Improving Far-field ASR Using Low-rank and Sparse Models

Pranay Dighe
Afsaneh Asaei, Hervé Bourlard

Outline

- Speech Enhancement using Low-rank and Sparse Models
  - Motivation
  - Approaches:
    - Enhancement using PCA
    - Enhancement using Dictionary Learning & Sparse Recovery
  - Multi-task Joint Dereverb. Acoustic Modeling Training
- Conclusions

Database

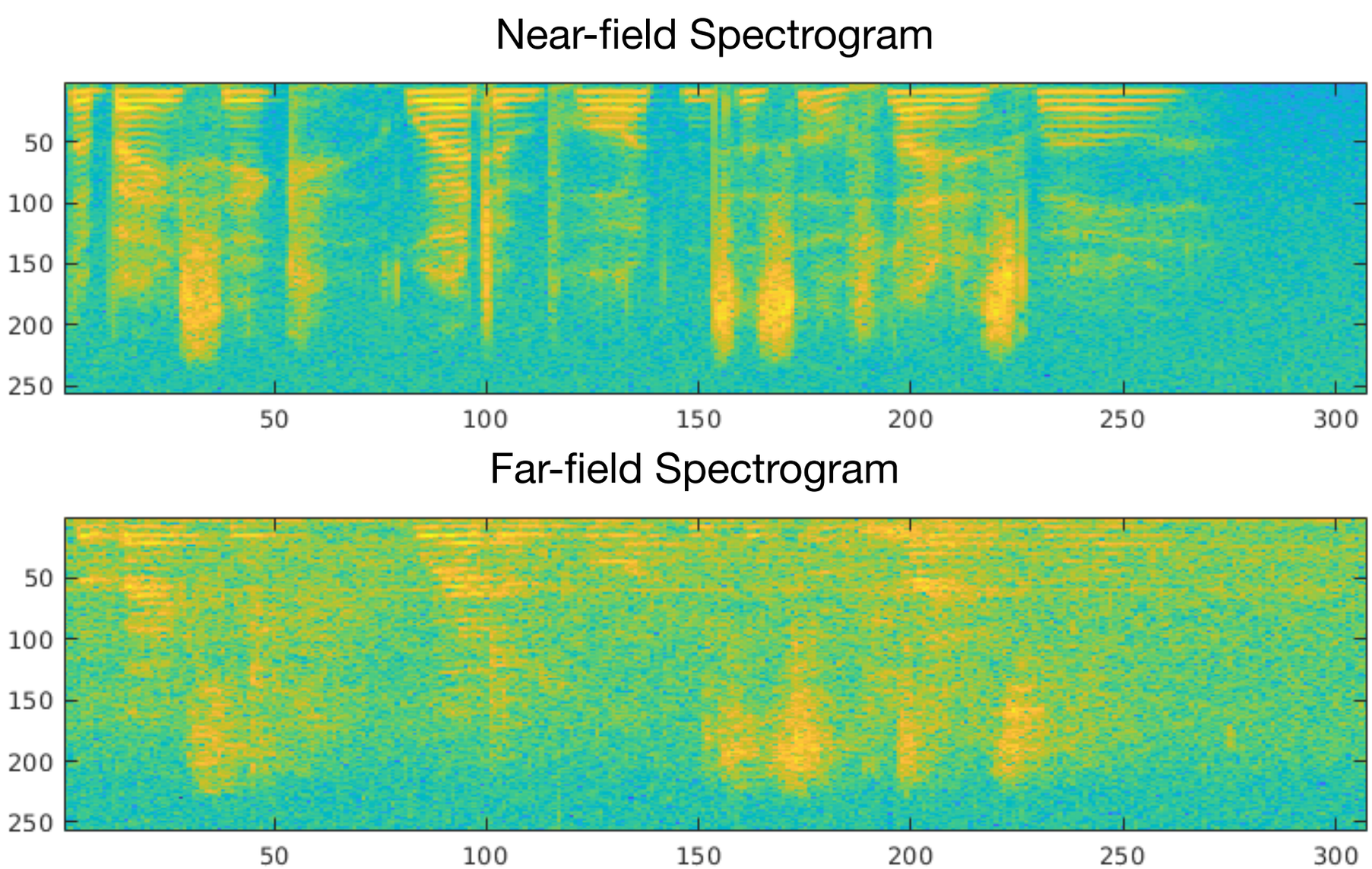
AMI Meeting Corpus

- Conversational Speech in Meeting Scenario
- Training: 80 hours
- Dev: 8 hours
- Test: 8 hours
- Parallely recorded far-field data

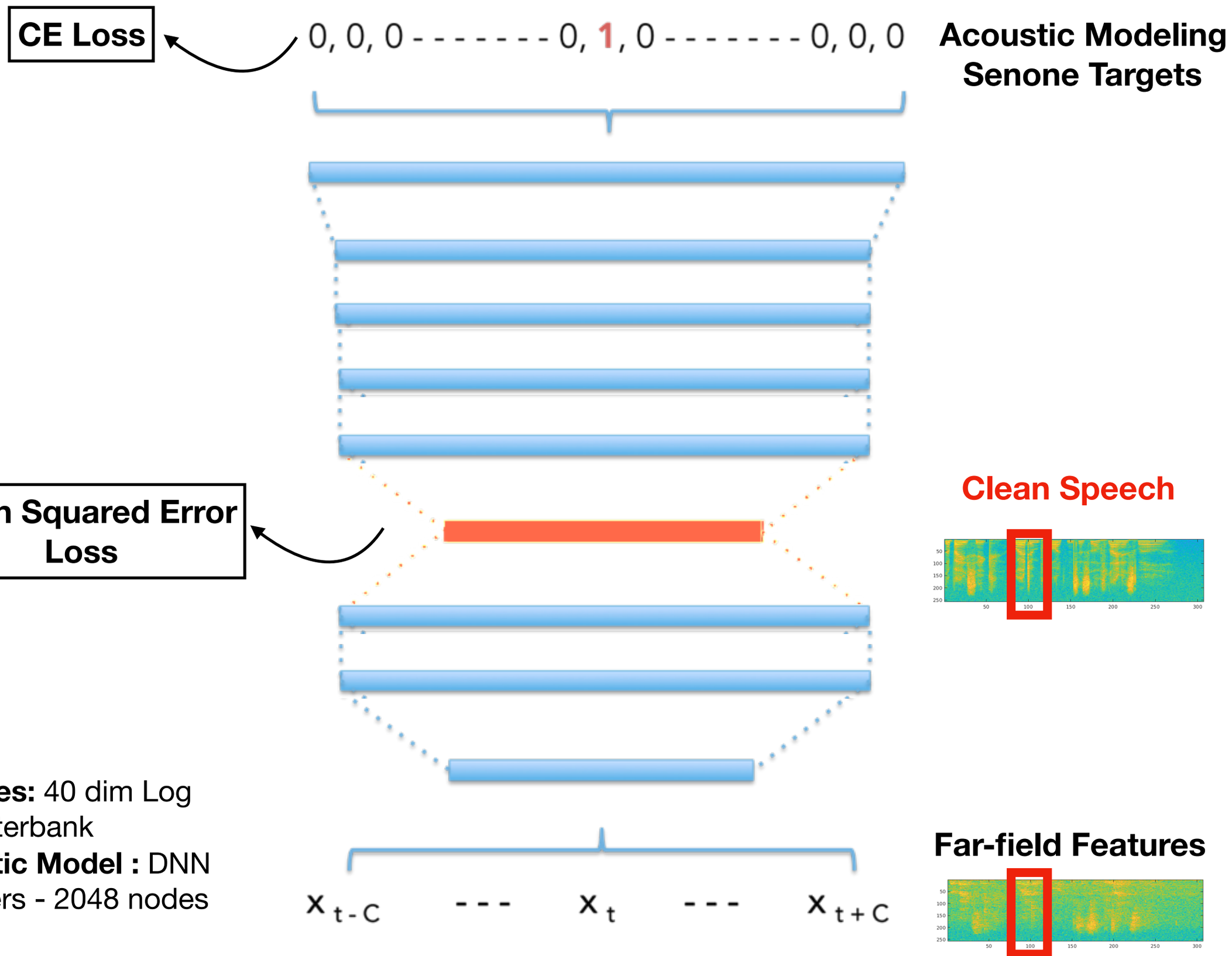
Near-field Condition: IHM (Individual Headset Mic.)
Far-field Condition: SDM (Single Distant Mic.)
Usually Very High WERs ~50%

Features for Speech Enhancement:
40 dimensional Log Mel Filterbank Features
Acoustic Model : DNN - 6 layers - 2048 nodes

Speech Enhancement

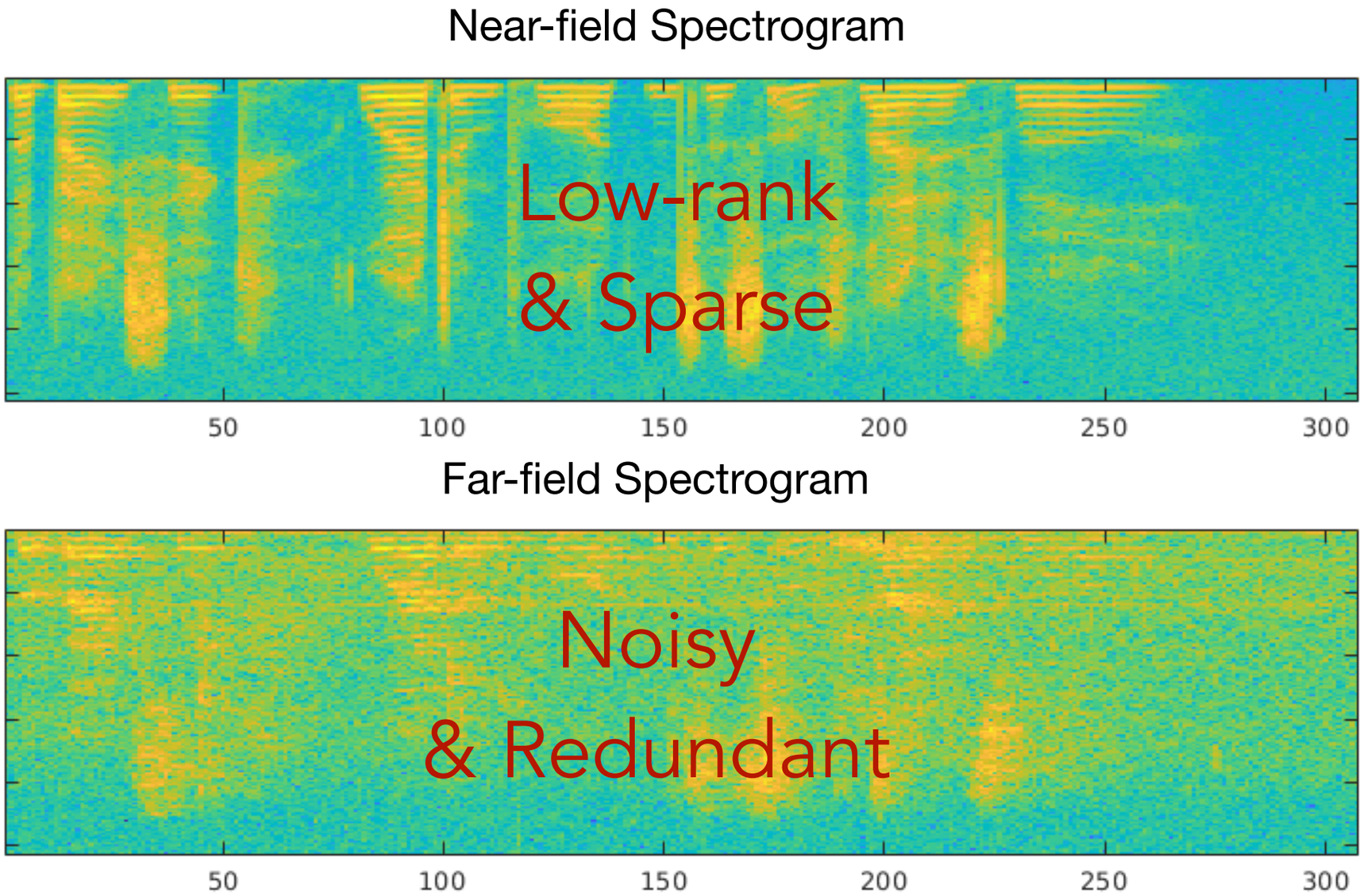


Joint Dereverb. and Acoustic Modeling



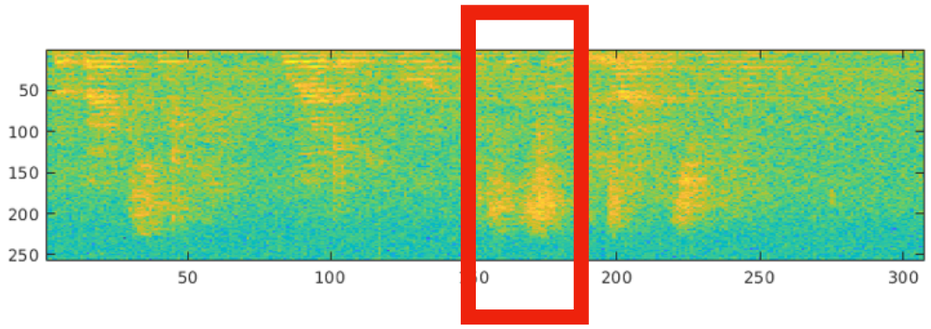
Features: 40 dim Log Mel-Filterbank
Acoustic Model : DNN - 6 layers - 2048 nodes

Motivation

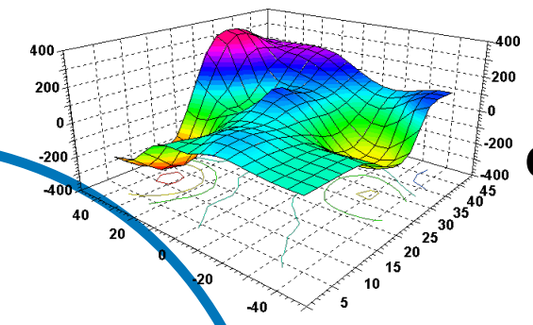


- **Similar acoustic information**

Speech Enhancement Using Low-rank & Sparse Models

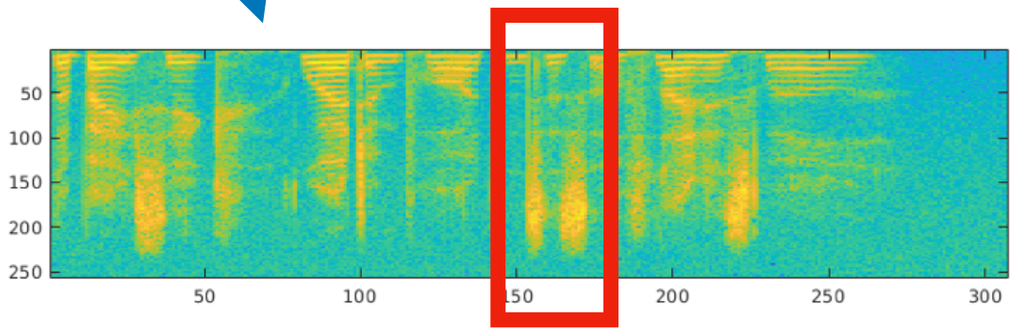


Far-field Features



Low-rank or Sparse Projection on Close-talk Speech

Learn Low-dimensional Subspace For Each Senones/Phones Separately

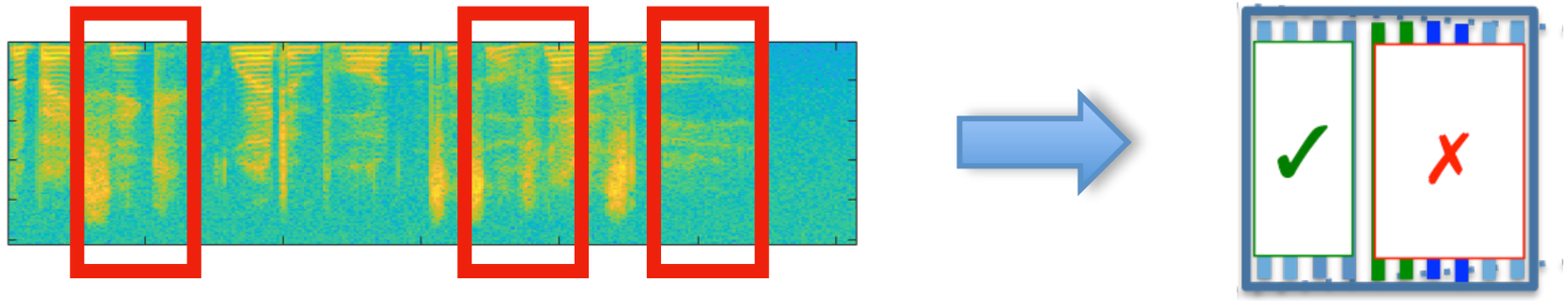


Projected Features (Estimate of Clean Near-field Speech)

- Dighe et al. "Exploiting low-dimensional structures to enhance dnn based acoustic modeling in speech recognition *ICASSP 2016*
- Dighe et al. “Sparse modeling of neural network posterior probabilities for exemplar-based speech recognition." *Speech Communication* 76 (2016): 230-244.

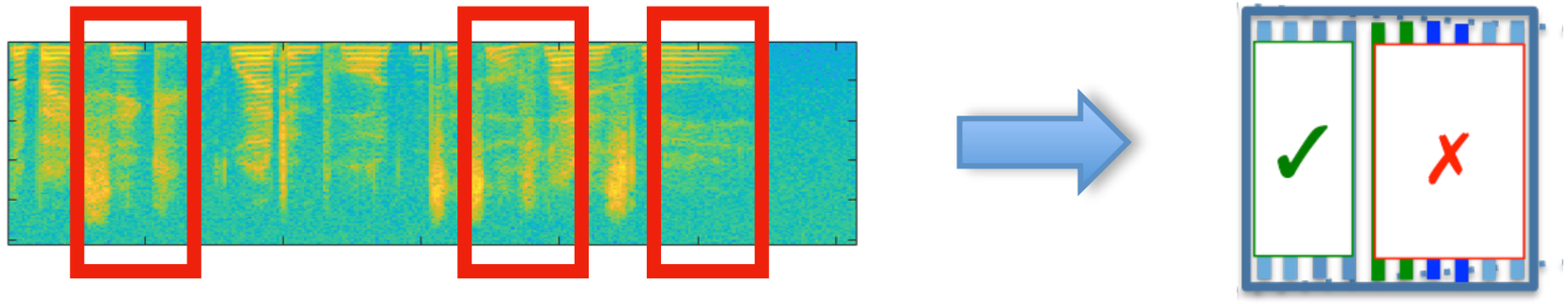
Approach 1: PCA

1. Learn PC Matrix from IHM Data

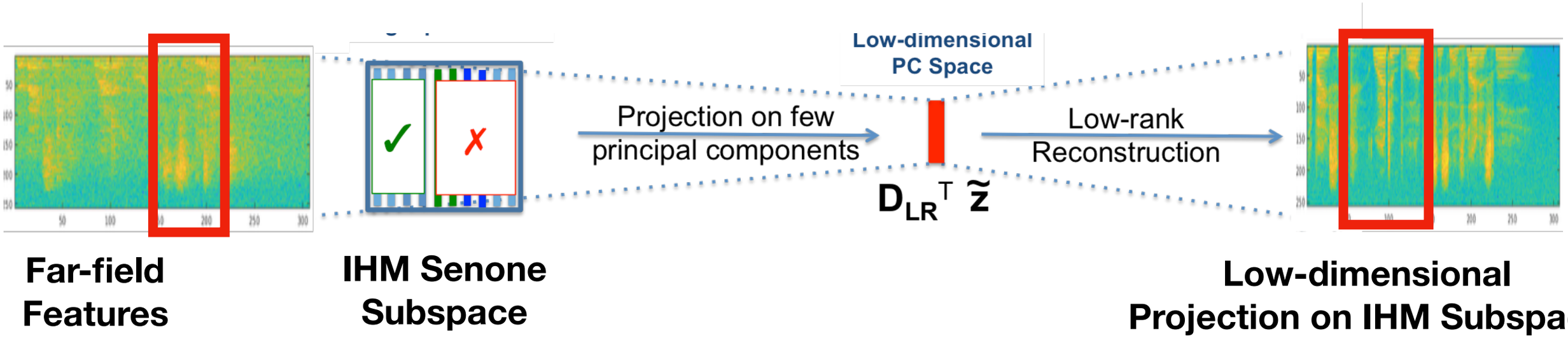


Approach 1: PCA

1. Learn PC Matrix from IHM Data



2. Project SDM Data on IHM Data (Uses GMM-HMM Alignments)



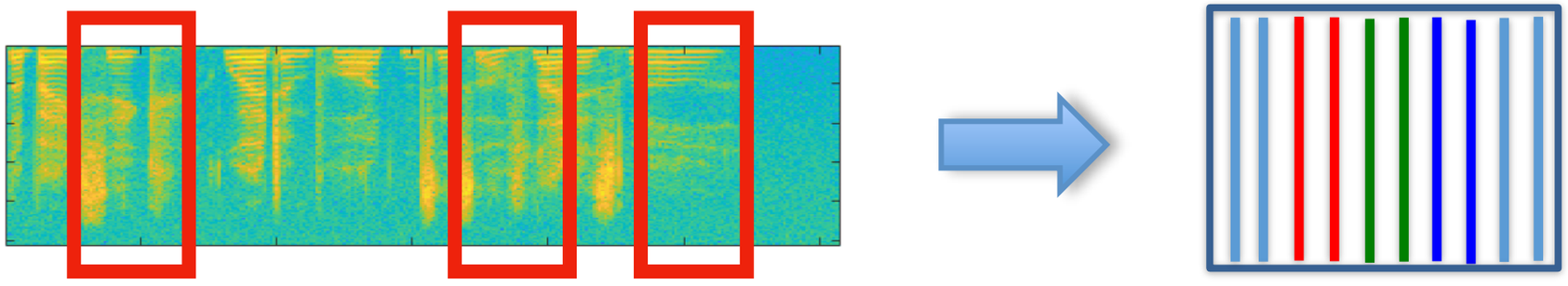
Select PCs which capture σ % variability

σ = 100% results in exact reconstruction

σ < 100% results in **preserving global patterns** and **discarding local errors**

Approach 2: Dictionary Learning +Sparse Recovery

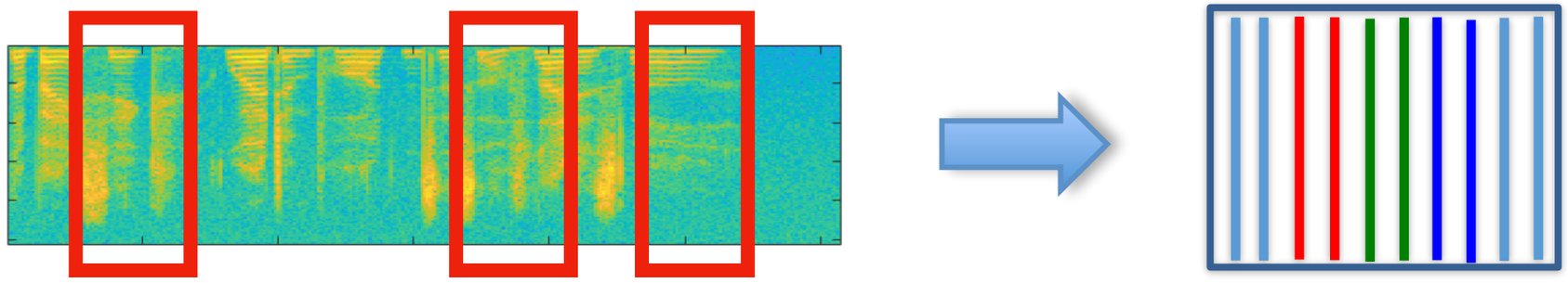
1. **Learn Dictionary from IHM Data**



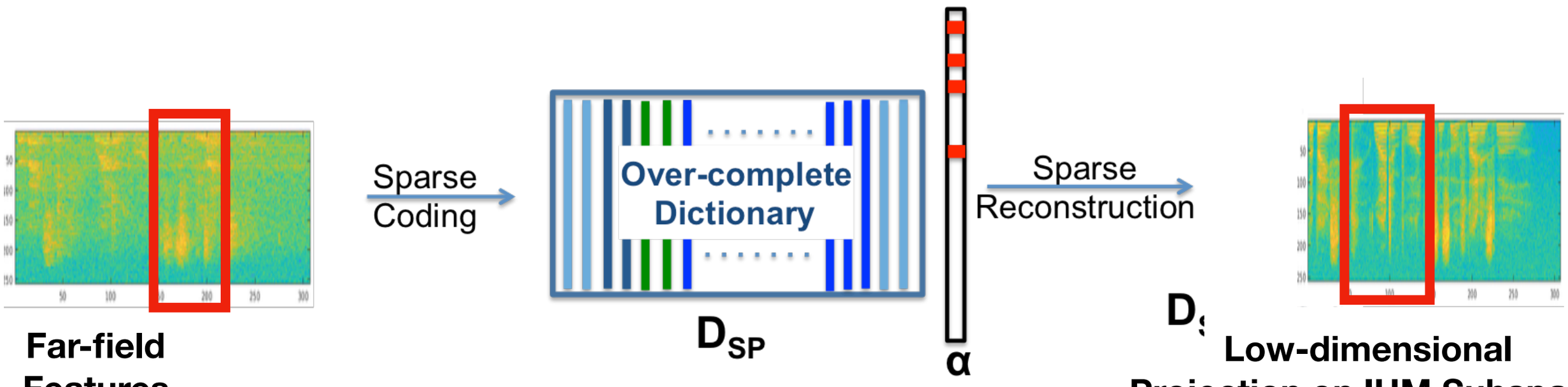
Over-complete Dictionary

Approach 2: Dictionary Learning +Sparse Recovery

1. **Learn Dictionary from IHM Data**



2. **Project SDM Data on IHM Data (Uses GMM-HMM Alignments)**

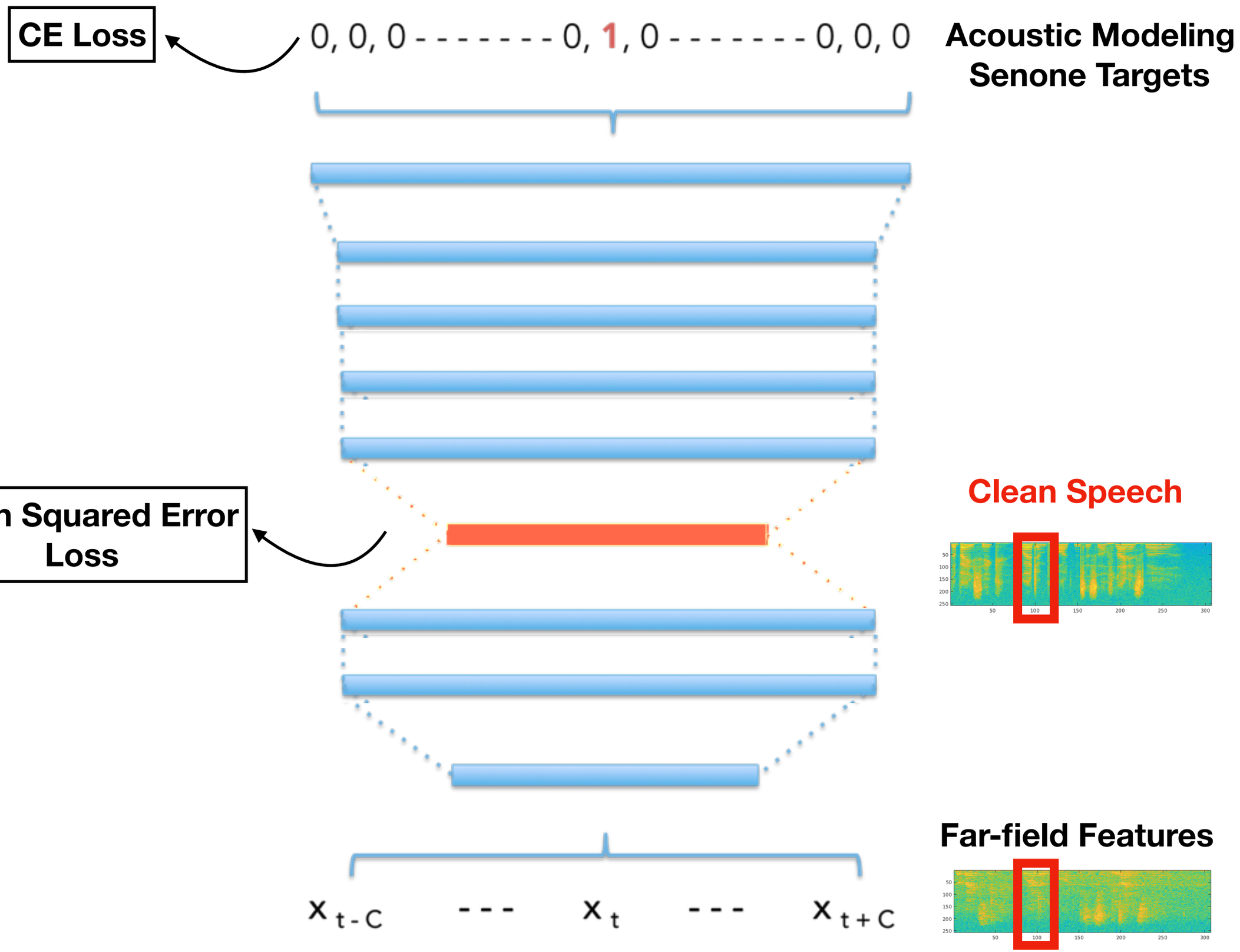


$$\min_{\alpha} \quad \frac{1}{2}\|z_{\square} - \mathbf{D}\,\alpha_{\square}\|_2^2 + \lambda\|\alpha_{\square}\|_1$$

Select λ such that sparsity discards **unwanted noise/local errors** but preserves **global patterns**

Lasso [Tibishirani 1996]
(Least Absolute Shrinkage and Selection Operator)

Joint Dereverb. and Acoustic Modeling

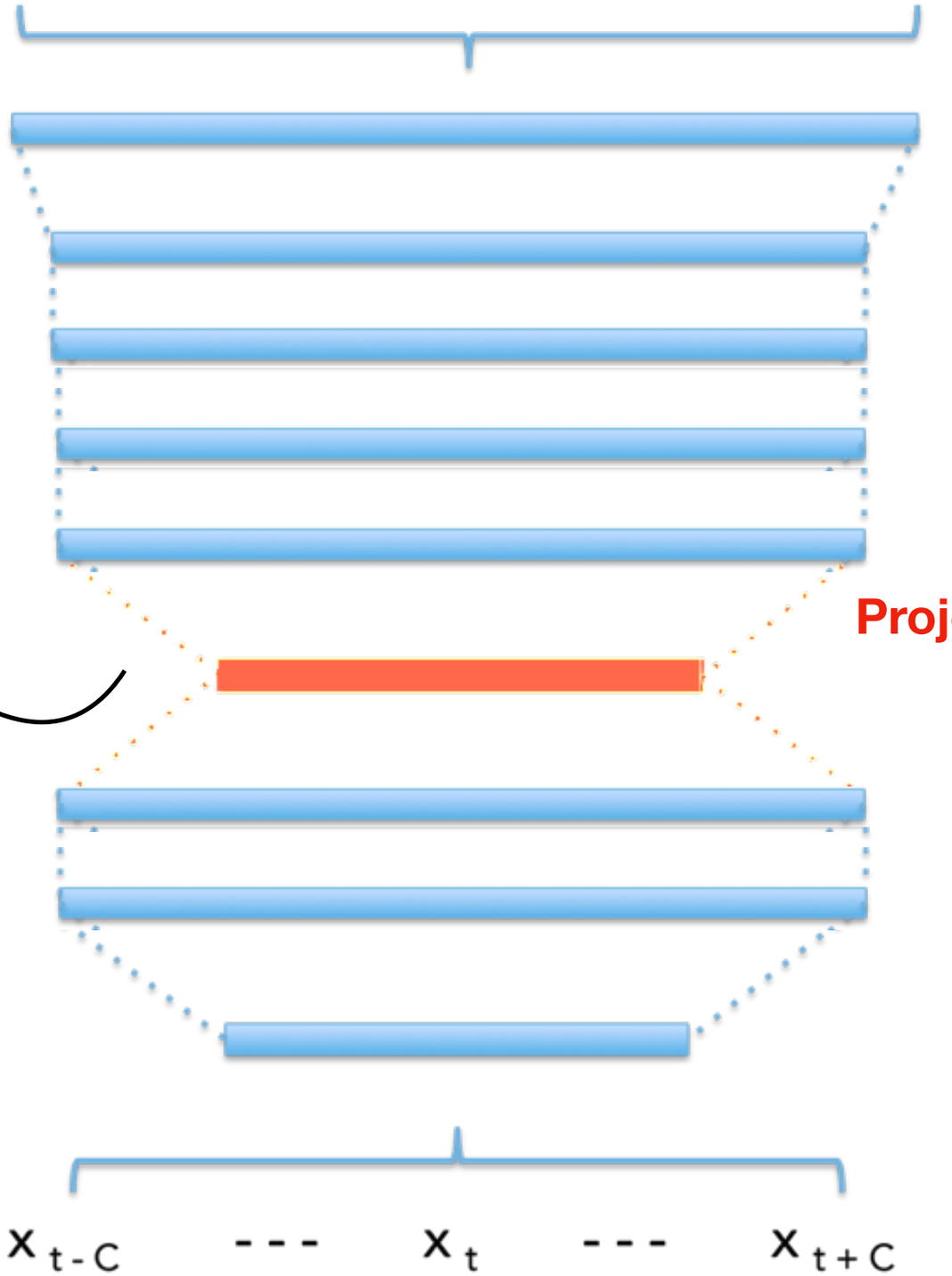


Joint Dereverb. and Acoustic Modeling

CE Loss

0, 0, 0 - - - - - - - 0, **1**, 0 - - - - - - - 0, 0, 0

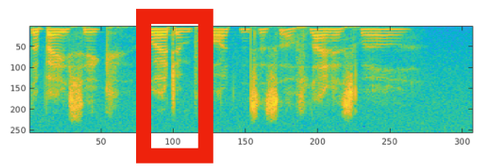
Acoustic Modeling Senone Targets

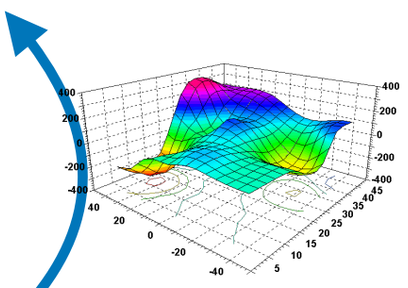


Mean Squared Error Loss

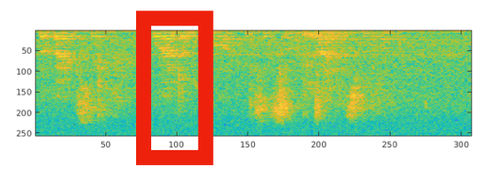
x_{t-C} - - - x_t - - - x_{t+C}

Low-dimensional Projection on IHM Subspace

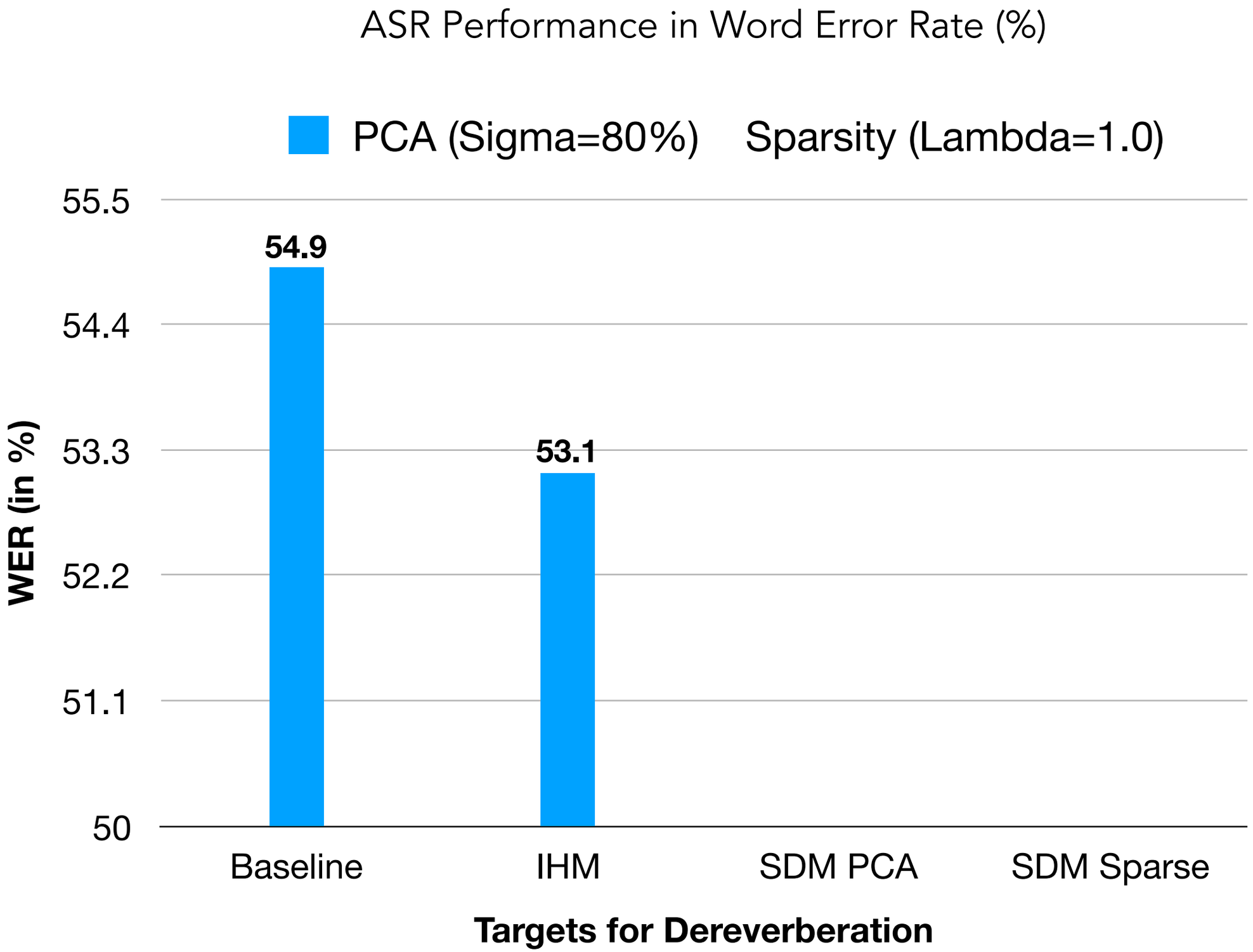




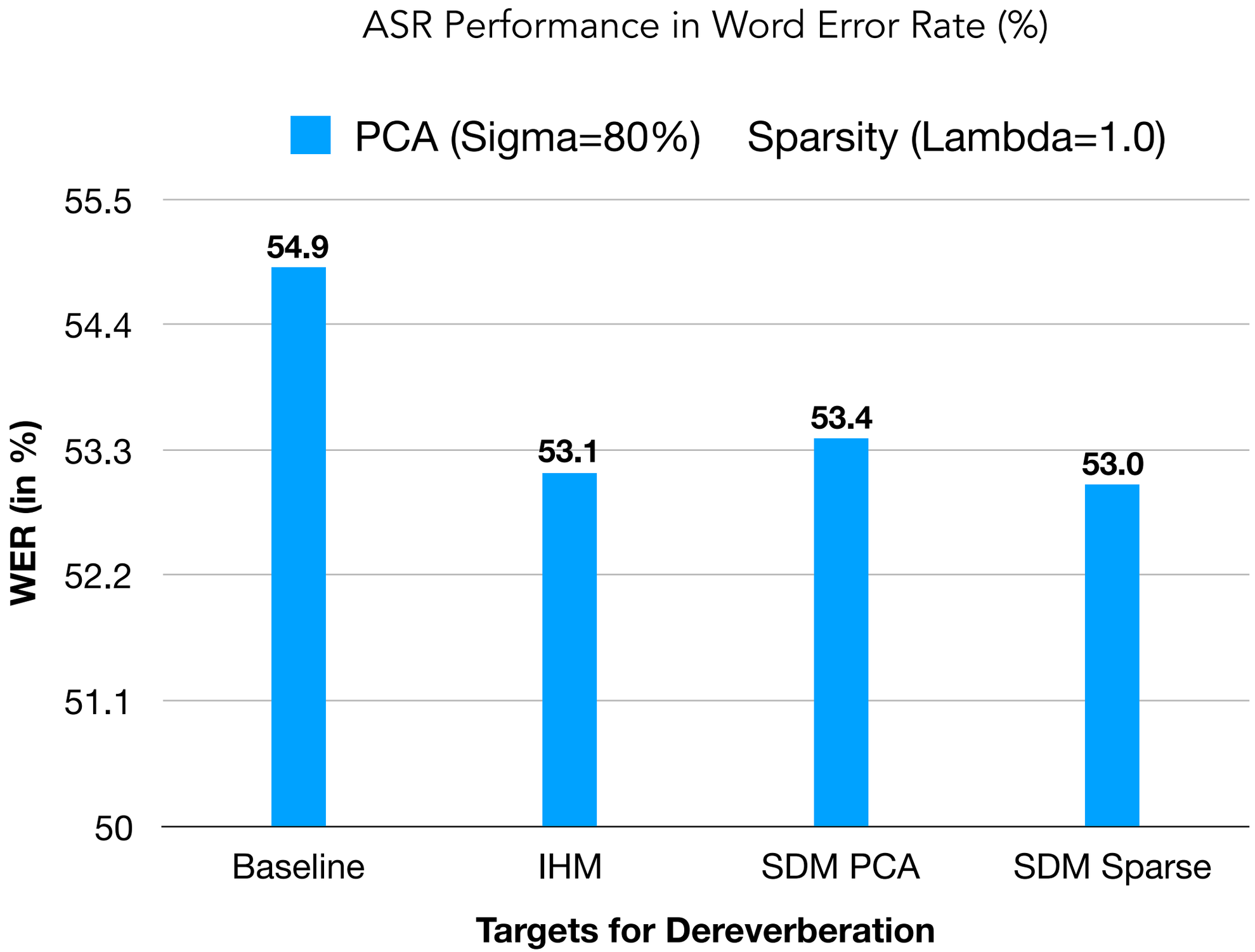
Far-field Features



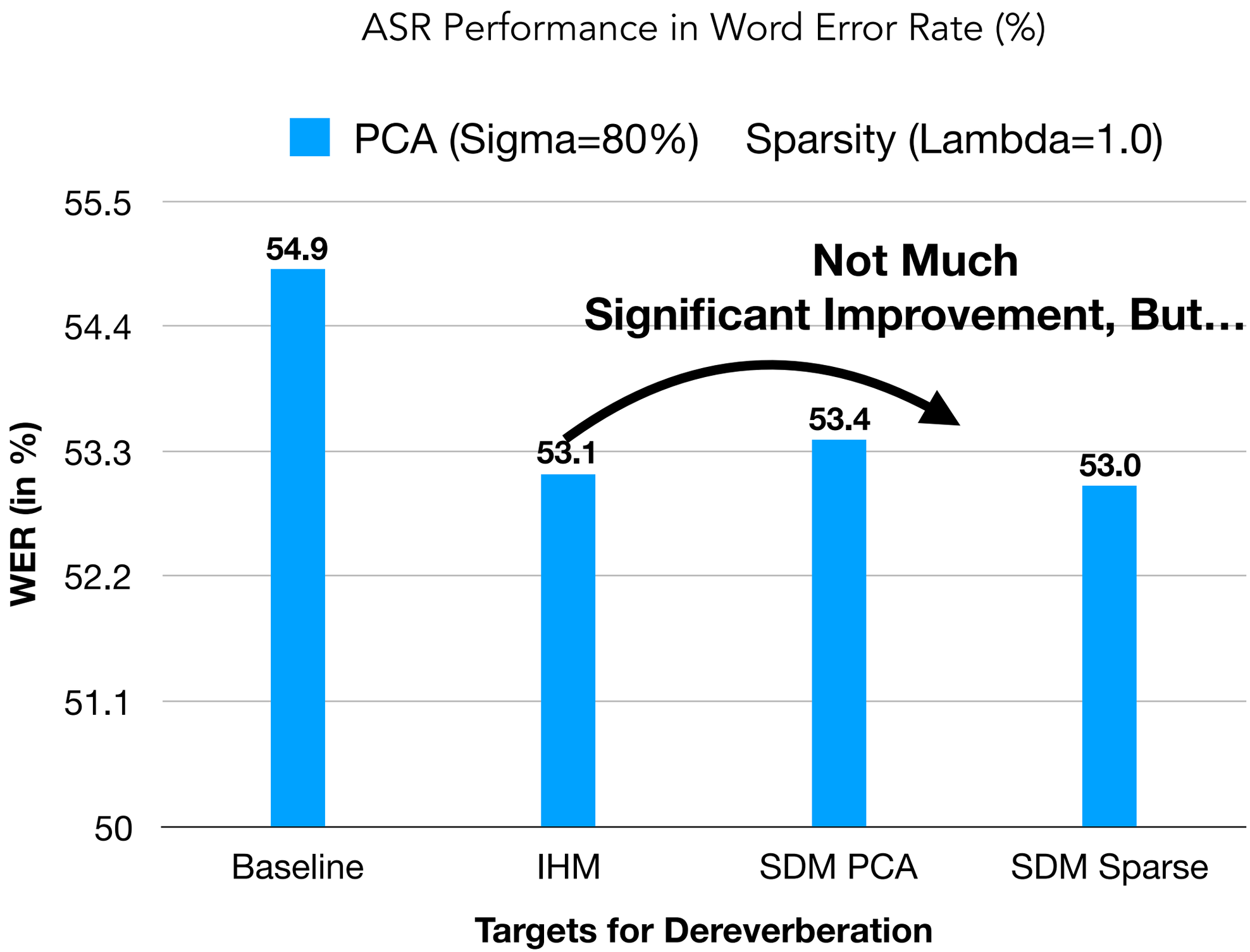
Results
Far-field Mic Speech ASR



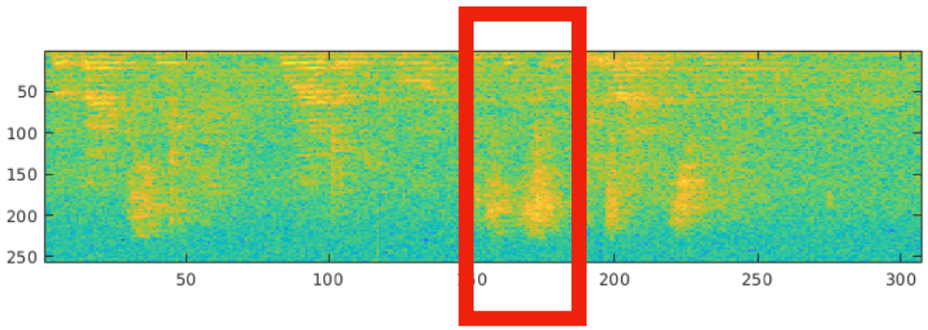
Results
Far-field Mic Speech ASR



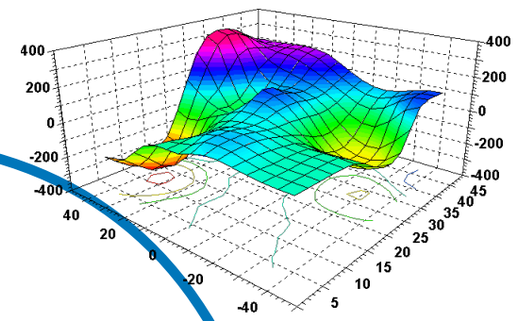
Results
Far-field Mic Speech ASR



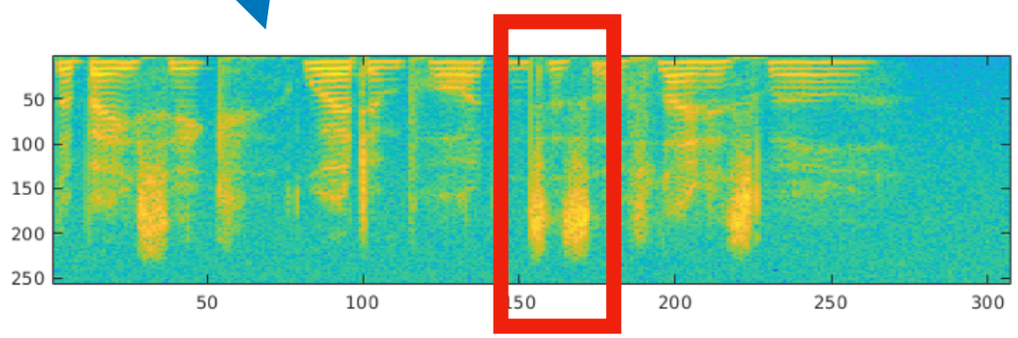
Joint Enhancement and Acoustic Modeling



Far-field Features



Low-rank or Sparse Projection on Close-talk Speech



Projected Features

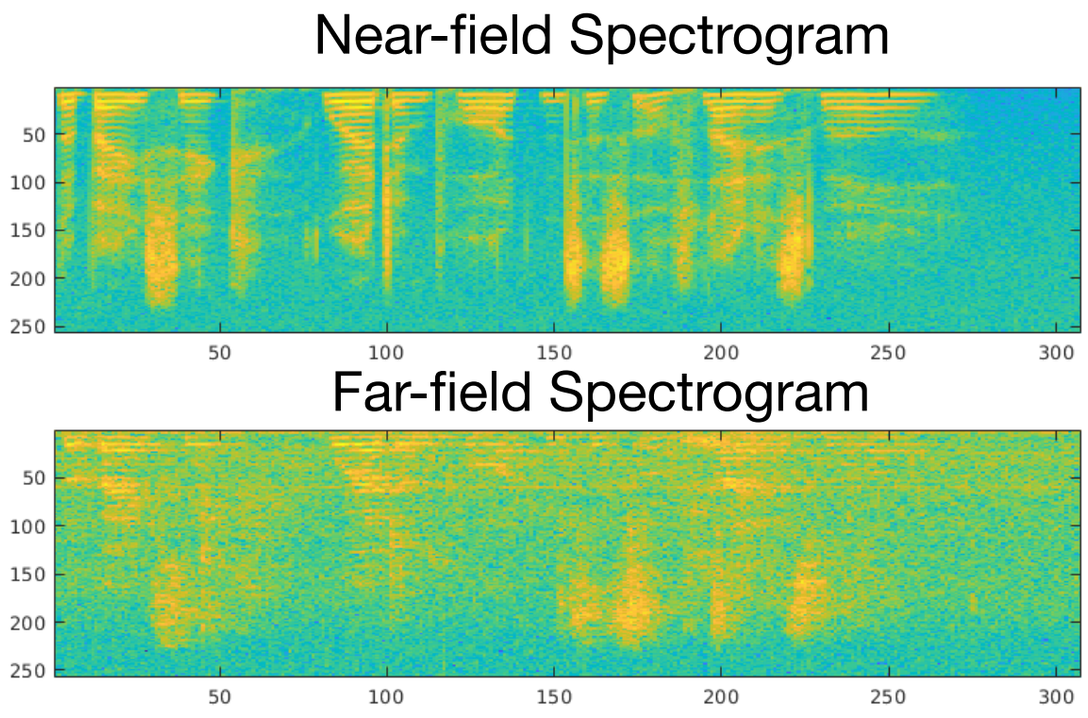
✓ **Actual Near-field/Clean Speech Not Needed**

✓ **Classwise PC/Dictionary Can Capture IHM Subspace Projection can act as Enhancement Targets**

✗ **Enhancement was supervised as clean speech alignments were still used**

Conclusions

- Low-rank and sparse transformations to map far-field feature to near-field features.



- Far-field speech enhanced using low-rank and sparse models acts as good target for speech enhancement

- **Future Work:** Explore Low-dimensional Modeling Approaches which do not need Classwise-modeling of Subspaces.

Thank You