

# LINEAR REGRESSION ON SPARSE FEATURES FOR SINGLE-CHANNEL SPEECH SEPARATION



Mikkel N. Schmidt and Rasmus K. Olsson



Informatics and Mathematical Modelling

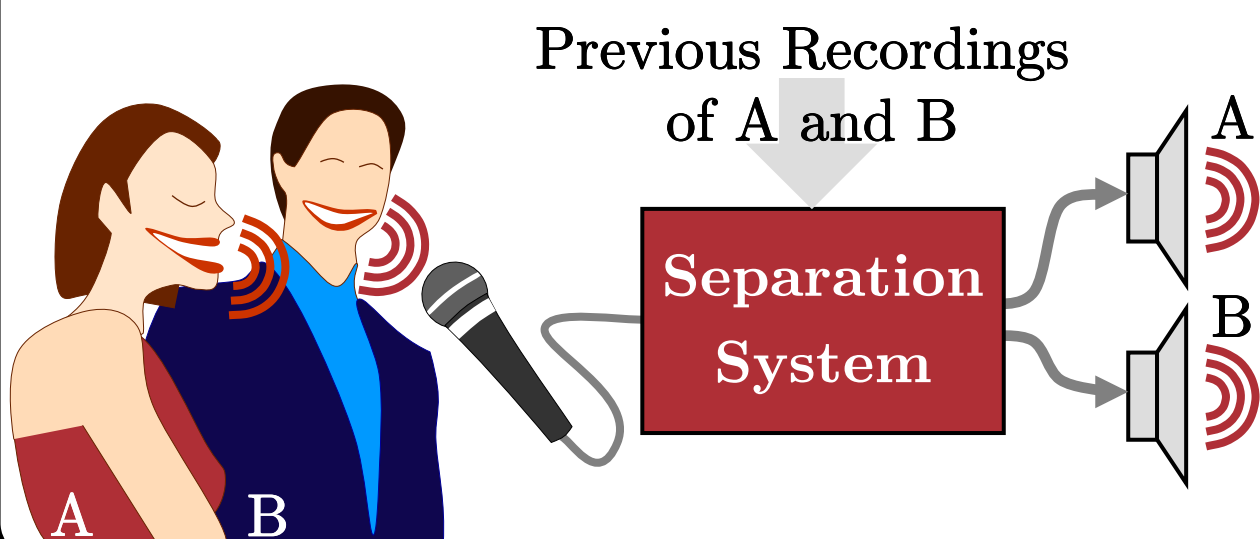
Technical University of Denmark

## Abstract

We use linear regression to separate multiple speech sources from a single channel recording. We show that using sparse non-negative matrix factorization features is significantly better than using spectral features.

## 1 Problem

Separate a single-channel mixture of speech from known speakers



## 2 Separation by Linear Regression

### Input

Mixed speech waveform.

### Features

- Amplitude compressed mel spectral magnitude features .
- Sparse non-negative matrix factorization features (see 3).

### Linear model

Clean speech estimated linearly from of features.

$$\hat{Y}_i^* = W_i^\top (X^* - \mu \mathbf{1}^\top) + m_i \mathbf{1}^\top + N$$

### Estimator

Maximum a posteriori estimate from speech mixtures.

$$W_i^\top = \Gamma_i \Sigma_i^{-1}$$

$$\Gamma_i = (Y_i - m_i \mathbf{1}^\top)(X - \mu \mathbf{1}^\top)^\top$$

$$\Sigma = (X - \mu \mathbf{1}^\top)(X - \mu \mathbf{1}^\top)^\top + \frac{\sigma_n^2}{\sigma_w^2} I$$

This requires training on speech mixtures. We make approximate the MAP estimator to train on clean speech.

$$\Gamma_i \approx (Y_i - m_i \mathbf{1}^\top)(X_i - \mu_i \mathbf{1}^\top)^\top$$

$$\Sigma \approx \sum_{i=1}^P (X_i - \mu_i \mathbf{1}^\top)(X_i - \mu_i \mathbf{1}^\top)^\top$$

## 3 Sparse Features

On training date we compute a sparse NMF decomposition for each speaker.

$$Y_i \approx D_i H_i$$

A mixture is mapped onto the concatenated basis matrices of its constituent speakers.

$$Y^* \approx [D_A D_B] H^*$$

The sparse encoding matrix,  $H^*$ , is used as features in the linear regression.

## 4 Experiments

### Training data

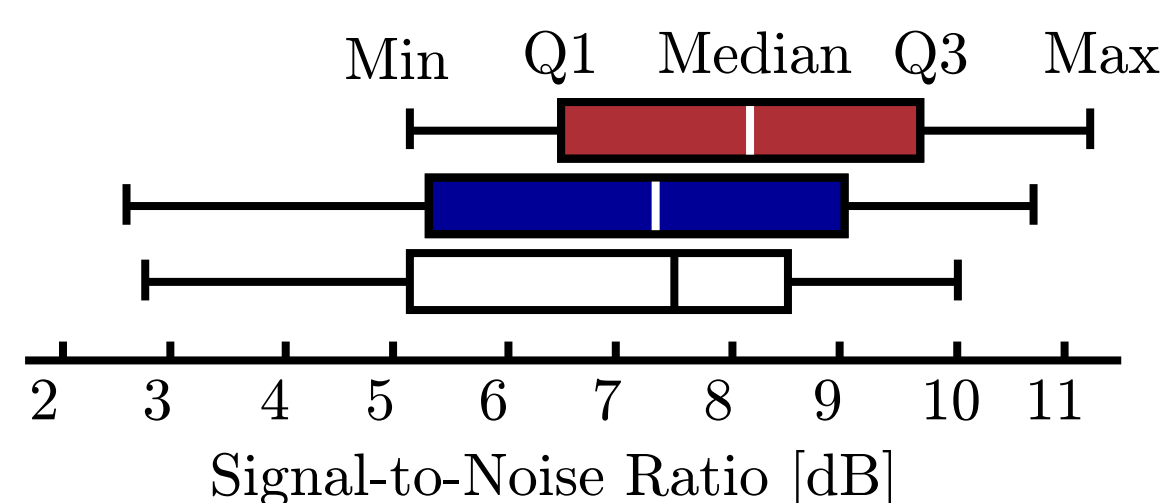
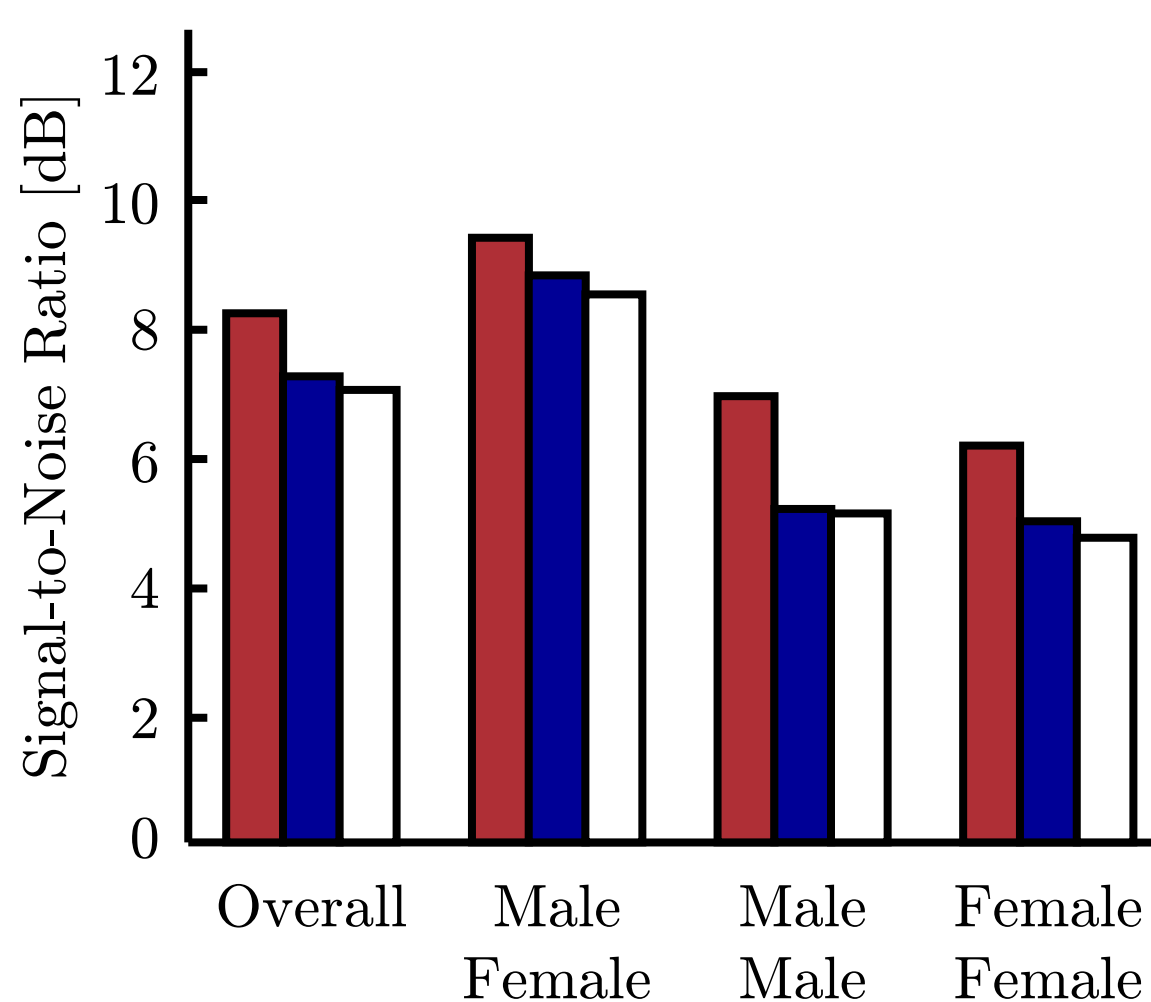
From GRID corpus: 40 minutes of speech from 8 speakers. 4 male + 4 female.

### Test data

9 minutes of 0 dB mixtures of all combinations of speakers. Same speakers as training set but different sentences.

## 5 Results

- Linear regression on NMF features
- Linear regression on spectral features
- Sparse NMF [4,5]



## 6 Conclusion

Linear regression on sparse NMF features leads to better signal separation than linear regression on spectral features or sparse NMF by itself.

## Sparse Non-negative Matrix Factorization

We optimize the following cost with respect to the matrices  $D$  and  $H$  [1,2]

$$C = \|Y - \bar{D}H\|_F^2 + \lambda \sum_{i,j} H_{ij}$$

$$\text{s.t. } D, H \geq 0$$

The cost balances the reconstruction error versus the sparsity of the solution. Sparse factorizations allow for meaningful solutions with large dictionaries.

Fast and simple multiplicative updates can be devised [3]

$$H_{ij} \leftarrow H_{ij} \cdot \frac{Y_i^\top \bar{D}_j}{R_i^\top \bar{D}_j + \lambda}$$

$$D_j \leftarrow D_j \cdot \frac{\sum_i H_{ij} [Y_i + (R_i^\top \bar{D}_j) \bar{D}_j]}{\sum_i H_{ij} [R_i + (V_i^\top \bar{D}_j) \bar{D}_j]}$$

where  $R=DH$ , bold operators are elementwise, and overbar denotes normalization.

## References

- D.D.Lee and H.S.Seung, [Learning the parts of object by non-negative matrix factorization](#), Nature, vol. 401, no. 6755, pp. 788-791, 1999
- P.O.Hoyer, [Non-negative sparse coding](#), in Neural Networks for Signal Processing, IEEE Workshop on, 2002, pp. 557-565.
- J.Eggert and E.Korner, [Sparse coding and NMF](#), in Neural Networks, IEEE International Conference on, 2004, vol. 4, pp. 2529-2533
- M.N.Schmidt, [Speech Separation using Non-negative Features and Sparse Non-negative Matrix Factorization](#), Submitted to Computer Speech and Language, 2008
- M.N.Schmidt and R.K.Olsson, [Single-Channel Speech Separation using Sparse Non-Negative Matrix Factorization](#), Interspeech, 2006