Paralinguistic Information Processing

- **Prosody**
  - **Pitch tracking**
  - **Intonation, stress, and phrase boundaries**
  - **Emotion**
- **Speaker Identification**
- **Multi-modal Processing**
  - **Combined face and speaker ID**
  - **Lip reading & audio-visual speech recognition**
  - **Gesture & multi-modal understanding**

Prosody

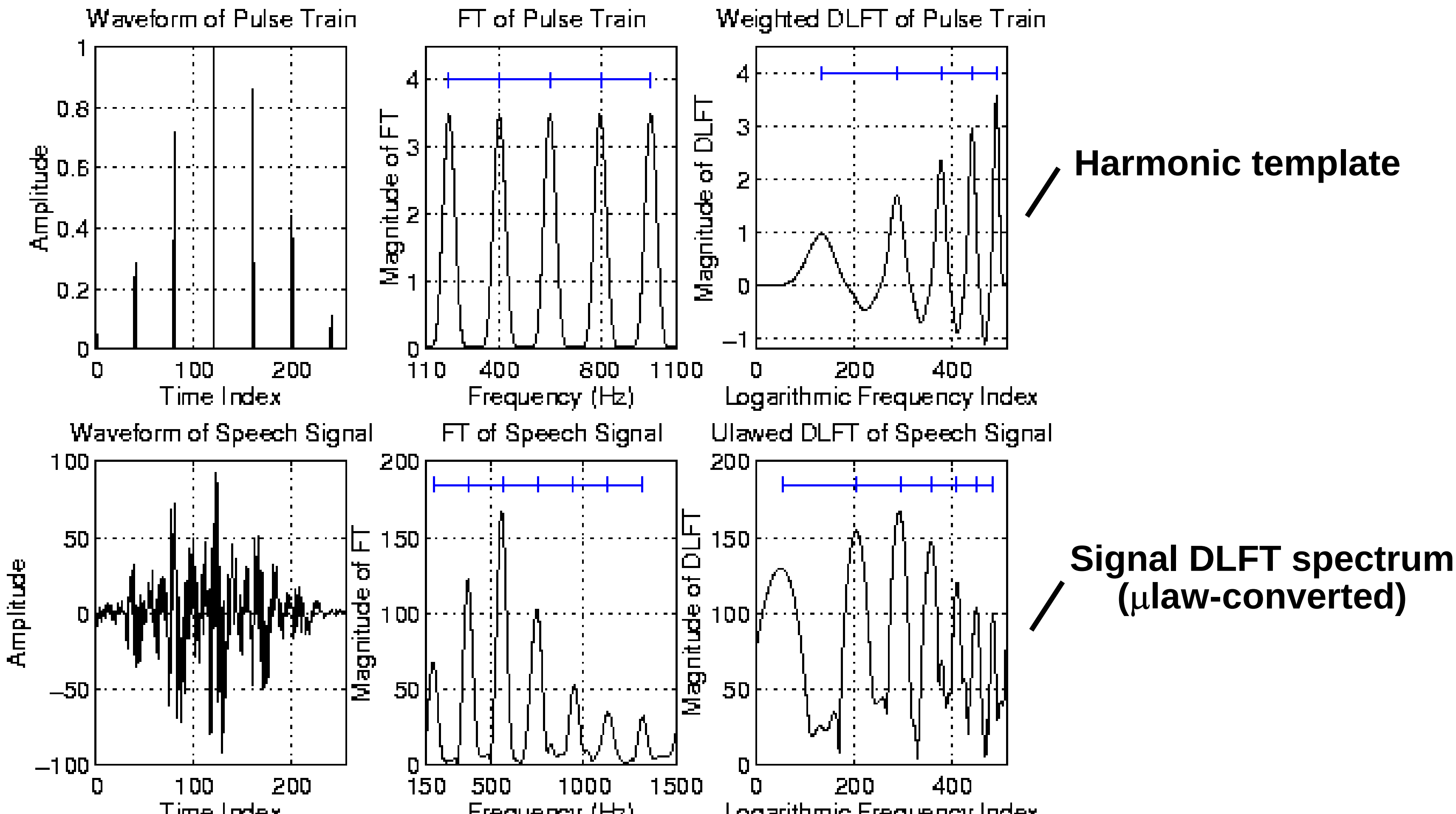
- Prosody is term typically used to describe the extra-linguistic aspects of speech, such as:
  - Intonation
  - Phrase boundaries
  - Stress patterns
  - Emotion
  - Statement/question distinction
- Prosody is controlled by manipulation of
  - Fundamental frequency (F_0)
  - Phonetic durations & speaking rate
  - Energy

Robust Pitch Tracking

- **Fundamental frequency (F_0) estimation**
  - **Often referred to as *pitch tracking***
  - **Crucial to the analysis and modeling of speech prosody**
  - **A widely studied problem with many proposed algorithms**
- **One recent two-step algorithm (Wang, 2001)**
  - **Step 1: Estimate F_0 and ΔF_0 frame each speech frame based on harmonic matching**
  - **Step 2: Perform dynamic search with continuity constraints to find optimal F_0 stream**

Discrete Logarithmic Fourier Transform

- **Logarithmically sampled narrow-band spectrum**
  - **Harmonic peaks have fixed spacing ($\log F_0 + \log N$)**
  - **Derive F0 and ΔF0 estimates through correlation**

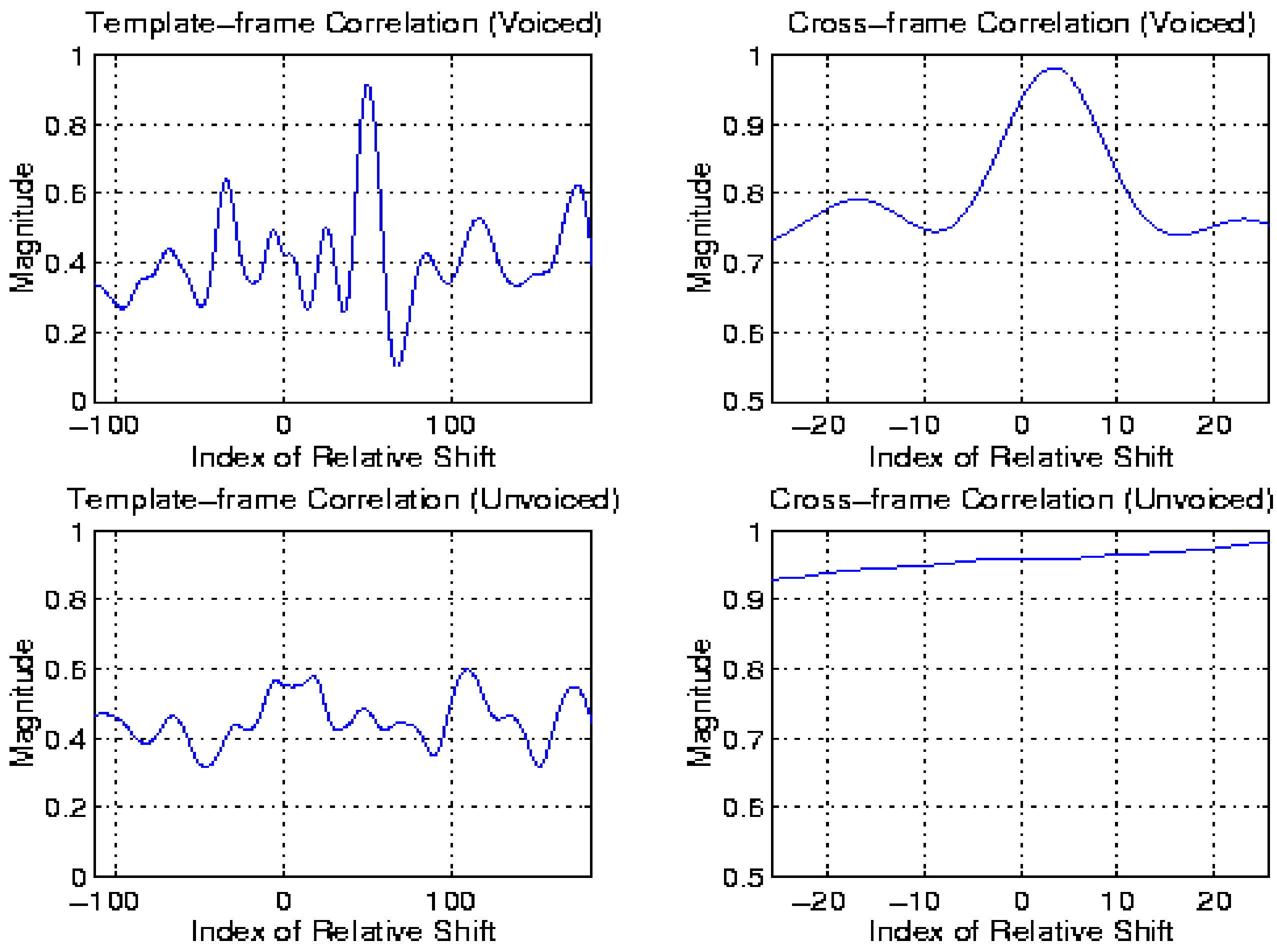


Two Correlation Functions

Template - Frame | Cross - Frame

Voiced

Unvoiced

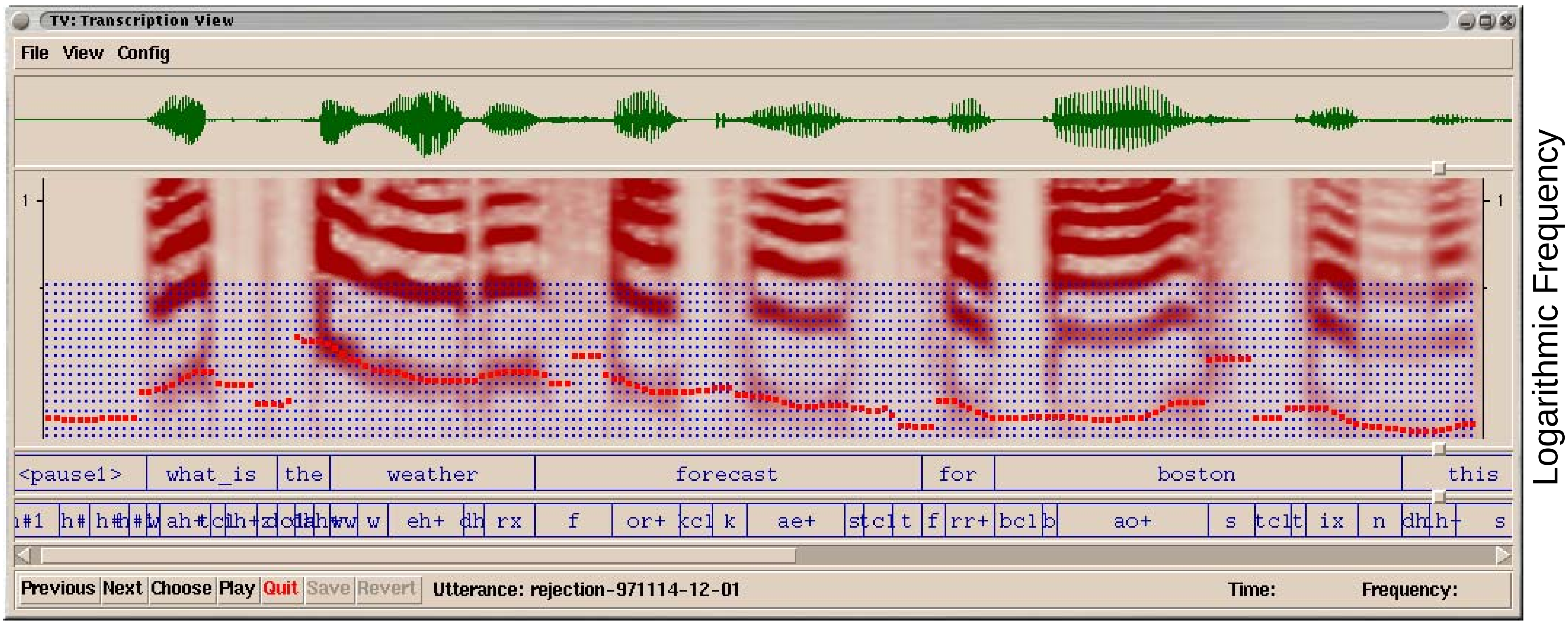


$$R_{Tx\,t}(n) = \frac{\sum_i T(i) X_t(i-n)}{\sqrt{\sum_i X_t(i)^2}}$$

$$R_{X_t X_{t-1}}(n) = \frac{\sum_i X_t(i) X_{t-1}(i-n)}{\sqrt{\sum_i X_t(i)^2}\sqrt{\sum_i X_{t-1}(i)^2}}$$

Dynamic Programming Search

- **Optimal solution taking into account F_0 and ΔF_0 constraints**
- **Search space quantized such that $\Delta F/F$ is constant**

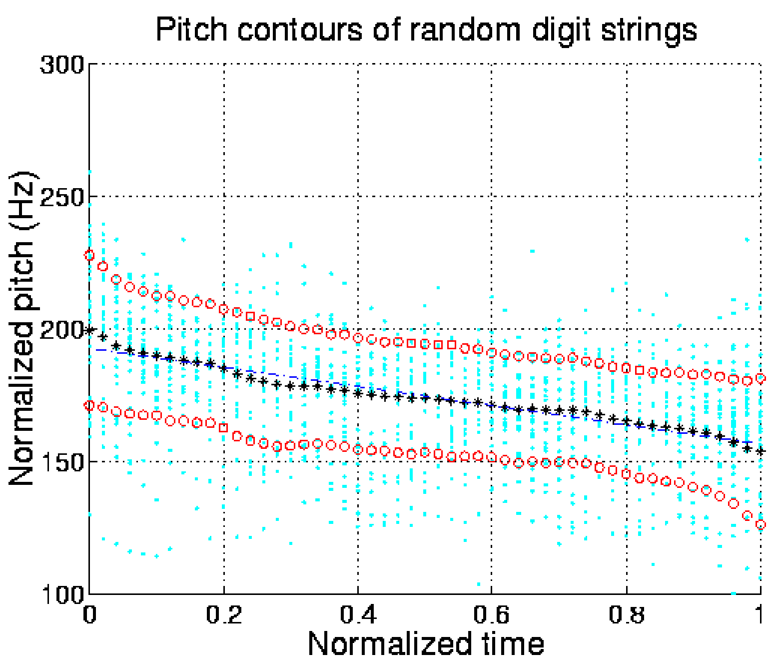


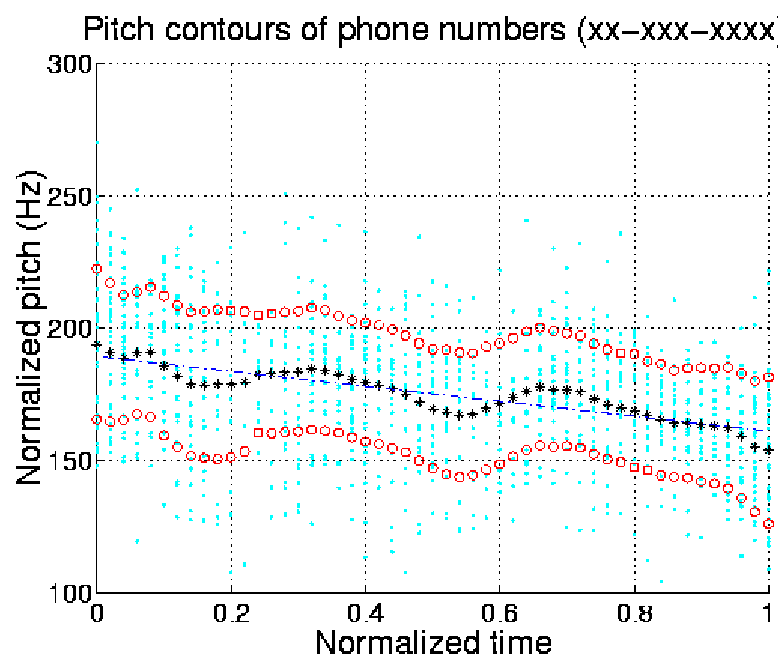
$$score_t(i) = \begin{cases} \max_j \{score_{t-1}(j) \cdot R_{X_t X_{t-1}}(i-j)\} + R_{TX_t}(i-c) & (t>0) \\ R_{TX_0}(i-c) & (t=0) \end{cases}$$

R_{XX} : cross-frame correlation , R_{TX} : template-frame correlation

The Rhythmic Nature of Speech

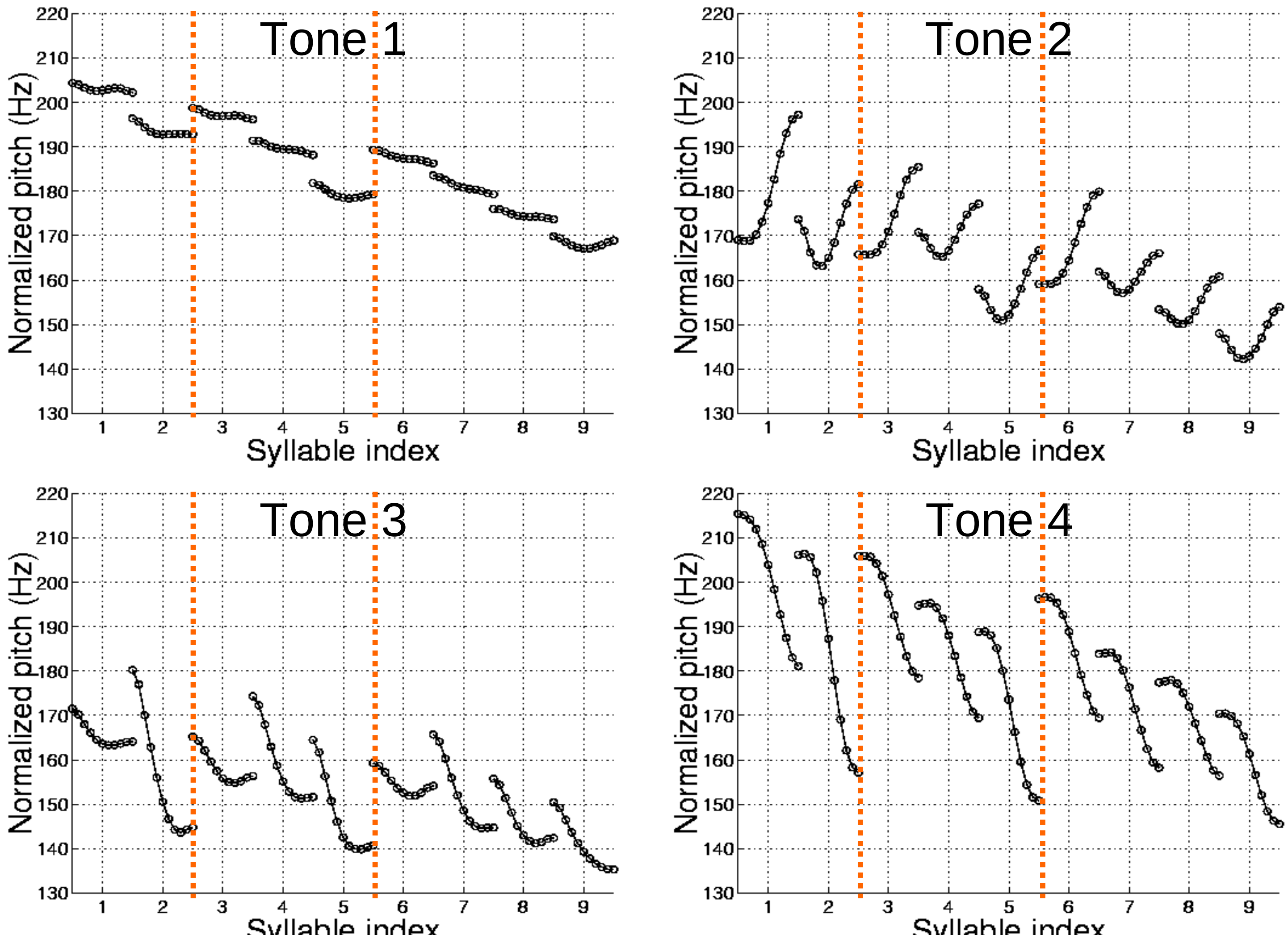
- Example using two read digit string types in Chinese
  - Random digit strings (5-10 digits per string)
  - Phone numbers (9 digits, e.g., 02 - 435 - 8264)
- Both types show a declination in pitch (i.e. sentence downdrift)
- Phone numbers show a predictable pattern or *rhythm*





Local Tones vs. Global Intonation

- **Position-dependent tone contours in phone numbers**



Characterization of Phrase Contours

- Phrases often carry distinction F_0 contours
- Canonical patterns for specific phrases can be observed
- Some research conducted into characterizing prosodic contours
  - Phrase boundary markers
  - TOBI (Tone and Break Indices) labeling
- Many unanswered questions
  - Do phrases have some set of predictable canonical patterns ?
  - How does prosodic phrase structures generalize to new utterances?
  - Are there interdependencies among phrases in the utterance ?
  - How can prosodic modeling help speech recognition and/or understanding ?

Pilot Study of Phrasal Prosody in JUPITER

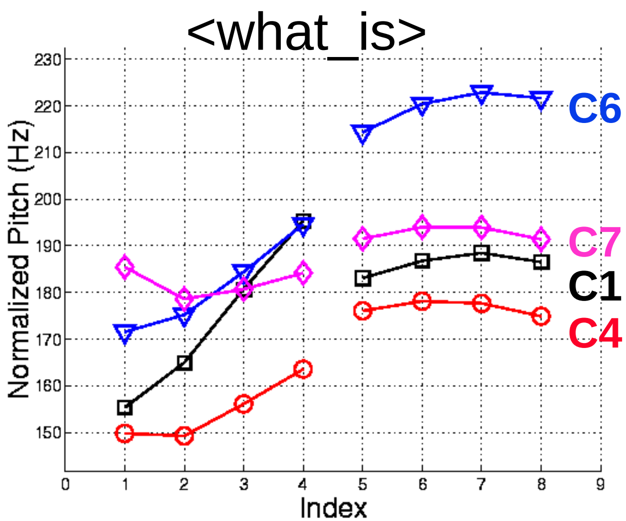
- **Five phrase types were studied:**
  - **<what_is>: what is, how is, …**
  - **<tell_me>: tell me, give me, …**
  - **<weather>: weather, forecast, dew point, …**
  - **<SU>: Boston, Monday, ...**
  - **<US>: Detroit, tonight, …**
- **Phrases studied with a fixed sentence template:**

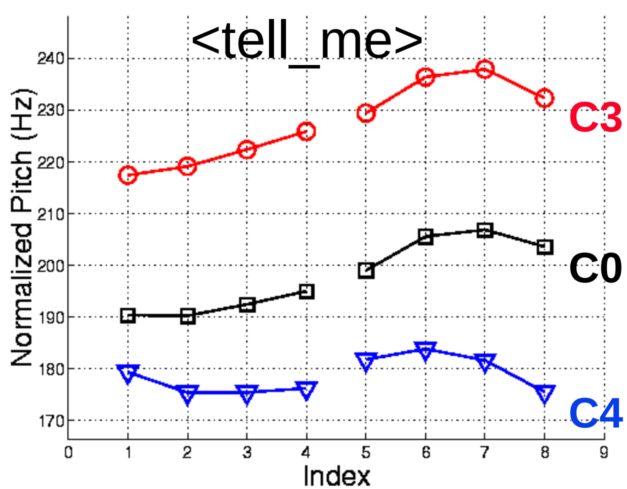
<what_is> | <tell_me> the <weather> in | for | on <SU> | <US>

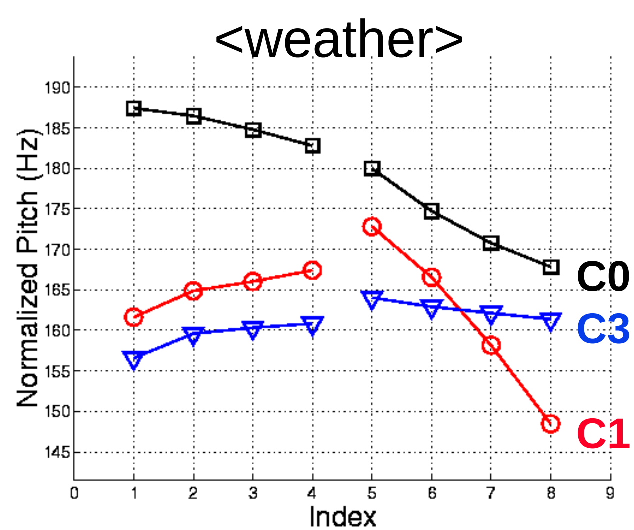
- **Pitch contours for each example phrase were automatically clustered into several subclasses**
- **Mutual information of subclasses can predict which subclasses are likely or unlikely to occur together in an utterance**

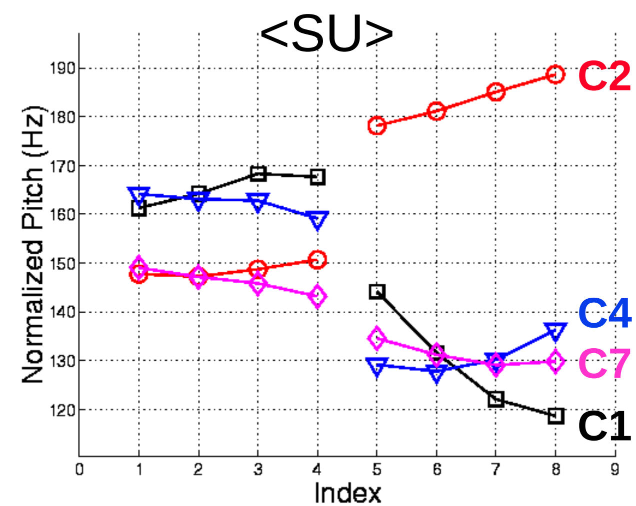
Subclasses Obtained by Clustering

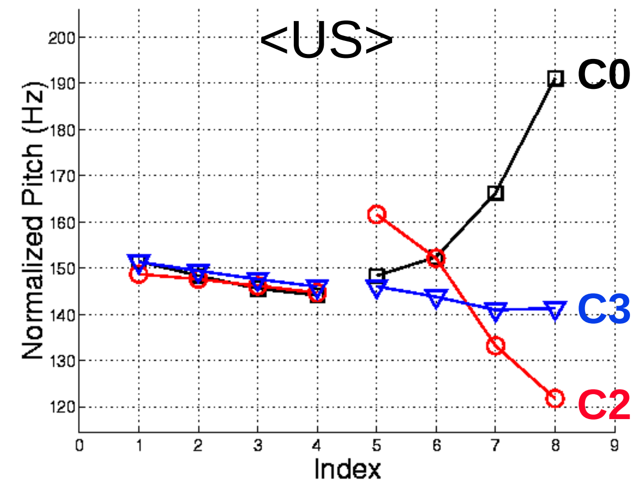
- ***K*-means clustering on training data followed by selection**



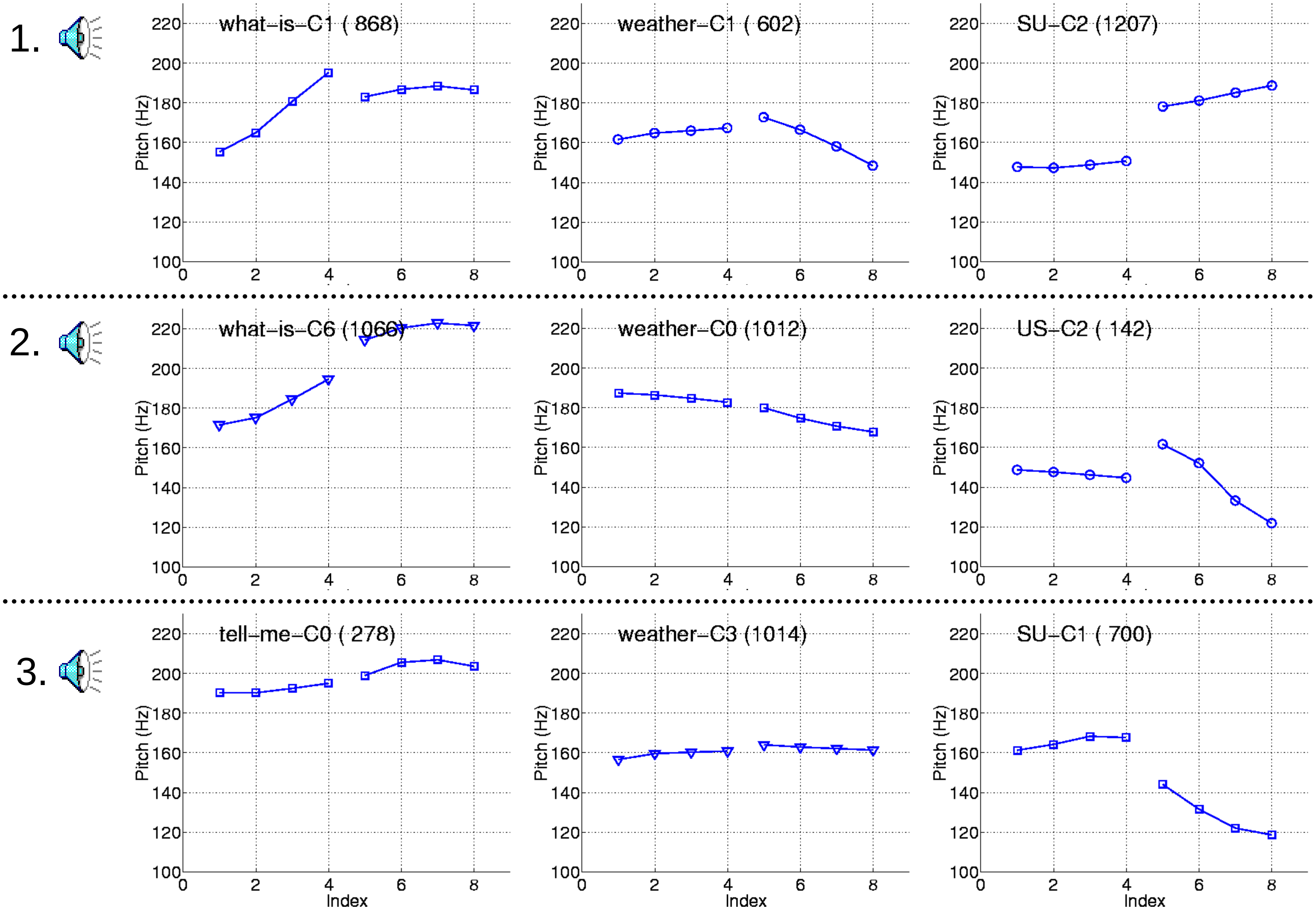




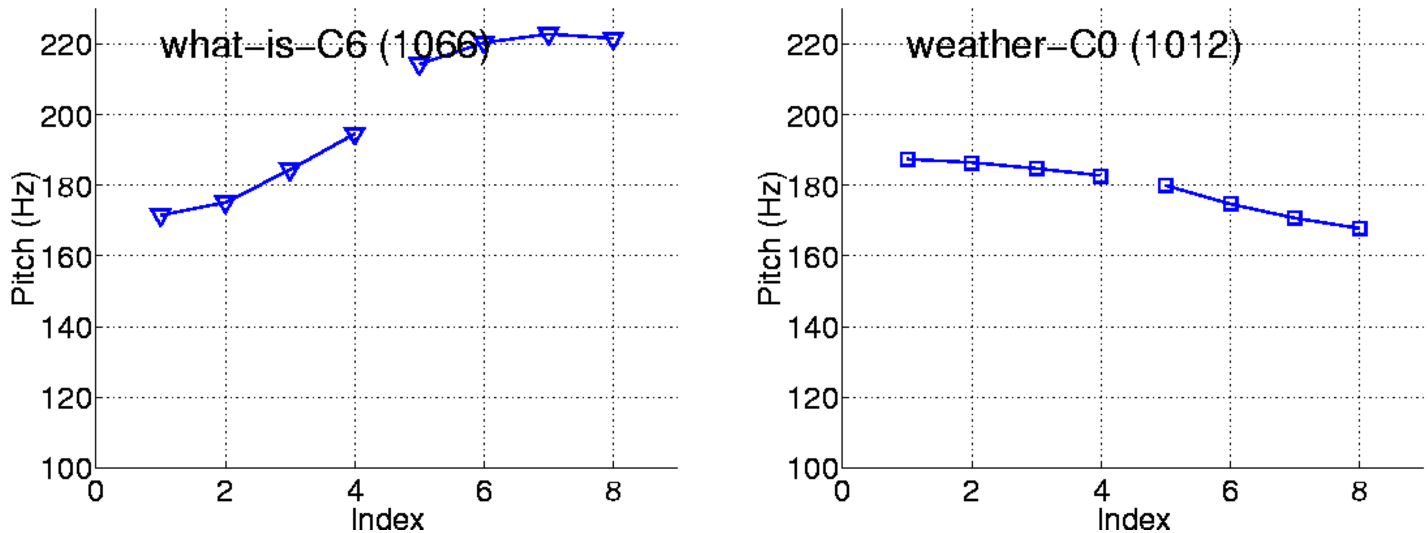




Example Utterances

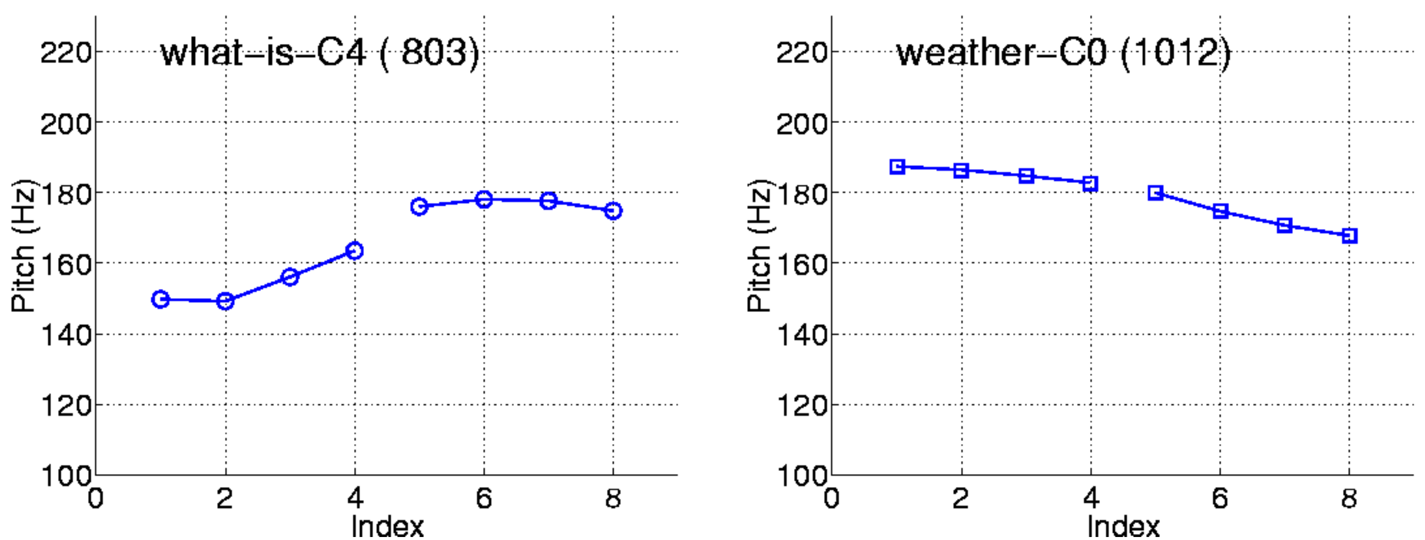


Mutual Information of Subclasses



MI = 0.67

Subclasses are commonly used together



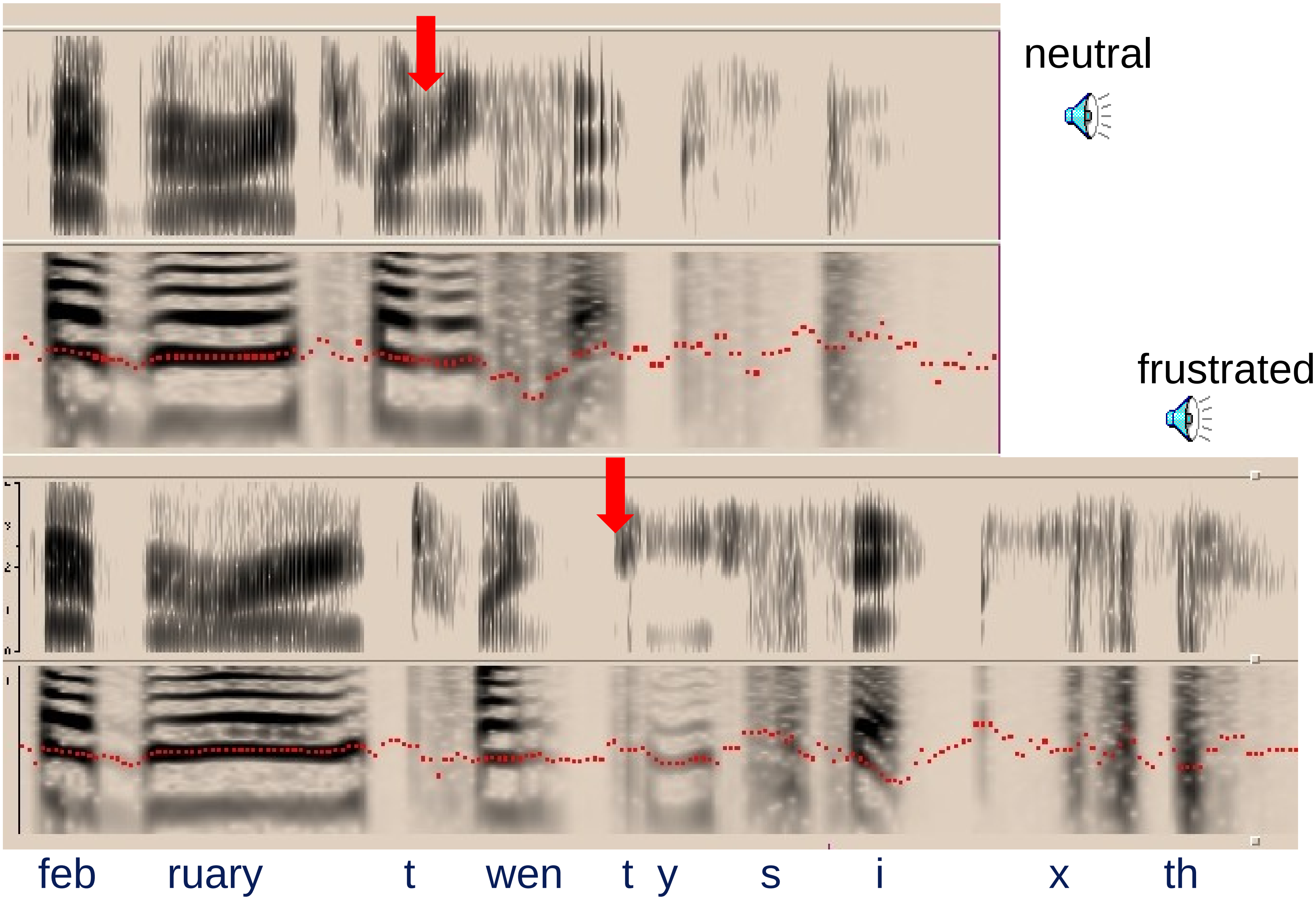
MI = -0.58

Subclasses are unlikely to occur together

Emotional Speech

- **Emotional speech is difficult to recognize:**
  - **Neutral speech word error rate in Mercury: 15%**
  - **WER of “happy” speech in Mercury: 25%**
  - **WER of “frustrated” speech: 33%**
- **Acoustic correlates of emotional/frustrated speech:**
  - **Fundamental frequency variation**
  - **Increased energy**
  - **Speaking rate & vowel duration**
  - **Hyper-articulation**
  - **Breathy sighs**
- **Linguistic content can also indicate frustration:**
  - **Questions**
  - **Negative constructors**
  - **Derogatory terms**

Spectrograms of an Emotional Pair



Emotion Recognition

- **Few studies of automatic emotional recognition exist**
- **Common features used for utterance-based emotion recognition:**
  - **F_0 features: mean, median, min, max, standard deviation**
  - **👎F_0 features: mean positive slope, mean negative slope, std. deviation, ratio of rising and falling slopes**
  - **Rhythm features: speaking rate, duration between voiced regions**
- **Some results:**
  - **75% accuracy over six classes (happy, sad, angry, disgusted, surprised, fear) using only mean and standard deviation of F_0 (Huang *et al*, 1998)**
  - **80% accuracy over fours classes (happy, sad, anger, fear) using 16 features (Dellaert *et al*, 1998)**

Speaker Identification

- **Speaker verification: Accept or reject claimed identity**
  - **Typically used in applications requiring secure transactions**
  - **Not 100% reliable**
    - Speech is highly variable and easily distorted
  - **Can be combined with other techniques**
    - Possession of a physical “key”
    - Knowledge of a password
    - Face ID or other biometric techniques
- **Speaker recognition: Identify speaker from set of known speakers**
  - **Typically used when speakers do not volunteer their identity**
  - **Example applications:**
    - Meeting transcription and indexing
    - Voice mail summarization
    - “Power users” of dialogue system

Speaker Identification Approaches

- **Potential features used for speaker ID**
  - **Formant frequencies (correlated with vocal tract length)**
  - **Fundamental frequency averages and contours**
  - **Phonetic durations and speaking rate**
  - **Word usage patterns**
  - **Spectral features (typically MFCCs) are most commonly used**
- **Some modeling approaches:**
  - **Text Independent**
    - Global Gaussian Mixture Models (GMMs) (Reynolds, 1995)
    - Phonetically-Structured GMMs
  - **Text/Recognition Dependent**
    - Phonetically Classed GMMs
    - Speaker Adaptive ASR Scoring (Park and Hazen, 2002)

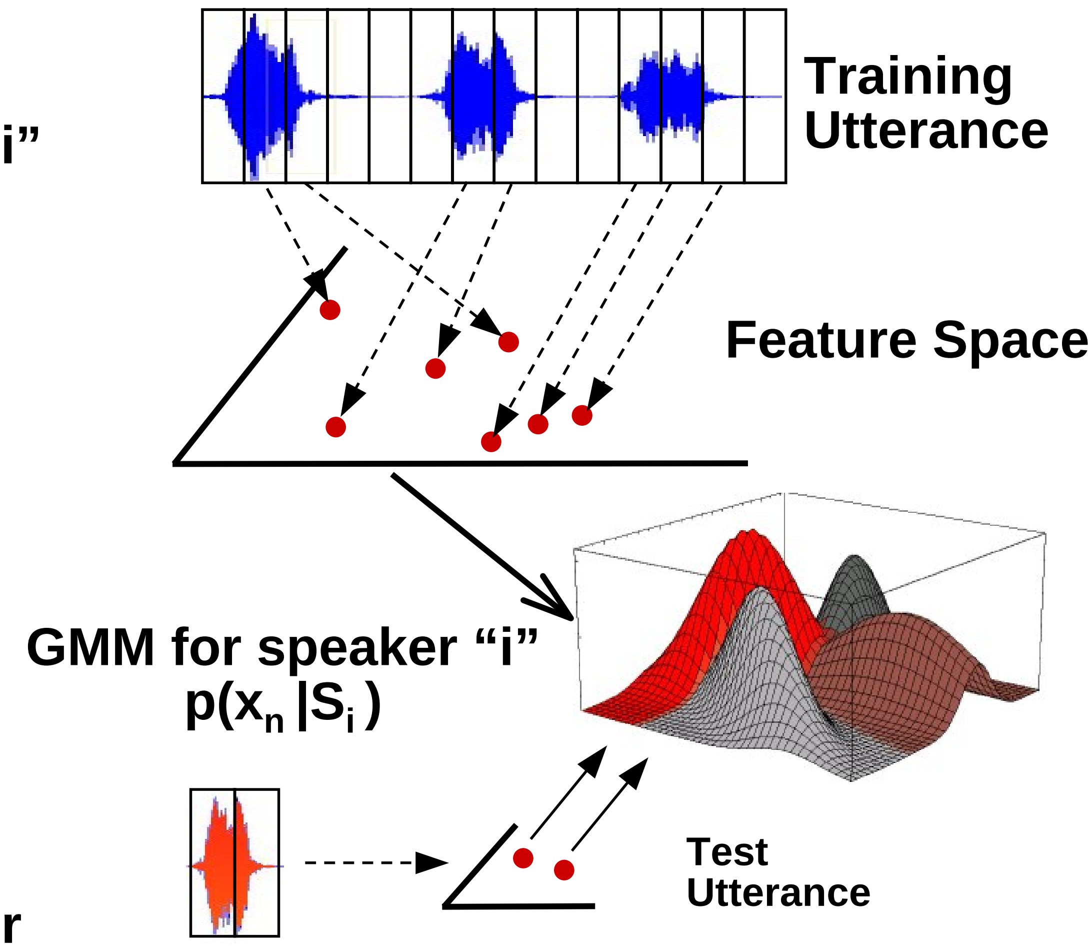
Global GMMs

Training

- Input waveforms for speaker "i" split into fixed-length frames
- Feature vectors computed from each frame of speech
- GMMs trained from set of feature vectors
- One global GMM per speaker

Testing

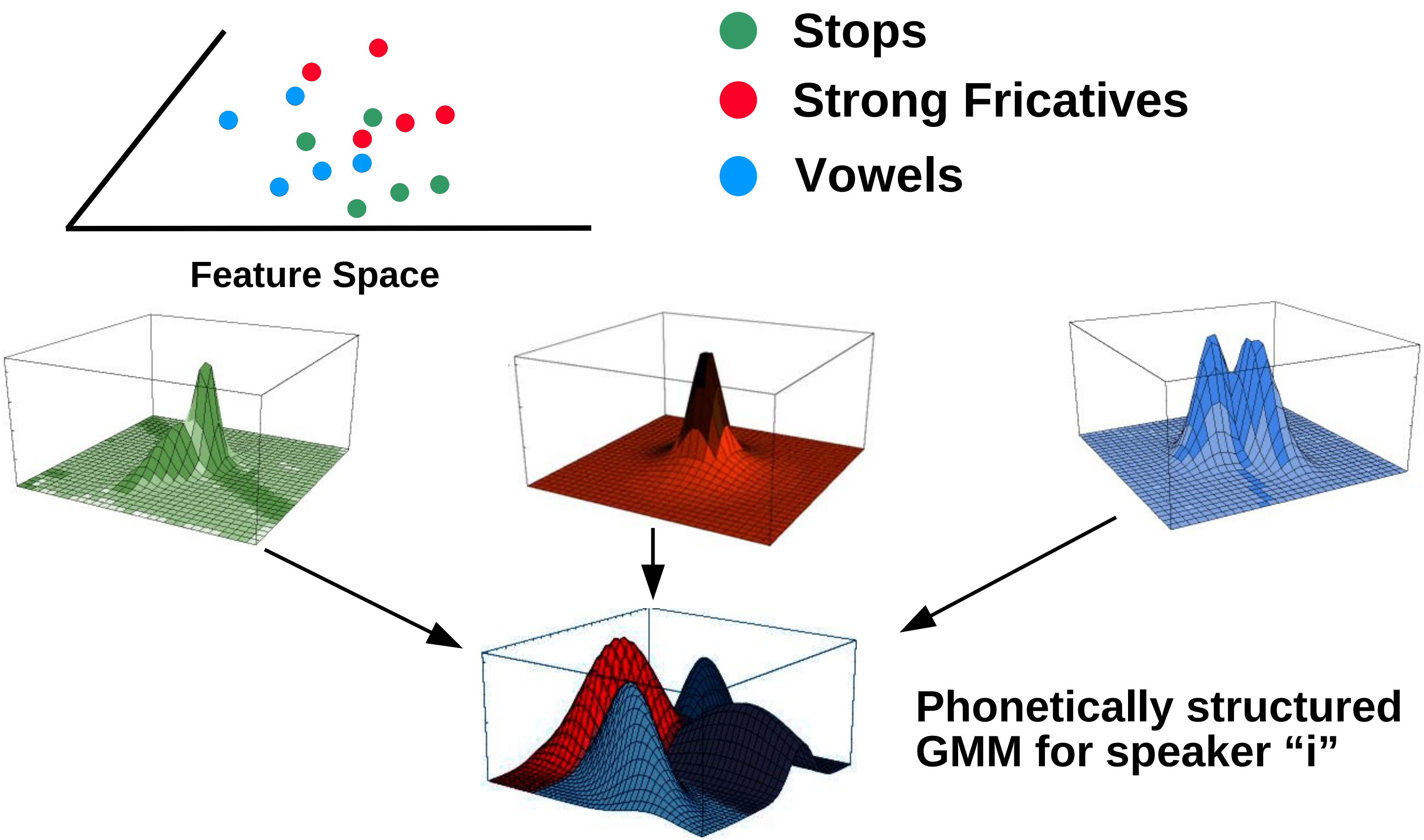
- Input feature vectors scored against each speaker GMM
- Frame scores for each speaker summed over entire utterance
- Highest total score is hypothesized speaker



$p(x_1 | S_i) + p(x_2 | S_i) =$ score for speaker "i"

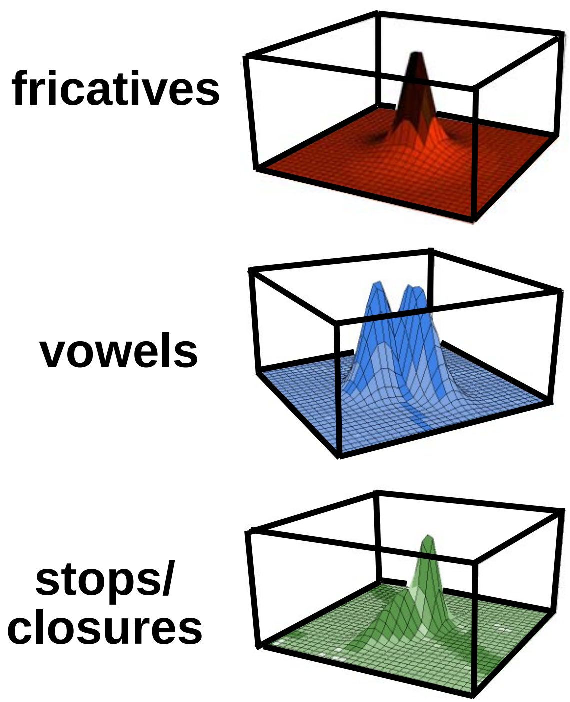
Phonetically-Structured GMM

- During training, use phonetic transcriptions to train phonetic class GMMs for each speaker
- Combine class GMMs into a single “structured” model which is then used for scoring as in the baseline system

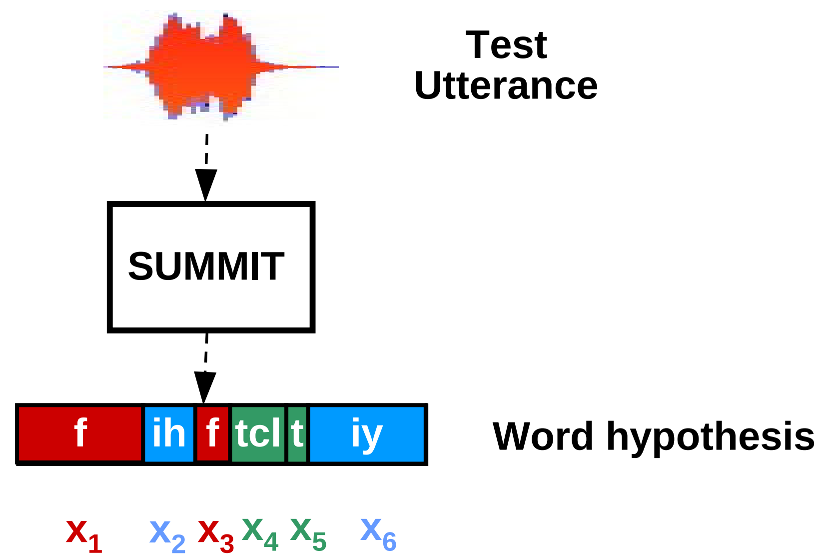


Phonetic Classing

- Train independent phone class GMMs w/o combination
- Generate word/phone hypothesis from recognizer
- Score frames with class models of hypothesized phones



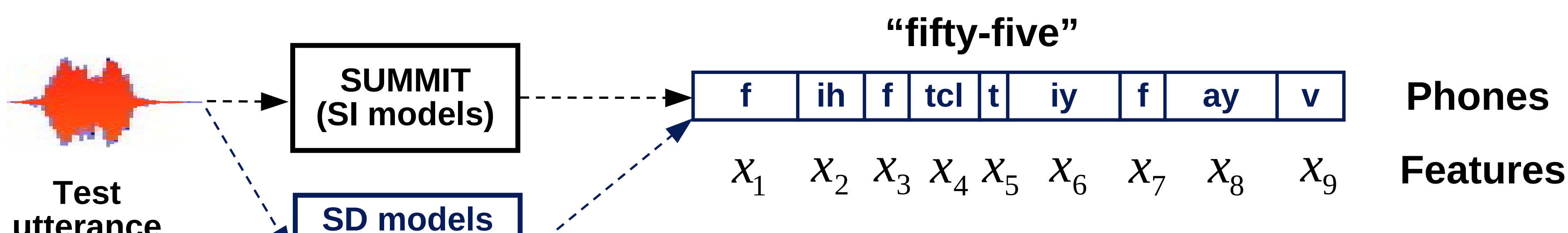
phonetic class GMMs
for speaker "i"



$$\sum p(x_n \,|S_i, \text{class}(x_n)) = \text{score for speaker “i”}$$

Speaker Adapted Scoring

- **Train speaker-dependent (SD) models for each speaker**
- **Get best hypothesis from recognizer using speaker-independent (SI) models**
- **Rescore hypothesis with SD models**
- **Compute total speaker adapted score by interpolating SD score with SI score**



$\mathrm{p}_{\mathrm{SI}}(x_n \mid u_n)$ **SI Model**

$\mathrm{p}_{\mathrm{SD}}(x_n \mid u_n, S_i)$ **SD Model**

$$\mathrm{p}_{\mathrm{SA}}(x_n \mid u_n, S_i) = \frac{\lambda_n \mathrm{p}_{\mathrm{SD}}(x_n \mid u_n, S_i) + (1-\lambda_n)\mathrm{p}_{\mathrm{SI}}(x_n \mid u_n, S_i)}{\mathrm{p}_{\mathrm{SI}}(x_n \mid u_n, S_i)} \qquad \lambda_n = \frac{c(u_n)}{c(u_n)+K}$$

Two Experimental Corpora

Corpus	YOHO	Mercury
Description	LDC corpus for speaker verification evaluation	SLS corpus from air-travel system
Type of Speech	Prompted Text “Combination lock” phrases (e.g. “34-25-86”)	Spontaneous conversational speech in air-travel domain
# Speakers	138 (106M, 32F)	38 (18M, 20F)
Recording Conditions	Fixed telephone handset Quiet office environment 8kHz band-limited	Variable telephone Variable environment Telephone channel
Training Data	96 utterances From 4 sessions (~3 seconds each)	50-100 utterances From 2-10 sessions (variable length)
Test Set Size	5520	3219

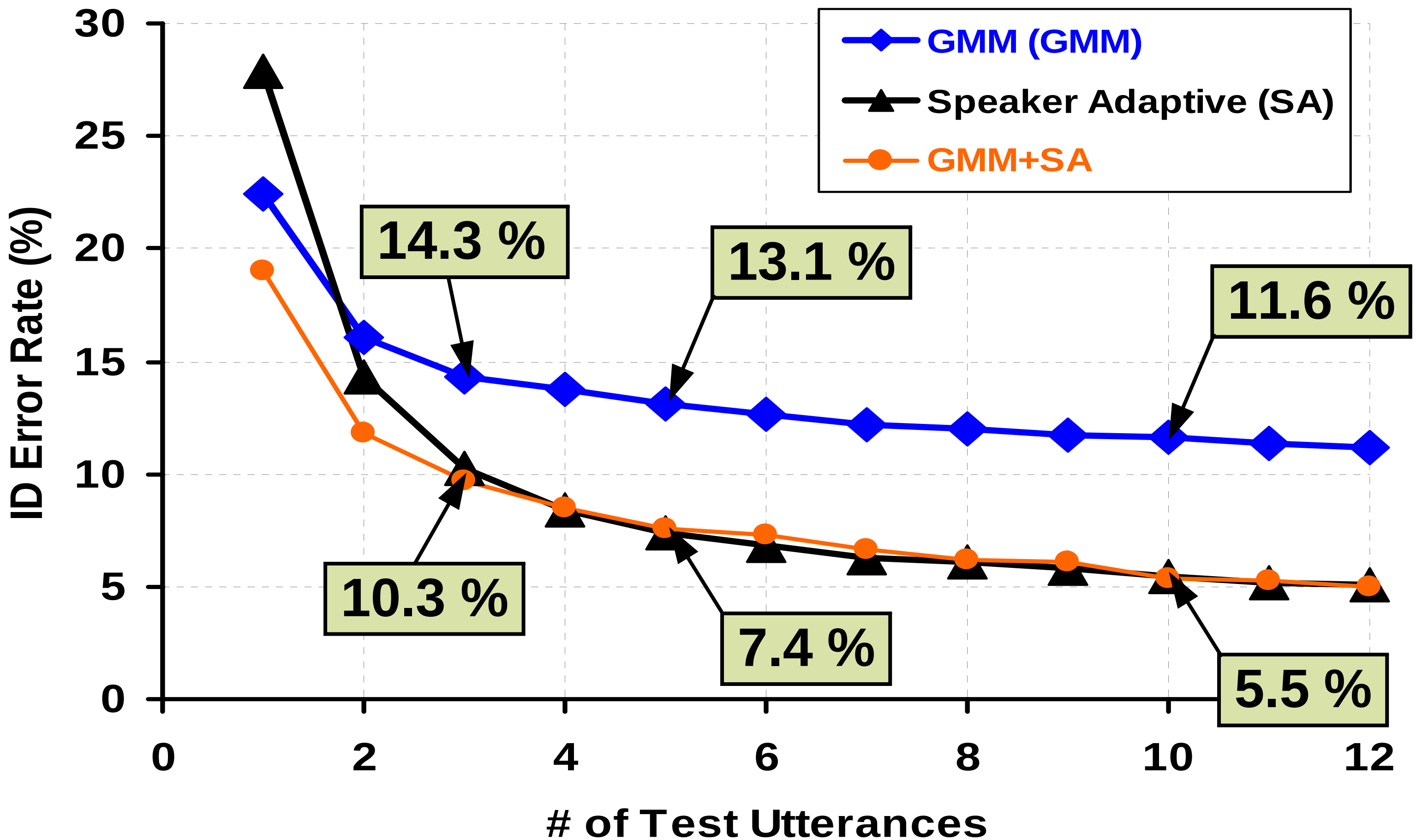
Single Utterance Results

- Experiment: closed set speaker recognition on single utterances
- Results:

System	Speaker ID Error Rate%	
	YOHO	Mercury
Structured GMM (SGMM)	0.31	21.3
Phone Classing	0.40	21.6
Speaker Adaptive (SA)	0.31	27.8
SA+SGMM	0.25	18.3

- All approaches about equal on YOHO corpus
- Speaker adaptive approach has poorest performance on Mercury
  - ASR recognition errors can degrade speaker ID performance
- Classifier combination yields improvements over best system

Results on Multiple Mercury Utterances



- **On multiple utterances, speaker adaptive scoring achieves lower error rates than next best individual method**
- **Relative error rate reductions of 28%, 39%, and 53% on 3, 5, and 10 utterances compared to baseline**

Multi-modal Interfaces

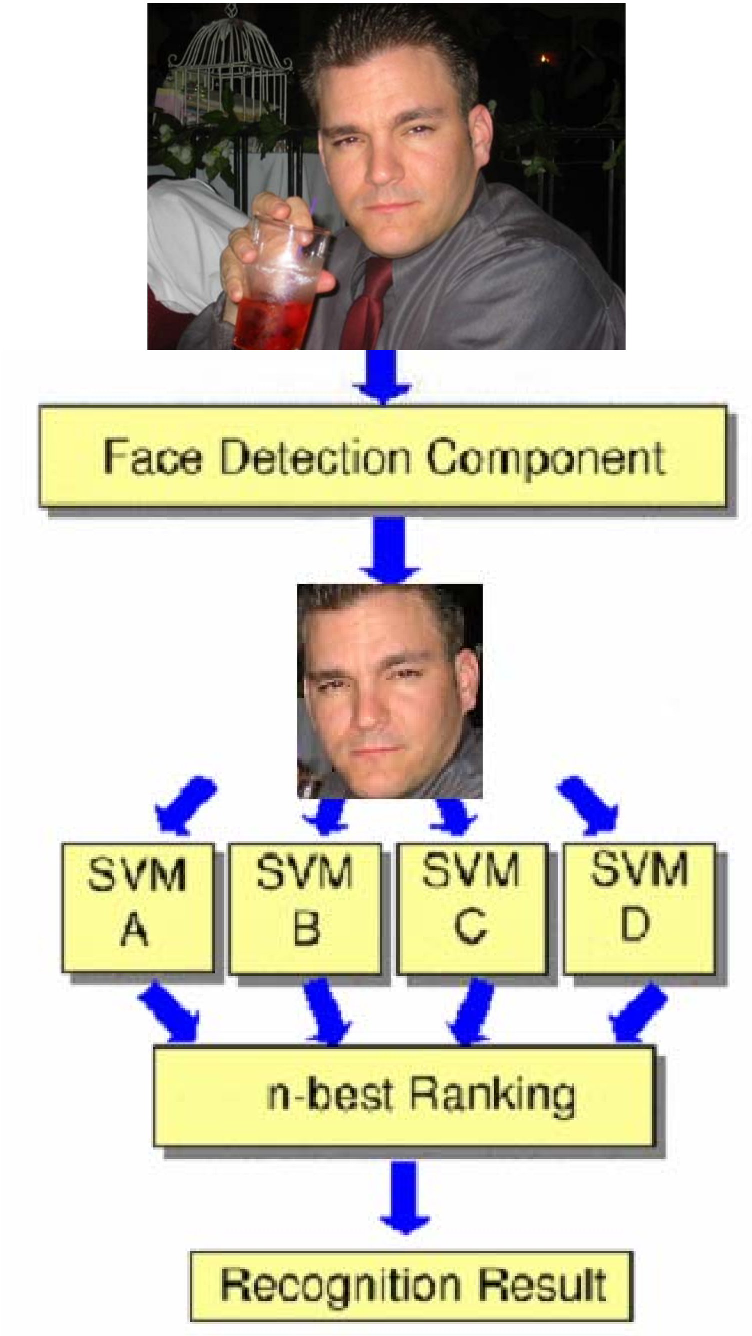
- Multimodal interfaces will enable more natural, flexible, efficient, and robust human-computer interaction
  - Natural: Requires no special training
  - Flexible: Users select preferred modalities
  - Efficient: Language and gestures can be simpler than in uni-modal interfaces (e.g., Oviatt and Cohen, 2000)
  - Robust: Inputs are complementary and consistent
- Audio and visual signals both contain information about:
  - Identity of the person: Who is talking?
  - Linguistic message: What are they saying?
  - Emotion, mood, stress, etc.: How do they feel?
- Integration of these cues can lead to enhanced capabilities for future human computer interfaces

Face/Speaker ID on a Handheld Device

- An iPaq handheld with Audio/Video Input/Output has been developed as part of MIT Project Oxygen
- Presence of multiple-input channels enables multi-modal verification schemes
- Prototype system uses a login scenario
  - Snap frontal face image
  - State name
  - Recite prompted lock combination phrase
  - System *accepts* or *rejects* user

Face Identification Approach

- **Face Detection by Compaq/HP (Viola/Jones, CVPR 2001)**
  - **Efficient cascade of classifiers**
- **Face Recognition by MIT AI Lab/CBCL (Heisele et al, ICCV 2001)**
  - **Based on Support Vector Machines (SVM)**
  - **Runtime face recognition: score image against each SVM classifier**
- **Implemented on iPaq handheld as part of MIT Project Oxygen (E. Weinstein, K. Steele, P. Ho, D. Dopson)**



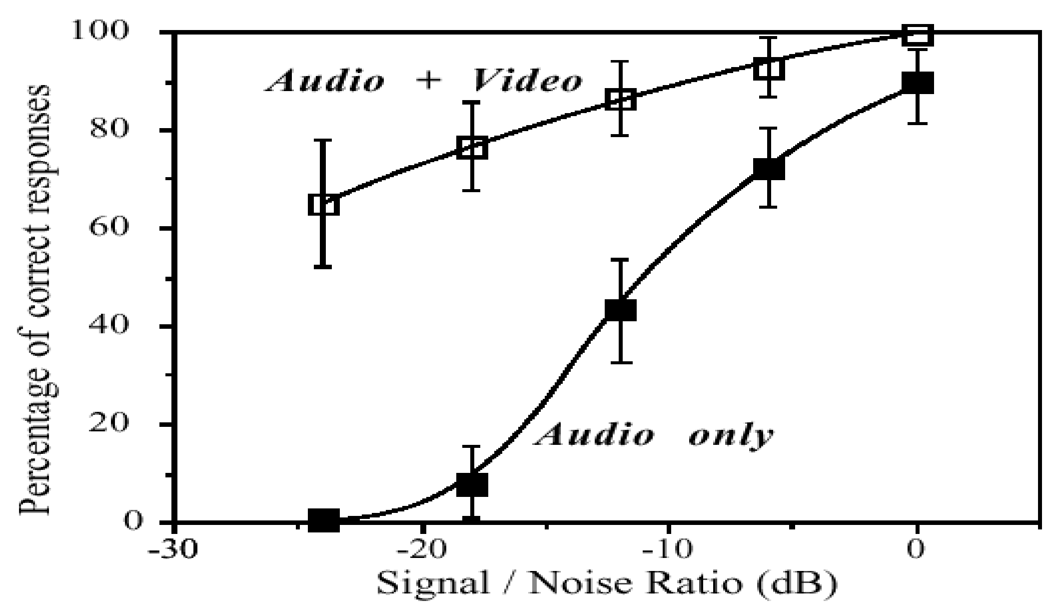
Combined Face/Speaker ID

- Multi-modal user login verification experiment using iPaq
- Enrollment data:
  - Training data collected from 35 enrolled users
  - 100 facial images and 64 lock combination phrases per user
- Test data:
  - 16 face/image pairs from 25 enrolled users
  - 10 face/image pairs from 20 non-enrolled imposters
- Evaluation metric: verification equal error rate (EER)
  - Equal likelihood of false acceptances and false rejections
  - Fused system reduces equal error rate by 50%

System	Equal Error Rate
Face ID Only	7.30%
Speech ID Only	1.77%
Fused System	0.89%

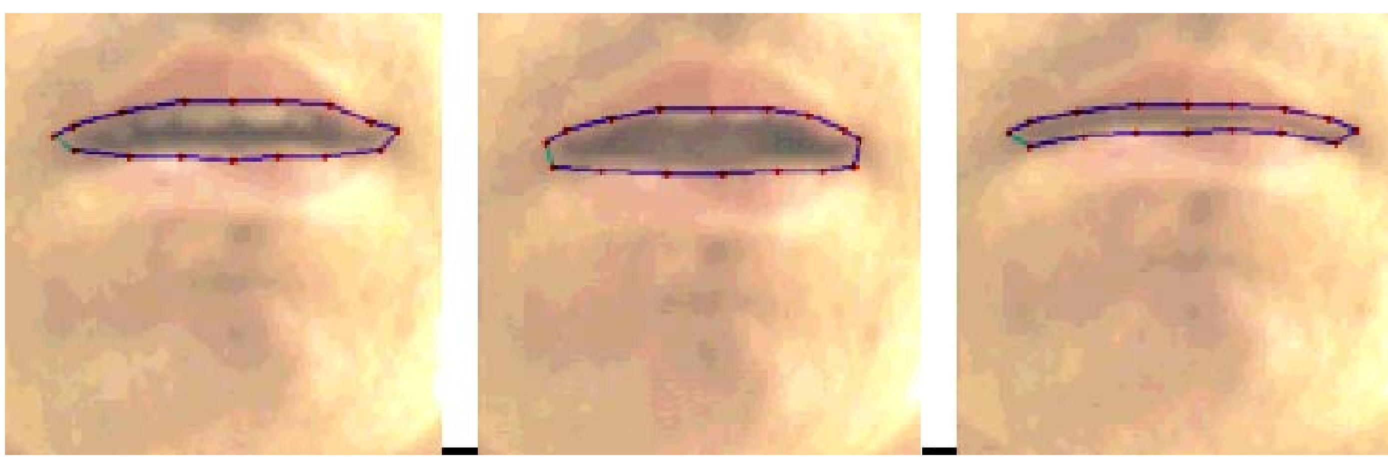
How can we improve ASR performance?

- Humans utilize facial expressions and gestures to augment the speech signal
- Facial cues can improve speech recognition in noise by up to 30 dB, depending on the task
- Speech recognition performance can be improved by incorporating facial cues (e.g., lip movements and mouth opening)
- Figure shows human recognition performance
  - Low signal-to-noise ratios
  - Presented with audio with video and audio only
  - Reference: Benoit, 1992



Audio Visual Speech Recognition (AVSR)

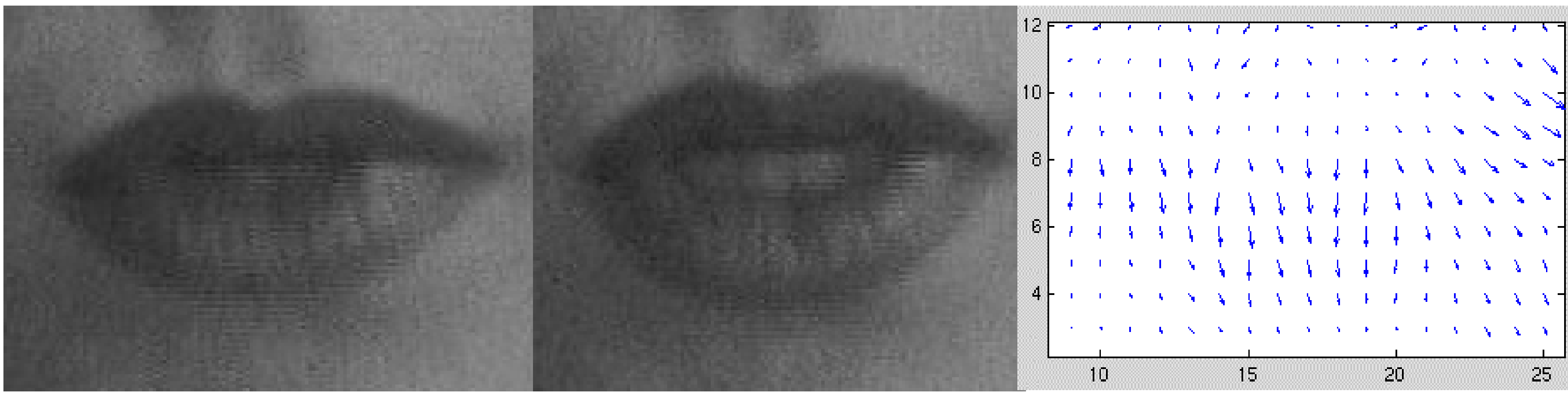
- Integrate information about visual mouth/lip/jaw features with features extracted from audio signal
- Visual feature extraction:
  - Region of Interest (ROI): mostly lips and mouth; some tracking
  - Features: pixel-, geometric-, or shape-based
  - Almost all systems need to locate and track landmark points
  - Correlation and motion information not used explicitly



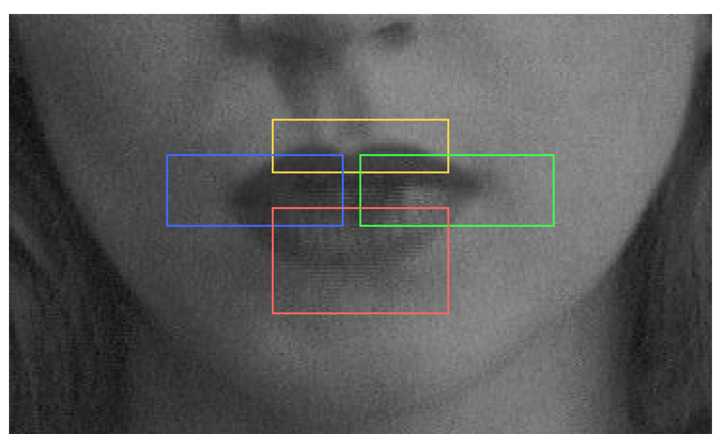
Example of pixel-based features (Covell & Darrell, 1999)

AVSR: Preliminary Investigations

- **Goal: integration with SUMMIT ASR system**
- **Visually-derived measurements based on optical flow**



- **Low-dimensional features represent opening & elongation**



Issues with Audio/Visual Integration

- **Early vs. Late Integration**
  - **Early: concatenate feature vectors from different modes**
  - **Late: combine outputs of uni-modal classifiers**
    - Can be at many levels (phone, syllable, word, utt)
- **Channel Weighting Schemes**
  - **Audio channel usually provides more information**
  - **Based on SNR estimate for each channel**
  - **Preset weights by optimizing the error rate of a dev. set**
  - **Estimate separate weights for each phoneme or viseme**
- **Modeling the audio/visual asynchrony**
  - **Many visual cues occur before the phoneme is actually pronounced**
  - **Example: rounding lips before producing rounded phoneme**

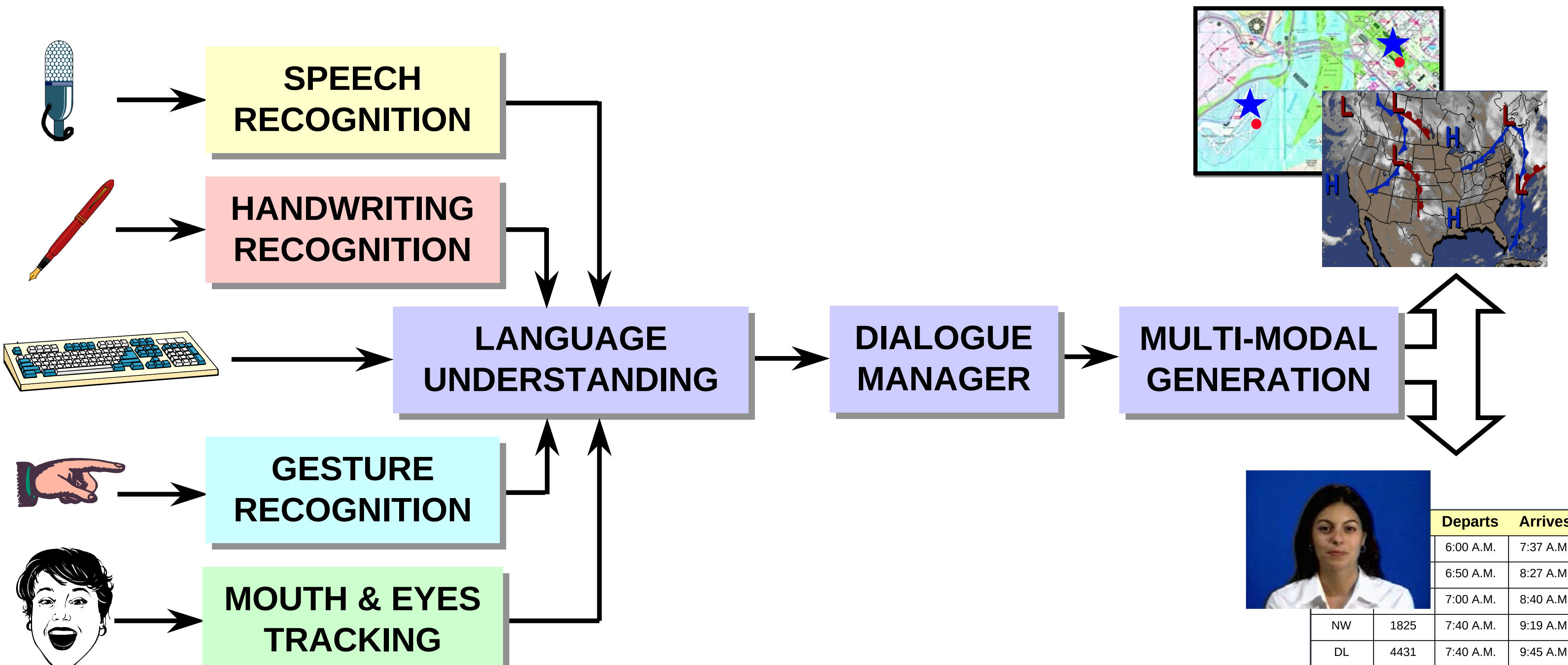
AVSR: State of the Art

- **Example: Neti *et al*, 2000 (JHU Summer Workshop)**
  - **>10K word vocabulary**
  - **Training and development data: 264 subjects, 40 hours**
  - **Test data: 26 subjects, 2.5 hours**
  - **Quiet (19.5 dB SNR) and noisy (8.5 dB SNR) conditions**

Conditions	Clean WER (%)	Noisy WER (%)
Audio Only	14.4	48.1
AVSR	13.5	35.3

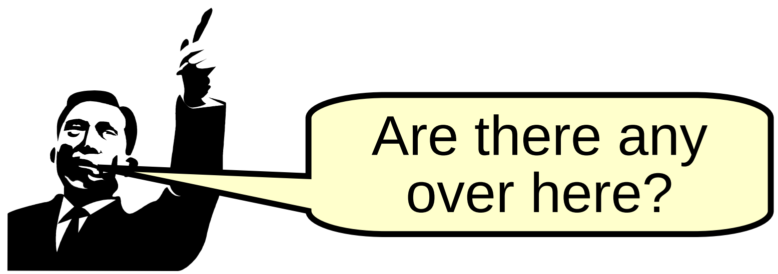
Multi-modal Interaction Research

- **Understanding the science**
  - **How do humans do it (e.g. expressing cross modality context)?**
  - **What are the important cues?**
- **Developing an architecture that can adequately describe the interplays of modalities**



Multi-modal Interfaces

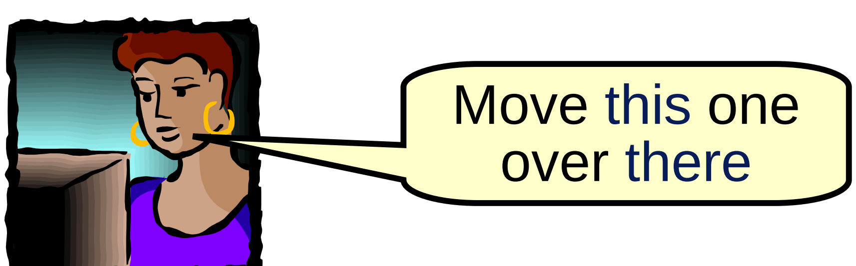
- **Inputs need to be understood in the proper context**



What does he mean by “any,” and what is he pointing at?

Does this mean “yes,” “one,” or something else?

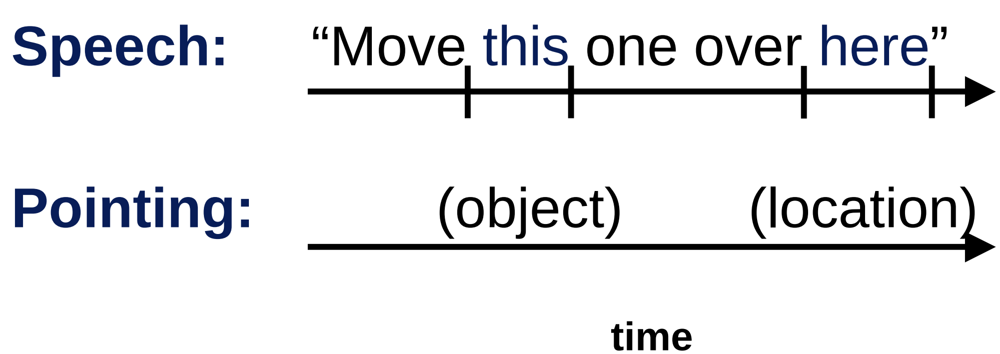
- **Timing information is a useful way to relate inputs**



Where is she looking or pointing at while saying “this” and “there”?

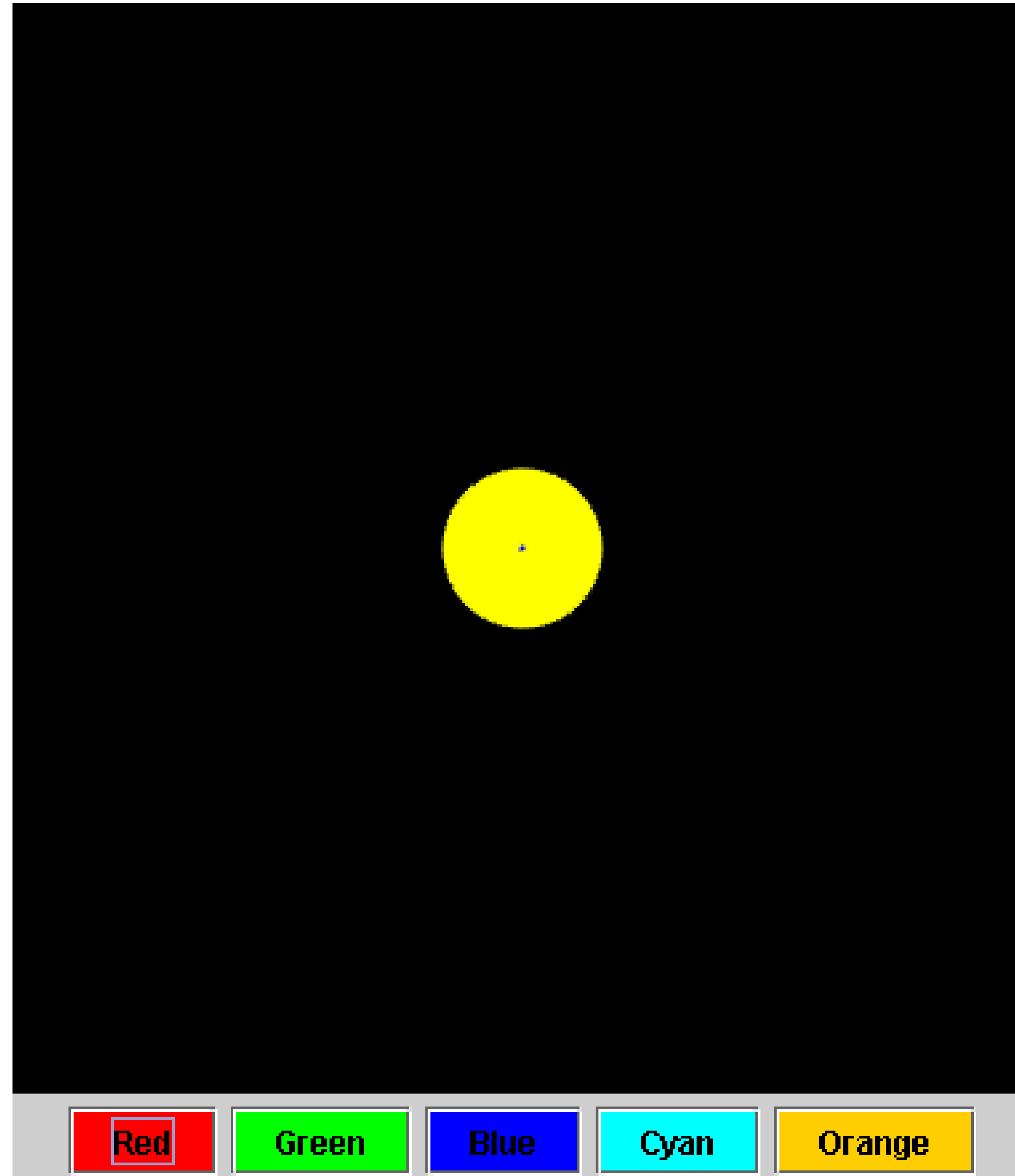
Multi-modal Fusion: Initial Progress

- **All multi-modal inputs are synchronized**
  - **Speech recognizer generates absolute times for words**
  - **Mouse and gesture movements generate {x,y,t} triples**
- **Speech understanding constrains gesture interpretation**
  - **Initial work identifies an object or a location from gesture inputs**
  - **Speech constrains what, when, and how items are resolved**
  - **Object resolution also depends on information from application**

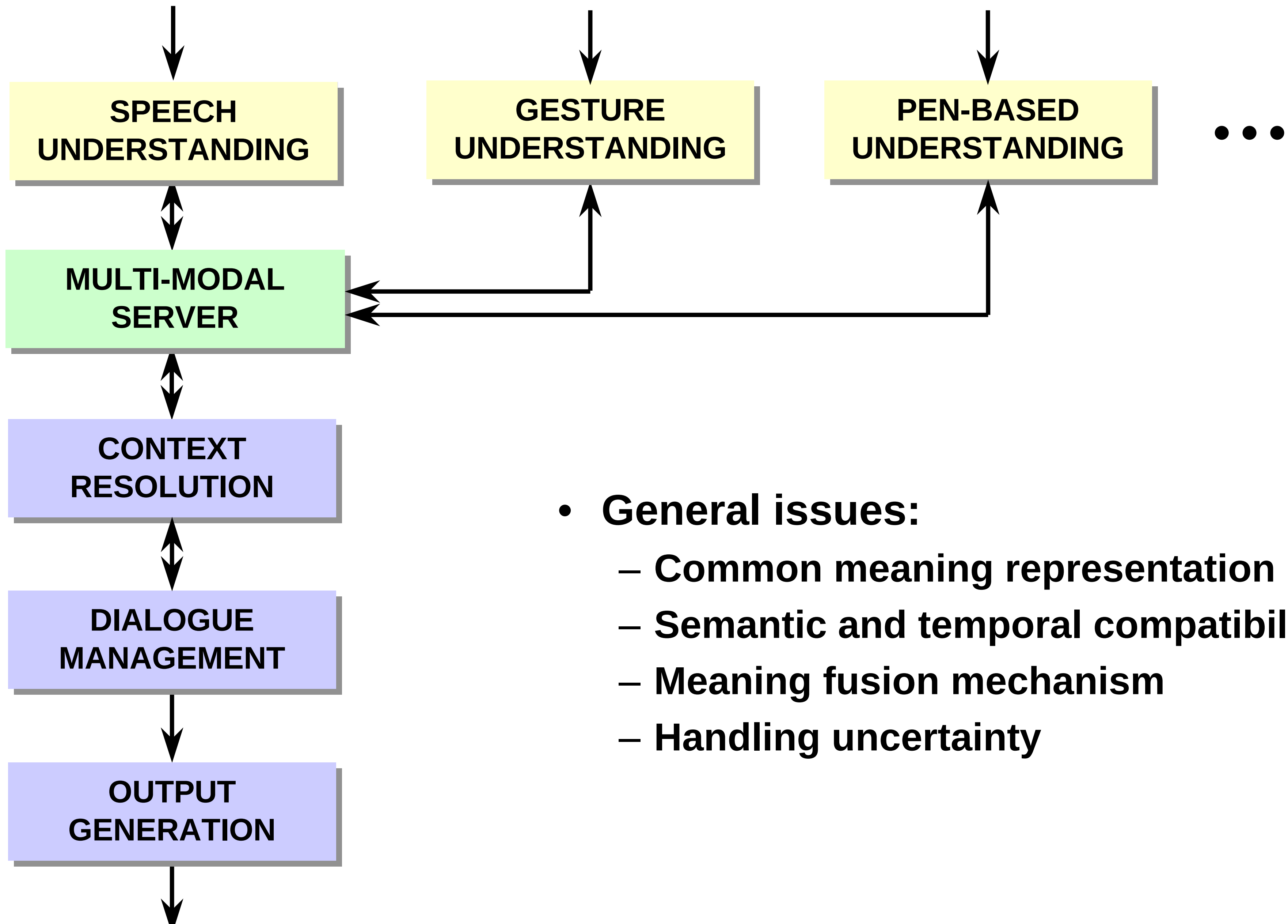


Multi-modal Demonstration

- **Manipulating planets in a solar-system application**
- **Continuous tracking of mouse or pointing gesture**
- **Created w. SpeechBuilder utility with small changes (Cyphers, Glass, Toledano & Wang)**
- **Standalone version runs with mouse/pen input**
- **Can be combined with gestures from determined from vision (Darrell & Demirdjien)**



Recent Activities: Multi-modal Server



- **General issues:**
  - **Common meaning representation**
  - **Semantic and temporal compatibility**
  - **Meaning fusion mechanism**
  - **Handling uncertainty**

Summary

- Speech carries paralinguistic content:
  - Prosody, intonation, stress, emphasis, etc.
  - Emotion, mood, attitude,etc.
  - Speaker specific characteristics
- Multi-modal interfaces can improve upon speech-only systems
  - Improved person identification using facial features
  - Improved speech recognition using lip-reading
  - Natural, flexible, efficient, and robust human-computer interaction

References

- **C. Benoit, "The intrinsic bimodality of speech communication and the synthesis of talking faces," *Journal on Communications*, September 1992.**
- **M. Covell and T. Darrell, "Dynamic occluding contours: A new external-energy term for snakes," *CVPR,* 1999.**
- **F. Dellaert, T. Polzin, and A. Waibel, "Recognizing emotion in speech," *ICSLP*, 1998.**
- **B. Heisele, P. Ho, and T. Poggio, "Face recognition with support vector machines: Global versus component-based approach," *ICCV*, 2001.**
- **T. Huang, L. Chen, and H. Tao, "Bimodal emotion recognition by man and machine," *ATR Workshop on Virtual Communication Environments*, April 1998.**
- **C. Neti, *et al*, "Audio-visual speech recognition," *Tech Report CLSP/Johns Hopkins University*, 2000.**
- **S. Oviatt and P. Cohen, "Multimodal interfaces that process what comes naturally," *Comm. of the ACM*, March 2000.**
- **A. Park and T. Hazen, "ASR dependent techniques for speaker identification," *ICSLP*, 2002.**
- **D. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, August 1995.**
- **P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," *CVPR*, 2001.**
- **C. Wang, "Prosodic modeling for improved speech recognition and understanding," PhD thesis, MIT, 2001.**
- **E. Weinstein, *et al*, "Handheld face identification technology in a pervasive computing environment," *Pervasive 2002.***