Using LSF Features for Speaker Verification in Noise

Pujita Raman and Dr. A. A. (Louis) Beex
DSP Research Laboratory
Wireless @ Virginia Tech

IEEE Global Conference on Signal and Information Processing
Tuesday, December 15th, 2015
Orlando, Florida

Outline

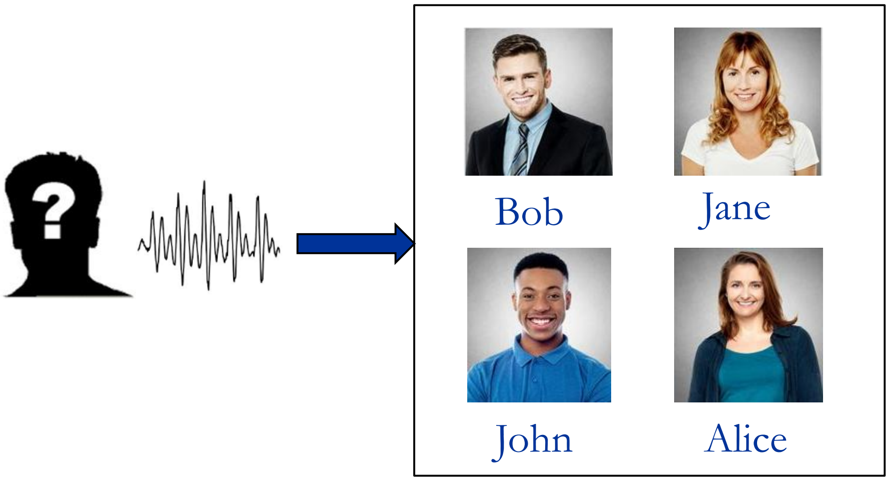
- **Introduction**
- Motivation
- Feature Extraction
- Speaker Verification Framework
- Speaker Verification Results
- Conclusions

Speaker Recognition

The task of determining a speaker's identity using information extracted from his/her voice.

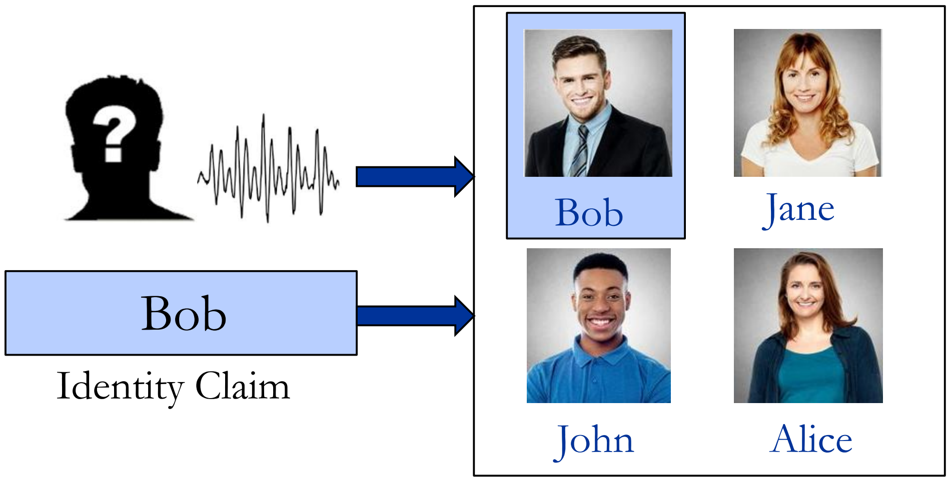
Speaker Recognition

- Speaker Identification
- Speaker Verification



Speaker Database

Whose voice is this?

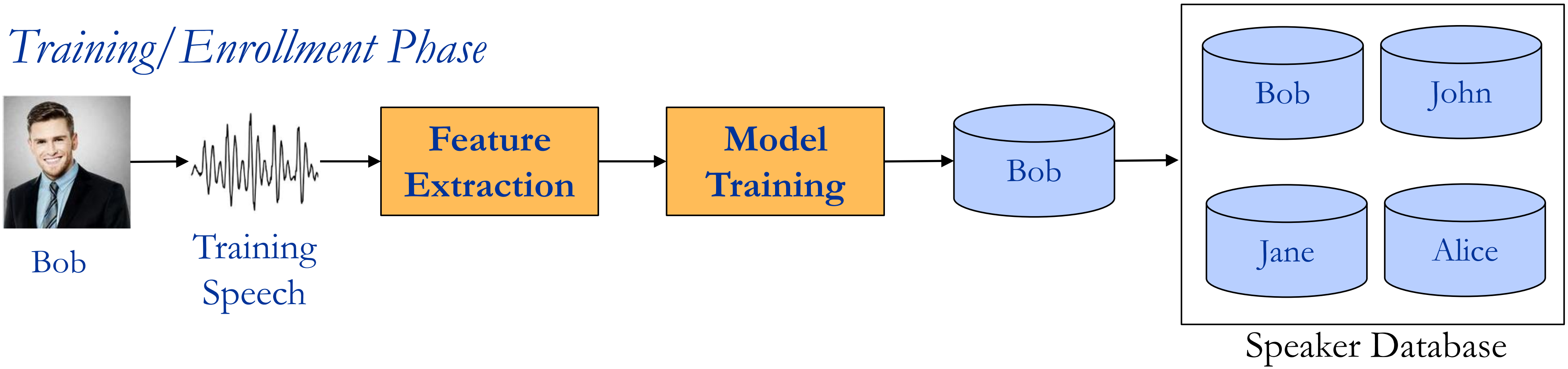


Speaker Database

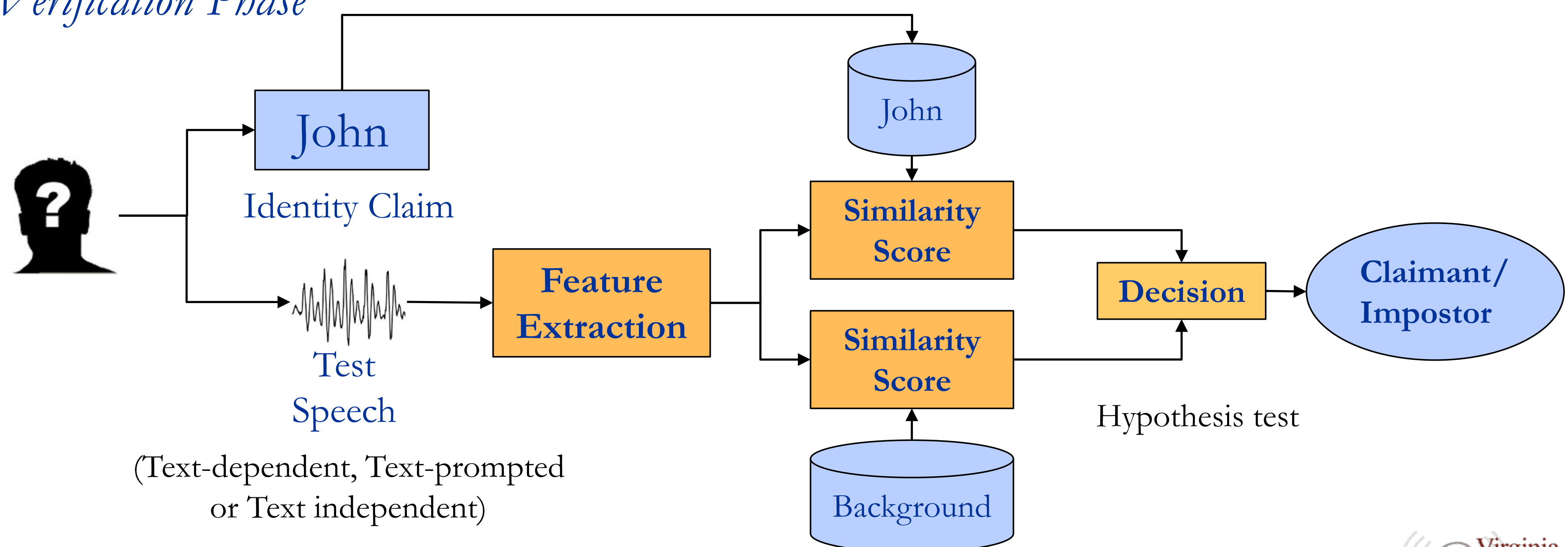
Is this Bob's voice?

Speaker Verification System

Training/Enrollment Phase



Verification Phase



Outline

- Introduction
- **Motivation**
- Feature Extraction
- Speaker Verification Framework
- Speaker Verification Results
- Conclusions

Motivation

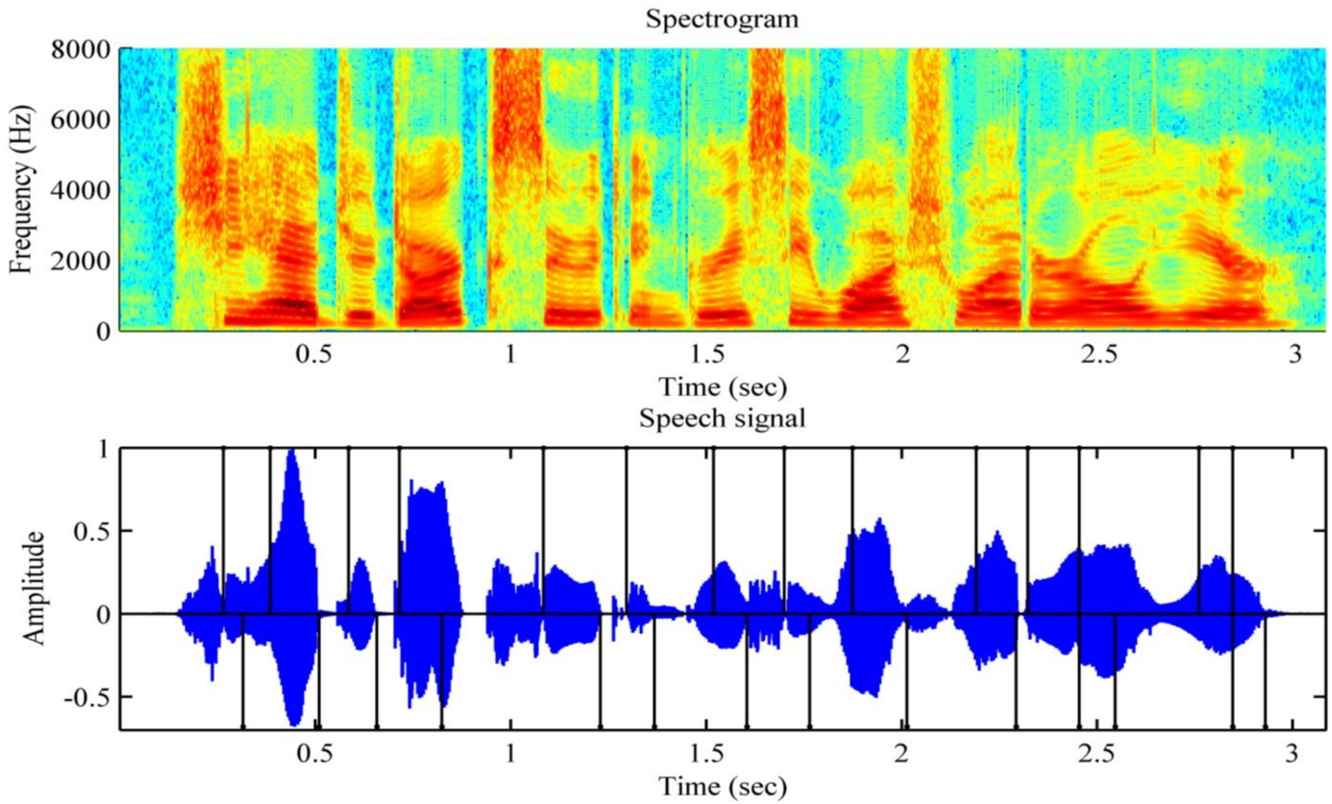
- State-of-the-art SV systems provide near perfect performance under clean conditions.
- Performance deteriorates in the presence of background noise.
- Noise-robustness improved by feature/model compensation and signal enhancement techniques.
- Drawbacks:
  - –Require extensive training
  - –Computationally expensive
  - –Make assumptions about the noise characteristics.

Can we improve performance by utilizing only important zones of speech, and discarding less important zones during verification?

Relative Importance of Speech Zones

Relative Importance = Amount of speaker-specific information

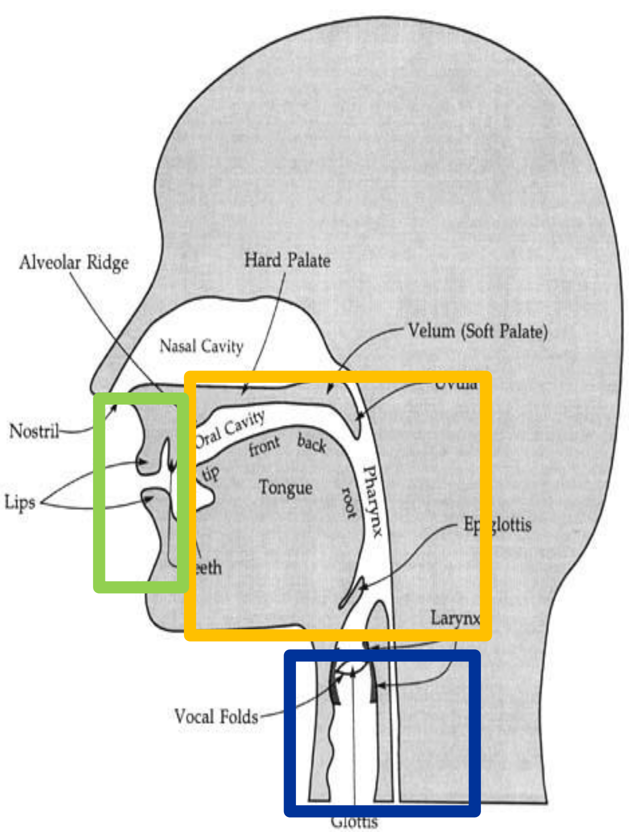
- Co-articulation: the way a speaker moves from one sound to another is speaker specific.
- Dynamic transition regions are more speaker-specific than steady regions.
- We consider consonant-vowel (CV) and vowel-consonant (VC) transitions as vowels are easy to identify under noise.

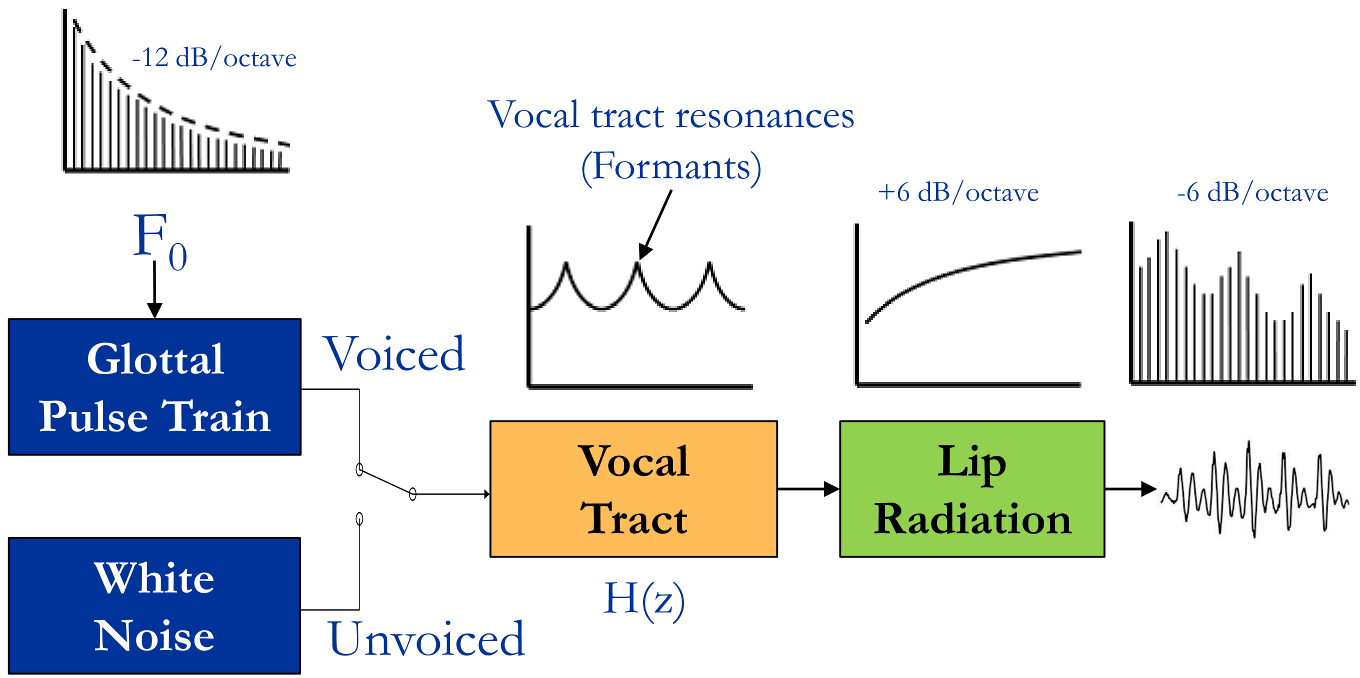


Outline

- Introduction
- Motivation
- **Feature Extraction**
- Speaker Verification Framework
- Speaker Verification Results
- Conclusions

Source-Filter Model of Speech Production



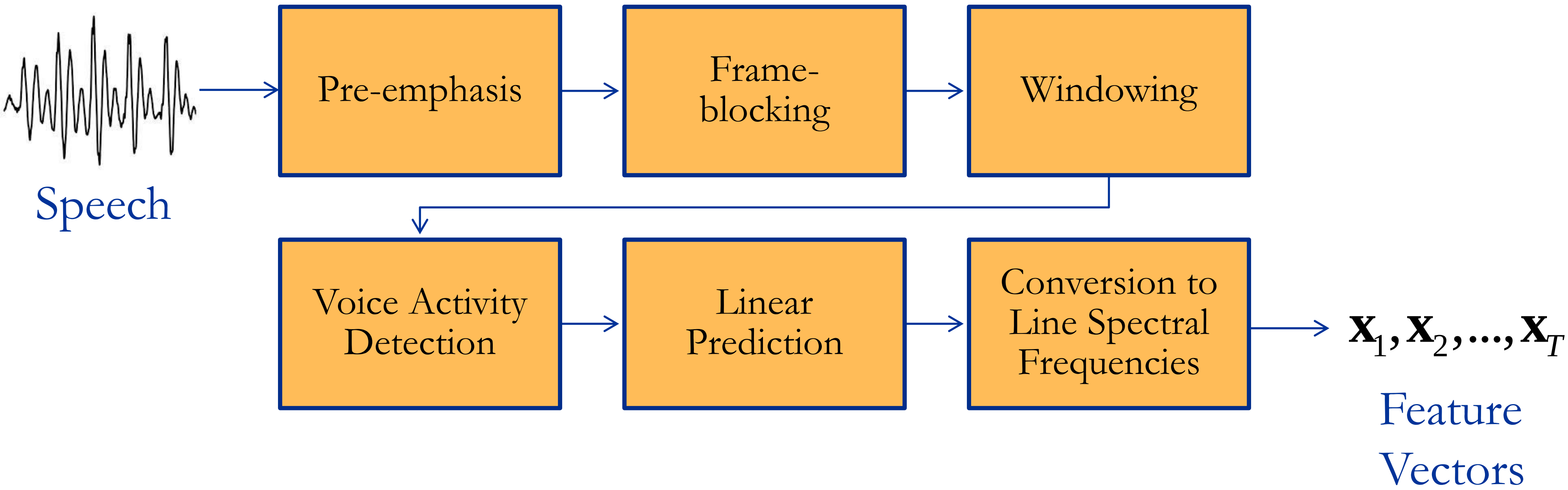


$$H(z)=\frac{1}{A_p(z)}=\frac{1}{1+\sum_{k=1}^{p}a_k z^{-k}}$$

Linear Prediction Coefficients

Image sources: [1] http://jcarreras.homestead.com/rrphonetics1.html [2] http://www.splab.net/APD/G500/index-e.html

Feature Extraction



Line Spectral Frequencies

Speech is a combination of two resonance conditions – vocal tract closed at the glottis and vocal tract open at the glottis.

$$A_p(z) = 1 + \sum_{k=1}^{p} a_k z^{-k} \Longrightarrow A_p(z) = \frac{P(z) + Q(z)}{2}$$

Closed glottis:

$$P(z) = A_p(z) + z^{-(p+1)} A_p(z^{-1})$$

(Symmetric)

$$\theta_{Pk} = e^{j\omega_{Pk}}, \quad 1 \le k \le p+1$$

Open glottis:

$$Q(z) = A_p(z) - z^{-(p+1)} A_p(z^{-1})$$

(Anti-symmetric)

$$\theta_{Qk} = e^{j\omega_{Qk}}, \quad 1 \le k \le p+1$$

LSF Feature: $\mathbf{x} = \left[\omega_{P1}\, \omega_{Q1}\, \omega_{P2}\, \omega_{Q2} \ldots \omega_{\frac{Pp}{2}}\, \omega_{\frac{Qp}{2}} \right]$

- Efficient representation
- Good quantization properties
- Can be interpolated

Interlacing : $0 < \omega_{P1} < \omega_{Q1} < \omega_{P2} < \omega_{Q2} \ldots < \omega_{\frac{Pp}{2}} < \omega_{\frac{Qp}{2}} < \pi$

Visualizing LSFs

- Line Spectral Frequencies (LSFs) are spectral features.
- Every formant is bracketed by an LSF pair
- If a pair of LSFs are far from each other, the magnitude response will be relatively flat around the two LSF.

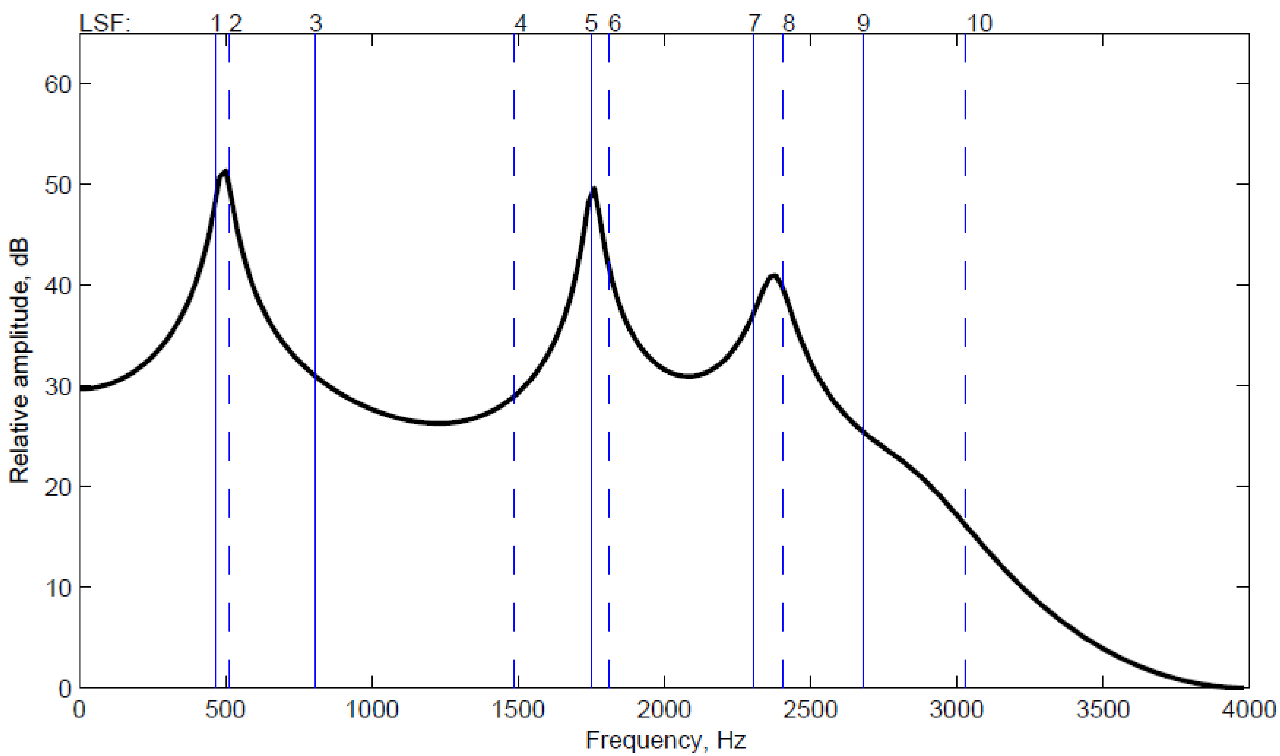


Image source: McLoughlin, Ian Vince. "Line spectral pairs." Signal processing 88.3 (2008): 448-467, http://www.lintech.org/webpapers/lsp_paper.pdf

Outline

- Introduction
- Motivation
- Feature Extraction
- **Speaker Verification Framework**
- Speaker Verification Results
- Conclusions

Gaussian Mixture Model

- A Gaussian Mixture Model (GMM) λ is a linear weighted sum of M Gaussian components

$$p(\mathbf{x}|\lambda) = \sum_{m=1}^{M} p_m g_m(\mathbf{x} \,|\, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) \qquad \sum_{m=1}^{M} p_m = 1 \qquad \boldsymbol{\Sigma}_m \in \mathbb{R}^{D\times D} \quad \boldsymbol{\mu}_m \in \mathbb{R}^{D}$$

$$X = \{\mathbf{x}_t \in \mathbb{R}^D : 1 \le t \le T\} \longrightarrow \lambda = \{p_m, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\}, \quad m = 1, 2, \ldots, M$$

- Maximize GMM Likelihood: $p(X \,|\, \lambda) = \prod_{t=1}^{T} p(\mathbf{x}_t \,|\, \lambda)$
- **Expectation-Maximization Algorithm:** $\boxed{p\left(X \,|\, \lambda^{(i+1)}\right) \ge p\left(X \,|\, \lambda^{(i)}\right)}$

Speaker Verification – Enrollment Phase

H_0 : Speech is from the hypothesized speaker – Speaker Model
H_1 : Speech is not from the hypothesized speaker – Background Model

- The Universal Background Model (UBM) is a 256 component GMM.
- Trained by pooling speech from 462 speakers in the TIMIT corpus.
- Speaker Models - GMMs obtained by Maximum a posteriori (MAP) adaptation of the UBM means
- Tighter coupling– better performance, faster scoring.

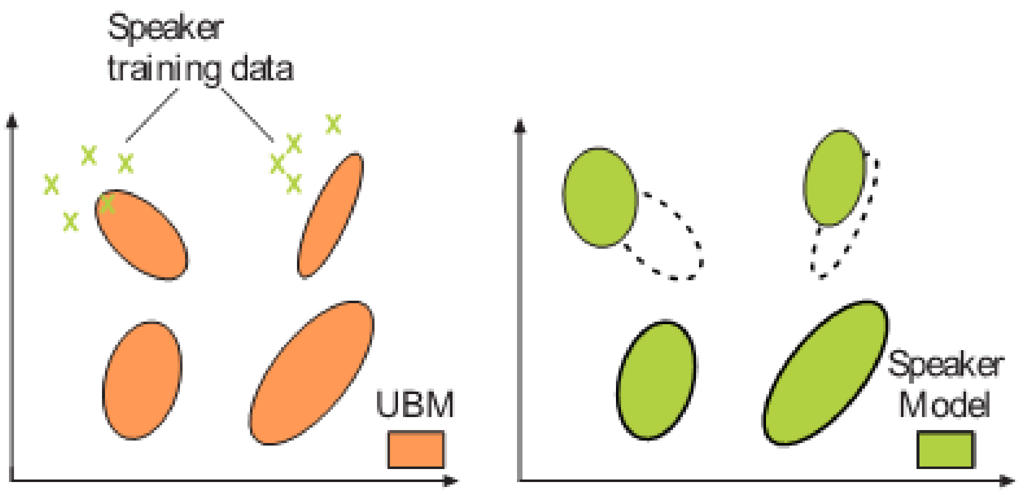


Image Source: Reynolds, Douglas A., Thomas F. Quatieri, and Robert B. Dunn. "Speaker verification using adapted Gaussian mixture models." *Digital signal processing* 10.1 (2000): 19-41.

Speaker Verification – Testing Phase

The speaker verification task is a simple hypothesis test

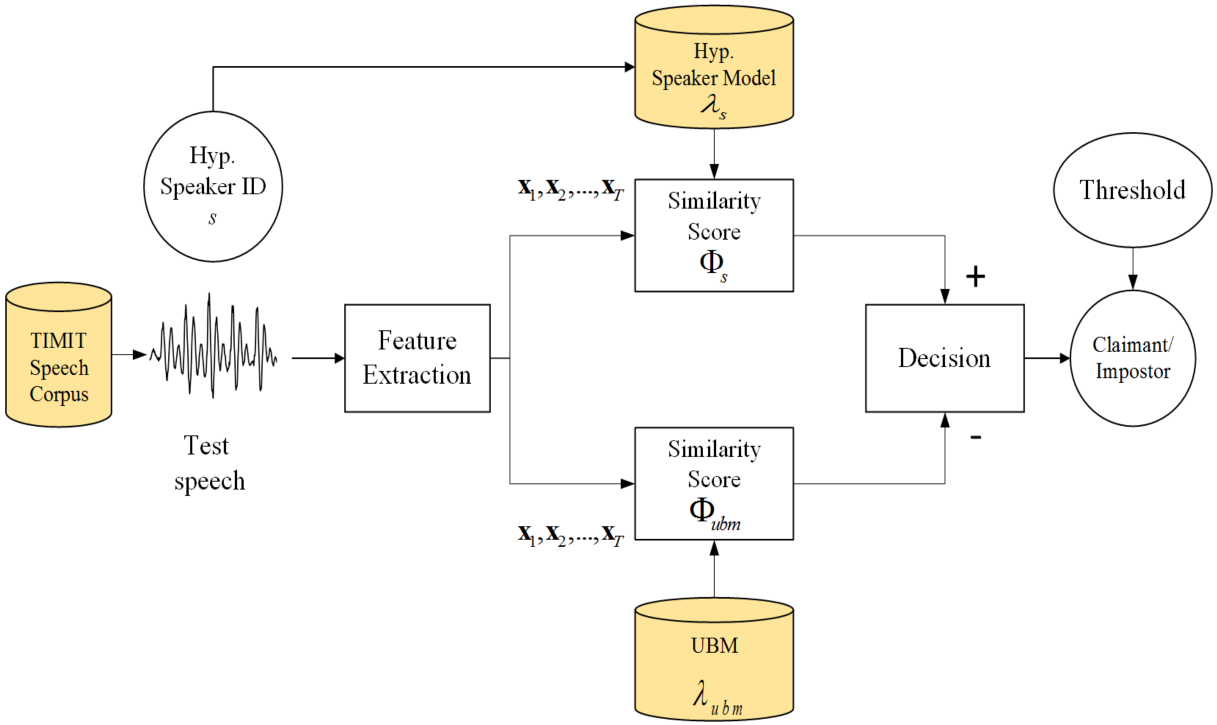
Given a set of test features $X = \{ \mathbf{x}_t \in \mathbb{R}^D : 1 \le t \le T \}$

H_0 : X is from speaker S

$$\Phi_s = \sum_{t=1}^{T} \log p(\mathbf{x}_t \mid \lambda_s)$$

H_1 : X is not from speaker S

$$\Phi_{ubm} = \sum_{t=1}^{T} \log p(\mathbf{x}_t \mid \lambda_{ubm})$$

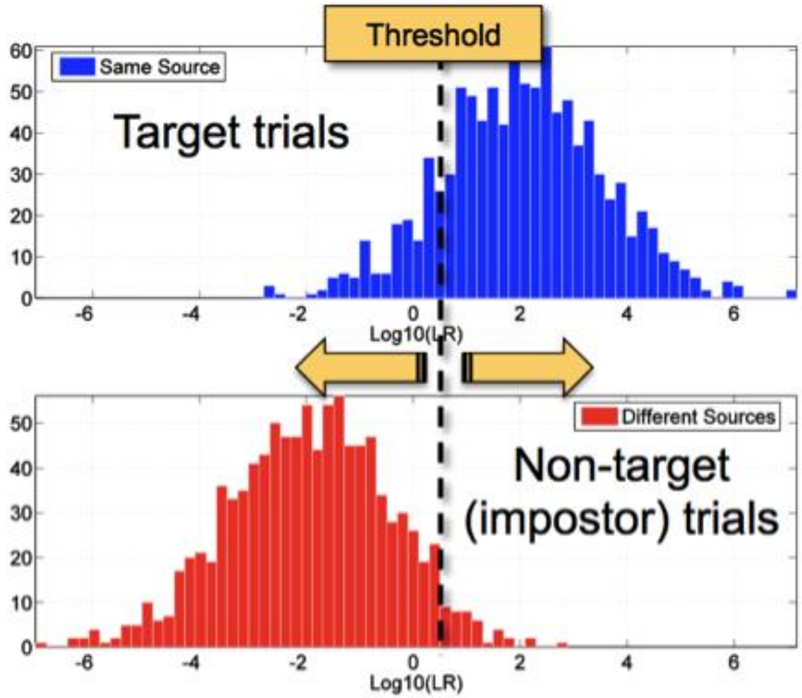


Log-likelihood ratio:

$$\Lambda(X) = \Phi_s - \Phi_{ubm}$$

$$\Lambda(X) \begin{cases} \ge \theta & accept\ H_0 \\ < \theta & reject\ H_0 \end{cases}$$

Speaker Verification – Performance Evaluation



Miss: Rejecting a target trial

$$E_{miss} = n_{miss} / n_t$$

False Alarm: Accepting an impostor trial

$$E_{fa} = n_{fa} / n_i$$

Equal Error Rate (%): Point at which probability of miss equals probability of false alarm.

Image source: http://loquens.revistas.csic.es/index.php/loquens/article/viewArticle/9/20

Outline

- Introduction
- Motivation
- Feature Extraction
- Speaker Verification Framework
- **Speaker Verification Results**
- Conclusions

Experimental Setup

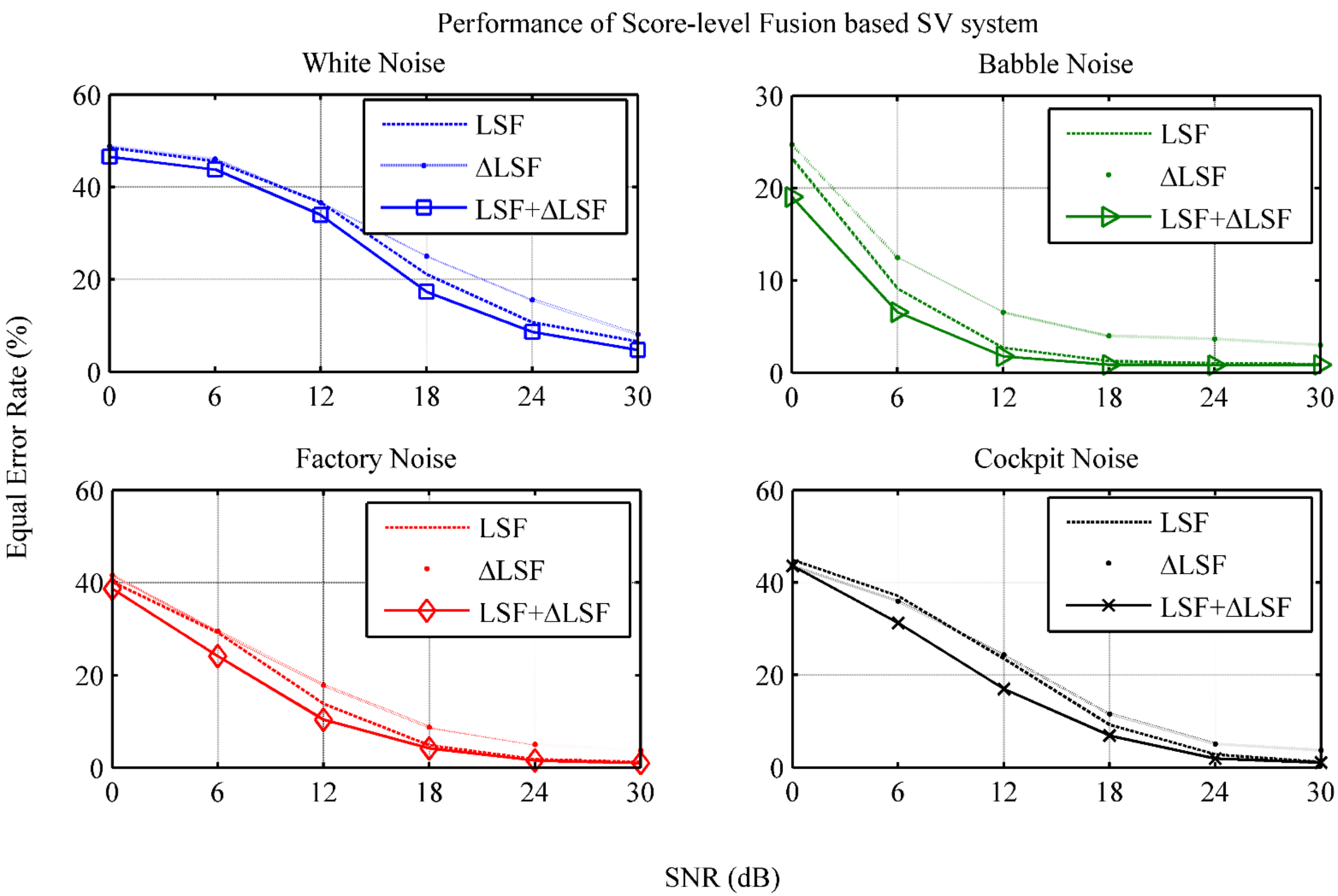
Parameter	Description
Number of speakers (S)	168
Training set of each speaker	All SX, SI sentences from TIMIT corpus (~3 seconds x 8)
Test set of each speaker	SA sentences from TIMIT (~3 seconds x 2)
Feature Type	LSF
Feature Dimension/Order (p)	20
Frame Length (L)	20 msec
Frame Shift (δ)	10 msec
Number of GMM Components (M)	256 (UBM adapted GMM)
GMM Covariance Type	Nodal and Diagonal
Noise Corpus	SPIB Noise Dataset

# Speakers	# Target Trials	# Impostor Trials	# Total Trials
168	168 x 2 = 336	168 x 167 x 2 = 56112	56112 + 336 = 56448

SV System Performance

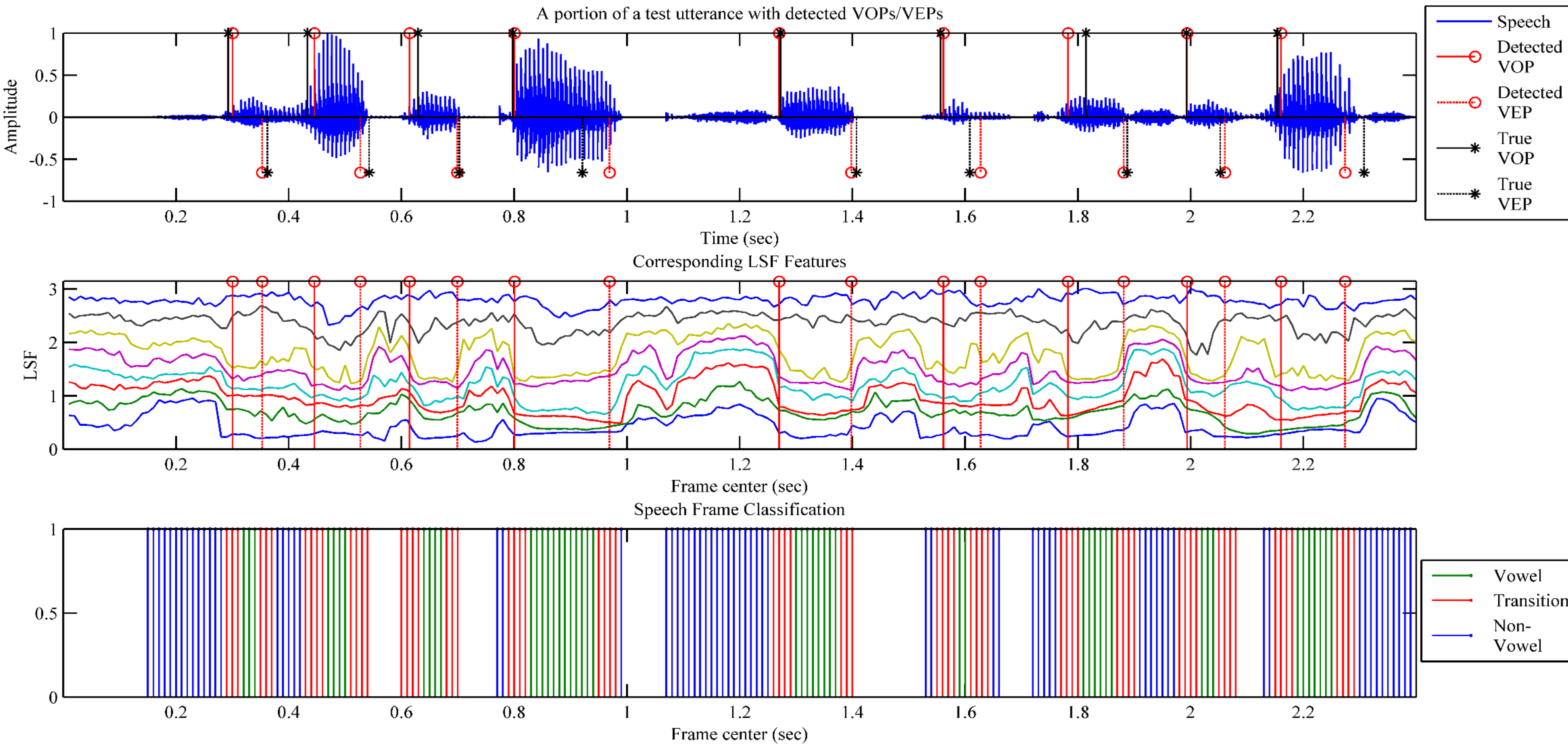
1) Static features - LSF
2) Dynamic features - ΔLSF
3) Score Level Fusion - LSF + ΔLSF

$$\Lambda_f(X) = \alpha\Lambda_{LSF}(X) + (1-\alpha)\Lambda_{\Delta LSF}(X)$$



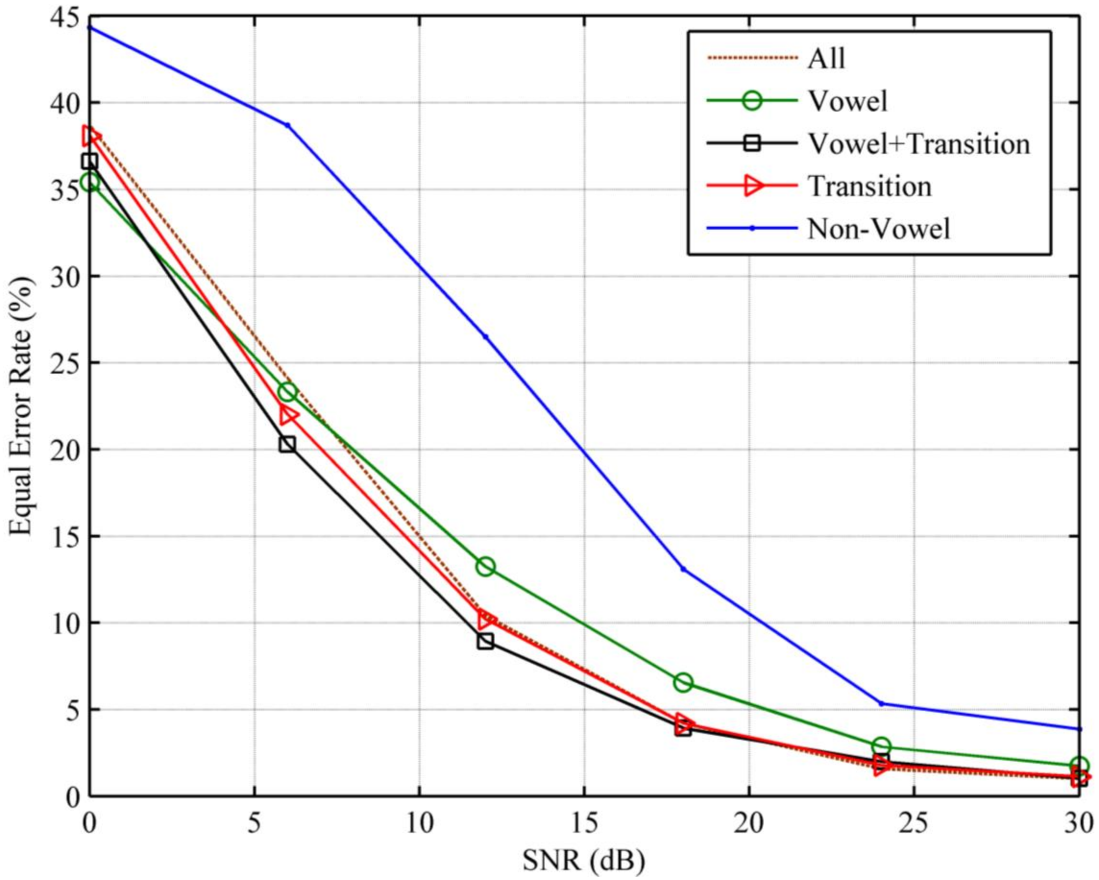
- Baseline EER=0.86%
- Score-level fusion improves performance under noise

Discriminative Power of Speech Zones



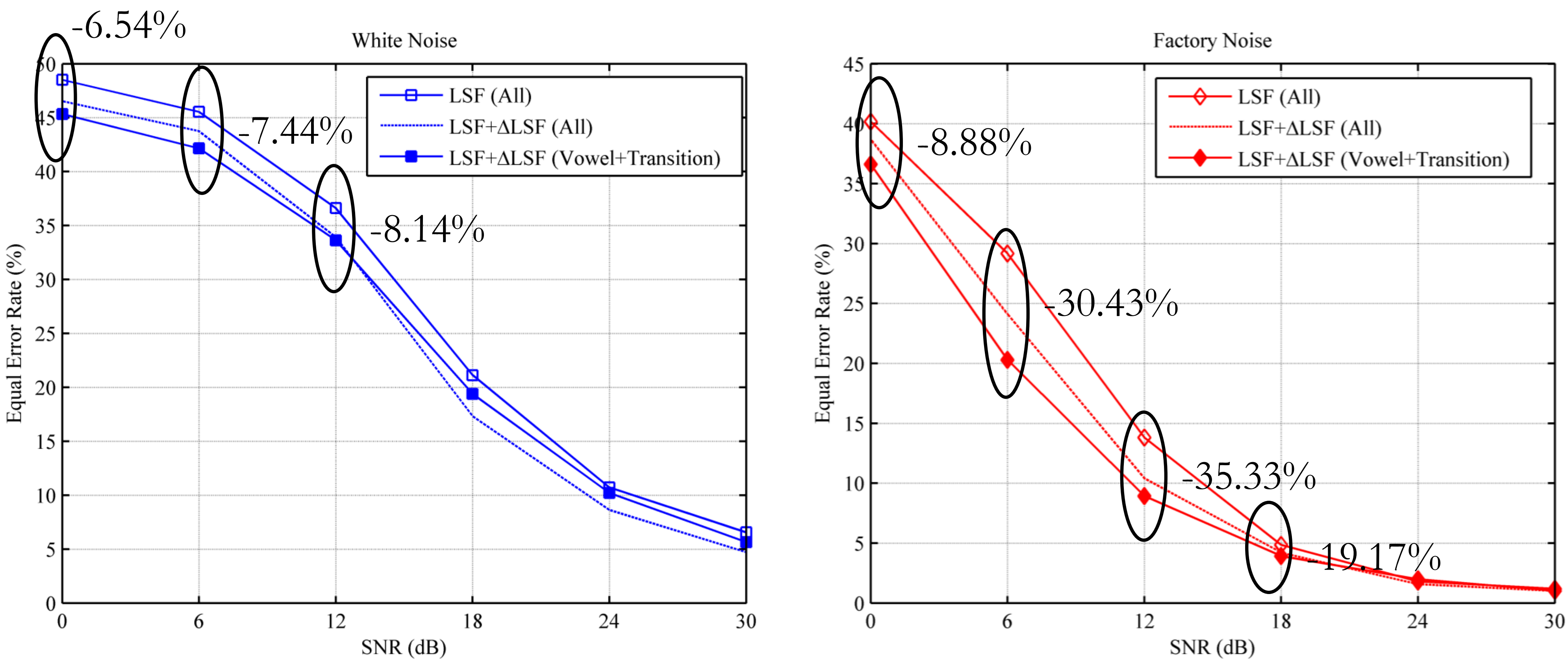
Discriminative Power of Speech Zones

X_{tr} is the set of features from transition frames $\Rightarrow$ $\Phi_{s,X_{tr}} = \sum_{\mathbf{x} \in X_{tr}} \log p(\mathbf{x} \mid \lambda_s)$



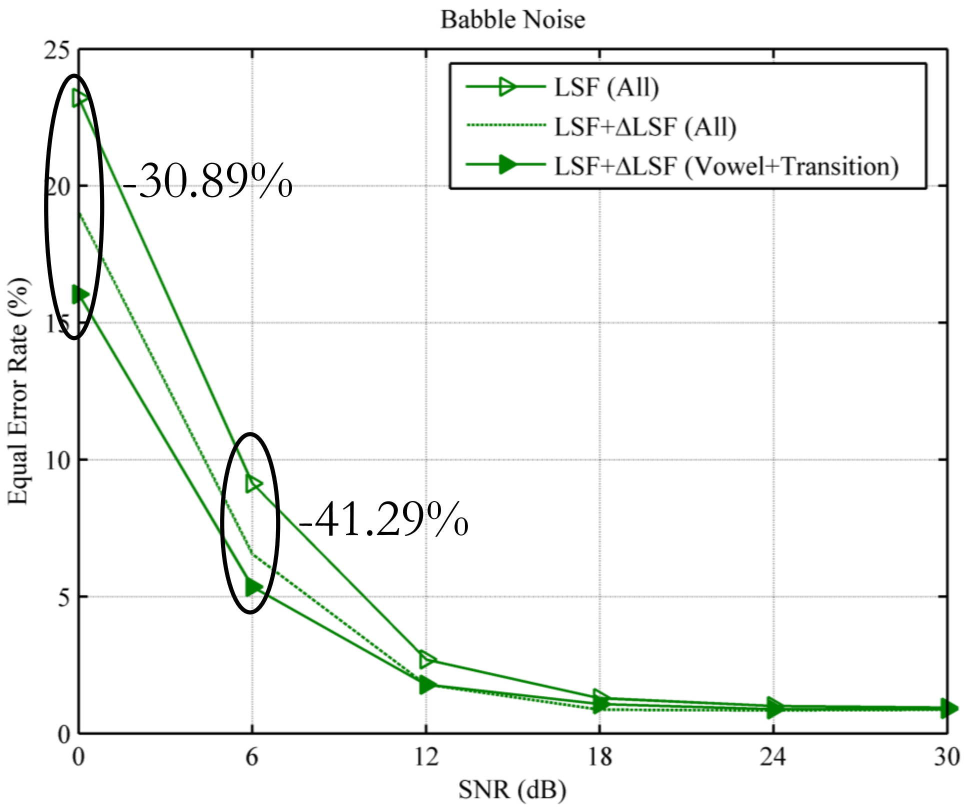
- High SNR – transitions are most speaker discriminative
- Low SNR – vowels are most speaker discriminative
- Frame-level selection- not much benefit in high SNR
- Scoring on **vowel + transition** frames improves performance in low SNR.

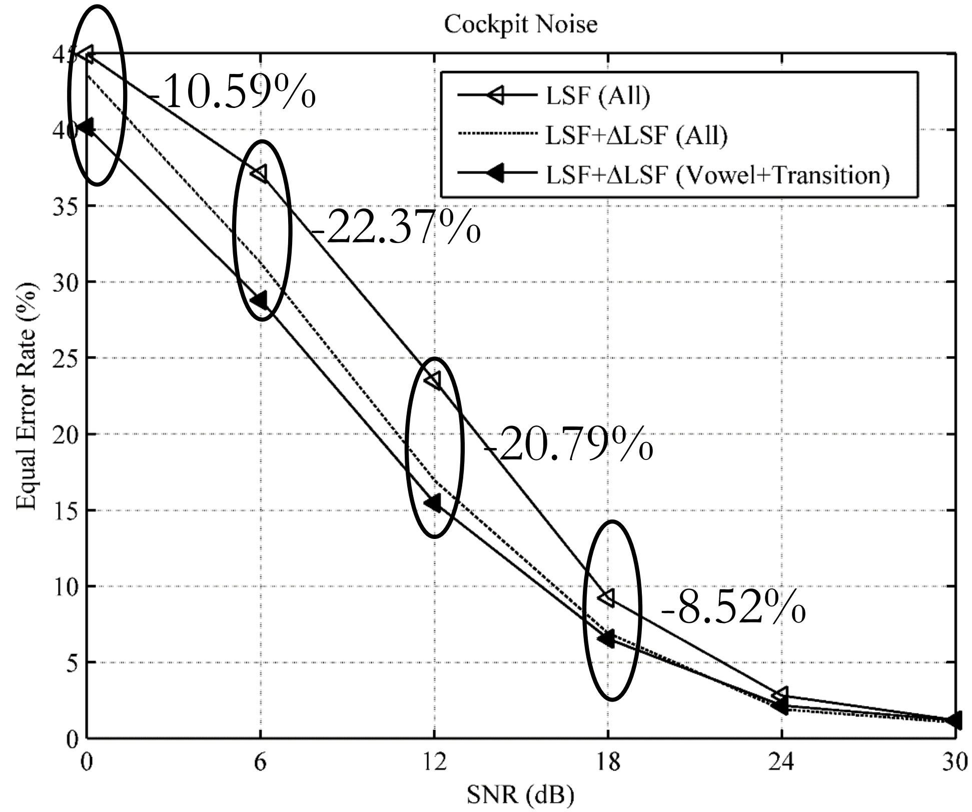
Performance Improvement



Wireless @ Virginia Tech

Performance Improvement





Outline

- Introduction
- Motivation
- Feature Extraction
- Speaker Verification Framework
- Speaker Verification Results
- **Conclusions**

Conclusions

- An automatic, text-independent speaker verification (SV) system was developed using Line Spectral Frequency (LSF) features.
- The performance of the SV system was evaluated under noise.
- Score-level fusion was used to combine complementary information from static and dynamic LSF features.
- Speaker-discriminative power of vowel, transition and non-vowel regions were investigated.
- Transition regions are the most speaker-discriminative under high SNR conditions
- High-energy vowel regions are most speaker-discriminative under low SNR conditions.
- Under noisy conditions, the performance of the score-level fusion based SV systems can be improved substantially by scoring exclusively on a combination of transition and vowel frames.
- Future work
  - Investigate the effect of training speaker models using transition zones.
  - Improve the algorithm used to localize transition zones.