Physical Modeling of Linguistic Features

Greg Kochanski

Chilin Shih (Bell Labs)

Tan Lee (CUHK)

Hongyan Jing (Bell Labs)

Jiahong Yuan (Cornell)

Yujia Lee (CUHK)

(Talk given atQueen's University, Kingston, Canada 30 Septemver 2002. It is available on the web at http://kochanski.org/gpk/papers/2002/Physical_Modelling_of_Linguistic_Features.pdf .

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

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

The prosodic channel

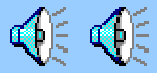
To prove and illustrate the prosodic channel, observe an elegant experiment by Stan Freberg (1950):

The text has no lexical information, but it still tells a story.

Physical implementations of prosody

- Intonation (pitch) is one of the more important components of prosody

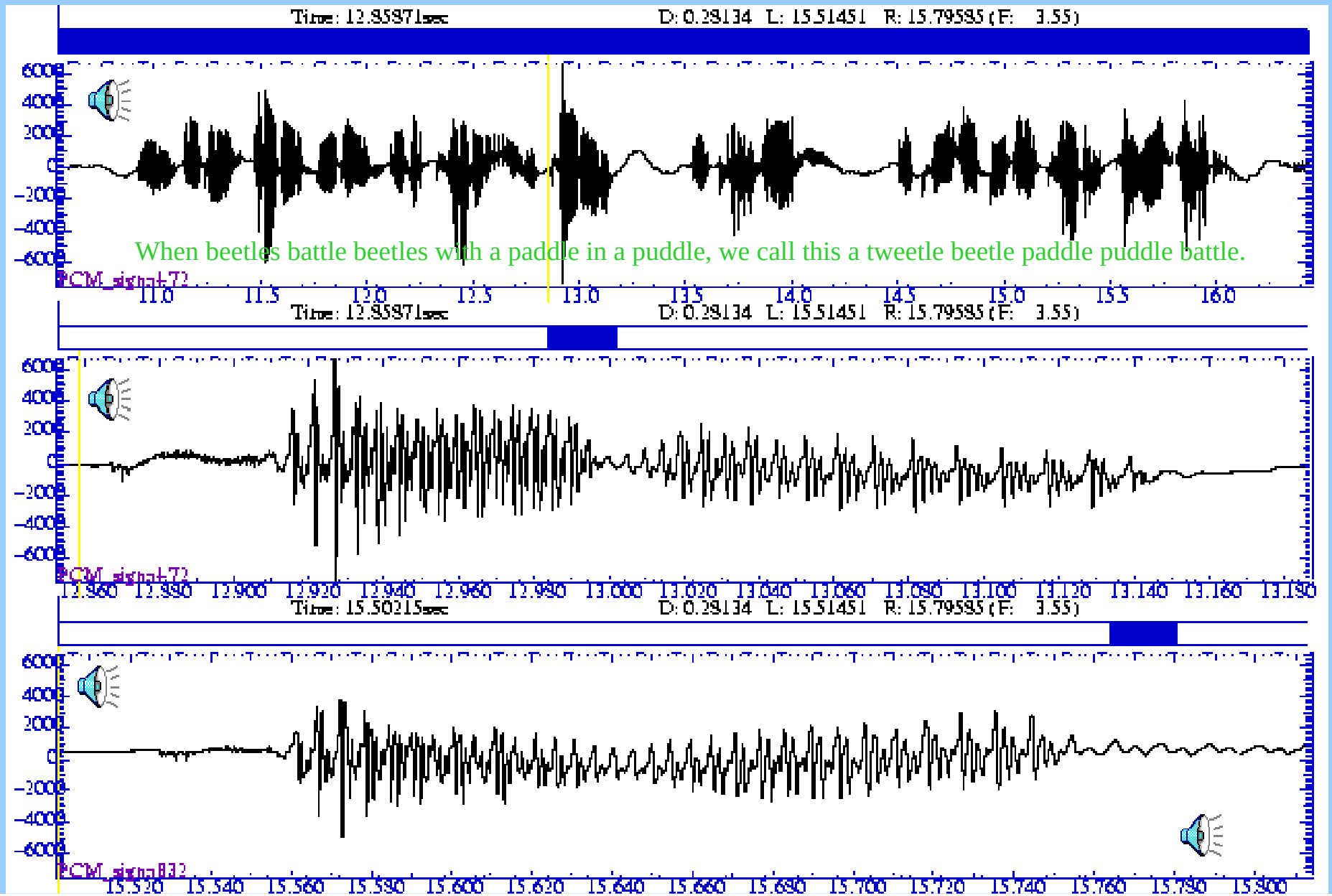




- Also duration, loudness, facial expressions.

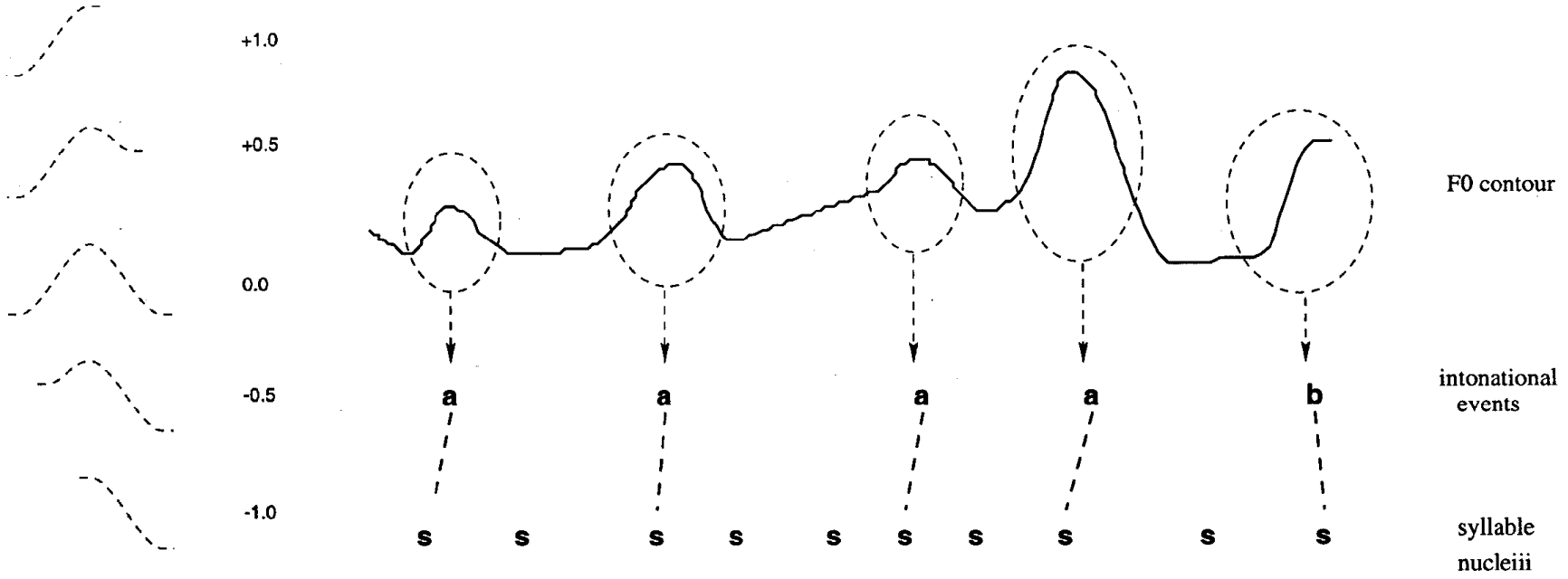


Suitable for Modeling



Existing work

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



Existing work

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

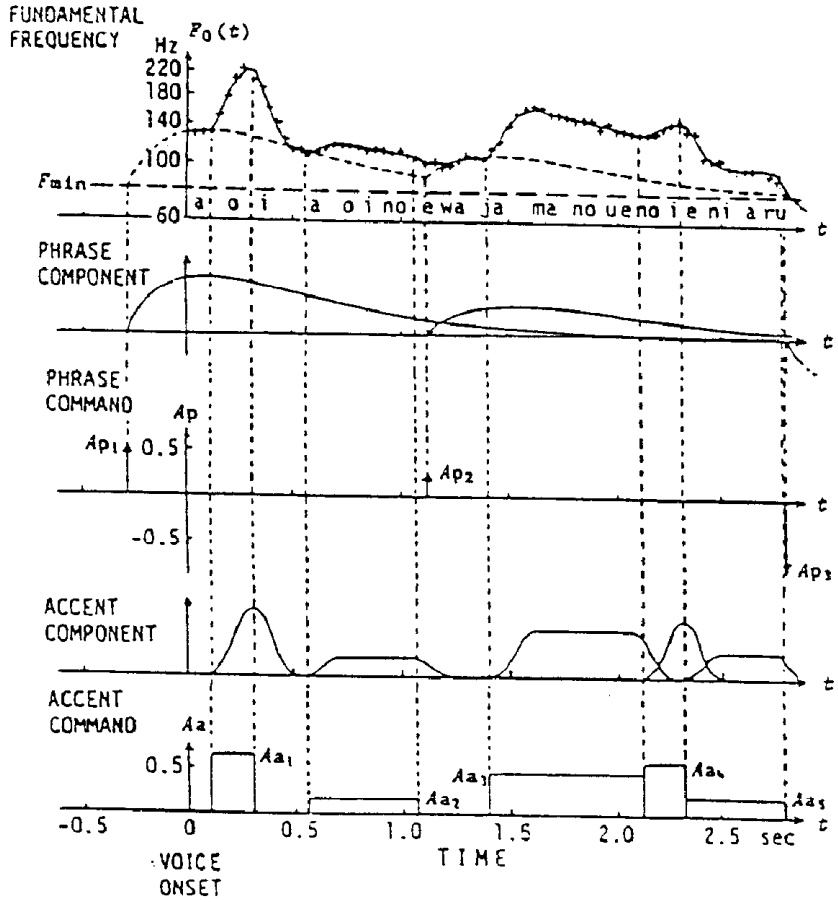
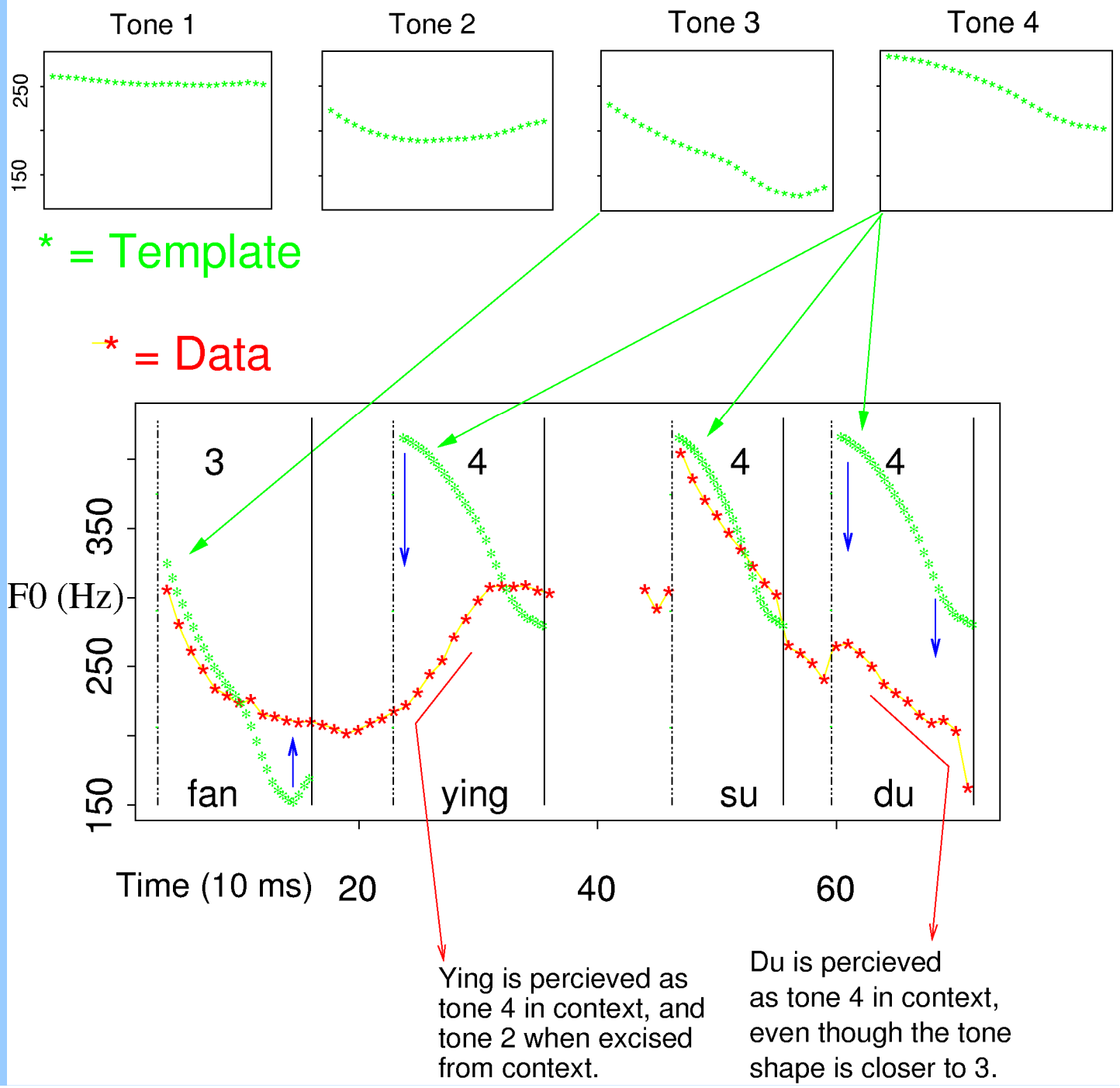
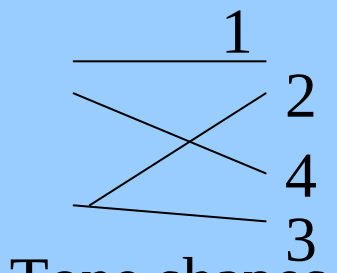


Fig. 4. Analysis-by-Synthesis of an F_0 contour of the Japanese declarative sentence: /aoiaoinoewajamanouenoieniaru/. The optimum decomposition of a given F_0 contour into the phrase and accent components is illustrated, and the underlying commands for these components are shown.

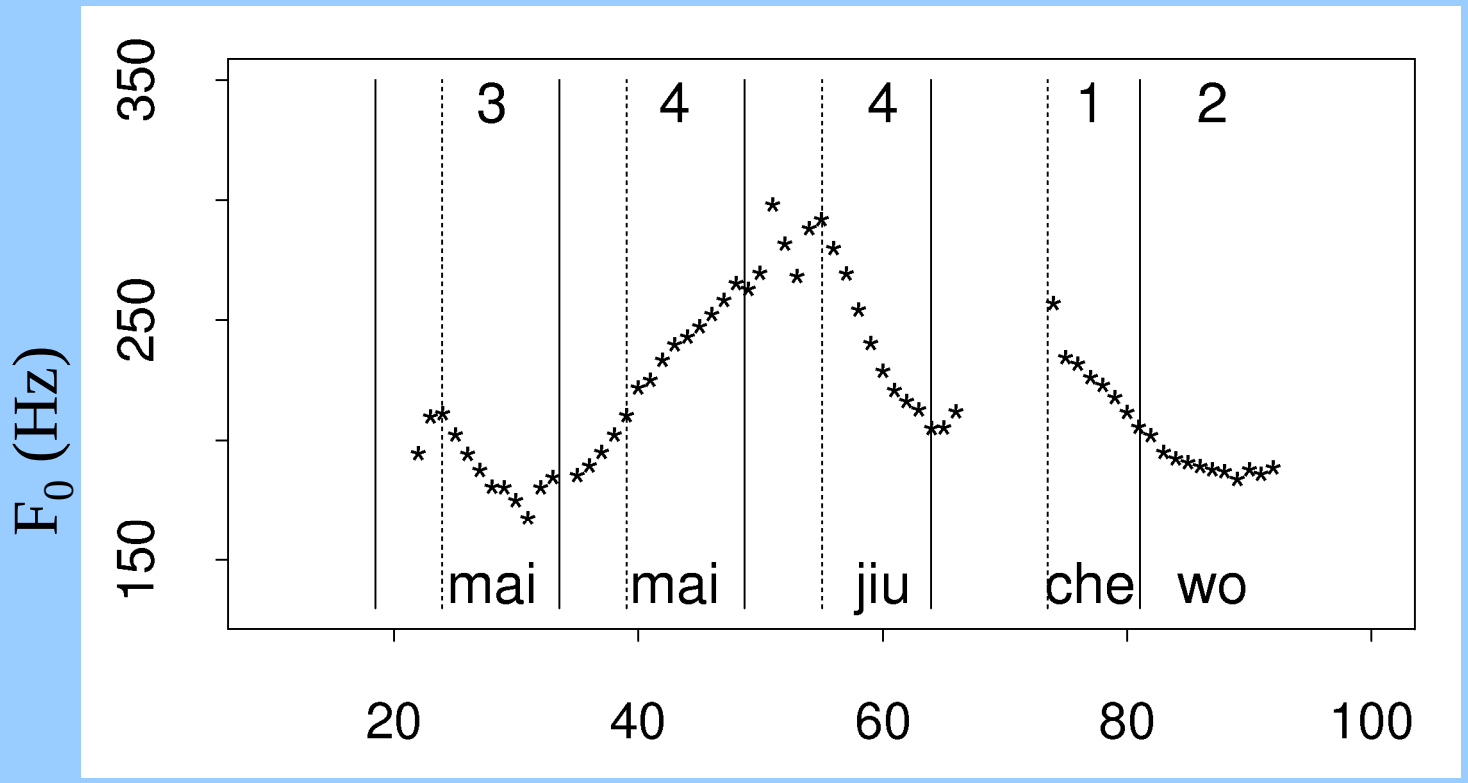
The Challenge



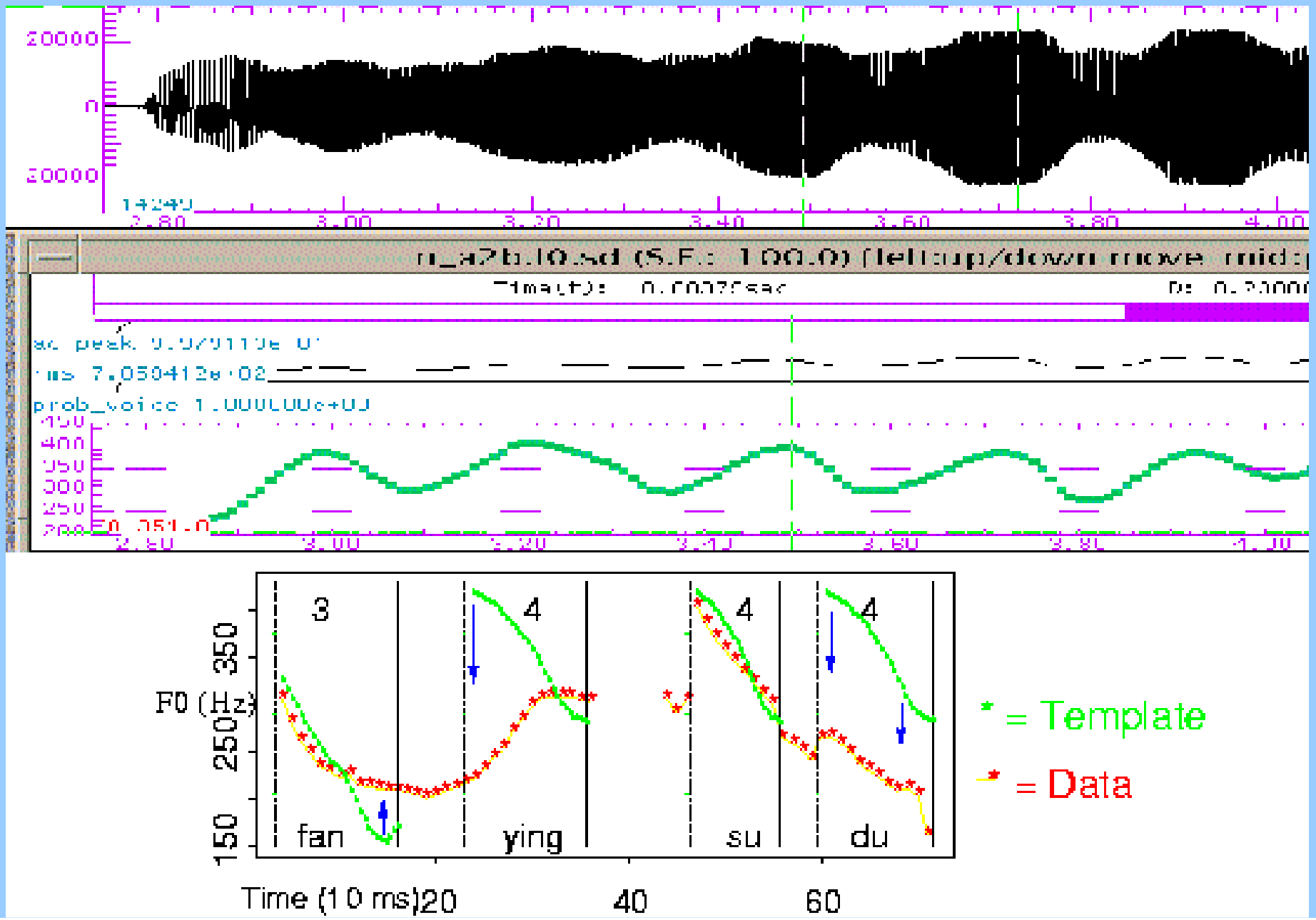


Tone shapes

Another Challenge



People talk nearly as fast as possible.



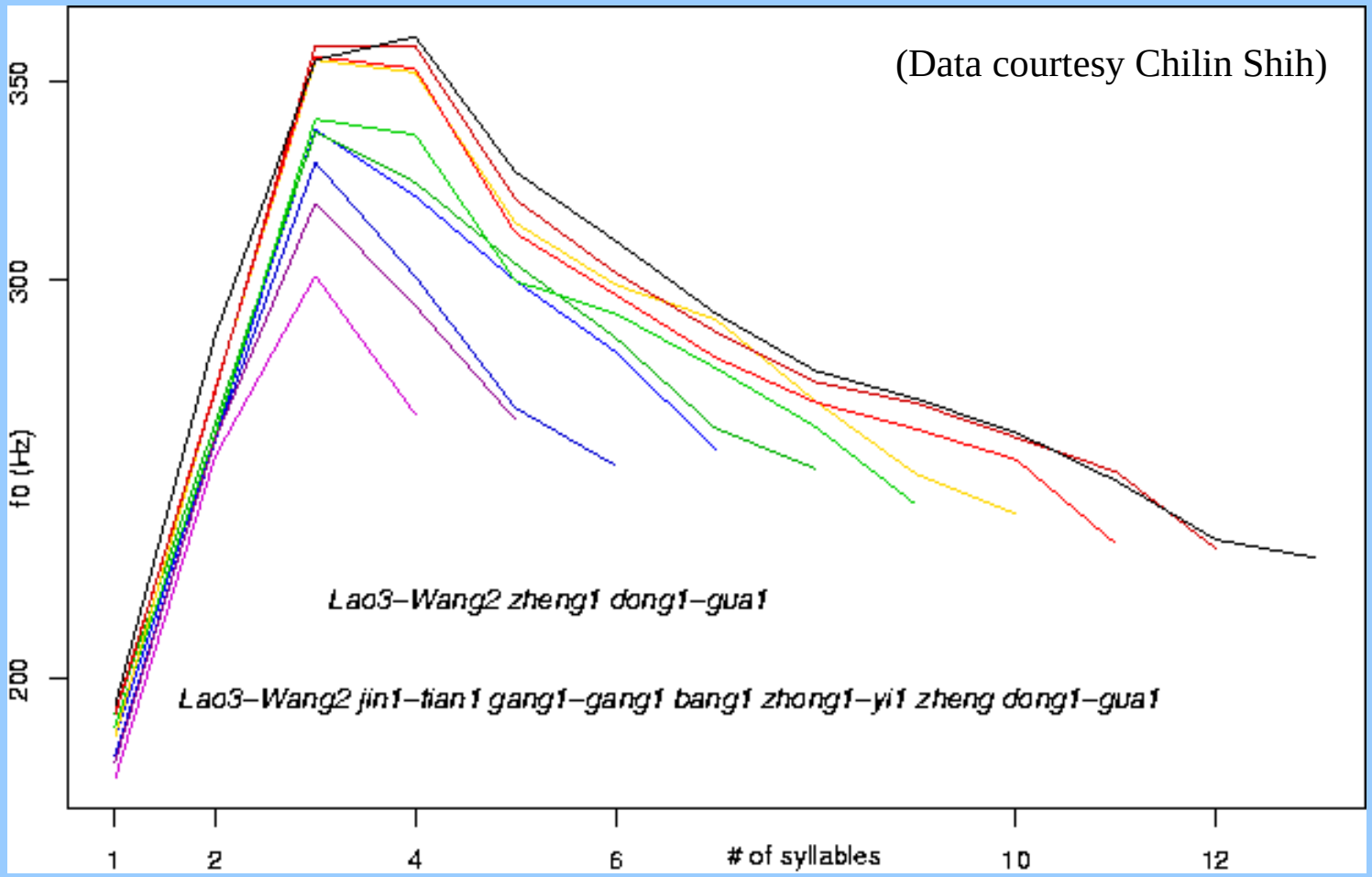
Basic assumptions used in modeling

- People plan their utterances several syllables in advance.
- People produce speech optimized to meet their needs.
- A dynamical model for the muscles that control f_0
- Approximate linearity in the muscle – f_0 mapping.
- …and what falls out of the model is…
- A strength parameter for each word.

Muscle – f_0 mapping is ~linear

- Try it yourself.
- I. R. Titze Principles of Voice Production (Prentice Hall, 1993).
- Two mass models (K. Stevens and others)
- Must subtract out “segmental effects” (i.e. effects of other articulators)

Speech is planned.



People start at a higher pitch when they begin longer sentences.
Also planning of inhaled air volume.
Therefore, there is some plan ~300 ms before start of speech.

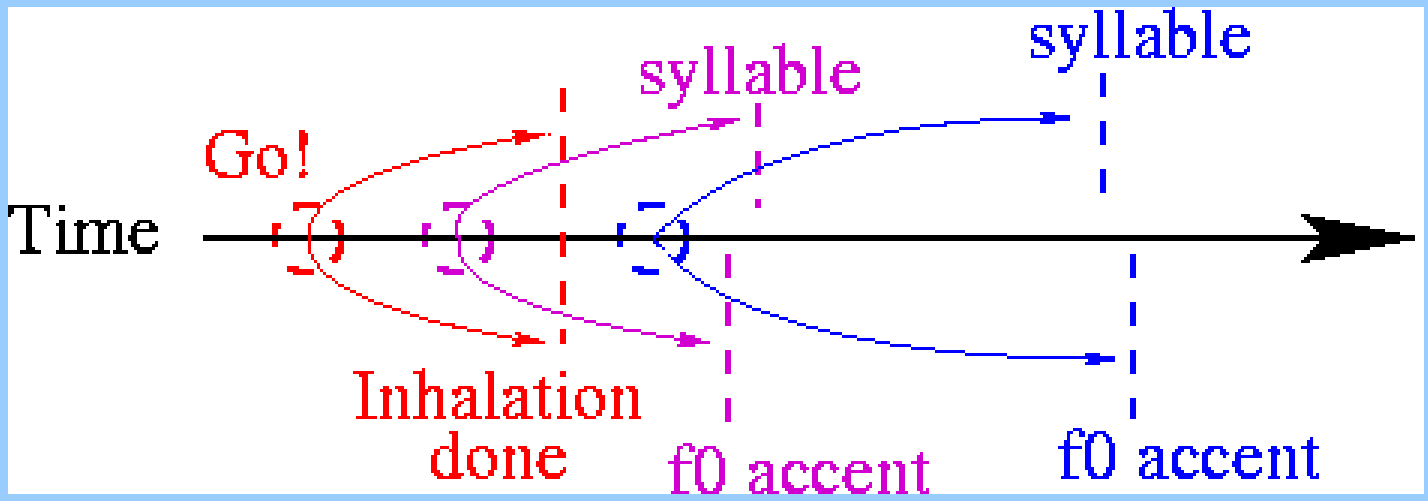
Speech is optimal

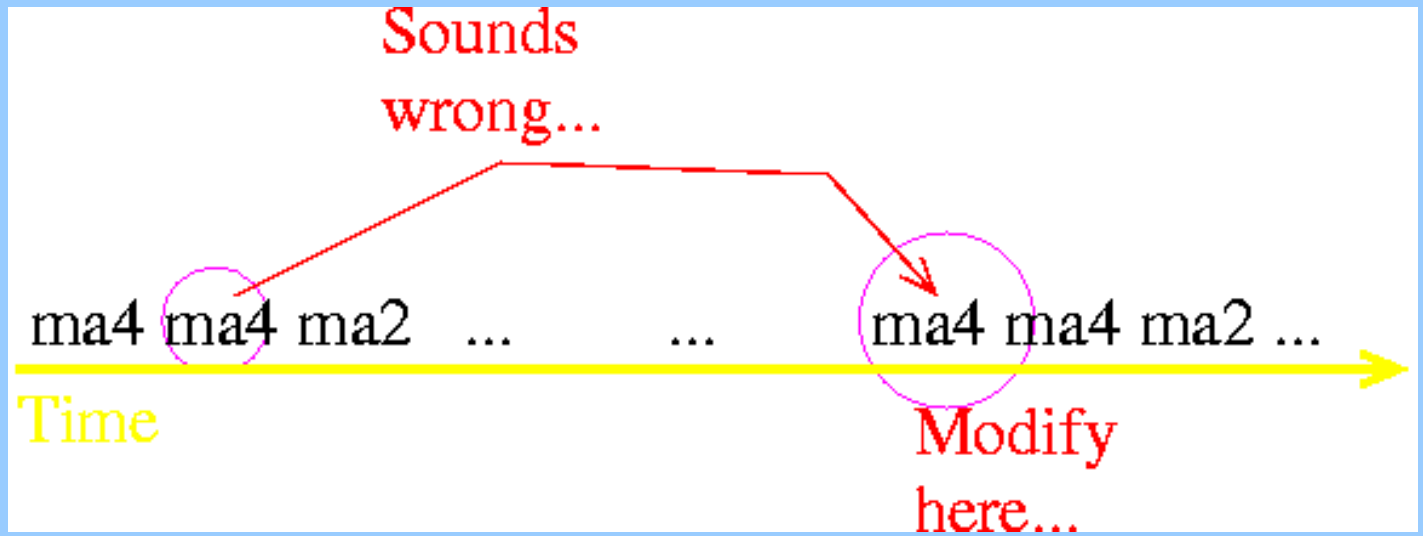
- Most of what we say is made from bits and pieces we've said before.
- There are only 4 (Mandarin) or 6 (Cantonese) tones to combine.
- A speaker has the chance to practice and optimize all the common 3- and 4- tone sequences.

Optimize what?

- People want to minimize the chance that they will be misunderstood.
  - The speaker's estimate of P(misinterpret)
  - The speaker's estimate of risk=P*cost
- People want to minimize effort and/or talk faster
  - Cars
  - Chairs
- How to combine the two?
  - A weighted sum.
  - We allow each syllable to have a different weight
  - Perhaps weight matches importance.

Planned, optimized speech can look non-causal.





Modeling math

$$p(t) = \arg\min_{p(t)}(G + R)$$

$$G = \int dt\left(\dot{p}^2 + \tau^2 \ddot{p}^2 + \eta^2 p^2\right)$$ "Effort"

$p(t)$ is the muscle tension (~frequency) at time *t*.

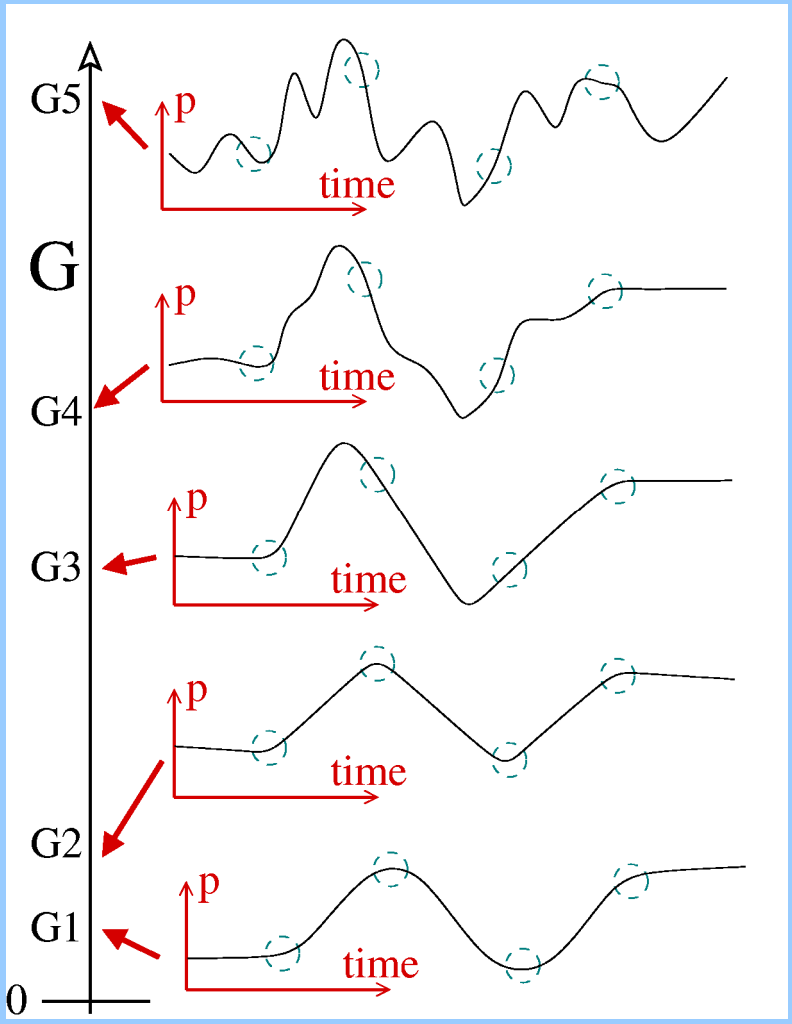
$$R = \sum_{i \in \text{targets}} s_i^2 r_i$$ Each target encodes some linguistic information, r_i is the error of the i^{th} target, and s_i is its importance.

$$r_i = \int_{t \in \text{target}_i} dt\left(\alpha((p - \bar{p}) - (y - \bar{y}))^2 + \beta(\bar{p} - \bar{y})^2\right)$$ "Error"

y is the i^{th} pitch target and a bar denotes an average over a target.

$$y \equiv y_i(t)$$

“Effort”



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

$$G = \int dt\left(\dot{p}^2 + \tau^2 \ddot{p}^2 + \eta^2 p^2\right)$$

Model behavior

- For s>>1, Error dominates, and pitch matches target.
- For s<<1, Effort dominates, both speaker and listener accept large deviations, and pitch smoothly interpolates.
- For s~1, everything compromises.

Where did this “strength” come from?

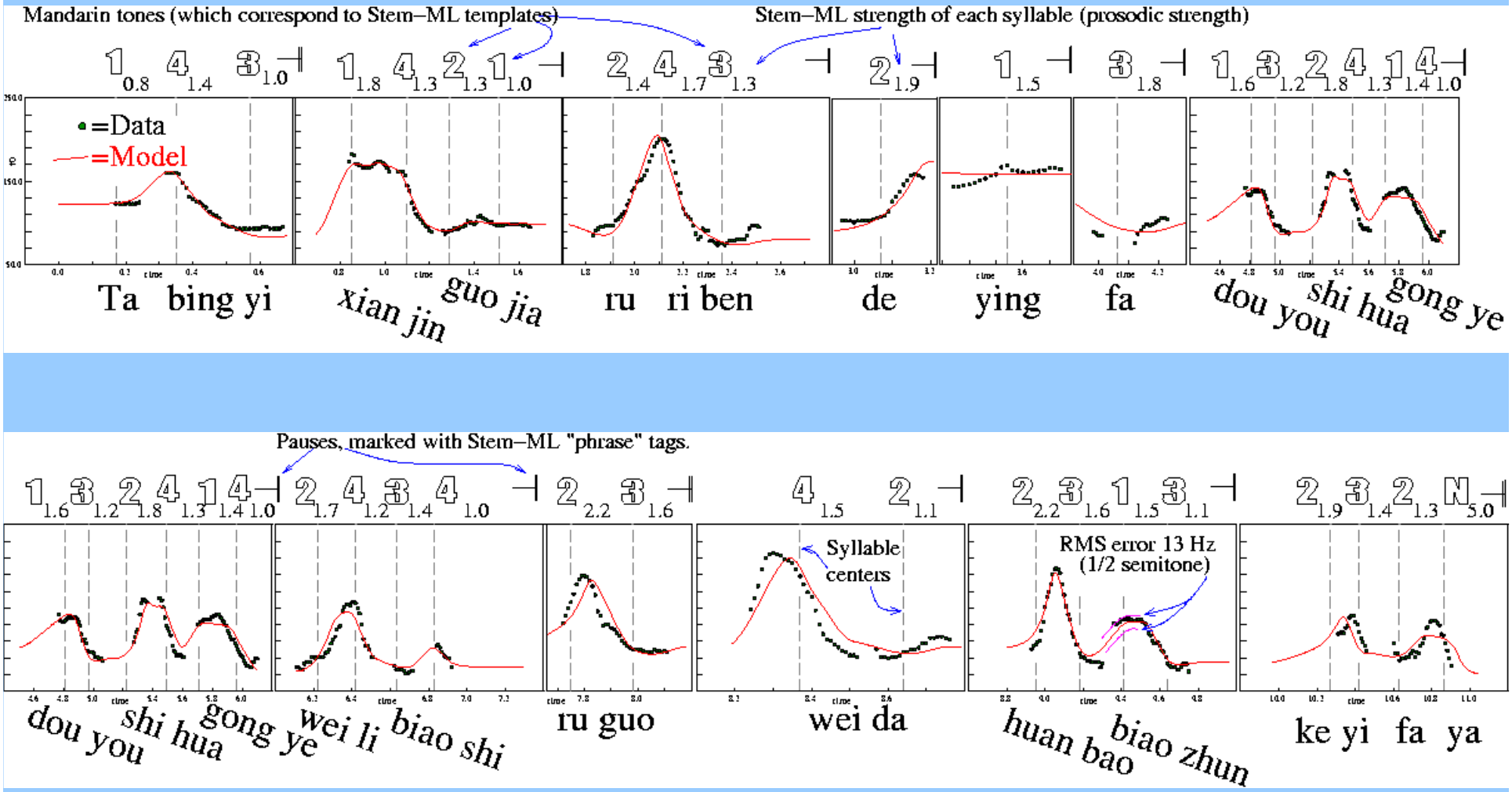
- A: What is 2 cm + 3 g ?

- “Effort” can have energy units.
- “Error” can be a pure number (error probability).
- A multiplier is needed to make the units agree.
- Nothing in the physics forces the multiplier to be the same from one word to the next.

The rest of the model.

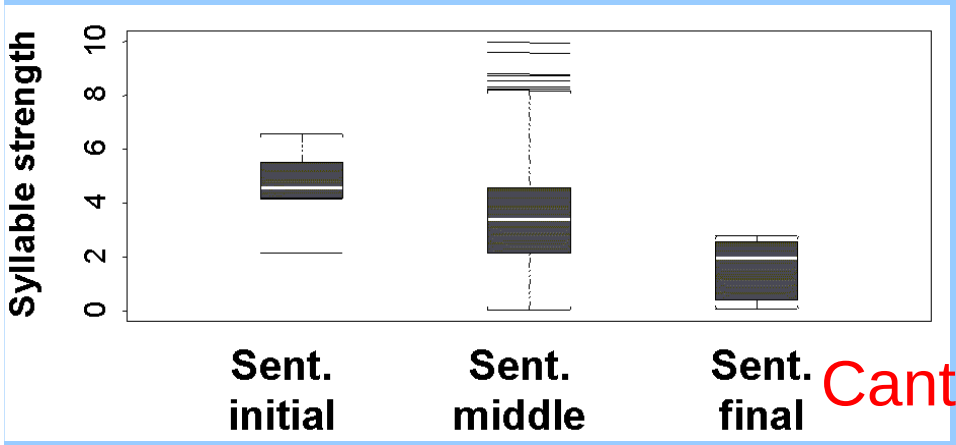
- A model is a sequence of targets
- Each target has a strength.
- For tone languages, there is one target per tone.
- Targets are stretched to fit syllable duration.

Model fits to Mandarin Chinese

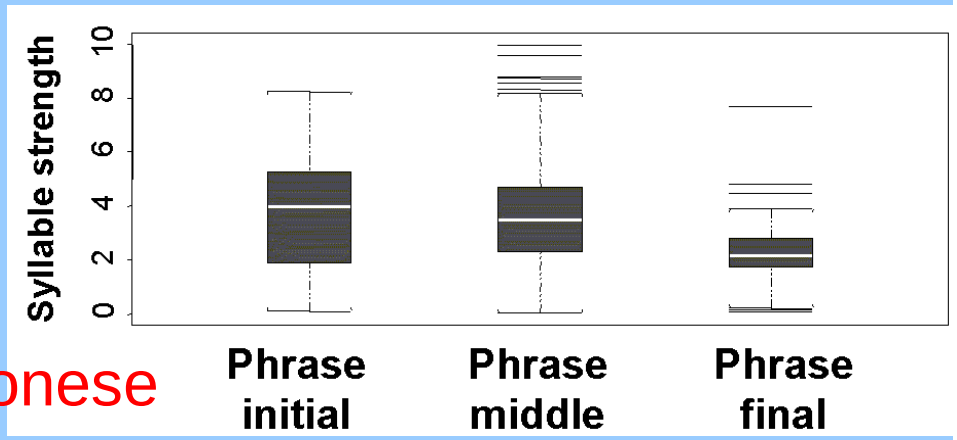


0.61 free parameters per syllable, 13 Hz RMS error.

Model parameters



Cantonese

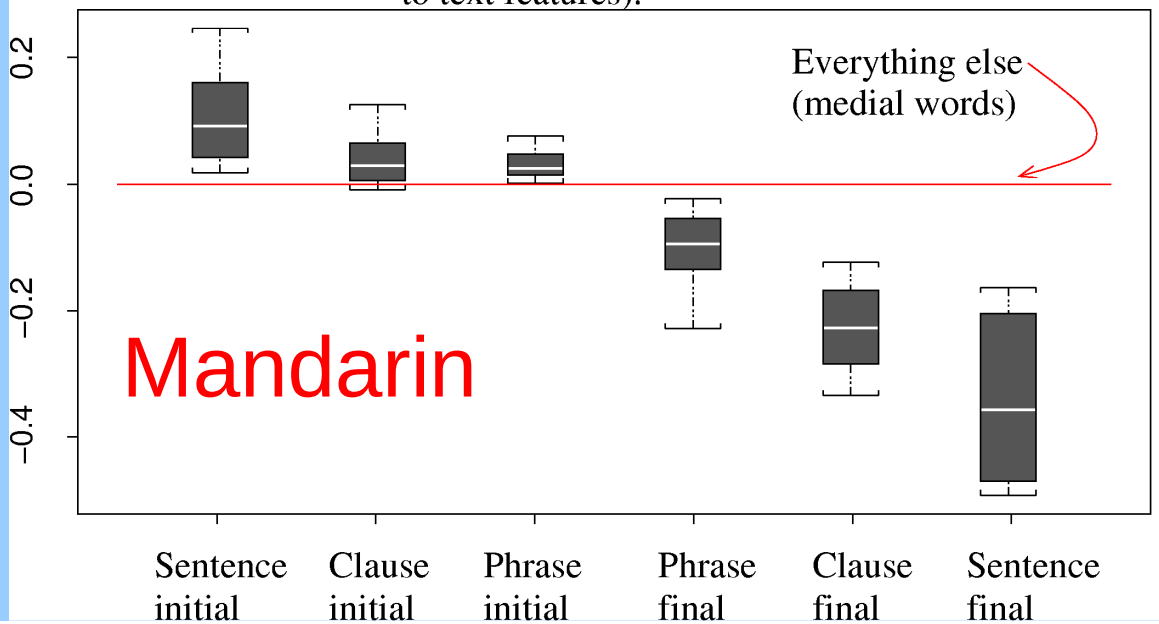


Phrasing is marked in speech.

Cantonese data courtesy of Prof. Tan Lee

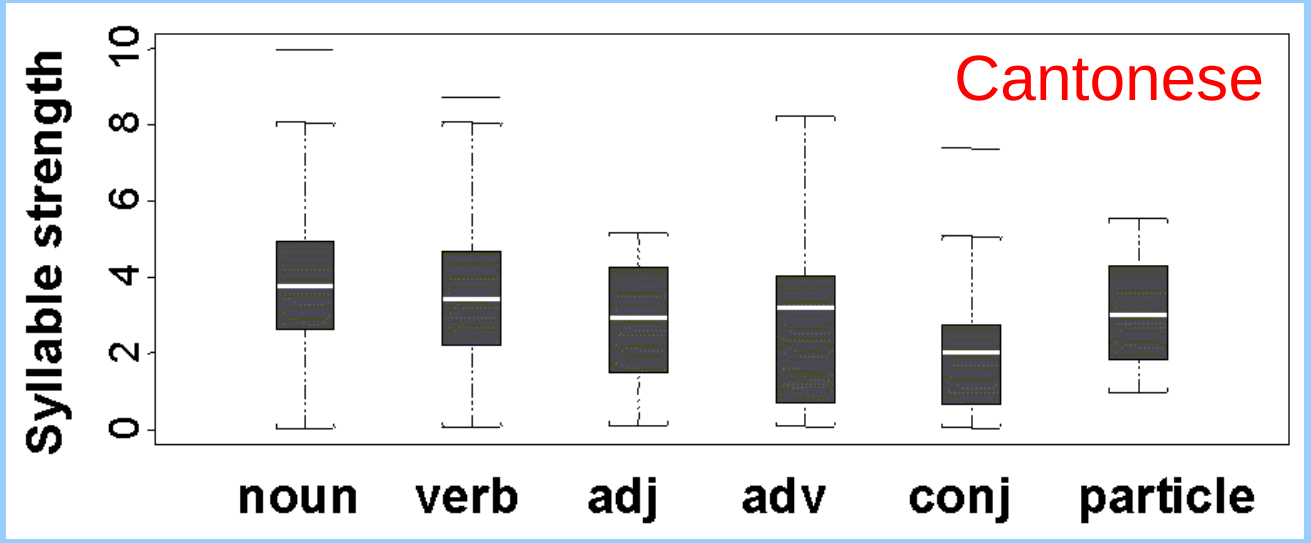
Strength coefficients vs. position in sentence

(results of linear additive model fitting strength to text features).

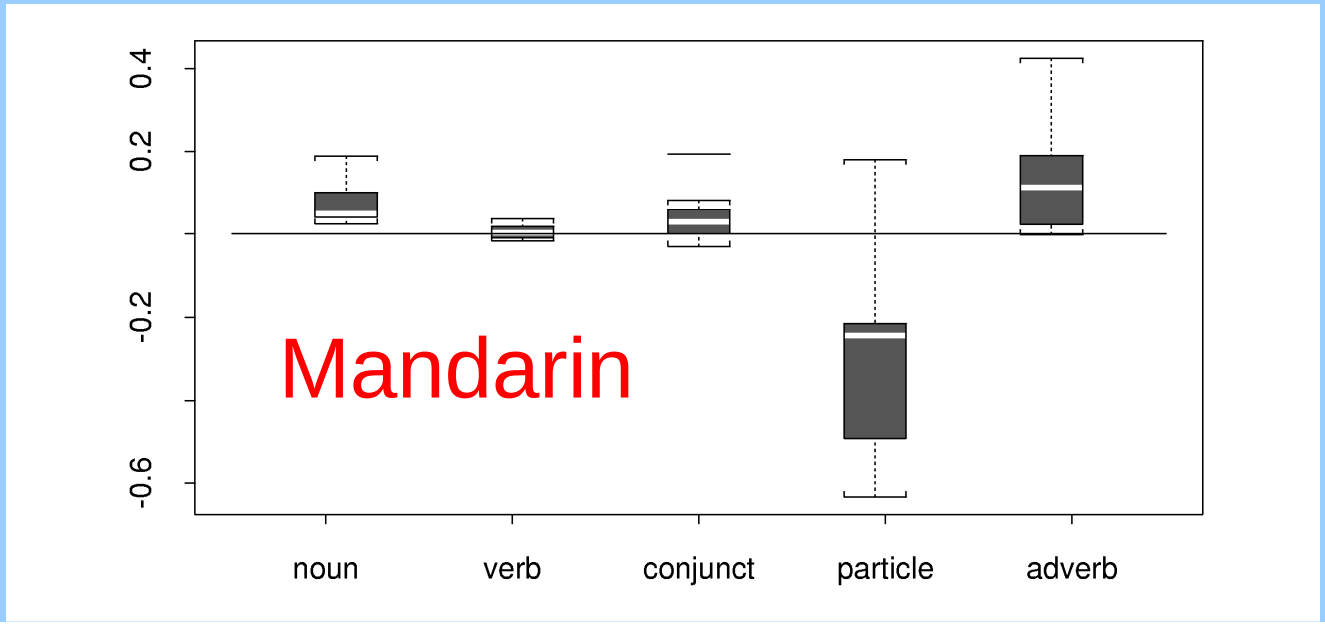


Mandarin

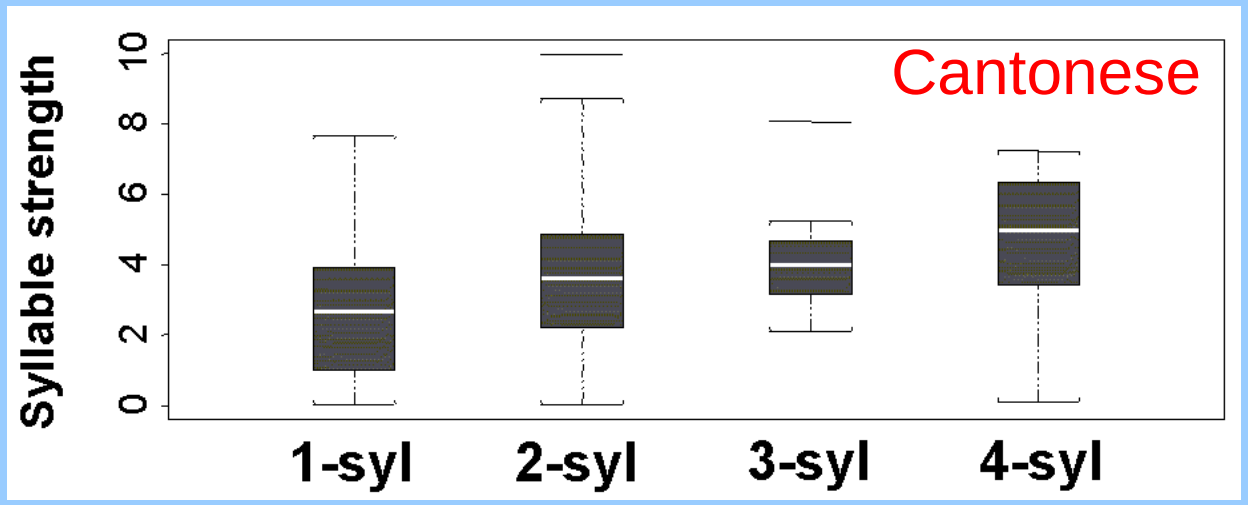
Model parameters

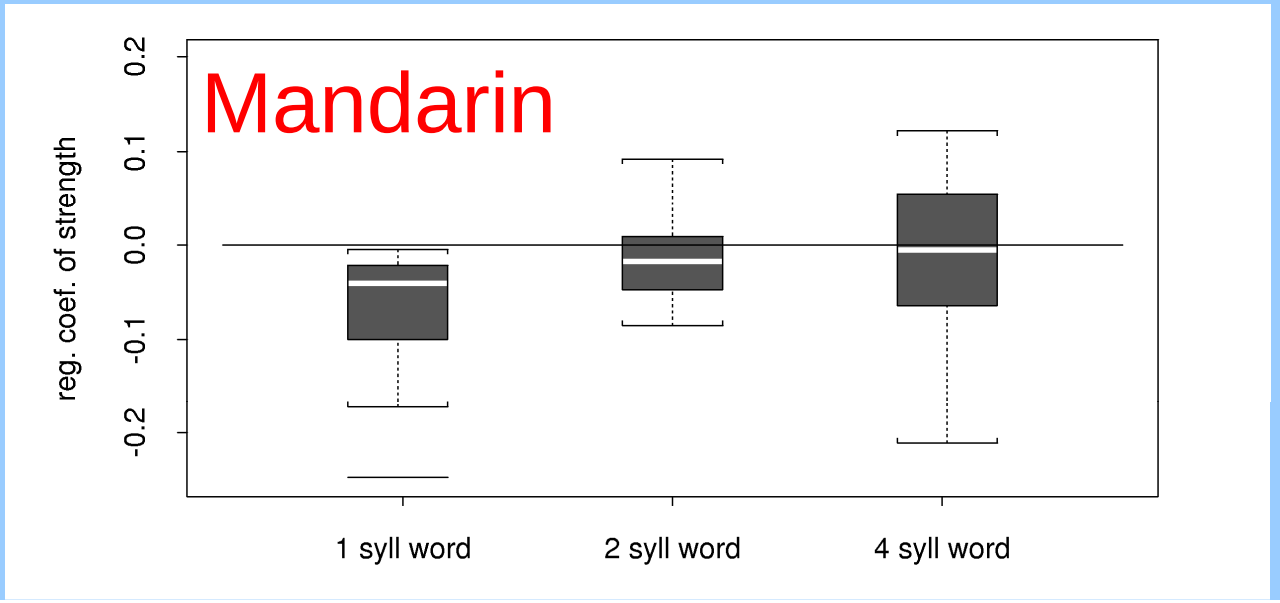


Nouns are relatively important.



Model parameters

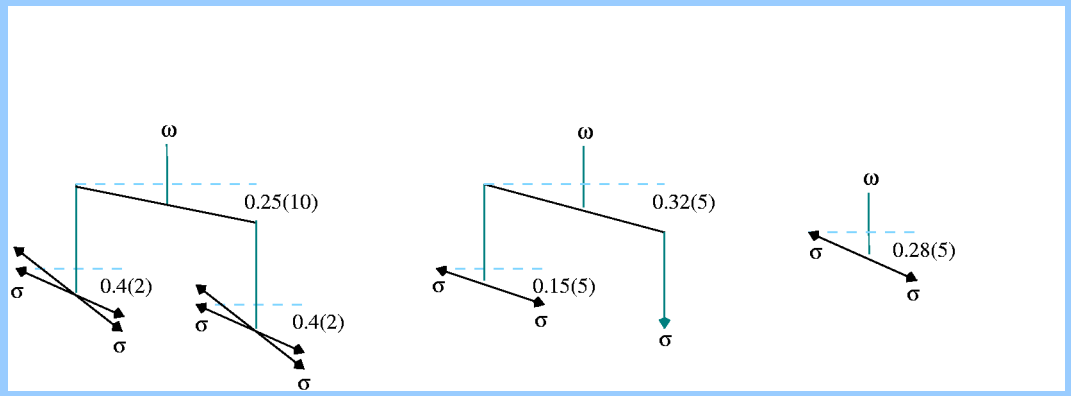




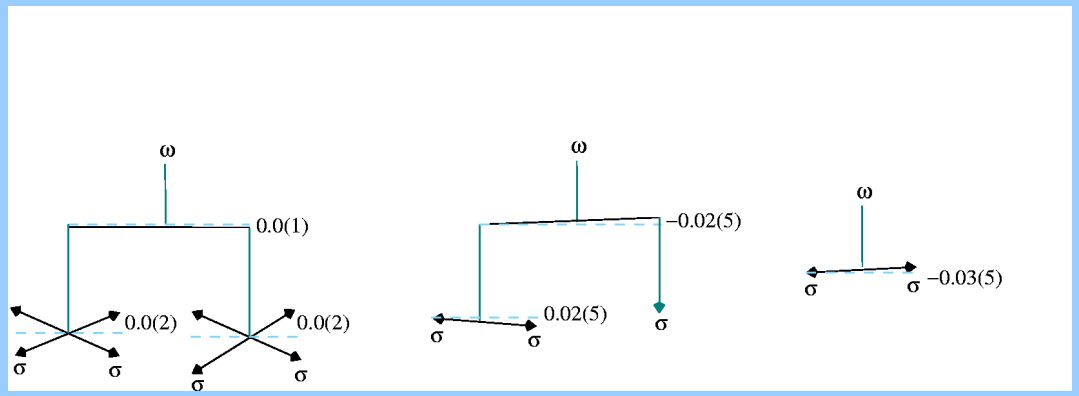
Longer words tend to be spoken more carefully.

Metrical patterns

Mandarin



"Normal" segmentation of characters into words.



Random segmentation of characters into words.

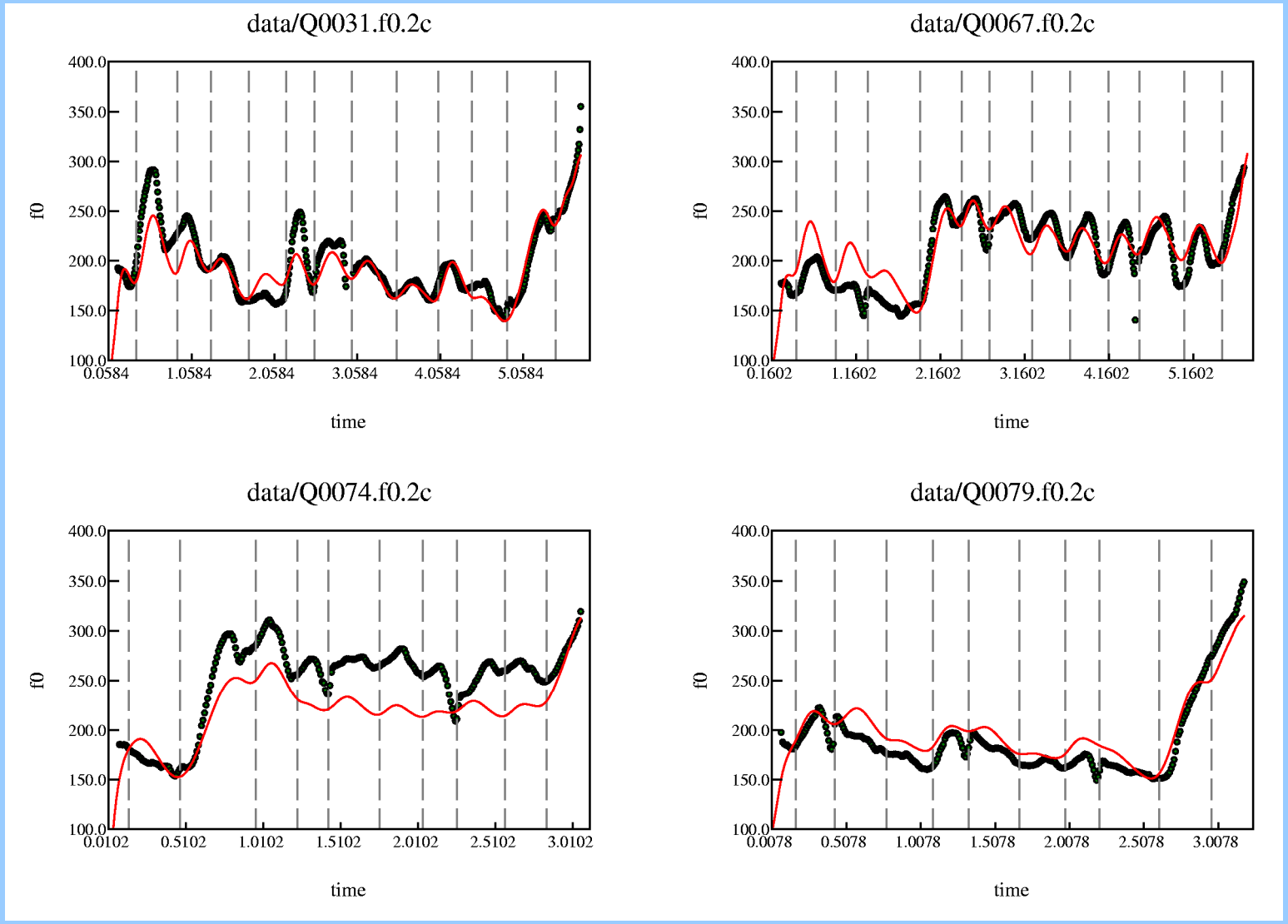
Chinese conclusions

- The strength parameters seem real
  - Similarities across language
  - Match linguistic expectations
- Can answer some outstanding questions: does Cantonese have 6 or 9 tones?
- Can quantify the data rate via intonation: ~16 bits/second
- Some insights into how infants acquire language.
- Better synthesizers and dialog systems.

English

- Sentences in the form "123-456-7890?"
- Speaker is trying to confirm a single digit.
- Models have just 1.1 parameters per sentence.
- Simple, linear forms for strengths in phases.
- Simple, additive linear model for effects near digit to be confirmed.

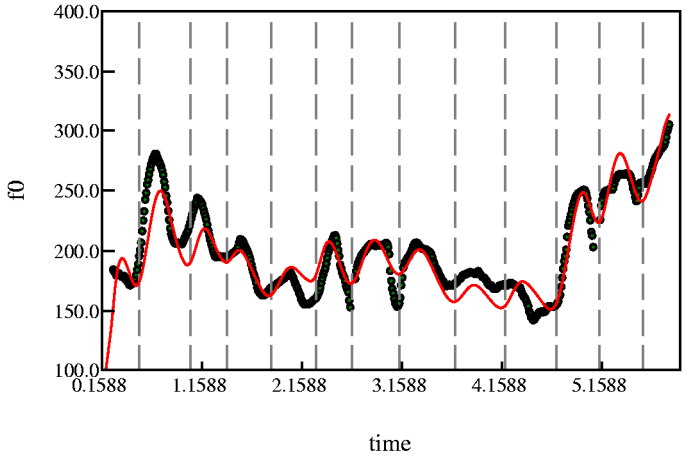
Model fits over a range of speeds.



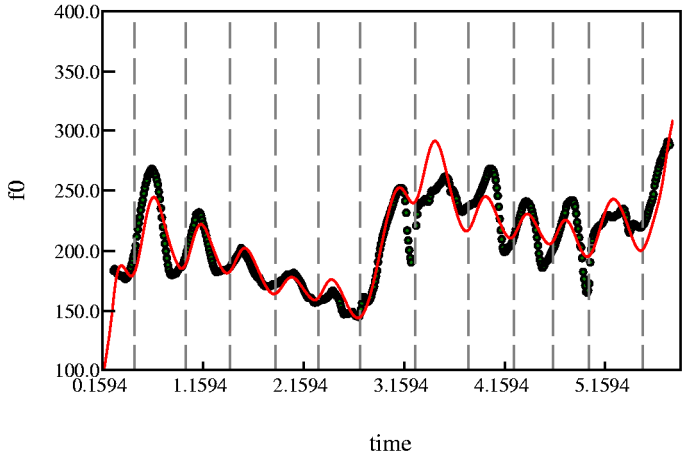
More fits - English confirming questions.

RMS deviation 0.212 Barks/21 Hz/1.7 semitones

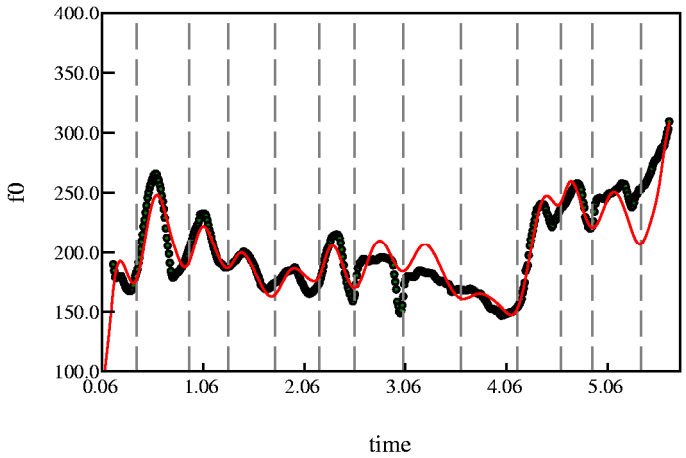
data/Q0034.f0.2c



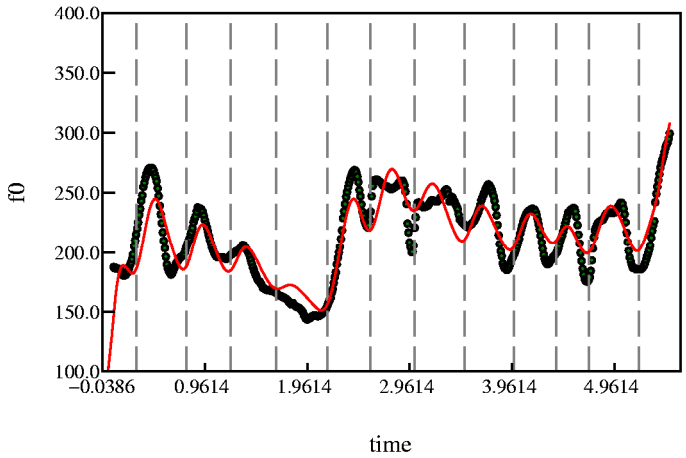
data/Q0039.f0.2c



data/Q0040.f0.2c



data/Q0041.f0.2c



Conclusion

- Modeling of muscle dynamics captures some important aspects of speech.
- It allows measurements of linguistic features.
- It can be applied broadly:
  - Two dialects of Chinese
  - Some aspects of English
  - Separating different singing and speaking styles from the content.
- If similar techniques were commonly used much of the linguistic fields of phonetics and phonology would be replaced