ICASSP
2011 International Conference on Acoustics Speech and Signal Processing
Prague, Czech Republic, May 26th, 2011

Model-based Compressive Sensing for Distant Multi-party Speech Recognition

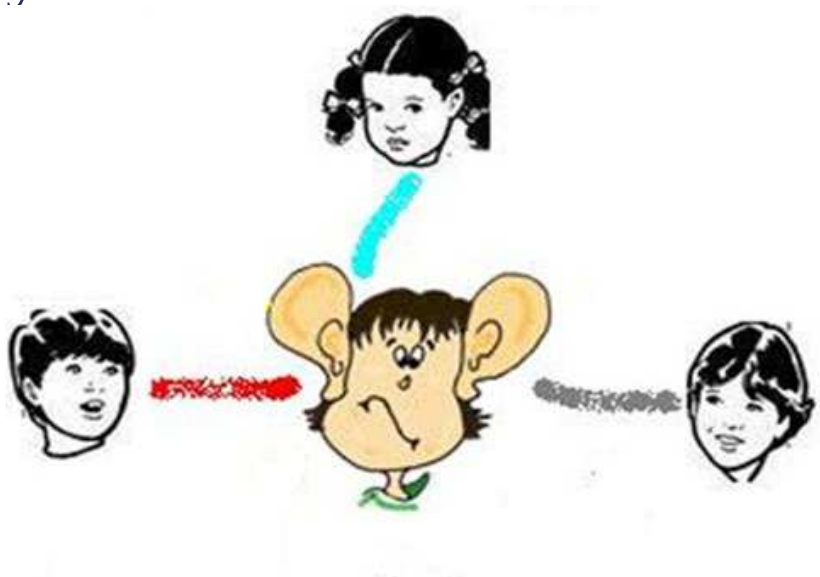
Afsaneh Asaei, Hervé Bourlard, Volkan Cevher
Idiap Research Institute,
École Polytechnique Fédérale de Lausanne

idiap RESEARCH INSTITUTE

EPFL ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

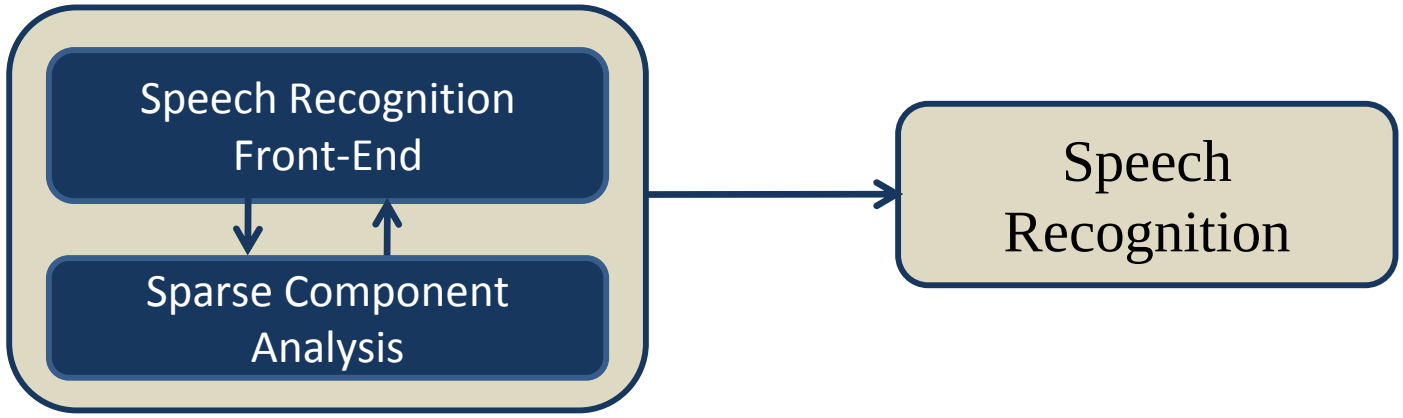
Problem of competing speech sources

- **Goal:** Speech separation to perform speech recognition
- **Challenge:** Fewer measurements than unknown sources
- **Approach:** Sparse Component Analysis



Key idea

- **We cast the under-determined speech separation problem as a sparse signal recovery where we leverage compressive sensing theory to solve it**
  - Integrate sparse component analysis into front-end processing of speech recognition system



Distant Speech Recognition Front-End

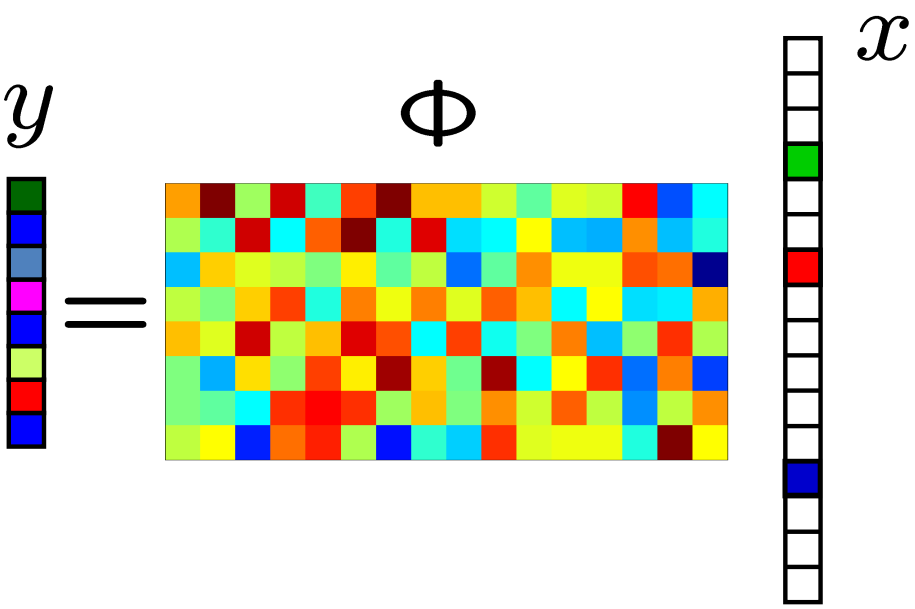
Agenda

- **A short review of compressive sensing**
- **Blind source separation via model-based compressive sensing (BSS-MSR)**
- **Speech recognition experiments**
- **Conclusion**

Compressive Sensing (CS)

In a nutshell

- **CS is sensing via dimensionality reduction**
  - Dimensionality reduction naturally happens in many problems. So, we can leverage the CS theory and algorithms.



CS Premises, in theory …

I. Sparse representation

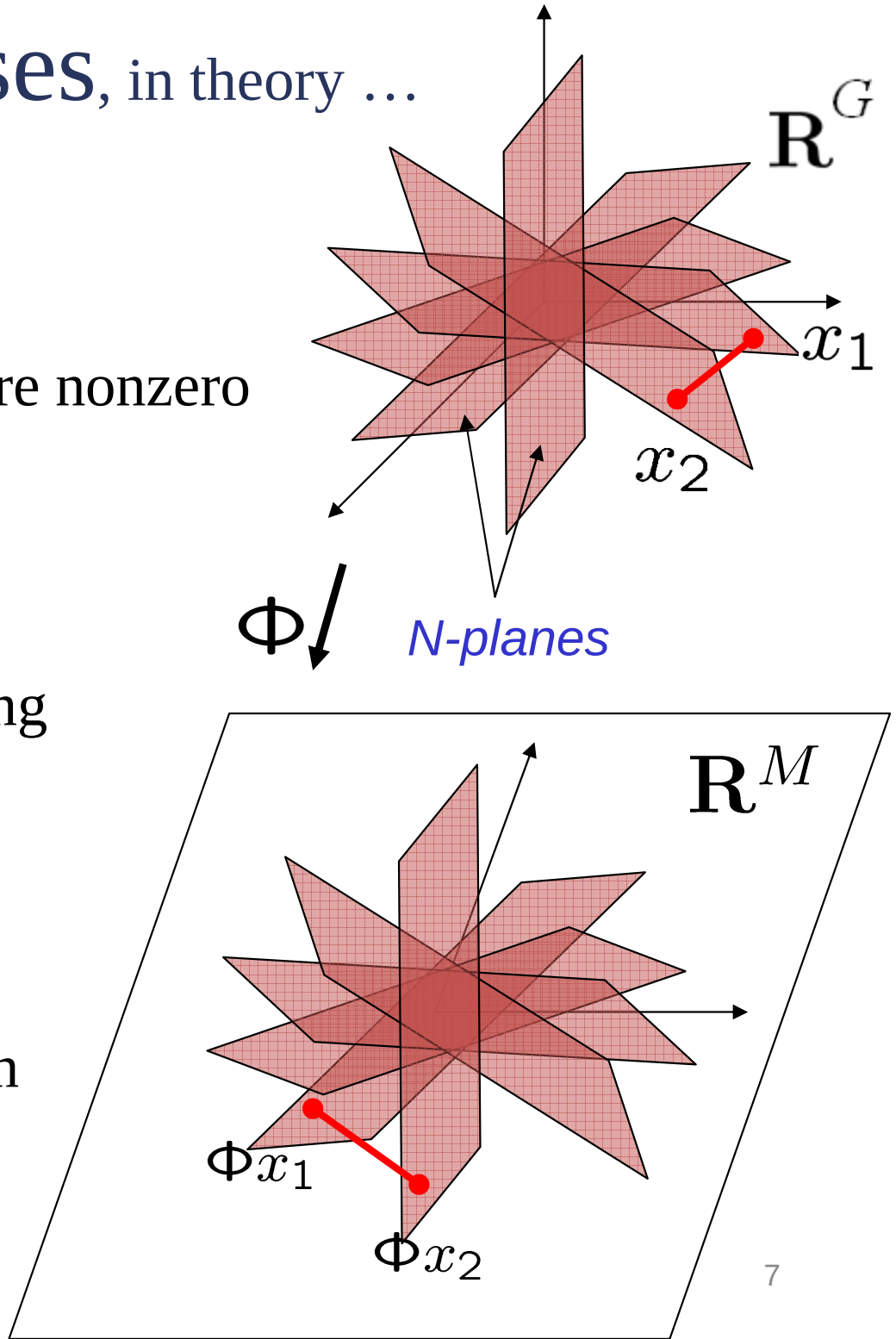
- Only *N* out of *G* coordinates are nonzero

II. Incoherent measurements

- Distance/information preserving

III. Signal recovery

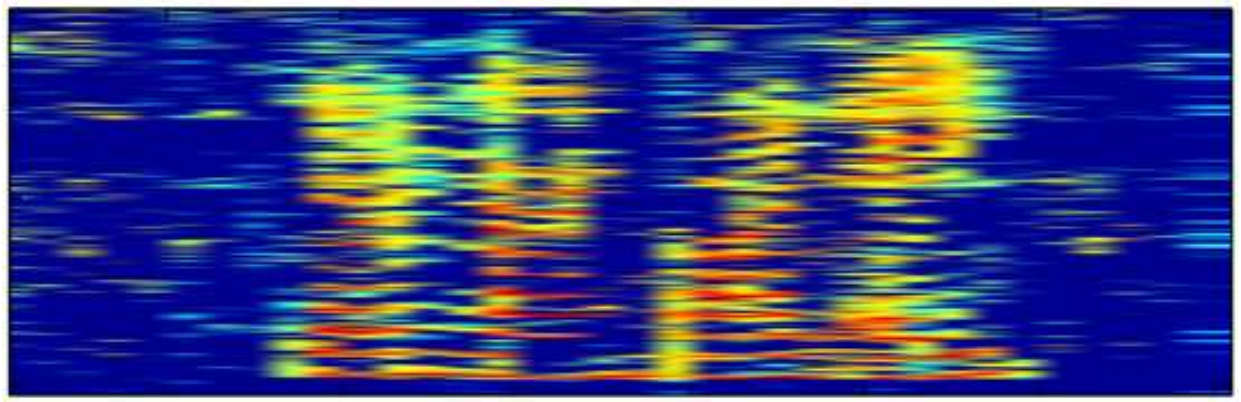
- Search for the sparsest solution

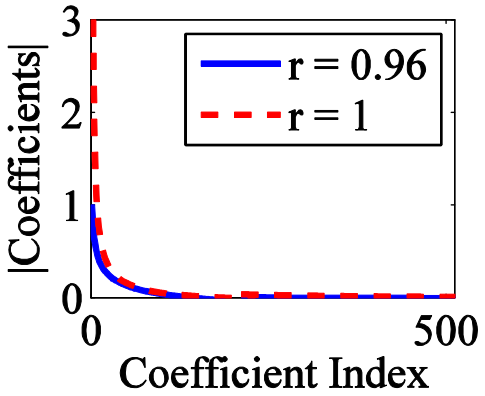


Model-based CS, in practice …

➢ Compressible representation

- Sorted coordinates decay according to the power-law with the rate $r < 1$
  - Sparse representation of speech is obtained by Gabor expansion





➢ Model-based signal recovery

- Leveraging the structure underlying the sparse coefficients improve the recovery performance and reduces the number of required measurements

Blind Source Separation via Model-based Sparse Recovery (BSS-MSR)

Insights from 2000's

➢ Sparse component analysis

[*Yilmaz , Rickard ; IEEE TSP'04* | Zibulevsky, Bofill; SP'01 | Saab et al. IEEE TSP'07 | Gribonval, ICASSP'02 | O'Grady, Pearlmutter; ICA'04 | Georgiev et al.; IEEE TNN'05]

➢ Source localization by sparse recovery

[*Cevher et al. IPSN'09* | Model and Zibulevsky; SP'06 | Malioutov, Cetin, and Willsky; IEEE TSP'05 | Guo et al. MSSP'10 | Chen et al.; Proc. of IEEE'03]

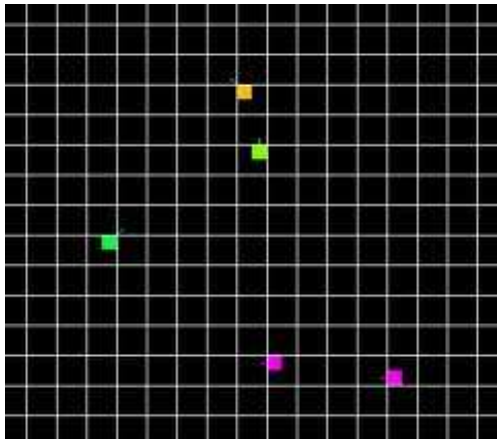
What is new about BSS-MSR?

- Model-based sparse recovery
- Convolutive mixtures
- New efficient and accurate recovery algorithm (ALPS)

I. Spatio-spectral sparse representation

➢ Spatial sparsity

- We discretize the room into G dense grids: only very few have speech activity



➢ Spectral sparsity

- We work with the time-frequency representation of speech

➢ Spatio-spectral representation

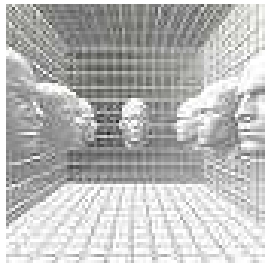
- It holds a **block-sparsity model**

$$X(\omega,\tau)=\begin{bmatrix} X_1(\omega,\tau) \\ \cdot \\ \cdot \\ X_N(\omega,\tau) \end{bmatrix}$$

II. Incoherent Measurement

- **Natural CS measurements are manifested by the media Green's function [Carin'09]**

- Image Model of multi-path effect



source at υ ; sensor at μ

$$\xi^{\omega}_{\upsilon\to\mu} : Y(\omega,\tau) = \sum_{r=1}^{R} \frac{\iota^{r}}{\left\|\mu-\upsilon_{r}\right\|} \exp(-j\omega \frac{\left\|\mu-\upsilon_{r}\right\|}{c}) X(\omega,\tau)$$

ι : Reflection coefficient $\quad$ c : Speed of sound

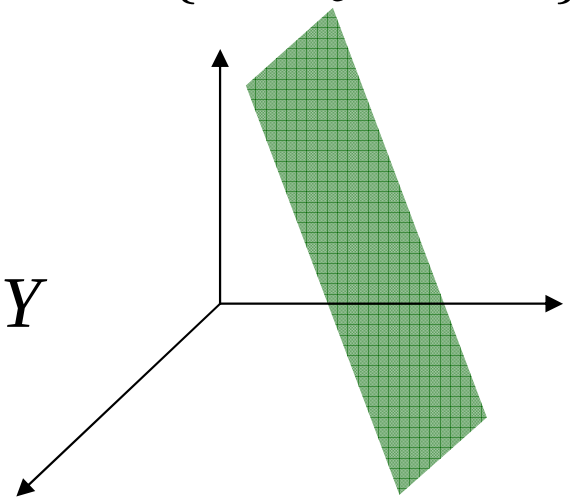
- Microphone array measurement matrix $\quad \Phi = \begin{bmatrix} \xi^{\Omega}_{\upsilon_1\to\mu_1} \cdots \xi^{\Omega}_{\upsilon_1\to\mu_N} \\ . \\ \xi^{\Omega}_{\upsilon_M\to\mu_1} \cdots \xi^{\Omega}_{\upsilon_M\to\mu_N} \end{bmatrix}$

III. Sparse signal recovery

- **Objective:** recover K-sparse signal $X_{G\times 1}$, $N << G$
  - Array observation: $Y_{M\times 1}$
  - Measurement matrix: $\Phi_{M\times N}$, $M < G$

Challenge: nullspace of Φ

$\forall V \in \text{kernel}(\Phi), X' = X + V \Rightarrow Y$



Sparsity gives enough prior information to overcome the ill-posed nature of the inverse problem

$\Rightarrow$ The recovery algorithm seeks the sparsest solution

III. Model-based Sparse Recovery

- **We estimate the signal using ALPS sparse recovery algorithm***
  - A model approximation is performed along with a gradient calculation at each iteration by thresholding (H_N) the energy of the blocks

$$X_{i+1} = H_N\left(X_i + \kappa \Phi^t (Y - \Phi X_i)\right)$$

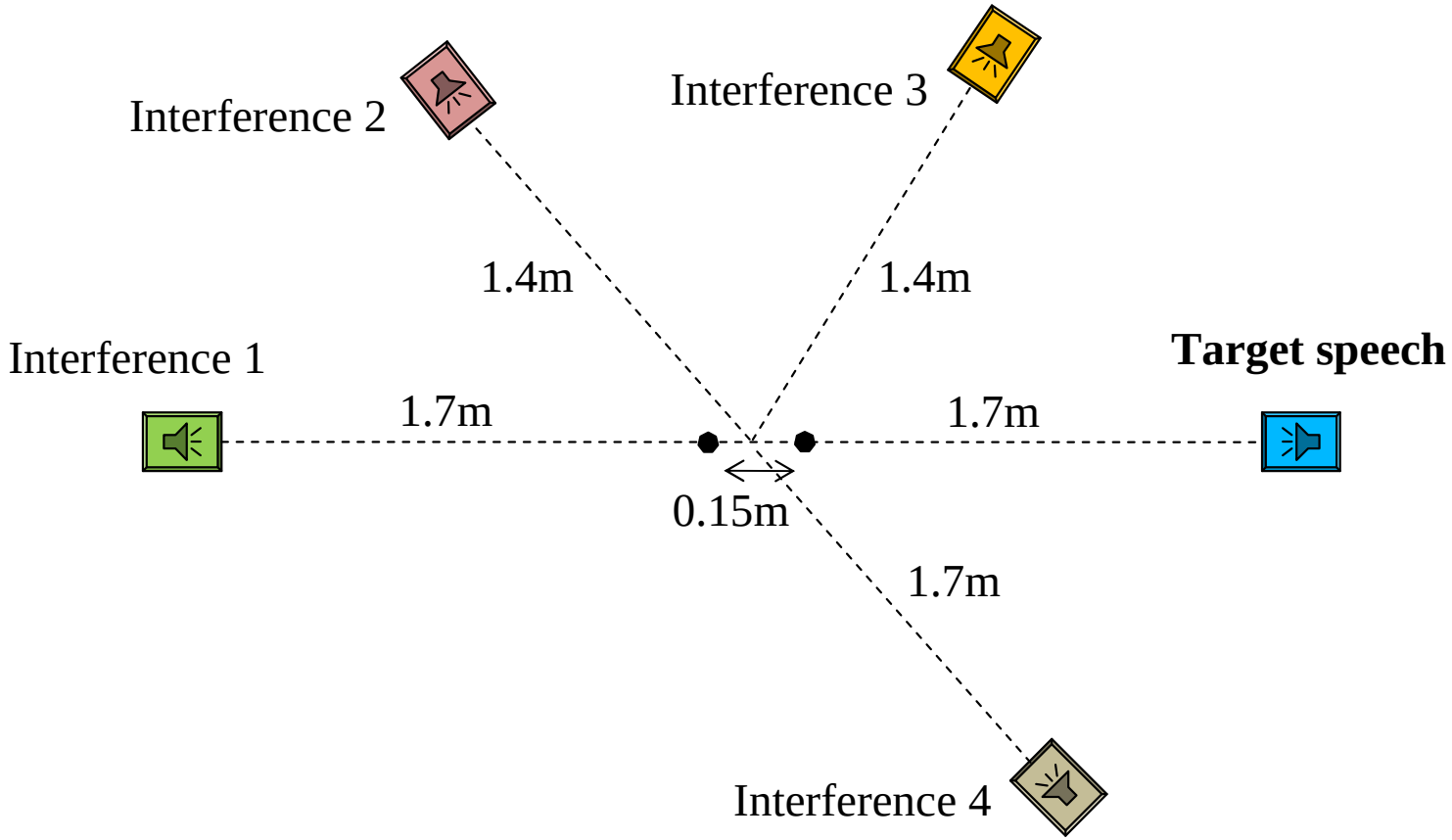
- Volkan Cevher, "An ALPS View of Sparse Recovery", *ICASSP'2011*.

  Codes available at http://lions.epfl.ch/ALPS

Speech Recognition Experiments

Experiments

- Reverberation time: 200ms
- Grid resolution: 0.5m×0.5m
- Speech Corpus: AURORA2 overlapping with HTIMIT interferences



Speech recognition performance

- Word accuracy of the separated speech for stereo echoic mixtures of 3 sources

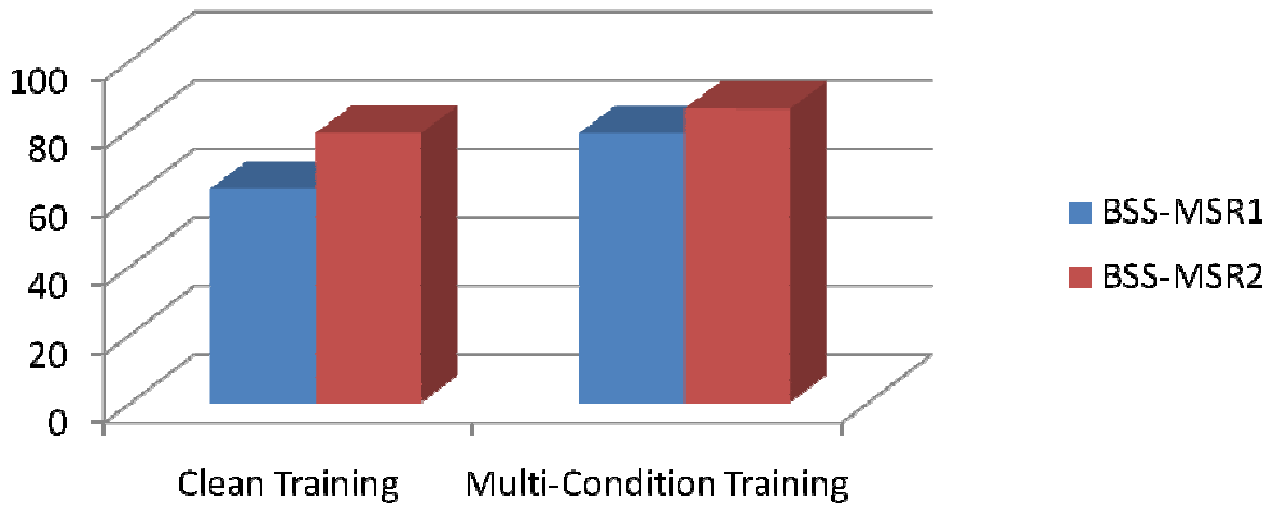
Training Condition	Aurora2 baseline	BSS-MSR
Clean	59.3	89.3
Multi-condition	61.78	92.7

- Relative improvement for recognition of 3 competing sources would be **73.7**% for clean-training and **80.9**% for multi-condition-training

Speech recognition performance

➢ Word accuracy of the separated speech for echoic mixtures of 5 sources
BSS-MSR[1] refers to the stereo recording and BSS-MSR[2] refers to 4-channel array

Training Condition	Aurora2 baseline	BSS-MSR[1]	BSS-MSR[2]
Clean	47.3	81.7	88.7
Multi-condition	58.19	91	94



Conclusion

- **Information bearing components for ASR are sparse**
- Sparse component analysis is a potential approach to deal with the problem of overlapping speech for realistic applications of ASR
- Model-based sparse recovery goes beyond sparsity
- Block-sparsity model exists in
  - Microphone array signal ensemble
  - Virtual source images due to multi-path effect

It motivates new approaches in multi-channel sparse component analysis in reverberant condition

Thank you!

Good … Bad

- **The good things about BSS-MSR are**
  - No constraint on the number of microphones, but
    - ✓ the more, the better
  - No restriction on the geometry of the array, but
    - ✓ the array must provide a reasonable spatial information
- **The bad things about BSS-MSR are**
  - Requires knowledge of microphones, but
    - ✓ automatic calibration is possible with the same framework
  - Requires knowledge of the room acoustic, but
    - ✓ initial evaluations on the simple Image model was promising