

Challenges to Using Prosody in Automatic Language Processing

ELIZABETH SHRIBERG

SRI International

International Computer Science Institute



TSD 2008, Brno, Czech Republic



Acknowledgments & Talk URL

Harry Bratt

Benoit Favre

Luciana Ferrer

James Fung

Martin Graciarena

Dilek Hakkani-Tur

Mary Harper

Julia Hirschberg

Sachin Kajarekar

Kornel Laskowski

Yang Liu

Nelson Morgan

Colleen Richey

Nicolas Scheffer

Andreas Stolcke

Jennifer Venditti

Talk URL:

www.speech.sri.com/people/ees/prosody

Introduction

- Most natural language is **spoken**
- Important aspect of spoken language is **prosody**
- Significant work on prosody for speech **generation**
- But less work for understanding. **WHY?**

IN THIS TALK:

- Why should you care about prosody in speech input?
- What are the challenges? Can they be overcome?
- Present a general framework for modeling prosody
- Describe some successes in a range of tasks

10 Common Reasons for Not Using Prosody

And Why They Deserve a Closer Look

1. I don't really know what prosody is.
 2. I don't need prosody; I use words.
 3. Prosody needs linguists to label data.
 4. Prosody requires exotic new tools.
 5. Prosody doesn't help that much.
 6. Prosody's only good for "fuzzy" tasks.
 7. It only works for contrived examples.
 8. Prosody will slow my system down.
 9. Prosody doesn't generalize.
 10. Prosody is too speaker-dependent.
- Feasibility
- Performance

Feasibility

1. I don't really know what prosody is



What Is Prosody?

- The rhythm and melody of spoken language
- “**Suprasegmental**” —variation that cannot be derived from the sequence of speech phones.

“no|seats are available”



“má tu být zítra|ráno|ale nebude mít na vás čas”



He should be here tomorrow morning. But he will have no time for you.

He should be here tomorrow. But he will have no time for you in the morning.









(Thanks to Jáchym Kolář)

What Is Prosody?

- Prosody conveyed through variation in:
 - Pitch (fundamental frequency)
 - Timing (durations, pausing, speaking rate)
 - Loudness (energy)
 - Voice quality
- Languages differ in prosody
 - This talk mostly on English, but some on other languages
- Prosody is central to language learning
 - Babies attend to
 - Difficult to master when learning a new language

(Bolinger, 1989; Hirschberg, 2002; Nooteboom, 1997; Hirst & Di Cristo, 1998; Cutler et al., 1997)

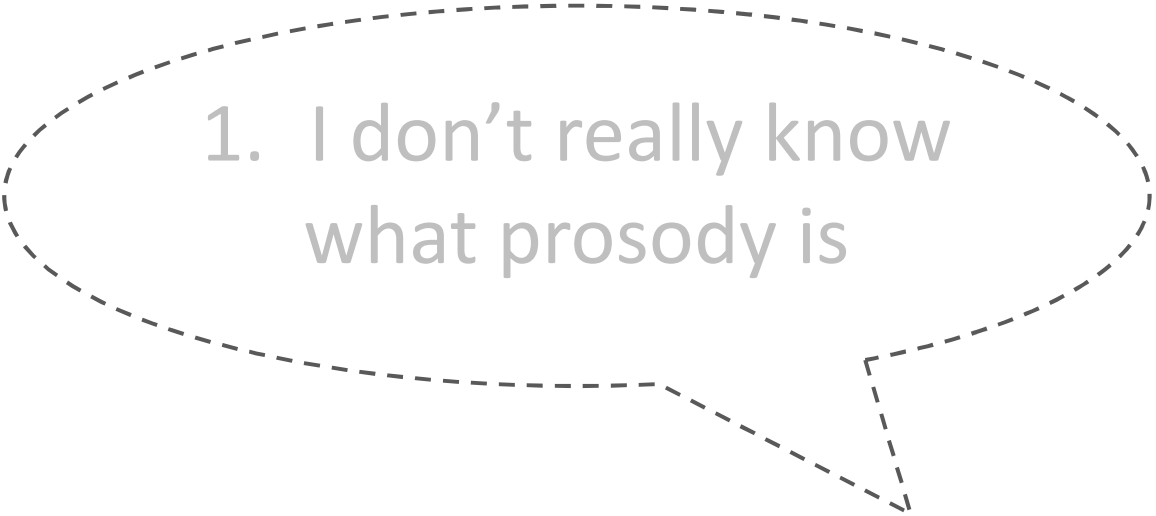
Prosody Is *Pervasive*

- Prosody conveys meaning at many different levels, including:
 - Syntax and phrasing Turn right. | Turn. Right.  
 - Focus, given/new Turn right. | Turn RIGHT.  
 - Pragmatic functions Turn right. | Turn right?  
 - Affect and emotion (annoyance, panic)  
 - Discourse phenomena (disfluencies, discourse markers)

(Ladd, 1980; Cutler et al., 1997; Nöth et al., 2000, 2002; Hirschberg, 2002)

Prosody Research

- Most work on prosody:
 - synthesis / TTS / generation (as in this TSD meeting)
 - descriptive linguistics
 - speech analysis
- Growing work in understanding, for specific tasks
- But prosody hasn't really diffused into language processing
 - No mainstream “off-the-shelf” technology (unlike for ASR)



1. I don't really know
what prosody is

2. I don't need prosody;
I use words



Language is Not Just Text (1)

First, real-world tasks increasingly involve **spoken language**

- **Voice-based dialog systems**
Call centers, navigation systems, gadgets, games, tutoring, etc.
- **Found data**
News broadcasts, phone calls, meetings, lectures, videos, etc.
- Even if you work on text, **your work may get applied to speech**
- Examples:
 - Process a request to a navigation system
 - Translate a news broadcast to another language
 - Summarize a meeting or web video
 - Determine sentiment in political speeches

Language is Not Just Text (2)

Second, spoken language is meant to be **heard**, not **read**

- When speech is automatically recognized, we lose the **prosody**
- Speech recognizers output only a stream of words
 - No overt structure (no punctuation, capitalization, formatting)
 - No “tone of voice”

→ Preserving prosody should aid future speech tasks



2. I don't need prosody;
I use words

3. Prosody needs linguists
to label data

4. Prosody requires
exotic new tools



Modeling Approach: Direct Modeling


Two general approaches

1. Symbolic /categorical

- Model discrete labels (e.g., pitch accents, boundary tones)
- Learn relationship between labels and classes of interest
- But requires hand annotation of prosody
- Lossy, relies on the discrete labels

2. Direct modeling

- Inputs are the classes of interest
- Models relate prosody with classes
- No human labeling of prosody
- Not lossy – allows feature discovery



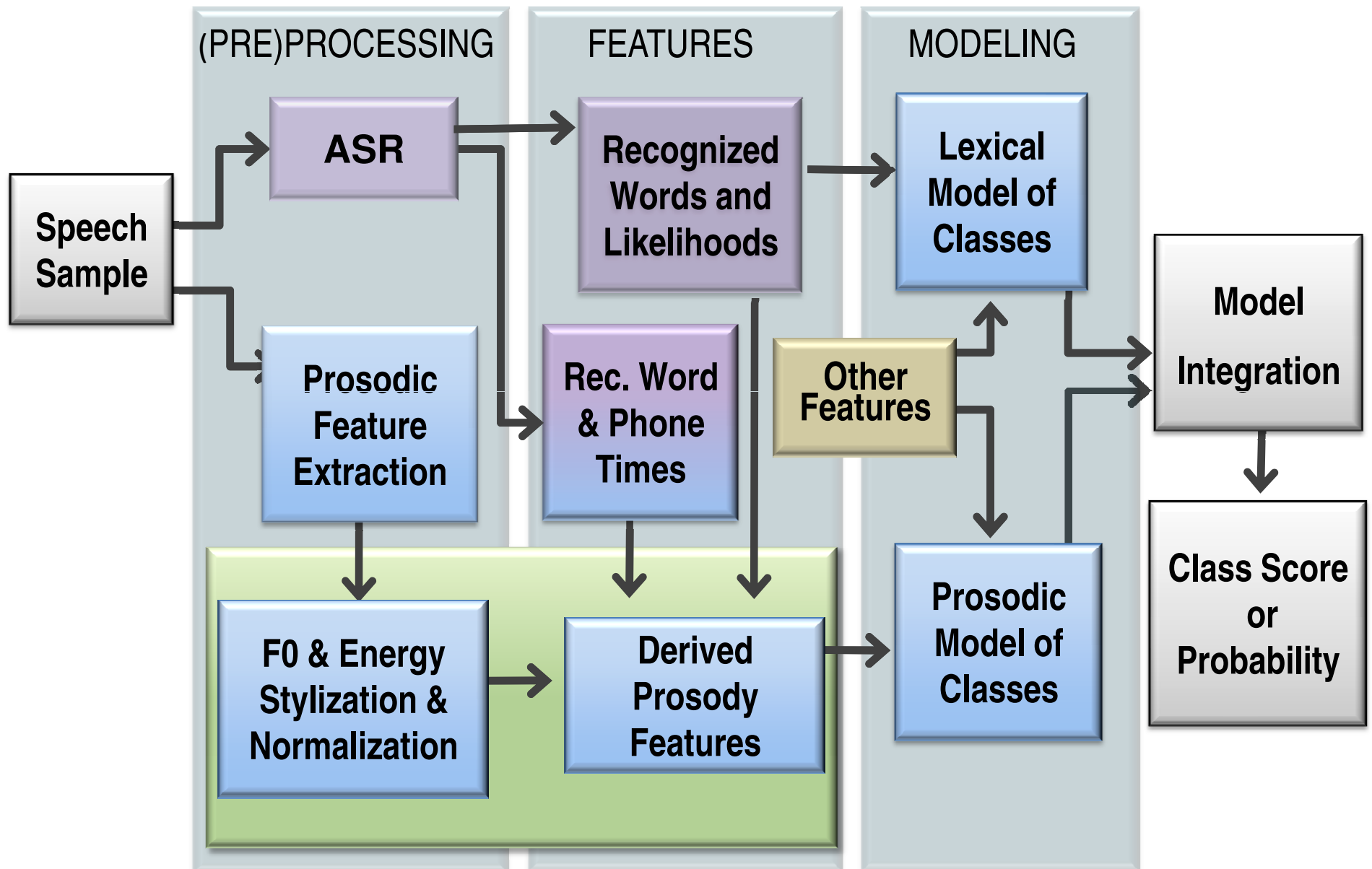
3. Prosody
needs linguists
to label data

(see web page for references)

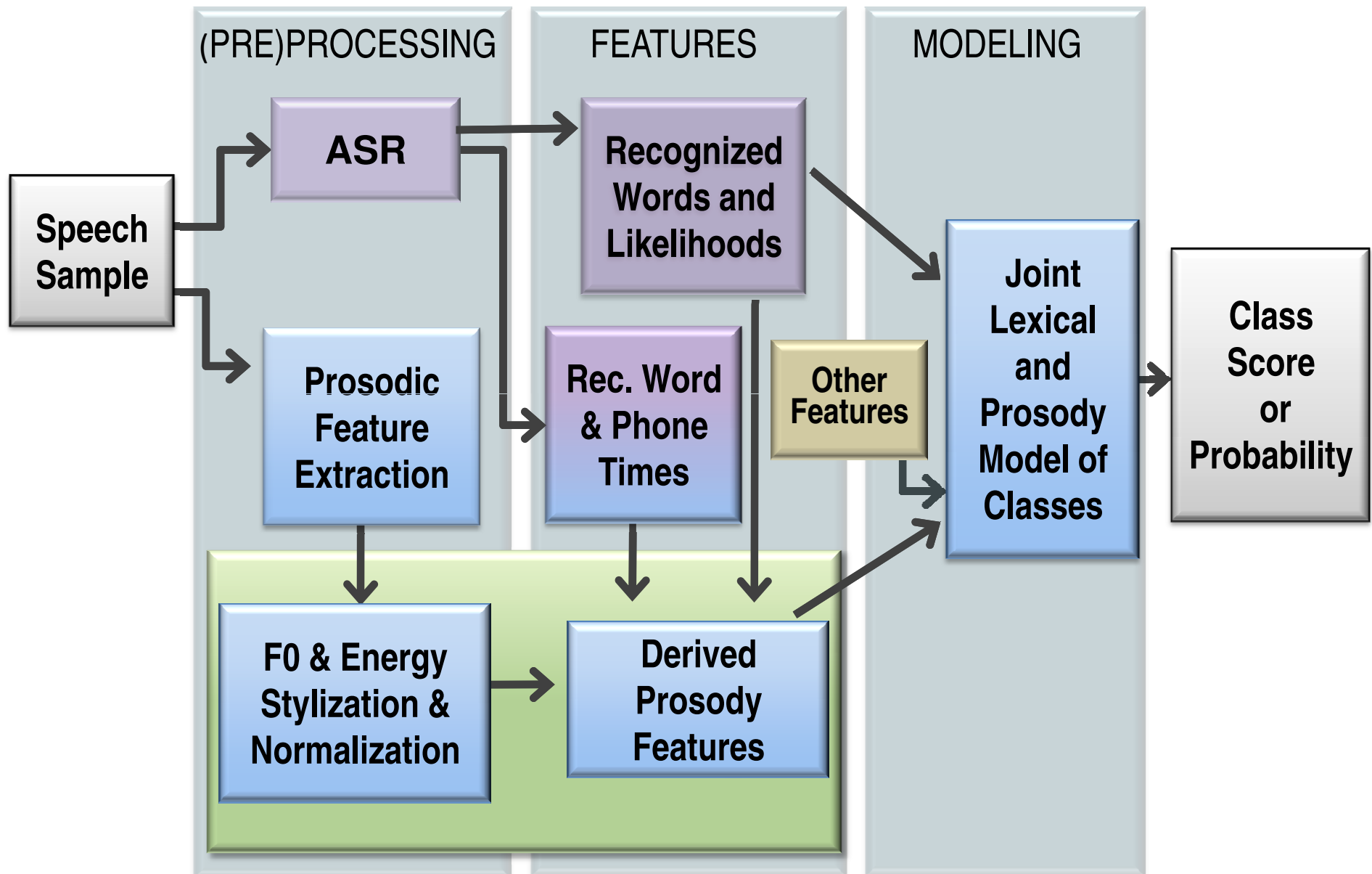
Aspects of Direct Modeling Approach

- **Automatic extraction of prosodic features**
- **Integration with lexical models**
- **Statistical modeling**
 - All models are probabilistic
 - can be configured for different tasks
 - Drawback: sensitive to train/test data mismatch
 - Benefit: important for coping with high variability
 - Many different prosodic ways to say similar things
 - Same basic features used to convey different meanings

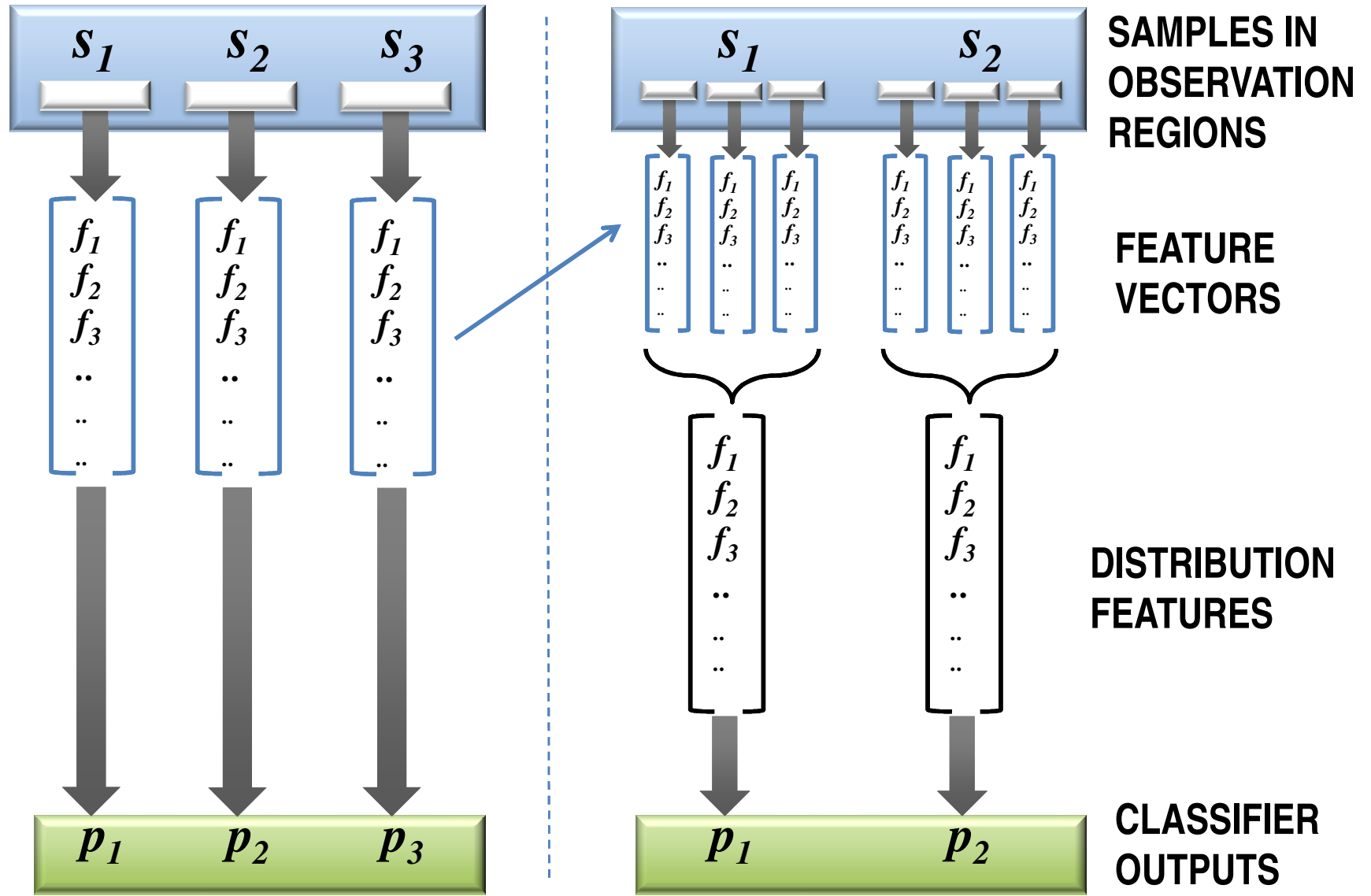
General Modeling Framework



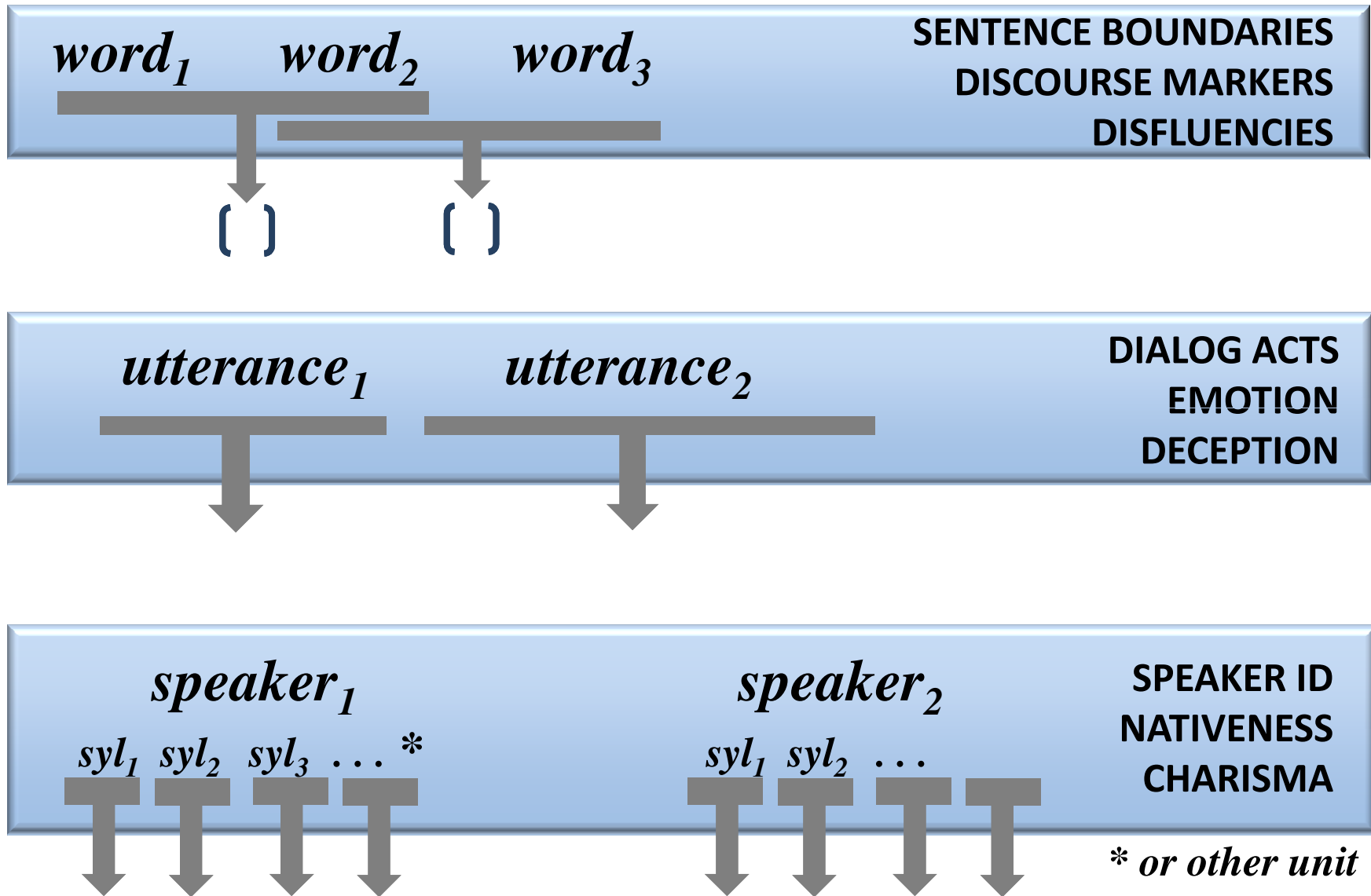
Joint Lexical and Prosody Modeling



From Samples to Classes

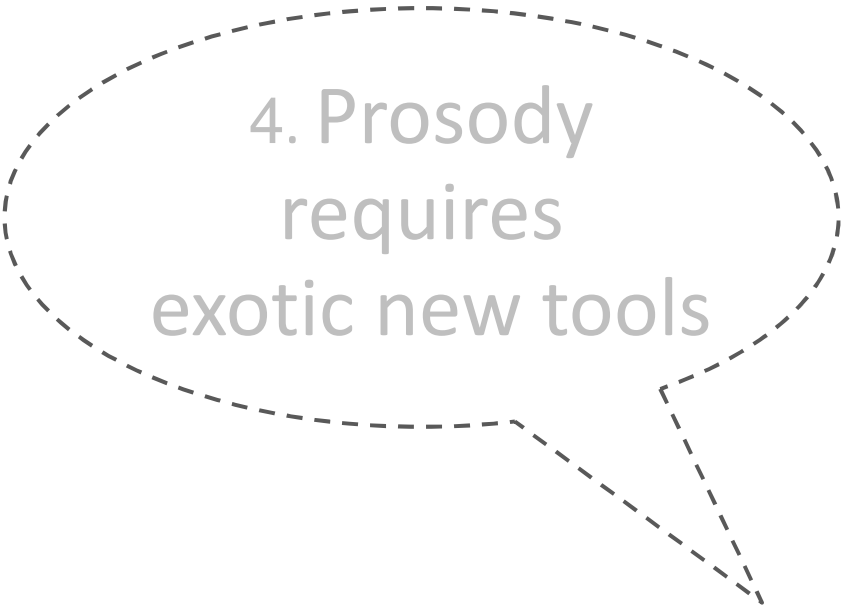


Task-Dependent Regions



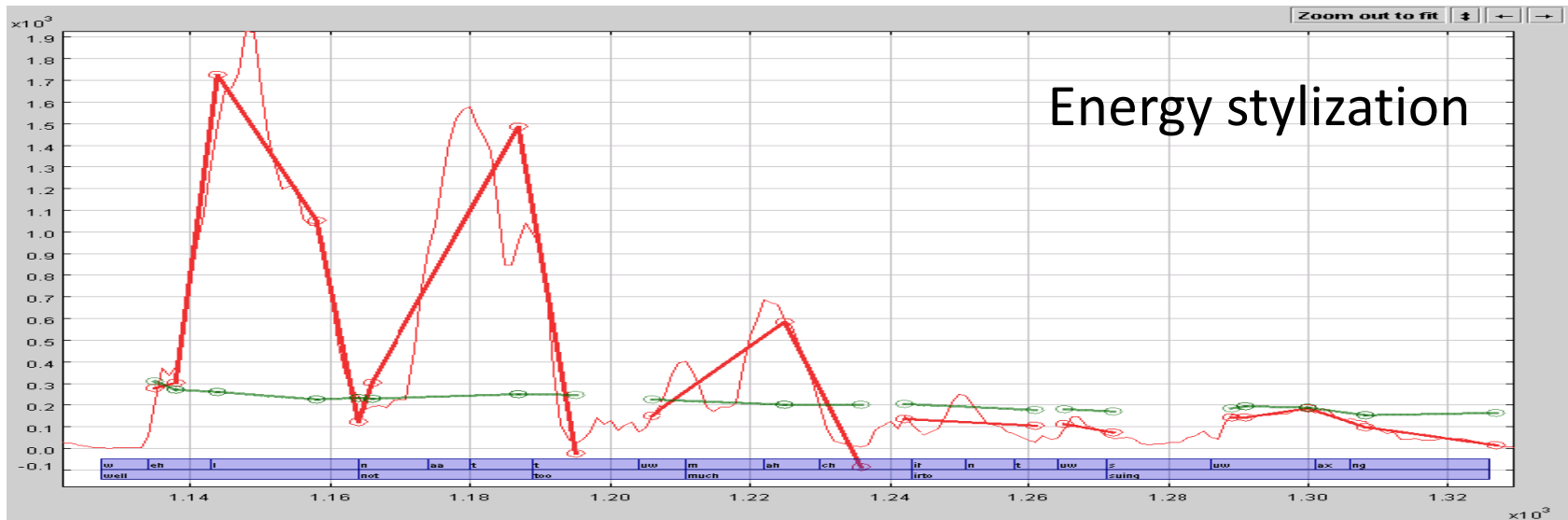
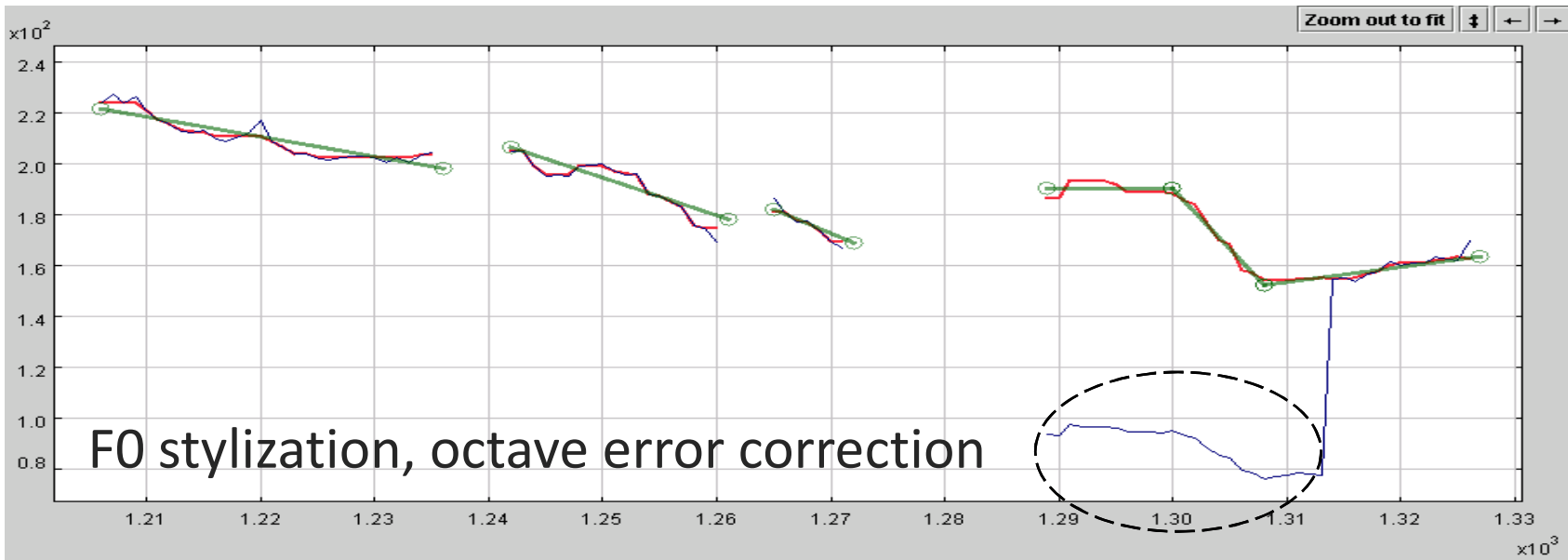
Prosodic Feature Extraction

- Extract F0 and energy, post-processing:
 - Smooth to remove microintonation
 - Estimate regions of pitch halving, doubling
 - Compute statistics for normalization
- We use an SRI-developed tool, graphical programming
- Public software available, e.g.:
 - Praat (U. Amsterdam)
 - Snack Sound Toolkit (KTH)
 - See web page for links



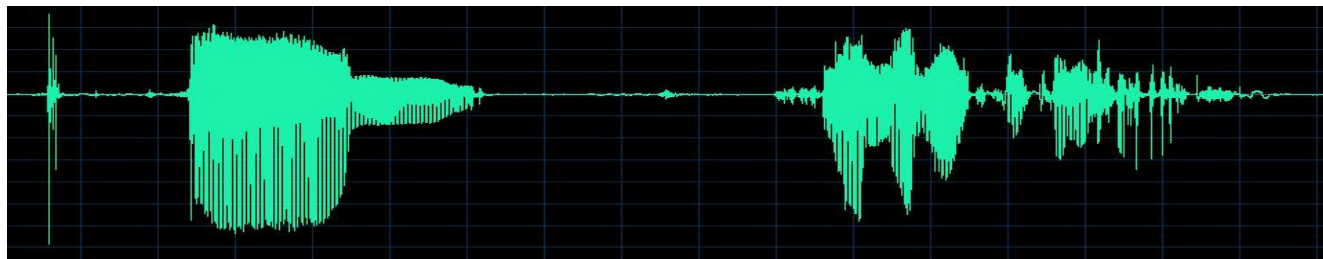
4. Prosody
requires
exotic new tools

F0 and Energy Stylization



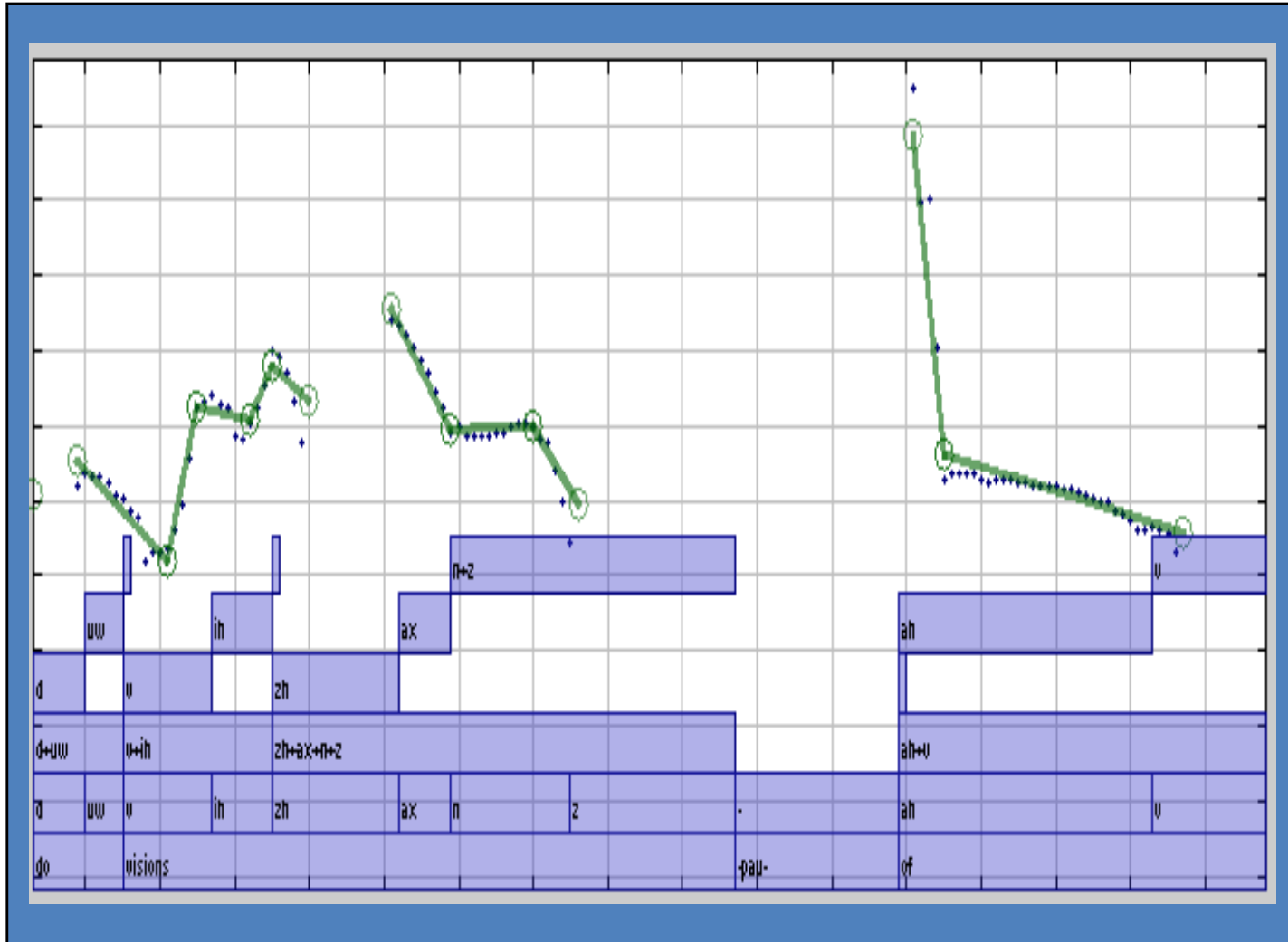
Alignment of Derived Features

- Align post-processed pitch and energy with speech recognizer word and phone time marks
- Compute derived features based on units of interest
- Example of word based regions:



	um	my	major
mean F0 in word:	140	148	157
final F0 slope:	-0.034	2.77	-4.531
mean energy:	834.9	652.1	1211.6

Other Types of Alignment Units



← Pitch

← Vowels

← Syllables

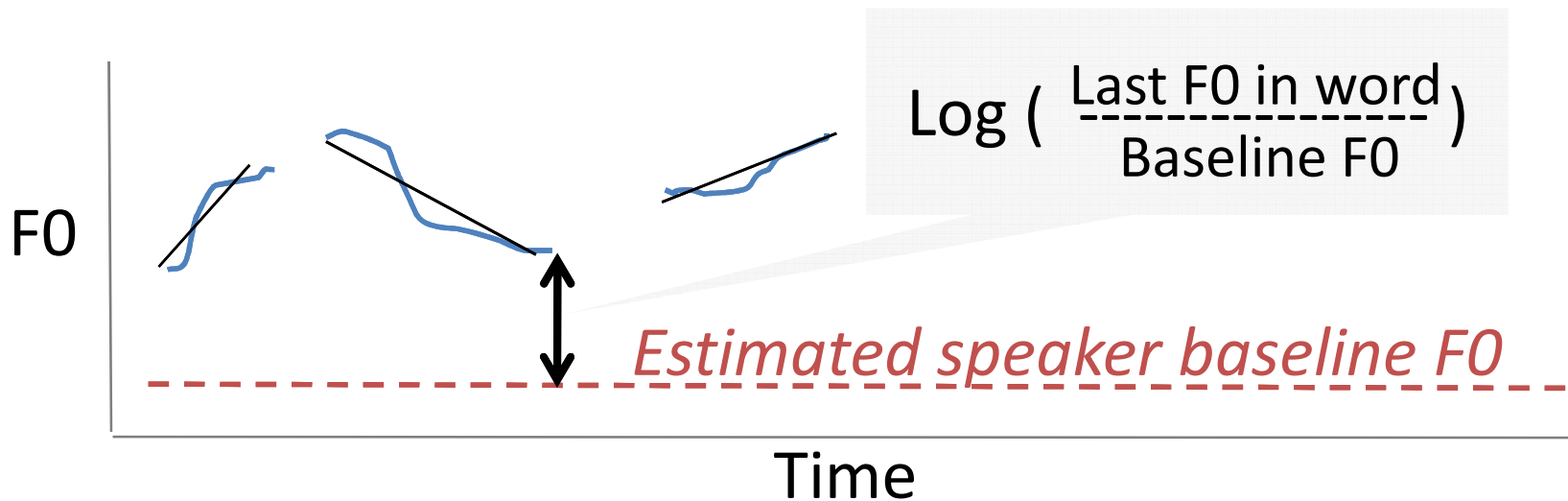
← Phones

← Words

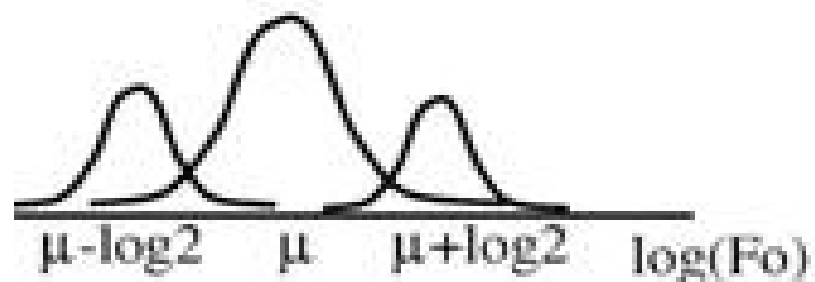
Derived Prosodic Features (1)

- **Pitch (F0)**

- Features capture range, slope, contour type, and discontinuity, e.g.:



- Speaker baseline from lognormal tied mixture of pitch



Derived Prosodic Features (2)

- **Duration, pause, and speaking rate**
 - Normalized for intrinsic phone duration
 - Lengthening, pause behavior, conditioning (e.g., pitch in longest phone)
- **Energy**
 - Stylized as for pitch; normalized for channel, often use vowel regions
 - Features capture magnitude and direction of energy changes
- **Voice quality**
 - Used for some tasks (emotion)
 - Highly speaker dependent

Modeling Approaches

Prosodic features modeled by a range of (familiar) classifiers:

- Generative models
 - GMMs, HMMs
- Discriminative models
 - Decision trees, Boosting, SVMs, CRFs
- Distribution modeling for variable-length sequences
- Integration with other (e.g., lexical) features
 - Posterior or score combination
 - Output of one as input feature to other
 - Feature concatenation (e.g., as SVM input)

Performance

5. Prosody doesn't help
that much



Prosody for Higher-Level Tasks

- Prosody helpful for tasks related to affect, emotion, user state
 - Emotion (See web page for SRI references)
 - Deception
 - “Hot spots” in meetings
 - Action items in meetings
 - Level of certainty in tutoring dialogs

Example from tutoring system (Liscombe, et al., 2005)



um because the f- the forces acting on it are th- are uh are
proportional to size so that they equal out to the same thing
no matter w- how how much mass the object has

Sample Task: Emotion Classification

- Emotion is a “hot” topic, lots of work and applications
 - Call center / customer service
 - Navigation systems
 - Speech-enabled toys and games
 - Automatic tutoring
 - Health monitoring
- Problem: lack of large, public data with real emotion
- Real emotions occur only in real applications, but
 - Data is proprietary
 - Privacy issues for speakers
- Therefore, most work is on acted emotion
- Easier to obtain, but doesn't scale to real emotion

(Leading group at Erlangen; Cowie et al., 2001; Batliner et al., 2003)

Large Study of Natural Emotion

- Example of natural (not acted) emotion
- Mock air travel application, DARPA “Communicator” project
- Large data set (22,000 utterances)
- Emotion labels: “neutral”, “annoyed”, “frustrated”, other
 - 5 labelers covered subsets of the data
 - 2 labelers later created “consensus” labels for disagreements
- Decision tree classifier
 - Based on prosody and lexical features
 - Based on automatically recognized words
 - Uses sampling to equate class priors

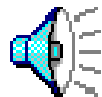
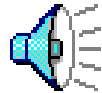
(Ang et al., 2002)

Emotion Samples

Neutral

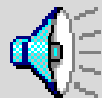
—*July 30*

—*Yes*



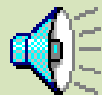
Disappointed/Tired

—*No*



Amused/Surprised

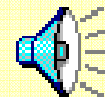
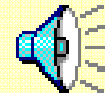
—*No*



Annoyed

—*Yes*

—*Late morning*



Frustrated

—*No*

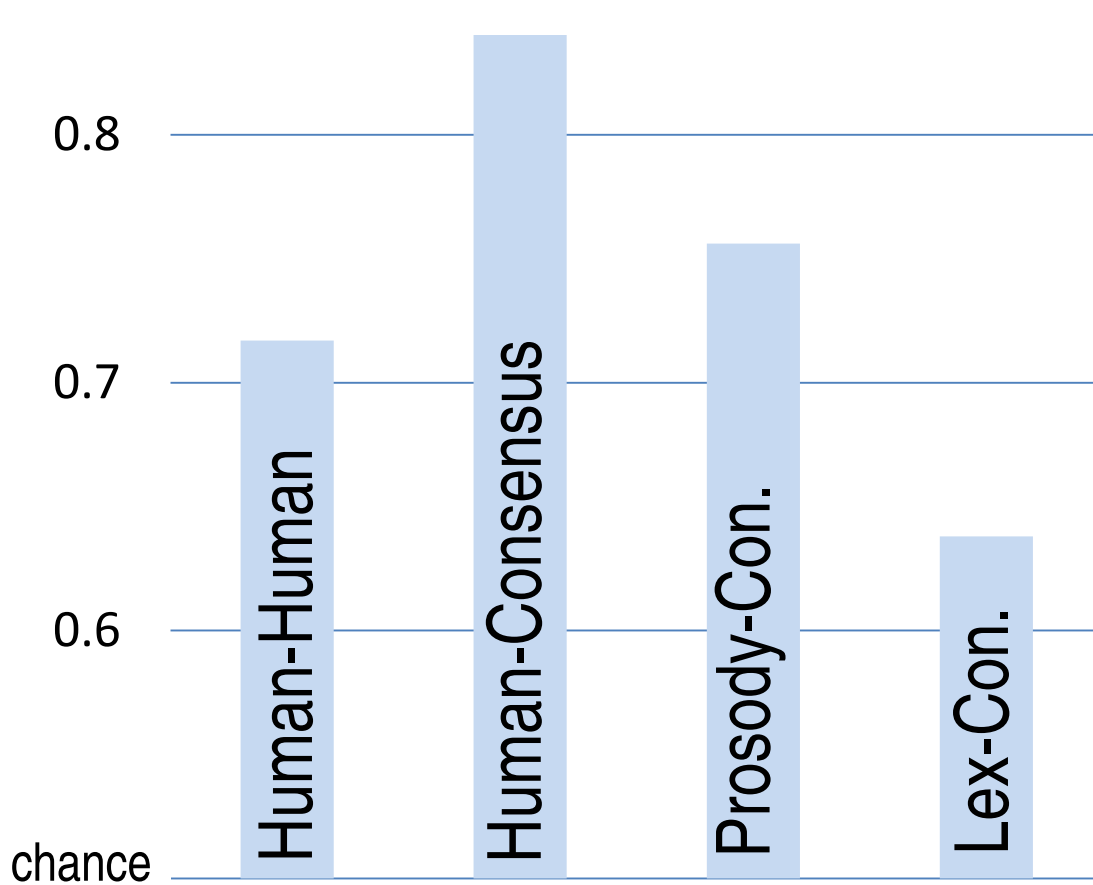
—*I am dePARTing...*



Difficulties: short utterances, speaker differences

(Ang et al., 2002)

Emotion Classification Results



(Ang et al., 2002)

- Prosody agrees with consensus better than humans with each other
- Best features: max norm. F0 in lengthened vowels, duration
- In this domain, lexical cues less helpful

5. Prosody doesn't help that much

6. Prosody's useful only for
“fuzzy” tasks, like emotion

7. Prosody matters only for
contrived examples, like
ambiguous sentences



Importance of Sentence Boundaries

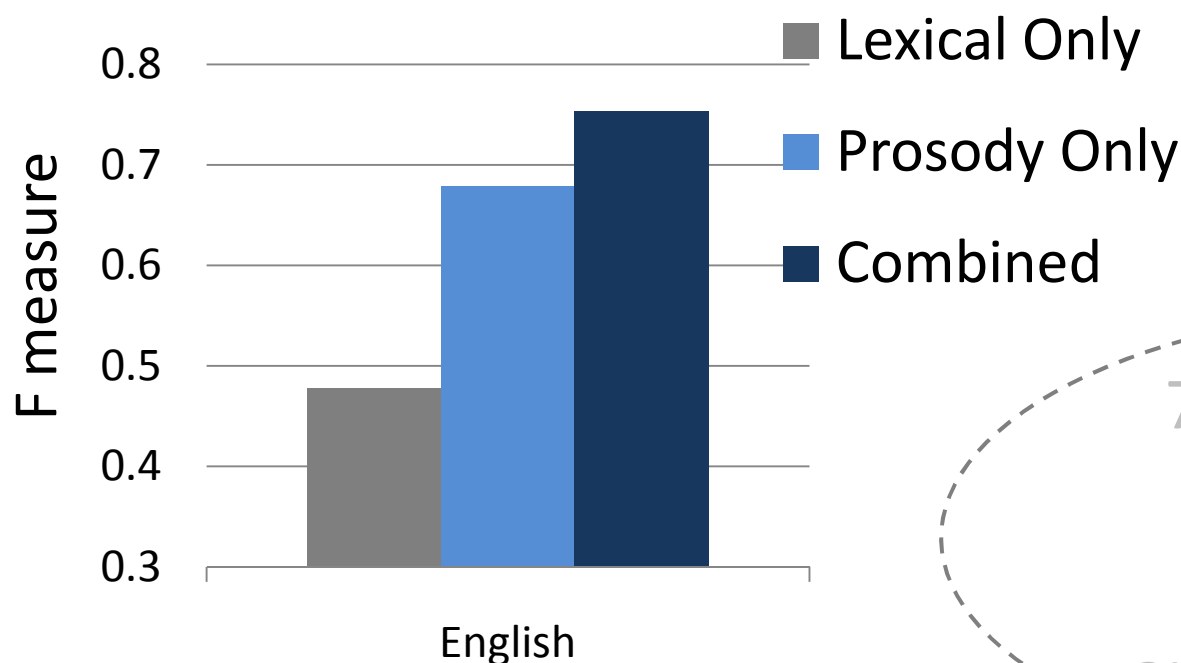
- **Human reading comprehension** (Jones et al., 2005)
- **Parsing, machine translation** (Harper et al., 2005; Matusov et al., 2006)
 - Both operate on sentence level units
 - Need short units; otherwise too expensive
 - Studies show benefit of linguistic boundaries
 - Optimal sentence length depends on task
- **Information extraction** (Makhoul et al., 2005, Favre et al., 2007)
 - Short units can break up entities; longer can merge them
 - In addition, help predict location of names
- **Extractive summarization** (Murray et al., 2005)
 - Relies on sentence units

Computational Models for Punctuation

- Typically involve combining lexical and prosodic cues
- Prosody model
 - Features: pauses, duration, F0, turn taking
 - Models: decision trees, neural networks
- Language model (LM):
 - Baseline: N-grams over words and punctuation; POS tags
 - More recently: parser-based (Favre et al., submitted)
 1. Segment using prosody and word N-grams
 2. Rescore segmentation lattices with lexicalized CFG
- Prosody and LM features often combined via
 - HMMs, maximum entropy models, CRFs, boosting

Sentence Segmentation Results (ICSI/SRI)

- English news speech (TDT4); ASR output
 - Lexical model: word N-grams
 - Prosodic model: pause, pitch, duration, energy patterns
 - Boosting, 1K iterations



6. Prosody's useful only for "fuzzy" tasks

7. Prosody matters only for contrived examples, like ambiguous sentences

(Fung et al., 2007; Hakkani-Tur, Favre)

8. Prosody will slow
my system down



Speed: Two Arguments

1. **Prosody is fast compared to ASR** (much faster than real time)
2. **Prosody can actually speed up a dialog system**
 - End-of-utterance detection (“endpointing”)
 - Example from driving simulator

What do I dooooo [2.4 sec] after I cross the river?

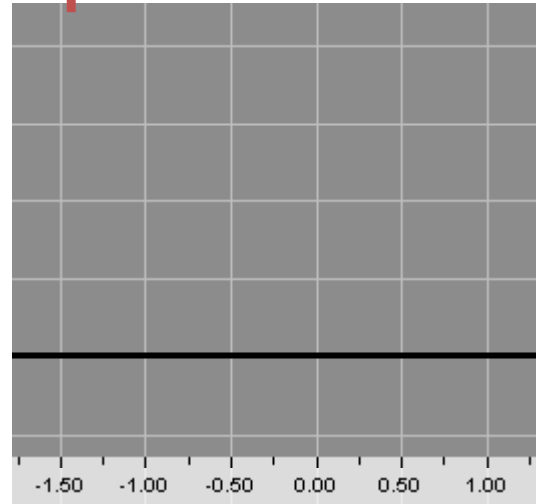


- If endpointer cuts off:
 - Miss crucial content (post-hesitation → higher entropy)
 - Annoy user
- If increase pause threshold, more wait time at true boundaries
- Need causal model: prosody **before the pause**

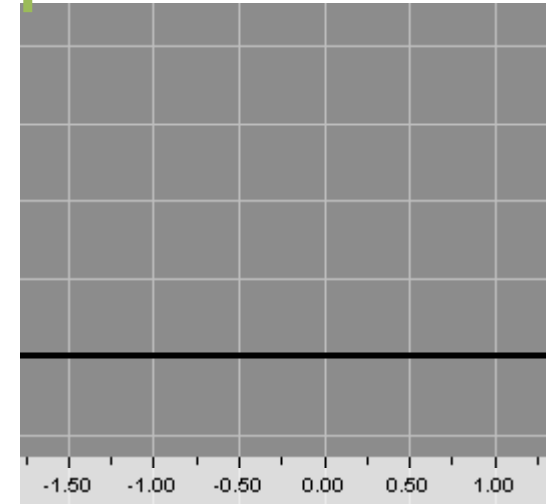
Improved Endpointing: Desired Outcomes

**Standard
Endpointer**
(fixed
nonspeech
threshold)

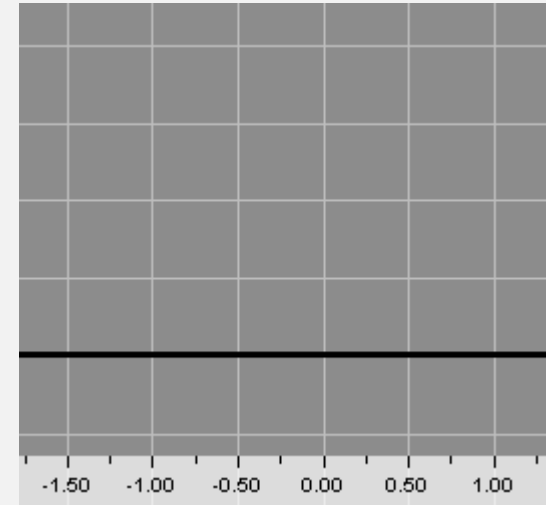
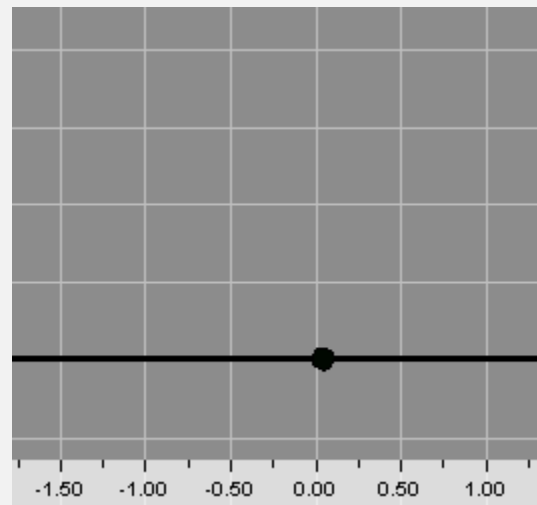
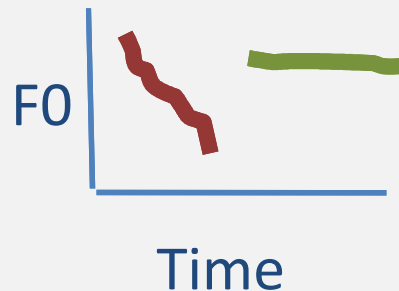
Speaker is Done



Speaker is Not Done



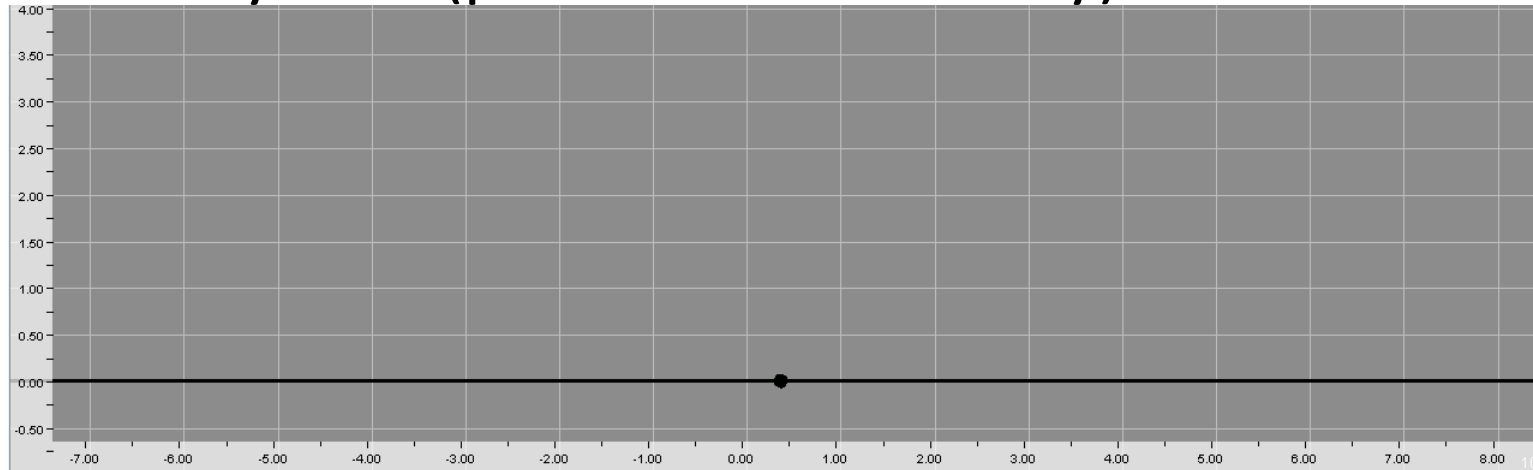
**With Pitch
Modeling**



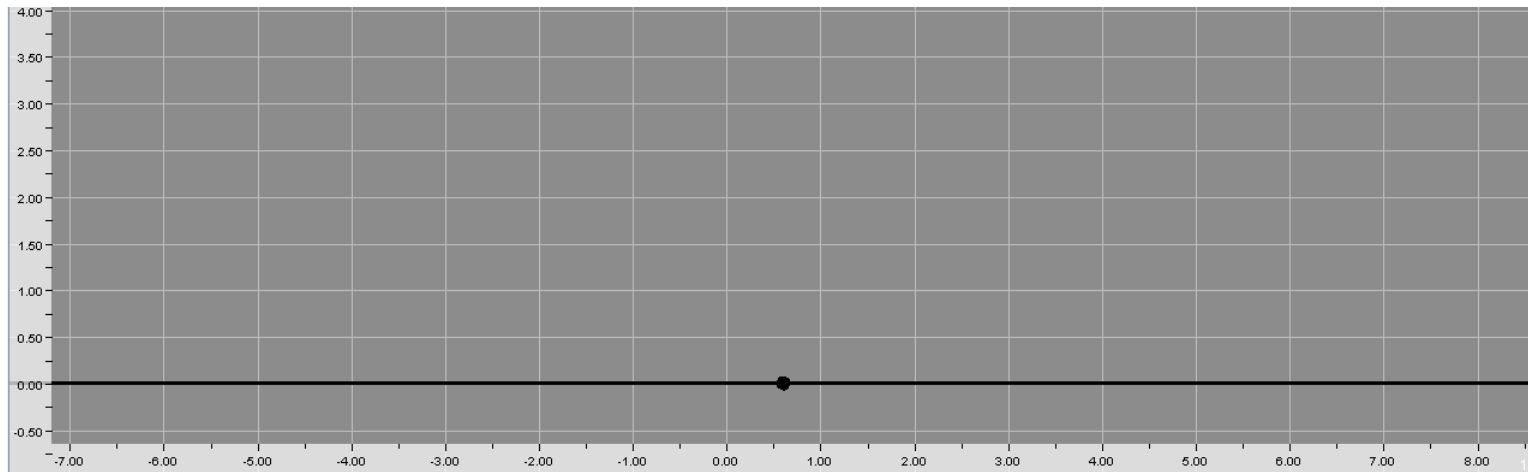
(Movies courtesy of Harry Bratt)

Improved Endpointing

Standard System (pause threshold only)



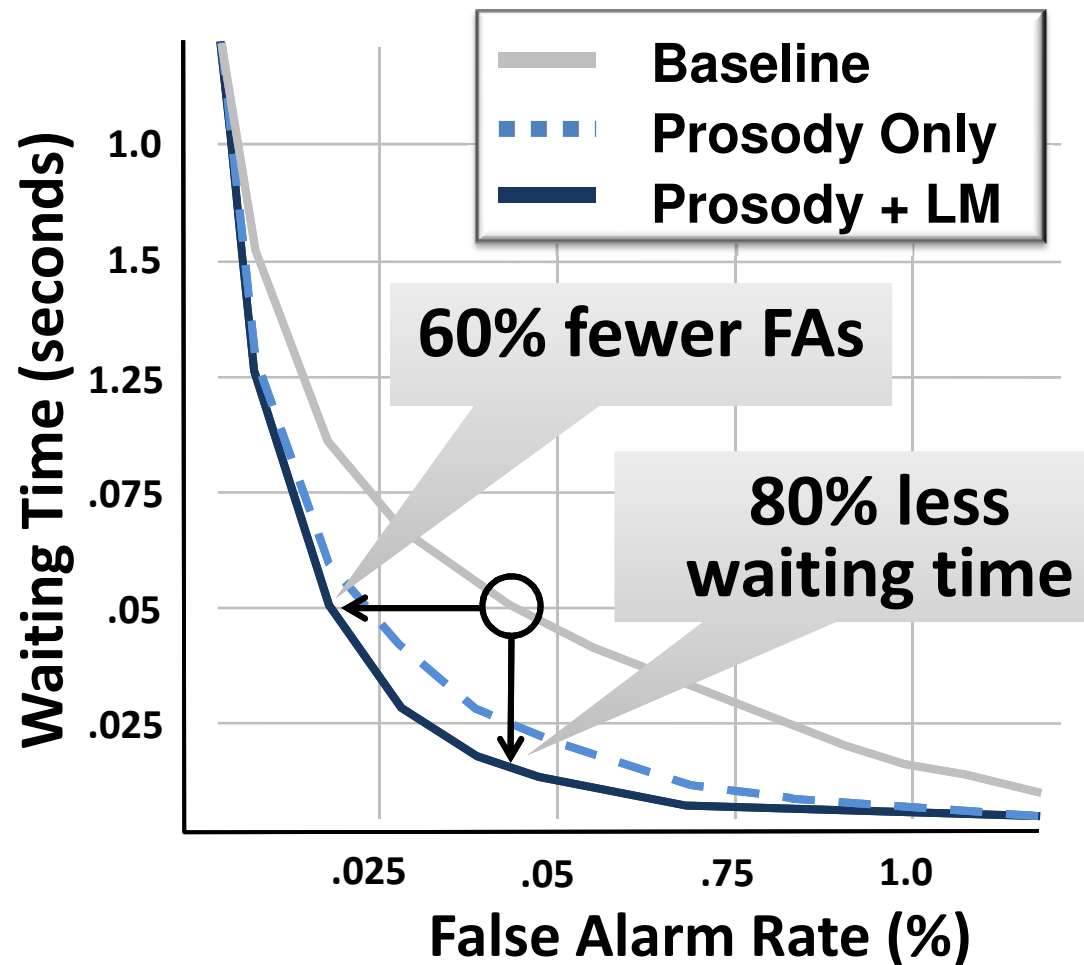
With Pitch Model



(Movies courtesy of Harry Bratt)

Prosodic Endpointing: Results

- Corpus of human-computer dialogs
 - Dramatically reduced
 - False alarm rate
 - Waiting time
 - Most of the benefit is from prosody
- Prosody can make your system faster. Adds minimal time if already using ASR



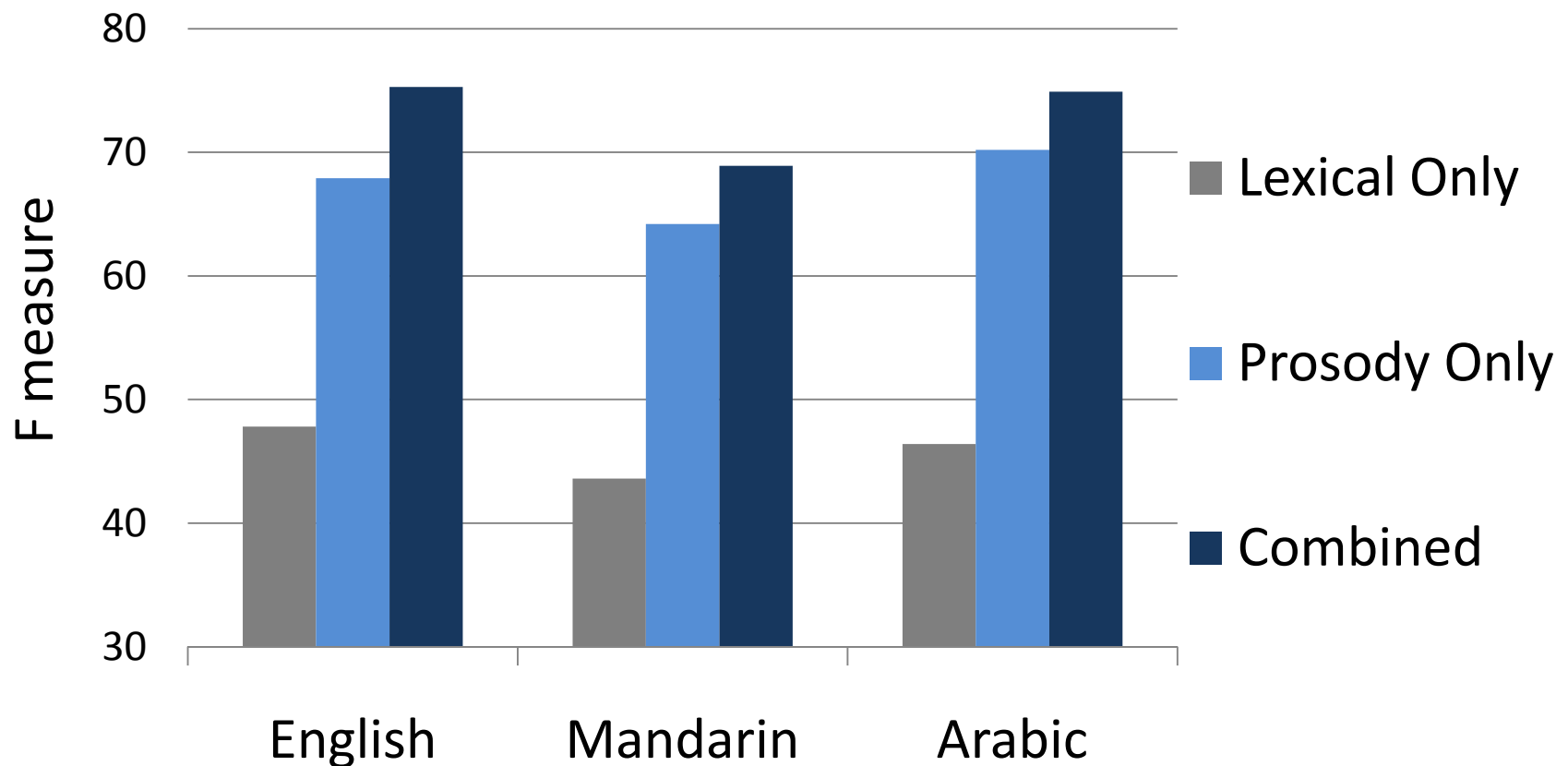
(Ferrer et al., 2002)

9. Prosody doesn't generalize
(across tasks, languages,
speaking styles)



Across Languages (Sentence Task)

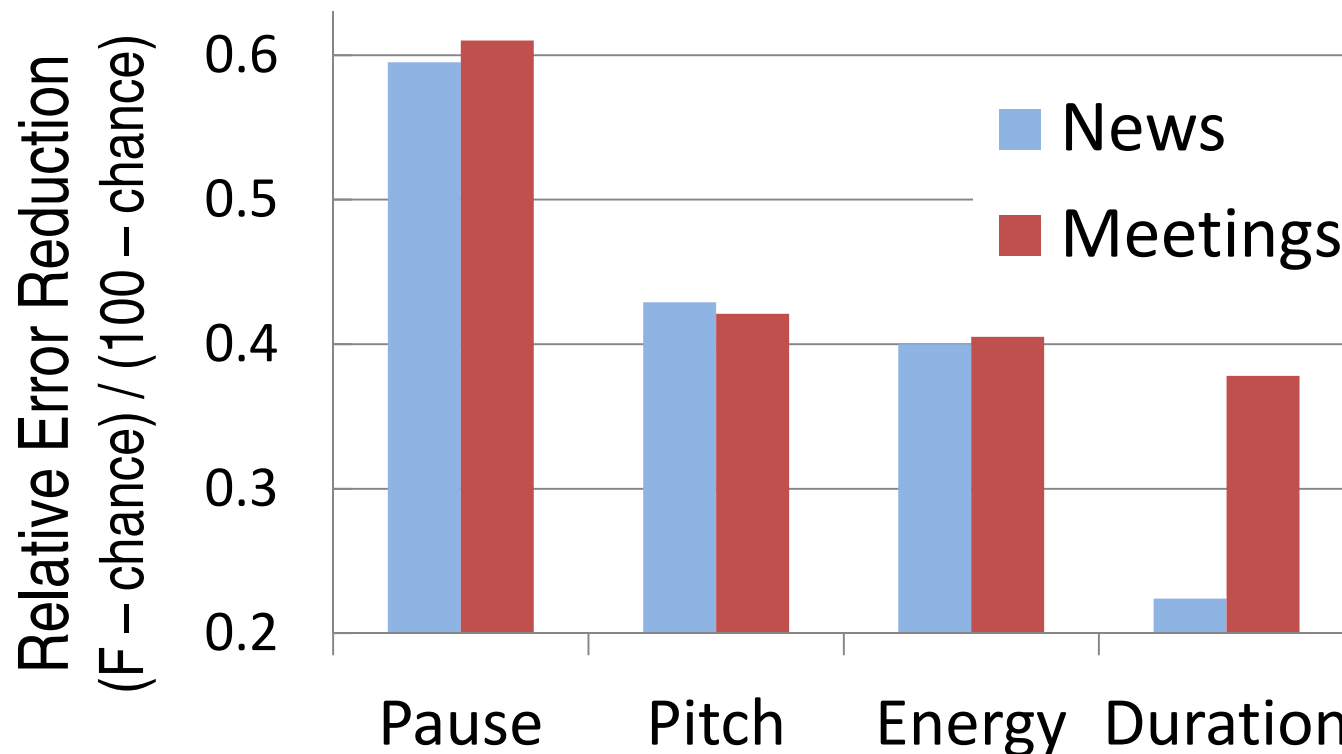
- Same experiment seen earlier; adding Mandarin and Arabic
- Modeling and features held constant over languages



(Fung et al., 2007; Hakkani-Tur, Favre)

Across Speaking Styles (Sentence Task)

- Styles differ in prosody, but some strikingly similar results
- Example: sentence segmentation (English)
 - News (TDT4), Meetings (ICSI); different speakers, recordings
 - Same available features, boosting, reference transcripts



10. Prosody is too speaker-dependent



Prosody and Speaker Dependence

Prosody is highly dependent on the speaker. Two approaches:

(1) **Normalize out the speaker effects**

- Using speaker statistics
- Evaluate by giving speaker as feature to classifier

(2) **Model the speaker-specific behaviors**

- Despite less training data, better matched to the speaker
- First experiments (dialog act segmentation in meetings): many speakers improved from combination of SI and SD prosody models (Kolář et al., 2006)
- Long-term goal: speaker clusters

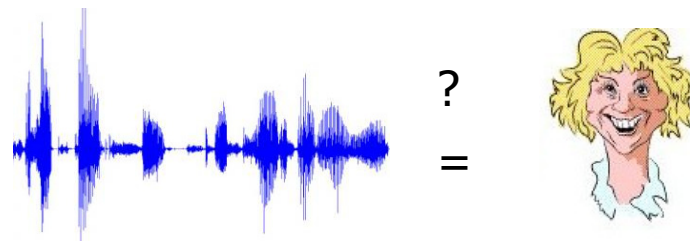
Prosody for Speaker Classification

- But: speaker dependence is a double-edged sword
Can **use** prosodic differences to **help classify speakers**
 - Speaker ID (see web page)
 - Diarization (Friedland, submitted)
 - Nonnativeness ID (Shriberg et al. , 2008)
 - Level of charisma (Rosenberg & Hirschberg, 2005)

Prosody and Speaker Recognition

- Task: determine whether speech is from a known speaker

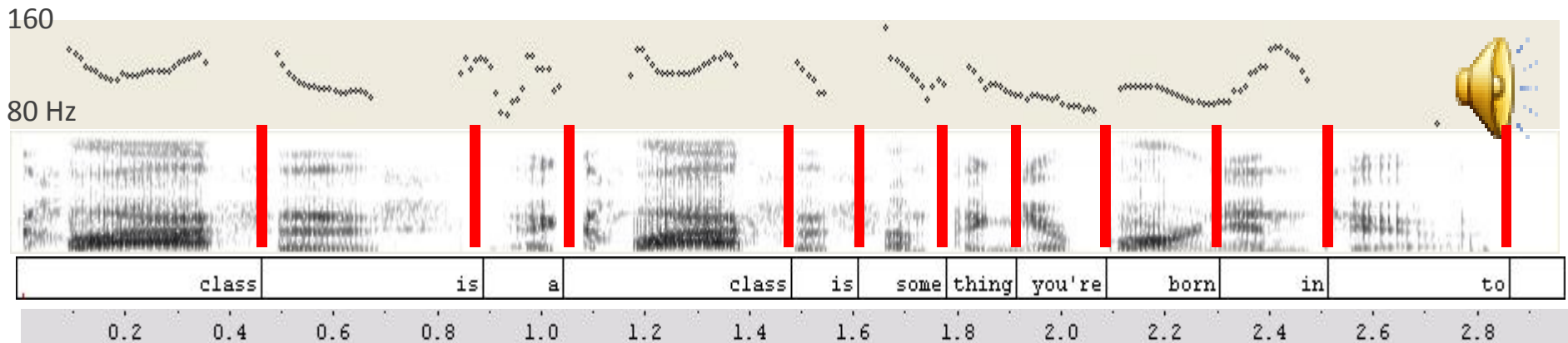
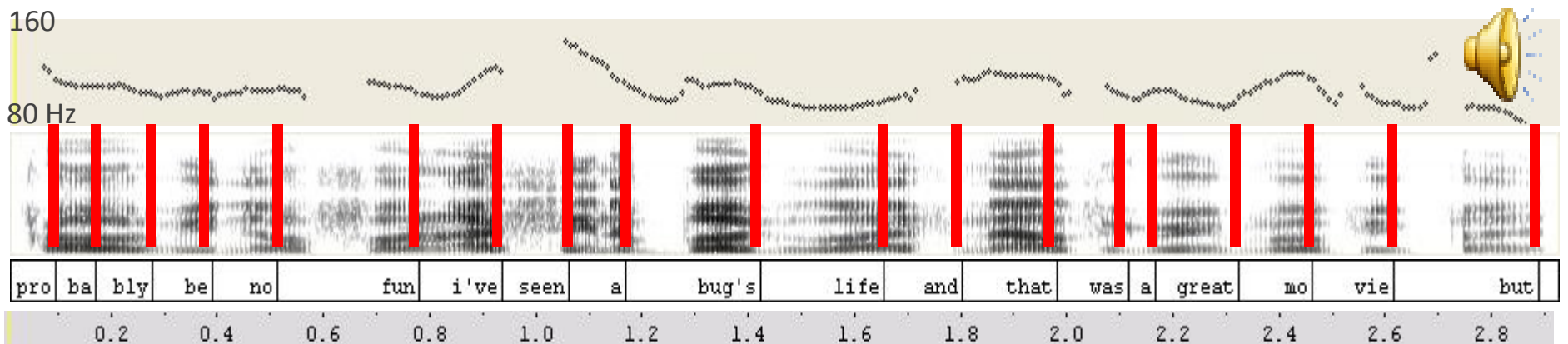
Speaker Verification
Is this Mary's voice?



- Standard systems model information from short time slices
- Cepstral features: energy in different frequency bands
- A speaker is modeled essentially as a “bag of frames”
- Approach doesn't capture longer-range behaviors

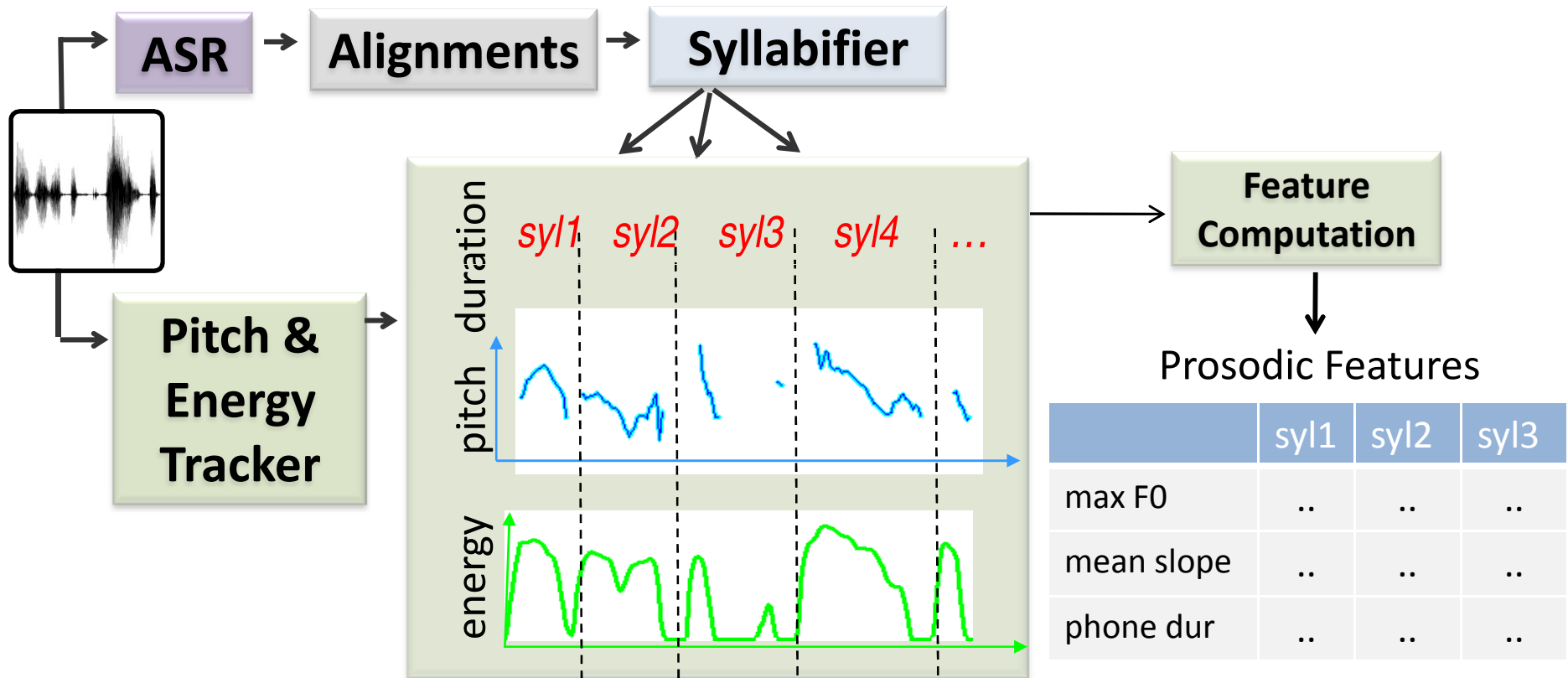
Two Confusable Speakers

- Two male speakers confused by SRI (SOA) cepstral system
- Very similar pitch range. Same elapsed time shown for each
- But: 1st speaker has nearly twice the word/syllable rate as 2nd



Prosodic Speaker ID System

- Syllabify ASR output; compute prosodic features per syllable
- (F0 polynomial feats also extracted based on energy valleys)

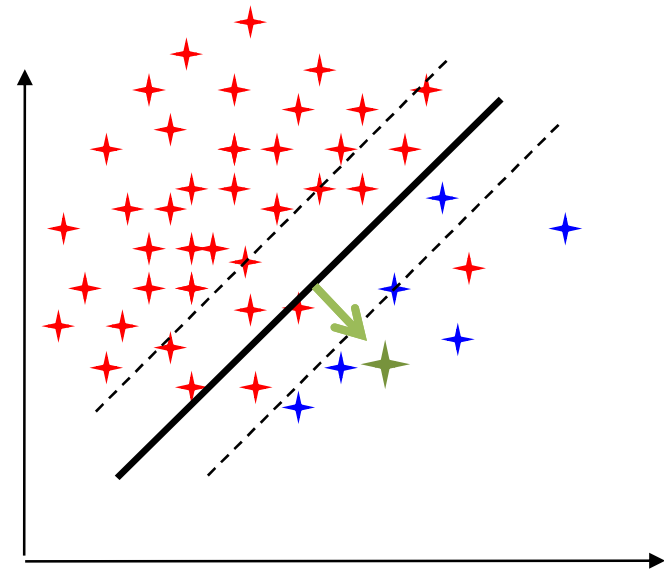


(Ferrer et al., 05, 07; Shriberg et al., 05, 07; Dehak et al., 2007)

Prosodic Speaker ID: Steps

1. Extract syllable-level features for syllabified ASR output
2. Form N-grams of syllable-level features
3. Transform features (vector of posterior weights for GMM)
4. Apply rank normalization and NAP
5. Concatenate transformed features
6. Train SVM. Testing: score = signed distance from hyperplane

- + Background sample
- + Target sample
- + Test sample



(Ferrer et al., 05, 07; Shriberg et al., 05, 07)

Prosodic Speaker ID: Results

- NIST 2008 evaluation system, English telephone data
- Training: 1 / 8 conversation sides (side ~ 2.5 min). Test: 1 side
- Prosody complementary; benefit increases with more data

Error Rate (Actual DCF x 10)

System	1 conv. side	8 conv. sides
Baseline (cepstral GMM)	.105	.080
Baseline + Prosody	.089	.046
Improvement from prosody	10%	40%

(Graciarena et al., 2008 [SRI NIST SRE08 Evaluation System])

Summary

Feasibility:

- Prosody important in natural spoken language
- General modeling framework relates prosody to classes of interest
- Completely automatic; does not use hand labeling of prosody
- Public feature extraction tools available; standard machine learning

Performance:

- Sample results show prosody helpful for range of tasks
 - From sentence boundaries, to emotion, to speaker classification
- Prosody has minimal compute demands compared to ASR
 - And can even save time in a dialog system (endpointing)
- Although tasks, languages, styles, and speakers differ in prosody
 - Remarkable generalizations for some tasks (sentence boundaries)
 - Can make use of differences for others (speaker ID)

Conclusions

- Spoken language is not complete without its prosody
- Although complex, prosody can be harnessed for more intelligent computational processing of spoken data

Thank You

More Information

Links to papers, references, and general resources can be found at:

www.speech.sri.com/people/ees/prosody