Implications of Prosody Modeling for Prosody Recognition

Chilin Shih, Greg Kochanski, Eric Fosler–Lussier (Bell Laboratories)
Melody Chan (Yale University), and Jia–Hong Yuan (Cornell University)

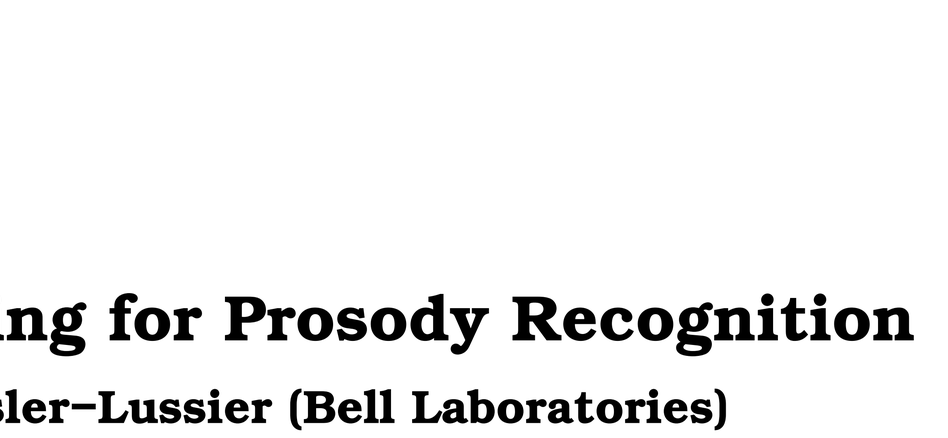
{cls,gpk,fosler}@research.bell–labs.com

Lucent Technologies
Bell Labs Innovations

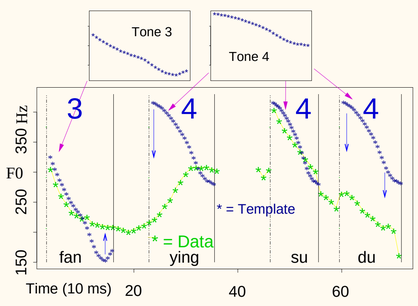
Abstract

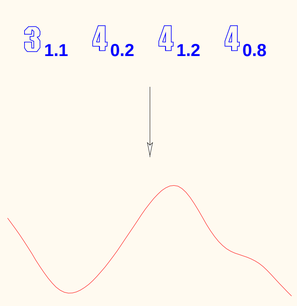
Recognition of prosody is difficult because many linguistically meaningful gestures of intonation are not obvious from the surface intonation contour. Stem-ML (Soft TEMplate Markup Language) is a model of prosody planning and execution, grounded in physiology and communications theory that has some ability to untangle the interactions among accents. The underlying accents are simpler, relatively independent, and more predictable. This model may impact ASR in three areas:

1. Recogniton of prominence: why unstressed words can have high f_0 values.
2. Recognition of questions: a few underlying patterns account for diversity on the surface.
3. Prosodic classification of phrases: Accent shape modeling over English noun phrases.



Prosodic Model: Soft Template Markup Language



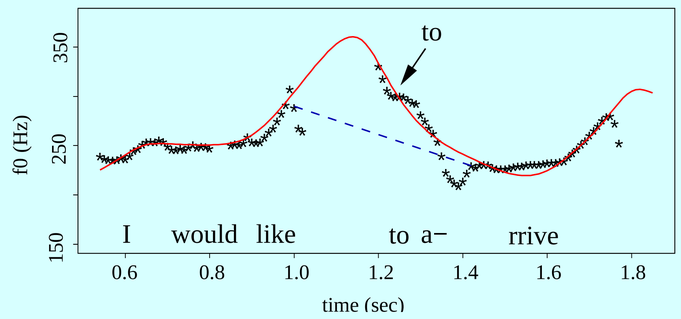


Blue curves show tone templates taken in a neutral environment. The green curve shows the same tones in conversation.

With Stem-ML, the distorted tone shape on the second syllable is accounted for with a low strength value. Four tags, along with global parameters that define pitch range and lexical tone templates, reproduce the observed f_0 contour.

Prosodic Strength

- Unimportant function words often have higher f_0 than their neighbors.
- Normalizing for sentence effects and discourse factors cannot solve the problem of local interpretation of f_0 height relative to nearby words: This complicates any algorithm designed to derive information from prosody.
- Stem-ML can account for the high f_0 of these unaccented words.



Example of a high-pitched function word in "I would like to arrive . . . " (DARPA Communicator database). Data is shown as "*". Dashed line is predicted from ToBI label interpolation; the value is too low which would imply "to" to be prominent. Solid line shows Stem-ML model, where "to" has zero strength, and therefore lies on a smooth curve connecting the neighboring accents.

Stem-ML

Stem-ML combines several ideas into a model of intonation:

- People plan their utterances several syllables in advance.
- People produce speech that is optimized to meet their needs: Speech balances between accurate communication and ease of production.
- A physically reasonable model for the dynamics of the muscles that control pitch: f_0 is smoothly related to muscle tensions.
- A prosodic strength associated with each syllable:
  - High strength → careful articulation → nearly ideal shape & expanded pitch range.
  - Low strength → minimum effort → pitch is controlled by neighborhood.

Stem-ML

Stem-ML calculates a pitch curve by finding the curve that has the smallest sum of *effort*+*error*, where *effort* behaves like the physiological effort: it is zero if muscles are stationary in a neutral position, and increases as motions become faster and stronger. The *error* term measures how far the pitch curve deviates from an ideal template.

In Stem-ML, a "tag" is a tone template, along with a few parameters that describe the scope of the template and how the template interacts with its environment. It corresponds to the mathematical description of an intonation event (*e.g.*, a tone or an accent). Tags also have parameters that control how they interact.

$$error = \sum_{k \in \text{tags}} s_k^2 r_k \tag{1}$$

where

$$r_i = \sum_{t \in \text{tag } i} (1-\alpha^2)^{1/2}(p_t - y_t)^2 + \alpha(\bar{p} - \bar{y})^2. \tag{2}$$

$$effort = \sum_t \dot{p}_t^2 + \tau^2 \ddot{p}_t^2 \tag{3}$$

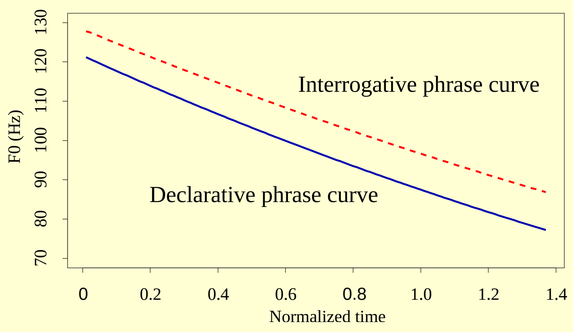
Where y is the template, and p is the pitch, and τ is the muscle response time, and α controls whether the shape or the average value of an accent is most important. The strength of the k^{th} accent is s_k. The index k identifies a tag, and t covers the tag's scope.

Question Intonation

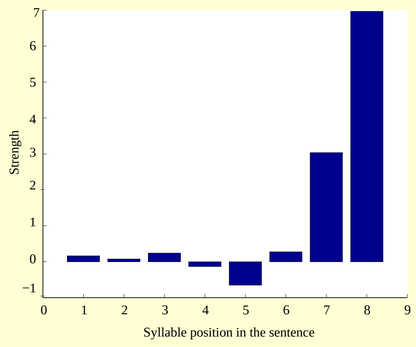
Mandarin question intonation shows an interesting diversity due to tone and intonation interaction. A sentence with a final rising tone (tone 2) has higher ending, while a sentence with a final falling tone (tone 4) has higher peak.

The optimal models trained by Stem-ML explain the differences with two simple mechanisms:

1. Higher phrase curve for question



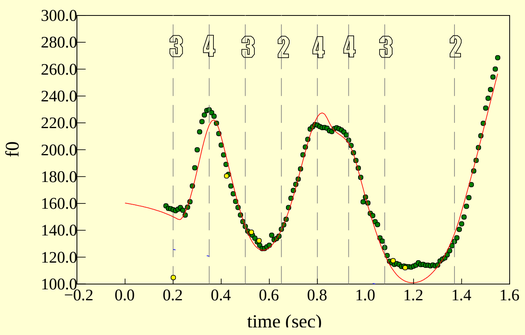
2. Increasing tonal strength near the end of the sentence.



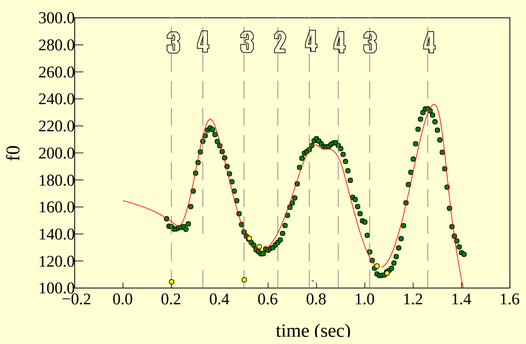
The bars shows strength differences of question and declarative sentences.

The following two plots show the fit of Stem-ML models (red curves) to the observed contour of Mandarin questions (green circles).

A sentence ending in a final rising tone (2). The high ending signals question intonation.



A sentence ending in a final falling tone (4). The high peak signals question intonation.



Noun Phrases in English Dialogues

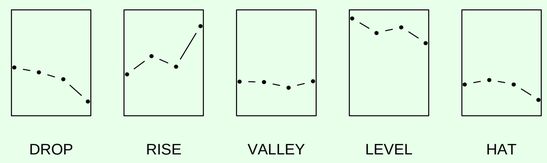
Can consistent prosodic templates be found for English noun phrases?

- Studied noun phrases in DARPA Communicator corpus
  - Human-computer dialogues in travel reservation task
  - Subcorpus: 103 noun phrases from 57 utterances, 26 speakers
- Through data analysis, five basic prosodic classes were found:

Pattern	Code	Freq	Description
DROP	δ	40	primarily falling
RISE	ρ	38	primarily rising
LEVEL	λ	9	no movement
HAT	η	9	initial rise, terminal fall
VALLEY	υ	7	initial fall, terminal rise

o = all words outside the noun phrase β = boundary tones

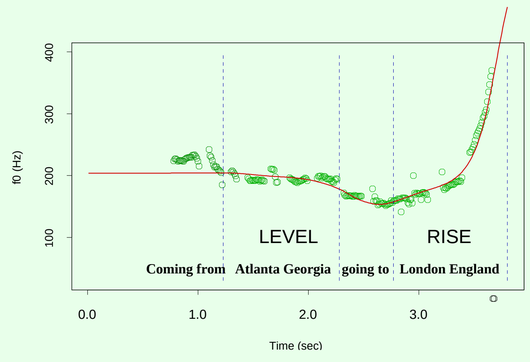
Noun phrase templates trained by Stem-ML.



The following three plots show Stem-ML models (red) of English sentences (green circles). The coding (blue) for each sentence is given below the text.

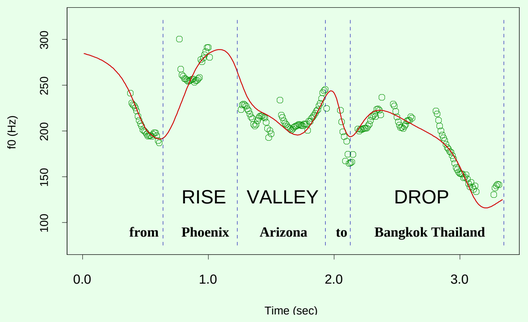
from Atlanta Georgia going to London England

$o_{0.09}$ $\lambda_{0.35}$ $o_{0.75}$ $o_{0.57}$ $\rho_{0.62}$ $\beta_{0.85}$



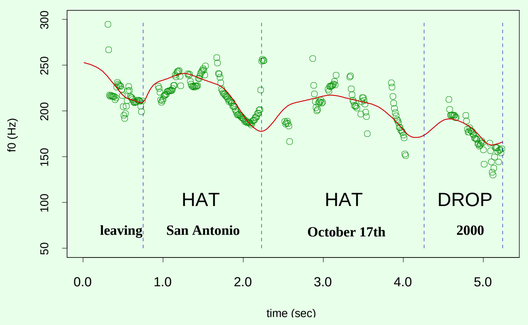
from PhoenixArizona to Bangkok Thailand.

$o_{6.65}$ $\rho_{2.1}$ $\upsilon_{3.07}$ $o_{6.92}$ $\delta_{2.19}$ $\beta_{0.85}$



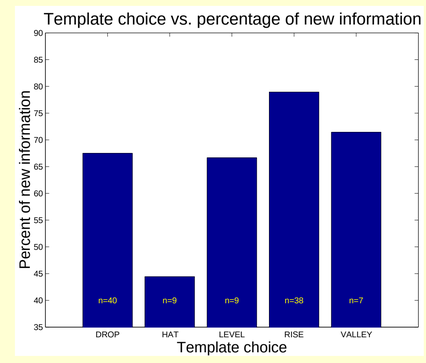
Leaving San Antonio, October 17th, 2000.

$o_{11.67}$ $\eta_{5.15}$ $\beta_{0.85}$ $\eta_{3.7}$ $\beta_{0.85}$ $\delta_{1.7}$ $\beta_{0.85}$



Template Patterns and Linguistic Events

- As conversation progresses:
  - Initially, users are often polite, give requests to system
    - Rising intonation on NPs with flight origin, destination, date, time
  - After failed system recognition, pattern changes:
    - Speakers slow down, pause more, and switch to HAT and DROP patterns
  - Consistent with other studies
- Modest correlation between f_0 patterns and frustration
  - Labeled each utterance with the frustration level of the user
    - 1=no frustration, 2=frustation, 3=extreme frustration
  - Knowing the prosodic pattern gives 0.3 bits of information in indicating frustration level
- Small correlation between f_0 patterns and new vs. old information
  - Labeled each NP with whether it contained new or old information in context of the dialogue
  - Knowing the prosodic pattern gives 0.1 bits of information in indicating new or old information
    - RISE indicates new information more than average
    - HAT indicates old information more than average
- No correlations found with strength parameters
- More data is needed to confirm these results



Implications for ASR and Dialogue Systems

- There has been significant work in integrating prosody into spoken language systgems, e.g.:
  - Detecting errors made by dialogue systems
  - Detecting user dialogue acts
- Using templates and strength parameters can provide a good basis for building these detectors
  - Mandarin experiment: can separate effects of tone and phrase curve with a simple model
    - $\Rightarrow$ can improve dialogue act detection between interogative/declarative sentences
    - Would require coordinating initial ASR word/tone hypotheses with Stem-ML model, joint best-path estimation
  - English experiment: template patterns carry information for discourse analysis
    - Choice of template can be indicative of frustration, repeated information
    - Implementation in recognition requires search over possible template patterns
    - Search space can be narrowed by first pass recognition