Automated modelling of Chinese intonation in continuous speech.

Greg Kochanski, Chilin Shih

Bell Laboratories, Lucent Technologies

{gpk,cls}@research.bell-labs.com

in Proceedings of 7th European conference on speech communication and technology (Eurospeech), Aalborg, Denmark, September 2001.

Abstract

We built and trained a model of intonation in continuous Mandarin speech based on the Stem-ML model of interacting accents. With this model, we found that we can accurately reproduce the intonation of the speaker using only one accent template for each lexical tone category. The resulting parameters are interpretable, and we find that the fitted model is consistent with linguistic expectations. Stem-ML is a phenomenological model of the muscle dynamics and planning process that controls the tension of the vocal folds. It describes the interactions between nearby tones or accents.

1. Introduction

Tonal languages, such as Chinese, use variations in pitch to distinguish otherwise identical syllables. Thus, good pitch prediction in a text-to-speech system is important not just for natural sounding speech but also for good intelligibility.

The challenge of tonal languages is that the realized f_0 contour sometimes bears little obvious relationship to the concatenation of the tones. Figure **??** shows a Mandarin phrase *fan3 ying4 su4 du4* "reaction time", along with the tones from which it is constructed [**?**]. The last three syllables are all recognized as tone 4 by native speakers, but have drastically different f_0 contours. Our model can simply explain these changes of shape.

Our view of intonation starts from a small collection of tone classes, each of which implements some linguistic function. We then calculate the surface realization of the pitch as a function of the shapes of nearby accents and their strengths. This function must be consistent with muscle dynamics (it must be smooth and continuous) and also with the neural mechanisms behind speech (pre-planning on the phrase level).

The automatic fitting is done by way of Stem-ML tags [**?**]. We parameterize a set of tags, then find the parameter values that accurately reproduce a training corpus. A positive feature of Stem-ML is that the representation is understandable, adjustable, and can be transported from one situation to another. One can even generate acceptable speech by using the templates of one speaker with parameters from another[**?**]. Parameters trained on one corpus should bear a reasonable relationship to the same parameters trained on another corpus, and also to linguistic expectations. Stem-ML also allows one to mix hand-fitting and machine fitting. Few machine learning systems have these properties.

In some previous approaches[**?**, **?**], a neural network (NN) is trained to select a segment of intonation from a large number of discrete classes. However, adding new classes obscures the relationship between the surface form and the linguistic driving information. Occam's razor [**?**] directs that "What can be done with fewer [entities] is done in vain with more." Further, NNs are, mathematically, an interpolater whose structure is unrelated to the system that they model, so one cannot expect the output of a NN to behave well in situations outside its training set.

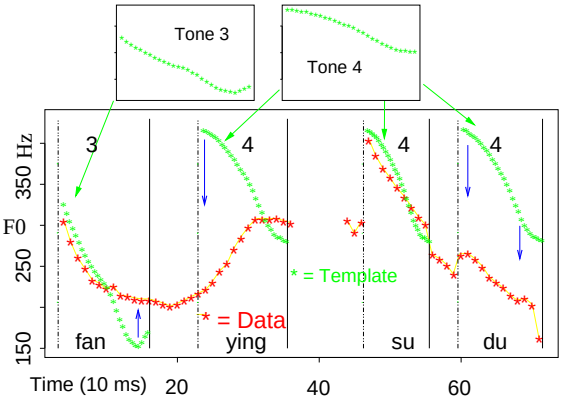

Figure 1: *Tones vs. realization. The upper panels show shapes of tones 3 and 4 taken in a neutral environment and the lower panel shows the realization of an actual sentence containing those tones. The grey curves show the templates, and the black curve shows the f_0 vs. time data.*

2. Modeling Intonation

Stem-ML introduces several ideas into intonation prediction:

- we assume that people plan their utterances several syllables in advance,
- we assume that people produce speech that is optimized to meet their needs,
- we introduce a linguistically reasonable concept of a strength that is associated with each syllable, and
- we apply a physically reasonable model for the dynamics of the muscles that control pitch [**?**].

Pre-planning in speech was first shown in terms of the control of inhaled air volume [**?**, **?**]: people will inhale more deeply when confronted with longer phrases. This fact implies that at least a rough plan for the utterance has been constructed $\approx 500ms$ before speech begins. Figure 8 in Bellegarda *et al.* [**?**] shows evidence of pre-planning of f_0, over a $\approx 1.5s$ range, at least in practiced, laboratory speech. It shows that in an upwards pitch motion, the rate of the motion is reduced as the interval lengthens, presumably to avoid running above the speaker's comfortable pitch range.

Next, we assume that speech is optimized for the speaker's purposes. A speaker has the opportunity to practice and optimize all the common 3-tone or perhaps 4-tone sequences, even

if one assumes that each tone needs to be practiced at several distinct strength levels.

The question then arises, "optimal in what sense?" We propose that optimality be defined by a balance between the ability to communicate accurately and the effort required to communicate[**?**]. Specifically that the optimal pitch curve is the one that minimizes the sum of effort plus a scaled error term. Certainly, when we speak, we wish to be understood, so the speaker must consider the error rate on the speech channel to the listener. Likewise, much of what we do physically is done smoothly, with minimum muscular energy expenditure, so minimizing effort in speech is also a plausible goal.

The error term behaves like a communications error rate: it has its minimum if the prosody exactly matches an ideal tone template, and it increases as the prosody deviates from the template. The choice of template encodes the lexical information carried by the tones. The speaker tries to minimize the deviation, because if it becomes large, the speaker will expect the listener to mis-classify the tone and possibly misinterpret the utterance.

The effort expended in speech can be approximated from knowledge about muscle dynamics [**?**]. Qualitatively, our effort term behaves like the physiological effort: it is zero if muscles are stationary in a neutral position, and increases as motions become faster and stronger. Accordingly, Stem-ML makes one physically motivated assumption. It assumes that f_0 is closely related to muscle tensions. There must then be smooth and predictable connections between neighboring values of f_0 because muscles cannot discontinuously change position. Most muscles cannot respond faster than 150ms, a time which is comparable to the duration of a syllable, so we expect the intonation of neighboring syllables to affect each other. In this sense, our model is an extension of those of [**?**, **?**, **?**].

Effort is ultimately measured in physical units, while the communication error probability is dimensionless, so a scale factor is needed to make the two compatible for addition. This scale factor varies from syllable to syllable, and we identify it with the linguistic strength, or importance of each syllable. If a syllable's strength is large, the Stem-ML optimal pitch contour will closely approximate the tone's template, and the communication error probability will be small. In other words, a large strength indicates that the speaker is willing to expend the effort to produce precise intonation. On the other hand, if the syllable is unimportant and its strength is small, the produced pitch will be controlled by other factors: neighboring syllables and ease of production. The listener then may not be able to reliably identify the correct tone on that syllable. Presumably, the listener is either able to infer the tone from the surrounding context or the speaker doesn't care if the listener can unambiguously identify the tone.

We then write simple approximations to the effort and error terms, so that the model can be solved efficiently as a set of linear equations.

3. Experiment

3.1. Data Collection

The corpus was obtained from a male native Mandarin speaker reading 423 sentences from newspaper articles, selected for broad coverage of prosodic factors. We fit two subsets (10 sentences each, 347 and 390 syllables), randomly chosen from the corpus. The speaking rate was 4 ± 1.4 syllables per second, with a phrase duration of 1.2 ± 0.7s.

Tones were identified by text analysis, and checked by two native speakers to find neutral tones. Phone and phrase boundaries were hand-segmented and were used to define syllables.

We computed f_0 with an automatic pitch tracker, then cleaned the data by hand, primarily repairing regions where the track was an octave off.

3.2. Optimization

The Stem-ML model is built by placing tags on syllables, with adjustable parameters defining the tag shapes and positions (details below). The two subsets have 388 and 431 free parameters in their respective models, or an average of 1.1 parameters per syllable.

The algorithm learned the parameters by minimizing the sum of the squared frequency difference between the data and the Stem-ML model. Unvoiced regions were excluded. We fit separately for each subset, to allow comparisons.

We used a Levenberg-Marquardt algorithm[**?**, **?**] with numerical differentiation to find the parameters that give the best fit. The algorithm required $\approx$ 30 steps before the RMS error and parameters stabilized. The *strength* parameters are local, and correlated only with a few neighbors, but there are a core of 37 global parameters that are correlated with everything.

3.3. Mandarin-specific Model

Our model for Mandarin is a more predictive, stronger model than bare Stem-ML.

We give each syllable a *strength* parameter, as described above. These strengths are the only place in our model where linguistic information can influence the f_0 contour (beyond selection of the lexical tone).

We also assume that each of the five lexical tones classes are described by one template; this template is merely stretched (in time) and shifted (in pitch) to describe all syllables which have that tone. A template is described by 5 (2 for neutral tones) pitch values, spaced across its scope, a *type* parameter, which describes whether the shape or the average pitch is more important, and *stype*, which shifts the entire accent up and down by *stype* times the syllable's strength.

There are several global parameters that affect all tones and syllables. Two parameters describe the scope of templates relative to syllable lengths: *ctrshift*, and *wscale*. Phrases are described by a straight-line phrase curve, which is controlled by three global parameters: the declination rate, the dependence of the declination on the sentence length, and the dependence of the initial height of the phrase curve on sentence length. Four other Stem-ML parameters control global properties: *adroop*, *add*, *smooth*, and *base*. To complete the model, Stem-ML *phrase* tags were placed in the centers of silences at phrase boundaries.

4. Discussion

4.1. Results of Fit

Overall, our model fits the data with a 12 Hz RMS error, or approximately 1.5 semitones. Much of that error is probably accounted for by phoneme-dependent segmental effects. In Figure **??**, we show a typical phrase, and in Figure **??**, the phrase that contains the worst-fit syllable (at $t = 1.52s$). Generally, the worst-fitting syllables tend to be the ones with the largest and fastest pitch excursions. These are conditions where Stem-ML's approximation to muscle dynamics may break down, or where

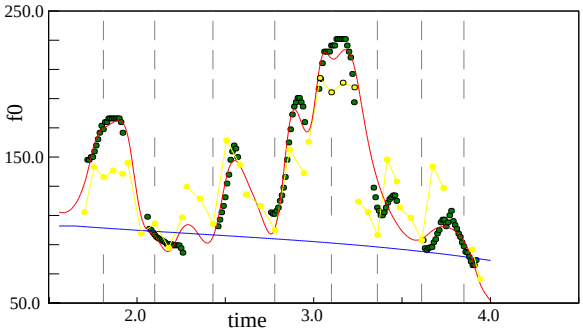

Figure 2: *Typical fit (solid) vs. data (dots). Syllable centers are marked with dashed lines, and Stem-ML templates are shown in grey. The nearly-horizontal line is the phrase curve.*

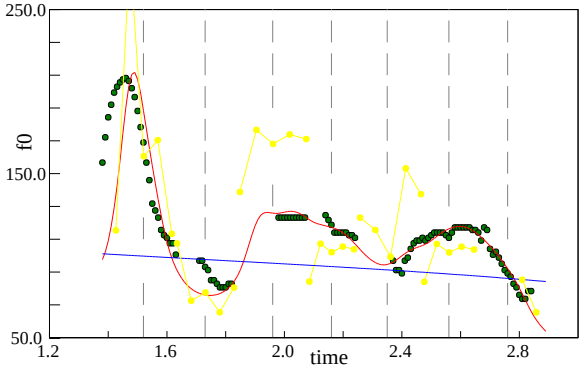

Figure 3: *Phrase containing the worst-fit syllable ($t = 1.52s$.) Displayed as above.*

the simple approximation that we use to estimate the error between templates and the realized pitch curve may be furthest from the actual perceptual metric. These figures also display the best-fit tone templates, which are similar to those found in [**?**].

In the optimization, it is found that some of the parameters are strongly correlated, so there are trade-offs among parameters that yield almost identical f_0 curves. In other words, the data should be able to be explained by fewer than the 1.1 parameters per syllable that we use here. An result of our over-parameterization is that individual parameter values become "noisier". We checked the magnitude of this problem by starting the optimizer with two different, randomly chosen sets of parameters for one of the subsets. The median absolute difference in strengths from the two repetitions is 0.67, and the median of the log of the strength ratio is 0.54. The differences are small enough that we should still be able to see some linguistic structure in the strength values, even though it will be partially hidden.

4.2. Information content of prosody

Since we have a quantitative model that accurately reproduces intonation, we can estimate the information that pitch can carry from the speaker to the listener. We do this by quantizing the strengths with different numbers of bits, and observing how much the fit degrades. Encoding strengths with 2 bits increased the RMS error by just 5% , while 1 bit per strength raised the error by 40% , and 0 bits (setting all the strengths to 1) raised the RMS error by 110% .

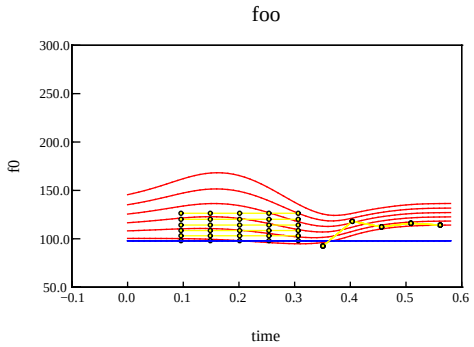

Figure 4: *Simulation showing that the best-fit tone 1 is primarily defined by its shape. The grey lines are 5 pitch curves for different levels of the initial tone; the second tone has the best-fit tone shape and parameters, with the corpus median strength of* 1.07*. Because the curves are roughly parallel, we can see that the tone shifts as a unit, and does not tie the pitch strongly to any particular value.*

While an actual measure of the information capacity requires perceptual tests to measure the probability of misinterpretation, inspection of the f_0 curves indicates that two bits per syllable is probably sufficient. This leads to a data rate of ≈ 16 bits per second for the intonation channel, including tone identity.

4.3. Analysis of parameters

For Stem-ML to be a model of a language, instead of just a scheme for efficiently coding f_0 contours, we should be able to correlate the results of the fit with linguistically important features. In the following sections, we will discuss the results of the fit and see how they correlate with linguistic expectations.

First, the fitted scope of the templates is very close to a syllable. The best fit templates are just $15\% \pm 3\%$ longer than their syllable, and their centers are offset by $6\% \pm 7\%$ after the center of the syllable. This implies that the tone is equally important at both ends of the syllable, and matches well with the intuition that tones are associated with syllables.

Next, we see that the *type* parameters are fairly small, implying that tones are defined primarily by their shape, and only weakly by their position relative to the baseline. The mean of *type* over all tone classes is 0.20. To show the implication of this, consider the end of a phrase: we can imagine pushing the pitch of the penultimate syllable up and down, and measuring how much the pitch at the tail of the last syllable changes. Figure **??** shows the relevant model. This simulation can be summarized by a transfer ratio, R, which is calculated by taking the ratio of the spread among different pitch curves at the tail of the last syllable ($t = 0.57$s) divided by the spread at the tail of the syllable before ($t = 0.34$s). The average of the end-to-end transfer ratio, R, is 60% : the tail moves a little more than half as much as the head, and is not strongly anchored to a particular pitch. Shape is most important to tone 1, with $R = 83\%$, and the average pitch is most important for tones 3 and 4, where $R = 41\%$ and 39%, respectively.

Another parameter which has a simple interpretation is *adroop*. *Adroop* controls the end-to-end transfer ratio for very weak ($strength << 1$) tones. Following the above procedure, one can construct a sequence of zero-strength tones, adjust the pitch at one end, and measure how much of that perturbation

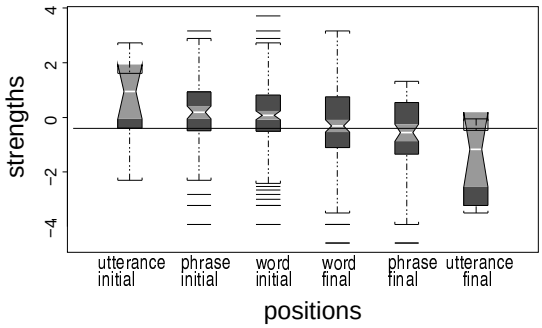

Figure 5: *Box plots of* $\log(strength)$ *of syllables in different positions. The horizontal line shows the median of the corpus. Each box shows the median,* 25^{th} *and* 75^{th} *percentiles and outlying points; notches show the 95% confidence limits for the median.*

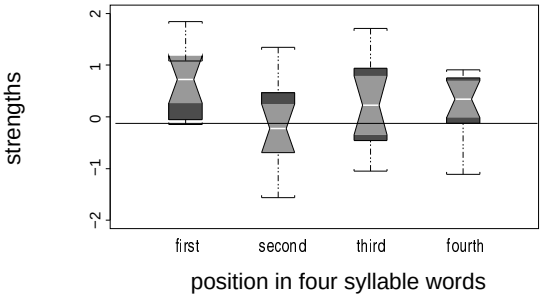

Figure 6: *Box plots of* $\log(strength)$ *values of syllables in four syllable words. The horizontal line shows the median of the corpus. Plotted as above.*

propagates across a 220 ms syllable. We get an end-to-end transfer ratio of 30% , which indicates that, while a tone may have a strong effect on it's nearest neighbor, it will have a relatively weak effect on the next-nearest-neighbor, and very little beyond that.

The fits also yield a value for the *add* parameter, and the result, $add < 0.3$ is consistent with Fujisaki's[**?**] exponential scaling.

In our model, the phrase curve is a simple linear declination model, which depends only on time and the sentence length. Both subsets have reasonable declination rates for these short phrases, $-20\text{Hz} \cdot \text{s}^{-1}$ and $-60\text{Hz} \cdot \text{s}^{-1}$.

4.4. Analysis of Strengths

The strengths that result from this fitting process can be correlated with linguistically important features. For instance, Figure **??** shows the distribution of strengths as a function of position. Word-initial, phrase-initial, and utterance-initial positions have greater strengths than average, while final positions have reduced strengths. These results are consistent with expectations [**?**].

Metrical structures in words are also apparent in the fitted strengths. Figure **??** shows strength vs. position in four syllable words. Our results are consistent with the rhythmic stress patterns in Liberman and Prince [**?**].

5. Conclusion

For more information, see http://www.bell-labs.com/project/tts/stem.html .

The simplicity and compactness with which one can describe Mandarin using this representation implies that it captures some important aspects of human behavior during speech.