorized Tensors:

- After separation in the modulation envelope domain, reconstruction is carried out in a Wiener filtering-like reconstruction approach.
- A component synthesis tensor, γ is generated by taking the STFT of the output of each auditory filterbank channel when filtering the original mixture signal

$$|\hat{\gamma}|_{r,p,m} = \sum_{k=1}^{K} G_{r,k} \; B_{p,k} \; S_{m,k}$$

- Defining $\mathcal{E} = |\gamma|/|\hat{\gamma}|$ allows repeated application of update rule:

$$B_{p,k} \leftarrow B_{p,k} \frac{\sum_{r,m} \mathcal{E}_{r,p,m} G_{r,k} S_{m,k}}{\sum_{r,m} G_{r,k} S_{m,k}}$$

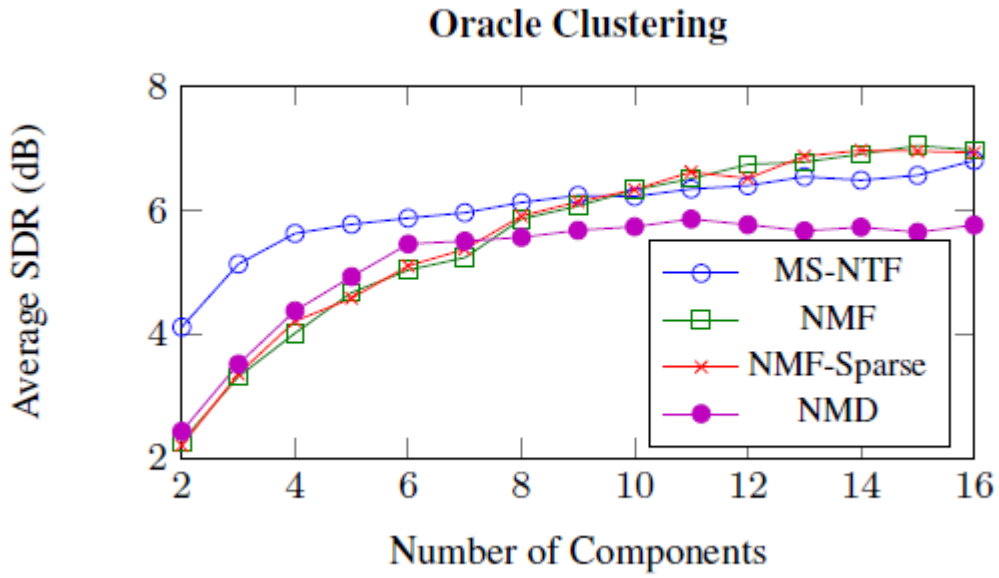
- The Wiener filter formed from G, B and S is applied to γ, to produce K separated components, $\hat{\gamma_k}$ in the STFT domain thus:

$$\hat{\gamma}^{k}_{r,p,m} = \gamma_{r,p,m} \frac{G_{r,k} B_{p,k} S_{m,k}}{\sum_{k'} G_{r,k'} B_{p,k'} S_{m,k'}}$$

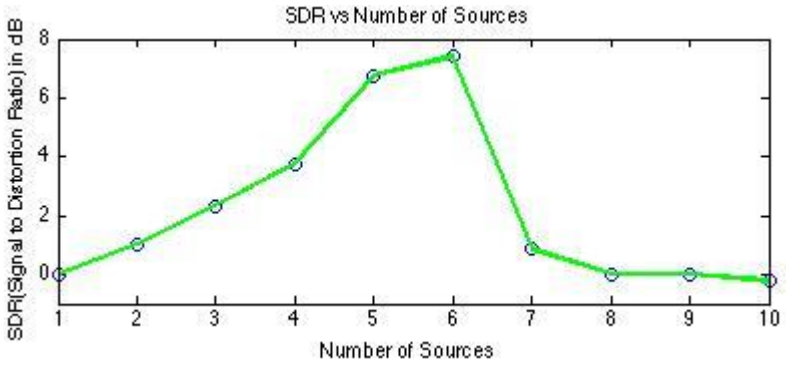
- Conversion of each of the K sets of STFTs back to the time domain frames is performed by the inverse DFT of the p dimension.

Simulation experiments

- We have used the oracle clustering of components to simulate the performance of our algorithm.
- Under this clustering approach every separated component is compared against the original components of the mixture using the signal distortion ratio (SDR) of BSS toolkit and assigned the source producing the highest SDR figure.
- Here is the SDR versus number of components graph of oracle clustering performed in the paper:



- here is our SDR versus number of components graph:



- As we can see from the two graphs that the maximum SDR obtained in the paper and in our simulation are comparable(both around 6 or 7).
- Also this approach provides better separating performance with large number of bases because the increasing numbers of bases reduces the minimal unit from which the sources can be reconstructed.

Conclusions

- A novel method proposed by Barker et.al. has been implemented in this report.
- The model makes use of redundancy in spectral similarities across frequencies during factorization of a mixture of signals into its components.
- The proposed algorithm was tested for different number of source mixtures and SDR was compared to the values provided in.
- The results are positive and imply the algorithm was implemented with reasonable amount of accuracy and works as well as it is described.
- The SDR attains a maximum for number of sources around 5 in our implementation, whereas the accuracy of NTF in is a strictly monotone which increases at a slow rate.
- Finally, the results also implies that NTF outperforms NMF but the drawback in NTF is its increased time-complexity.