SINGLE-CHANNEL SPEECH SEPARATION USING SPARSE NON-NEGATIVE MATRIX FACTORIZATION

Mikkel N. Schmidt and Rasmus K. Olsson

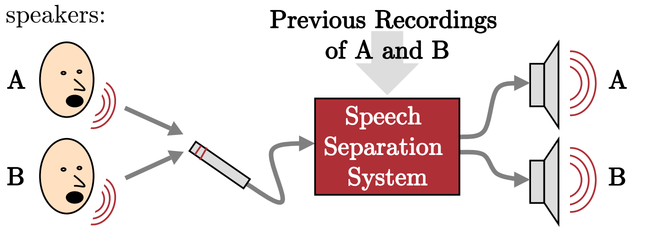
Informatics and Mathematical Modelling Technical University of Denmark

Abstract

We use sparse non-negative matrix factorization to separate multiple speech sources from a single channel recording. We show that computational savings can be achieved by segmenting the training data on a phoneme level using a conventional speech recognition system.

1 Problem

Separate a single-channel mixture of speech from known speakers:



2 Speech Separation

We assume an additive mixing model

$$\mathbf{Y} \approx \mathbf{Y}_A + \mathbf{Y}_B = \begin{bmatrix} \mathbf{D}_A & \mathbf{D}_B \end{bmatrix} \begin{bmatrix} \mathbf{H}_A \\ \mathbf{H}_B \end{bmatrix}$$

The objective is to estimate the speech sources $\mathbf{Y}_A, \mathbf{Y}_B$

Keeping $\mathbf{D} = \begin{bmatrix} \mathbf{D}_A & \mathbf{D}_B \end{bmatrix}$ fixed, the sparse NMF algorithm is used to estimate $\hat{\mathbf{H}}^\top = \begin{bmatrix} \hat{\mathbf{H}}_A^\top & \hat{\mathbf{H}}_B^\top \end{bmatrix}$ from the mixture, $\mathbf{Y}$.

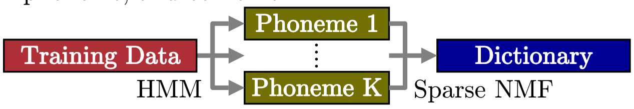
The individual speech sources can the be resynthesized as e.g. $\hat{\mathbf{Y}}_A = \mathbf{D}_A \hat{\mathbf{H}}_A$.

3 Speeding up dictionary learning

a) Learn the dictionary directly from training data.

Training Data
Sparse NMF
Dictionary

b) Seperate the training data into phonemes using a hidden Markov model, learn the dictionary for each phoneme, and combine.


Training Data
HMM
Phoneme 1
⋮
Phoneme K
Sparse NMF
Dictionary

This divides the problem into smaller sub-problems and reduces the computational complexity by a factor K.

Sparse Non-negative Matrix Factorization

We optimize the following cost with respect to the matrices $\mathbf{D}$ and $\mathbf{H}$

$$E = ||\mathbf{Y} - \bar{\mathbf{D}}\mathbf{H}||_F^2 + \lambda \sum_{i,j} \mathbf{H}_{ij} \quad \text{s.t.} \quad \mathbf{D}, \mathbf{H} \geq 0$$

where $\bar{\mathbf{D}}$ is the column-wise normalized dictionary matrix.

The cost balances the reconstruction error (L2 norm) versus the sparsity of the solution (L1 norm). Sparse factorizations allow for meaningful solutions with large dictionaries.

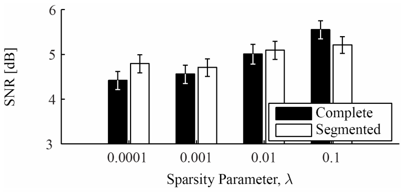
Fast and simple multiplicative updates can be devised [1]

$$\mathbf{H}_{ij} \leftarrow \mathbf{H}_{ij} \bullet \frac{\mathbf{Y}_i^\top \bar{\mathbf{D}}_j}{\mathbf{R}_i^\top \bar{\mathbf{D}}_j + \lambda} \qquad \mathbf{D}_j \leftarrow \mathbf{D}_j \bullet \frac{\sum_i \mathbf{H}_{ij} \left[\mathbf{Y}_i + (\mathbf{R}_i^\top \bar{\mathbf{D}}_j)\bar{\mathbf{D}}_j\right]}{\sum_i \mathbf{H}_{ij} \left[\mathbf{R}_i + (\mathbf{V}_i^\top \bar{\mathbf{D}}_j)\bar{\mathbf{D}}_j\right]}$$

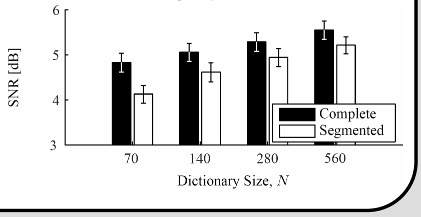
where $\mathbf{R} = \mathbf{DH}$ and the bold operators indicate pointwise multiplication and division.

5 Results

- On a test set, the signal-to-noise (SNR) of the reconstruction is 9.0 ± 1.4 dB for opposite gender and 6.5 ± 1.4 dB for same gender mixtures.
- A certain level of sparsity was found to be optimal.



- Larger dictionaries are better.

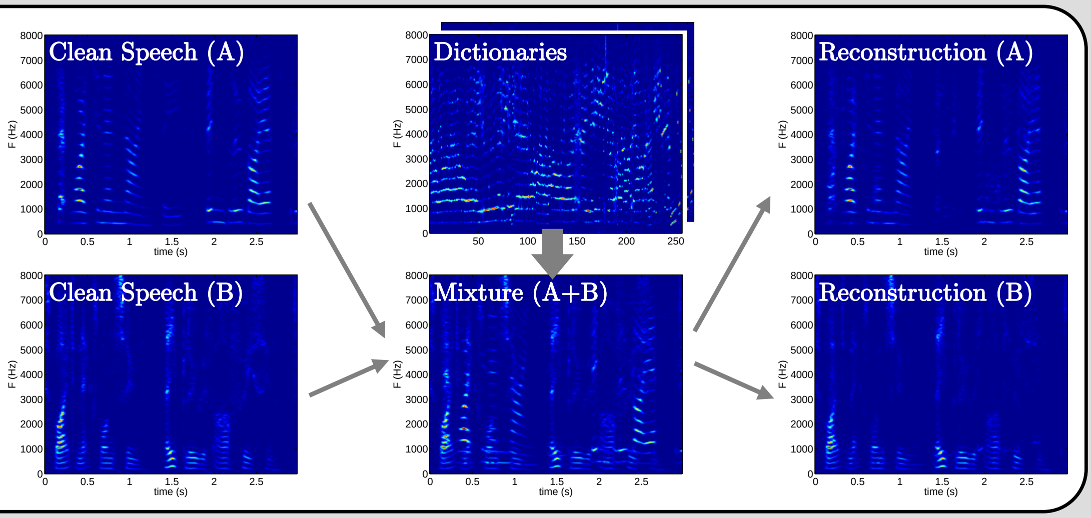


References

[1] J. Eggert and E. Körner, “Sparse coding and NMF,” Proceedings of Neural Networks, vol. 4, pp. 2529–2533, 2004.

4 Example

1) Personalized dictionaries are learned from magnitude spectograms for speakers A and B.
2) Clean test speech is mixed at 0 dB.
3) The mixture is separated into its components, A and B, using the pre-learned dictionaries.


Clean Speech (A)
Clean Speech (B)
Dictionaries
Mixture (A+B)
Reconstruction (A)
Reconstruction (B)
F (Hz)
time (s)