# Discriminative Non-negative Matrix Factorization for Single-Channel Speech Separation

USC Viterbi School of Engineering

Zi Wang

Mentors: Dingchao Lu, Fei Sha

## Introduction

**Speech separation (Cocktail-party problem)**

◆ **Goal:**

Segregating each stream of sound from mixed speech of many speakers.

◆ **Application:**

- Robust speech recognition: preprocessing noisy or multi-speaker speech data
- Improve speech quality: boosting signal noise ratio for targeted speech

## Current approaches

◆ Non-negative matrix factorization (NMF)

- Model non-negative data using parts-based, additive representations
- Exploit speaker-specific parts to separate mixed speech

◆ Sparse Non-Negative Matrix Factorization (SNMF)

- Extend NMF by sparsely combining parts
- Estimate over-complete dictionaries

◆ Limitations

- Learn parts independently
- Does not adapt to other speakers' interference

## Our approaches

● **Main Idea**

◆ Discriminative non-negative matrix factorization (DNMF)

- Learn parts jointly for all speakers
- Optimize parts to be maximally effective in segregating from other speakers

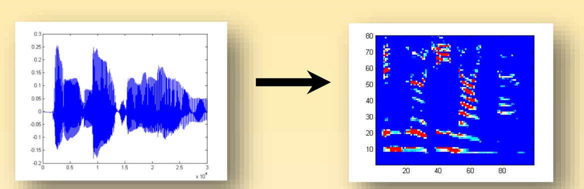
◆ Pairwise DNMF

- Extend DNMF by distinguishing only pairwise speakers
- Reduce computational cost

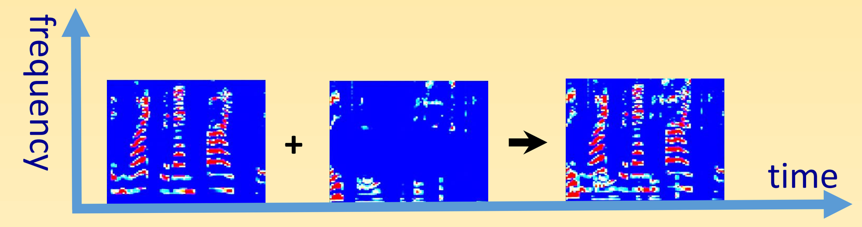
## Nonnegative matrix factorization

◆ **Intuitions**

- Represent speech signals with nonnegative magnitudes of their mel spectrum
- Model mixed signal's spectrum as additive sum of each individual source's spectrum



wav to mel spectrum




Female speaker a (high energy mainly in high frequency bins)

Male speaker b (high energy mainly in low frequency bins)

Mixed speech c (energy in both high and low frequency bins)

◆ **Models**

- Let $D_{i1}, D_{i2}, \ldots, D_{iK}$ denote speaker i's speech prototypes (e.g., one for each phoneme's spectrum), $S_i$ denotes the input signal's spectrum of speaker i
- Minimize the difference between input signals and linear combinations of those prototypes for each speaker

$$F(D,H) = \sum_i KL(S_i \parallel D_i H_i)$$

◆ **Learning** **How to learn prototypes D without knowing h?**

- Iteratively learn D and H for each speaker
- Update rules*

$$H_i \leftarrow H_i \cdot \frac{D_i^T S_i / D_i H_i}{\sum D_i} \qquad D_i \leftarrow D_i \cdot \frac{S_i H_i^T / D_i H_i}{\sum H_i}$$

## Discriminative NMF

◆ **Intuitions**

- Reconstructed speech from clean conditions should also be optimal under other interfering conditions
- Learning jointly all prototypes and consider the sparsity of H

◆ **Models**

- Let $S_{ij}$ denotes the mixed signal's spectrum of speaker i and j
- Let $\hat{S}_{ij}$ denotes the reconstruction of the mixed signal's spectrum

$$F(D,H) = \sum_i KL(S_i \parallel D_i H_i) + \sum_{i,j} KL(S_{ij} \parallel \hat{S}_{ij}) + \lambda \sum H_i \quad \text{where } \hat{S}_{ij} = [D_i \quad D_j] \times \begin{bmatrix} H_i \\ H_j \end{bmatrix}$$

◆ **Pairwise speakers**

- Limit the speakers involved during training
- Easily adapt to multiple speakers

◆ Optimization algorithm

- Optimize each speaker's prototypes alternatively

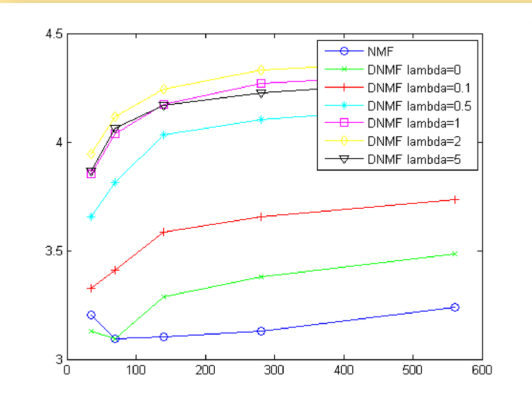
$$H_i \leftarrow H_i \bullet \frac{(D_i)^T \sum_j S_{ij} / \widehat{S_{ij}}}{\sum D_i + \lambda}$$

$$D_i \leftarrow D_i \bullet \frac{\sum_j \frac{S_{ij}}{\widehat{S_{ij}}} H_i^T + U\left(V H_i^T \bullet D_i\right) \bullet D_i}{U\left(\sum_j \frac{S_{ij}}{\widehat{S_{ij}}} H_i^T \bullet D_i\right) \bullet D_i + V H_i^T}$$

## Results

◆ DNMF vs. NMF

| | NMF | DNMF | | | | | |
|---|---|---|---|---|---|---|---|
| component \ λ | 0 | 0 | 0.1 | 0.5 | 1 | 2 | 5 |
| 35 | 3.20 | 3.13 | 3.33 | 3.66 | 3.85 | 3.95 | 3.87 |
| 70 | 3.10 | 3.10 | 3.41 | 3.81 | 4.04 | 4.12 | 4.06 |
| 140 | 3.10 | 3.29 | 3.59 | 4.03 | 4.17 | 4.24 | 4.17 |
| 280 | 3.13 | 3.38 | 3.66 | 4.10 | 4.27 | 4.33 | 4.23 |
| 560 | 3.24 | 3.48 | 3.74 | 4.17 | 4.32 | 4.38 | 4.29 |



- Outperform NMF in improving SNR

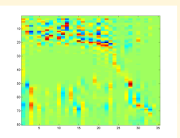
◆ DNMF vs. SNMF

- No significant improvement

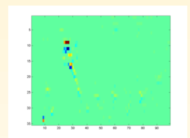
| | DNMF | | | SNMF | | |
|---|---|---|---|---|---|---|
| component | best lambda | best beta | snr | best lambda | best beta | snr |
| 35 | 2 | 1 | 3.94 | 2 | 1 | 4.03 |
| 70 | 2 | 1 | 4.11 | 2 | 1 | 4.13 |
| 140 | 2 | 1 | 4.25 | 5 | 2 | 4.26 |
| 280 | 2 | 1 | 4.33 | 5 | 2 | 4.31 |
| 560 | 2 | 1 | 4.37 | 2 | 1 | 4.37 |

◆ Why similar results?

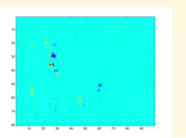
- the dictionary D from training and the activity H from testing are different
- the reconstruction, DH, are similar.



difference of D



difference of H



difference of DH

◆ Pairwise DNMF vs. DNMF

- Compare one pair of speaker (different gender)
- Slightly better than DNMF, need further investigation

| component | 35 | | 140 | | 560 | |
|---|---|---|---|---|---|---|
| model | pair | dnmf | pair | dnmf | pair | dnmf |
| SNR | 5.57 | 5.42 | 5.72 | 5.70 | 5.96 | 5.97 |

## Conclusion

◆ We have developed a new method for speech separation. The key idea is to learn speaker-specific parts discriminatively.

◆ Our method yields promising results, improving the popular approach NMF.

◆ Our method is applicable to other problems where NMF is used.

## Experiment setup

◆ The Grid Corpus

- 34 speakers and 1000 sentences per speaker
- half of the 1000 sentences for each speaker are used for training and the other half for evaluation

◆ Evaluation

- tune parameters and validate on development set(half of the evaluation set)
- evaluate the performance of the model on test set
- Signal noise ration (SNR): the ratio of signal power from reconstructed speech to the residual signals after subtracting reconstructed speech

◆ Analyze the prototypes and reconstruction coefficients to gain further insight

## Selected References

[1] Seung D, Lee L. Algorithms for non-negative matrix factorization[J]. Advances in neural information processing systems, 2001, 13: 556-562.

[2] Schmidt M, Olsson R. Single-channel speech separation using sparse non-negative matrix factorization[J]. 2006.

[3] Eggert J, Korner E. Sparse coding and NMF[C]//Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on. IEEE, 2004, 4: 2529-2533.

[4] Smaragdis, P. and Brown, J.C. Non-negative matrix factorization for polyphonic music transcription. In IEEE Workshop on Applications of Signal Process- ing to Audio and Acoustics (WASPAA), pp. 177 { 180, oct. 2003

[5] Emad M. Grais, , Hakan Erdogan. Regularized nonnegative matrix factorization using Gaussian mixture priors for supervised single channel source separation

* Divisions and dot-multiplications in update rules are element wise