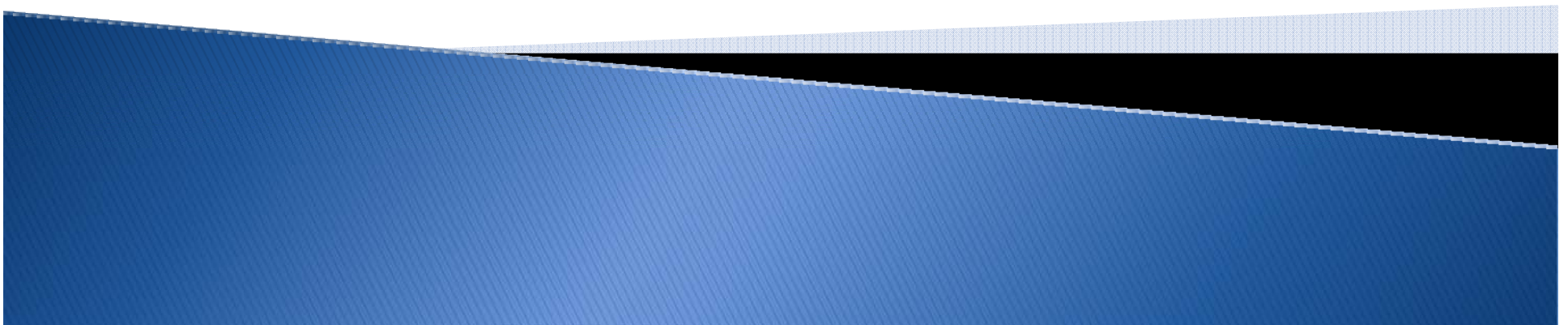


Utilizing Compressibility to Reconstruct Unreliable Time-Frequency Representations: An Application to Speech Recognition

Bengt J. Borgstrom and Abeer Alwan

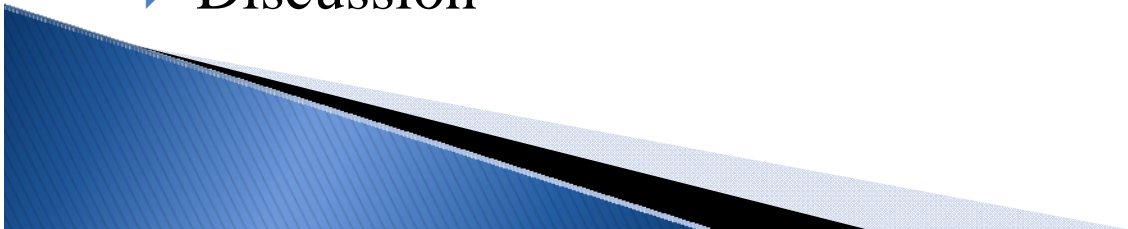
UCLA Annual Research Review

4/24/09



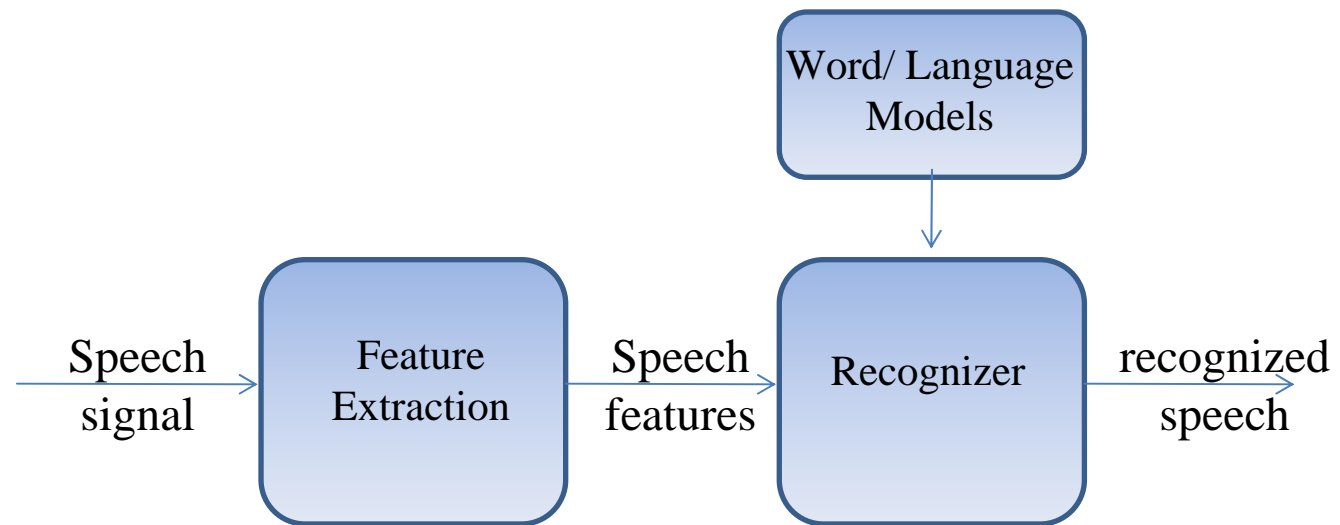
Presentation Outline

- ▶ Noise robustness in automatic speech recognition (ASR)
- ▶ Overview of compressive sensing
- ▶ The compressibility of spectro-temporal speech data
- ▶ Spectral reconstruction utilizing compressibility of spectro-temporal data
- ▶ Experimental ASR results
- ▶ Discussion



Overview of Automatic Speech Recognition Systems

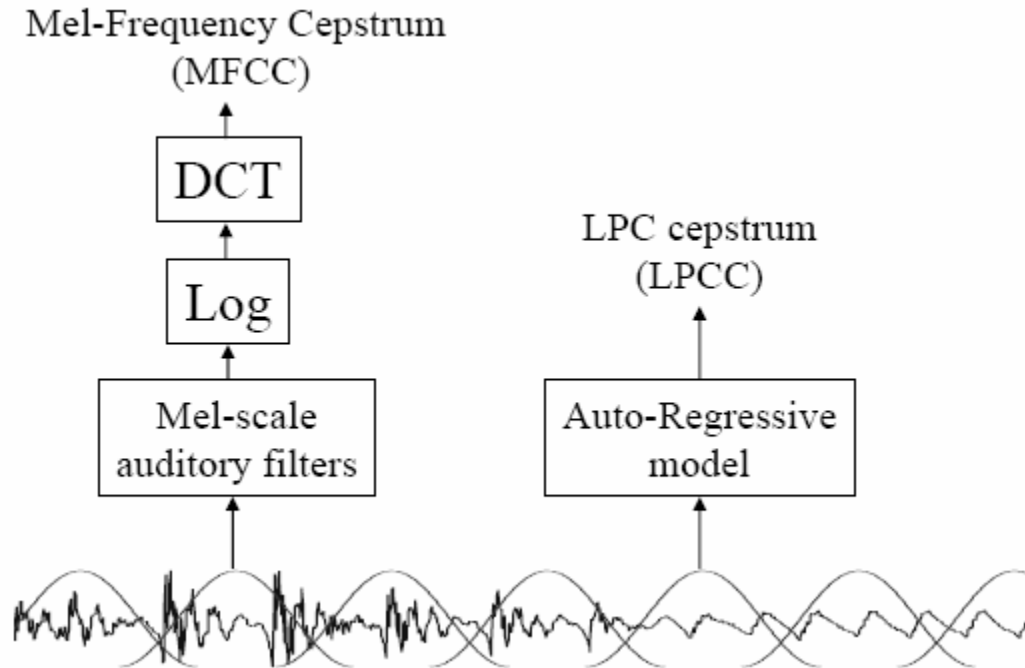
- In ASR systems, discriminative features are extracted from speech signals and passed to the recognizer.
- The recognizer uses word & language models to recognize the given feature sets.



Feature Extraction

- Common features used for recognition include Mel-Frequency Cepstral Coefficients (MFCCs) and Linear Predictive Cepstral Coefficients (LPCCs)

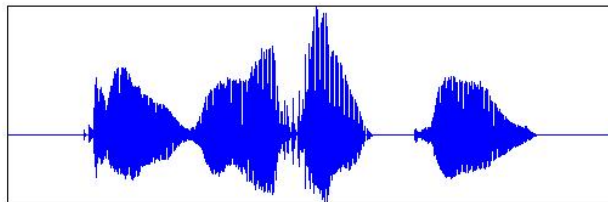
- DCT decorrelates features →
- Log decreases dynamic range →
- Mel-scale filters warp frequency scale →
- Windows speech signal →



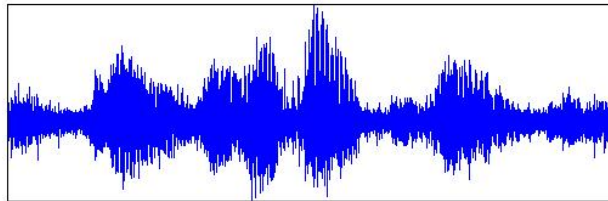
Noise Robust Automatic Speech Recognition (ASR)

- The presence of background acoustic noise degrades the performance of automatic speech recognition (ASR) .
- Estimating clean spectra is nontrivial since background noise generally occurs in channels of important speech activity.

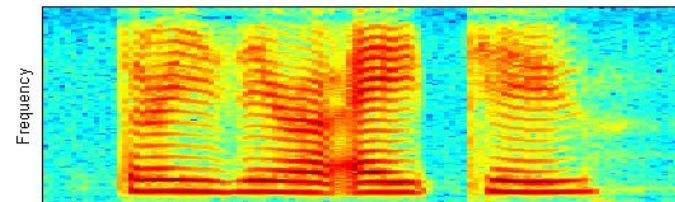
Time domain signals of “three
zero eight two”
clean



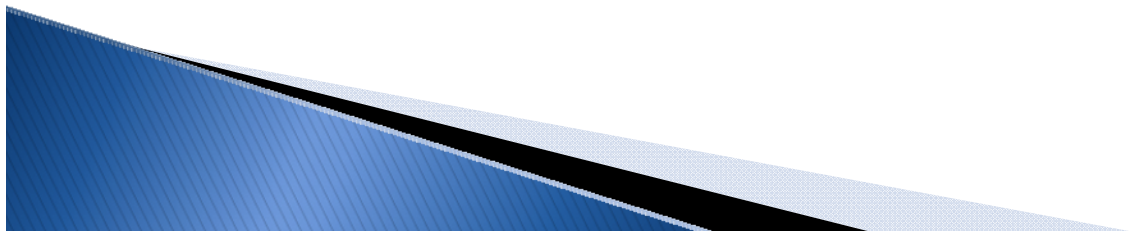
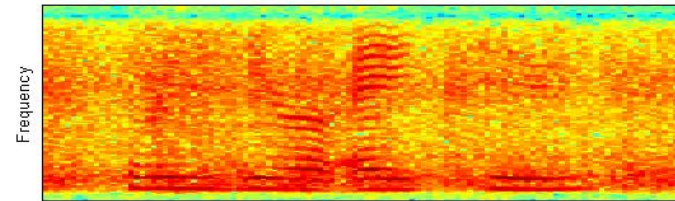
5 dB SNR (Subway Noise)



Time-frequency representations
(spectrograms)
clean

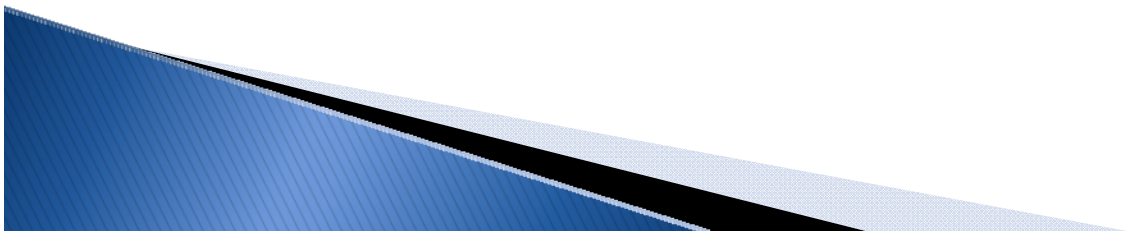


5 dB SNR (Subway Noise)



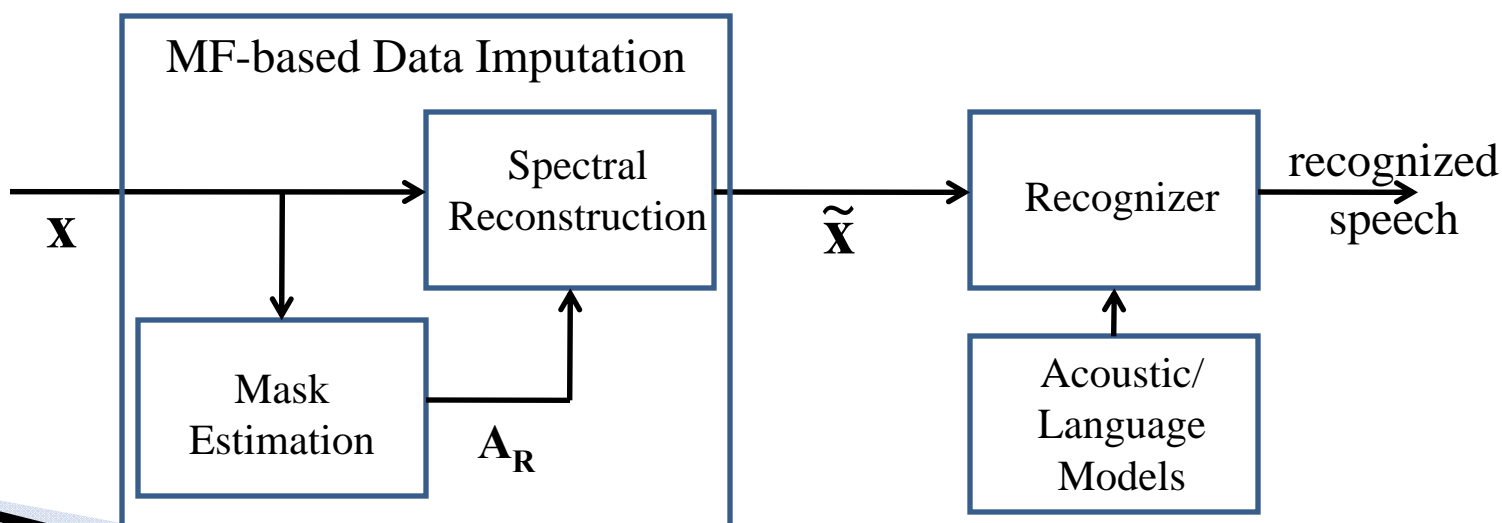
Approaches to Noise Robust ASR

- Noise robust feature extraction
- Missing feature approaches aim to locate unreliable components in the feature domain, and reconstruct them based on neighboring reliable features
- Recognizer-based techniques aim to weight speech features according to reliability during the recognition process
- Multimodal recognition techniques aim to integrate other speech-related modes such as visual information with acoustic features



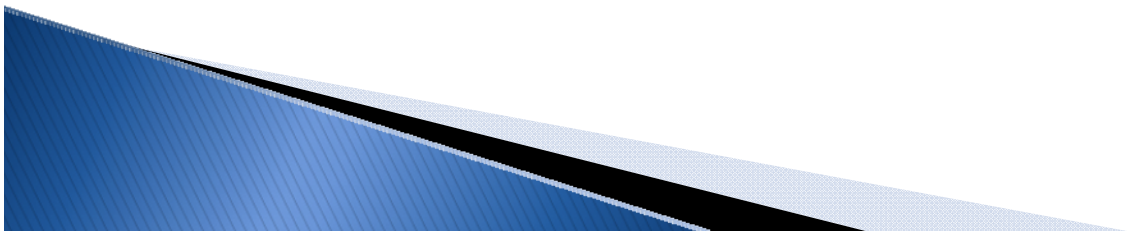
Missing Feature-Based ASR

- Missing feature (MF)-based data imputation can be used in ASR to estimate clean spectro-temporal speech data prior to recognition.
- This process involves mask estimation to determine the reliability of spectral components, followed by spectral reconstruction.
- Previous studies have performed spectral reconstruction using iterative correlation-based methods, or cluster-based methods (Raj et al., 2004).
 - These methods are relatively complex and require data-dependent training.



Compressive Sensing (CS): Problem Formulation

- ▶ CS theory states that perfect reconstruction can be achieved when sampling below the Nyquist rate by exploiting the *sparsity* of signals.
- ▶ Let $\mathbf{f} \in \mathbf{R}^N$ represent a signal of interest, and let $\Phi \in \mathbf{R}^{K \times N}$ represent a matrix of sensing functions, such that incomplete observations of \mathbf{f} are obtained via $\mathbf{y} = \Phi \mathbf{f}$.
- ▶ Let $\Psi \in \mathbf{R}^{N \times N}$ be a suitable transformation that reveals a sparse representation of \mathbf{f} : $\mathbf{v} = \Psi \mathbf{f}$.
- ▶ The aim of compressive sensing is to reconstruct the signal \mathbf{f} with little or no distortion, based on the incomplete observation set \mathbf{y} , where $K \ll N$.

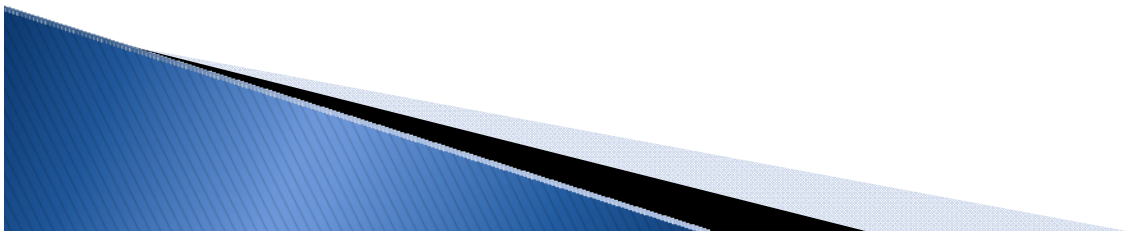


Compressive Sensing: Solution

- ▶ The term *sparsity* describes \mathbf{v} as being comprised of few nonzero terms.
- ▶ CS recovers the original signal efficiently via a linear program:

$$\min_{\tilde{\mathbf{v}} \in \mathbb{R}^N} \|\tilde{\mathbf{v}}\|_1 \quad \text{subject to:} \quad \mathbf{y} = \mathbf{\Phi}\mathbf{\Psi}^* \tilde{\mathbf{v}}$$

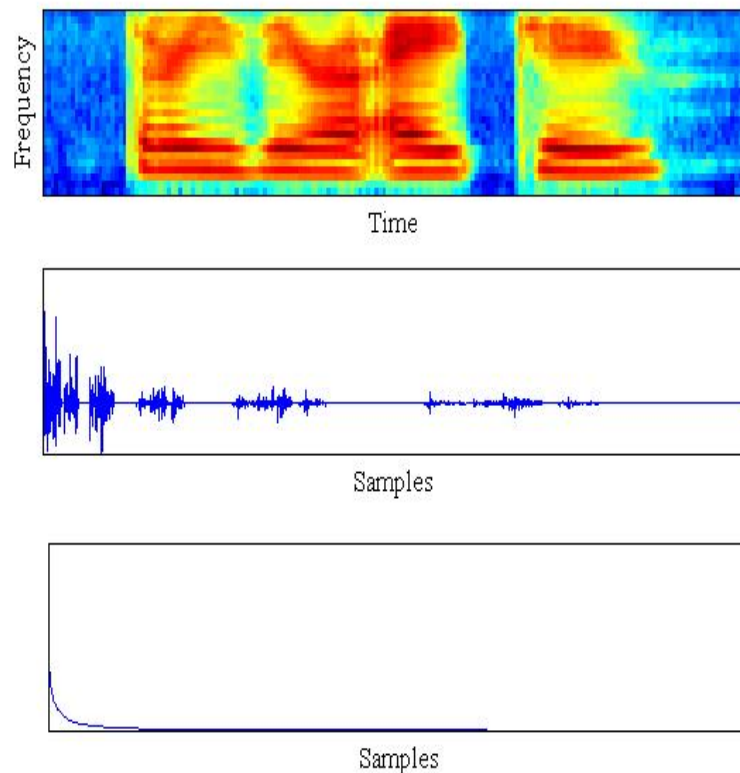
- ▶ If \mathbf{v} is truly sparse, compressive sensing can achieve perfect reconstruction.
- ▶ However, \mathbf{v} can be compressible, so that term magnitudes decrease rapidly. In this case, CS can achieve near-perfect reconstruction.



The Compressibility of Speech Data

- We apply the discrete Haar transform (DHT) to vector-form Mel-filtered spectrograms.
- The middle panel shows the DHT of the spectrogram, revealing few nonzero components.
- The bottom panel shows the sorted magnitudes of the DHT, revealing high compressibility of spectro-temporal speech data.

“three zero eight two”



The Effect of Induced Sparsity on ASR

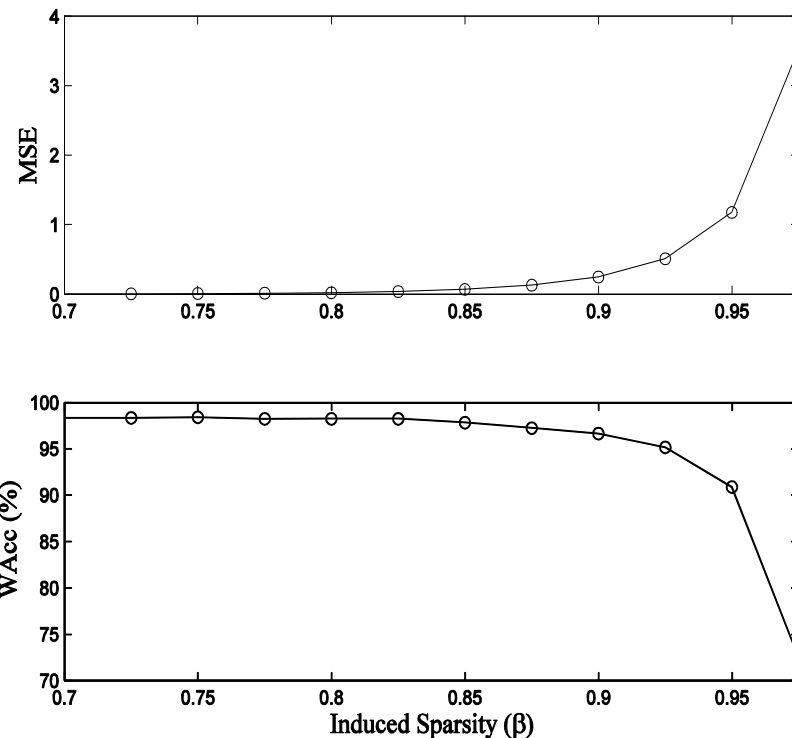
- We force N_s of the N terms in \mathbf{v} to zero, thus inducing sparsity in \mathbf{f} .

- The portion of zeroed terms is: $\beta = \frac{N_s}{N}$

- The top panel shows the MSE of spectro-temporal data as a function of induced sparsity.

- The bottom panel shows the word accuracy as a function of induced sparsity.

- As can be observed, β can be increased to ~ 0.9 without noticeable performance degradation.



Reconstruction of Unreliable Spectro-Temporal Data

- The additive noisy speech model can be approximated in the spectral domain as:

$$\mathbf{x} = \mathbf{s} + \mathbf{d}$$

where \mathbf{x} is the observed signal, \mathbf{d} is noise, and \mathbf{s} is the clean underlying speech.

- It is assumed that certain components of \mathbf{x} are deemed unreliable due to the corruptive effect of noise.

- We define the selection matrix \mathbf{A}_R as:

$$\mathbf{A}_R = \begin{cases} 1, & \text{if } \mathbf{x}(j) \text{ is the } i^{\text{th}} \text{ reliable term in } \mathbf{x} \\ 0, & \text{else} \end{cases}$$

- In this way, the vector $(\mathbf{A}_R \mathbf{x})$ gives the set of reliable components of \mathbf{x} .



Reconstruction of Unreliable Spectro-Temporal Data

- Letting Ψ be the DHT basis, we can use CS theory to reconstruct the unreliable components of \mathbf{x} :

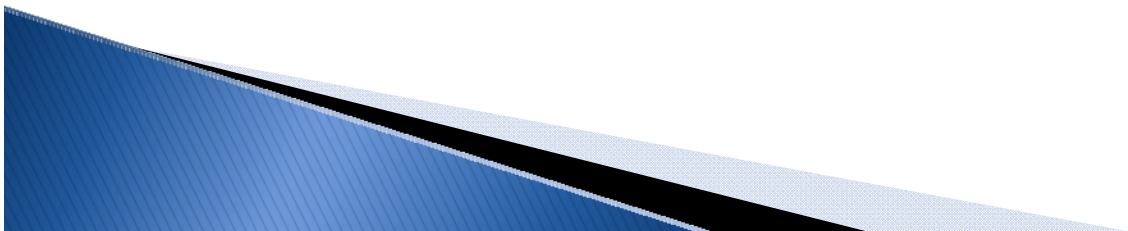
$$\min_{\tilde{\mathbf{v}} \in \mathbb{R}^N} \|\tilde{\mathbf{v}}\|_1 \quad \text{subject to:} \quad \mathbf{A}_R \mathbf{x} = \mathbf{A}_R \Psi^* \tilde{\mathbf{v}}$$

- We can apply two constraints which follow intuitively from the additive speech model to achieve a more accurate reconstruction:

$$\min_{\tilde{\mathbf{v}} \in \mathbb{R}^N} \|\tilde{\mathbf{v}}\|_1 \quad \text{subject to:} \quad \mathbf{A}_R \mathbf{x} = \mathbf{A}_R \Psi^* \tilde{\mathbf{v}}$$

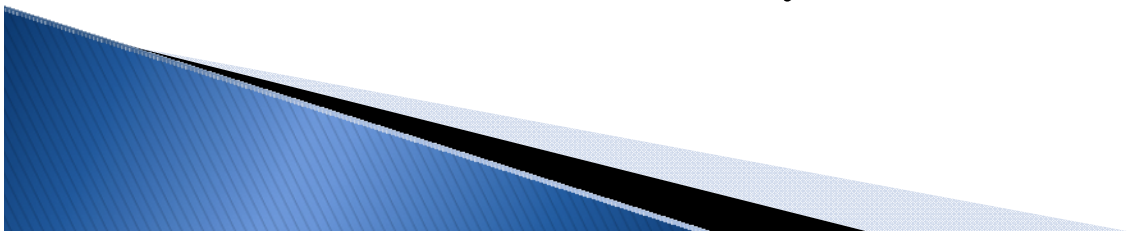
$$\Psi^* \tilde{\mathbf{v}} \geq 0$$

$$\Psi^* \tilde{\mathbf{v}} \leq \mathbf{x}$$



Comparisons with Compressive Sensing

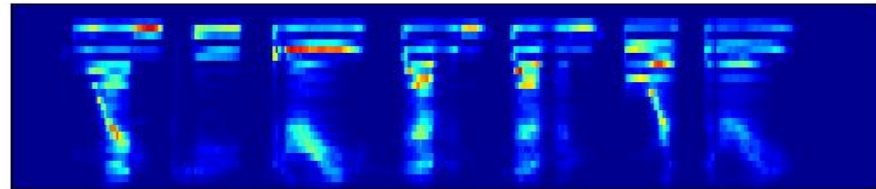
- As can be interpreted from the previous slides, there exists great similarity between CS theory and the proposed spectral reconstruction algorithm.
 - Both techniques aim to reconstruct an original signal based on an incomplete set of observations.
 - Both techniques exploit the sparsity of the underlying signal.
- However, the concept of *sensing* is different.
 - In CS, sensing matrices, Φ , are actively designed to minimize the coherence with the representation matrix, Ψ .
 - In the proposed algorithm, the sensing matrix, \mathbf{A}_R , is not actively designed, and is instead determined by the effect of noise on the input speech signal.



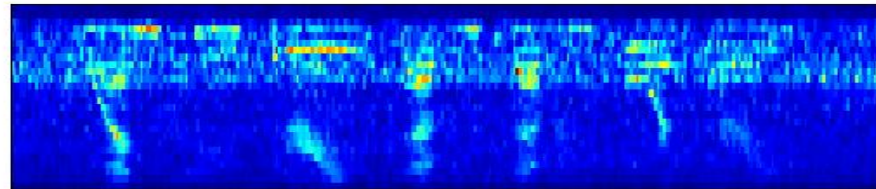
Mask Estimation

- Mask estimation is required for spectral reconstruction.
- Oracle masks, which use the clean signal version, are used to provide a performance bound for MF ASR.
- Masks can also be estimated based on statistical models of speech and noise, eg. SPP masks.

“one two three seven seven four three”



0 dB SNR, vehicular noise



Oracle Mask

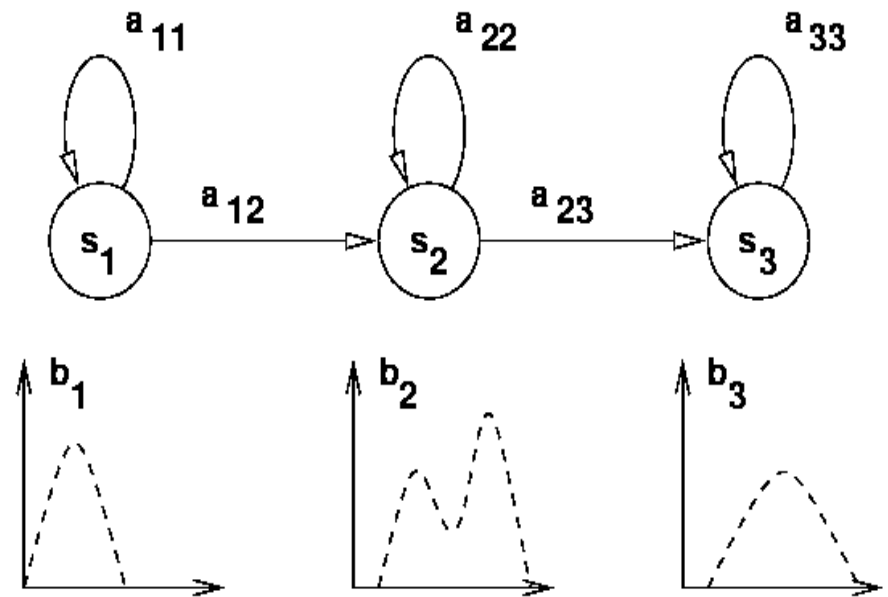


Mask using statistical models

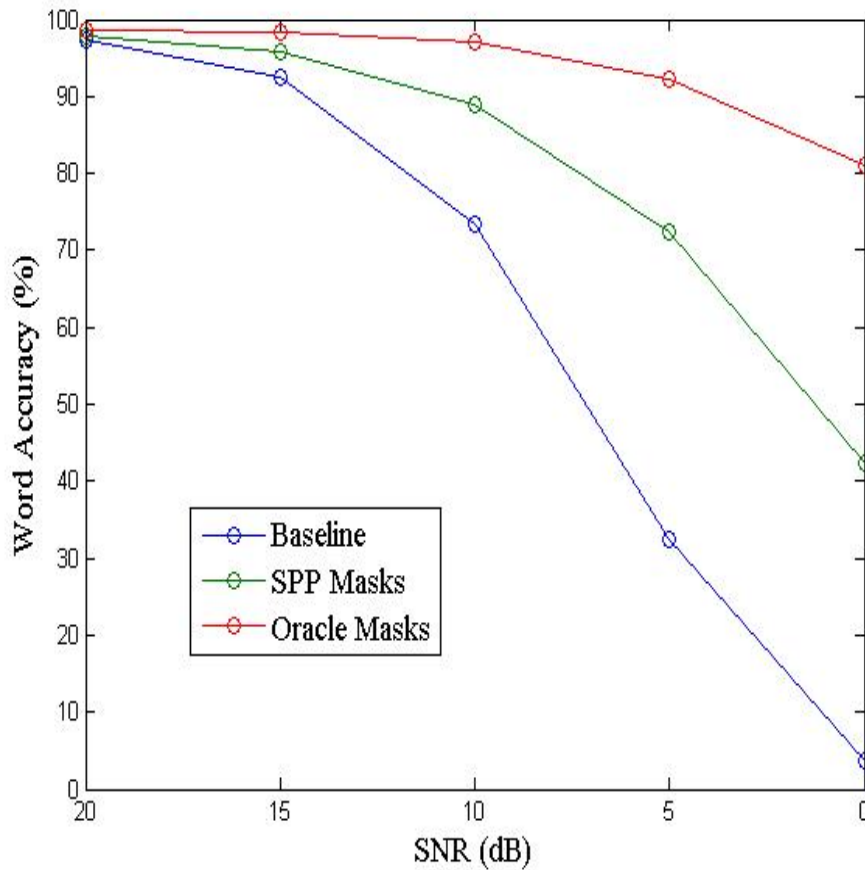


Experimental Methods

- The proposed MF-based ASR system was applied to the Aurora-2 database, comprised of connected digit utterances.
- 8000 files for training, 1000 files for testing at 20 noise conditions/levels each
- 16-state, 3-mixture HMMs were used
- 13 dimensional MFCCs, with 1st and 2nd derivatives (39 dimensional feature vectors)



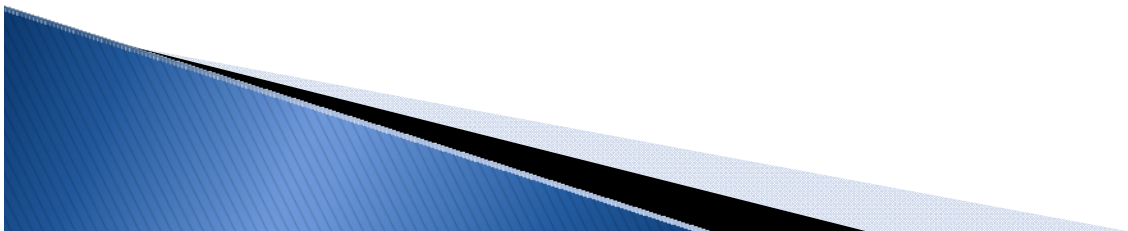
Experimental Results



- The proposed framework was tested on the Aurora-2 database.
- Oracle masks were used to provide an upper performance bound.
- Masks based on Speech Presence Probability (SPP) were also used.
- The proposed framework provides significant performance improvements, both using SPP and oracle masks.

Discussion

- We have explored the compressibility of spectro-temporal speech data
- We have observed the effect of induced sparsity on speech recognition.
- We have derived an algorithm for reconstructing unreliable spectro-temporal components given an incomplete set of reliable observations.
- We have applied our proposed algorithm to missing feature-based data imputation for noise robust ASR, and achieved significant performance improvements.
- Future work includes optimizing statistical reliability masks in hopes of nearing the performance bound provided by use of oracle masks.



References

B. J. Borgstrom and A. Alwan 2009, *Utilizing Compressibility in Reconstructing Spectrographic Data, with Applications to Noise Robust ASR*, IEEE Signal Processing Letters, Vol. 16, Issue 5, pp. 398-401.

B. Raj, M. L. Seltzer, and R. M. Stern 2004, *Reconstruction of Missing Features for Robust Speech Recognition*, *Speech Communication*, vol. 43, pp. 275-296.

D. L. Donoho 2006, *Compressed Sensing*, IEEE Trans. on Information Theory, Vol. 52, No. 4, pp. 1289-1306.

E. J. Candes and M. B. Wakin 2008, *An Introduction to Compressive Sampling*, IEEE Signal Processing Magazine, Vol. 25, No. 2, pp. 21-20.

S. Boyd and L. Vandenberghe 2004, *Convex Optimization*, Cambridge University Press.

