Hierarchical Structure and Word Strength Prediction of Mandarin Prosody

Greg Kochanski, Chilin Shih, and Hongyan Jing
Bell Laboratories, Lucent Technologies
Murray Hill, NJ
{gpk,cls,hjing}@research.bell-labs.com

November 25, 2002

A version of this document is scheduled for publication as (Kochanski et al., 2003).

Abstract

We use Stem-ML to build an automatic learning system for Mandarin prosody that allows us to make quantitative measurements of prosodic strengths. Stem-ML is a phenomenological model of the muscle dynamics and planning process that controls the tension of the vocal folds. Because Stem-ML describes the interactions between nearby tones or accents, we were able to use a highly constrained model with only one accent template for each lexical tone category, and a single prosodic strength per word. The model accurately reproduces the intonation of the speaker, capturing 87% of the variance of the speech's fundamental frequency, f_0. The result reveals strong alternating metrical patterns in words, and suggests that the speaker uses word strength to mark a hierarchy of sentence, clause, phrase, and word boundaries.

1 Introduction

Intonation production has generally been considered a two-step process: an accent or tone class is predicted from available information, and then the accent is used to generate f_0 as a function of time. Historically, most attention has been paid to the first, high level, step of the process. We show here that by focusing on the control of the fundamental frequency, f_0, one can build a model that starts with acoustic data and reaches far enough up to predict directly from linguistic concepts.

Specifically, we present a model of Mandarin Chinese intonation that makes quantitative f_0 predictions, in terms of the lexical tones and the prosodic strength of each word. The model is able reproduce f_0 accurately in continuous Mandarin speech, with a 13 Hz RMS error. We fit this model to acoustic data and show that the strengths, tone shapes, and metrical patterns of words that result can be associated with linguistic concepts. We note that this and previous work with Stem-ML (Shih et al., 2001) may explain why Mandarin tone recognition has been so difficult to add to speech recognition systems: the tone shapes can be severely distorted by interactions with their neighbors.

Further, we will show here that parameters trained on one corpus (with a properly designed model) will match equivalent parameters trained on another corpus, and also correspond to linguistic expectations. We see effects correlated with the part of speech of words, and with the beginning and ending of the sentence, clause, phrase, and word levels of the linguistic hierarchy.

We note that this is not a typical train and test experiment, because there is no known way to adequately predict the prosodic strength of each word, based on the text. Indeed, there may not be any way to reliably predict word strengths from a reasonably small window of text: sentence focus can often be placed on any of several words in a sentence, depending on what the speaker considers important. Consequently, errors in the f_0 modeling will likely be swamped by much larger errors in trying to predict which words the speaker considers important.

Our approach is, instead, to think of the analysis (fitting Stem-ML tags to the f_0 curves) as a measurement of the prosodic strength of words in the sentence. We then check to see if the measured strengths behave in linguistically reasonable manners. If so, it is evidence (though not proof) that the strengths we measure are indeed real, and that the modeling technique captures some important aspects of the way people employ f_0 in Chinese, to transmit information.

The automatic fitting is done by way of Stem-ML tags (Kochanski and Shih, 2000; Kochanski and Shih, 2002). We parameterize a set of tags, then find the parameter values that accurately reproduce a training corpus.

2 Chinese Tones

Tonal languages, such as Chinese, use variations in f_0 to distinguish otherwise identical syllables. Mandarin Chinese has four lexical tones with distinctive shapes: high level (tone 1), rising (2), low (3), and high falling (4). The syllable *ma* means *mother* with a high level tone but *horse* with a low tone. Thus, in a text-to-speech (TTS) system, good f_0 prediction is important not just for natural sounding speech but also for good intelligibility. There is a fifth tonal category, traditionally named *neutral tone*, or tone 0, which refers to special syllables with no lexical tone assignment. The f_0 values of such syllables depends primarily on the tone shape of the preceding syllable.

Superficially, the modeling of Chinese tones seems straightforward. One

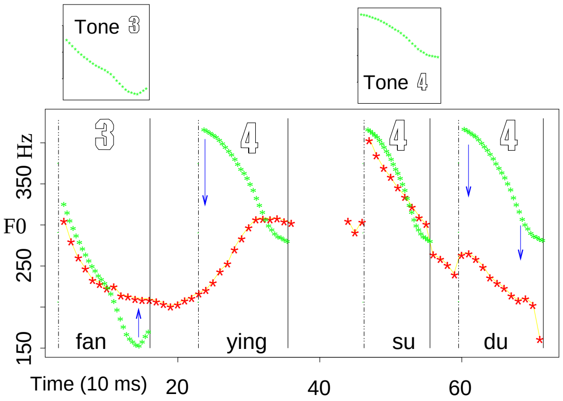

Figure 1: *Tones versus realization. The upper panels show shapes of tones 3 and 4 taken in a neutral environment and the lower panel shows the realization of an actual sentence containing those tones. The grey curves show the templates, and the black curve shows the f_0 versus time data.*

might concatenate lexical tones to generate continuous speech. The challenge is that the realized f_0 contour sometimes bears little obvious relationship to the concatenation of the tones. Figure 1 shows a Mandarin phrase *fan3 ying4 su4 du4* ("reaction time"), along with the tones from which it is constructed (Shih and Kochanski, 2000). The last three syllables are all recognized as tone 4 by native speakers, but have drastically different f_0 contours. Our model explains these changes of shape.

We explain the phenomenon displayed in Figure 1 as a natural consequence of articulatory constraints interacting with prosodic strengths. These severely distorted tone shapes occur when the shape of a weak tone is contradictory to the trajectory defined by strong neighbors. In those cases the weak tone accommodates the shapes of neighboring strong tones to maintain smooth surface f_0 contours.

Our model of Chinese intonation starts with the concatenation of lexically determined tonal templates. From these, we calculate f_0 at each time as a function of the nearby templates and their prosodic strengths.

Assuming that the lexical tone is known, the task of learning the Chinese prosody description given surface f_0 curves involves the learning of the lexical tone templates and the prosodic strengths of the templates.

3 Modeling Intonation

We build our model for Mandarin on top of Stem-ML because it captures several desirable properties. A positive feature of Stem-ML is that the representation is understandable, adjustable, and can be transported from one situation to another.

Unlike many machine-learning approaches, one can generate acceptable speech by using the templates of one speaker with parameters from another (Shih and Kochanski, 2000), where tone templates from a female speaker were used as part of a model to predict a male speaker's f_0 contours. Unlike some descriptive models, we predict numerical f_0 values, and so our model is subject to quantitative test, and can be extended to testing linguistic theories. Few other approaches to intonation have these properties.

Stem-ML introduces several ideas into intonation modeling:

- 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 apply a physically reasonable model for the dynamics of the muscles that control f_0 (Hollien, 1981), and
- we introduce a linguistically reasonable concept of a strength that is associated with each syllable.

Pre-planning in speech was first shown in terms of the control of inhaled air volume (Wilder, 1981; Winkworth et al., 1995): 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 about 500 ms before speech begins. As another example, Figure 8 in (Bellegarda et al., 2001) shows that in an upwards pitch motion, the rate of the motion is reduced as the motion becomes longer, presumably to avoid running above the speaker's comfortable pitch range. We take this as evidence for pre-planning of f_0 over a 1.5 s range, at least in practiced, laboratory speech.

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, the optimal f_0 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 the probability that the listener will misinterpret the word's lexical tone: essentially zero if the prosody exactly matches an ideal tone template, and increasing 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 (Stevens, 1998). 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 150 ms, a time which is comparable to the duration of a syllable, so we expect the intonation of neighboring syllables to affect each other. Because our model derives a smooth f_0 contour from muscle dynamics, our model is an extension of those of (Öhman, 1967; Fujisaki, 1983; Xu and Sun, 2000).

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 f_0 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 resulting f_0 will be controlled by other factors: neighboring syllables and ease of production. The listener then may not be able to identify reliably the correct tone on that syllable. Presumably, the listener either can infer the tone from the surrounding context or he/she doesn't care if the listener misidentifies the tone.

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

4 Experiment

4.1 Data Collection

The corpus was obtained from a male native Mandarin speaker reading 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. There were 265 distinct syllables, 211 counting only segmental differences (i.e., ignoring tone differences). The speaking rate was 4 ± 1.4 syllables per second, with a phrase duration of 1.2 ± 0.7 s. We define phrase as speech materials separated by a pause.

Tones were identified by automatic text analysis, including the tone sandhi rule (Shih, 1986) that converts tone 3 to tone 2 in the presence of tone 3. The tones were then checked by two native speakers. Neutral tones were manually identified. Phone, syllable, and phrase boundaries were hand-segmented, based on acoustic data.

We computed f_0 with an automatic pitch tracker, then cleaned the data by hand, primarily to repair regions where the track was an octave off. If uncorrected, the octave errors would have doubled the ultimate error of the fit, and systematically distorted tone shapes.

Because word boundaries are not marked in Chinese text, different native speakers can assign word boundaries differently. Even so, the concept of a word is present, and is reflected in the prosody. We obtained word boundaries independently from three native Mandarin speakers: A, J, and S (J and S are authors). All three had generally consistent segmentation of the text into words. Pairwise comparison indicates that J and S have the highest level of agreement: J identified 395 word boundaries, S identified 370 boundaries, 99% of which were also identified by J. A identified 359 word boundaries, of which 98% agree with J's boundaries and 92% agree with S's boundaries.

Most disagreements were related to the granularity of segmentation: whether longer units were treated as single words or multiple words, and whether neutral tone syllables were attached to the preceding words. The labelers exhibited strong and consistent personal preferences on words that could be segmented more than one way. A had the longest words, 2.04 syllables on average. J and S divided words at a finer granularity: S's words averaged 1.98 syllables, and J's words averaged 1.86 syllables per word. One labeler (A) consistently cliticized neutral tone syllables to the preceding word, while the other two labelers rarely did so.

We also created a random word segmentation (called R). The random segmentation provides a check that the metrical patterns we found are indeed significant.

4.2 Optimization

The Stem-ML model is built by placing tags on syllables, with adjustable parameters defining the tag shapes and positions (details below). We built several different models: each model has one parameter (prosodic strength) for each word, plus a set of 36, 39, or 42 shared parameters (corresponding to the "wAT", "wA", and "w" parameterizations described below). The models discussed here have between 210 and 246 free parameters, or an average of 0.6 parameters per syllable. The parameters that define the strength of words correlated only with a few neighbors, but the core group of shared parameters is correlated with everything.

The algorithm obtains the parameters's values by minimizing the RMS frequency difference between the data and the model. Unvoiced regions were excluded. We fit the two subsets separately, to allow comparisons.

We used a Levenberg-Marquardt algorithm (Levenberg, 1944; Marquardt, 1963) with numerical differentiation to find the parameters that give the best fit. The algorithm requires about 30 steps before the RMS error and parameters stabilize.

Levenberg-Marquardt, like many optimization algorithms, can become trapped in a local minimum of the Chi-squared error measure, χ^2, and may miss the global optimum. If we start the optimization with parameters randomly chosen from 'reasonable' ranges, it will converge to what we believe to be the global minimum in about 1 in 4 tries. Consequently, we believe there are only a small number of minima. The global minimum seems to be characterized by values of *adroop* < 1 (*adroop* is a Stem-ML parameter), and its χ^2 is often 10% smaller than the next best minimum. Convergence to the global minimum seems fairly reliable if an optimization is started with values of the shared parameters taken from a previous successful optimization, even if the model or data subset differ, and even if the strengths are initialized randomly.

4.3 Mandarin-specific Model

Our model for Mandarin is a more predictive, stronger model than bare Stem-ML, and is stronger even than that of (Kochanski and Shih, 2001).

The model consists simply of a Stem-ML *stress* tag on each syllable. We assume that each of the five lexical tone classes is described by one template. A template is defined by 5 (2 for neutral tones) pitch values, spaced across its scope. It is merely stretched (in time) and scaled (changing the range of f_0) to describe all syllables which have that tone. Each tone class has a Stem-ML *type* parameter. Tone classes also have an *atype* parameter, which controls how the template scaling depends on each syllable's strength. The f_0 excursions of the template are scaled by a factor $atype \cdot s_i{}^{|atype|}$ before the Stem-ML tag is generated, so that if $|atype| > 1$, the pitch range of the generated Stem-ML tag will change a lot for a small change in strength, while if $|atype| < 1$, the pitch range of the tag will be relatively independent of strength.

We give each word a *strength* parameter, S_w and derive strengths for each syllable via

$$s_{w,i} = S_w \cdot M_{L(w),i} \ , \qquad (1)$$

where $s_{w,i}$ is the strength of the i^{th} syllable of word w, $M_{L,i}$ is the metrical strength of the i^{th} position in a word of L syllables, and $L(w)$ is the length of word w. These word strengths, S_w, are the only place in our model where linguistic information can influence the f_0 contour, beyond selection of the lexical tone.

There are several parameters that are shared by all syllables. Two parameters describe the scope of templates: *ctrshift* is the offset of the template's center from the syllable's center, and *wscale* sets the length of the template relative to the syllable. Phrases are described by a straight-line phrase curve:

$$p(t) = P \cdot L - (D \cdot L^d) \cdot t \ , \tag{2}$$

where t is time, $p(t)$ is the phrase curve, and L is the length of the phrase (in seconds). All phrase curves share three parameters: D, the declination rate; d, the dependence of the declination on the sentence length; and P, which tells how the initial height of the phrase curve depends on sentence length. To complete the model, We used Stem-ML *step_to* tags to implement the phrase curve, and *phrase* tags were placed on phrase boundaries. Four other Stem-ML parameters control overall properties: *adroop*, *add*, *smooth*, and *base*.

We created and fit 24 different models to the data in a factorial design. We used two subsets of the corpus times the four different word segmentations (A, J, S, R) times three different parameterizations. We refer to the three parameterizations as "w", "wA", and "wAT". These form a nested set of models with a decreasing number of parameters. In the "w" parameterization, each tone class has its own *atype* and *type* parameters: we allow tone templates to scale differently as the strength increases, and we allow some tones to be defined by their shape while others are defined by their position relative to the phrase curve. In the "wA" parameterization, we force all tone classes to share one *atype* parameter, so that all tone templates scale with the same function of strength. Finally, in the "wAT" parameterization, we force all tones to share the *type* parameter, so all tone classes exercise the same trade-off between control of shape and control of average f_0.

5 Discussion

5.1 Results of Fit

Overall, our word-based models fit the data with a 13 Hz RMS error, or approximately 1.5 semi-tones. Much of that error may be accounted for by phoneme-dependent segmental effects. We show an entire sentence in Figure 2, then a typical phrase in Figure 3, and in Figure 4, the the phrase containing the worst-fit pair of syllables in the worst model. Generally, the worst-fitting syllables tend to be the ones with the largest and fastest f_0 excursions. These are conditions where Stem-ML's approximation to muscle dynamics may break down, or where the simple approximation that we use to estimate the error between templates and the realized f_0 curve may be furthest from the actual perceptual metric.

These models explain 87% of the variance of the data; thus, the errors between the model and the data are small compared to normal swings in f_0. Much of the remaining error may be explainable by phoneme-dependent segmental effects (Lea, 1973; Silverman, 1987). Thus, nearly all the prosodic information in the f_0 contour must be captured by the parameters we obtain from the fits.

One sentence from the automated fit of a Stem−ML model to the corpus.

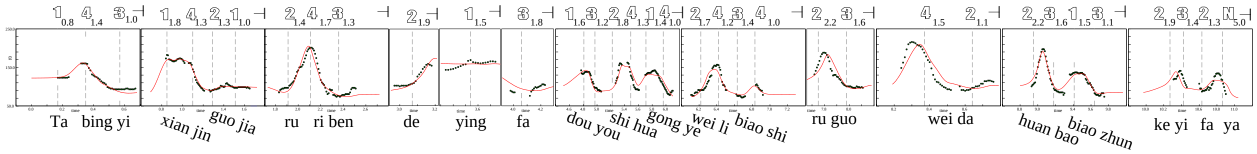

Figure 2: *Typical fit (solid) versus data (dots), for model subset1-J-A. The large, open-font numbers show the lexical tones, and their subscripts show the strengths of each syllable, and phrase boundaries are marked with "⊣". The text is shown below the plot, and syllable centers are marked with vertical dashed lines.*

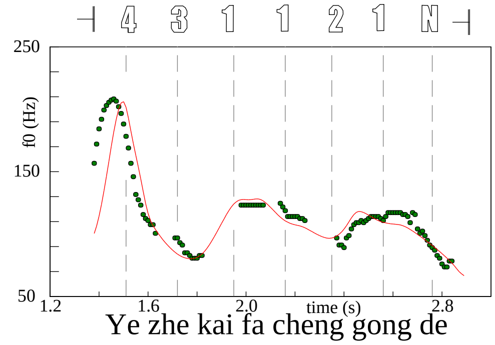

Figure 3: *Typical fit (solid) versus data (dots), for model subset1-J-A. Syllable centers are marked with vertical dashed lines.*

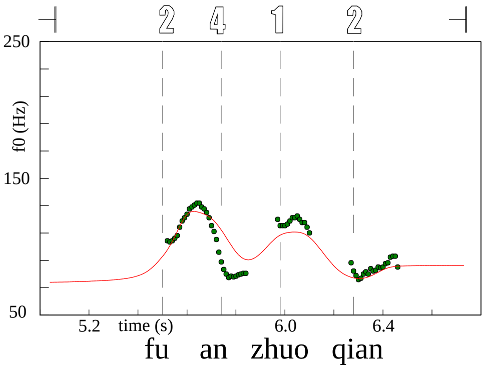

Figure 4: *Phrase containing the worst-fit pair of syllables in the worst model (subset2-S-AT). Displayed as above.*

Of the parameters, only the word strengths have localized effects so that only they can capture localized prosodic features like emphasis, focus, and marking of sentence structure. We expect, then, that the word strengths resulting from the Stem-ML analysis are nearly a complete description of Mandarin prosody. The rest of the paper will attempt to show that they are simple, useful descriptions of prosody in addition to being complete.

We can show that the strength values that we obtain are robust against small changes in the assumptions that define the model. For example, Figures 5 and 6 shows comparisons of syllable strengths obtained from different models plotted against each other. Despite the different word segmentations and the different sets of shared parameters the strength values are quite consistent. Comparisons between different models using the same segmentation are even closer. All the values fit on a narrow band about a smooth curve that maps the strength from one fit to the other. This mapping summarizes differences of shared parameters (most importantly *atype*) among the fits.

The strength values that are least reproducible are single syllable words, especially single syllable neutral tones.

5.2 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.

Our phrase curve is Equation 2: simple linear declination. We see no evidence that the phrase curve is important, and no systematic declination. Neither $P = -4(3)$ Hz·s^{-1} nor $D = 0(4)$ Hz·s^{-1} is very large, and neither is substantially different from zero (error bars are shown in parentheses, and are derived from the differences between models).

In our model of Mandarin, a positive D would correspond to a systematic decrease in f_0 during a phrase. This is distinguishable from a systematic decrease in strength, which would cause the magnitude of f_0 swings to become smaller as the phrase progresses.

5.3 Analysis of Tone Shapes

First, the fitted scope of the templates is close to a syllable. The best fit templates are just 15(5)% shorter than their syllable, and their centers are offset by 18(8)% after the center of the syllable. This matches well with the intuition that tones are associated with syllables (but see (Xu, 2001)).

Figure 7 shows the shapes of the four main Mandarin tones in isolation, calculated for each of our 24 models. The tone shapes are consistent among different models, and across subsets. Overall, the shapes match standard descriptions of Mandarin tones. The symmetry between tones 1 and 3 and tones

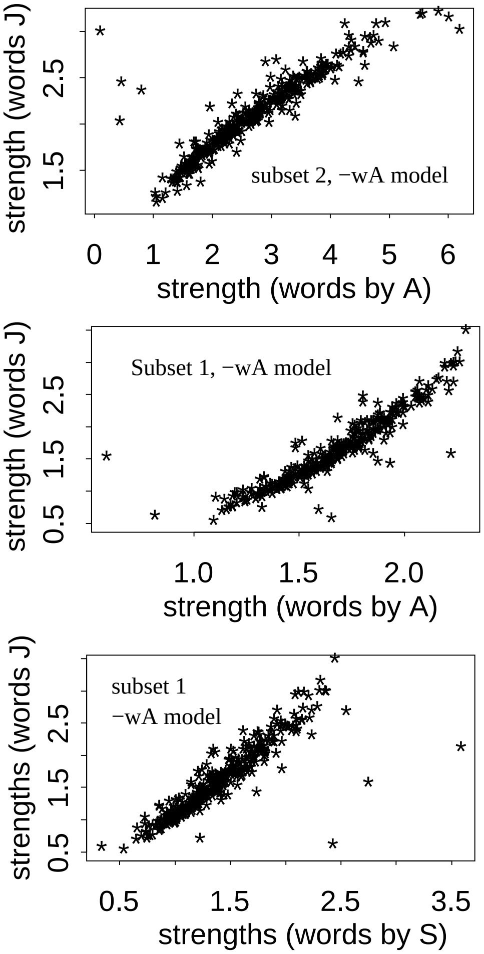

Figure 5: *Comparisons of* $\log(strength)$ *values of syllables, between models where the words are defined by different labelers.*

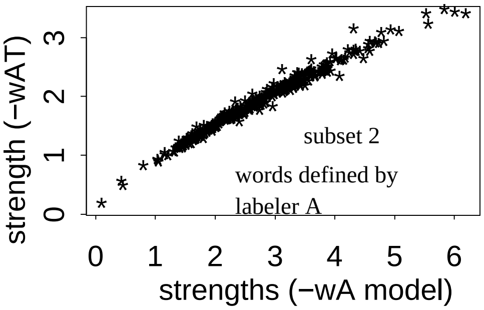

Figure 6: *Comparisons of* log(*strength*) *values of syllables between two models with different parameterizations.*

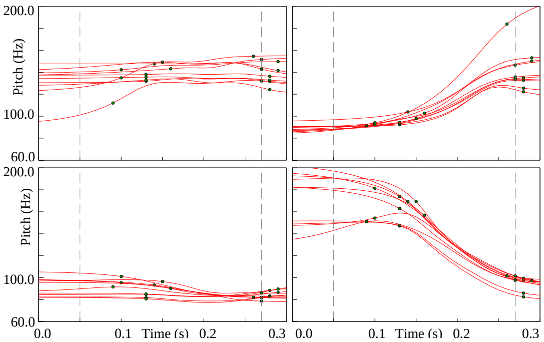

Figure 7: *Modeled shapes of isolated tones. The shapes match standard descriptions, and interact to reproduce continuous speech. The two dashed vertical bars mark the syllable boundaries, and dots mark the boundaries of the tone's template in each of the 24 models. Each tone was calculated with a strength set to the median of all the strengths in the model.*

2 and 4 is striking, and was in no way imposed by the analysis procedure. The four tones appear to have evolved to be nearly as different as possible.

5.4 Analysis of Metrical Patterns

The RMS error from these word-based models, 13 Hz, compares well with the 12 Hz RMS error we obtain from similar models (Kochanski and Shih, 2001), with nearly twice as many parameters, that allow the strength of each syllable to vary independently, and do not impose a metrical pattern. Clearly, the metrical patterns in the words are successful at capturing much of the strength variation from syllable to syllable.

Metrical structures in words are also apparent in the fitted strengths. Figure 8 shows a tree diagram of the metrical patterns we observe. Figure 9 shows

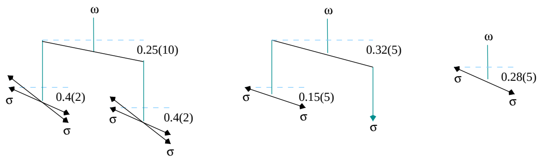

Figure 8: *Metrical patterns for the J and S segmentations of 4, 3, and 2 syllable words. The words are plotted as trees, where the height of the* i^{th} *leaf is proportional to the metrical strength of the* i^{th} *syllable:* $\log(M_{L,i}) \cdot atype^{1/2}$. *Differences of* $\log(M)$ *among leaves and nodes are shown numerically, with the parenthesized number showing the uncertainty in the last digit, as determined from the scatter among different models. The patterns for four syllable words have larger errors, as they are rare: they are drawn with double arrows to display the range of fitted solutions.*

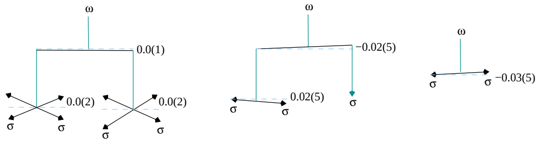

Figure 9: *Metrical patterns for random word segmentation, plotted as above. As expected, the residual patterns are weak and inconsistent.*

the corresponding pattern for a random word segmentation (R). As expected, the R-segmentation does not yield a strong metrical pattern, because there is no consistent relationship between the spoken words and the random model. Further, the R-segmentations are not as good of a fit to the data: the χ^2 for R-segmentations are 11% to 21% above the corresponding models with real (A, J, or S) segmentations. This change in χ^2 is substantial: at least an order of magnitude larger than necessary for 99% significance, even if one makes allowance for correlations among the f_0 measurements.

All the real segmentations (A, J, S), show a clear strong-weak pattern for two syllable words. This means that the initial syllable's tone is realized more precisely, and the f_0 swings will tend to be larger. Although the details are strongly dependent on the circumstances, our results indicate that RMS swings on the first syllable should be 30% larger than the second syllable. While it has

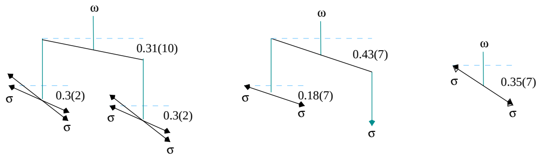

Figure 10: *Metrical patterns for the A-segmentation, plotted as above.*

been generally expected that Mandarin words would show a consistent metrical pattern, previous expectations tended more to a weak-strong pattern, based primarily on evidence from duration and perceptual judgments (Lin and Yan, 1983).

In the A, J, and S segmentations, three-syllable words are predominantly left-branching. Because of this, we applied the same metrical pattern to all three-syllable words, and did not attempt to see if words with different internal structure had different metrical patterns. Again, we see strong-weak patterns at both levels of the metrical hierarchy, though the patterns are weaker than the two-syllable case.

All of the four-syllable words could be broken up into pairs of two-syllable words. We know this from comparison of the J and S segmentations, where the primary difference was just such a splitting and from plausibility judgments of the labelers. Consequently, we adopted the metrical tree shown in Figure 8. Expressed on that tree, we again get strong-weak patterns at both levels.

In Figure 10, we show the metrical trees from the A-segmentation. While the patterns differ in detail, because of A's tendency to attach particles to words, the pattern is similar to the J and S segmentations.

Our results are broadly consistent with the alternating rhythmic stress patterns in (Liberman and Prince, 1977).

5.5 Analysis of Word Strengths

The strengths that result from the above fitting process can be correlated with linguistically important features. We considered three features: the number of syllables in the word, the position of the word in the utterance, and the part of speech of the word, and fit the strengths with a trimmed linear regression (MathSoft, Inc.i, 1995) to separate out the effects of the different factors. We then ran this regression on our models, and plotted the coefficients of the factors. We found that:

(1) **Words at the beginning of a sentence, clause, or phrase have greater strengths than words at the final positions.** Figure 11 shows the regression coefficients at different positions. We define a sentence as a grammatical utterance that is marked with a period at the end, a clause as a subset of a sentence that is marked by a comma, and a phrase as a group of words that are separated by pause. Note that the models we used (or Stem-ML) makes no distinction between clause and phrase boundaries: the fact that clause boundaries are marked more strongly emerged strictly from the data.

The hierarchy of linguistic units is displayed with strengths that increase with the size of the unit. Note that the zero line corresponds to the average of words that are not at a boundary, and that this line neatly divides the initial words of units from the final words of the units. These results are consistent with (Hirschberg and Pierrehumbert, 1986).

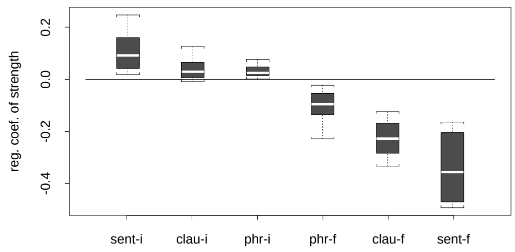

Figure 11: Correlation between strength and word positions. Each box shows the range of the data (the shaded region extends from the 25*th* and 75*th* percentiles), the median (white stripe in the box), and outlying points (brackets on the border).

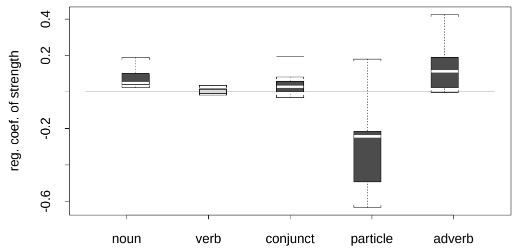

Figure 12: Correlation between part of speech and strength.

(2) **Nouns and adverbs typically have more strength than words of other parts of speech, and particles have the lowest strengths.** Figure 12 shows the regression coefficients for different part of speech. As we can see, adverbs on average have a greater strength than words of other part of speech. The strengths for nouns, verbs, and conjunctions are slightly weaker than for adverbs and their strengths are close to each other. In contrast, the strength for particles (e.g., neutral tones) are much weaker than for other parts of speech.

(3) **Words with more syllables have greater strength than words with smaller number of syllables.** Figure 13 shows the regression coefficients for strengths for words of different lengths. It indicates that 3-syllable and 4-syllable words have a larger strength value than 2-syllable words, and that multi-syllable words are stronger than 1-syllable words (the 1-syllable word average is the zero line in the figure).

Overall, predicting word strengths via this linear model reduces the median absolute deviation of the residuals by 17%. If the strength distribution were

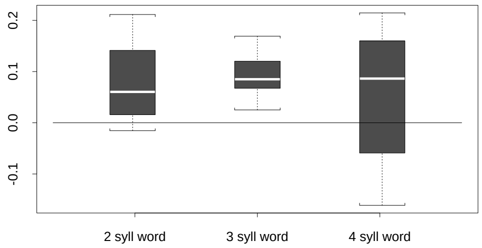

Figure 13: Correlation between strength and the number of syllables in a word.

Gaussian, this would correspond to Pearson's $r = 0.31$. We use robust estimators like a trimmed regression because the distribution of strengths has about 2% of outliers.

This linear model does not provide an accurate prediction of the strength, nor a particularly accurate prediction of f_0. However, the correlations between strength in our Stem-ML models and the above linguistic features suggest that the strengths indeed represent the prosody importance of syllables and words. On one hand, we can use the strengths from Stem-ML models to test linguistic theories; on the other hand, we can use features such as position, part of speech, and number of syllable in word to predict the strength of a word, and thus improve prediction of f_0.

6 Conclusion

We have used Stem-ML to build a model of continuous Mandarin speech that connects the acoustic level to text analysis results (part-of-speech information, and word, phrase, clause, and sentence boundaries). When fit to a corpus, the model implies that prosody may be used in a consistent way to mark divisions in the text: sentences, clauses, phrases, and words all start strong and end weak. Our prosodic measurements also show a useful correlation with word length and the part of speech of words.

The simplicity and compactness with which one can describe Mandarin using this representation suggests that it captures some important aspects of human behavior during speech. For more information, see `http://prosodies.org` .

References

Bellegarda, J., Silverman, K., Lenzo, K., and Anderson, V. (2001). Statistical prosodic modeling: from corpus design to parameter estimation. *IEEE Transactions on Speech and Audio Processing*, 9(1):52–66.

Fujisaki, H. (1983). Dynamic characteristics of voice fundamental frequency in speech and singing. In MacNeilage, P. F., editor, *The Production of Speech*, pages 39–55. Springer-Verlag.

Hirschberg, J. and Pierrehumbert, J. (1986). The intonational structuring of discourse. In *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics*, volume 24, pages 136–144.

Hollien, H. (1981). In search of vocal frequency control mechanisms. In Bless, D. M. and Abbs, J. H., editors, *Vocal Fold Physiology: Contemporary Research and Clinical Issues*, pages 361–367. College-Hill Press, San Diego, CA.

Kochanski, G. and Shih, C. (2001). Automated modelling of Chinese intonation in continuous speech. In *Proceedings of Eurospeech 2001*, Aalborg, Denmark. International Speech Communication Association.

Kochanski, G. and Shih, C. (2002). Prosody modeling with soft templates. http://prosodies.org/papers/SpeechComm1_2001.pdf. Accepted for publication in Speech Communication, 2003.

Kochanski, G., Shih, C., and Jing, H. (2003). Hierarchical structure and word strength prediction of Mandarin prosody. *International Journal of Speech Technology*, 6:33–43.

Kochanski, G. P. and Shih, C. (2000). Stem-ML: Language independent prosody description. In *Proceedings of the International Conference on Spoken Language Processing*, volume 3, pages 239–242, Beijing, China.

Lea, W. (1973). Segmental and suprasegmental influences on fundamental frequency contours. In Hyman, L., editor, *Consonant Types and Tones*, pages 15–70. University of Southern California, Los Angeles.

Levenberg, K. (1944). A method for the solution of certain problems in least squares. *Quart. Applied Math.*, 2:164–168.

Liberman, M. Y. and Prince, A. (1977). On stress and linguistic rhythm. *Linguistic Inquiry*, 8:249–336.

Lin, M.-C. and Yan, J. (1983). The stress pattern and its acoustic correlates in Beijing Mandarin. In *Proceedings of the 10th International Congress of Phonetic Sciences*, pages 504–514.

Marquardt, D. (1963). An algorithm for least-squares estimation of nonlinear parameters. *SIAM J. Applied Math*, 11:431–441.

MathSoft, Inc.i (1995). *Splus Online Documentation*, 3.3 edition. Subroutine *ltsreg()*, set to exclude the 5 most extreme data points from the objective function.

Öhman, S. (1967). Word and sentence intonation, a quantitative model. Technical report, Department of Speech Communication, Royal Institute of Technology (KTH).

Shih, C. (1986). *The Prosodic Domain of Tone Sandhi in Chinese.* PhD thesis, University of California, San Diego.

Shih, C. and Kochanski, G. (2000). Chinese tone modeling with stem-ml. In *Proceedings of the International Conference on Spoken Language Processing*, Beijing, China. See http://prosodies.org/papers/tonemodel_2000.pdf.

Shih, C., Kochanski, G. P., Fosler-Lussier, E., Chan, M., and Yuan, J.-H. (2001). Implications of prosody modeling for prosody recognition. In Bacchiani, M., Hirschberg, J., Litman, D., and Ostendorf, M., editors, *Proceedings of the ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding*, pages 133–138. International Speech Communication Association. Red Bank, NJ.

Silverman, K. E. (1987). *The Structure and Processing of Fundamental Frequency Contours.* PhD thesis, University of Cambridge, UK.

Stevens, K. N. (1998). *Acoustic Phonetics.* The MIT Press.

Wilder, C. N. (1981). Chest wall preparation for phonation in female speakers. In Bless, D. M. and Abbs, J. H., editors, *Vocal Fold Physiology: Comtemporary Research and Clinical Issues*, pages 109–123. College-Hill Press, San Diego, CA. ISBN 0-933014-87-2.

Winkworth, A. L., Davis, P. J., Adams, R. D., and Ellis, E. (1995). Breathing patterns during spontaneous speech. *Journal of Speech and Hearing Research*, 38(1):124–144.

Xu, Y. (2001). Pitch targets and their realization: Evidence from Mandarin Chinese. *Speech Communication*, 33:319–337.

Xu, Y. and Sun, X. J. (2000). How fast can we really change pitch? maximum speed of pitch change revisited. In *Proceedings of the Sixth International Conference on Spoken Language Processing (ICSLP)*, Beijing, China.