SOUND SEPARATION OF MONAURAL SPEECH VIA NON-NEGATIVE TENSOR FACTORIZATION

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

Indian Institute of Technology, Kanpur
Email: {chahuja,avijit,krmohit,rbala,shahidee}@iitk.ac.in

ABSTRACT

Sound Separation has been a very challenging problem since the origin of multimedia processing. Monaural Sound processing poses a bigger challenge as all the information is contained in the same signal. Precision modelling of the mixed signal is a must to ensure accurate separation of source from the given signal. We have implemented the method proposed by Tom Barker and Tuomas Virtanen in [1]. The signal is modelled as a tensor, which is an element-wise product of 3 matrices. The method involves estimation of the 3 matrices and then reconstruction of the separated signal using wiener like filtering.

1. INTRODUCTION

Monaural Sound Separation from a mixture of signals poses a challenge in the research areas of multimedia processing. There are currently numerous techniques for performing sound source separation, providing different performances for differing optimal condi- tions. Non-negative matrix factorisation (NMF) provides state- of-the-art single-channel blind source separation. Basic NMF techniques decompose a mixture signal into a sum of components having a fixed spectrum and time-varying gain, which when multiplied approximate the mixture spectrogram.One drawback of NMF is that it fails to utilise the redundancies of sounds across frequencies. Another elegant solution was proposed in [1] in 2013. This method involves modelling the Modulation Spectrogram (MS) as a tensor and its 3 factors are estimated by minimizing the Kullback-Leibler (KL) divergence D.

The rest of the report is as follows. Section 2 introduces the modelling of the MS. Section 3 explains the algorithm for estimating the tensor factors. Section 4 contains reconstruction of separated signals. In Section 5 Signal to Distortion Ratio has been calculated to demonstrate effectiveness of the algorithm. Concluding remarks are made in Section 6

2. MODELLING OF MODULATION SPECTROGRAM

The Generation of the modulation spectrogram is based on the computational model of the cochlea. Components of the mixed signal must be significantly different for accurate separation. Components having the same dominant harmonics will lie in the same auditory channel hence separation will not be possible.

A gammatone filter bank is used to divide the signal into different auditory channels. The increase in resolution of the filter bank will increase the probability of separation but will also increase the amount of resources required for computation. This gives rise to a trade-off between separation accuracy and time/resources consumption.

The gammatone filter bank is implemented using Slaneys Auditory Toolbox [2].Each band is passed through a hilbert-transform generated envelope detector. The rectified signal is filtered using alow-pass filter with -3dB bandwidth of approximately 26Hz. An envelope detector is used to convert the signal to a positive quantity, hence being true to the word "Non-Negative" in Non-Negative Tensor Factorization (NTF). The envelopes are segmented into a series of overlapping frames, windowed by a hamming window and finally converted to frequency domain using Fast-Fourier Transform (FFT). The output of each window is stacked on top of each other to give a matrix which represents the Short-Time Fourier Transform (STFT) and the image is known as MS.

The STFT is truncated to 150 bins as the signal is passed through the low pass filter. This implies that the significant information of the source-signals is contained in the low frequency bins and thus processing on higher frequency bins yields no significant change in the final results. Also as the algorithm has to operate on lesser number of points, there is an improvement in the time complexity.

The tensor factorization model is described by χ which has dimensions of R x M x N (number of filterbank channels, truncated STFT length and number of observation frames respectively). Modelling χ as the sum of K components, we get $\hat{\chi}$ as an estimate of χ:

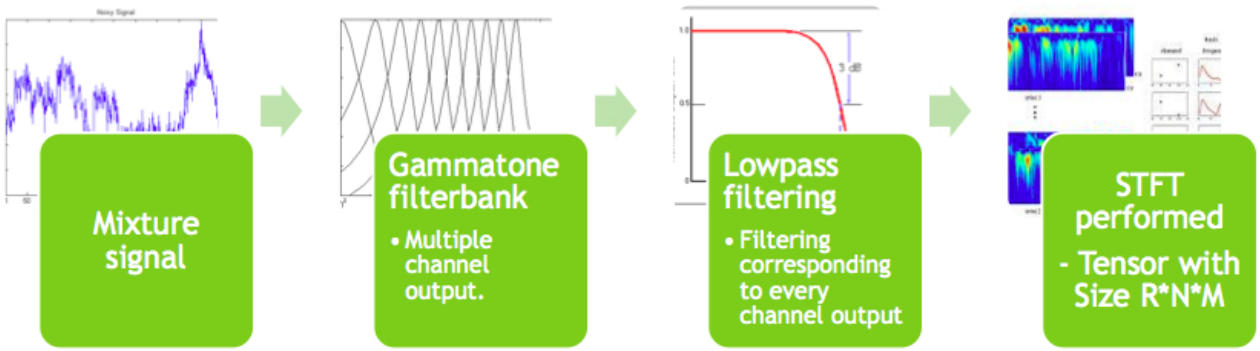

Fig. 1. Schematic of the NTF algorithm implemented in the report

$$\chi_{r,n,m} \approx \hat{\chi}_{r,n,m} = \sum_{k=1}^{K} \mathbf{G}_{r,k}\mathbf{A}_{n,k}\mathbf{S}_{m,k} \tag{1}$$

where $\mathbf{G}_{R,K}$ contains the gain, $\mathbf{A}_{N,K}$ contains the frequency basis function which models the spectral content and $\mathbf{S}_{M,K}$ is the time-varying activation of the component. These model parameters are estimated by minimizing the Kullback-Leibler (KL) divergence D,

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

Minimizing of the cost function D is achieved by the following update sequence. This update sequence is based on gradient descent method and hence finds the global minimum. **G**, **A** and **S** are intialised to random values and then iterated through the following steps using $(C) = \chi/\hat{\chi}$:

$$\mathbf{G}_{r,k} \leftarrow \mathbf{G}_{r,k} \frac{\sum_{n,m} \mathcal{C}_{r,n,m}\mathbf{A}_{n,k}\mathbf{S}_{m,k}}{\sum_{n,m} \mathbf{A}_{n,k}\mathbf{S}_{m,k}} \tag{3}$$

and similarly,

$$\mathbf{A}_{n,k} \leftarrow \mathbf{A}_{n,k} \frac{\sum_{r,m} \mathcal{C}_{r,n,m}\mathbf{A}_{r,k}\mathbf{S}_{m,k}}{\sum_{r,m} \mathbf{G}_{r,k}\mathbf{S}_{m,k}} \tag{4}$$

and

$$\mathbf{S}_{m,k} \leftarrow \mathbf{S}_{m,k} \frac{\sum_{r,n} \mathcal{C}_{r,n,m}\mathbf{A}_{n,k}\mathbf{G}_{r,k}}{\sum_{n,r} \mathbf{A}_{n,k}\mathbf{G}_{n,k}} \tag{5}$$

$\mathcal{C}$ is re-evaluated after every single update and the number of iterations required in the convergence is around 200.

3. SYNTHESIS OF COMPONENTS FROM TENSOR

After factorization of the mixture signal into components we reconstruct it using an approach like Wiener-filtering. Since Wiener-filtering approach requires the parameters in the STFT domain of the mixture signal rather than in the modulation envelope domain, we generate a component synthesis vector by taking the STFT of the output of each auditory filterbank channel. The tensor we get is of the dimension R x P x M as we havent performed the truncation of frequency bins. Also we estimate the matrix of signal reconstruction basis function B using the minimization of Kullback divergence between $|\nu|$ and its approximation $|\hat{\nu}|$. $|\hat{\nu}|$ is calculated as:

$$|\hat{\nu}_{r,n,m}| = \sum_{k=1}^{K} \mathbf{G}_{r,k}\mathbf{B}_{p,k}\mathbf{S}_{m,k} \tag{6}$$

We define $\epsilon = |\nu|/|\hat{\nu}|$ and then repeatedly update B as:

$$\mathbf{B}_{p,k} \leftarrow \mathbf{B}_{p,k} \frac{\sum_{r,m} \epsilon_{r,p,m}\mathbf{A}_{r,k}\mathbf{S}_{m,k}}{\sum_{r,m} \mathbf{G}_{r,k}\mathbf{S}_{m,k}} \tag{7}$$

We initialized B to be some random non-negative value and iterate 200 times. The K separate components $\hat{\nu}^k$ in the STFT domain by applying the obtained Wiener-filter to ν as:

$$\hat{\nu}^k_{r,p,m} = \nu^k_{r,p,m} \frac{\mathbf{G}_{r,k}\mathbf{B}_{p,k}\mathbf{S}_{m,k}}{\sum_{k'} \mathbf{G}_{r,k'}\mathbf{B}_{p,k'}\mathbf{S}_{m,k'}} \tag{8}$$

Then we perform the p-dimensional inverse DFT to convert these K sets of STFTs back to the time domain. By application of overlap and add we have the separated components.

Algorithm for the whole process is explained as follows:

Algorithm 1 NTF for source separation

1: **Initialization**:Choose the mixed signal for analysis
2: Pass the mixture through a Gammatone Filter-Bank to generate multiple channel inputs
3: Perform half-wave rectification and Low Pass Filtering to obtain a non-negative input.
4: Take STFT of the multiple channel rectified input to generate the Feature Tensor χ
5: Estimate **G**, **A** and **S** by using Eq. 3-5.
6: For reconstruction of the separated signals, find ν which is calculated using Steps 1,2,4,5.
7: Keeping **G** and **S** fixed NTF is applied to estimate **B** using Eq. 7
8: K Separated Signals are estimated by applying Eq. 8.

3.1. Simulation Results

Oracle Clustering was performed to estimate the Signal to Distortion Ratio (SDR). From Figure 2 we can see that the maximum SDR was obtained for number of Sources close to 5. This implies that the separation of sources is achieved optimally for more number of sources. This feature is also clear the work by Tom Barker and Tuomas Virtanen in [1] whose results is shown in Figure 3.

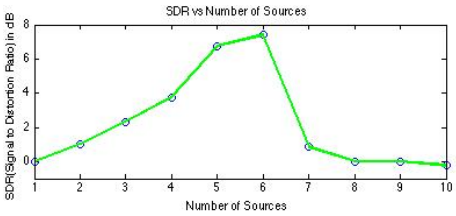

Fig. 2. SDR obtained using oracle clustering for various number of source mixtures

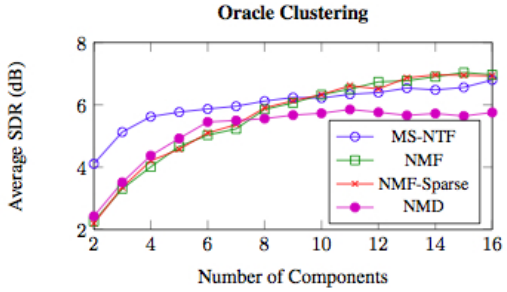

Fig. 3. SDR obtained from various methods in literature. Courtesy: Barker et. al.

4. CONCLUSION

A novel method proposed by Barker et.al.[1] 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 [1]. 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 [1] 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.

5. ACKNOWLEDGMENTS

We would like to thank our mentor, Karan Nathwani, for the immense support that he has provided throughout the project. His constant motivation is what lead us to the final implementation. We would also like to thank Prof. Rajesh Hegde for giving us the opportunity to work on this project.

6. REFERENCES

[1] Tom Barker and Tuomas Virtanen, "Non-negative tensor factorisation of modulation spectrograms for monaural sound source separation," in *Proceedings of INTERSPEECH*, 2013.

[2] Malcolm Slaney, "Auditory toolbox," *Interval Research Corporation, Tech. Rep*, vol. 10, pp. 1998, 1998.