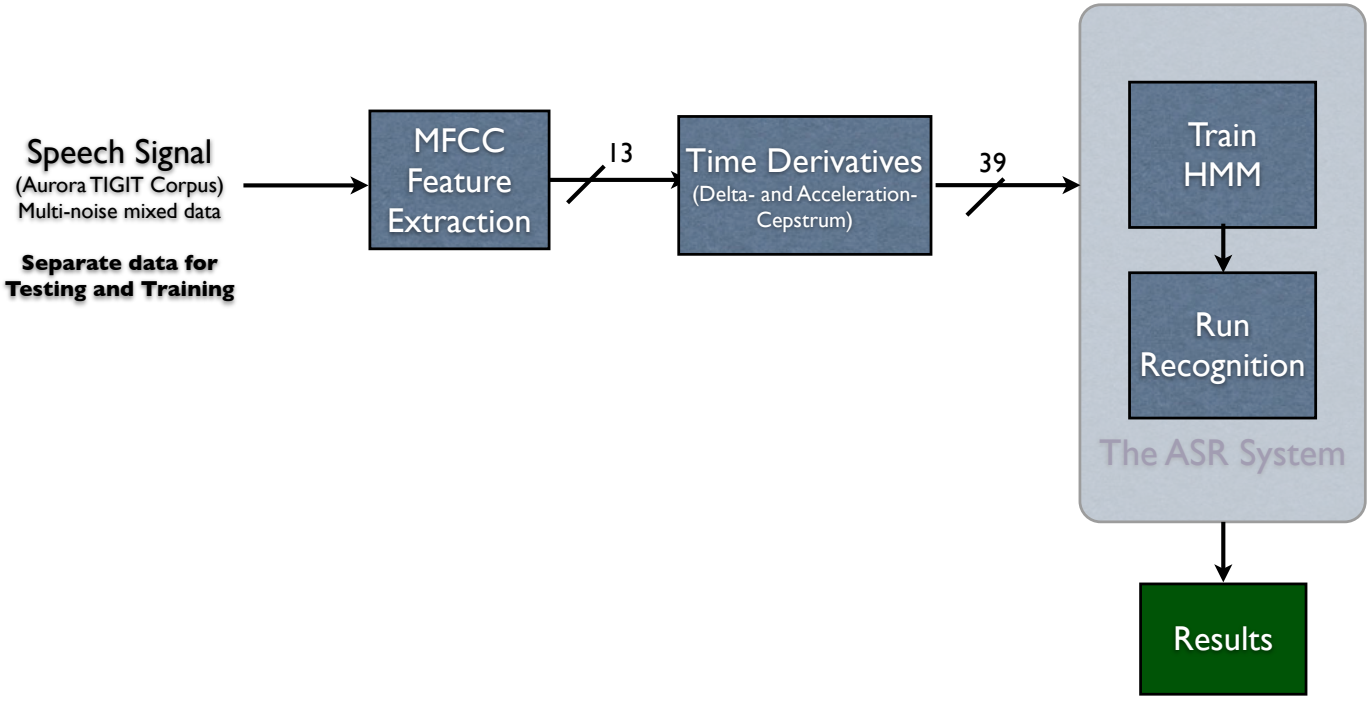
Discriminant Feature Space Transformation for Automatic Speech Recognition

Vikrant Tomar

Department of Electrical & Computer Engineering
McGill University
Montreal, Canada

December 15, 2010

Traditional ASR– The Big Picture



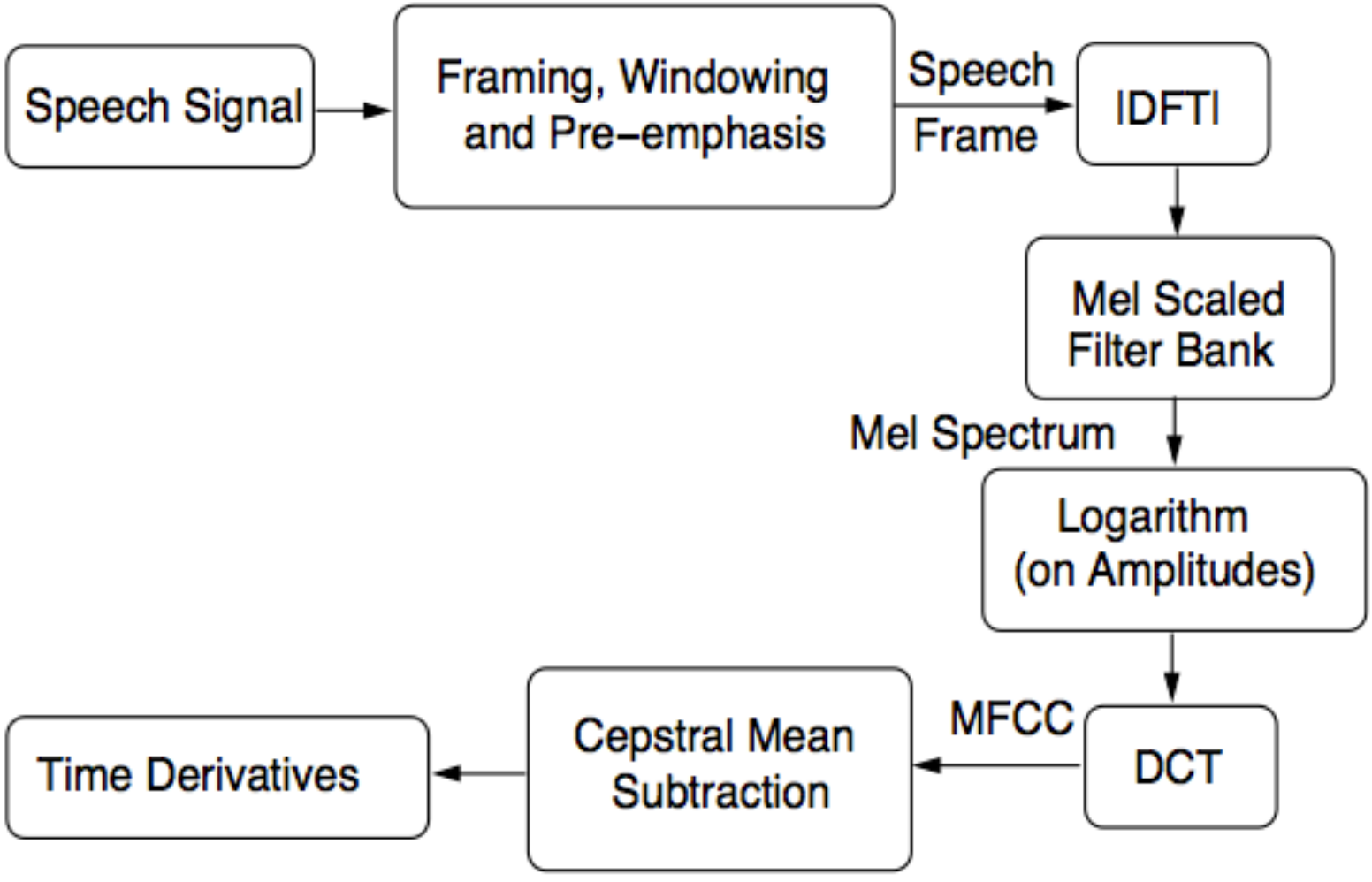
ASR – A Machine Learning Perspective

- Objective: Find the *optimal* sequence of words $\hat{\mathbf{W}}$ from all possible word sequences $\mathbf{W}$ that yields the highest probability given acoustic feature data $\mathbf{X}$ [3].

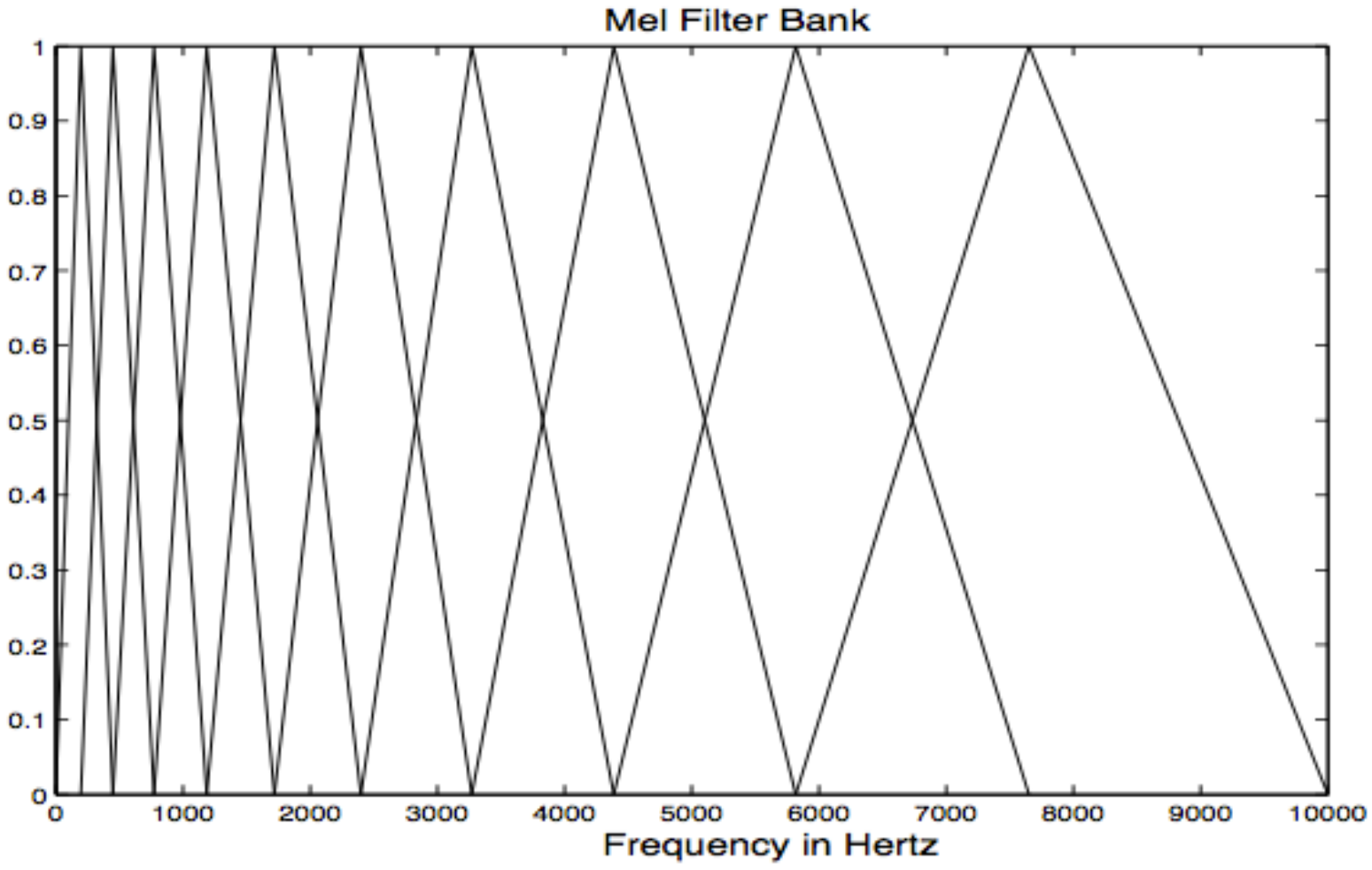
$$\begin{aligned}\hat{\mathbf{W}} &= \arg\max_{\mathbf{W}} p(\mathbf{W}|\mathbf{X}) \\ &= \arg\max_{\mathbf{W}} \frac{p(\mathbf{X}|\mathbf{W})p(\mathbf{W})}{P(\mathbf{X})} \\ &= \arg\max_{\mathbf{W}} \left\{ \overbrace{p(\mathbf{X}|\mathbf{W})}^{acoustic\ \ model} \times \underbrace{p(\mathbf{W})}_{language\ \ model} \right\}\end{aligned}$$

- $p(\mathbf{X}|\mathbf{W})$: probability of observing $\mathbf{X}$ under the assumption that $\mathbf{W}$ is a true utterance.
- $p(\mathbf{W})$: probability that word sequence $\mathbf{W}$ is uttered.
- $p(\mathbf{X})$: average probability that $\mathbf{X}$ will be observed.

MFCC Feature Extraction



Mel-filter bank



Problem Statement

- MFCC cepstrum gives us a fairly accurate static information in speech. However, we are also interested in the time-evolution of speech.
- For the same, the static Cepstrum vector is augmented with Δ-, and $\Delta\Delta$-Cepstrum features. But ...
  - The components of Augmented Feature vector are strongly correlated [5].
  - The obtained features are not necessarily the most parsimonious representation for capturing dynamics of speech.
  - It's not even clear that the linear operator (discrete cosine transform) used in the final step of Cepstrum coefficient extraction, is an optimal choice.

Problem Statement

- MFCC cepstrum gives us a fairly accurate static information in speech. However, we are also interested in the time-evolution of speech.
- For the same, the static Cepstrum vector is augmented with Δ-, and $\Delta\Delta$-Cepstrum features. But ...
  - The components of Augmented Feature vector are strongly correlated [5].
  - The obtained features are not necessarily the most parsimonious representation for capturing dynamics of speech.
  - It's not even clear that the linear operator (discrete cosine transform) used in the final step of Cepstrum coefficient extraction, is an optimal choice.
- To avoid that, LDA based methods of capturing dynamics of speech have been suggested. [1, 4].

ASR as a classification problem

- Divide the speech feature vectors into various classes, and apply class based discrimination algorithms.
  - The best definition of classes is not clear.
  - Once can use words, phones, HMM states or some other arbitrarily specified notion of class.

ASR as a classification problem

- Divide the speech feature vectors into various classes, and apply class based discrimination algorithms.
  - The best definition of classes is not clear.
  - Once can use words, phones, HMM states or some other arbitrarily specified notion of class.
- Use Linear Discriminant Analysis (LDA) for Discriminant feature-space transformation on an augmented super vector (dim. $117 \rightarrow 39$) .
- Advantages:
  - a proper mathematical basis for extracting information about the time evolution of speech frames
  - Maximizes the separability of different classes (however defined).

LDA sounds good, but...

- ASR models are often trained with the assumption that the data has diagonal covariance.
- LDA produces a projected space whose dimensions might be highly correlated (full covariance matrix).

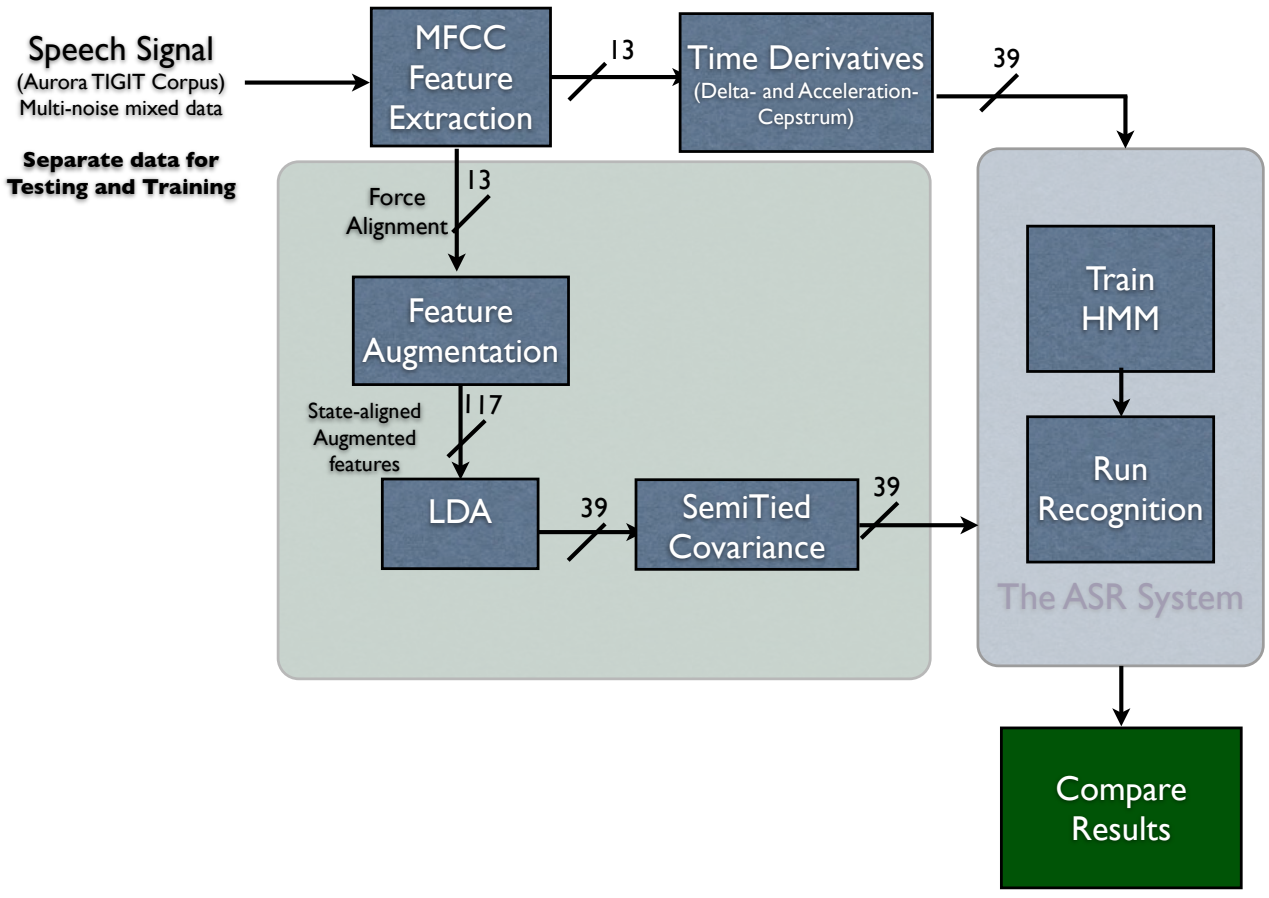
LDA sounds good, but...

- ASR models are often trained with the assumption that the data has diagonal covariance.
- LDA produces a projected space whose dimensions might be highly correlated (full covariance matrix).
- To tackle that we can use separate covariance matrix of each class. Or,
- Semitied Covariance Transform [2].

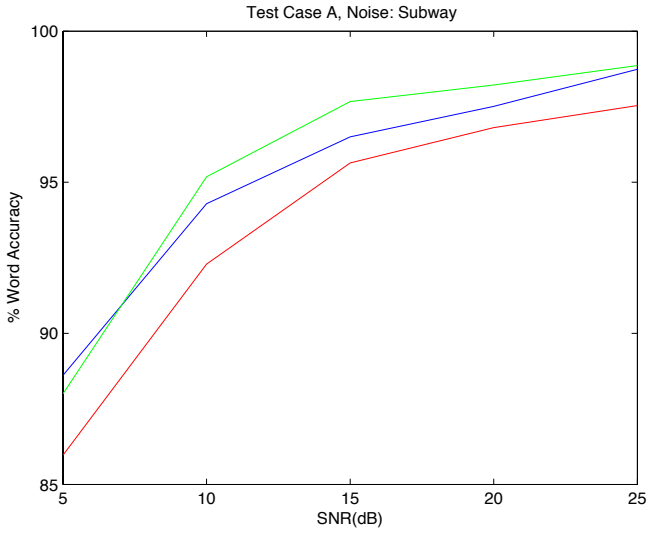
Semitied Covariance Transform

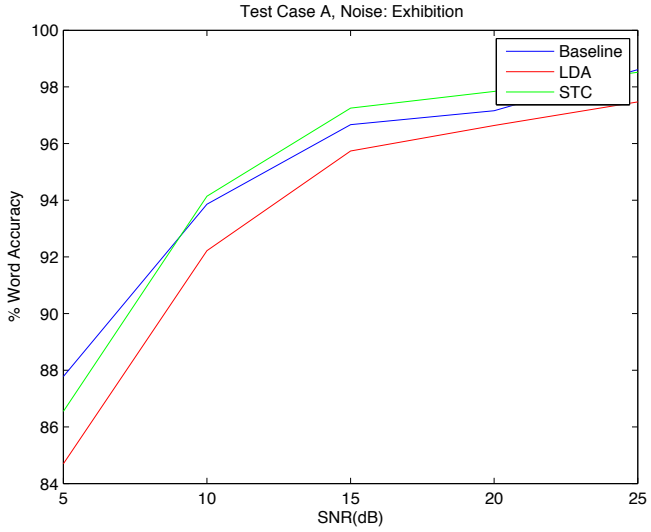
- Instead of having a distinct covariance matrix for every component in the recognizer, each covariance matrix consists of two elements:
  - a component specific diagonal covariance element
  - a semi-tied class-dependent, non-diagonal matrix
- Basically, transforms LDA's full covariance matrix into a semi-diagonal matrix, which results in increased ASR performance.

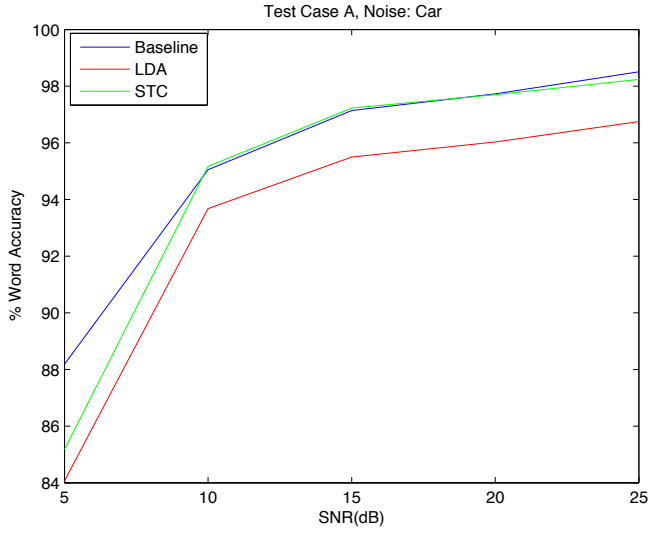
LDA based ASR– The Big Picture

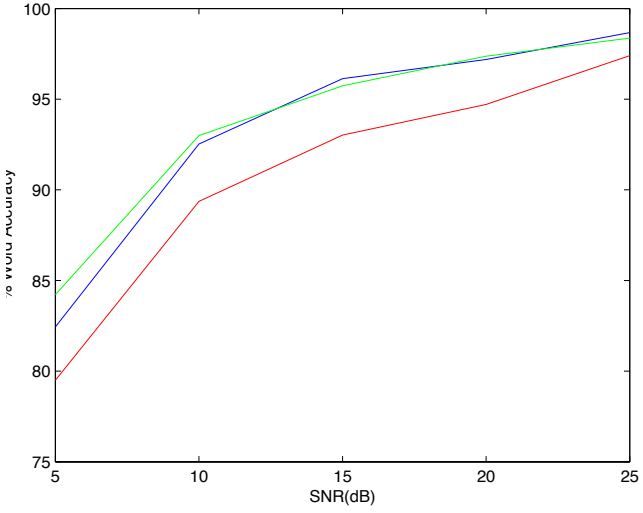


Results-Test Case A

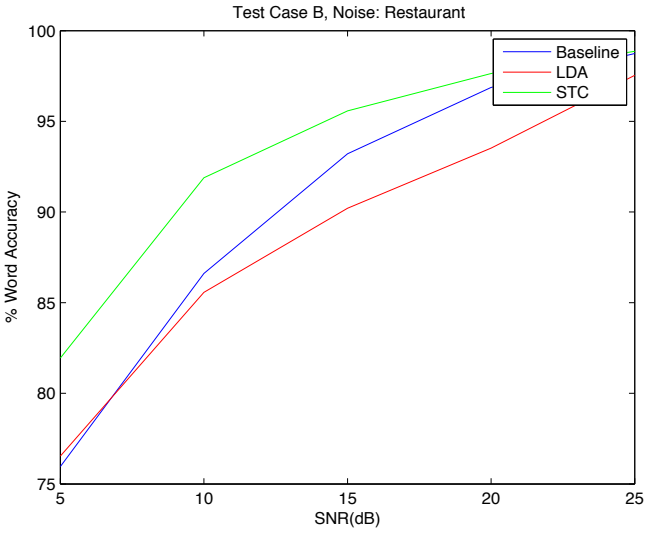


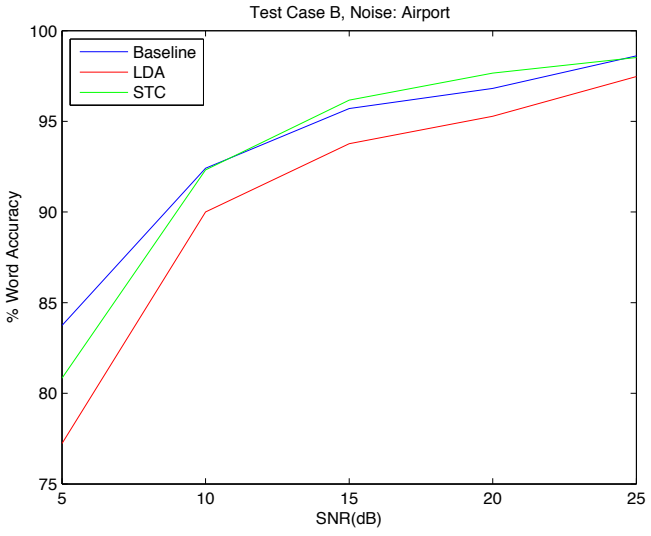


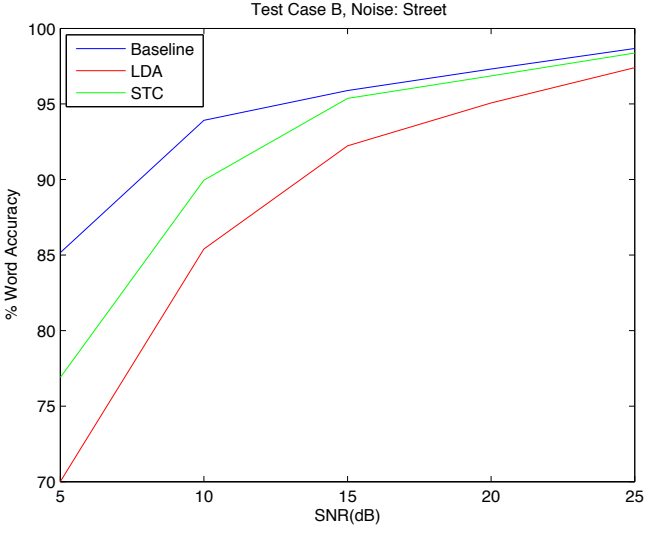


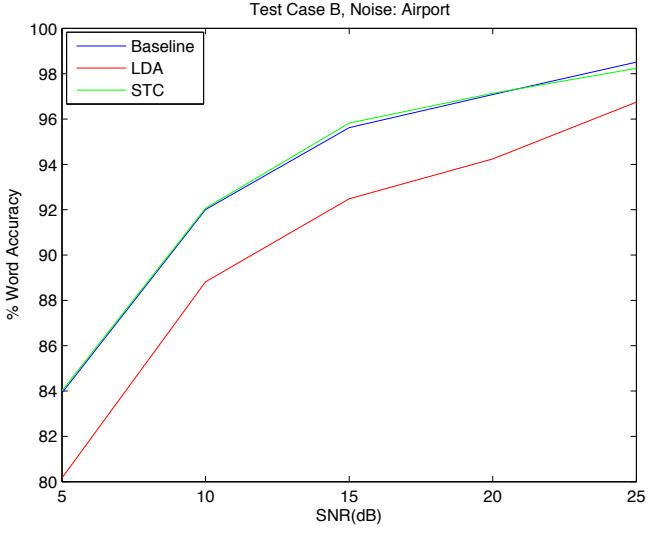


Results-Test Case B

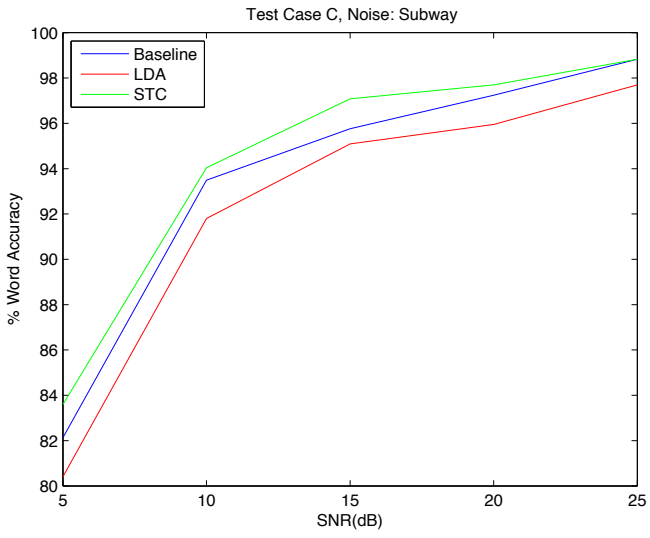


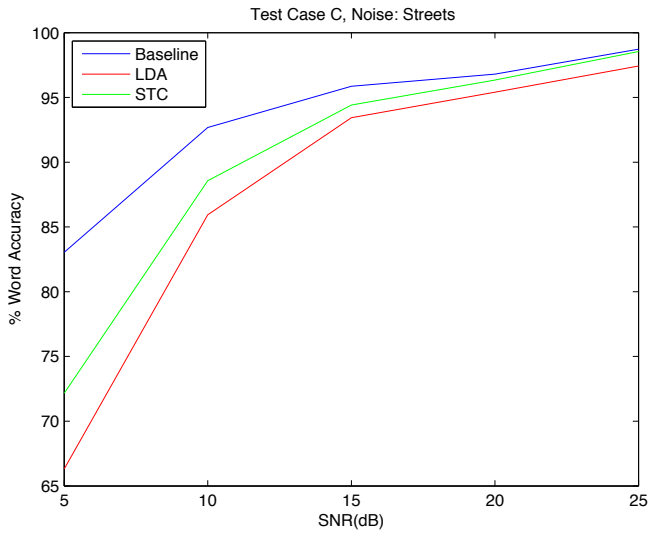






Results-Test Case C





References

P. F. Brown.

The Acoustic-Modeling Problem in Automatic Speech Recognition.
PhD thesis, Carnegie Mellon University, May 1987.

Mark J. F. Gales.

Semi-tied covariance matrices for hidden markov models.
IEEE Transactions on Speech and Audio Processing, 7(3):272 – 281, May 1999.

Xuedong Huang, Alex Acero, and Hsiao-Weun Hon.

Spoken Language Processing.
Prentice Hall PTR, 2001.

G. Saon, M. Padmanabhan, R. Gopinath, and S. Chen.

Maximum likelihood discriminant feature spaces.
Technical report, IBM T. J. Watson Research Center, 2000.

Yun Tang and Richard Rose.

A study of using locality preserving projections for feature extraction in speech recognition.
In *ICASSP: IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2008.