

The Noisy Speech Chain

Abeer Alwan

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

<http://www.icsl.ucla.edu/~spapl>

alwan@ee.ucla.edu

Improving Intelligibility of 'Competing Messages'

- Staggering onsets (Webster et al., 1954)
- Localization (Spieth et al., 1954)
- Pitch differences (Treisman, 1964)
- Filtering (Spieth et al., 1954)
- Differences in level and voice characteristics (Brungart and Simpson, 2001)

Can more explicit knowledge of speech perception and production be exploited to improve intelligibility?

Clear Speech

‘Clear’ speech is characterized by a **reduced speaking rate** (Picheny, 1986.)

Krause and Braida (2004): with training, speakers can produce clear speech characterized by an energy **increase in the 1-3 kHz range**. Some speakers also increase the depth of LF modulations in the intensity envelope and/or manifest phonetic differences (e.g., VOT).

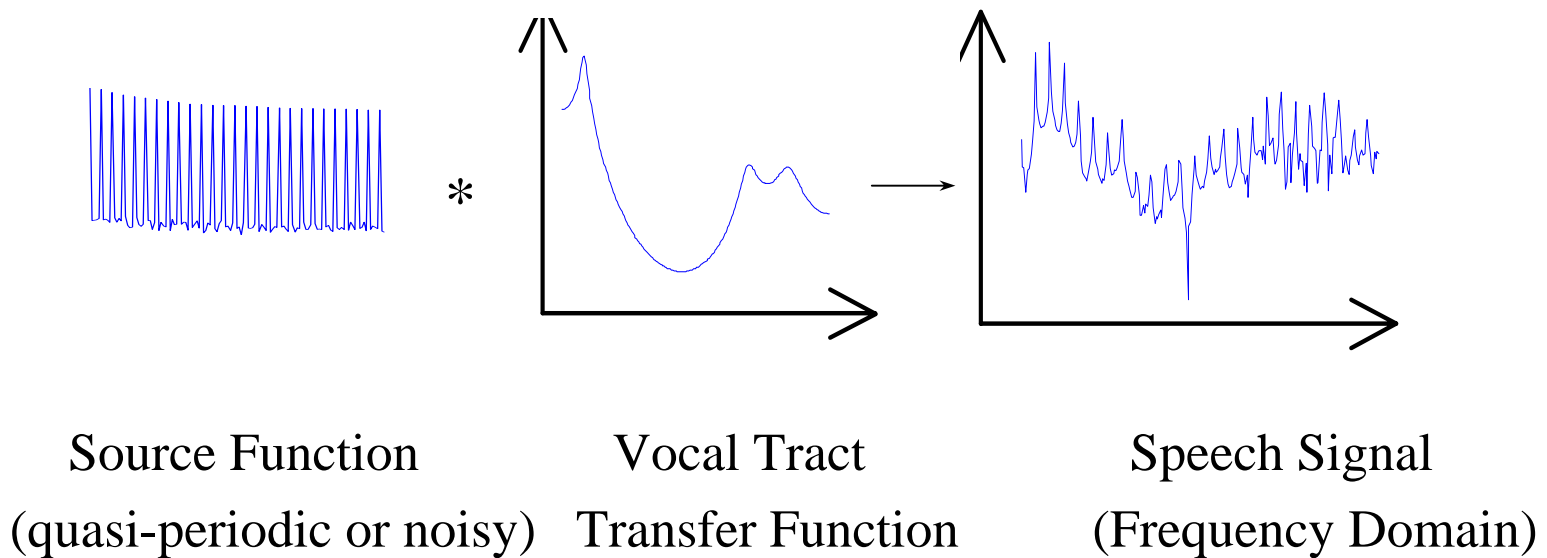
Greenberg and Arai (2004): intelligibility depends on the integrity of **modulation spectrum at 3-10 Hz** (core range of the syllable).

Speaker Differences

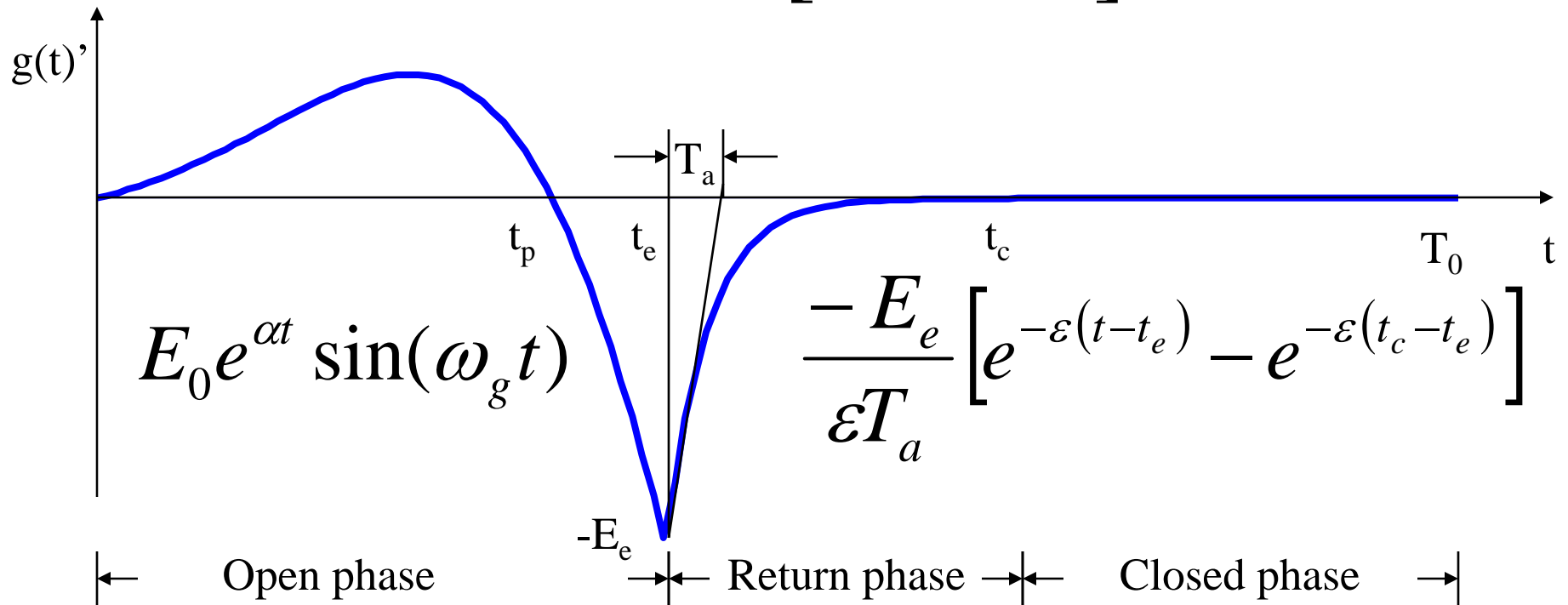
- **Physiological:** related to properties of the vocal folds and vocal tract
- **Behavioural and Linguistic:** dialect/accents/pronunciation, choice of words, relative frequency of disfluencies, laughter, prosodic patterns (**energy, pitch, and duration, phone- and syllable-based**)

Prosody/accents affect temporal and spectral cues. Speaker recognition by humans and machines exploits these differences.

LTI Model of Speech Production



The Liljencrants-Fant (LF) Source Model [Fant85]



$$F_0 = \frac{1}{T_0}$$

$$OQ = \frac{t_e + T_a}{T_0}$$

Fundamental Frequency


- Fundamental Frequency (F0) reflects the quasi-periodicity of vocal folds' vibration for voiced sounds



$$T_0 = 1/F_0$$

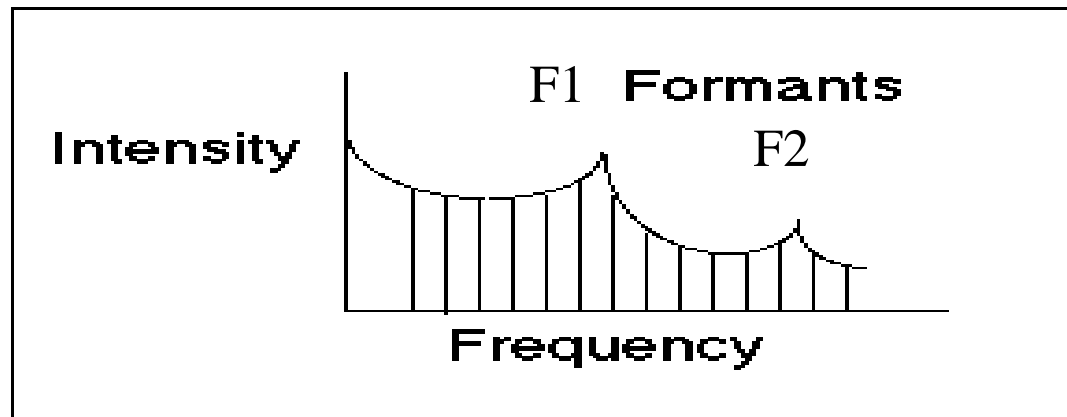
	Male	Female	Child
F0 (Hz)	125	225	300

Source Parameters

- F0 is correlated with age, gender, and emotion
- Other source parameters are related to the voice quality but are not well understood. OQ is related to breathiness of the voice. 
- Temporal aspects of the source are also important (jitter and shimmer)
- Some of the properties of the glottal shape/derivative have been used in speaker recognition experiments (Plumpe et al., 1999)

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.
- Relative locations of the formants are related to voice quality (Story et al., 2003).

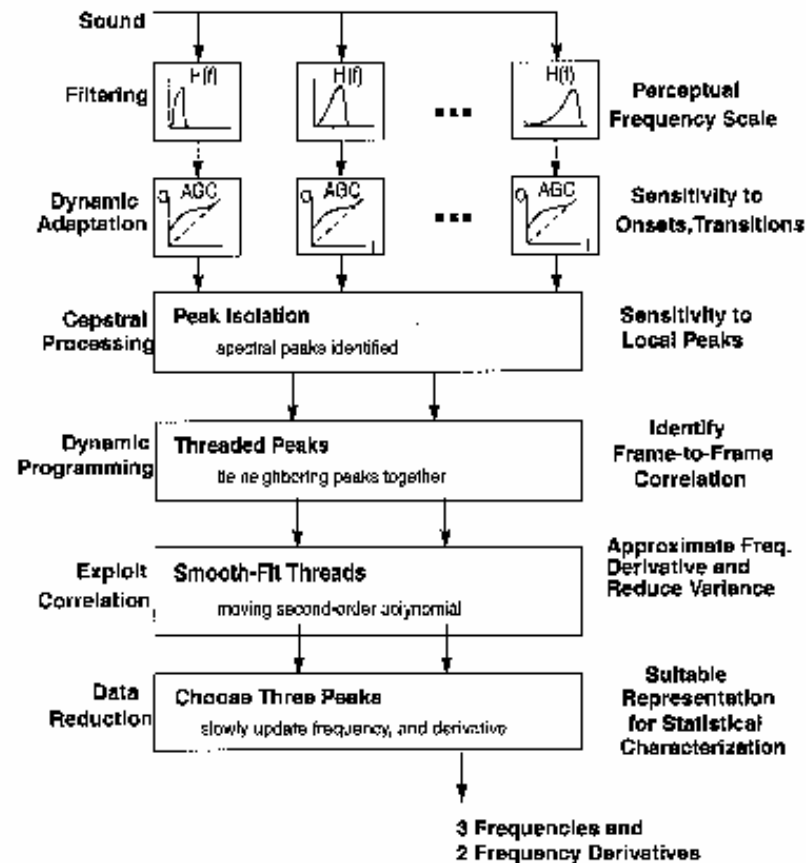


Zeros

Zeros in the transfer function occur when energy is trapped in the back, side, or sublingual cavities of the vocal tract, or in the front cavity in the case of nasals.

Since the articulators move at a slow rate, expect the VTTF to change slowly.

Overall System



Focus: adaptation and temporal correlations of local spectral peaks.

(Strope and Alwan, 1997)

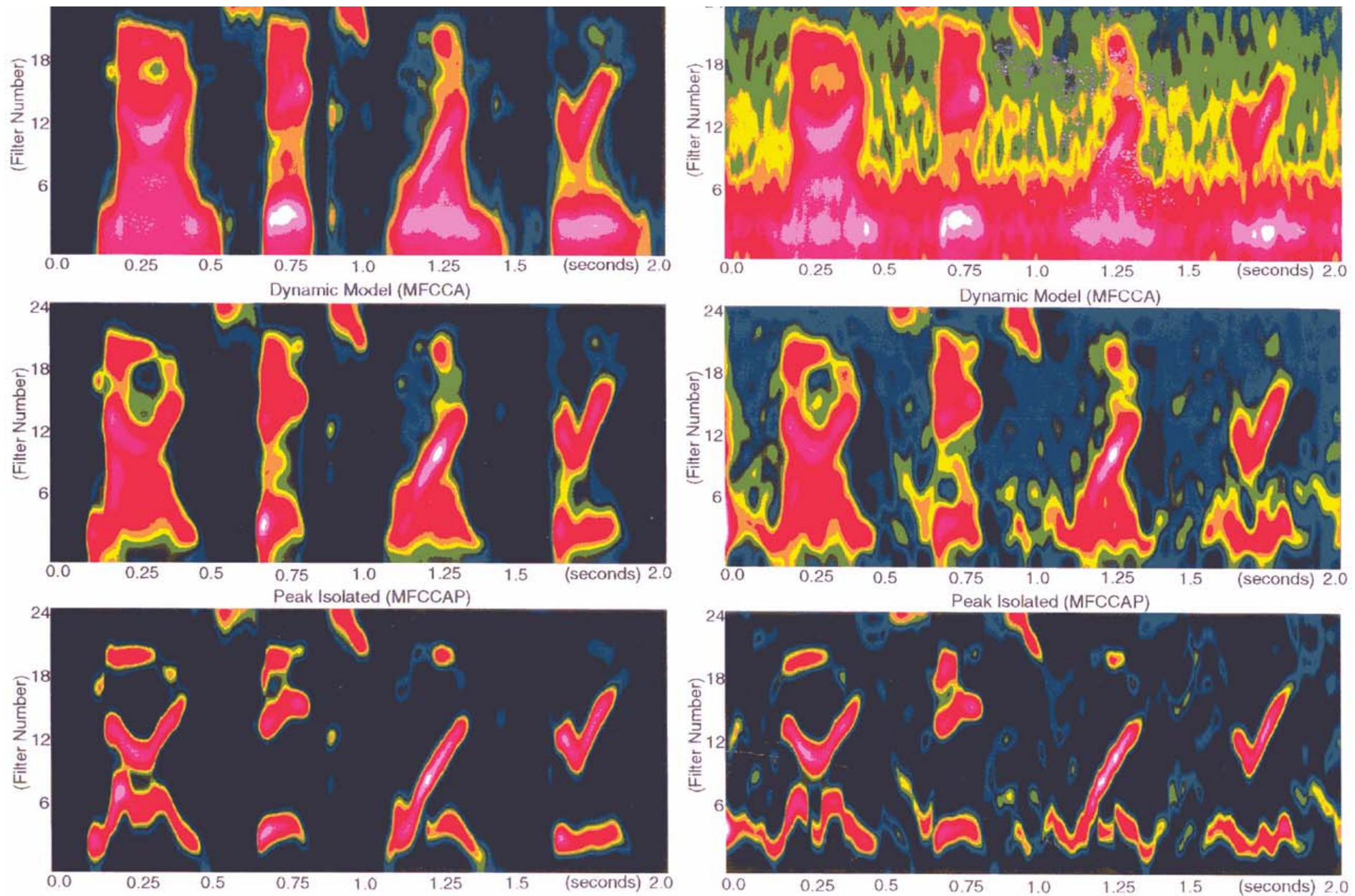


Fig. 11

Strope and Alwan, 1997

(Strope and Alwan, 1998)

These techniques improved ASR in noise significantly.

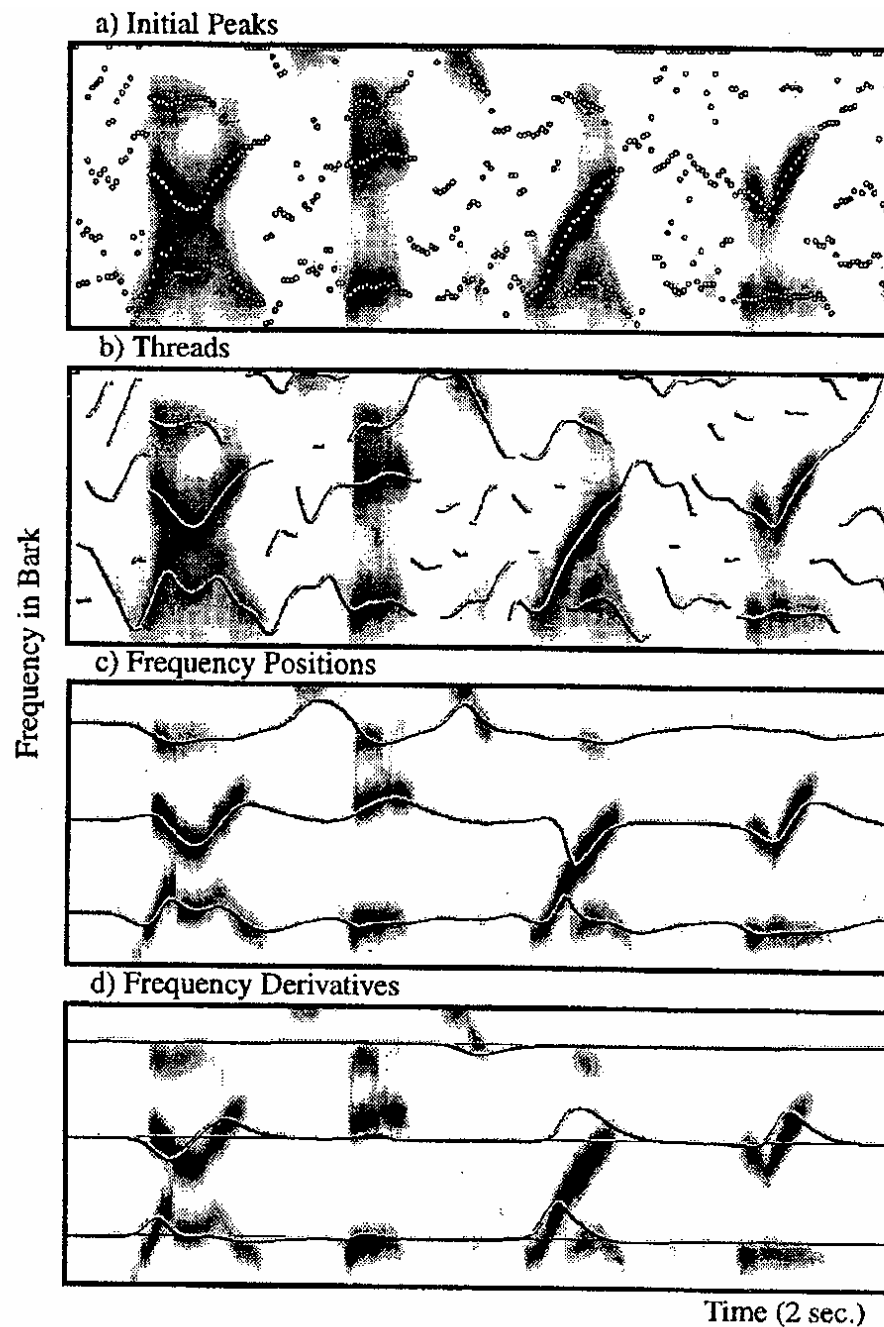
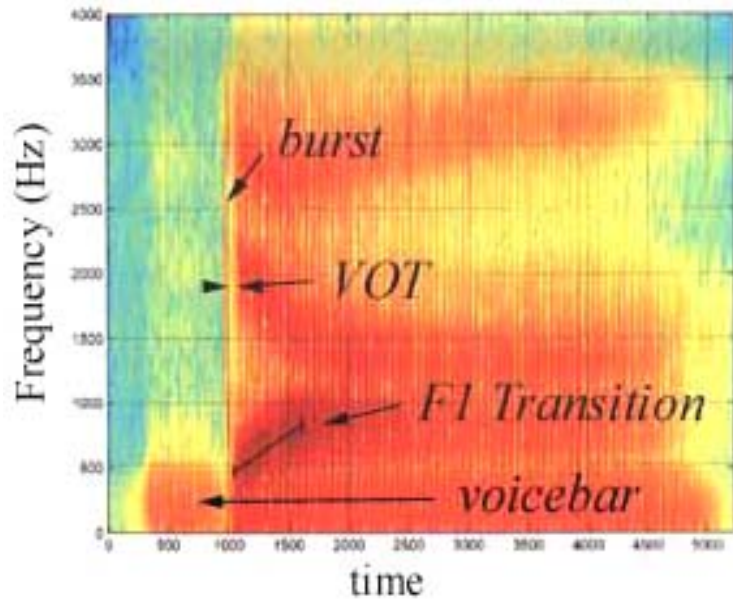


Figure 2. Peak positions and motion.

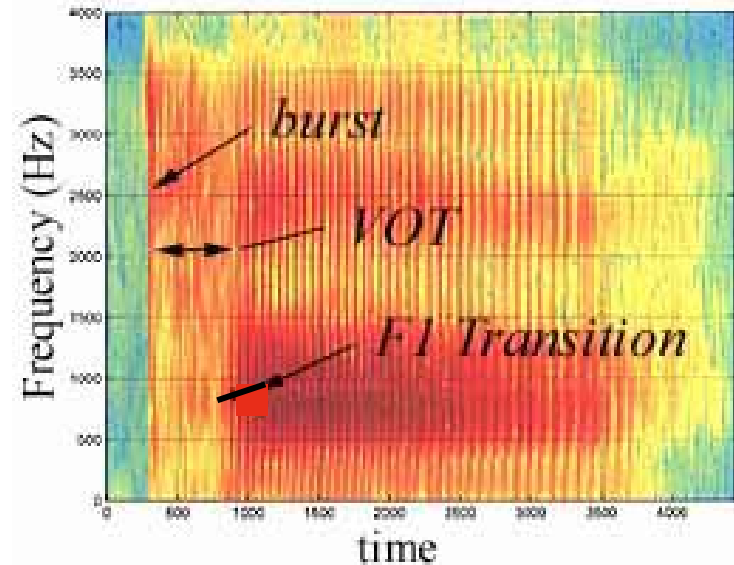
Phonological Features

- 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.
- Q:
 - **Which acoustic cues account for differences, if any, in perceptual thresholds?**
 - **How does the perception of a feature vary with noise level (SNR)?**
 - **Does the threshold for perceiving a consonantal feature in noise vary with vowel context?**

Case Study I: Voicing in Syllable-Initial Plosives (M. Chen and Alwan, 2001)

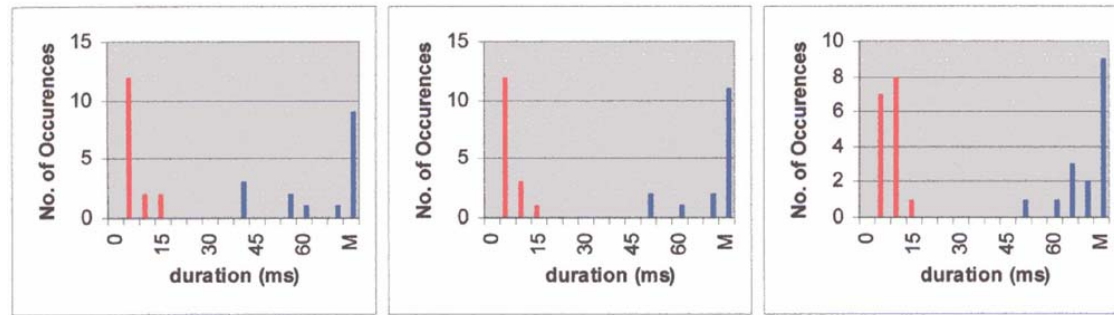


/da/



/ta/

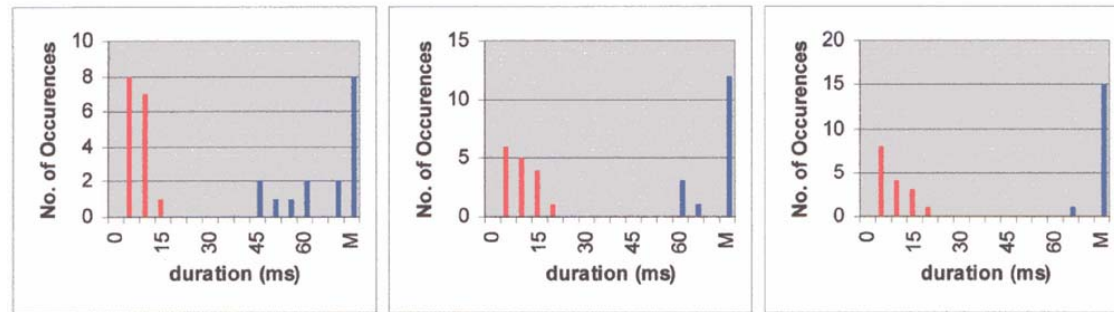
HISTOGRAMS: VOT DURATION



ba-pa

bee-pee

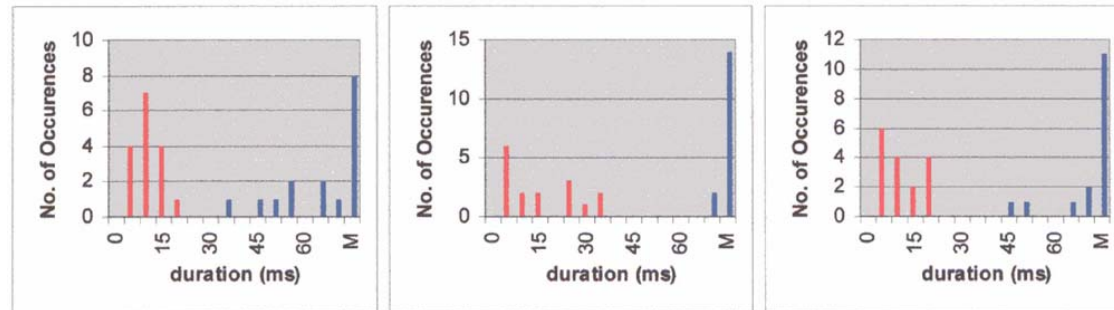
boo-poo



da-ta

dee-tee

doo-too



ga-ka

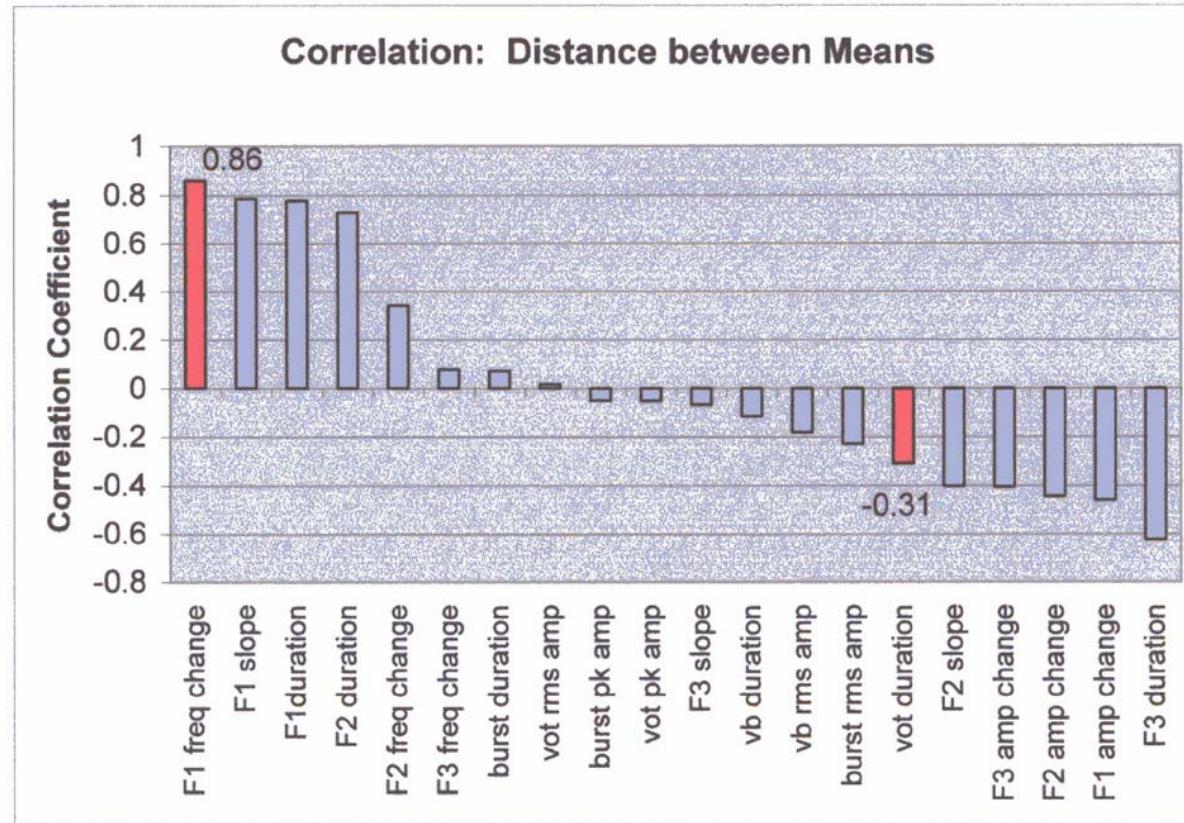
gee-kee

goo-koo

■ Voiced (0 – 20 ms)

■ Unvoiced (45 -120 ms)

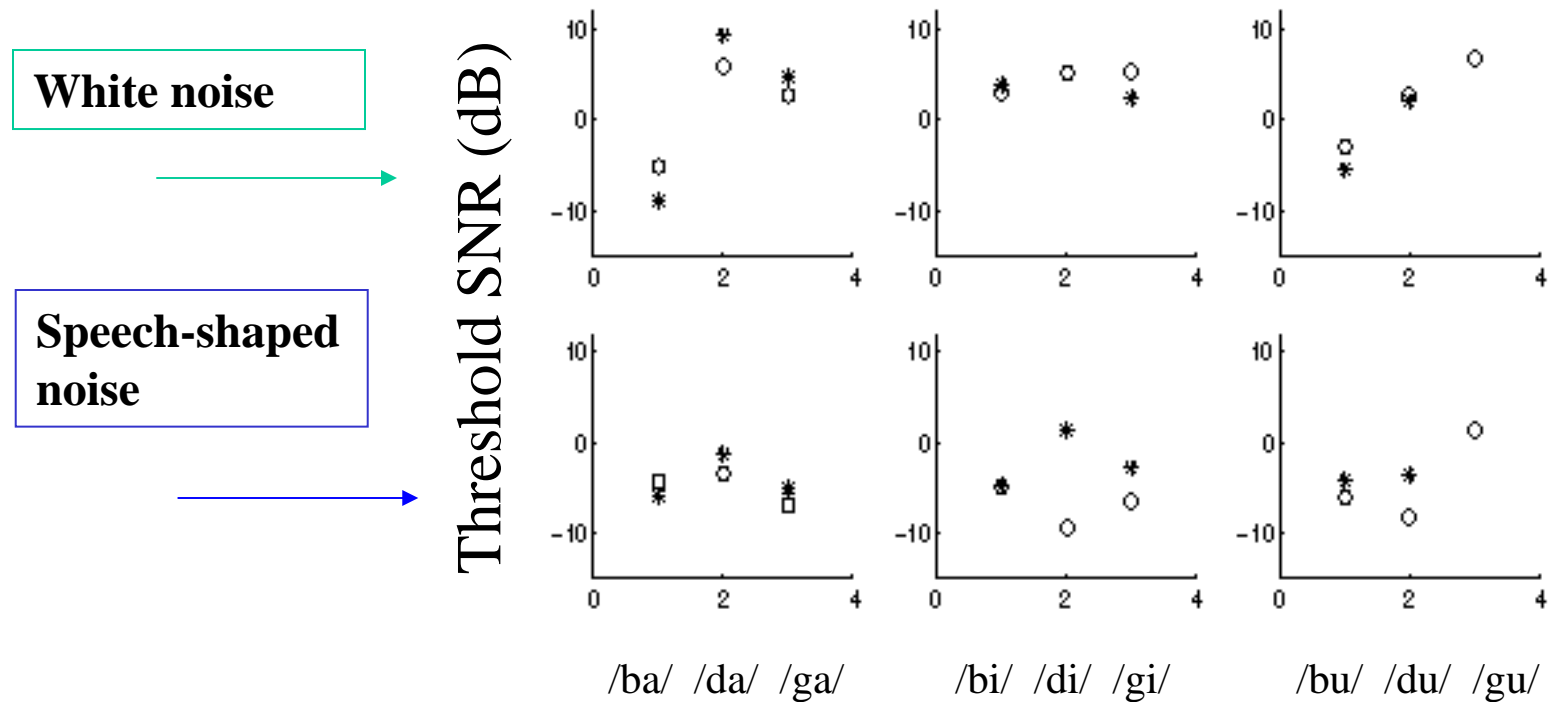
CORRELATION BETWEEN ACOUSTIC FEATURES & PERCEPTUAL THRESHOLDS



(M. Chen
and
Alwan,
2001)

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

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

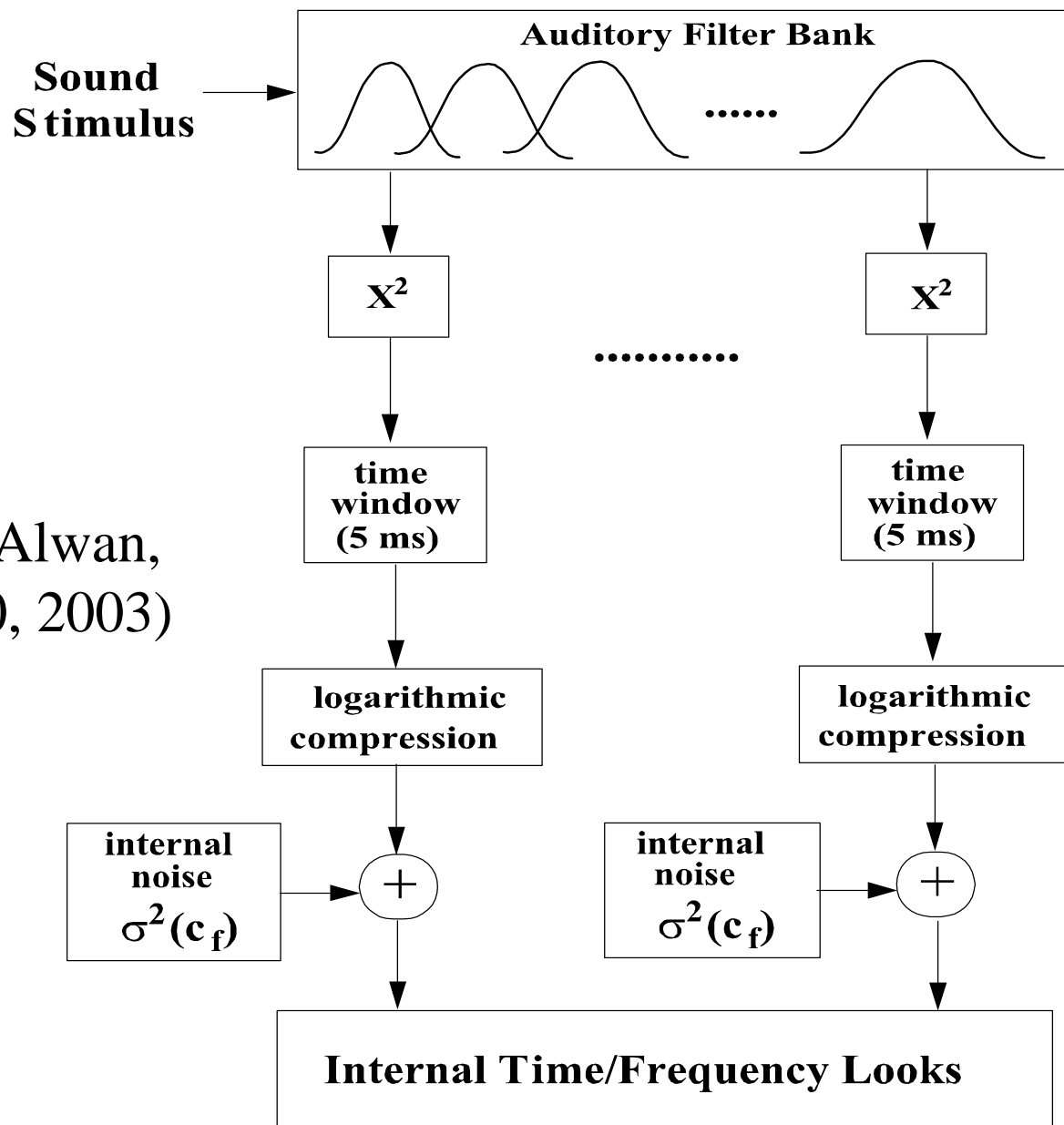


O CVs with burst

* CVs with no burst

Case Study II: Discriminative Acoustic Features and Perceptual Thresholds for the Place Feature (W. Chen and Alwan, 2003)

	/ba,da/	/bi,di/	/bu,du/
Feature	F2 Δ , 100%	Av-Ahi, 93.75%	Av-Ahi, F3 Δ , 90.63%
Percept. threshd.	-7.3	4	-1.6
	/pa,ta/	/pi,ti/	/pu,tu/
Feature	Burst Dur., 96.88%	Ahi-A23, 96.88%	Av-Ahi, 100%
Percept. threshd	6.7	.12	0
	/va,za/	/vi,zi/	/vu,zu/
Feature	F1 onset, 100%	Av-Anoise, 96.88%	Av-Anoise, 96.88%
Percept. threshd.	-4.5	-1.2	-3.4
	/fa,sa/	/fi,si/	/fu,su/
Feature	F1 onset, 100%	Av-Anoise, 93.75%	Av-Anoise, 100%
Percept. threshd.	-5	-3.8	-5



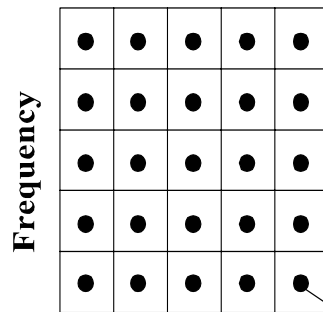
(Hant and Alwan,
1999, 2000, 2003)

100 Examples of
Signal + Masker

100 Examples of
Masker

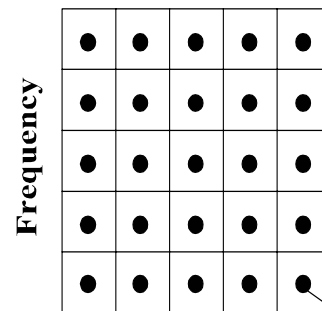
Auditory Front End

Signal + Masker Distribution
S + M



$\mu_{S_{ij}}, \sigma_{S_{ij}}$

Masker Distribution
M



$\mu_{M_{ij}}, \sigma_{M_{ij}}$

1 ERB
5 ms


(Hant and
Alwan, 1999,
2000,2003)

Summary

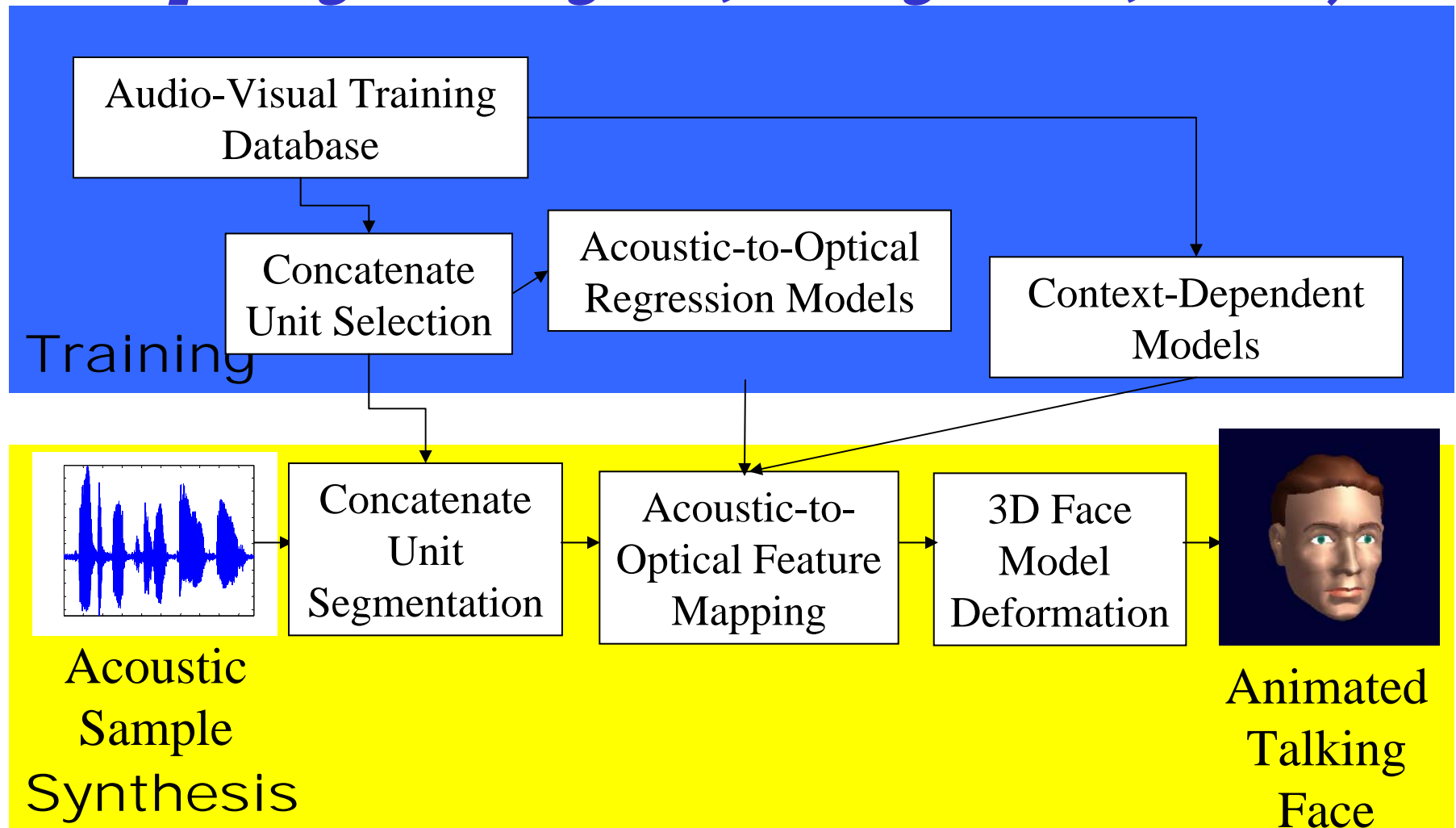
Acoustic cues which classify sounds accurately are not necessarily predictors of the noise robustness of corresponding features. Perceptual noise robustness of a feature depends on:

- **noise masker shape and level**
- **extent and amplitude of formant-frequency transitions (hence the large effect of vowel context and voicing)**
- **duration and relative amplitude of the burst and noise segments (hence, the effect of manner and place)**

Improving Intelligibility of Competing Messages

- Alter the source: whisper, creak, falsetto, period doubling. Would not recommend whisper in noisy environments.
- Alter the VTTF: extra nasality, gender change -if preserving speaker ID is not an issue-. 
- Vary prosodic cues: use a different dialect or an intelligible foreign accent. Vary speaking rate.
- Manipulate the modulation spectrum

Acoustically-driven Visual Speech Synthesis (note that not all faces are equally intelligible; Jiang et al., 2002)



Summary

- Capturing prosodic information (beyond F0) and fine-detailed characteristics can be further exploited as well as AV perception.
- Need to know whether perceiving speech monaurally or binaurally, and the SNR.
- Other relevant speech processing literature/techniques:
 - i. analysis-by-synthesis techniques
 - ii. voice transformation/morphing
 - iii. speaker recognition

Summary

iv. Lombard speech (speech spoken in the presence of background noise)

Acknowledgements: Former and Current Students: Willa Chen, Marcia Chen, James Hant, Markus Iseli, Jintao Jiang, Brian Strobe, and Jane Xue.

Work supported in part by the NSF and the NIH.

SPAPL References

- W. Chen and A. Alwan, "Perception of the Place of Articulation Feature for Plosives and Fricatives in Noise," in Proc. ICPHS, Barcelona, August, 2003
- J. Hant and A. Alwan, "[A Psychoacoustic-Masking Model to Predict the Perception of Speech-Like Stimuli in Noise](#)," *Speech Communication*, Vol. 40, May 2003, pp. 291-313.
- Q. Zhu and A. Alwan, "[Non-linear feature extraction for robust recognition in stationary and non-stationary noise](#)," *Computer, Speech, and Language*, 17(4): 381-402, Oct. 2003
- J. Jiang, A. Alwan, P.A. Keating, E.T. Auer, and L.E. Bernstein, "[On the relationship between face movements, tongue movements, and speech acoustics](#)," special issue of *EURASIP Journal on Applied Signal Processing* on joint audio-visual speech processing, Nov. 2002, pp.1174-1188.
- M. Chen and A. Alwan, "[On the Perception of Voicing for Plosives in Noise](#)," Proc. Eurospeech 2001, Aalborg, Denmark, Vol. 1, pp. 175-178.
- J. Hant and A. Alwan, "[Predicting the Perceptual Confusion of Synthetic Stop Consonants in Noise](#)," 6th International Conference on Spoken Language Processing, ICSLP 2000. Vol. 3, pp. 941-944
- J. Hant and A. Alwan, "Modeling the Masking of Formant Transitions in Noise," Proc. Eurospeech 1999, Vol. 4, pp. 1895-1898.

References Cont'd.

- B. Strobe and A. Alwan, "[Robust Word Recognition Using Threaded Spectral Peaks](#)," Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Seattle, Vol. II, pages 625-629, May 1998
- B. Strobe and A. Alwan, "[A model of dynamic auditory perception and its application to robust word recognition](#)," IEEE Transactions on Speech and Audio Processing, Vol. 5, No. 5, pp. 451-464, September 1997

Other publications can be found on www.icsl.ucla.edu/~spapl