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

Zi Wang, 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

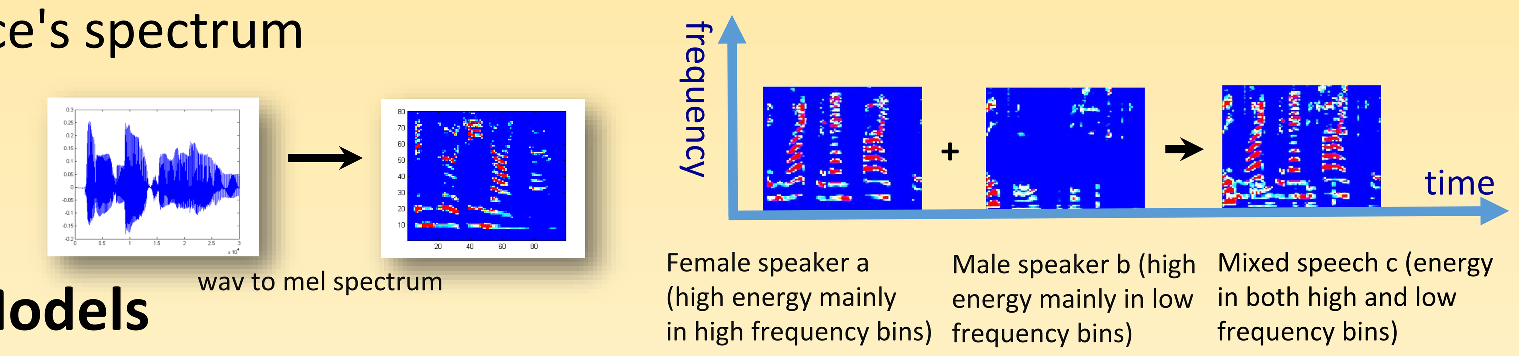
◆ Pairwise DNMF

- 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



◆ **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

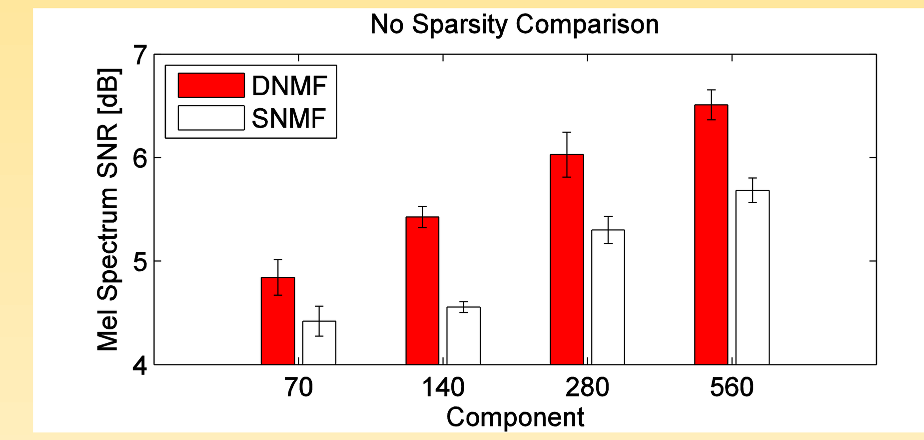
- Optimize each speaker's prototypes alternatively

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

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

Results

◆ DNMF vs. NMF

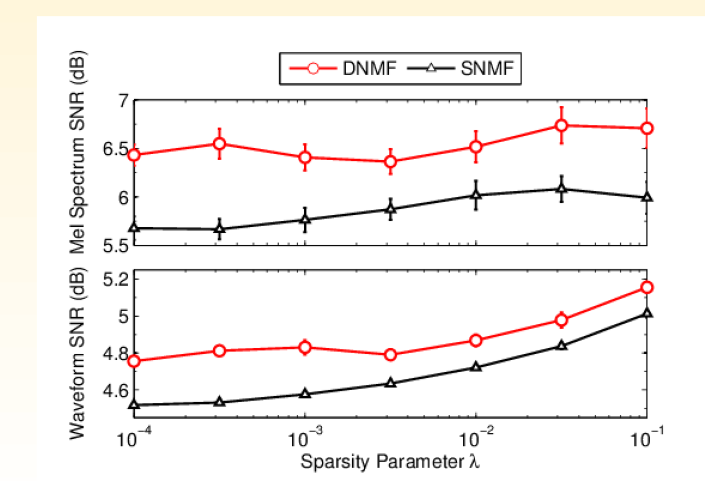


- Outperform NMF in improving SNR

◆ DNMF vs. SNMF

- Outperform SNMF in most parameter settings

SNR^MEL	K \ λ	0	0.0001	0.001	0.01	0.1
DNMF	70	4.84	4.69	4.70	5.07	5.28
	140	5.43	5.43	5.40	5.44	5.72
	280	6.03	6.08	6.05	6.19	6.44
	560	6.51	6.43	6.41	6.52	**6.71**
SNMF	70	4.42	4.42	4.42	4.59	4.91
	140	4.56	4.55	4.61	5.13	5.69
	280	5.30	5.30	5.31	5.67	5.80
	560	5.68	5.68	5.76	**6.02**	5.99
SNR^WAV						
DNMF	70	4.49	4.46	4.39	4.51	4.76
	140	4.54	4.53	4.53	4.67	4.93
	280	4.71	4.67	4.72	4.76	5.08
	560	4.82	4.76	4.83	4.87	**5.16**
SNMF	70	4.29	4.29	4.30	4.39	4.71
	140	4.25	4.25	4.27	4.48	4.87
	280	4.41	4.41	4.44	4.64	4.91
	560	4.52	4.52	4.58	4.72	**5.01**



◆ Gender Difference

- 14% for same gender
- 8.7% for different genders

	λ	0	0.0001	0.001	0.01	0.1
DNMF	SS	4.79	4.77	4.71	4.91	**5.28**
SNMF	SS	4.31	4.30	4.34	4.50	**4.63**
DNMF	OS	7.80	7.67	7.68	7.72	**7.78**
SNMF	OS	6.71	6.71	6.83	**7.16**	7.02

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.

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