# The Noisy Speech Chain


*Abeer Alwan*

*Speech Processing and Auditory Perception Laboratory*
*Department of Electrical Engineering, UCLA*

*http://www.icsl.ucla.edu/~spapl* *alwan@ee.ucla.edu*

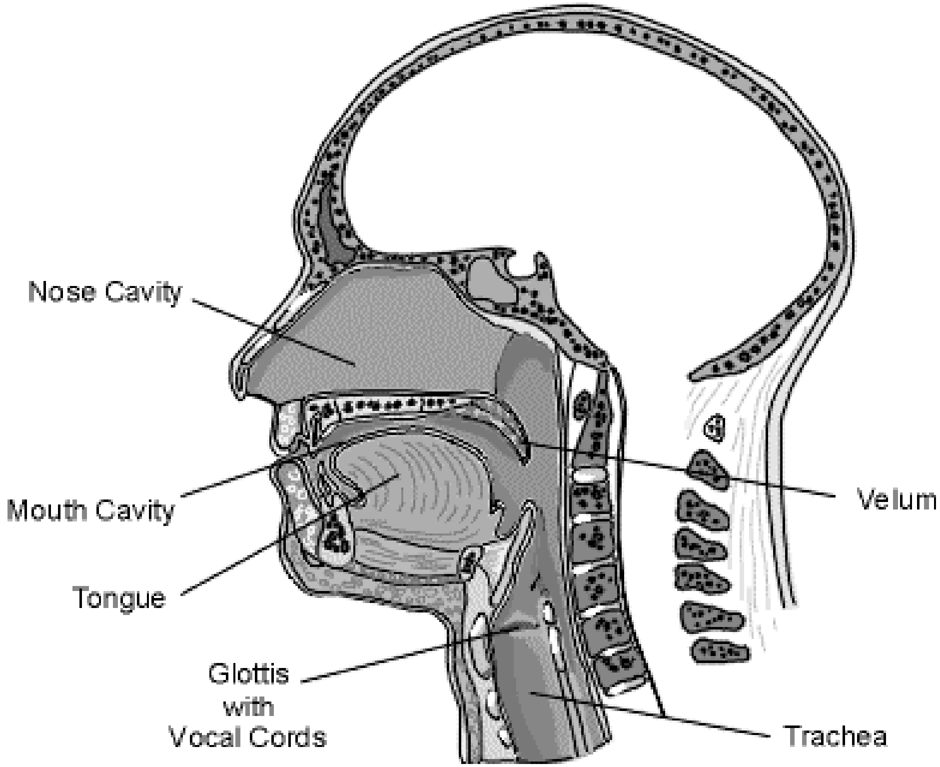
# Outline

- Part I: auditory-based representations and their use in automatic speech recognition (ASR)
- Part II: Speech perception experiments to illustrate the importance of different acoustic cues in noise
- Part III: A time-frequency model to predict the perception of  stimuli of various durations and bandwidths including synthetic speech sounds in noise

# Terminology

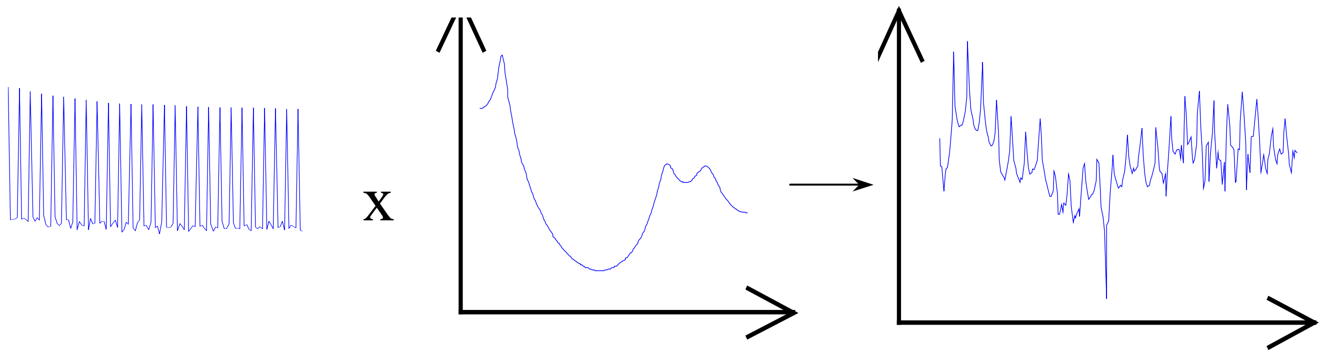
- Speech Production (how sounds are produced)
- Psychoacoustics (perception of a wide range of acoustic stimuli)
- Speech Perception (how speech-like sounds are perceived)

# Speech Production Organs




Ref: Technical University of Berlin

# *LTI Model of Speech Production*




Source Function (periodic and/or noisy)

Vocal Tract Transfer Function

Speech Signal (Frequency Domain)

# *Fundamental Frequency*

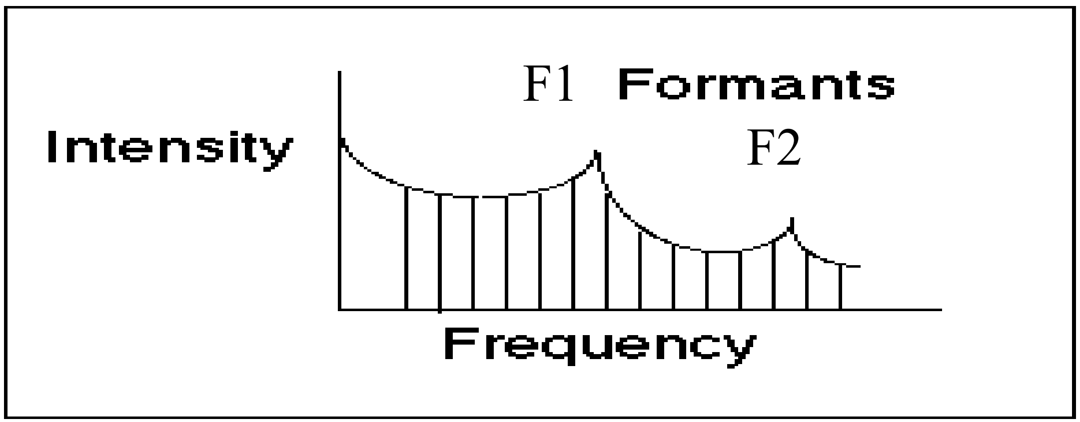
- Fundamental Frequency (F0) is the periodicity induced by the vocal cords for voiced sounds

$T_0 = 1/F_0$

| | Male | Female | Child |
|---|---|---|---|
| F0 (Hz) | 125 | 225 | 300 |

# Pole-Zero Patterns in the Vocal Tract Transfer Function (VTTF)

- Resonances of the vocal tract (formants) are critical to sound identification are correlated with the size of the vocal tract.

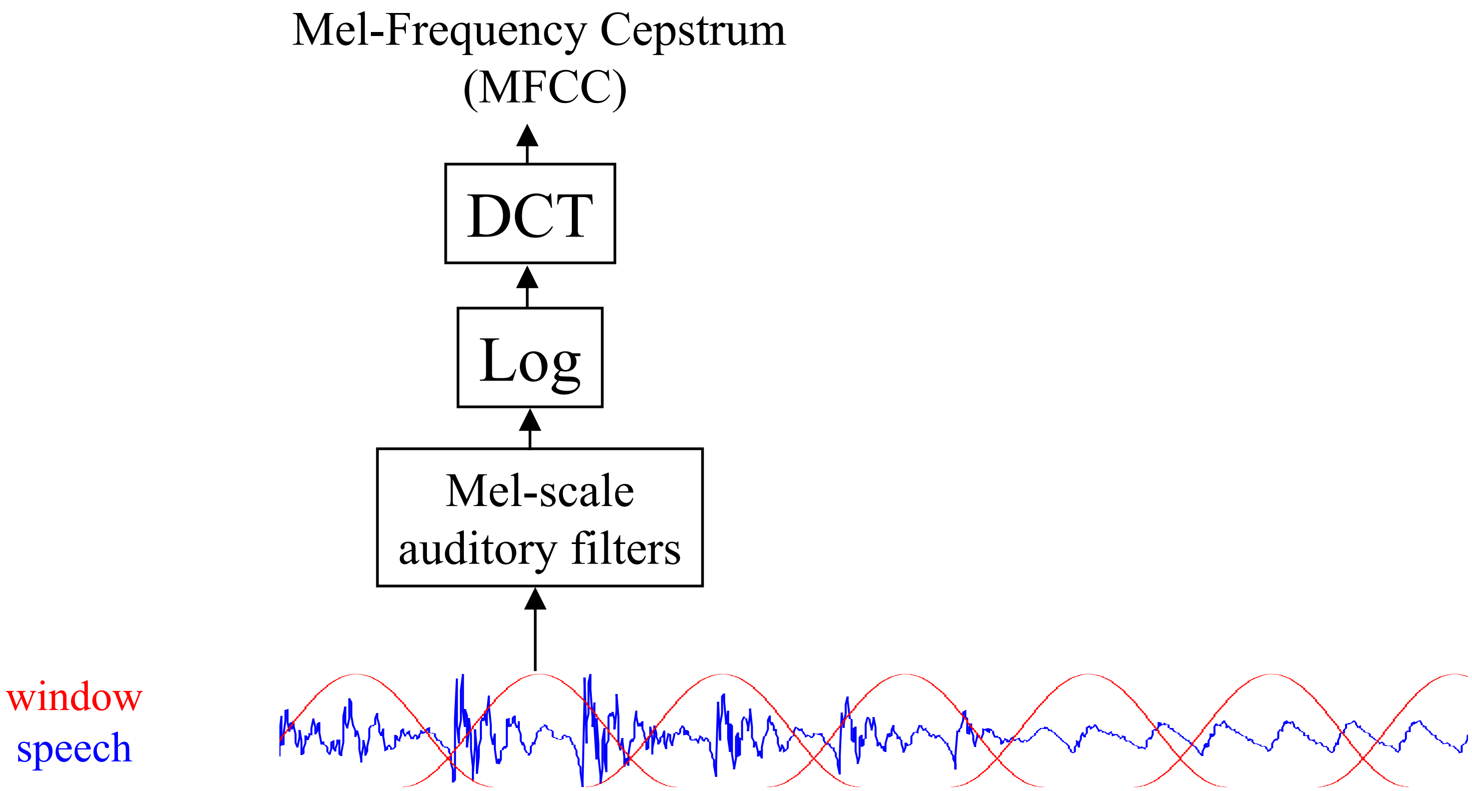


# *Speech in Noise*

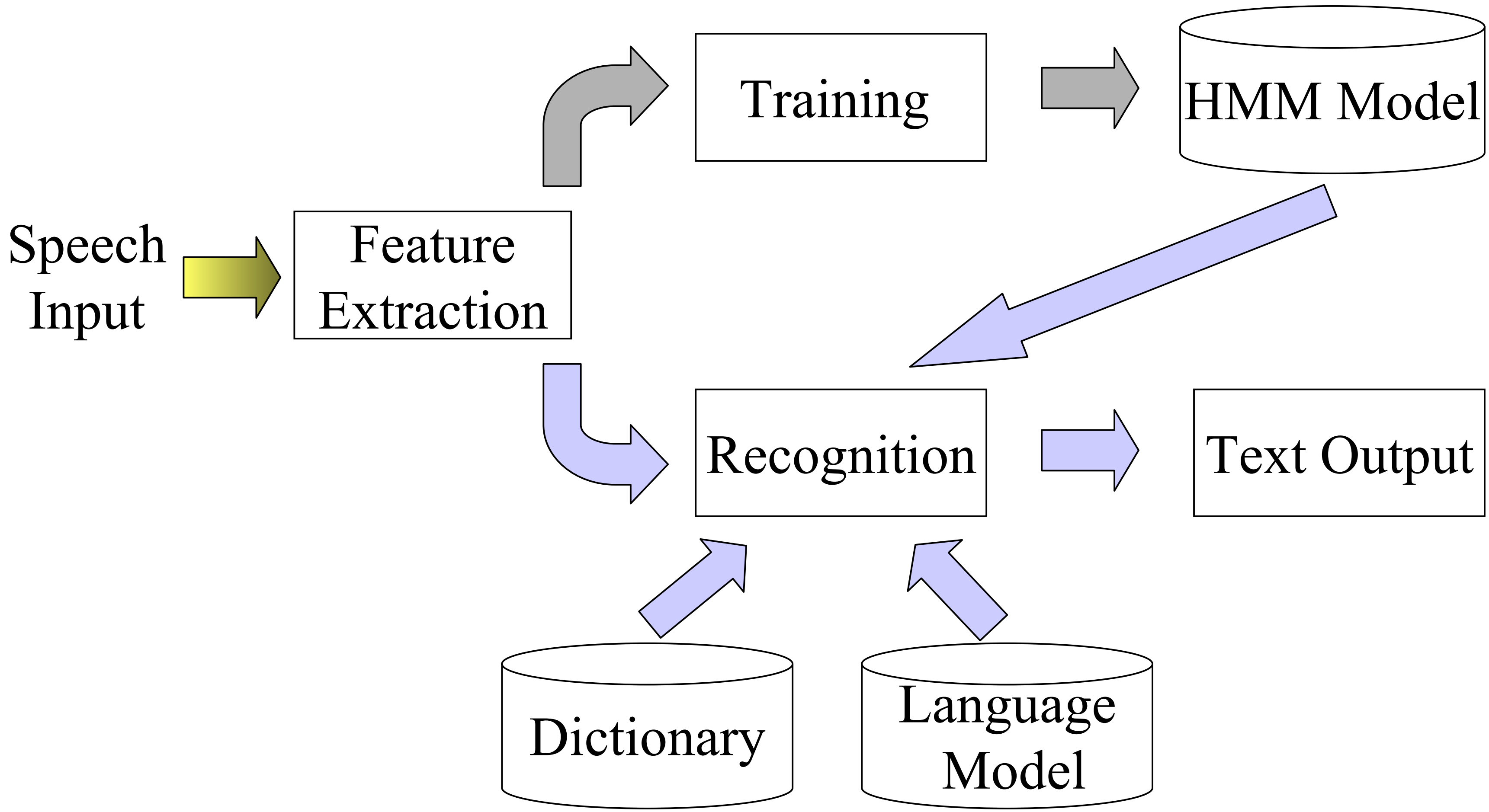
**Healthy-hearing are remarkably adept at perceiving speech in noise. However,**

- The most common complaint of hearing-aid users is listening to speech in naturally noisy environments.
- The performance of automatic speech recognition (ASR) systems degrades significantly in the presence of noise.

# Front-end: feature extraction



# Typical ASR Systems



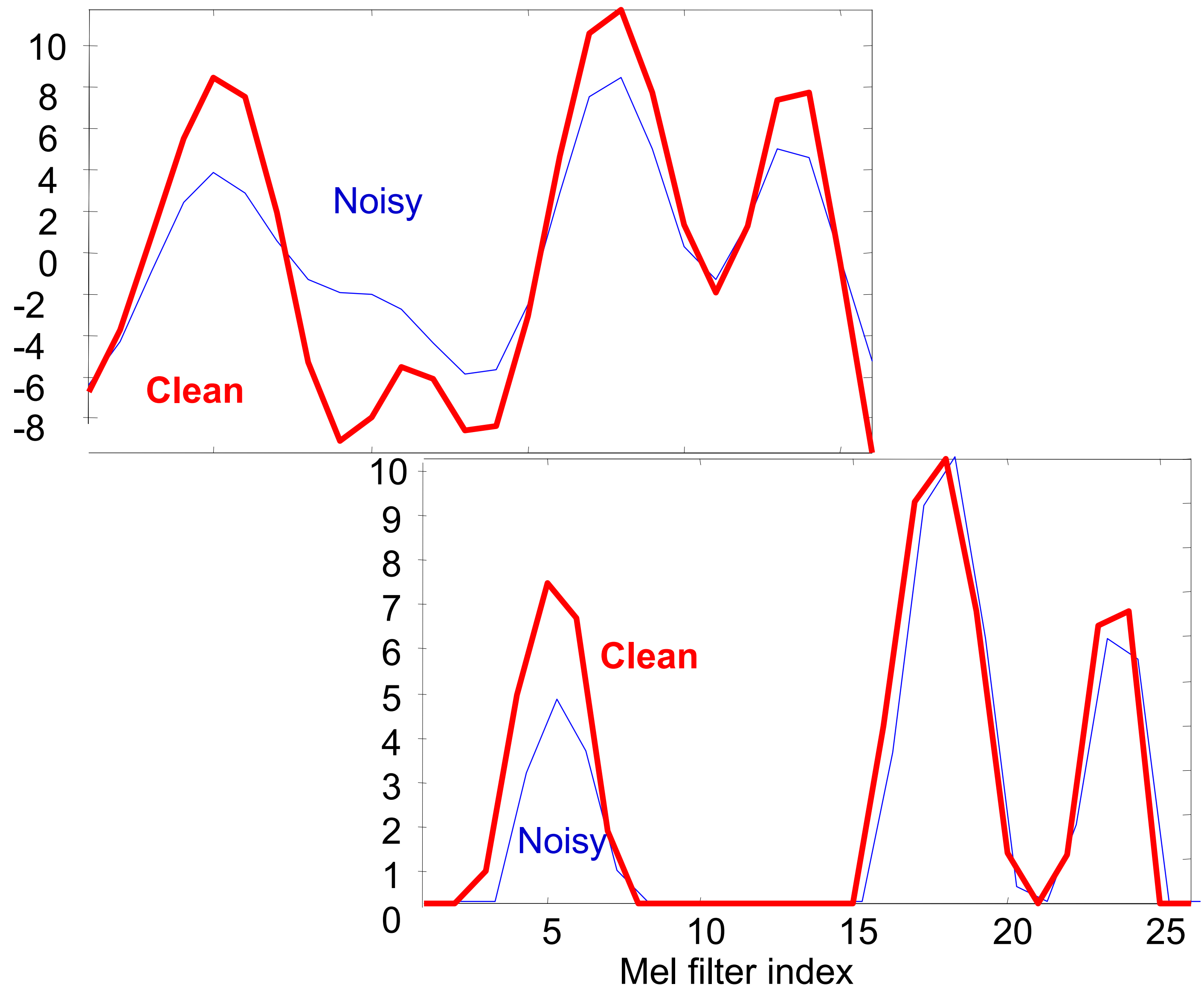
# The ‘Robust’ Human Auditory System

***The auditory system is extremely robust to noise due to both:***

- ***“Intelligent” High-Level Processing***
- ***Inherently Robust Auditory Representation***

# Auditory-based signal representations (spapl, 1997-present)

- Adaptation (sensitivity to onsets and offsets): modeled after FM experiments
- Spectral sharpening: physiological and perceptual evidence
- Exploiting the fact that the VTTF moves slowly in time
- Not all ‘uniform’ segments are equally-important




Clean and noisy log Mel spectra before (upper panel) peak isolation and peak-to-valley ratio locking, and after (lower panel).

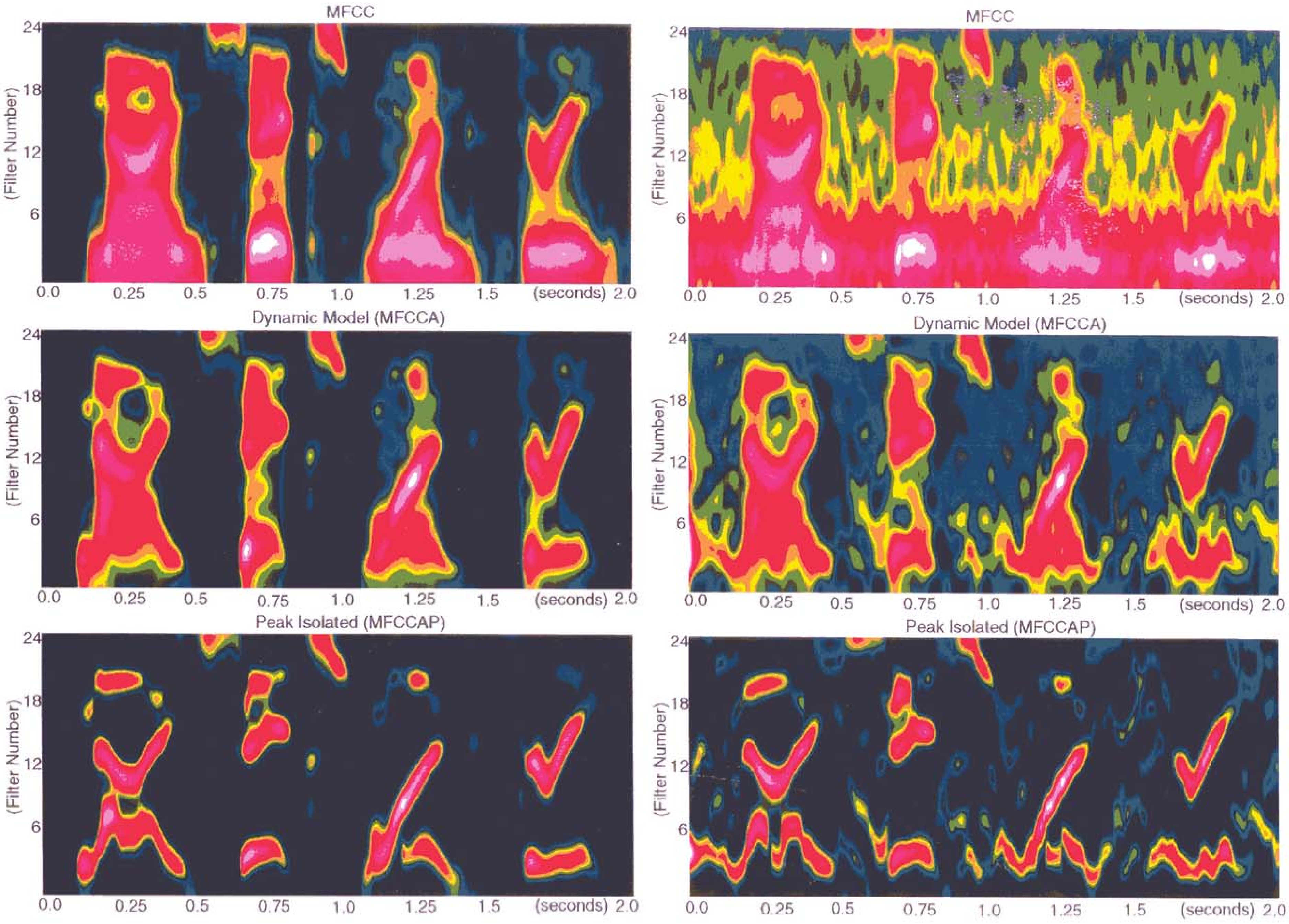

MFCC
Dynamic Model (MFCCA)
Peak Isolated (MFCCAP)
(Filter Number)
24
18
12
6
0.0
0.25
0.5
0.75
1.0
1.25
1.5
(seconds) 2.0

*Fig. 11*

(Strope and Alwan, 1997,1998)

These techniques improved ASR in noise significantly.

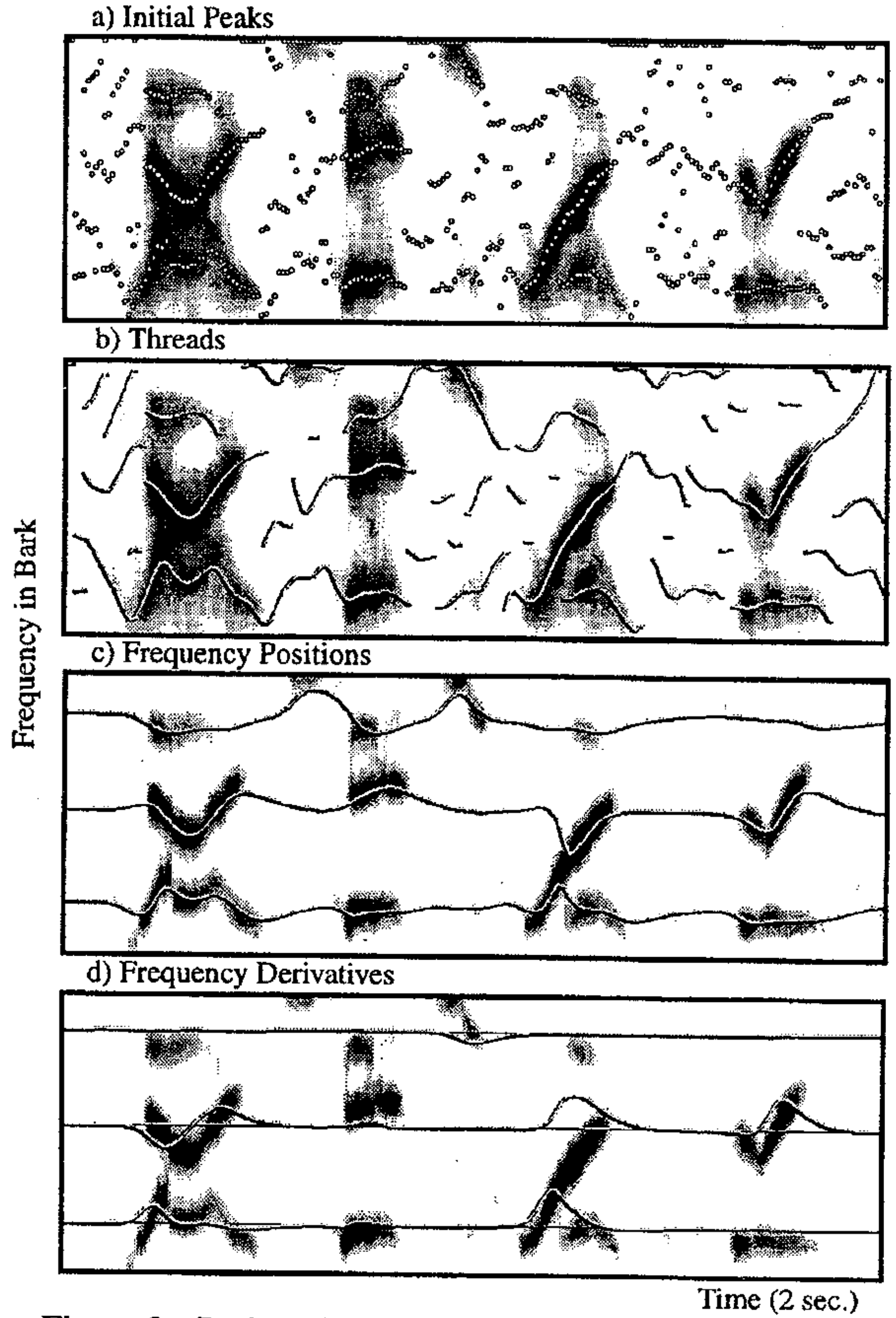
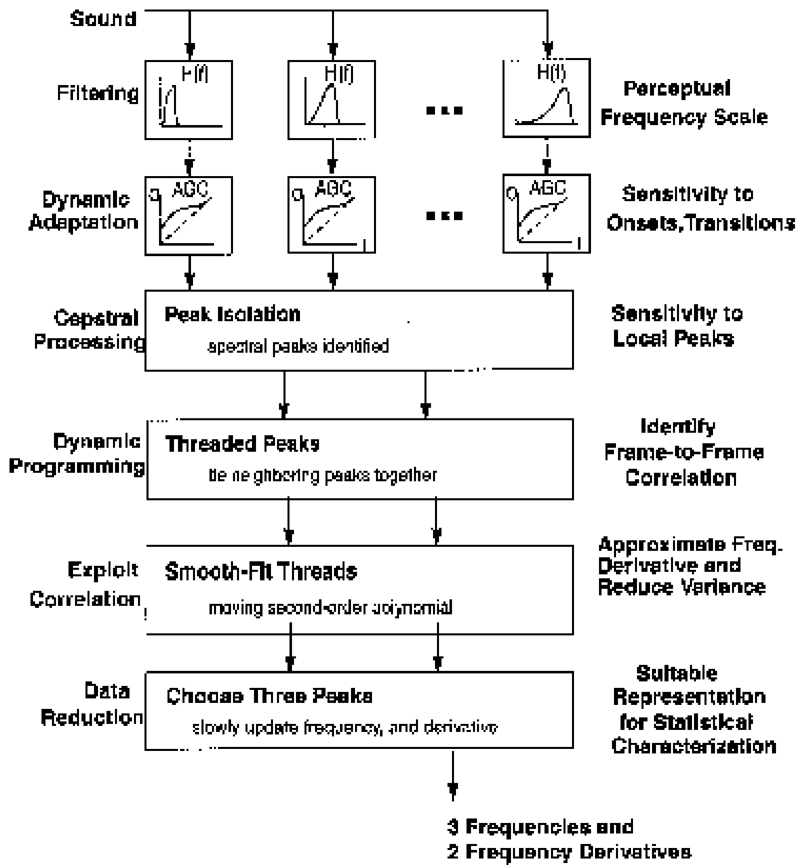


**Figure 2.** Peak positions and motion.

# *Overall System*



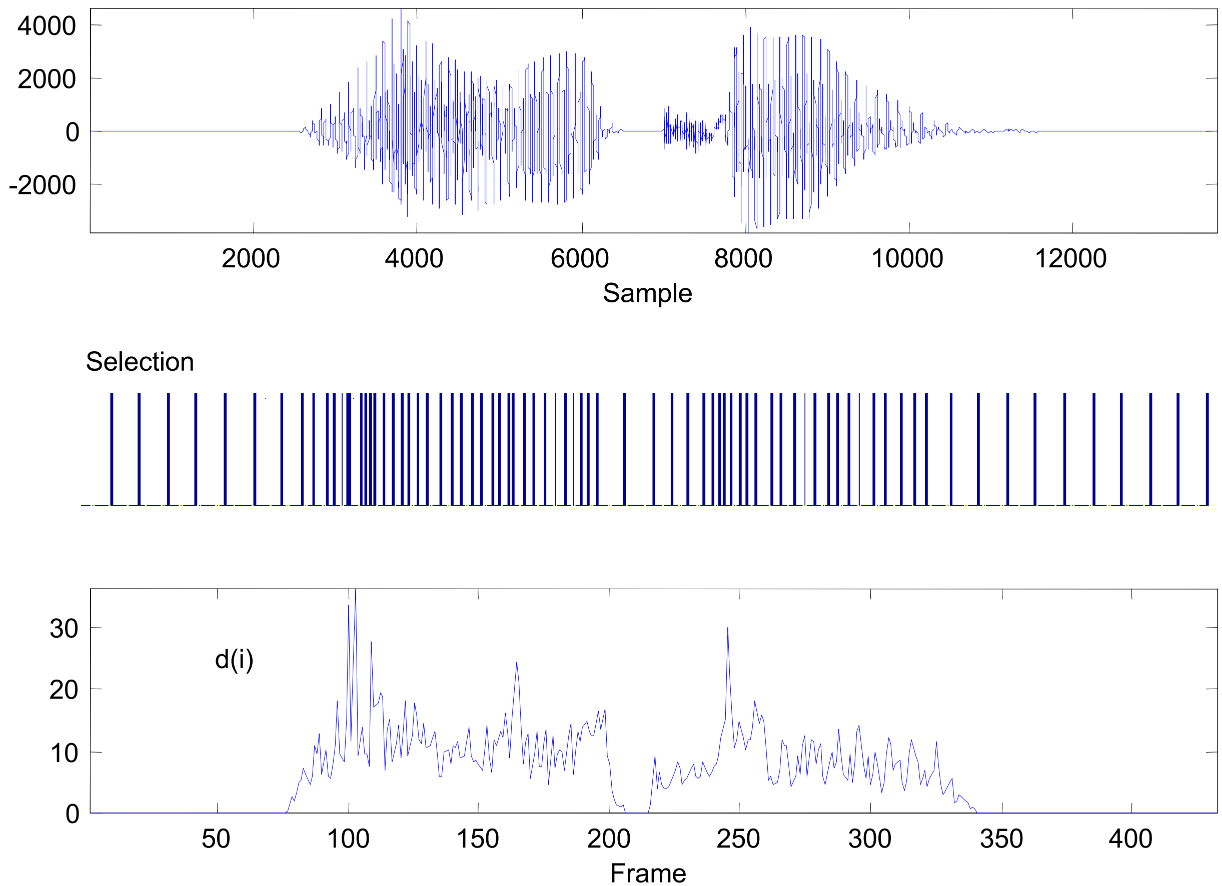

**Focus: adaptation and temporal correlations of local spectral peaks.**

**(Strope and Alwan, 1997)**

# *Variable Frame Rate Analysis (VFR) (Zhu and Alwan, 2000)*

- Spectral changes are important perceptual cues for discrimination. Such changes can occur over very short time intervals.
- Computing frames every 10 ms, as commonly done in ASR, is not sufficient to capture such dynamic changes.

# An Example of VFR



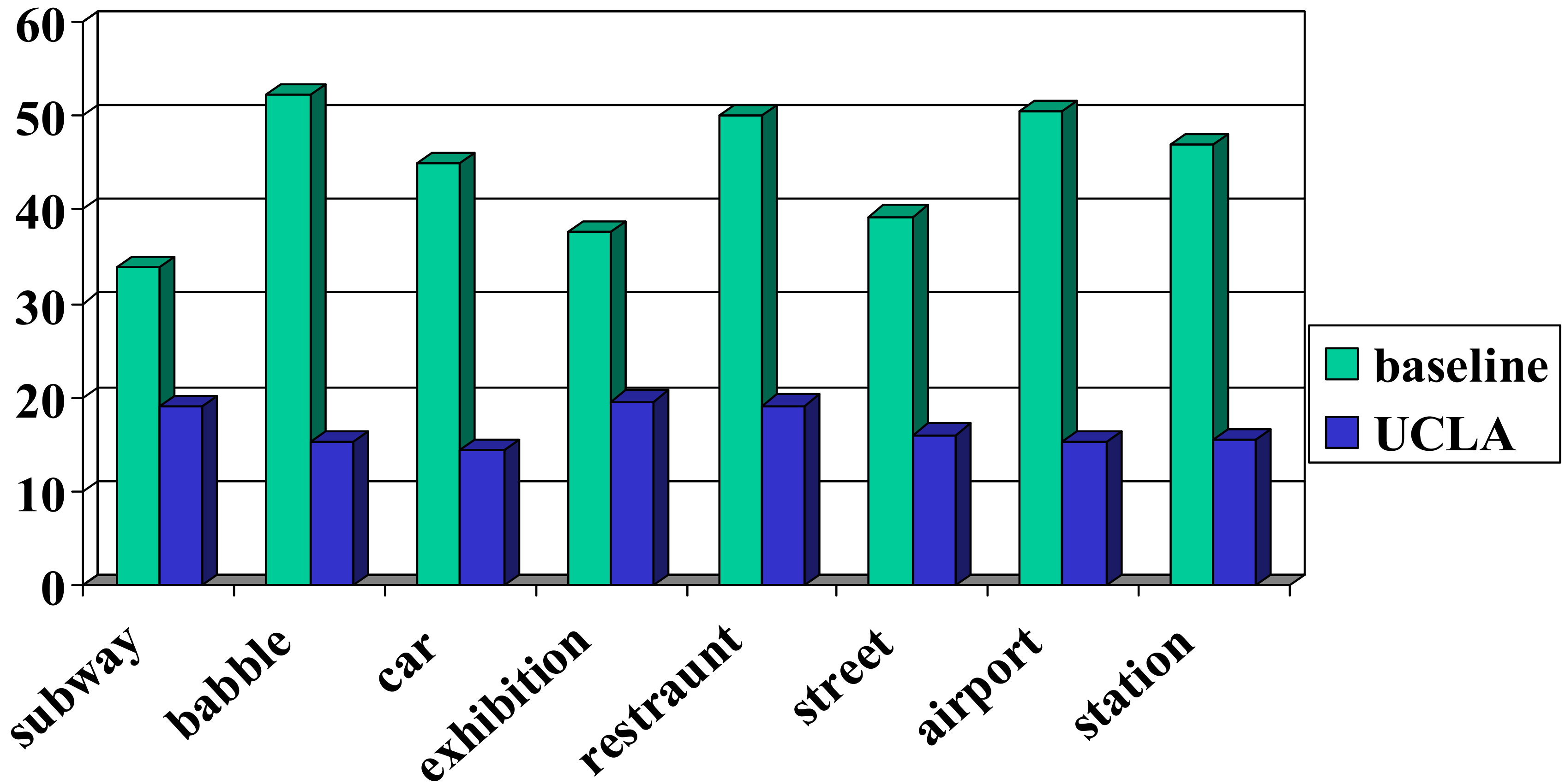

Speech waveform

Frames selected

Inter-frame distance d(i)

Frame selection in VFR for a digit string “one two” with silence.

# Aurora II Clean Training Results in Word Error Rate (%)



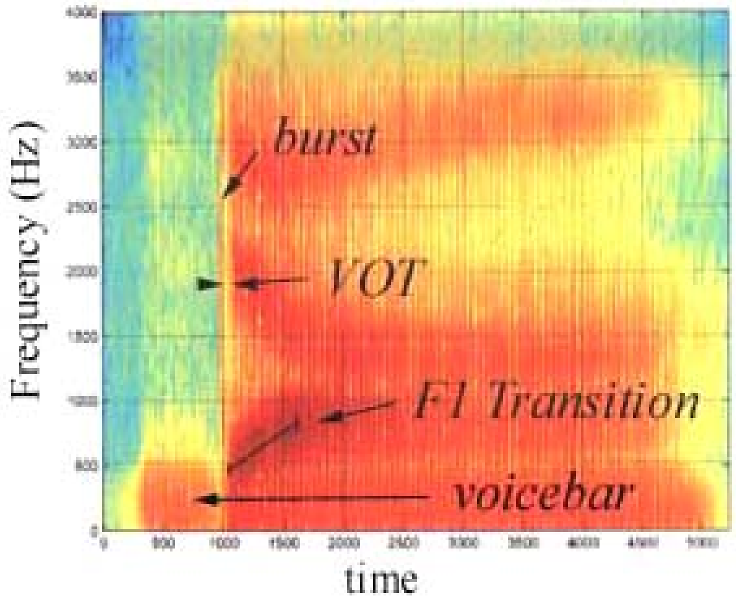
# *Part I Summary*

- Modeling aspects of human audition and knowing what is important in the speech signal can improve ASR in noise. However, we are yet to match human performance in noise!
- Speaker adaptation techniques that utilize formant-like information have also been successful (Cui and Alwan, in revision)

# *Part II: Phonological Features*

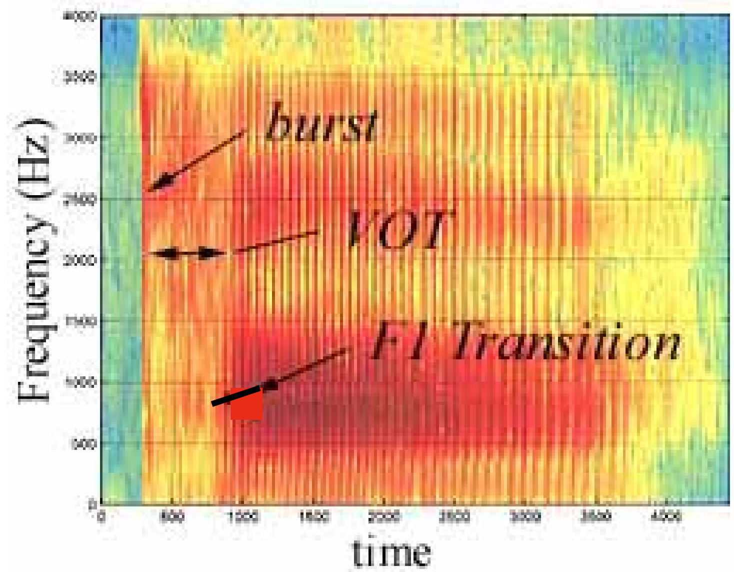
- Sounds can be characterized by a small number of constituents or features (Jakobson et al., 1963; Chomsky and Halle, 1968).
- The mapping from the linguistic domain to the acoustic domain is not necessarily one-to-one.
- Miller and Nicely (1954) presented /Ca/ syllables in noise to listeners and examined how different consonants were perceived. They analyzed confusion matrices using information theoretic approaches assuming that the underlying features: voicing, place of articulation, nasality, affrication, and duration.

# Case Study: Voicing in Syllable–Initial Plosives (M. Chen and Alwan, 2000)




/da/

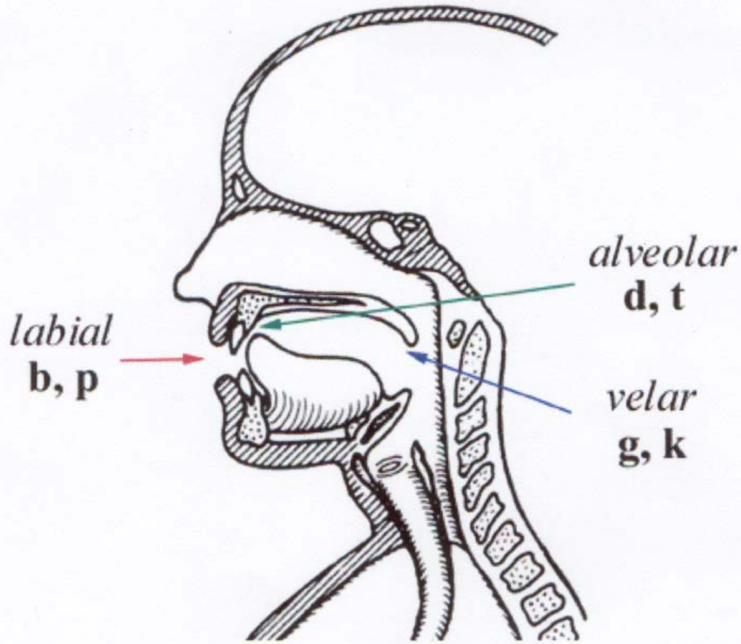
/ta/



# SPEECH TOKENS

Manner of Articulation: *Plosives*

3 Places of Articulation:




Consonant-Vowel Tokens across 3 Vowels:
**/a/ /i/ /u/**

9 Pairs with Distinguishing Feature : **Voicing**

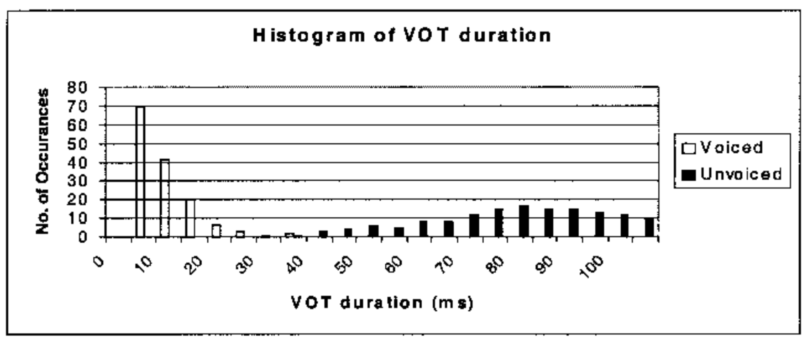
2 Male & 2 Female Talkers, 4 Repetitions Each per CV

Total: 16 Tokens per CV

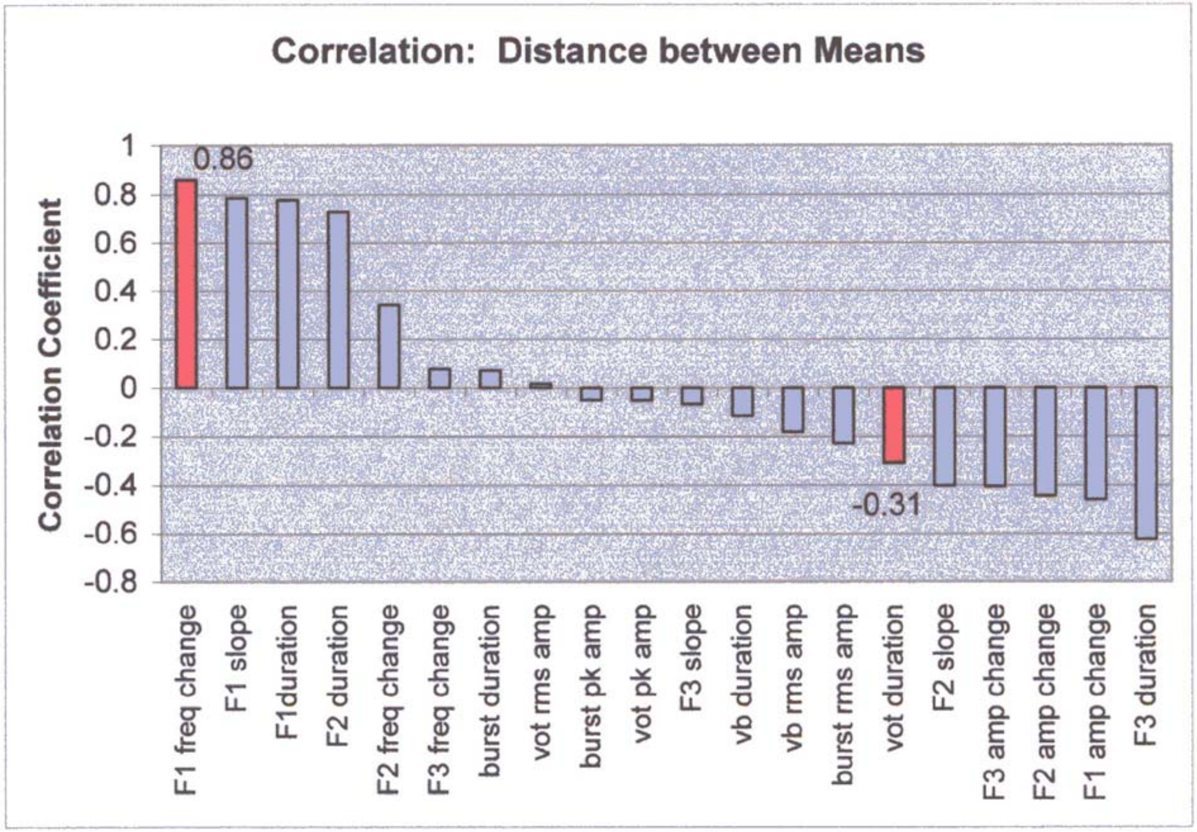
# Percent correct classification

| | voicebar duration | burst duration | VOT duration | F0 freq change |
|---|---|---|---|---|
| bapa | 71.9% | 68.8% | **100.0%** | 65.6% |
| data | 75.0% | 75.0% | **100.0%** | 68.8% |
| gaka | 65.6% | 68.8% | **100.0%** | 65.6% |
| bipi | 75.0% | 62.5% | **100.0%** | 65.6% |
| diti | 75.0% | 65.6% | **100.0%** | 71.9% |
| giki | 75.0% | 56.3% | **100.0%** | **81.3%** |
| bupu | 68.8% | 62.5% | **100.0%** | 68.8% |
| dutu | 75.0% | 59.4% | **100.0%** | 75.0% |
| guku | 78.1% | 59.4% | **100.0%** | 62.5% |

| | F1 | | | F2 | | | F3 | | |
|---|---|---|---|---|---|---|---|---|---|
| | duration | freq chg | slope | duration | freq chg | slope | duration | freq chg | slope |
| bapa | 75.0% | **100.0%** | **96.9%** | 65.6% | 62.5% | 59.4% | 65.6% | 71.9% | 65.6% |
| data | **93.8%** | **100.0%** | **87.5%** | **81.3%** | **84.4%** | 75.0% | 62.5% | **87.5%** | **81.3%** |
| gaka | **81.3%** | **100.0%** | **93.8%** | **87.5%** | **100.0%** | 78.1% | 65.6% | 62.5% | 59.4% |
| bipi | 75.0% | **84.4%** | 71.9% | 62.5% | **87.5%** | **87.5%** | **81.3%** | **90.6%** | **93.8%** |
| diti | **84.4%** | **81.3%** | 75.0% | 71.9% | **84.4%** | **87.5%** | 68.8% | 78.1% | **87.5%** |
| giki | 65.6% | 59.4% | 62.5% | 59.4% | 71.9% | 71.9% | 68.8% | 59.4% | 62.5% |
| bupu | **84.4%** | 78.1% | 75.0% | 62.5% | 75.0% | 75.0% | 68.8% | 68.8% | 62.5% |
| dutu | 68.8% | 65.6% | 59.4% | 59.4% | 75.0% | 75.0% | 68.8% | 62.5% | 62.5% |
| guku | 71.9% | 65.6% | 75.0% | 56.3% | 56.3% | 59.4% | 62.5% | 62.5% | 56.3% |



# CORRELATION BETWEEN ACOUSTIC FEATURES & PERCEPTUAL THRESHOLDS

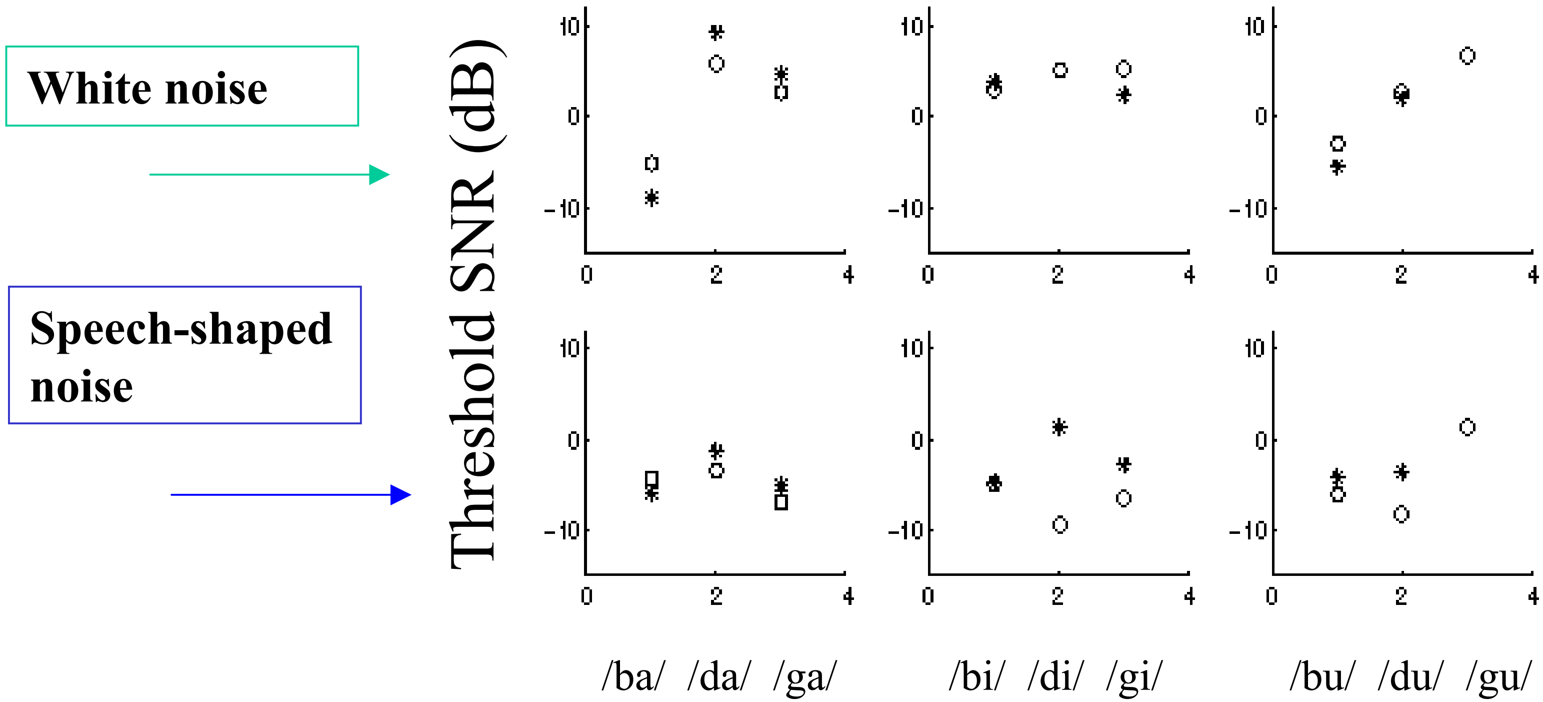



- Highest correlation with F1transition
  *(0.86 for F1 frequency change)*
- No apparent correlation with VOT
  *(-0.31 for VOT duration)*

# Summary

- **For syllable-initial plosives, voicing is clearly manifested by differences in the VOT. Differences in F1 results in perfect classification for only the /Ca/ syllables.**
- **In noise, robustness of the voicing feature is dependent on the vowel context. Lower thresholds were highly correlated with differences in the F1 transition which are most dominant for the /Ca/ syllables.**
- **Temporal cues, such as differences in burst duration and amplitude appear to be secondary cues**

# *The effect of the noise masker shape (Alwan, 1992; Hant and Alwan, 2000)*

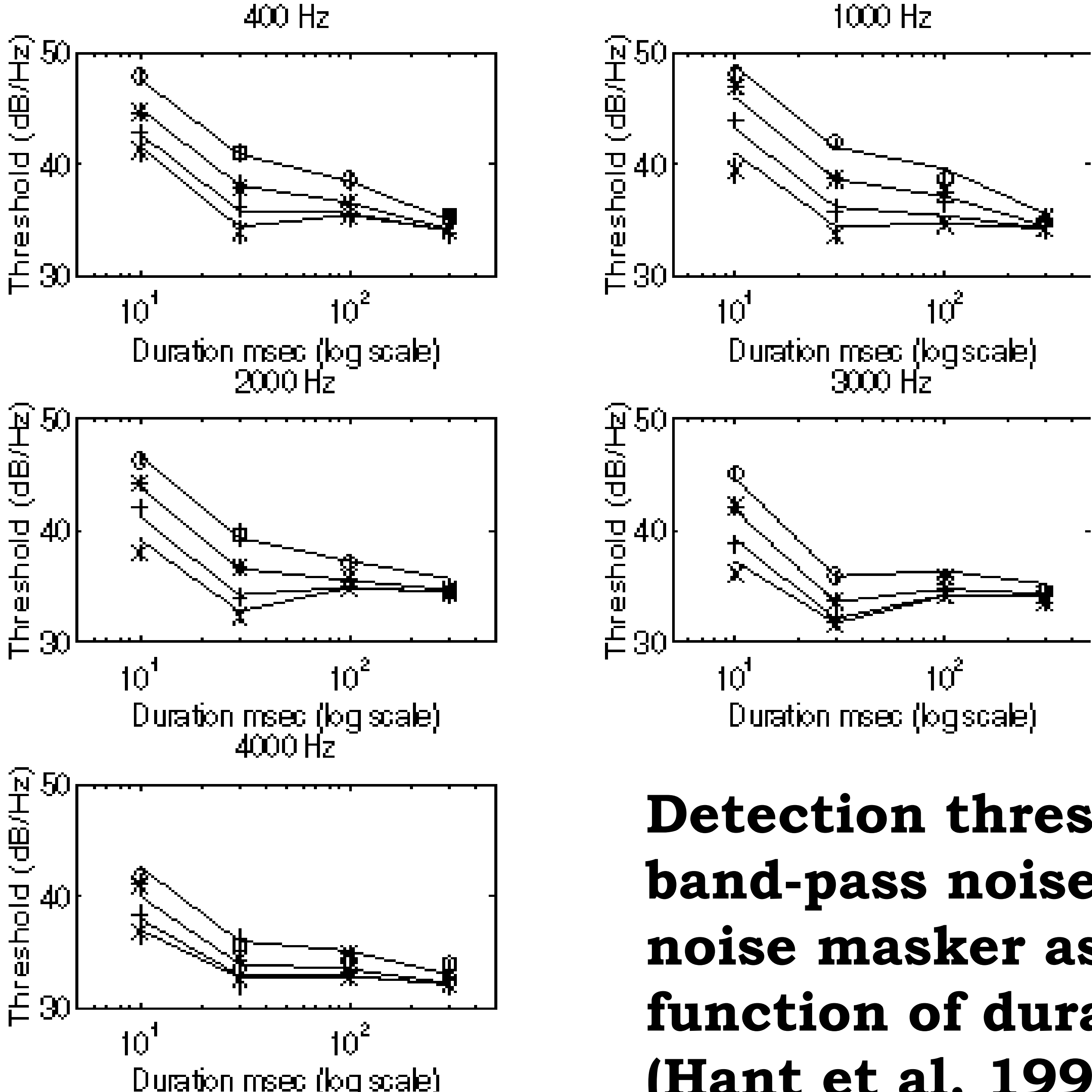



O CVs with burst

* CVs with no burst

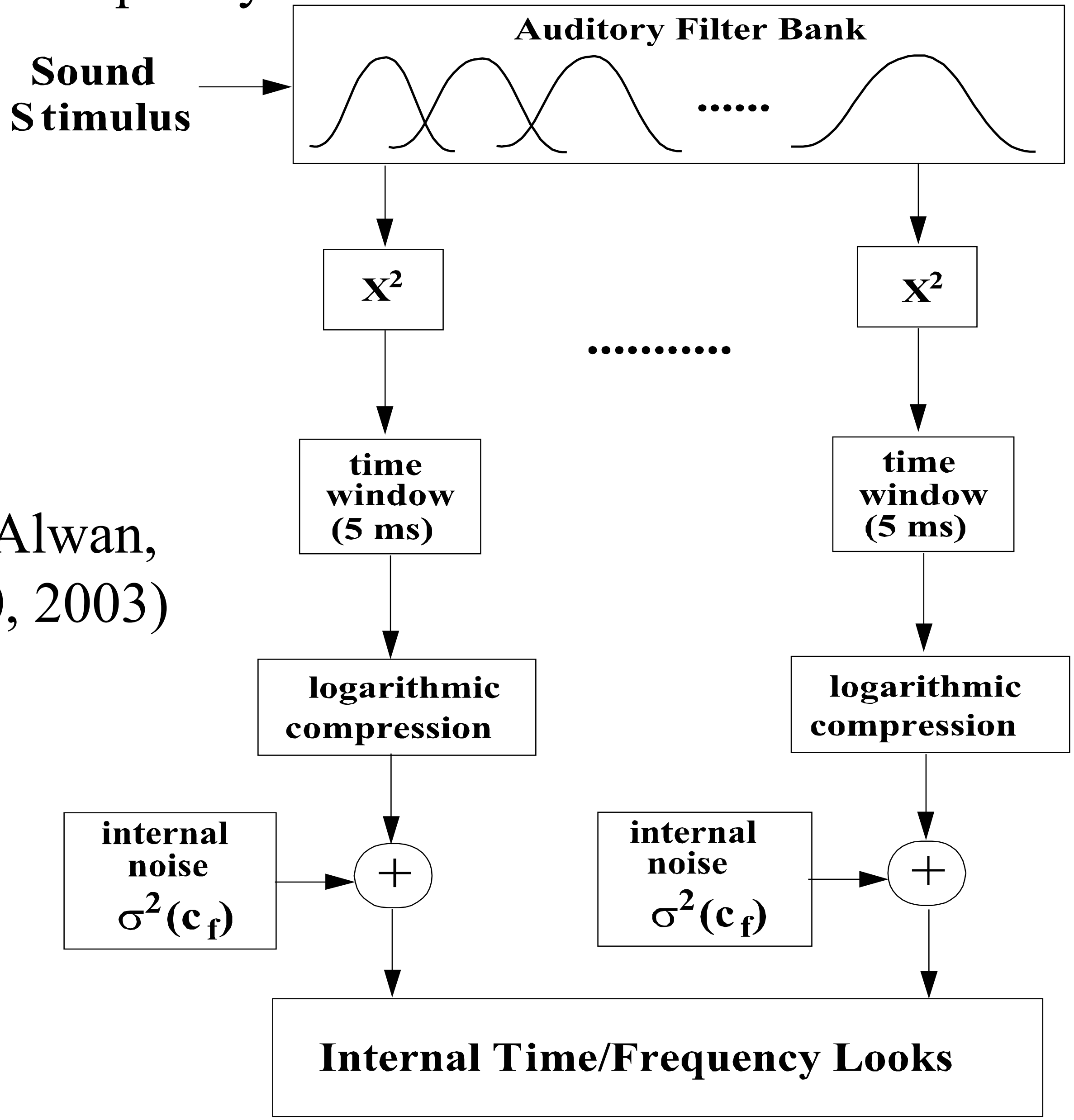
# *Part III*

- Traditional masking models focus on long-duration, narrow-band signals
- To predict noise-masking of wide-band and non-stationary signals such as speech, the effects of signal duration and bandwidth need to be taken into account
- Previous models: temporal integration (Plomp et al., 1959), multi-band excitation model (Plomp, 1970), multi-look in time model (Viemeister & Wakefield, 1991), duration-dependent filters (Hant et al, 1997), and multi-look in time and frequency model (van Schijndel et al., 1999).

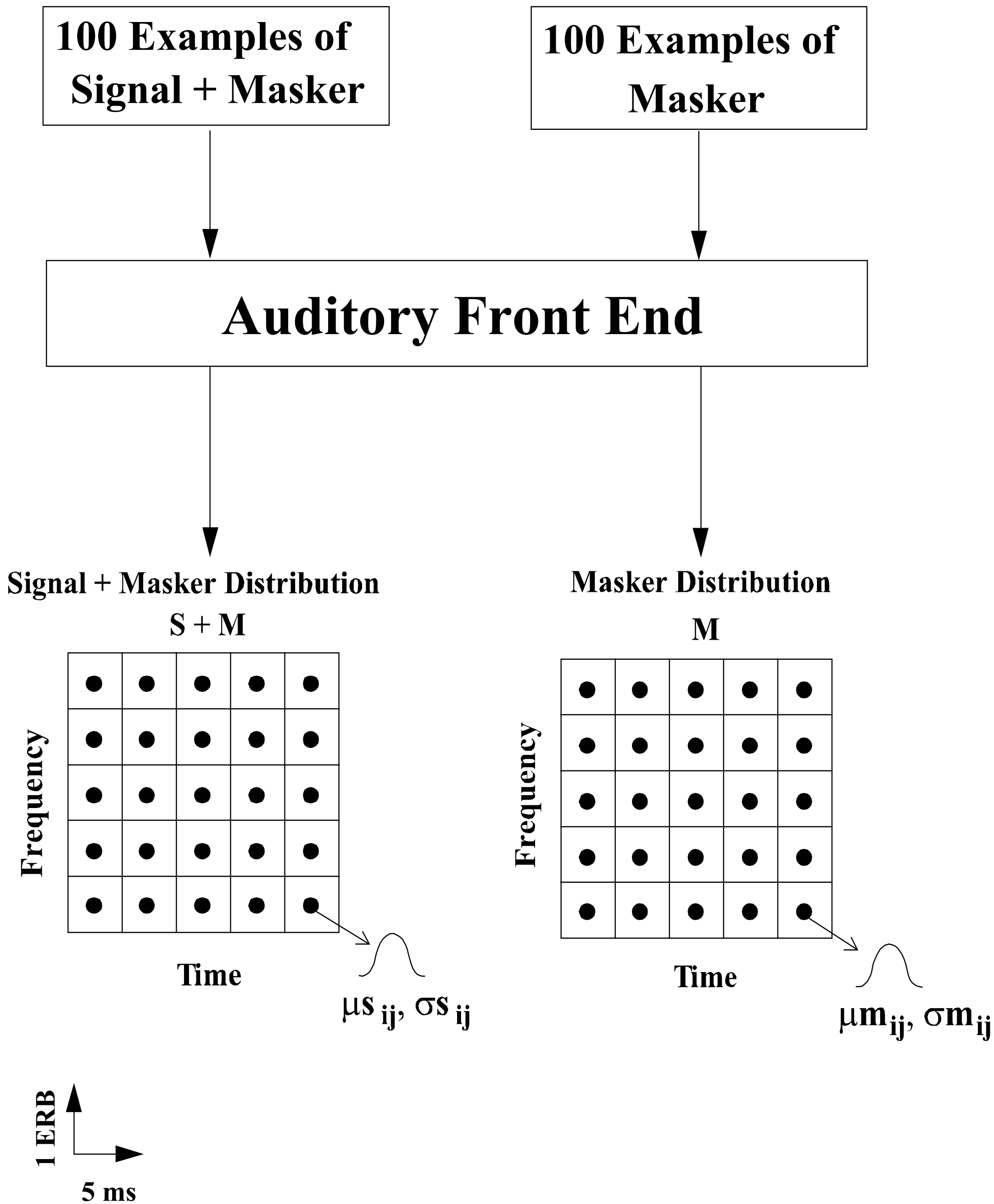



**Detection thresholds of band-pass noises in a noise masker as a function of duration (Hant et al. 1997)**

# A time-frequency model




(Hant and Alwan, 1999, 2000, 2003)


100 Examples of Signal + Masker
100 Examples of Masker
Auditory Front End
Signal + Masker Distribution
S + M
Frequency
Time
μs ij, σs ij
Masker Distribution
M
Frequency
Time
μmij, σmij
1 ERB
5 ms

# *Model Predictions*

Model predicts well the noise masking of:

• bandpass noises of various durations (10-300 ms), bandwidths (1-8 CB), and center frequencies (400-4000 Hz),

•tone glides and synthetic formant transitions with durations varying between 10-100 ms, and frequencies between 300-4000 Hz (except for 1500 Hz, 100 ms glides), and

•stop bursts.

The model also predicted well the discrimination of synthetic CV syllables (/bV, dV, gV/ in 3 vowel contexts (/a/, /i/, /u/) and 2 noise maskers (exception: /bi,di/).

# Acknowledgements

Former and Current Students: Alexis Bernard, Willa Chen, Marcia Chen, Xiaodong Cui, James Hant, Markus Iseli, Jintao Jiang, Brian Strope, Hong You, and Qifeng Zhu.

Work supported in part by the NSF and the NIH.