Aalto University
School of Electrical Engineering

Recognition of Reverberant Speech by Missing Data Imputation and NMF Feature Enhancement

Heikki Kallasjoki*, Jort F. Gemmeke, Kalle J. Palomäki, Amy V. Beeston, Guy J. Brown

Department of Signal Processing and Acoustics
Aalto University, School of Electrical Engineering
heikki.kallasjoki@aalto.fi
http://research.spa.aalto.fi/speech/robust/kallasjoki-reverb14/

May 10, 2014

Outline

Introduction

- Two lines of investigation:
  - Missing data methods for dereverberation
  - Extending NMF-based feature enhancement
- Both turn out to be beneficial for reverberant speech (even with multi-condition training, CMLLR adaptation)

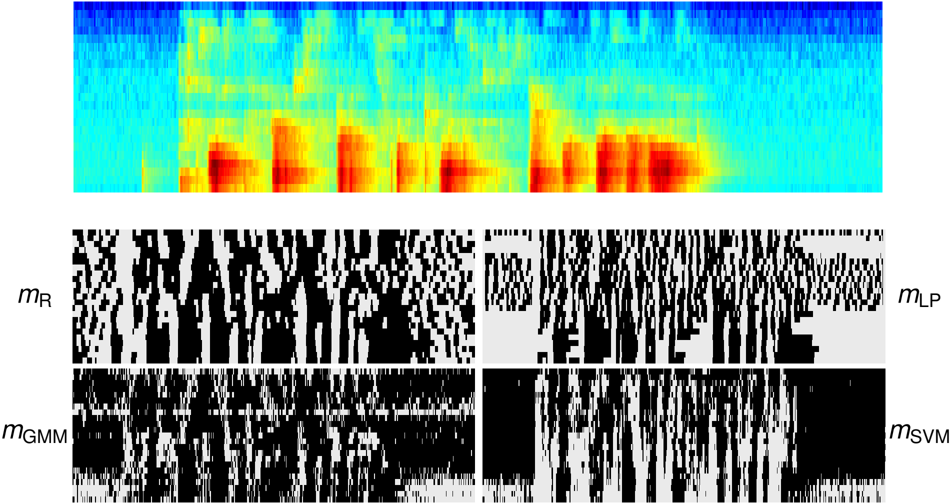
Outline

Missing Data Framework

- Essential idea: focus on spectro-temporal regions dominated by the speech signal
- Estimate reliability (soft or hard decision)
- Use the estimates to improve speech recognition (e.g. by marginalization, imputation...)
- Can make minimal assumptions about the distortion

- In this work: feature imputation with binary masks

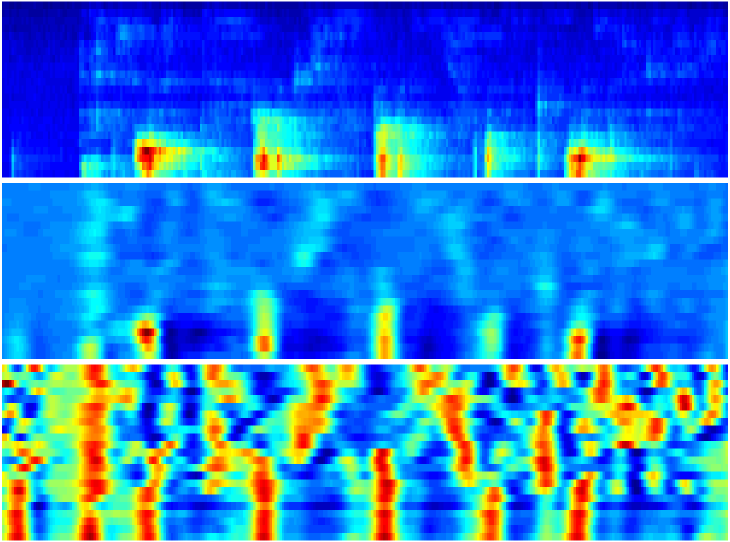
Mask Estimation



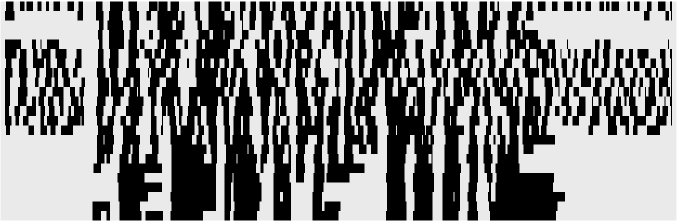
Mask Estimation: m_{R}

- Based on mel-spectral features compressed to $x^{0.3}$
- Band-pass modulation filter, 1.5 . . . 8.2 Hz
- Followed by an AGC and normalization
- Threshold based on "blurredness" metric: ratio of channel mean and channel max

Mask Estimation: m_{R}, illustrated

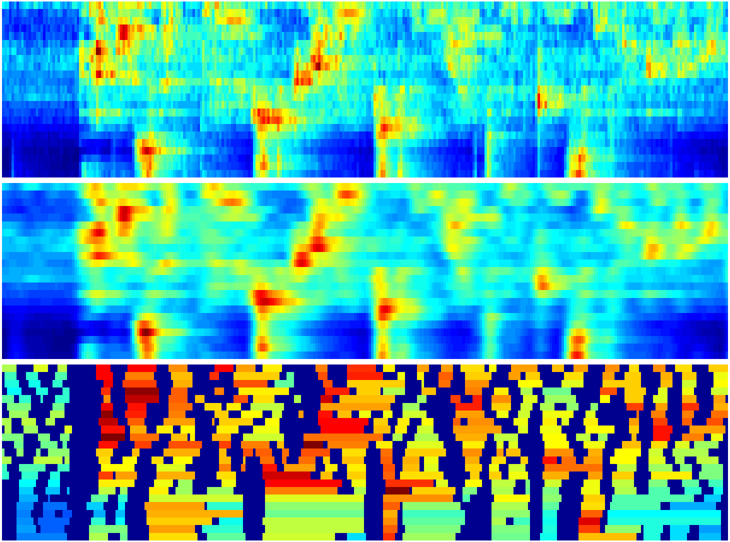


Mask Estimation: m_{LP}

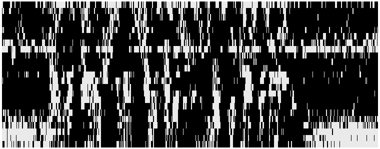
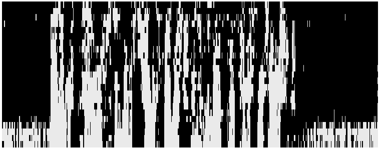


- Based on normalized $x^{0.3}$ mel-spectral features
- Low-pass modulation filter with cutoff at 10 Hz
- Means of each contiguous region where $y' < 0$

Mask Estimation: m_{LP}, illustrated



Mask Estimation: m_{GMM} & m_{SVM}

- Oracle mask:
  threshold difference between clean and reverberant
- Features: spectra, gradient, “blurredness”, m_{R}, m_{LP}
- Train a (GMM or SVM) classifier for each channel

Bounded Conditional Mean Imputation

Conditional Mean Imputation

- Model distribution of clean speech $\mathbf{x}$ with a GMM
- Estimate missing $\mathbf{x}_u$ by conditioning on reliable $\mathbf{x}_r$:

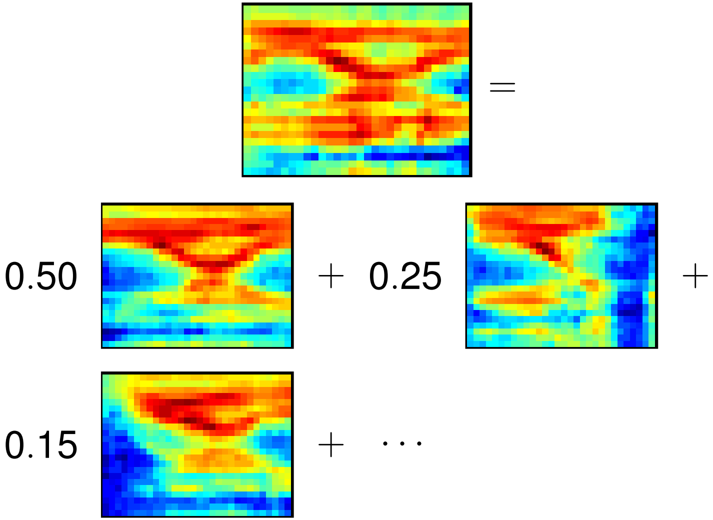
$$\hat{\mathbf{x}}_u = \int_{\mathbf{x}_u} \mathbf{x}_u \, p(\mathbf{x}_u \mid \mathbf{x}_r)$$

Bounded Conditional Mean Imputation

- Use observation as upper bound: $\hat{\mathbf{x}}_u < \mathbf{x}_u^{\text{obs}}$
- In this work:
  truncated $p(\mathbf{x}_u \mid \mathbf{x}_r)$ approximated with a parametric model

Outline

NMF Signal Model



Using NMF for Speech Feature Enhancement

Example: source separation for noisy speech

- Fixed dictionary of clean speech and noise samples (also called *exemplars*)
- After solving coefficients, reconstruct clean speech only
- A lot of flexibility here

Using NMF for Speech Feature Enhancement

Example: source separation for noisy speech

- Fixed dictionary of clean speech and noise samples (also called *exemplars*)
- After solving coefficients, reconstruct clean speech only
- A lot of flexibility here

What about reverberation?

- Source separation approach not directly applicable

Accounting for Reverberation

$$\mathbf{Y} \approx \mathbf{S}\,\mathbf{A}$$

$TC \times W$
stacked
observation

$TC \times N$
dictionary
matrix

$N \times W$
activation
matrix

Accounting for Reverberation

$$\mathbf{Y} \approx \mathbf{R}\ \mathbf{S}\ \mathbf{A}$$

$T_r C \times W$	$T_r C \times TC$	$TC \times N$	$N \times W$
stacked observation	filter matrix	dictionary matrix	activation matrix

- (**RS**) **A**: modeling with a reverberated dictionary
- **R** (**SA**): reverberating the NMF approximation

The Filter Matrix R

$$\mathbf{R} = \underbrace{\left.\left(\begin{array}{cccccccc}
r_{1,1} & 0 & 0 & & & & & \\
0 & r_{1,2} & 0 & \cdots & & & & \\
0 & 0 & r_{1,3} & & & & & \\
 & \vdots & & \ddots & & & & \\
r_{2,1} & 0 & 0 & & r_{1,1} & 0 & 0 & \\
0 & r_{2,2} & 0 & \cdots & 0 & r_{1,2} & 0 & \cdots \\
0 & 0 & r_{2,3} & & 0 & 0 & r_{1,3} & \\
 & \vdots & & \ddots & & \vdots & & \ddots \\
\multicolumn{4}{c}{\underbrace{\hphantom{r_{1,1}\quad r_{1,2}\quad r_{1,3}\quad \cdots}}_{C}} & & & &
\end{array}\right)\right\} T_r C}_{TC}$$

Issues

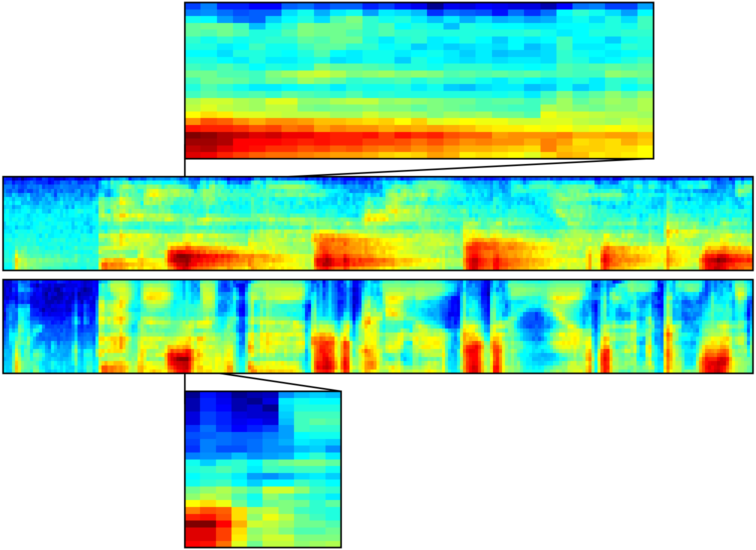
- Does not want to converge to a useful solution

- Sliding-window approach not so suitable for reverberation

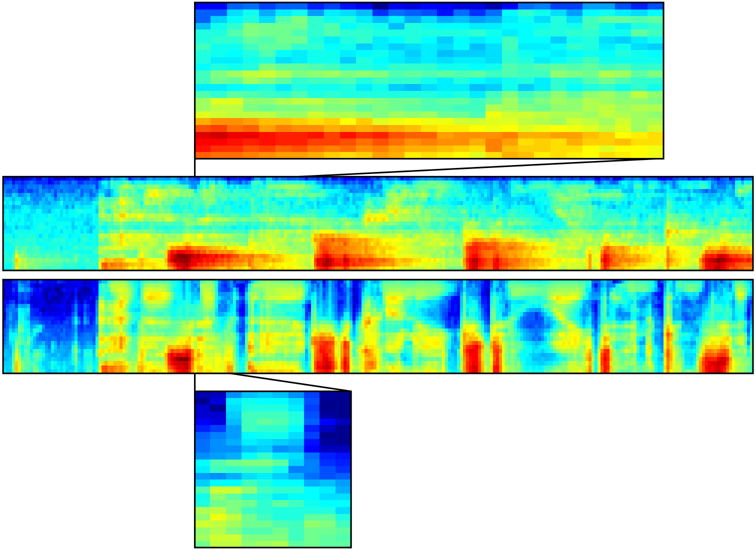
Issues

- Does not want to converge to a useful solution
  - Initialization with missing-data imputation
  - Tuning of iteration scheme
  - Activation matrix filtering
- Sliding-window approach not so suitable for reverberation
  - Sum overlapping windows in multiplicative updates
  - (Or do convolutive NMF)

The Case for Convolutional NMF



The Case for Convolutional NMF



NMF Feature Enhancement Process

1. Estimate $\tilde{\mathbf{X}}$ using BCMI
2. Iteratively update $\mathbf{A}$ in $\tilde{\mathbf{X}} \approx \mathbf{RSA}$ with identity $\mathbf{R}$
3. Filter $\mathbf{A}$ to suppress consecutive nonzero activations
4. Initialize $\mathbf{R}$ to contain filter $\frac{1}{T_f}\left[1 \ldots 1\right]$ on all channels
5. Iteratively update $\mathbf{R}$ in $\mathbf{Y} \approx \mathbf{RSA}$ with fixed $\mathbf{A}$
   (under constraints $r_{t+1,b} < r_{t,b}, \;\; \sum_{t,b} r_{t,b} = C$)
6. Iteratively update $\mathbf{A}$ in $\mathbf{Y} \approx \mathbf{RSA}$ with fixed $\mathbf{R}$

- Then use $\hat{\mathbf{X}} = \mathbf{SA}$ and $\hat{\mathbf{Y}} = \mathbf{RSA}$ for feature enhancement, with a per-frame Wiener filter in the mel-spectral domain

Outline

Further Processing

Channel Normalization

- Mean of the $\frac{1}{L}$ largest-valued samples on each channel
- Reduces mismatch between NMF dictionary and test data

Beamforming

- Simple delay-sum beamformer
- TDOA estimation with PHAT-weighted cross-correlation

Outline

Setup

- REVERB Challenge HTK recognizer
- Four sets of acoustic models:

Clean	WSJCAM0 clean speech training set
MC	REVERB Challenge multi-condition training set
MC+ad.	...with CMLLR adaptation over a test condition
8-ch.	...on audio preprocessed with the PHAT-DS beamformer

Results for Mask Estimation Methods

- Development set, clean speech acoustic models

		SimData	RealData
Baseline		51.81	88.51
BCMI	mask m_{R}	40.07	**67.88**
	mask m_{LP}	48.01	73.06
	mask m_{GMM}	**39.94**	70.87
	mask m_{SVM}	40.78	74.14
NMF (with m_{R})		*28.26*	*58.84*

Results for Mask Estimation Methods

- Development set, clean speech acoustic models

		SimData	RealData
Baseline		51.81	88.51
BCMI	mask m_{R}	40.07	**67.88**
	mask m_{LP}	48.01	73.06
	mask m_{GMM}	**39.94**	70.87
	mask m_{SVM}	40.78	74.14
NMF (with m_{R})		*28.26*	*58.84*

Results for Feature Enhancement

Model	FE	SimData	RealData
Clean	Baseline	51.82	89.04
	BCMI	39.14	71.67
	NMF	**29.74**	**59.13**
MC	Baseline	29.60	56.58
	BCMI	27.25	51.31
	NMF	**24.11**	**47.06**
MC+ad.	Baseline	25.37	48.88
	BCMI	24.58	46.05
	NMF	**21.91**	**41.41**
8-ch.	Baseline	19.76	40.21
	BCMI	19.40	38.28
	NMF	**17.80**	**34.79**

Results for Feature Enhancement

Model	FE	SimData	RealData
Clean	Baseline	–	–
	BCMI	−24.5%	−19.5%
	NMF	**−42.6%**	**−33.6%**
MC	Baseline	–	–
	BCMI	−7.9%	−9.3%
	NMF	**−18.5%**	**−16.8%**
MC+ad.	Baseline	–	–
	BCMI	−3.1%	−5.8%
	NMF	**−13.6%**	**−15.3%**
8-ch.	Baseline	–	–
	BCMI	−1.8%	−4.8%
	NMF	**−9.9%**	**−13.5%**

Outline

Conclusions

Main results

- Both methods are beneficial in reverberant environments, also in conjunction with MC training, CMLLR, beamforming
- NMF approach outperforms the missing data methods
- Activation filtering degrades performance for clean speech

Future plans

- Missing data: improving the mask estimation
- NMF: convolutional NMF, activation matrix filtering
- Tackling both noise and reverberation with NMF
- Use of uncertainty information

References

- K. J. Palomäki, G. J. Brown, and J. P. Barker, “Techniques for handling convolutional distortion with ‘missing data’ automatic speech recognition,” Speech Communication, vol. 43, no. 1-2, pp. 123–142, 2004.
- U. Remes, “Bounded conditional mean imputation with an approximate posterior,” in Proc. INTERSPEECH, 2013, pp. 3007–3011.
- K. J. Palomäki, G. J. Brown, and J. P. Barker, “Recognition of reverberant speech using full cepstral features and spectral missing data,” in Proc. ICASSP, 2006.

Samples and sources

`http://research.spa.aalto.fi/speech/robust/kallasjoki-reverb14/`

Questions

?