Prosody and Prosodic Models

ICSLP 2002 – September 16, 2002

Denver Colorado

Chilin Shih and Greg Kochanski

Bell Laboratories

http://prosodies.org/tutorial2002/

Section 1: What is Prosody

Theme: Prosody encodes different information simultaneously

What is the implication for prosodic modeling?

- Lexical information
- Intonation type
- Discourse functions
- Emotion
- Other physiologically based effects

Prosody Encodes Lexical Information

- Stress Language

  Example: English, Russian

  Stress location is part of the lexical entry, but the pitch contour (accent type) on the stressed syllable may vary

- Accentual Language

  Example: Japanese, Swedish

  The location of the accent is lexically marked. Accent type in a word is typically fixed

- Tone Language

  Example: Chinese, Navajo, Igbo

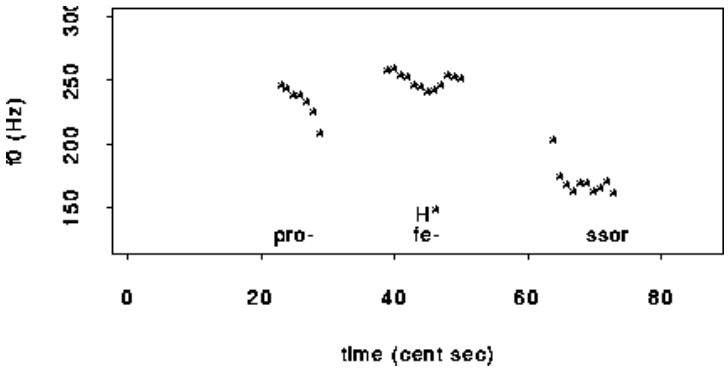
  Lexically determined tone on every syllable or every word

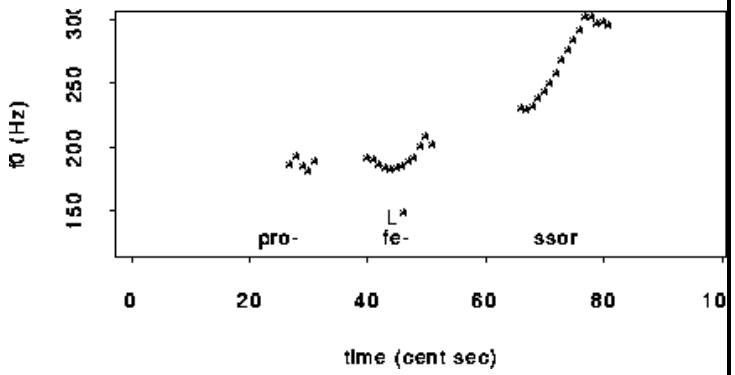
Stress Language: English and Russian

Fixed stress location: proféssor

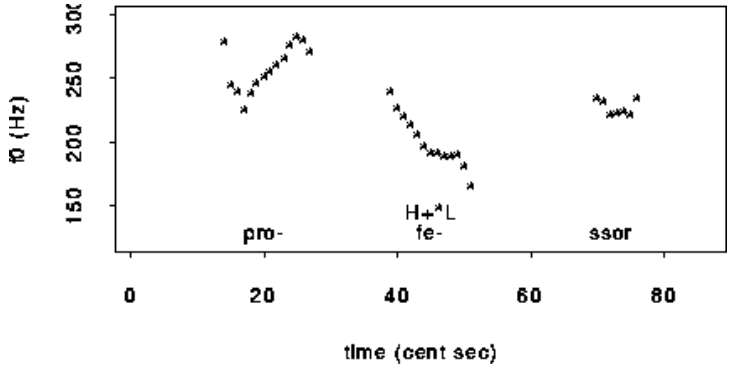
Many possible pitch contours:

English:



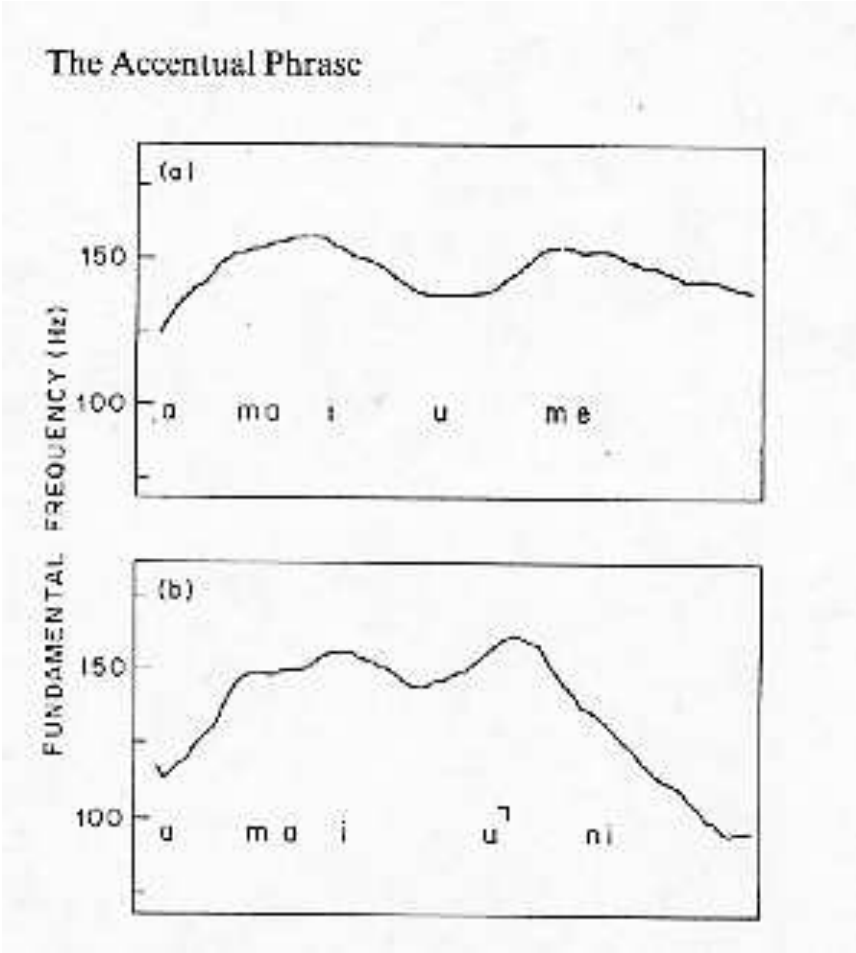


Russian:



Accentual Language: Japanese

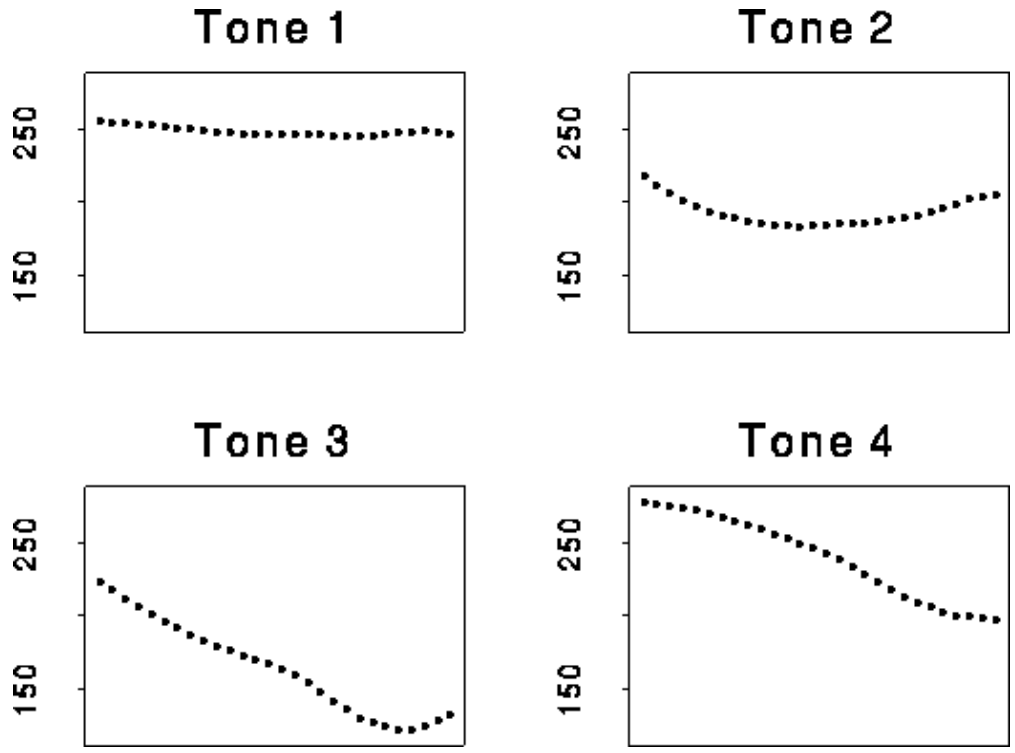
Pierrehumbert and Beckman (1988) Japanese Tone Structure. MIT Press. p. 27.



Tone Language: Mandarin Lexical Tones

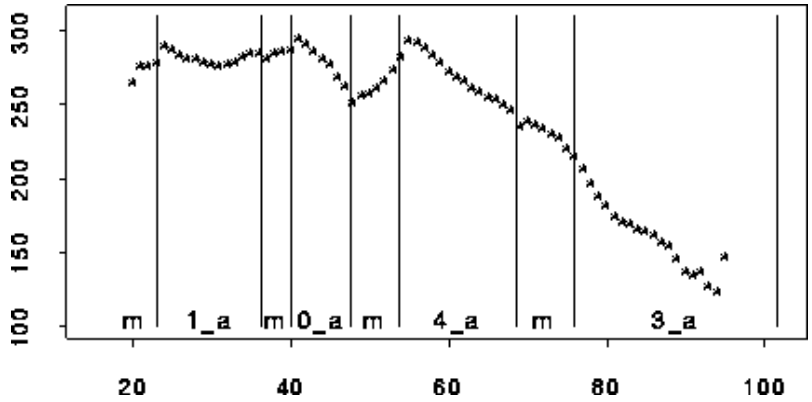
Mandarin has four lexical tones.

In addition, the lack of tone is lexically distinctive (neutral tone).

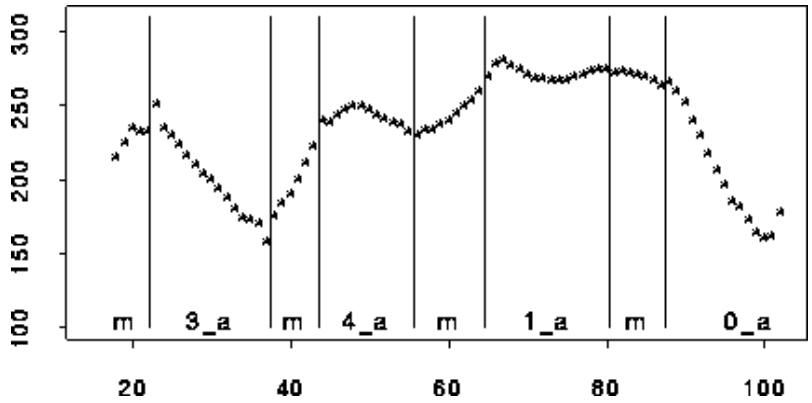


Tone Language: Chinese Sentences

Ma1-ma0 ma4 ma3 “Mother scolds the horse.”



Ma3 ma4 ma1-ma0 “The horse scolds mother”



Prosody Expresses Intonation Type

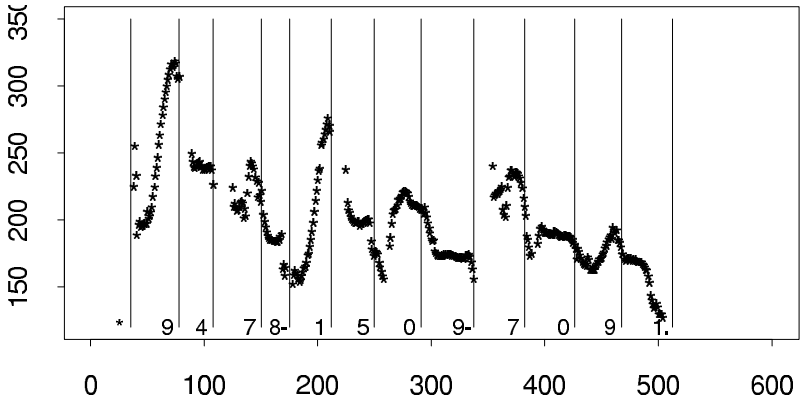
Declarative, question ...

Language specific patterns which may reflect universal features

- English declarative:

  Final fall (H* L- L%)

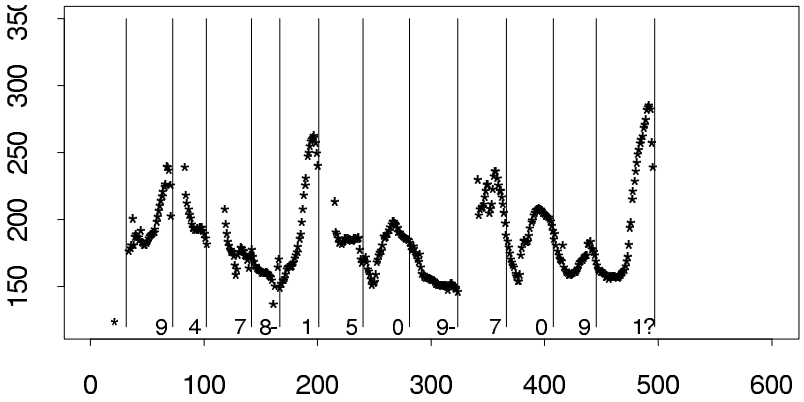
  9478-1509-7091.



- English question:

  Final rise (L* H- H%)

  9478-1509-7091?



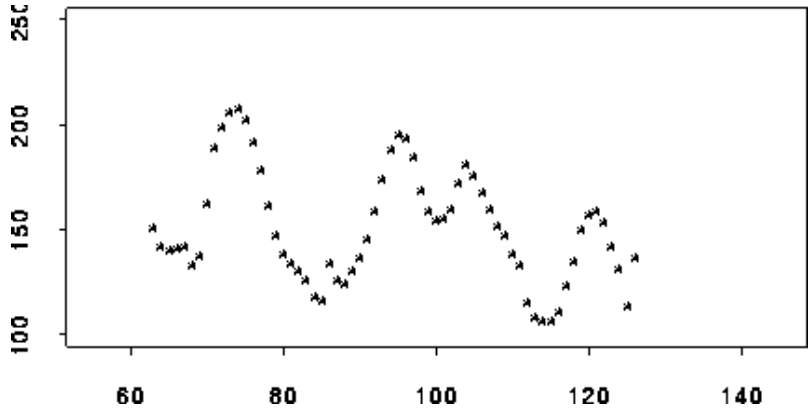
Chinese Intonation Type

Preserve lexical tone shapes. Question uses higher pitch near the end of the sentence (Yuan, Kochanski, Shih 2002)

- Chinese declarative:

  ... mai3 lu4.

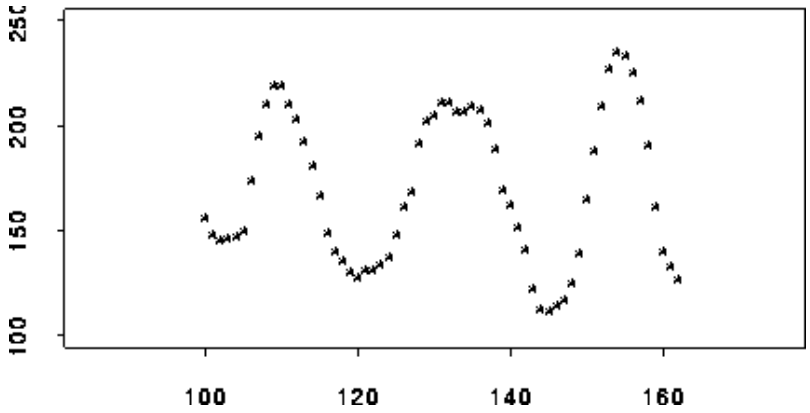
  ... buy deer.



- Chinese question:

  ... mai3 lu4?

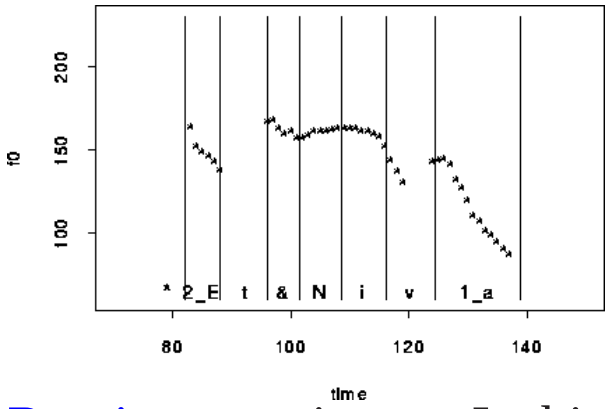
  ... buy deer?



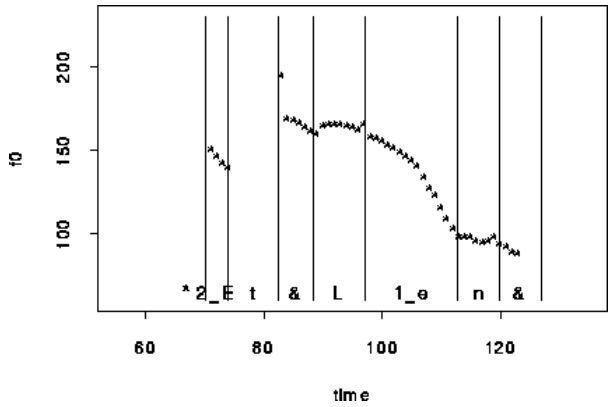
Russian Intonation Type

In questions, pitch rises on the last stressed syllable and falls afterwards.

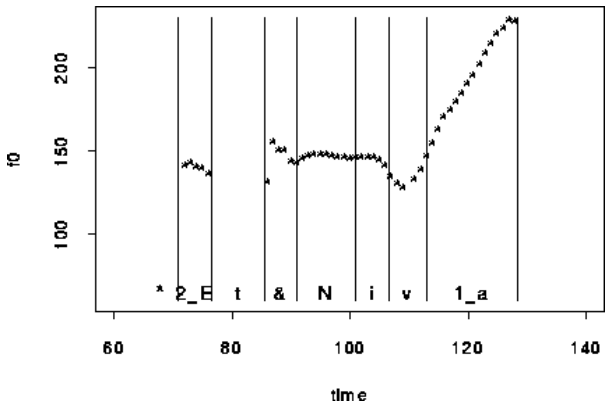
Russian declarative: This is Nevá.



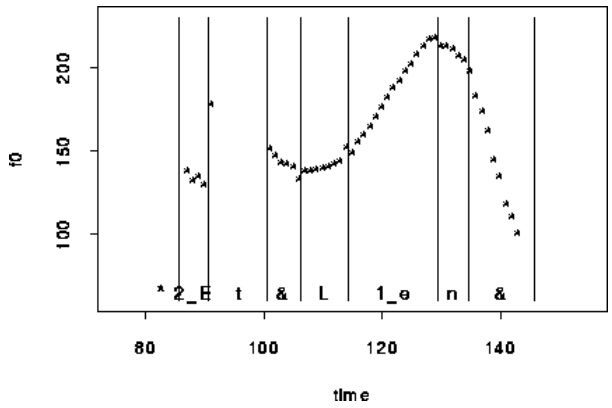
This is Léna.



Russian question: Is this Nevá?



Is this Léna?



Prosody Marks Discourse Functions

- Topic initialization (Hirschberg and Pierrehumbert 1986)
- Discourse structure
- Phrasing
- Emphasis
- New vs. old information: Emphasize new information, de-accent old information
- Other communicative means

Prosody of Emotion

Different emotions, such as excitement, anger, suspicion, fear, sad, sarcasm, have characteristic prosodic patterns.

Features: duration, f_0 mean, f_0 range, loudness, jitter, spectral tilt, accent shape

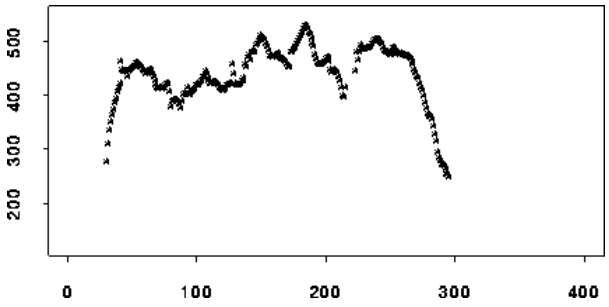
- Excitement: fast, very high pitch, loud
- Hot anger: fast, high pitch, strong, falling accent, loud
- Fear: jitter
- Sarcasm: prolonged accent, late peak
- Sad: slow, low pitch

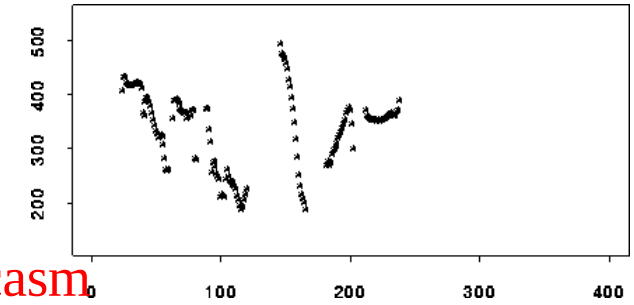
Example of Emotional Speech

Marylyn won 9 million dollars

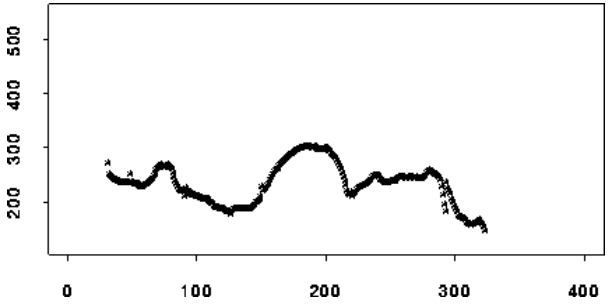
Dirty rats are the best, aren't they?

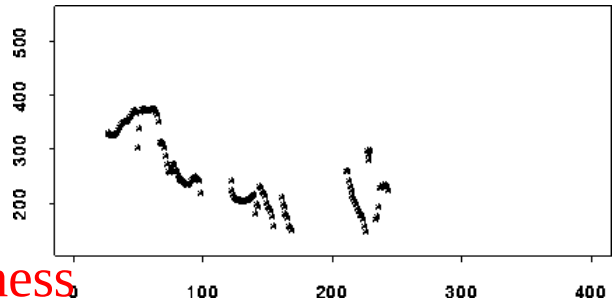
Excitement



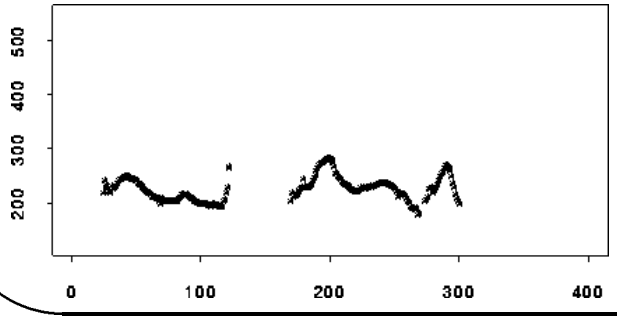


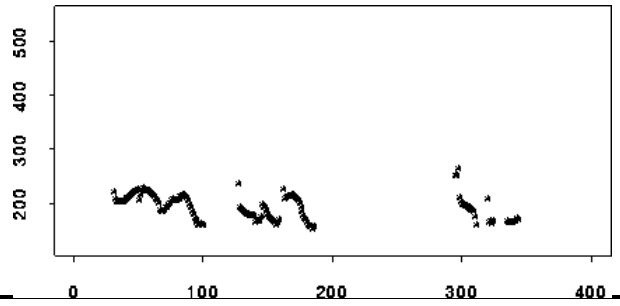
Sarcasm





Sadness



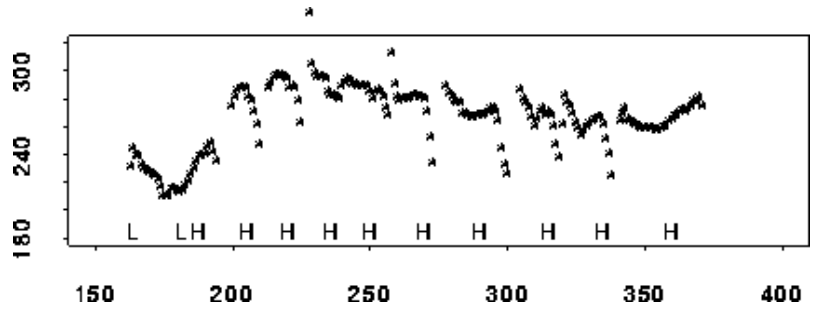


Physiological Effects on Prosody

- Declination: Gradual decline of pitch over the course of an utterance
- Perception test (Pierrehumbert 1979)

  Given two accented words of the same pitch height, the second one is perceived as more prominent— listener compensates for the declination effect

- Example (Shih 2000)



Summary

- Prosody encodes different information simultaneously
- Different signals, scopes and locations
  - lexical tones vs. pitch range
  - local effect vs. global effect
  - strong beginning vs. strong ending
- Modeling prosody requires understanding of individual components

Intonation Schools

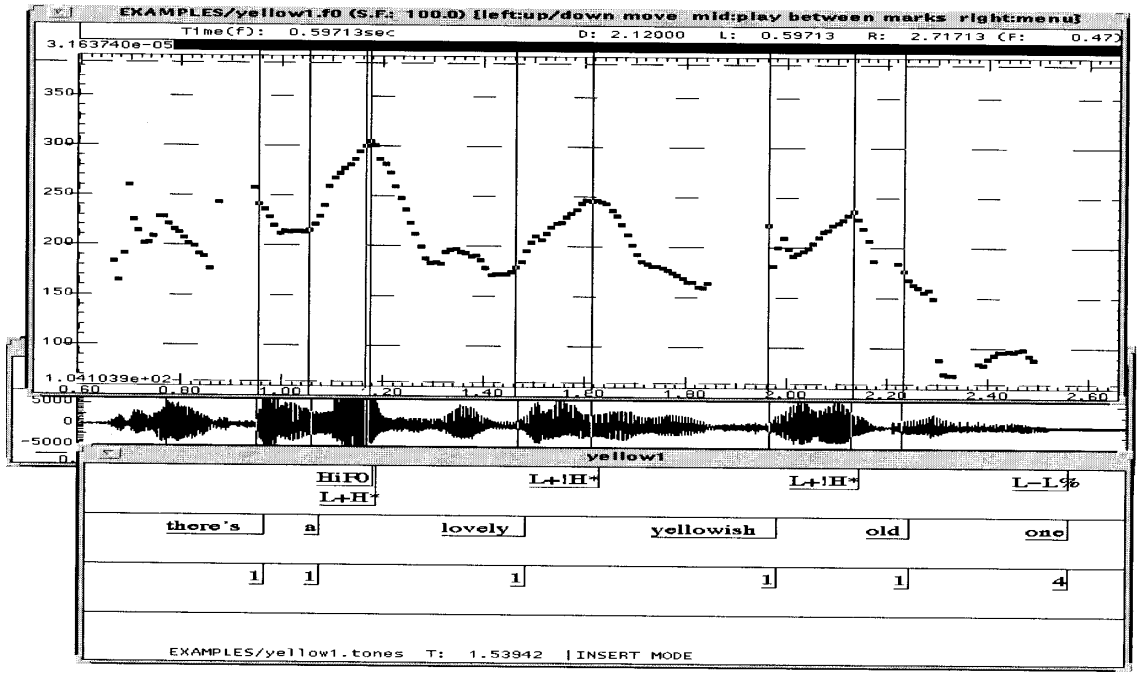
	Under-specified			Fully Specified
Single Component	INTSINT	ToBI, Xu	IPO, Tilt	Olive, Machine Learning
Two Component	Gronnum		Fujisaki	
Multiple Component				Van Santen

ToBI

Data from Mary Beckman

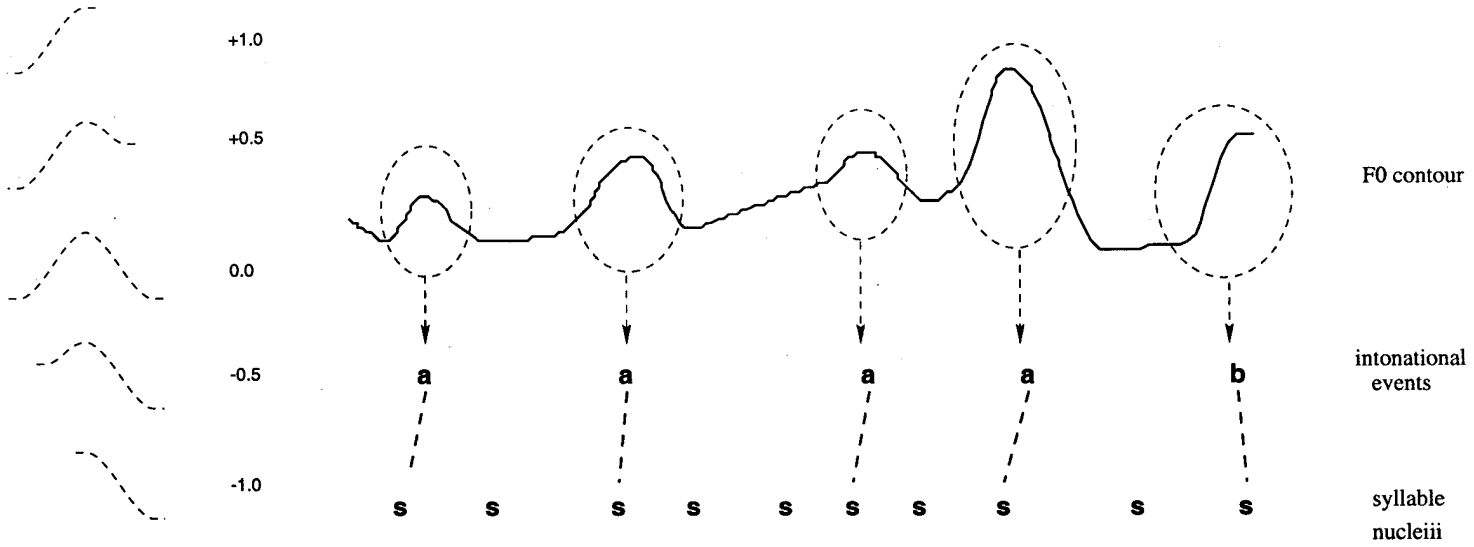
http://ling.ohio-state.edu/ tobi/ame_tobi/annotation_conventions.html

		L+H*	L+H*	L+H*	L- L%	
There's	a	lovely	yellowish	old	one.	



Tilt

Paul Taylor (2000). Analysis and Synthesis of Intonation using the Tilt Model JASA 107, 1697-1714.



Fujisaki's Model

Hiroya Fujisaki (1993). From Information to Intonation. Print from *Lecture at Laboratorio de Investigaciones Sensoriales*, Buenos Aires.

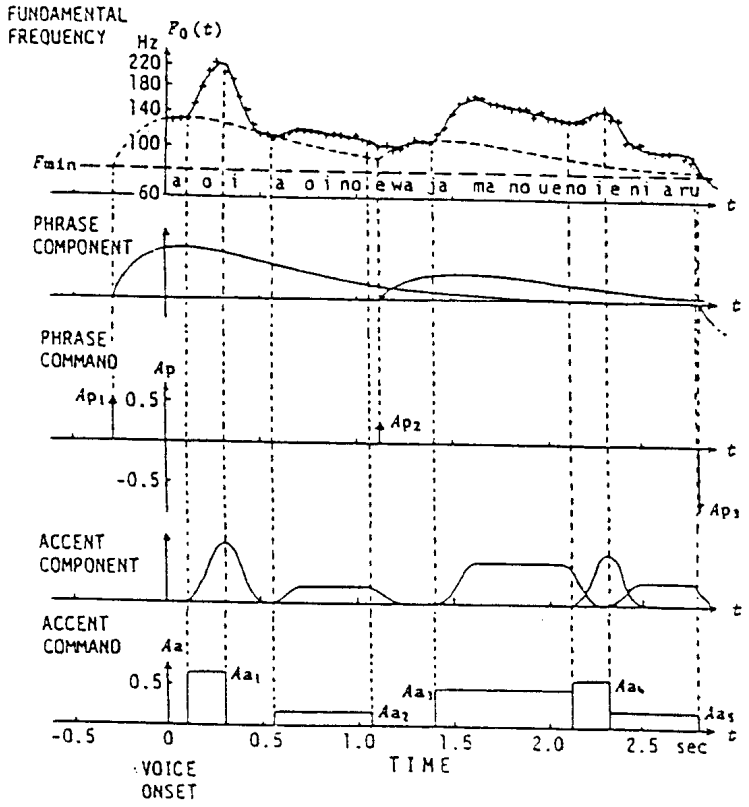


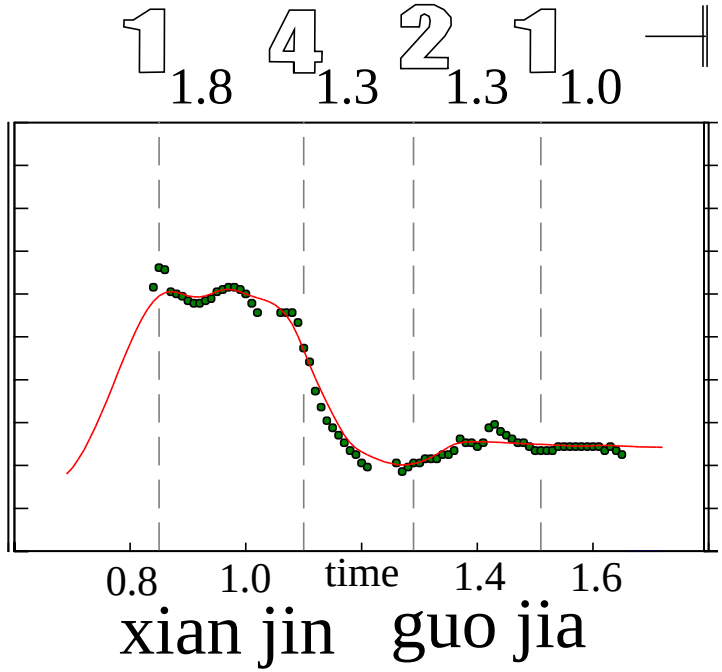
Fig. 4. Analysis-by-Synthesis of an F_0 contour of the Japanese declarative sentence: /aoiaoinoewajamanouenoieniaru/. The optimum decomposition of a given F_0 contour into the phrase and accent components is illustrated, and the underlying commands for these components are shown.

Section 2: Prosodic Models

- Intonation schools
- Tonal distortion data
- Stem-ML (Soft Template Markup Language)

Between Phonology and Phonetics

- The work we present here is a link between intonation phonology (abstract categories) and phonetics (surface f_0 contours).
- An example from Mandarin: Generating f_0 contours using lexical tone categories and their prosodic strengths.



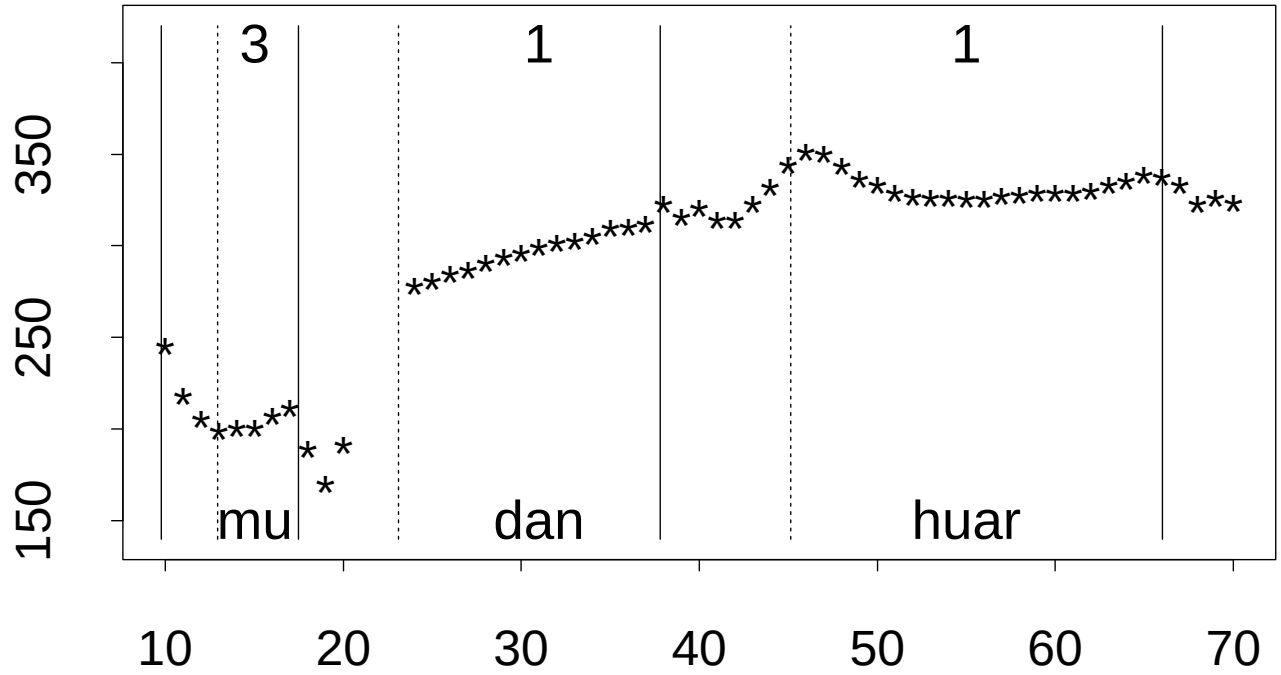
The Prosodic Model

- The prosodic modeling is based on Stem-ML (Soft Template Mark-up Language).
- Stem-ML consists of a set of mathematically defined tags with value attributes (tag_{value}), allowing user-defined accent shapes, phrase curves and other speaker specific parameters.
- Generation (tag $\rightarrow$ f_0):

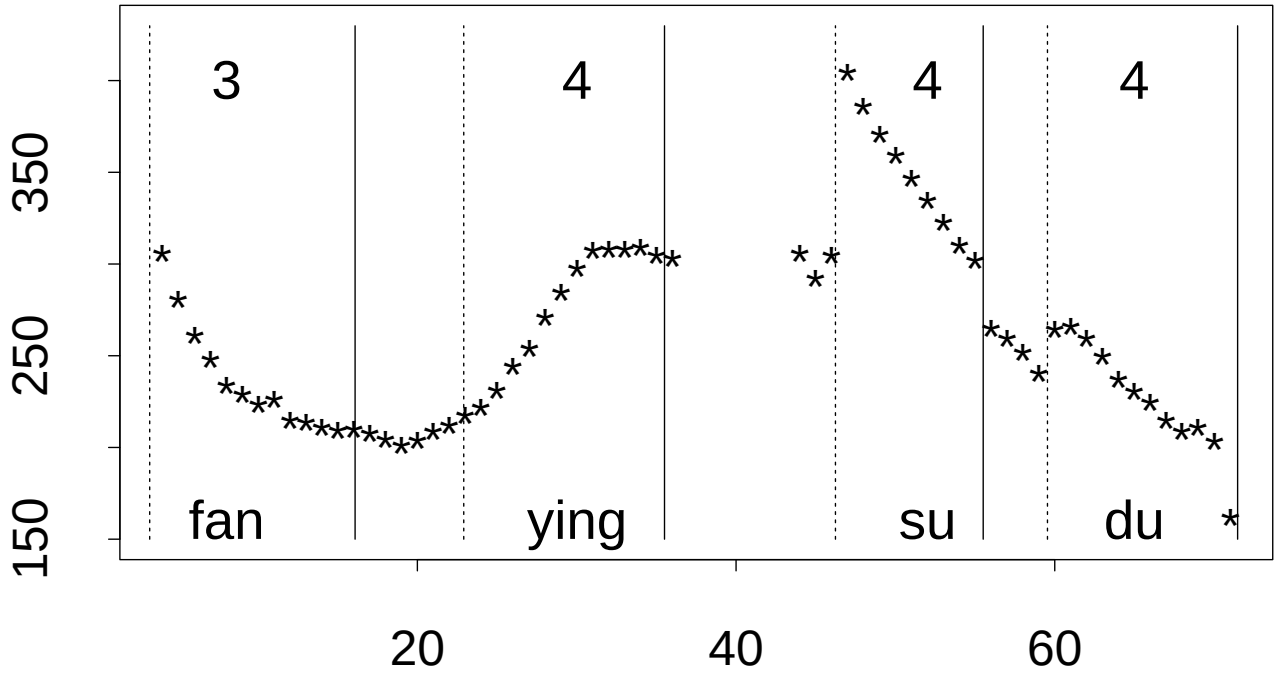
  Stem-ML calculates an intonation contour from a set of linguistically motivated tags that describe the intonation.
- Learning (f_0 $\rightarrow$ tag value):

  Given f_0 contour and abstract tags, the Stem-ML optimizer finds the optimal tag values that fit the contour.

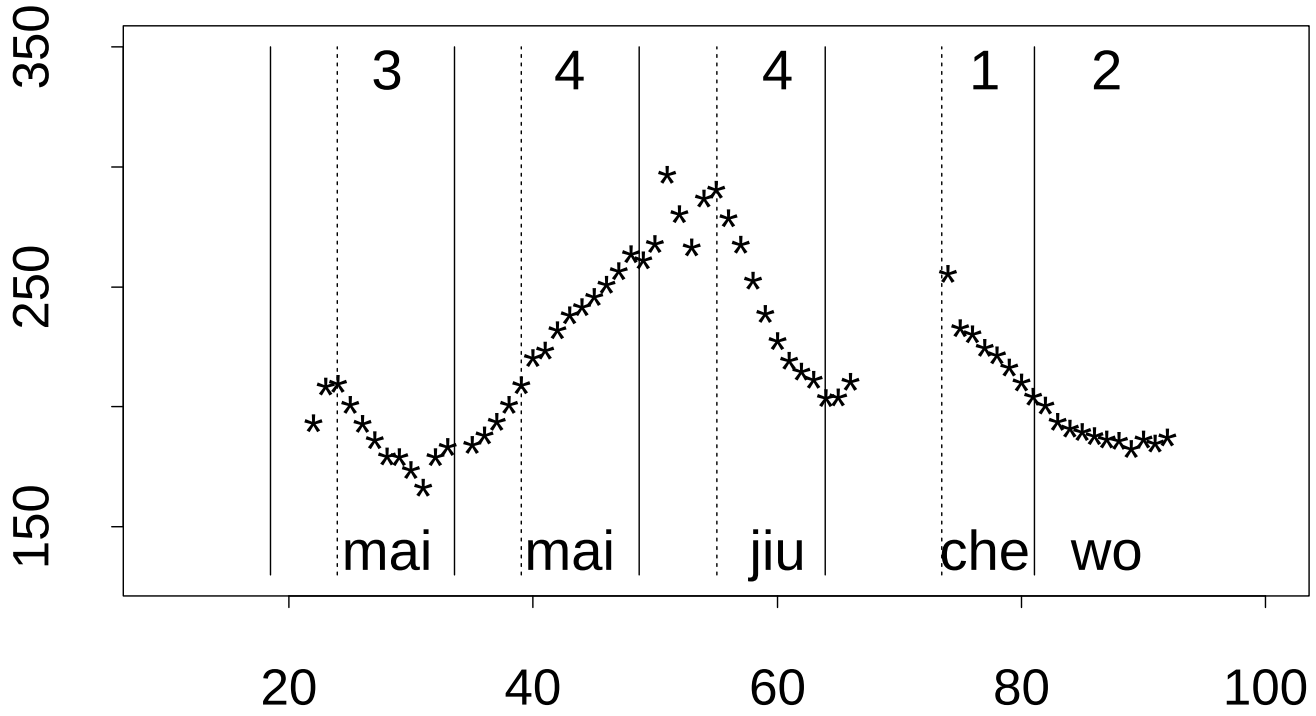
Tonal Distortion–Example 1



Tonal Distortion–Example 2



Tonal Distortion–Example 3

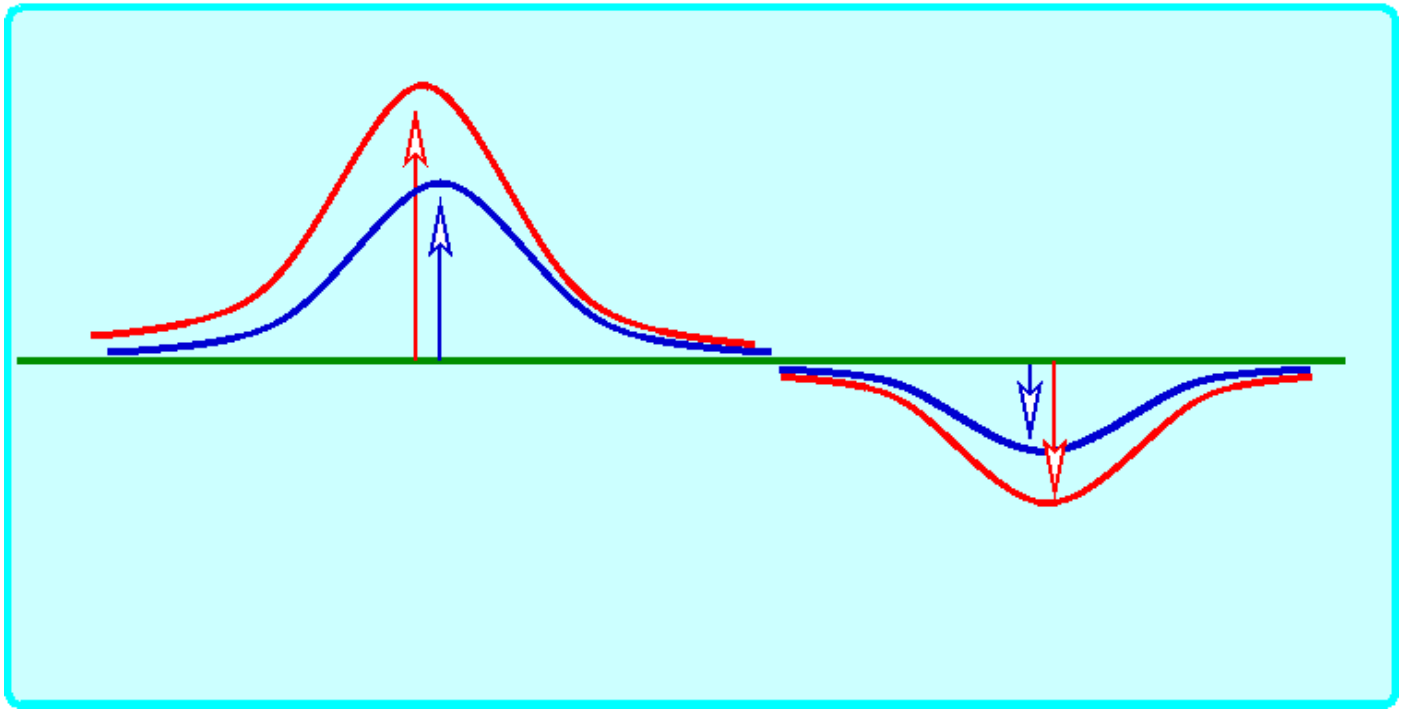


Representation of Prosodic Strength–I

ToBI Assumption

Strong: Further away from the reference line

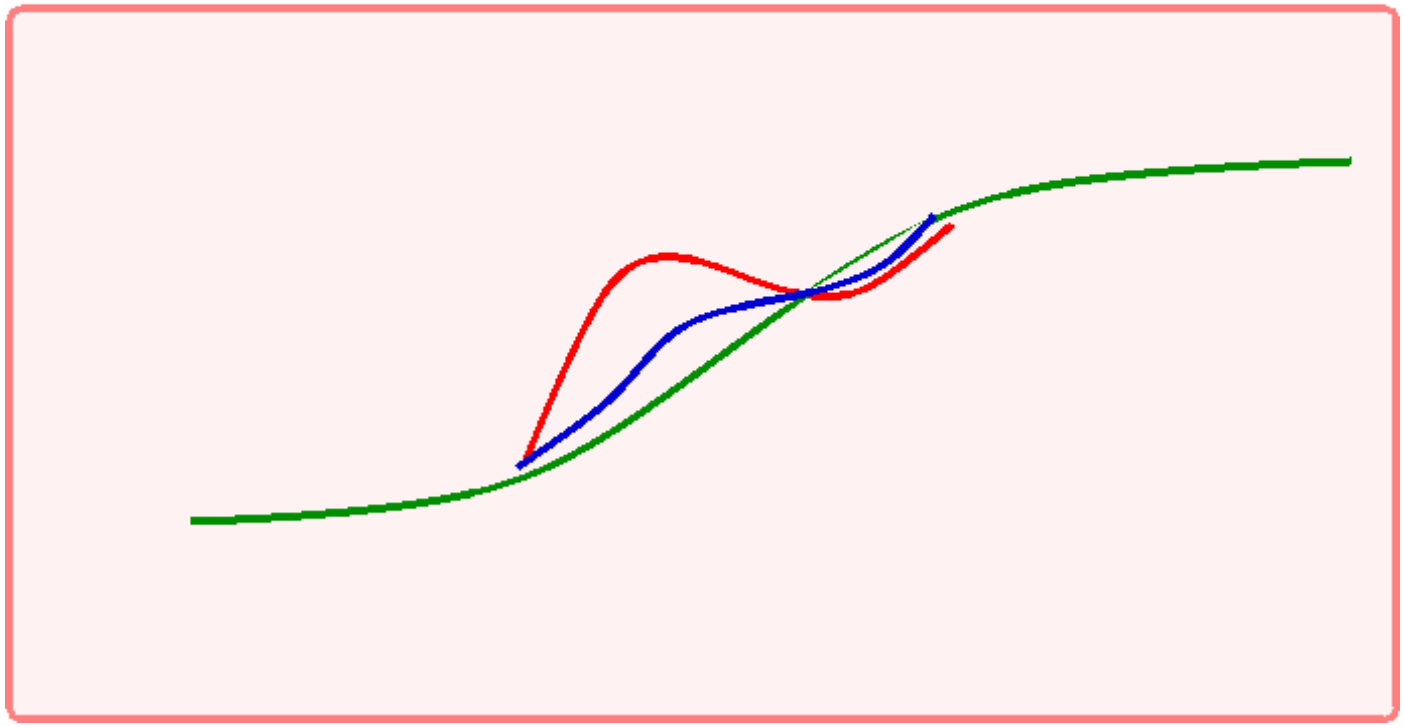
Weak: Closer to the reference line

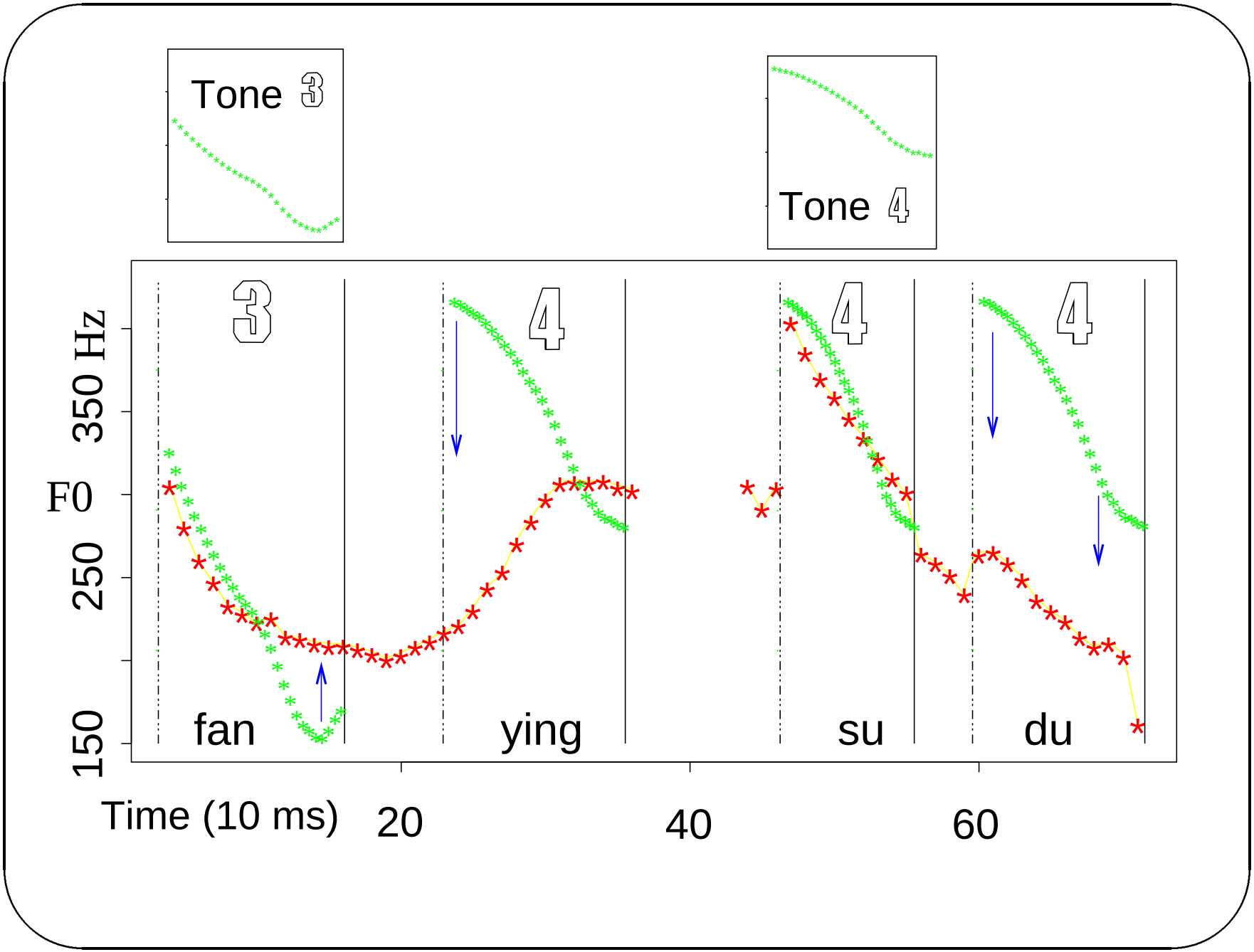


Representation of Prosodic Strength–II

Stem–ML Assumption

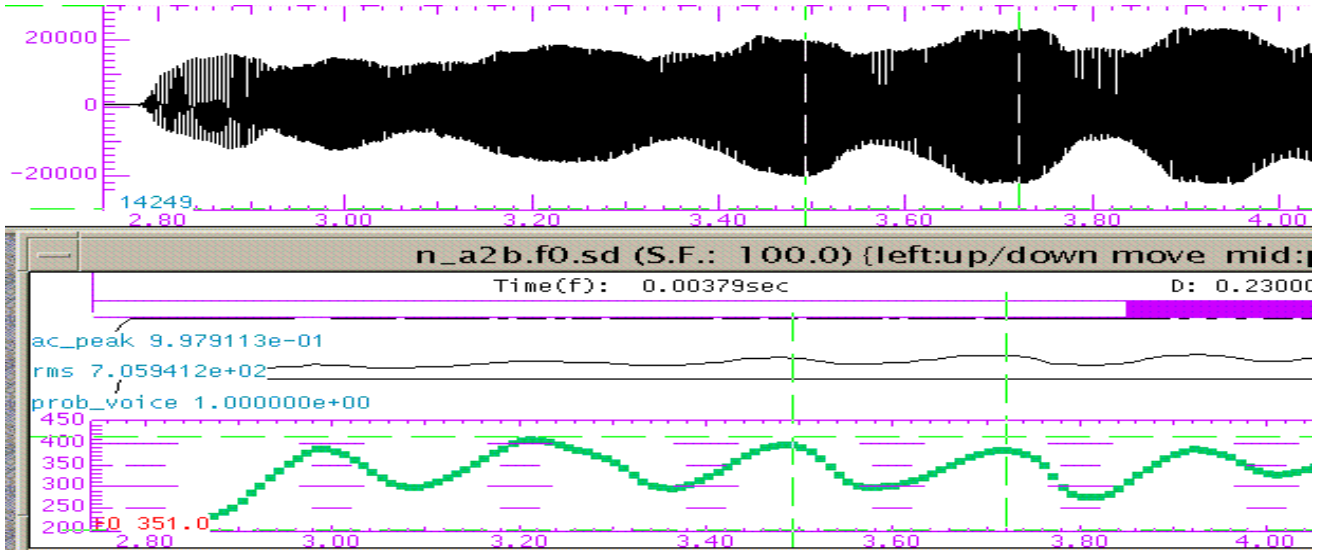
Strong: Maintain strong identity
Weak: Blend in into the environment



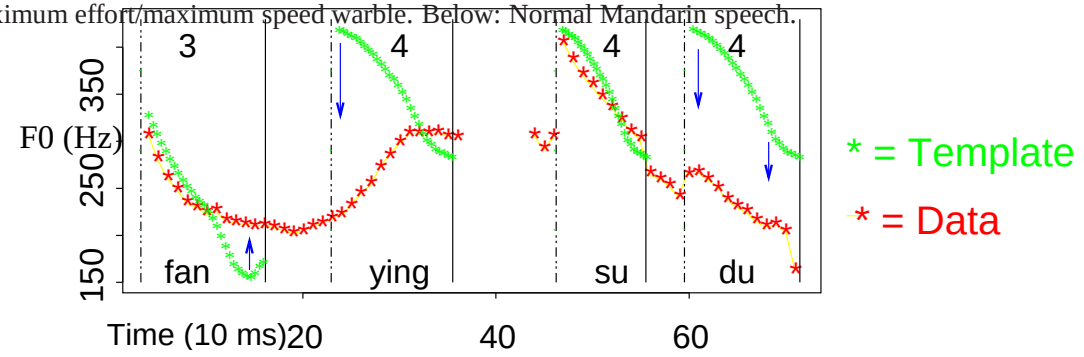

Tone 3
Tone 4
3
4
4
4
F0
350 Hz
250
150
fan
ying
su
du
Time (10 ms)
20
40
60

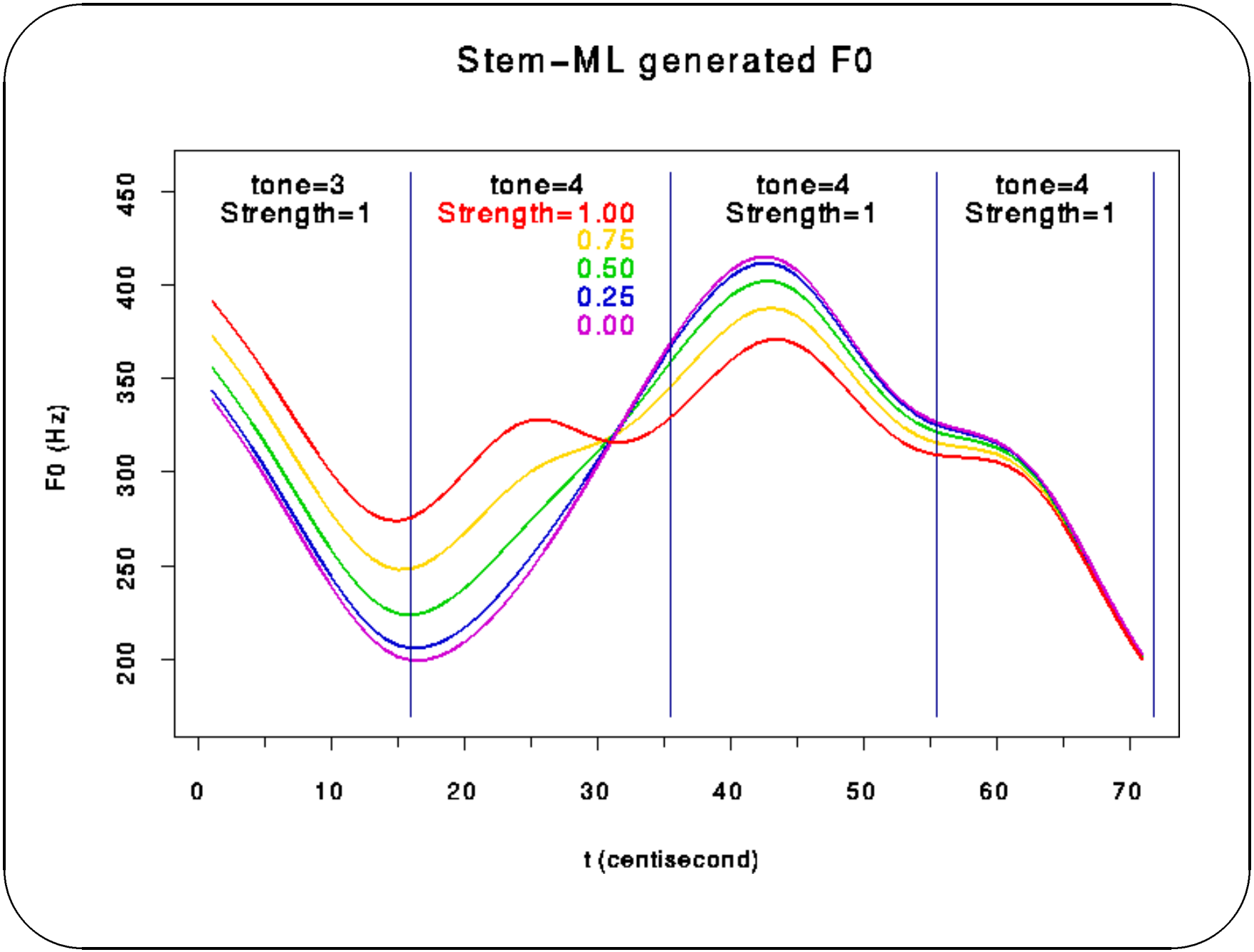
Physical Limits

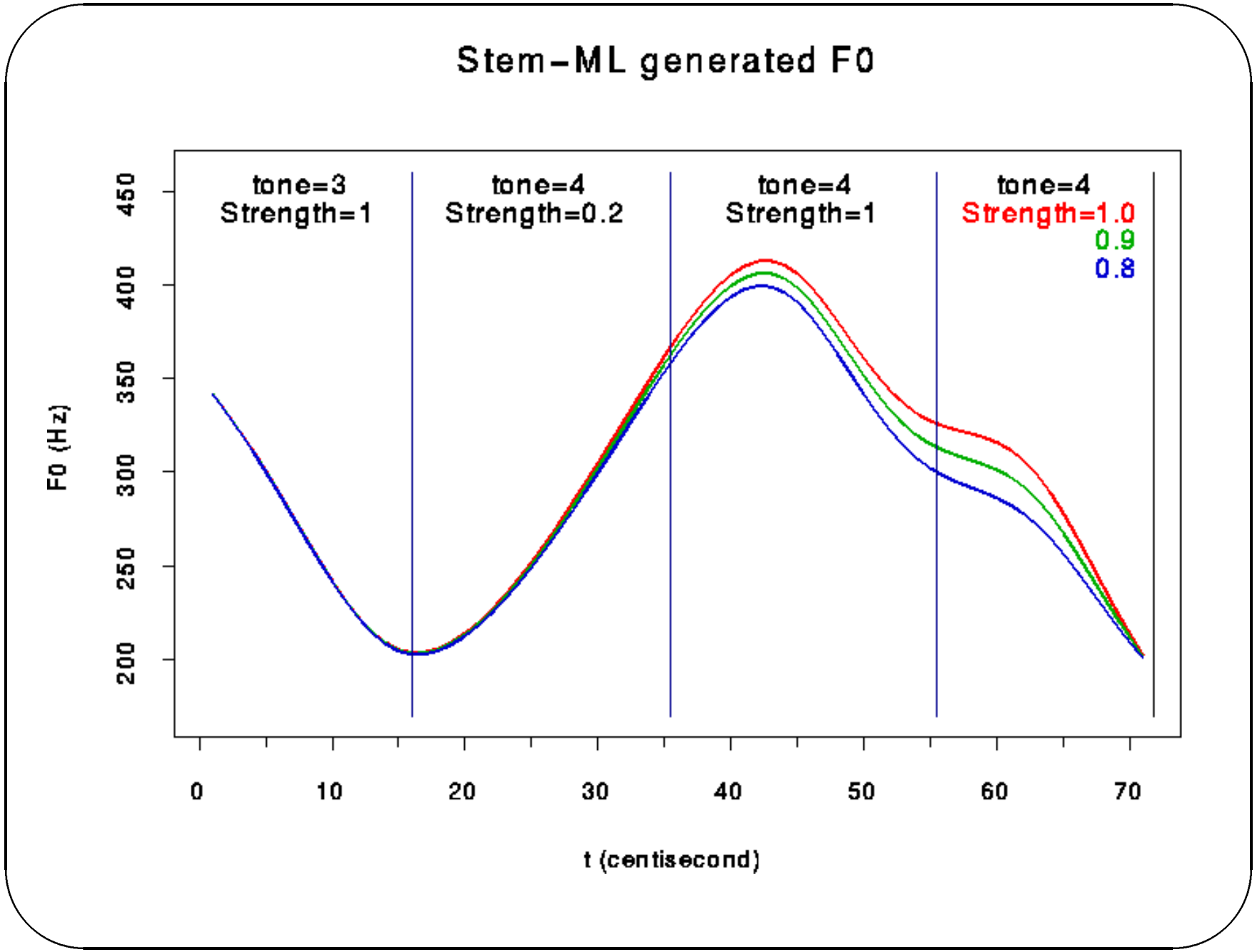
People talk nearly as fast as possible

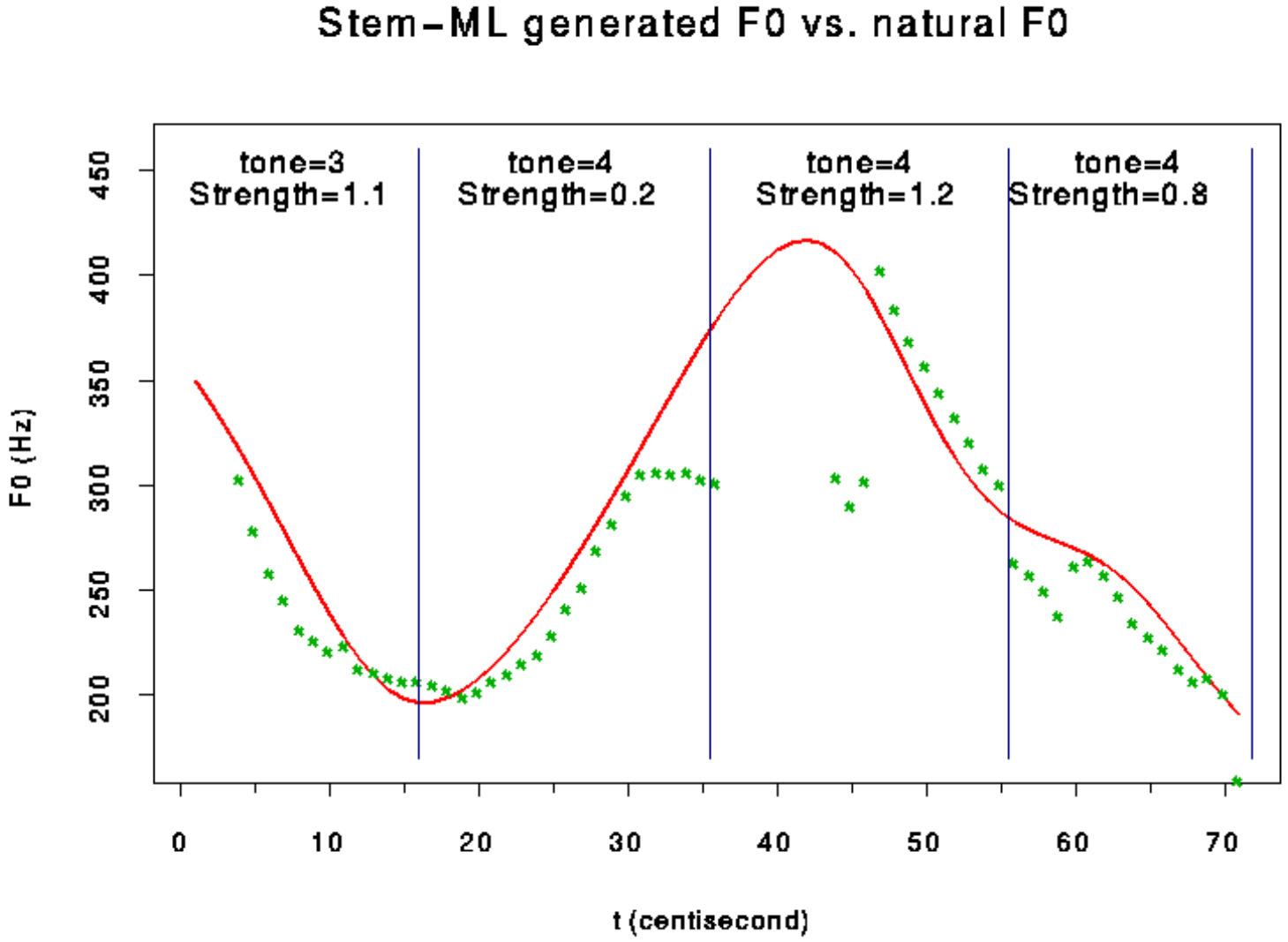


Above: maximum effort/maximum speed warble. Below: Normal Mandarin speech.




Stem-ML generated F0
tone=3
Strength=1
tone=4
Strength=1.00
0.75
0.50
0.25
0.00
tone=4
Strength=1
tone=4
Strength=1
F0 (Hz)
200
250
300
350
400
450
0
10
20
30
40
50
60
70
t (centisecond)


Stem-ML generated F0
tone=3
Strength=1
tone=4
Strength=0.2
tone=4
Strength=1
tone=4
Strength=1.0
0.9
0.8
F0 (Hz)
200
250
300
350
400
450
0
10
20
30
40
50
60
70
t (centisecond)


Stem-ML generated F0 vs. natural F0
tone=3
Strength=1.1
tone=4
Strength=0.2
tone=4
Strength=1.2
tone=4
Strength=0.8
F0 (Hz)
200
250
300
350
400
450
0
10
20
30
40
50
60
70
t (centisecond)

Section 3: A Modeling Example

English Intonation of Asking for Confirmation

The Problem

- Build a model of the intonation used to confirm a word in a question
- The examples are in the form 123-456-7890
- The speaker tries to confirmation (in this example) the digit 6

Why this is an Interesting Problem

- This intonation is often used in dialogue systems, confirming credit card numbers, telephone numbers, and others
- This intonation type involves the interaction of question and emphasis
- The intonation type generalizes to many other domains
- Many intonation theories cannot handle the basic phenomenon

Experiment–Data

200 digit sequences in 16 blocks

Variations in phrasing, speaking speed, and position

301-123-5045	301-123-5045
301-123-5045	301-123-5045
301-123-5045	301-123-5045
301-123-5045	301-123-5045
301-123-5045	301-123-5045

Reading Instructions

- Delarative intonation Instruction: This is your phone number. You are giving other people this information over the phone.

  Text presentation: 901-109-9091.

- Yes-no question Instruction: You are repeating a phone number back, asking whether you've got it right.

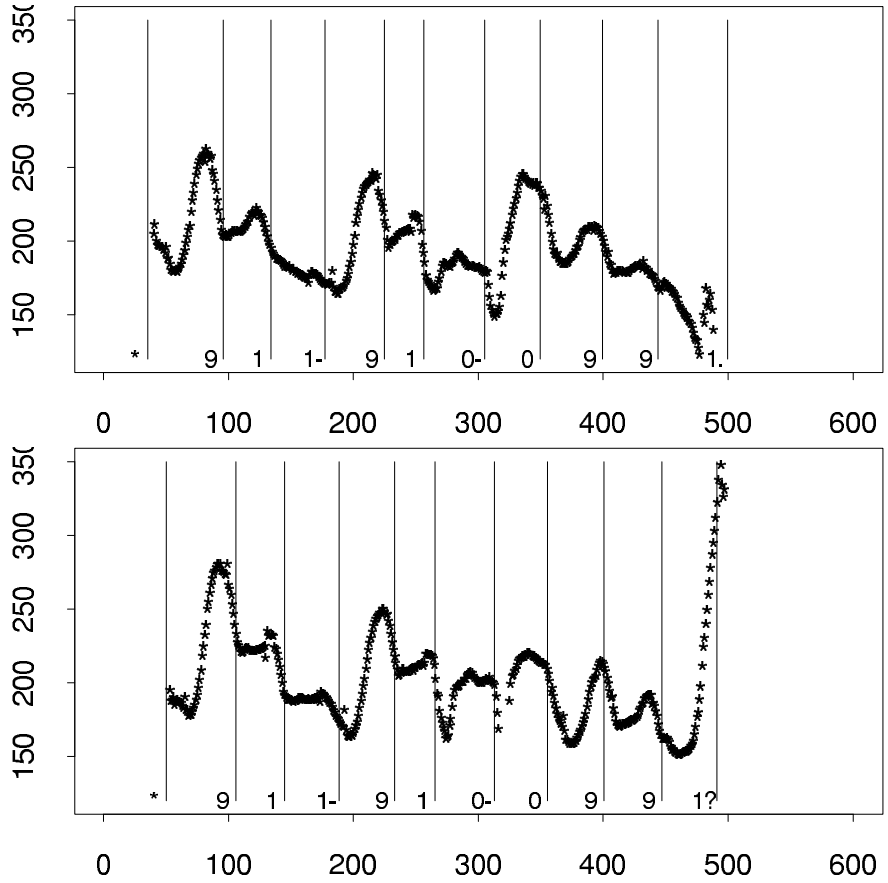
  Text presentation: 901-109-9091?

- Confirmation Instruction: You know you've got most of the numbers but are not sure about the one underlined in red. You are trying to confirm whether this digit is correct.

  Text presentation 901-109-9091?

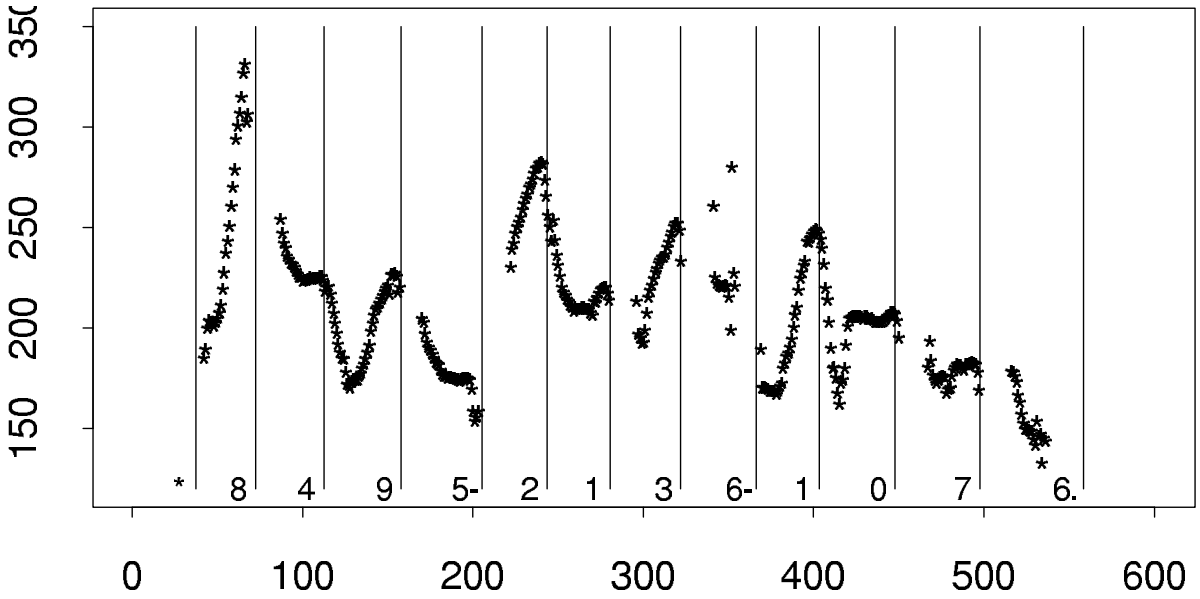
Observations–1

Not surprisingly, declarative sentences end with falling pitch and questions end with rising pitch. This is a consistent difference between yes-no question and declarative sentence.



Observations–2

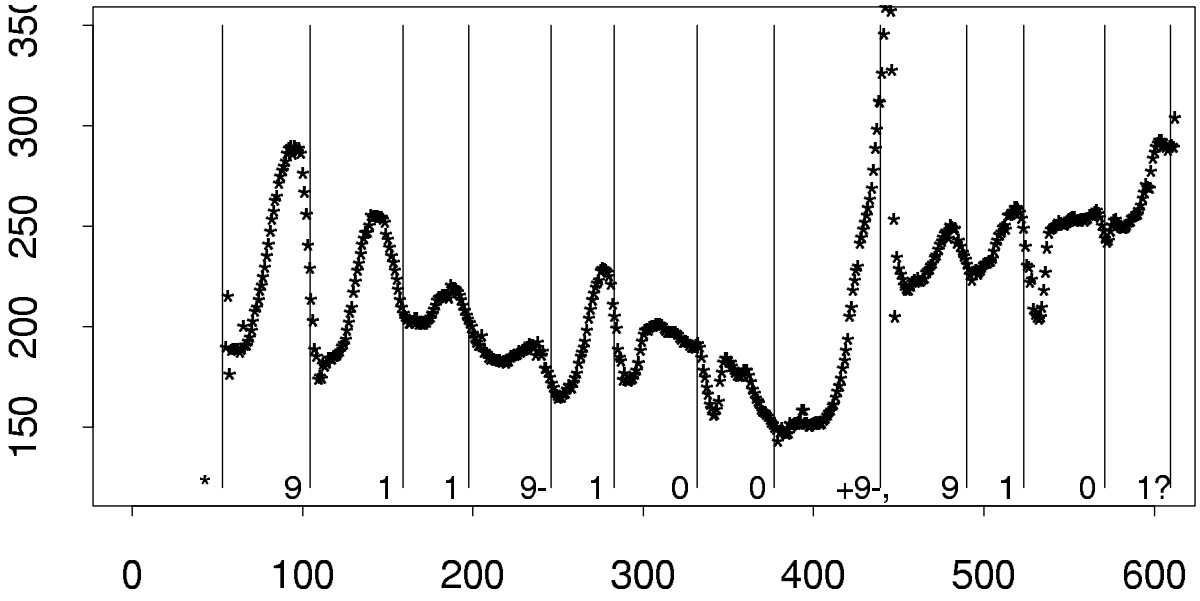
Phrasing as indicated by the dash is clearly marked in declarative sentences. Pitch rises on the phrase initial digit.



Observations–3

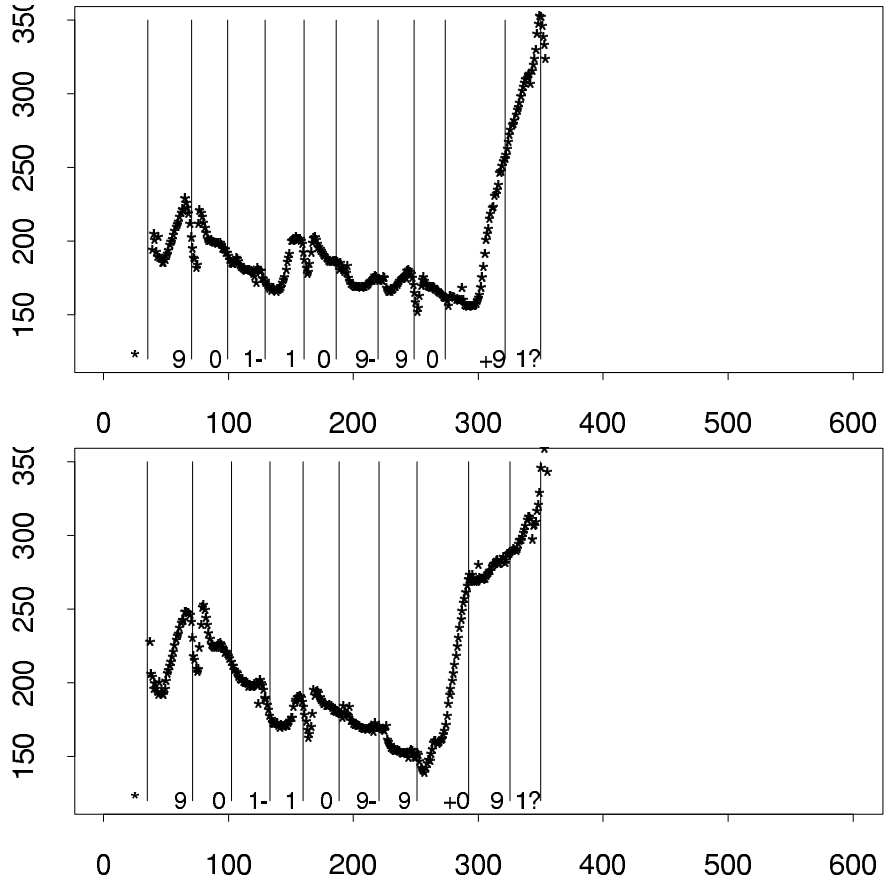
Digit confirmation is marked with a strong rise and longer duration on the digit being confirmed. Post-confirmation pitch remain high.
Pre-confirmation phrasing is similar to that of declarative and yes-no question sentences.

ToBI: L*+H H- H% (??)



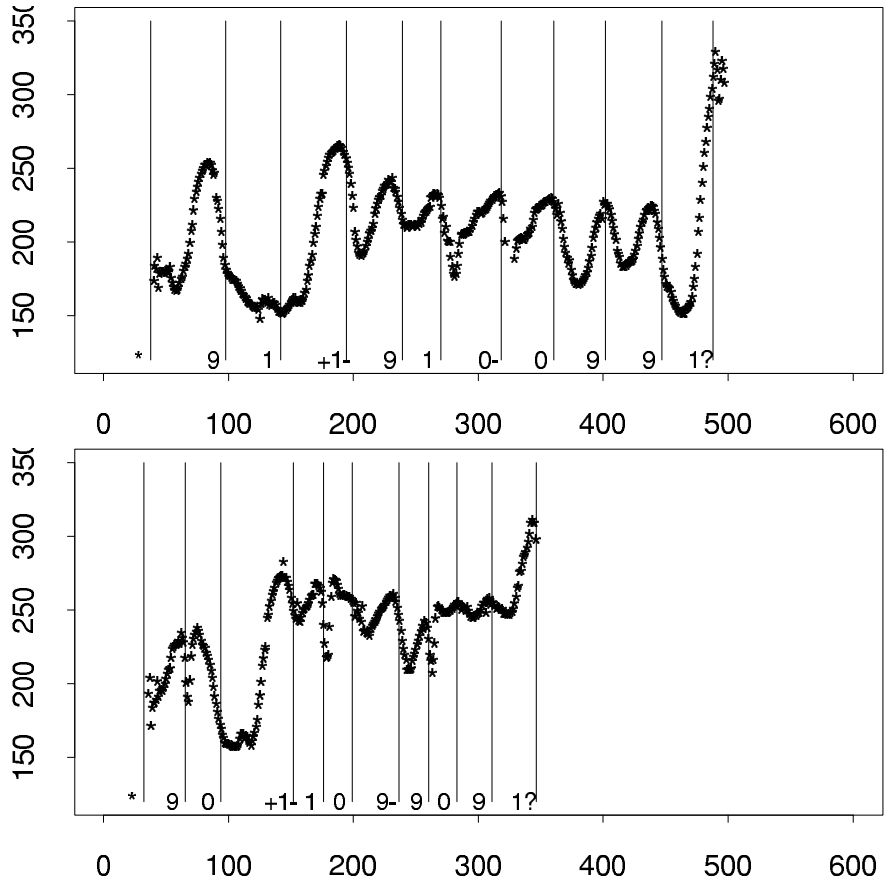
Observations–4

There is another final rise in the confirmation sentences. But when the confirmed number is very close to the end, the confirmation rise and the final rise fuse together.



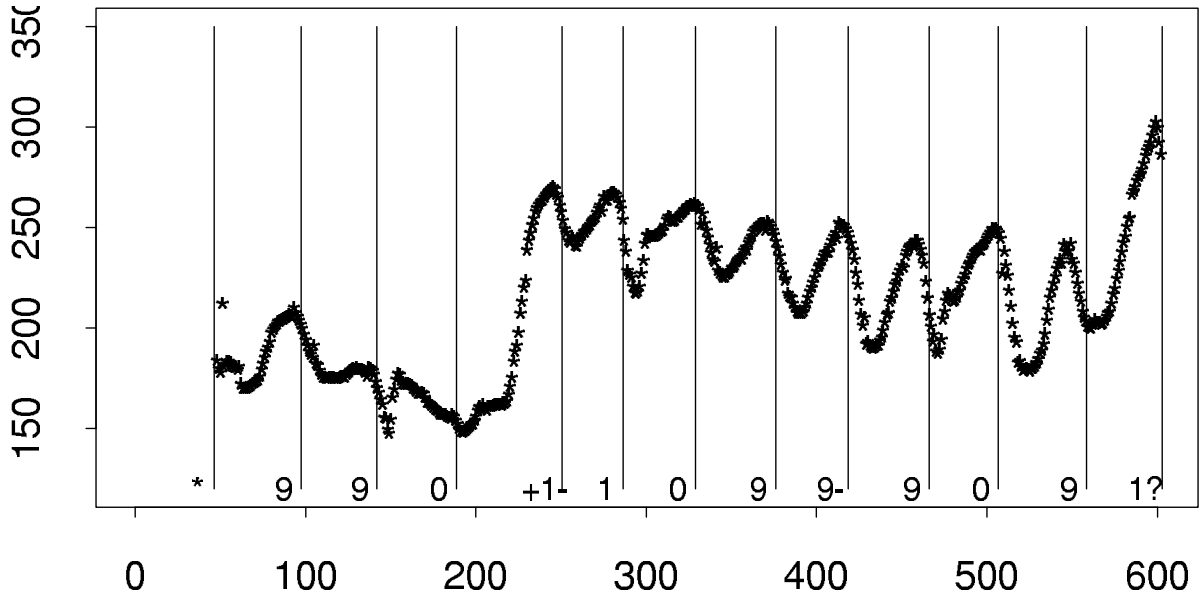
Observations–5

Post confirmation pitch tends to be flatter in fast speech than in slow speech.



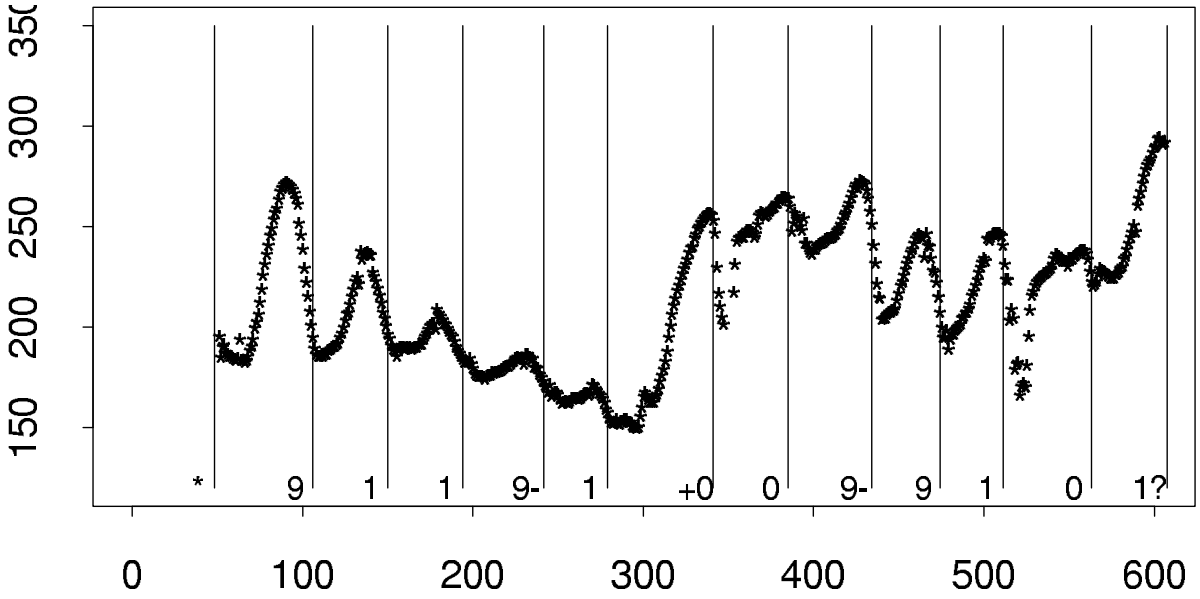
Observations–6

Post confirmation accent returns after a while, this is especially clear in slow speech.



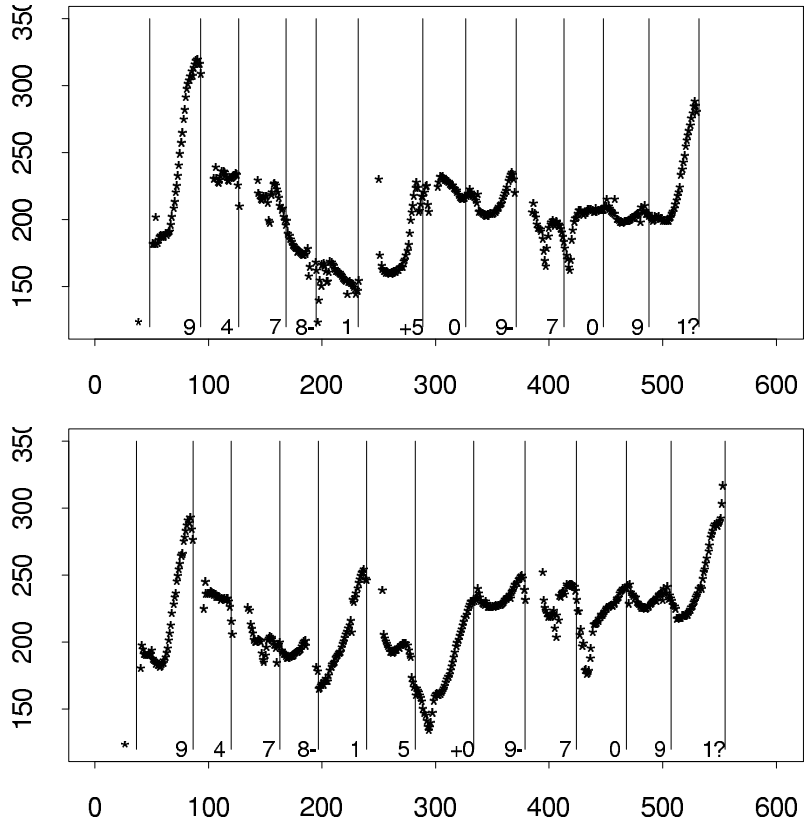
Observations–7

Post-confirmation phrasing is less obvious. However, when the phrasing structure is observable, new phrases are marked by pitch drop, in contrast to the pitch rise in declarative sentences.



Observations–8

The digit immediately before the confirmed digit tends to get de-accented, even when this digit starts a new phrase, where it would normally be marked with phrase initial high pitch.



The Model–Parameters

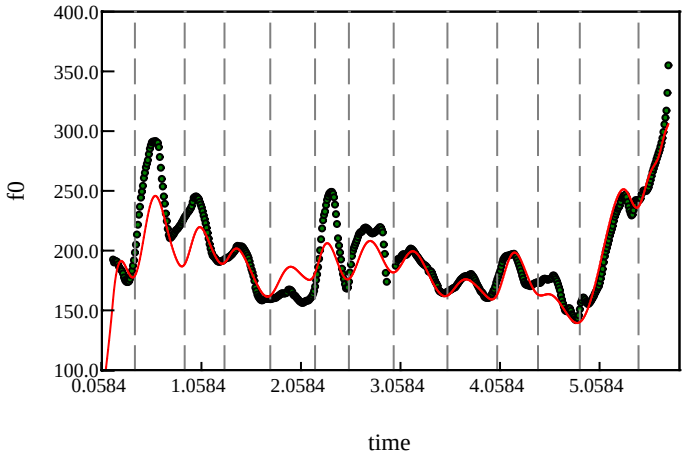
- Fit a subset of the data with sonorant digits
- 90% of the sentences used in the training set

  10% of the sentences are reserved as test set

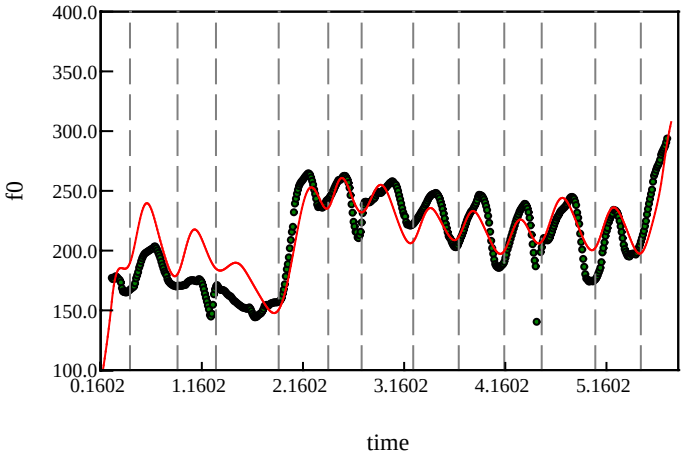
  8 runs
- 48 parameters for 39 sentences. Average 1.2 parameters per sentence
- All are global parameters which are shared among all sentences
- No parameters for speed or sentence-specific variation

Results–Slow and fast speech

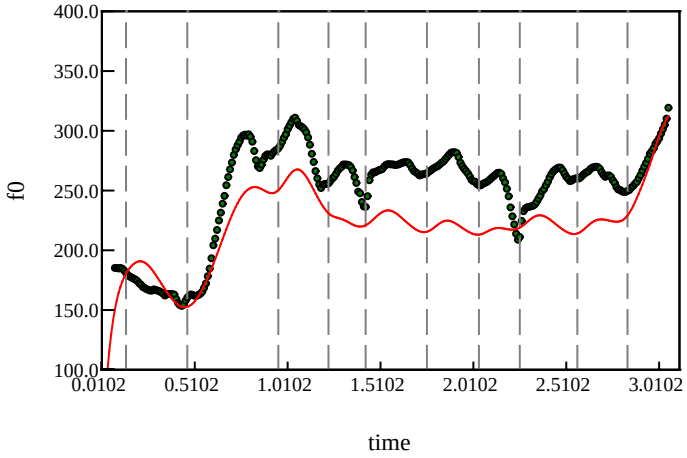
data/Q0031.f0.2c



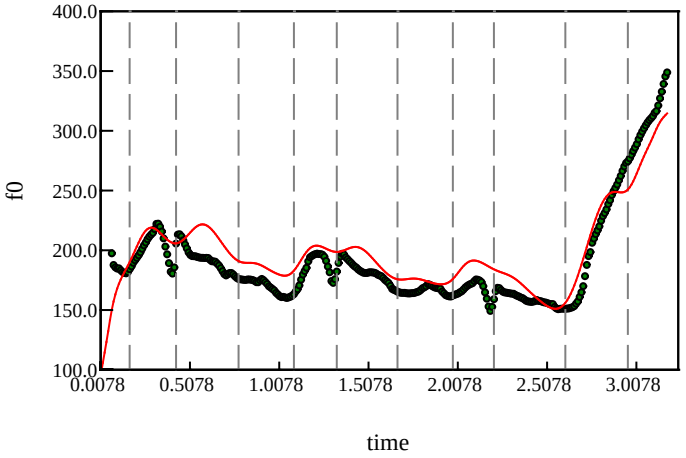
data/Q0067.f0.2c



data/Q0074.f0.2c



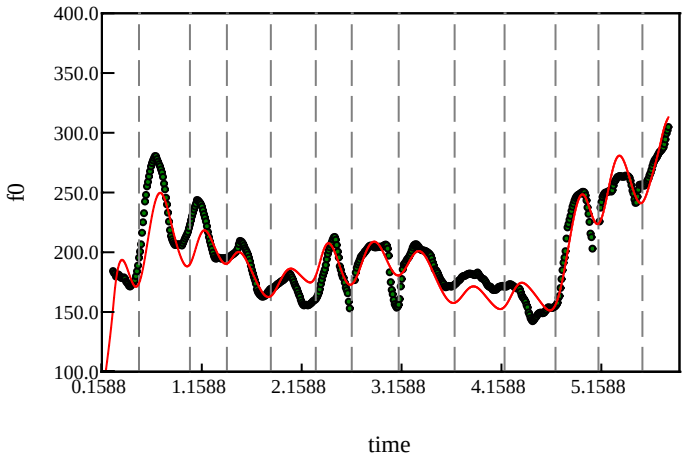
data/Q0079.f0.2c



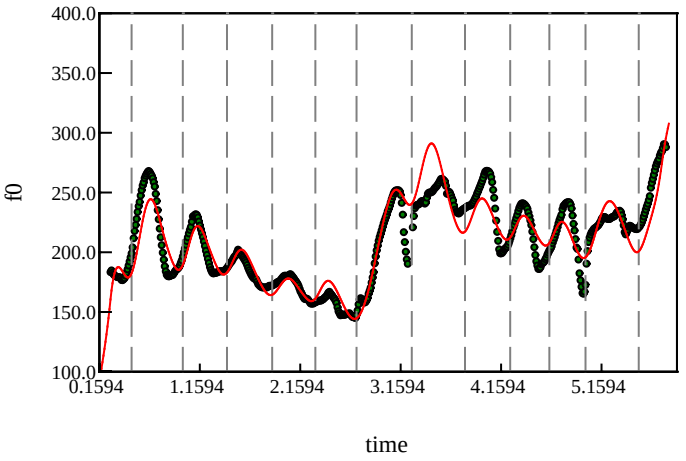
Results–De-accenting before emphasis

RMS deviation 0.212 Barks/21 Hz/1.7 semitones

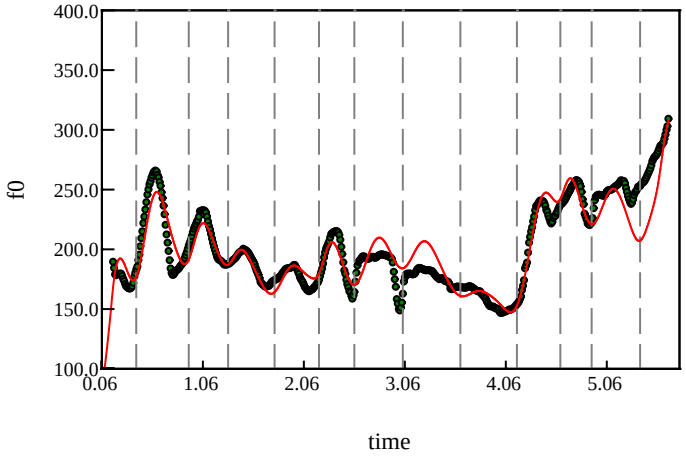
data/Q0034.f0.2c



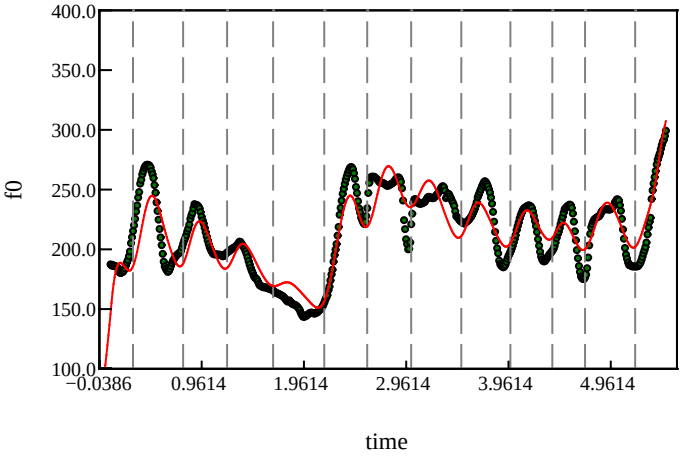
data/Q0039.f0.2c



data/Q0040.f0.2c



data/Q0041.f0.2c



Two basic channels: lexical and prosody

languages have two different ways of transferring information languages have a lexical channel which consists of a sequence of discrete symbols and a prosodic channel which modifies the words

(Punctuation is the written expression of prosody.)

Languages have two different ways of transferring information. Languages have a lexical channel, which consists of a sequence of discrete symbols, and a prosodic channel, which modifies the words.

The Lexical Channel

When reading text, the *Lexical channel* is the sequence of written words.

The Prosodic Channel

What is it?

In 1950 a comedian named Stan Freberg conducted an elegant controlled experiment to find out if information could be carried over the prosodic channel *vs.* the lexical channel.

Lexically, there is essentially no information, just endless repetitions of "John Marsha."

[John&Marsha]

Prosodically, it tells a story, but it is peculiarly hard to put one's finger on the details.

Section 2: Prosodic Models - Science

What are the properties of a good model?

Syntax in a small window cannot predict prosody:

- **I** did not eat the melon. (She did, instead.)
- I did not **eat** the melon. (I fed it to the dog.)

Directing a play by Shakespeare.

The Prosodic Channel

Prosody is carried by several different properties:

- Phoneme Duration: Robot R1D1 has defective speech rhythm:
  - Playful: R1D1 picks a random duration between 10 and 400 milliseconds for each phone. [R1D1r]
  - Serious: R1D1 assigns the same duration to each phone. [R1D2s]
- Pitch: Robot R1P1 has defective pitch control:
  - Playful: He creates a random melody for each sentence. [R1P1r]
  - Serious: R1P1 speaks in a monotone. [R1P1s]

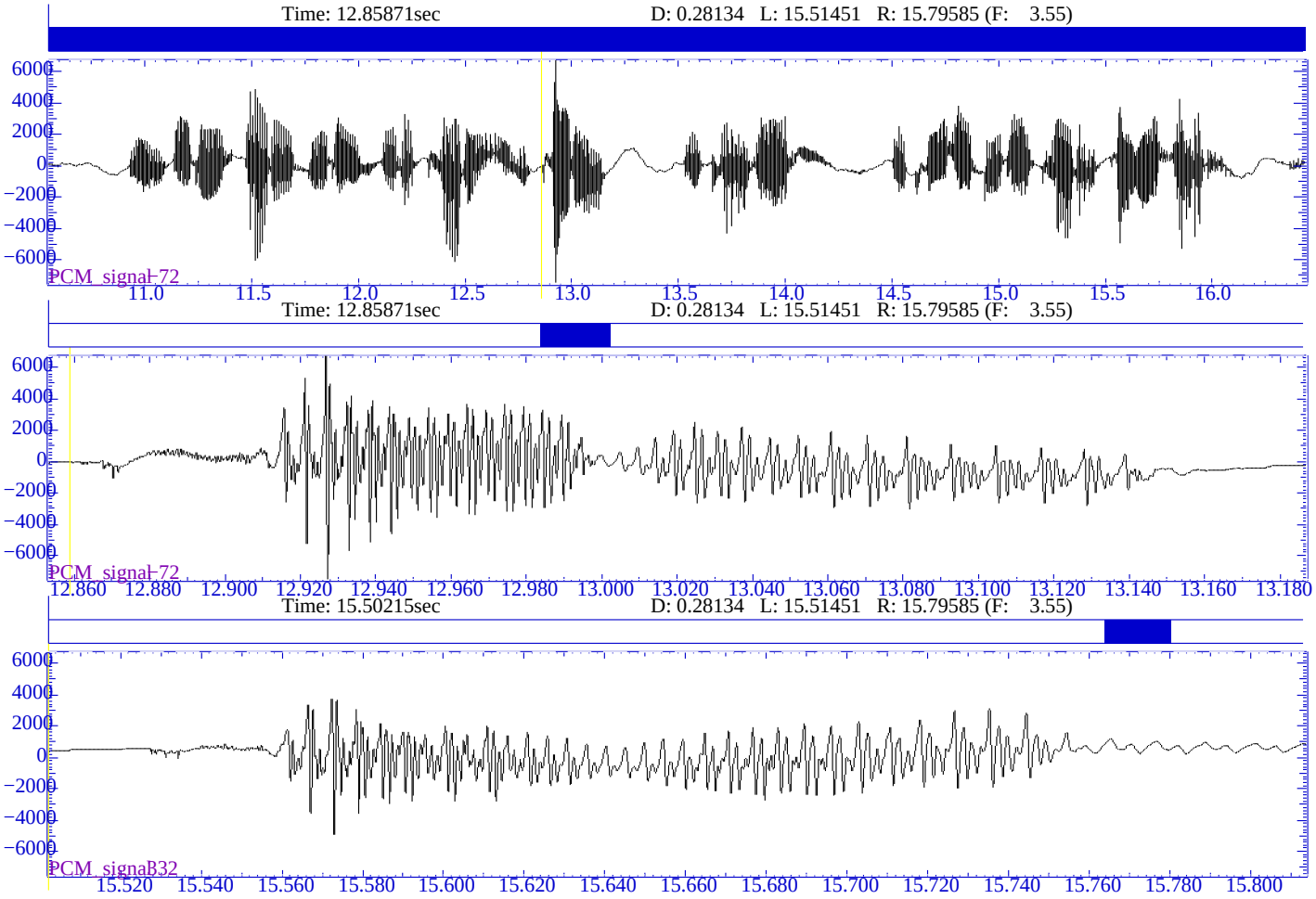
The Prosodic Channel

Prosody is carried by several different properties:

- Amplitude: Amplitude has strong social controls, because many people normally need to share the same acoustic channel. Temporary changes, though, have impact.
- Nonverbal prosody: facial expressions and gestures.

The Prosodic Channel

Prosody is visible in the acoustic data, and should be deducible from the speech. [PUDDLE]



Section 2: Prosodic Models

What are the properties of a good model?

- A model can be a scientific theory of:
  - How we communicate with each other.
  - What we communicate.
- A model can be an engineering approximation to be used in a
  - Speech synthesizer
  - Speech recognizer
  - Dialog system.

Section 2: Prosodic Models - Science

What are the properties of a good model?

- Quantitative and Falsifiable.
- Models need to connect to observables.
- Humans are Mechanical Systems.
- Unified prosody: similar mechanisms for many gestures.
- Prosody is about Communication.
- Separating Style and Substance.
- Similar speech implies similar parameters.
- A principled treatment of linguistic variation.

Section 2: Prosodic Models - Pragmatic

What are the properties of a good model?

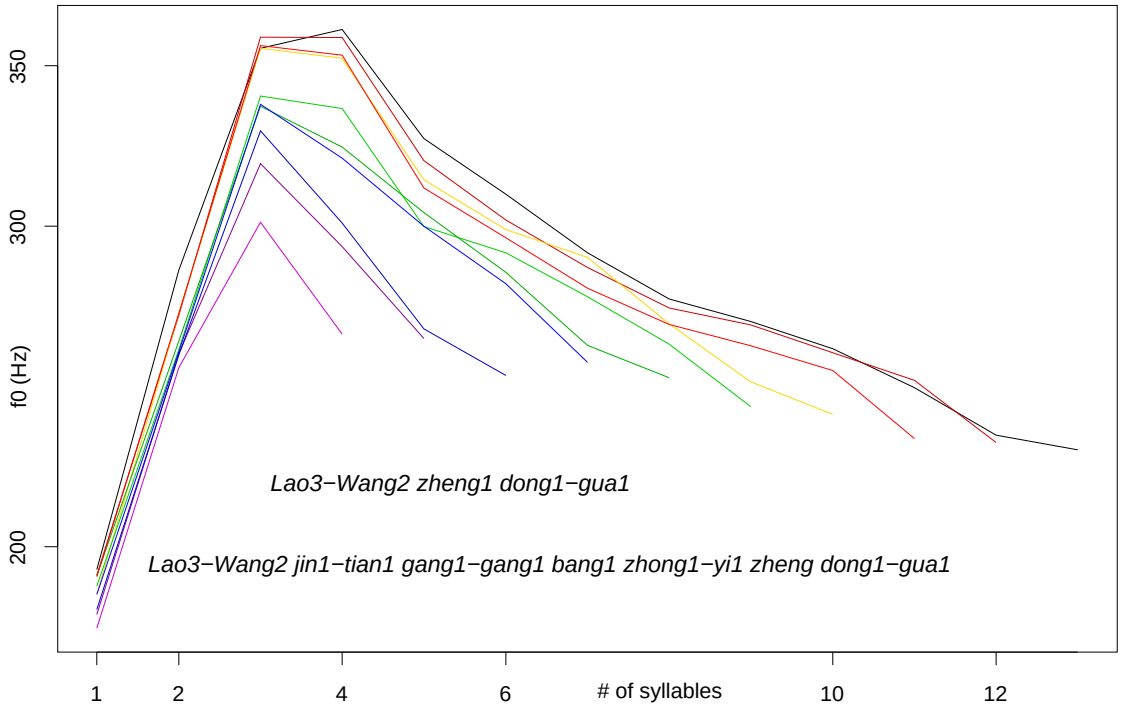
- Sensibly handles semantic mark-up.
- Covers all speech.
- The model is invertable.
- No Black Boxes.
- Relatively good predictability from text.

Section 2: Stem-ML basic concepts

- People plan utterances several syllables in advance.
- People produce speech optimized to meet their needs.
- A physical model for the muscles that control pitch.
- A linguistic strength for each word.

Techniques - Modeling Prosody

Speech is Planned:



People start at a higher pitch when they begin longer sentences.

Inhaled air volume is controlled by sentence length, too. Therefore, there is some plan 300 ms before the start of speech.

Techniques - Modeling Prosody

Intonation is Smoothly related to Muscle Dynamics

- All aspects of prosody are controlled by muscle actions.
- Relationship between muscle actions and percieved prosody is reasonably linear.
- Pitch is controlled by vocal fold tension and subglottal pressure.
- Register jumps are not frequent.
- Must subtract out or ignore segmental effects.

Techniques - Modeling Prosody

Speech is Optimal:

- Most of what we say is made from bits and pieces we've already practiced.
- Both Mandarin Chinese tones and ToBI English transcriptions have very few intonational symbols.
- A speaker has the opportunity to practice and optimize all the common 3-tone or perhaps 4-tone sequences.
- $\approx$ 1000 sequences to learn.
- Will practice each sequence every few hours.

Techniques - Modeling Prosody

Optimize what?

- People want to minimize the chance of being misunderstood.
  - The speaker's estimate of the probability that his/her idea will be misinterpreted.
- People want to minimize effort and/or talk faster.
  - cars
  - chairs
- How to combine the two? A weighted sum. (Each syllable can have a different weight, depending on its importance.)

Techniques - Modeling Prosody - The Math

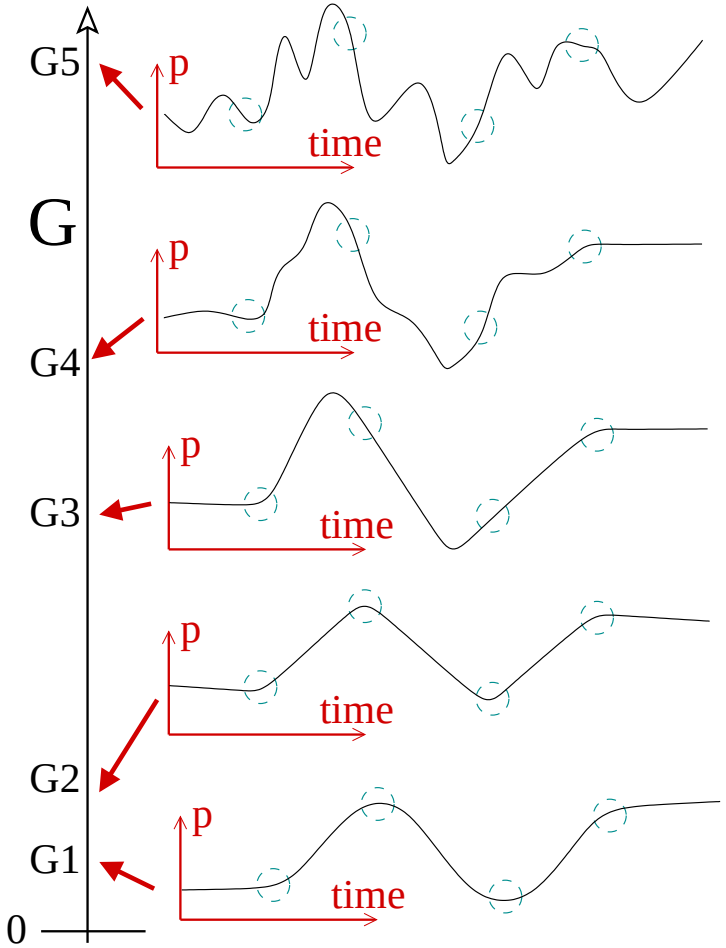
Effort:

$$G = \sum_t \dot{p}_t^2 + \tau^2 \ddot{p}_t^2 + \eta^2 p_t^2$$

where p_t is the muscle tension (pitch) at time t, γ and η define the dynamics.

Techniques - Modeling Prosody - The Math

How does G depend on the shape of the pitch curve?



$$G = \sum_t \dot{p}_t^2 + \tau^2 \ddot{p}_t^2 + \eta^2 p_t^2$$

Techniques - Modeling Prosody - The Math

Weighted error probability:

$$R = \sum_{i \in \text{targets}} s_i^2 r_i$$

each *target* is the encoding of some linguistic information, and s_i is the importance of the i^{th} target.

Error from i^{th} target:

$$r_i = \sum_{t \in \text{ targeti}} \alpha(p_t - y_{i,t})^2 + \beta(\bar{p} - \bar{y_i})^2$$

where $y_{i,t}$ is the i_{th} pitch target; $\bar{y}$ and $\bar{p}$ are averages over the target; α and β define whether the shape or postion of the target is important.

Techniques - Modeling Prosody - The Math

Finally, minimize $G + R$ over all possible p_t.

Recall, G (*Effort*) is small for smooth pitch changes, and R (*Error*) is small when the pitch matches the template.

- If strength is big, *Error* matters, and pitch matches target.
- If strength is small, *Effort* matters, and both speaker and listener accept large pitch errors.
- If strength is near 1, everything compromises.

Techniques - Modeling Prosody - The Math

Q: Where did this “strength” come from?

A: What’s two centimeters plus three hours?

- *Effort* involves physical motions: it has units of energy (*e.g.*, 0.003 Watt-seconds).
- *Error* is just the probability of being misinterpreted: it’s a pure number (*e.g.*, 0.23).
- You need a multiplier to make the units agree.
- There’s nothing in the physics that forces the multiplier to be the same from one word to the next.

Techniques - Modeling Prosody - The Math

Connection of a scalar strength to categorical accents.

Listeners *might* treat high-strength accents as categorically different from low-strength ones.

Low strength ? = no accent

High strength ? = accent

A Modeling Example

The model uses 4 different accent templates:

- An accent on the digit to be confirmed.
- A second kind of accent on every other digit.
- An initial boundary tone.
- A final boundary tone.

All the sentences share the same template shapes.

A Modeling Example

The accent templates are

- $\longleftarrow$stretched$\longrightarrow$ in time, to fit word duration, and
- scaled in $\begin{matrix}\Uparrow \\ pitch \\ \Downarrow\end{matrix}$, depending on each word's strength.

The same templates are used for both 1 and 2 syllable words, but the accent position is tied to the stressed syllable.

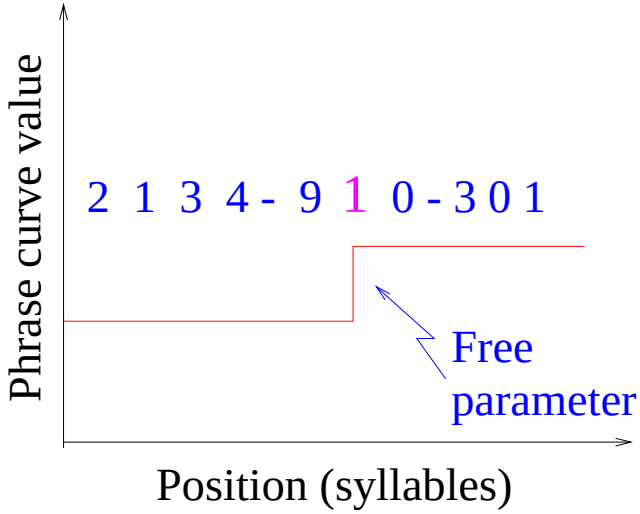
Each accent template is used many times.

A Modeling Example

The Phrase Curve.

We built the model to test whether a phrase curve was necessary to describe this data.

The phrase curve consists of a step, at the position of the digit to be confirmed.

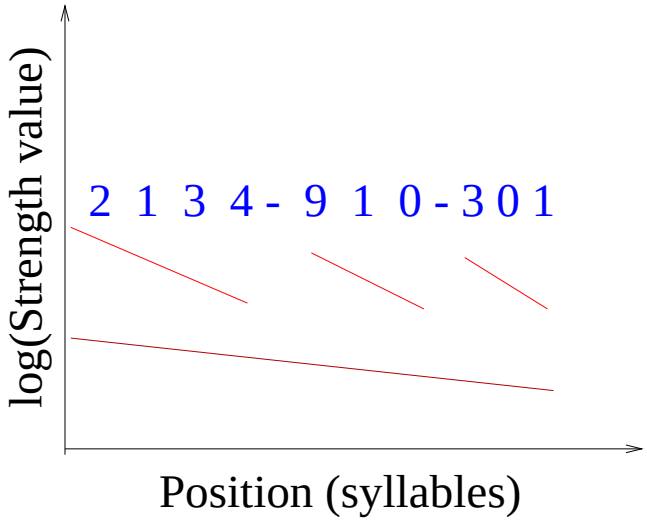


Null hypothesis: the step size is zero, therefore no phrase curve is necessary.

A Modeling Example

Strength of words.

The model we use puts phrasing information into the prosodic strength.



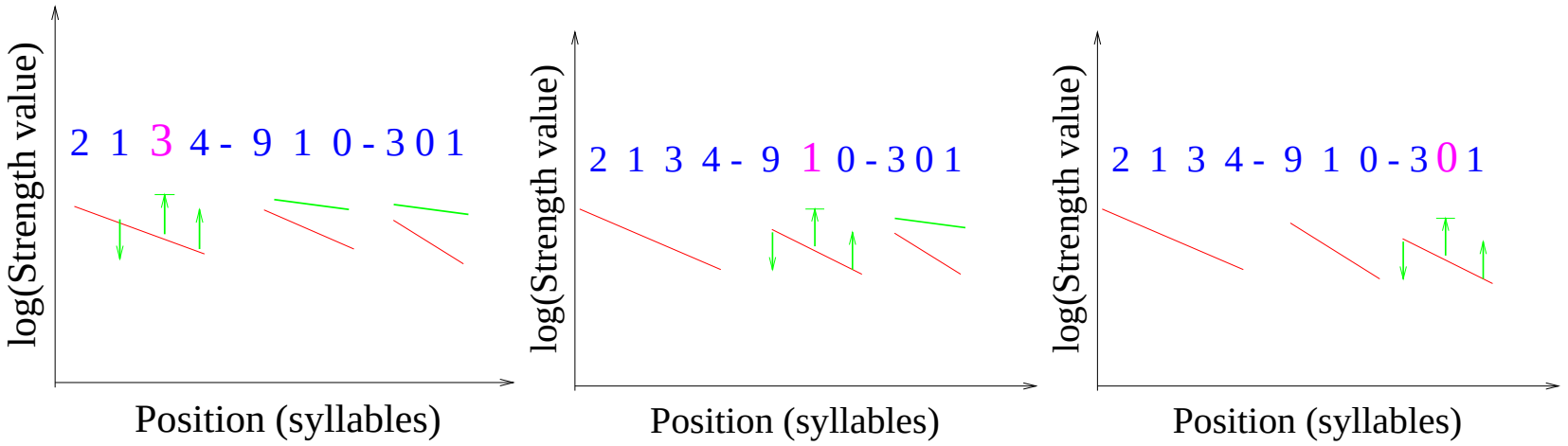
Generic question:

A Modeling Example

Strength of words.

When the question confirms a particular digit:

- Strength of all emphasized digits set to a shared value.
- Strength of preceeding digit reduced by a shared factor.
- All following digits has a weakened phrasing (One shared factor).
- All following digits have a strength boost (One shared boost factor).



A Modeling Example

What does a model actually look like?

header:

```
base=1*(1+BA); smooth=(0.1*exp(SM)); ...
```

utterance data/Q2.f0.2c:

```
ac_name=wat1; info=Nine;
pos=WCTR*0.520+(1-WCTR)*0.520+0.1*POFF;
strength=exp(STR+STRPF*0.125+STRUF*0.041667);
wscale=exp(WSCALE)*pow(0.41,WSE)*pow(0.41,SSE) ...
...
```

A Modeling Example

What does a model actually look like?

```
# boundary tones:
Cname=stress; ac_name=bts1; pos=0.3251;
strength=exp(SBTS1)*pow((exp(STR+STRPF*0.125+ ...
wscale=exp(FBTS1)*pow(0.41,BTDES)* ...
...
# phrase curve:
Cname=step_to; info=Start; pos=0.3251; to=PHRNOACC;
...
# Accent templates:
name=waxt1;
shape=[(-0.5,WAXT1S1),(-0.25,WAXT1S2), ...
type=WAXT1TY;
...
```

Conclusions

- Prosody is an important part of speech.
- It carries information that is *not* in the sequence of words.
- Multiple types of information are encoded in F_0.
- To separate them, you need a model.
- You need to test the model on real speech.

Conclusions

Building Models.

- You can build models where the parameters are tied to linguistic factors, so there is a reasonable hope that the model will match new data.
- We can build very simple descriptions that work.
- Can account for fast/slow speech variation without any adjustable parameters.
- We have a detailed model of a small corner of English.

Final Conclusions

This modeling technique can be applied to

- Tone languages (Mandarin).
- Non-tone languages (English confirmatory questions).
- Singing.
- Different speaking styles. [MLK&Schacht]