Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT®
Speech Communication 41 (2003) 625–645
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Quantitative measurement of prosodic strength in Mandarin

Greg Kochanski *, Chilin Shih [1], Hongyan Jing [2]

Bell Laboratories, Murray Hill, NJ, USA

Received 2 July 2003; received in revised form 2 July 2003; accepted 3 July 2003

Abstract

We describe models of Mandarin prosody that allow us to make quantitative measurements of prosodic strengths. These models use Stem-ML, which is a phenomenological model of the muscle dynamics and planning process that controls the tension of the vocal folds, and therefore the pitch of speech. Because Stem-ML describes the interactions between nearby tones, we were able to capture surface tonal variations using a highly constrained model with only one 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 f_0 with these strength parameters. The result reveals alternating metrical patterns in words, and shows that the speaker marks a hierarchy of boundaries by controlling the prosodic strength of words. The strengths we obtain are also correlated with syllable duration, mutual information and part-of-speech.

Keywords: Intonation; Tone; Tonal variation; Prosodic structure; Metrical pattern; Prosodic strength; Prosody modeling; Muscle dynamics; Text-to-speech

1. Introduction

Intonation production is generally considered a two-step process: an accent or tone class is predicted from available information, and then the tone class 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 here show that by focusing on f_0 generation, one can build a model that starts with acoustic data and reaches far enough up to connect directly to linguistic factors such as part-of-speech, word length and position in the text.

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 to generate tonal variations from a few tone templates that correspond to lexical tones, and accurately reproduce f_0 in continuous Mandarin speech with a 13 Hz RMS error. The result is comparable to machine learning systems that may use more than one hundred tone templates to account for Mandarin tonal variations.

We find that some parameters of the model can be interpreted as the prosodic strength of a tone.

* Corresponding author. Address: Phonetics Laboratory, Oxford University, 41 Wellington Square, Oxford OX1 2JF, UK. Tel.: +44-1865-270446.

E-mail addresses: gpk@kochanski.org, greg.kochanski@phon.ox.ac.uk (G. Kochanski).

[1] Present address: University of Illinois, Urbana-Champaign, IL, USA.

[2] Present address: IBM, T.J. Watson Research Center, Yorktown Heights, NY, USA.

0167-6393/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/S0167-6393(03)00100-6

We determine the prosodic strengths (and the values of the other global parameters) by executing a least-squares fit of the model to the time-series of f_0 from a corpus of speech data. The resulting best-fit strengths, tone shapes, and metrical patterns of words can be associated with linguistic properties. We show that strengths computed from the model exhibit strong and weak alternation as in metrical phonology (Liberman and Prince, 1977), and the values are correlated with the part-of-speech of words, with mutual information, and with the hierarchy of the prosodic structure (Ladd, 1996; Pierrehumbert and Beckman, 1988; Selkirk, 1984) such as the beginning and ending of sentences, clauses, phrases, and words.

We will also show that values of parameters from a fit to one half of the corpus match equivalent parameters fit to the other half of the corpus. Further, we can change the details of the model, and show that the values of many parameters are essentially unaffected by the change. This consistency is important because if we hope to interpret these parameters (and thus the models that contain them) as statements about the language as a whole, they must at least be consistent across the corpus and between similar models.

The model we use is described in Section 3. It is written in Soft Template Mark-up Language (Stem-ML) (Kochanski and Shih, 2003; Kochanski and Shih, 2000), and depends upon its underlying mathematical model of prosody control. We write a Stem-ML model in terms of a set of tags (parameters) then find the parameter values that best reproduce f_0 in a training corpus. Fitting the model to the data can be done automatically.

Stem-ML calculates an intonational contour from a set of tags. Some of the tags set global parameters that correspond to speaker characteristics, such as pitch range, while others represent intonational events such as lexical tone categories and accent types. The tags can contain adjustable parameters that can explain surface variations.

Stem-ML does not impose restriction on how one define tags. In our view, a meaningful way is to use the tags to represent linguistic hypotheses such as Mandarin lexical tones, or English accent types. We call tags that define tones or accents *templates* because they define the ideal shapes of f_0 in their vicinity. In this paper, our usage of tone tags (tone templates) corresponds directly to Mandarin lexical tone categories, and we interpret the Stem-ML *strength* parameters as the prosodic strengths of these tone templates. The actual realization of f_0 depends on the templates, their neighbors, and the prosodic strengths. We show in the paper that this treatment successfully generates continuous tonal variations from lexical tones.

Described another way, a Stem-ML model is a function that produces a curve of f_0 vs. time. The resulting curve depends on a set of adjustable (free) parameters which describe things like the shape of tones, how tones interact, and the prosodic strength of each syllable. When Stem-ML is generating a f_0 curve, one can set these parameters to any values, and each setting will get you a different curve. In reverse, one can find the best values for the parameters via data fitting procedures.

We use a least-squares fitting algorithm to find the values for the parameters that best describe the data. The algorithm operates iteratively by adjusting the parameter values, and accepting steps that reduce the sum of the squared differences between the model and the data. The values of the parameters that make the summed squared difference as small as possible, for a given model, are called the best-fit (or fitted) parameters.

2. Chinese tones

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

Superficially, modeling Chinese tones seems straightforward. One might concatenate lexical tones to generate continuous speech. The challenge is that tone shapes vary in natural speech to the extent that the realized f_0 contour sometimes bears no obvious relationship to the concatenation of the tones. Fig. 1 shows a Mandarin phrase *fan3 ying4 su4 du4* (“reaction time”), along with the tones from which it is constructed (Shih and Kochanski, 2000; Shih and Sproat, 1992). The last three syllables are all recognized as tone 4 by native speakers, but have drastically different f_0 contours. The second syllable *ying4* has an inverted tone shape while the last syllable *du4* is lower than expected.

In previous Chinese intonation generation models, variations of a lexical tone are either ignored, or are treated as discrete classes. These discrete classes may be linked to the lexical tone by rules (Lee et al., 1993; Shih, 1988), or by a machine learning method such as a neural network (Chen et al., 1992; Chen et al., 2000). It is not uncommon for these systems to use up to a hundred discrete classes to represent tonal variations. Both rule-based and machine learning methods link the lexical tone and their surface forms in an ad hoc manner, using factors such as lexical tones, tonal contexts, and positions in the sentence, yet neither method offers an explanation of the relations between lexical tone and their variations, or the relationship among discrete classes.

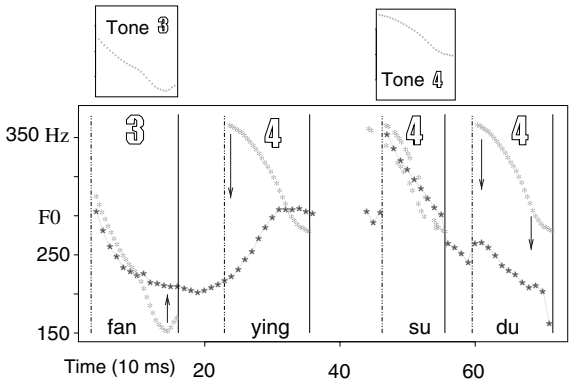

Fig. 1. Tones vs. realization. The upper panels show shapes of tones 3 and 4 taken in a neutral environment and the lower panel shows the realization of an actual sentence containing those tones. The grey curves show the templates, and the black curve shows the f_0 vs. time data.

We explain the phenomenon displayed in Fig. 1 as a natural consequence of tone shapes interacting via articulatory constraints. 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 a linguistically reasonable assumption: that all tonal variations of a lexical tone are generated from the lexically determined tonal templates. From these, we calculate f_0 at each time point as a function of the nearby templates and their prosodic strengths. We will show that this conceptually simple representation is capable of capturing the drastic tonal variations such as shown in Fig. 1.

Given surface f_0 curves, and assuming that the lexical tone is known, learning the Chinese prosody description reduces to learning 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 (Kochanski and Shih, 2003) 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 most engineering approaches, this model cleanly separates into local (word-dependent) and global (speaker-dependent) parameters. For instance, one can generate acceptable speech by using the templates of one speaker with prosodic strengths from another (Shih and Kochanski, 2000), where a female speaker’s tone templates 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. Few other approaches to intonation have all these properties.

3.1. Concepts behind the model

Stem-ML brings together 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 pitch and skilled movements (Hollien, 1981; Nelson, 1983), and
- we introduce the concept of prosodic strength, a continuous parameter associated with linguistic units such as syllable, tone, and word, to control variations.

Pre-planning in speech was first shown in terms of the control of inhaled air volume (McFarland and Smith, 1992; Whalen and Kinsella-Shaw, 1997; Wilder, 1981; Winkworth et al., 1995): people will inhale more deeply when confronted with longer phrases, hence we see a positive correlation of longer phrase and higher initial f_0 (Shih, 2000). This fact implies that at least a rough plan for the utterance has been constructed about 500 ms before speech begins. As another example, Fig. 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. The idea of representing muscle motions as the solution to an optimization problem has been developed in the biomechanics literature (Hogan and Winters, 1990; Seif-Naraghi and Winters, 1990; Zajac and Winters, 1990), and there have been comparisons of these models to actual movements (Flash and Hogan, 1985) and to electromyogram signals (Crowninshield and Brand, 1981). Nelson (1983) modeled jaw movement and arm movement during violin bowing and showed that skilled movements are influenced by minimum-cost solutions which balance performance trade-offs between competing objectives.

Speech is a skilled movement, and native speakers of Mandarin are skilled practitioners of tonal production. A speaker of Mandarin 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. For instance, if we count tone N-gram in the ROCLING Chinese Corpus (1993), we find that the most common 64 of the 179 tone 3-gram cover 90% of the corpus (we count phrase boundaries in the N-gram). Likewise, the most common 358 of the 881 4-gram cover 90% of the corpus. A speaker could practice the common tonal combinations in an hour of speech.

A more realistic model, such as the one we propose in this paper, would add a strength parameter to each tone, but one could then still expect to practice the common tonal combinations with several levels of strength in a short time.

The question then arises, "optimal in what sense?" It has been proposed that optimality be defined by a balance between the ability to communicate accurately and the effort required to communicate (Kochanski and Shih, 2003; Kochanski and Shih, 2000; Ohala, 1992), and such models have been applied by ourselves (Kochanski et al., 2003; Kochanski and Shih, 2000; Shih and Kochanski, 2001) and others (Perkell and Zandipour, 2002; Perkell et al., 2002).

Our works extend the concept of optimizing communication needs and the ease of articulatory efforts to account for tonal variations in continuous speech (Kochanski and Shih, 2003; Kochanski and Shih, 2000). 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. Different from most previous works, our view is that the trade-off relations between different objectives change dynamically during continuous speech. We introduce a scale

factor (the prosodic strength) to describe the shifting dynamics of how the speaker optimizes communication needs and articulatory efforts in continuous speech.

3.2. Mathematical definition of model

The assumption that pitch is produced to optimize the sum of effort plus error can be converted into a quantitative mathematical model. We will describe the equations below, and the variables involved will be defined in Table 1.

The effort expended in speech, G (Eq. (1)), is based upon the literature on muscle dynamics and energetics (Flash and Hogan, 1985; Stevens, 1998; Winters, 1990; Zahalak, 1990; Zajac, 1989). 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. Minimizing G tends to make the pitch curve smooth and continuous, because it minimizes the magnitude of the first and second derivatives of the pitch.

Note that we do not depend on the assumption that the effort term is an actual measurement of the energy expenditure in the muscle. The effort term could very well be a measure of competition for resources in the central nervous system, could be due to neural feedback loops local to the muscle (similar to the Equilibrium Point Hypothesis (Feldman et al., 1990; Laboissiè et al., 1996)) or could be entirely phenomenological. It does seem, however, that the effort term is not just a way to express the non-zero response time of a muscle fiber: measurements of single-fiber twitches (i.e., the force vs. time curve of a single muscle fiber triggered by a single nerve impulse) show a contraction time of ≈19 ms (MacNeilage et al., 1979), which is too short to account for inverted tone shapes and other phenomena that can last for 100 ms or more.

Table 1
Definitions of parameters and variables used in this paper

Symbol	Location	Meaning
add[a]	Eq. (6)	Controls the mapping between e and f_0. See $g(\ )$
adroop[a]	Eq. (1)	Rate at which e droops toward the phrase curve in the absence of a tag
base[a]	Eq. (6)	The speaker's relaxed f_0
smooth[a]	Eq. (1)	Response time of muscles
type[a]	Eq. (3)	Is tone defined by its shape (0) or f_0 value (1)
$M_{L,i}$	Eq. (8)	Metrical pattern of the ith syllable in a L syllable word
s_k[a]	Eqs. (2), (7) and (8)	Strength of syllable k
S_w	Eq. (8)	Strength of word w
atype	Eq. (7)	Controls how the size of the template depends on the strength of a syllable
ctrshift	Section 4.3	Position of center of template relative to center of syllable
wscale	Section 4.3	Width of a tone template, relative to a syllable
P, D, d	Eq. (9)	Parameters defining the phrase curve
f_0	Many places	Measured pitch
$\hat{f}_0$	Eq. (6)	Modeled pitch
p[a]	Eq. (9)	Phrase curve
e[a], e_t	Section 3.2	Emphasis, i.e., $\hat{f}_0$ relative to the speaker's range
$\bar{e}$[a]	Eqs. (3) and (4)	Mean emphasis over the scope of a tag
y[a], y_t	Section 3.2	Tone template
$\bar{y}$[a]	Eqs. (3) and (5)	Mean value of a tone template
G[a]	Eq. (1)	Effort expended in realizing the pitch contour
r_i	Eq. (3)	The summed error for syllable i between the template and the realized pitch
R[a]	Eq. (2)	The summed error for an utterance between the ideal templates and the realized pitch contour
$g(\ )$[a]	Eq. (6)	Function to map between subjective emphasis (e) and objective f_0

[a] Parameters defined more fully in (Kochanski and Shih, 2003).

The error term, R (Eqs. (2) and (3)), behaves like a communications error rate: it is zero 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 too large, the speaker will expect the listener to misclassify the tone and possibly misinterpret the utterance.

Stem-ML makes one physically motivated assumption. It assumes that f_0 is closely related to muscle tensions (Monsen et al., 1978). 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), Lindblom (1963), and Moon and Lindblom (1994), and is similar to that of Xu and Sun (2000).

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 have a parameter, *type*, which controls whether errors in the shape or average value of the pitch curve are most important. In this work, the targets, y, consist of a tone component riding on top of the phrase curve, p.

In order to efficiently solve the optimization problem, and calculate the surface realization of prosody, we write simple approximations to G and R so that the model can be solved efficiently as a set of linear equations:

$$G = \sum_t \dot{e}_t^2 + (\pi \cdot \mathrm{smooth}/2)^2 \ddot{e}_t^2 + \mathrm{adroop}^2 \cdot e_t^2, \tag{1}$$

$$R = \sum_{k \in \mathrm{tags}} s_k^2 r_k, \tag{2}$$

$$r_k = \sum_{t \in \mathrm{tag}\, k} \cos(\mathrm{type} \cdot \pi/2)(e_t - y_{k,t})^2 + \sin(\mathrm{type} \cdot \pi/2)(\bar{e}_k - \bar{y}_k)^2, \tag{3}$$

where

$$\bar{e}_k = \sum_{t \in \mathrm{tag}\, k} e_t \Big/ \sum_{t \in \mathrm{tag}\, k} 1 \tag{4}$$

and

$$\bar{y}_k = \sum_{t \in \mathrm{tag}\, k} y_t \Big/ \sum_{t \in \mathrm{tag}\, k} 1. \tag{5}$$

Finally, f_0 is e, scaled to the speaker's pitch range:

$$\hat{f}_0 = g(e, \mathrm{add}) \cdot \mathrm{range} + \mathrm{base}, \tag{6}$$

the scaling allows p and e to be dimensionless quantities, typically between 0 and 1. The function $g(\,)$ handles linear (add = 1) or log (add = 0) scaling, and has the properties that $g(e, 1) = e$ for any e, and that $g(0, \mathrm{add}) = 0$, and $g(1, \mathrm{add}) = 1$ for any *add*.

Fig. 2 shows how the G (effort) term depends on the shape of e. The curves we show all go through the same set of pitch targets (dashed circles). The G values increase with the RMS curvature and slope of e. In this case, optimal pitch curve has the smallest value of G, G_1.

Note that there are two distinct optimizations in this paper, and they should not be confused. First (Section 3.2), we represent the Stem-ML model as an optimization problem, minimizing *effort* + *error* to find f_0 as a function of the model parameters. This first minimization is actually done analytically, to convert the Stem-ML model into a set of linear equations that are solved by matrix techniques.

Second (Section 4.2), we adjust the parameters to minimize the difference between the model and the data. This gives us best-fit values for the parameters that best describe the data. This second minimization treats the evaluation of the Stem-ML model as a black box, calculating many models to find the best-fit.

As an additional complication, we then take some of the best-fit parameter values (specifically

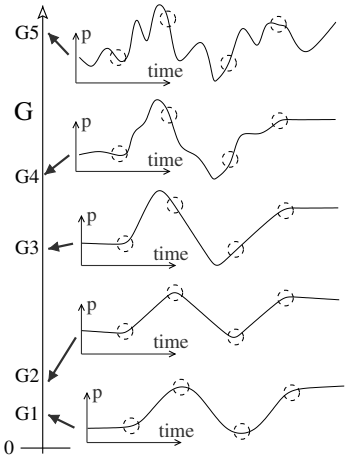

Fig. 2. Schematic diagram showing the dependence of G on the shape of the pitch curve. The large, left axis shows values of G (speech effort) for each of the displayed curves ($G_1, \ldots, G_5$). Each small axes show sample curves of pitch as a function of time. The resulting Stem-ML pitch curve is the one with the optimal (smallest) value of $G + R$. Because we have chosen $R = 0$ in this example, the solution here is G_1, the one with the smallest G.

the values of the prosodic strength parameters) and fit *them* with an additive linear model (Section 5.5). This final fit helps us to understand which linguistic features have the most influence on the strength of a syllable.

3.3. Prosodic strength

Effort is ultimately measured in physical units, while the communication error probability is dimensionless. Since one can only sensibly add numbers with the same units (e.g., 1 kg + 1 m = ?), a scale factor is needed to convert one into the units of the other. This scale factor, s_k (in Eq. (2)), can vary from a tone, a syllable or a word to the next, and we identify it with the prosodic strength.

If a syllable's strength is large, the Stem-ML 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 enough effort to produce precise intonation on a syllable. On the other hand, if the syllable is deaccented and its strength is small, the produced pitch will be controlled by other factors: neighboring syllables and ease of production. For prosodically weak syllables, minimizing the effort term will have the most effect: when s_k is small, smoothness becomes more important than accuracy. The listener then may not be able to reliably identify the correct tone on that syllable. Presumably, the listener can infer the tone from the surrounding context.

The concept that strength is related to how carefully speech is articulated was discussed by Browman and Goldstein (1990), in the context of phoneme changes in casual speech. Flemming (1997, 2001) discusses optimization models with continuous parameters (into which class this model falls), and their relationship with Optimality Theory (Prince and Smolensky, in press).

Traditionally, prosodic strength is expressed as abstract categories S (strong) and W (weak) in metrical phonology (Liberman and Prince, 1977), where one of the goals is to capture the rhythmic alternation in natural sentences even though words typically do not come in iambic or trochaic pairs. One can build a prosodic structure with strong and weak nodes to describe sentence prosody in relative terms.

Our model is related to conventional views of accents and intonation, except that we consider strength to be a continuous parameter associated with a word or a syllable. We suggest that listeners *might* treat strong tones as categorically different from weak tones, so these strength measurements *might* be equivalent to the presence or absence of accents (strong implies present). The strength numbers are associated with a particular rendition of the sentence. They vary somewhat even among utterances that were spoken with the same intent, but they seem to vary more between utterances where the sentence focus, the intonation type, or other prosodic features differ.

4. Experiment

4.1. Data collection

The corpus was obtained from a male native Mandarin speaker reading paragraphs from

newspaper articles, selected for broad coverage of factors in the text that are associated with prosodic effects, including tonal patterns in the beginning, medial, and final positions of utterances, phrases, and words. To select sentences from a corpus, we used the greedy algorithm described in (van Santen and Buchsbaum, 1997). Pause and emphasis were transcribed manually after text selection and recording. A complete description of the factors, procedures, and evaluation of the algorithm were described in (Shih and Ao, 1997).

We fit two subsets (10 sentences each, 347 and 390 syllables), that were randomly chosen from the corpus. The speaking rate was 4 ± 1.4 syllables per s, with a phrase duration of 1.2 ± 0.7 s. We define a phrase as speech materials separated by a perceptual pause. We measured these pauses acoustically, and found that the speech power dropped by at least 10 dB relative to a 50 ms window on either side in 94% of the pauses, and the median duration of pauses was 240 ms.

Tones were identified by automatic text analysis, including the tone sandhi rule in (Shih (1986)), then checked by two native speakers. Neutral tones were manually identified prior to fitting, because they cannot be reliably identified from a dictionary. Phone, syllable, and phrase boundaries were hand-segmented, based on acoustic data.

We computed f_0 with an automatic pitch tracker (Talkin and Lin, 1996), 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 a 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% were also marked by J and 92% were also marked by S.

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. Labeler 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. 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 (Section 5.4) we found are indeed significant.

4.2. Fitting

The Stem-ML model is built by placing tags on tone templates, with adjustable parameters defining the tag shapes and positions (details below). We built several different models, focusing on models with one parameter (prosodic strength) for each word, plus a set of 36, 39, or 42 shared parameters. 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 are correlated only with a few neighbors, but the shared parameters are 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 parameter values that give the best-fit. The algorithm requires about 30 iterations before the RMS error and parameter values stabilize.

Levenberg–Marquardt, like many data fitting algorithms, can become trapped in a local minimum of χ^2, and may miss the global best-fit. If we start the fit with parameter values randomly cho-

sen from "reasonable" ranges, it will converge to what we believe to be the global minimum in about one in four tries. Consequently, we believe there are only a small number of minima. The global minimum seems to be characterized by values of *adroop* < 1 (see Table 1), and its χ^2 is often 10% smaller than the next best minimum. Convergence to the global minimum seems fairly reliable if a fit is started with values of the shared parameters taken from a previous successful fit, even if the model or data subset differs, 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 than our previous works on Mandarin tone modeling (Kochanski and Shih, 2001) where an independent strength parameter is fitted for every syllable.

The current model, which is an extension of (Kochanski et al., 2003), starts with a Stem-ML *stress* tag specifying the lexical tone templates associated with the syllable. The syllabic strength is tied to the strength of the word via metrical patterns. This model fits less parameters but still achieve comparable results.

We assume that each of the five lexical tone classes is described by one template. A template is defined by five (two for neutral tones) pitch values, spaced across its scope. It is merely stretched (in time) and scaled (changing its pitch range) 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 pitch excursions of the template on syllable k are scaled by a factor

$$F_k = \text{atype} \cdot s_k^{|\text{atype}|}, \tag{7}$$

before the Stem-ML tag is generated. Thus, if $|\text{atype}| > 1$, the pitch range of the generated Stem-ML tag will change a lot for a small change in strength, while if $|\text{atype}| < 1$, the pitch range of the tag will be relatively independent of strength.

In the general Stem-ML model, each tone template has a strength value, which controls how it interacts with its environment. In a pitch generation process this gives us enough parameters to describe a pitch contour (Kochanski and Shih, 2003; Shih and Kochanski, 2000). In the reverse process of fitting the strength values from data (Kochanski and Shih, 2001), we found that the data cannot support the estimation of one parameter per syllable and that the fitting process was often trapped in a local minima. Increasing the size of the database would not help to disambiguate syllable strengths, since the number of strength parameters to be estimated increases with the number of syllables in the database.

However, we noticed that syllable strength within a word is not independent of each other, and that they tend to exhibit alternating metrical patterns. If there are consistent strength patterns within a word, then we should be able to describe the observed prosody with word-level strength and a few metrical patterns. In the current model, we allow words of different length to have different metrical patterns. This turns out to be a viable method. Compared to the syllable model, the word model reduces the number of parameters by 40% while maintaining a very good fit.

In the model, each word has its own adjustable *strength* parameter, S_w, and we derive strengths for each syllable via

$$s_{w,i} = S_w \cdot M_{L(w),i}, \tag{8}$$

where $s_{w,i}$ is the strength of the ith syllable of word w, $M_{L,i}$ is the metrical strength of the ith position in a word of L syllables, and $L(w)$ is the length of word w. That means we allow the strengths of words to vary independently [3] while restricting the strength relationship of syllables within the word. Each word is associated with a word strength and the strengths of the component syllables are

[3] One alternative to the assumption that each word has its own strength parameter would be to assume that (for example) all sentence-initial words have the same strength. Instead, we chose to let each word have its own strength so that we could search for relationships among the strengths we obtain by fitting the model to a corpus of data.

derived from the word strength and the metrical pattern. This metrical pattern is assumed to be the same for all words that have the same number of syllables. The 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. In Section 4.2, the word strengths will be adjusted to fit the model to the data.

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:

$$p(t) = P \cdot L - (D \cdot L^d) \cdot t, \tag{9}$$

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 a set of different models to the data, using 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 pitch.

Of these 24 models, 15 converged to comparably small χ^2 values, and we believe those sets of parameters to be globally optimal for their model. Of the remainder, several were not attempted, due to limits on the available CPU time, and the rest seemed to land on local minima, with χ^2 values more than 30% larger than the global minimum.

5. Analysis of best-fit parameters

5.1. Results of fit

Overall, our word-based models fit the data with a 13 Hz RMS error, approximately 1.5 semitones. In Fig. 3, we show the beginning of an utterance from the best-fit model (subset1-J-wA). In Fig. 4, we show the phrase with median error from that model, and in Fig. 5, the phrase containing the worst-fit pair of syllables in the worst of the converged models (subset2-S-wAT). 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 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 models explain 87% of the variance of the data, and much of the rest may be explainable by phoneme-dependent segmental effects (Lea, 1973; Silverman, 1987). Thus, essentially all the prosodic information in the f_0 contour is captured by the parameter values we obtain from the fits. 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. [4] The rest of the paper will attempt to

[4] Prosody as it affects intonation, not necessarily duration or articulation.

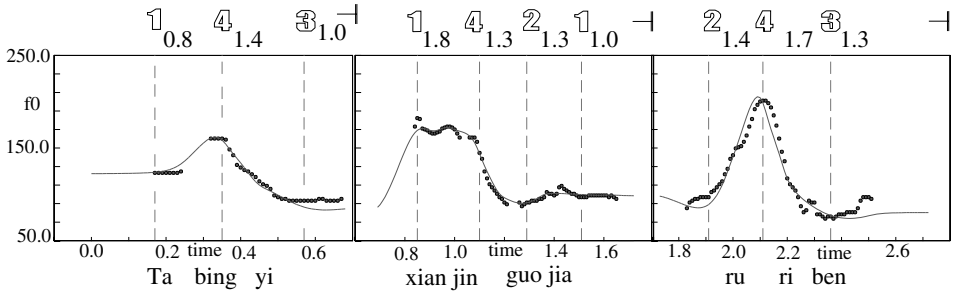

Fig. 3. The beginning of an utterance. Fit (solid) vs. data (dots). Syllable centers are marked with vertical dashed lines. The tones are marked above (in an open face font) and fitted prosodic strength, s_i, is marked as a subscript. (Syllable strength is calculated from word strength and metrical patterns.) The text is marked below. Stem-ML phrases, as defined by pauses, are marked with "⊣".

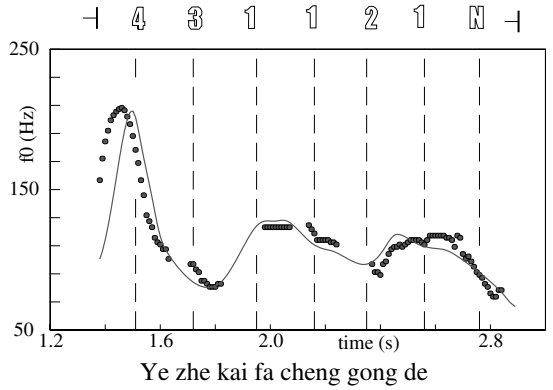

Fig. 4. Typical fit (solid) vs. data (dots), for model subset1-J-wA. Displayed as above.

show that they are simple, useful descriptions of prosody in addition to being nearly complete descriptions.

We can show that the strength values that we obtain are robust against small changes in the assumptions that define the model. For example, Fig. 6 shows a plot of syllable strengths obtained for the first subset with the S-wA model, plotted against strengths obtained from the J-wAT model. 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. Nearly all of the values fall on a narrow band about a smooth curve that maps the strength from one fit to the other. This mapping is the result of differences of shared parameters (most

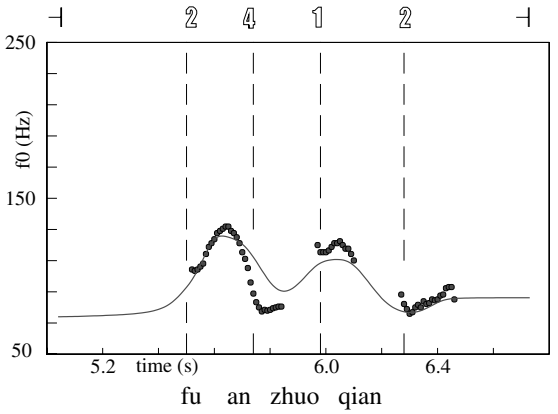

Fig. 5. Phrase containing the worst-fit pair of syllables in the worst model. Displayed as above.

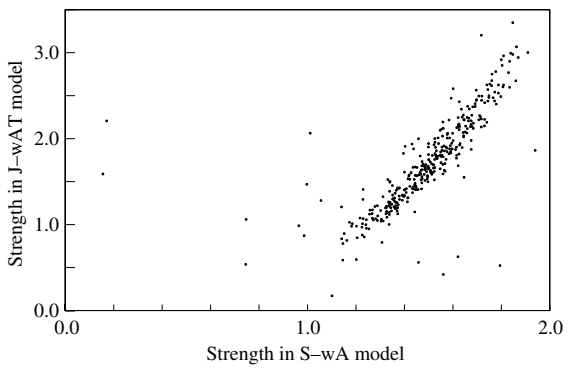

Fig. 6. Comparisons of strength values of syllables between the S-wA and J-wAT models. The strength of most syllables is measured nearly independently of the details of the model.

importantly *atype*) between the two fits. The strength values that are least reproducible are single syllable words, especially single syllable neutral tones.

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 linguistic features. In the following sections, we will discuss the results of the fit and see how they correlate with linguistic expectations.

5.2. Analysis of phrase curve

Our phrase curve is Eq. (9): simple linear declination. We included a phrase curve in the model and fit it, because phrase curves are a common feature in many qualitative descriptions of intonation. However, the data shows no evidence that the phrase curve is necessary, and we see no systematic declination. Neither $P = -4 \pm 3$ Hz s^{-1} nor $D = 0 \pm 4$ Hz s^{-1} is very large, and neither is substantially different from zero (the error bars are derived from the standard deviation of the values of equivalent parameters among the 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 causes the magnitude of f_0 swings to become smaller as the phrase progresses. Our phrase curve roughly corresponds to the reference line of Liberman and Pierrehumbert (1984), and our strength is similar to the difference between their base line and their top line.

5.3. Analysis of tone shapes

First, the fitted scope of the templates is well matched to a syllable. The best-fit templates are $68 \pm 4\%$ of the length of their syllable, and the centers of the tone templates are just $18 \pm 8\%$ of the length of the syllable after the center. This matches well with the intuition that tones are associated with syllables (but see Xu (2001)).

Fig. 7 shows the shapes of the four main Mandarin tone templates, calculated for each of our 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 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

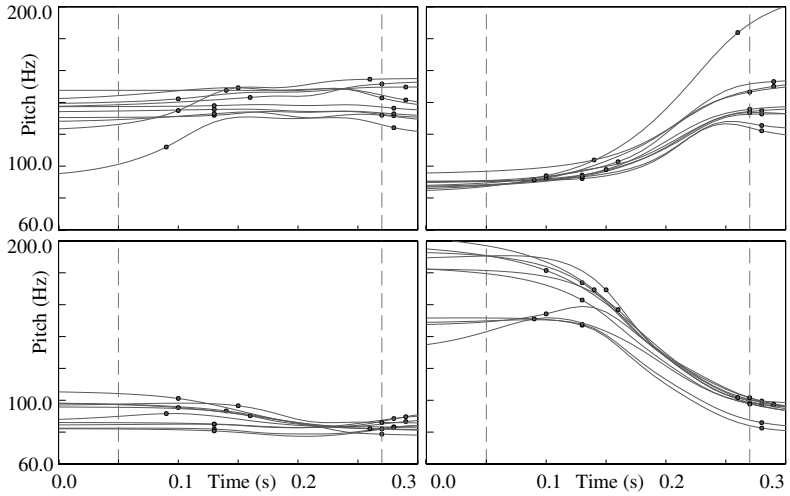

Fig. 7. Modeled shapes of isolated tones. The shapes match standard descriptions, and interact to reproduce continuous speech. The two dashed vertical bars mark the syllable boundaries, and dots mark the boundaries of the tone's template in each of the models (random segmentations were excluded). Each tone was calculated with its strength set to the median of all the strengths in the corpus.

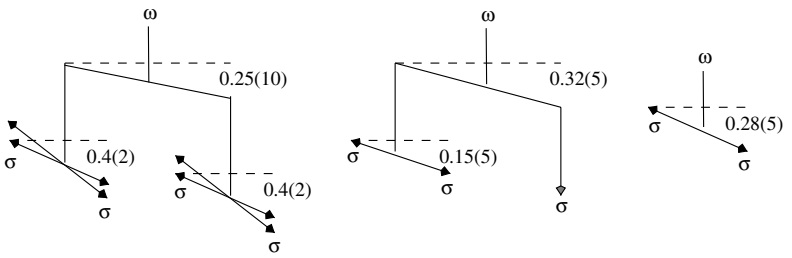

Fig. 8. Metrical patterns for the J and S segmentations of 4, 3, and 2 syllable words. The words (ω) are plotted as trees, and syllables (σ) are represented by the black arrowheads at the end of the lines. The vertical position of the ith arrowhead is proportional to the metrical strength of the ith syllable: $\log(M_{L,i}) \cdot \text{atype}^{1/2}$. 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.

possible, under the constraint that the pitch changes can be accomplished by human muscles within one syllable (Xu and Sun, 2000).

5.4. Analysis of metrical patterns

The RMS error from these word-based models, 13 Hz, is nearly the same as the 12 Hz RMS error we obtain from similar models (Kochanski and Shih, 2001) that do not impose a metrical pattern, but instead allow the strength of each syllable to vary independently. Clearly, the metrical patterns in the words are successful at capturing much of the strength variation from syllable to syllable within a word. The models in this paper have approximately half as many free parameters (and thus are more predictive) than our earlier models (Kochanski and Shih, 2001), and yet still provide an accurate representation of the actual speech.

Fig. 8 shows a tree diagram of the metrical patterns we observe. A direct comparison of the metrical patterns from different models is not useful, because *atype* differs from model to model. The metrical patterns are really measures of relative syllable strength, and *atype* controls how the strength is related to the amplitude of the template. Looking back at Eq. (7), we see that tags with a small value of *atype* will need a broad range of strengths to get a relatively small change in the pitch excursion, and vice versa. This happens because the pitch excursion is proportional to F_k (Eq. (7)), thus it increases at least as fast [5] as the strength raised to the power *atype*. Since the pitch excursions are fit to the data, we expect that models with a small *atype* will have the largest range of strengths. This correlation between *atype* and variance ($\log(s_k)$) is indeed strong. In order to make comparisons clearer, we scale the metrical patterns, $\log(M_{L,i})$, by $atype^{1/2}$ to make the strengths of different models comparable. Recall that *atype* is a global parameter, so that this scaling does not change the shapes or the metrical patterns, nor the relationship between metrical patterns for different words.

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 vary by model, and depend on the neighboring words, our results indicate that RMS swings on the first syllable should be about 30% larger than the second syllable. While it has been generally expected that Mandarin

[5] It will actually increase faster, because as the strength increases from zero, the f_0 curve will tend to follow the templates more and more closely. Note that this argument applies to typical pitch excursions, and is not necessarily true for each syllable: the excursion in a particular syllable depends on its tone class and the strengths and tone classes of its neighbors.

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

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

In Fig. 9, 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 overall picture is similar to the J and S segmentations.

Fig. 10 shows 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*-segmentation does not give as good of a fit to the data: the χ^2 values are 11–21% above the corresponding models with real (A, J, or S) segmentations. This change in χ^2 is substantial: it is an order of magnitude larger than necessary for statistical significance at the 1% level, even if one makes allowance for correlations among the f_0 measurements.

Our results are consistent with the prediction of metrical phonology (Liberman and Prince, 1977). We find an alternating strong/weak relation within bisyllabic words. This pattern repeats in a four syllable word with a higher order hierarchical relation that also shows strong/weak alternation.

5.5. Analysis of word strengths

The strengths that result from the above fitting process can be correlated with linguistic factors.

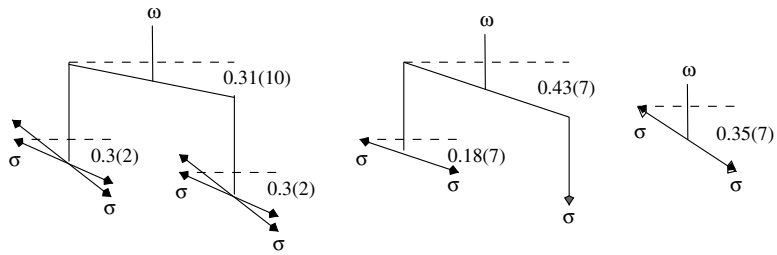

Fig. 9. Metrical patterns for the A-segmentation, plotted as above.

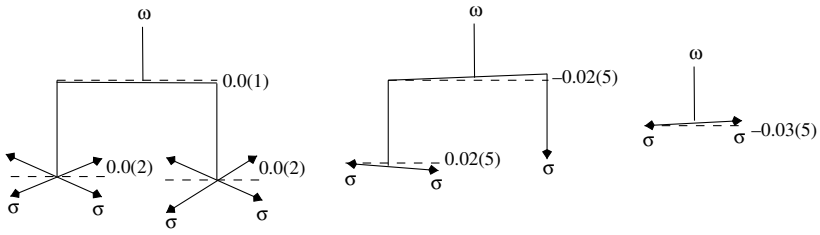

Fig. 10. Metrical patterns for random word segmentation, plotted as above. As expected, the residual patterns are weak and inconsistent.

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. We did not include any semantic features, and syntax was only included through part-of-speech and (to some extent) through our definition of pauses. Also, there was no feature in the model equivalent to the concept of "the focus of a sentence". We limited ourselves to features that could be derived from the text alone (with the exception of phrasal pauses). Phrasal pauses seem to be clear enough to a listener, and their perception seems relatively independent of the pitch, so we tolerated the slight circularity introduced by their use as features.

We then fit the strengths with a trimmed linear regression (MathSoft, 1995) to separate out the effects of the different factors. The model for the observed word strength, S_w, is

$$\widehat{S}_w = c_0 + \sum_i c_i \cdot f_{i,w}, \tag{10}$$

where $\widehat{S}_w$ is the modeled strength. In the sum, i ranges over the features described below, $f_{i,w}$ is 0 or 1, depending on whether the ith feature is present on word w, and c_i is the regression coefficient for the ith feature. Coefficient c_0 shows the strength of words without any features. In this trimmed linear regression, we find the regression coefficients that minimize $\sum'_w (S_w - \widehat{S}_w)^2$, where the primed sum excludes the five largest errors. Excluding a handful of wild points prevents the regression from being dominated by words whose strength cannot be accurately measured (i.e., monosyllabic words that have a neutral tone), and leads to a much more reliable result. Such outliers comprise about 2% of the strength measurements, and can be clearly seen in Fig. 6. We calculated this regression separately for each of our models. In Fig. 11, we plot the distribution of the regression coefficients across models for each factor.

Overall, predicting strength via this linear model reduces the median absolute deviation by 17%: these factors do not provide more than a partial prediction of the strengths or f_0. Again, we use a robust estimator like median absolute deviation instead of variance to reduce the effect of the outliers. If the strength distribution were Gaussian, this regression would have Pearson's $r = 0.31$.

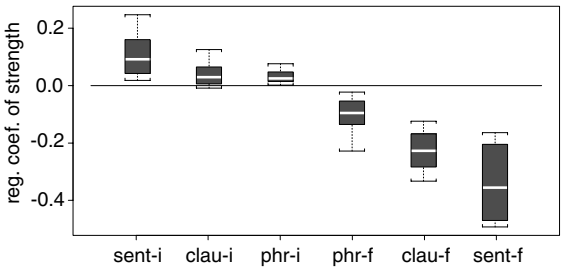

Fig. 11. Relation between strength and word positions. Each box shows the range of the data (the shaded region extends from the 25th and 75th percentiles), the median (white stripe in the box), and outlying points (brackets on the border). All boxes are referenced to words that are not at any kind of boundary, which are shown as the zero line.

We found that:

5.5.1. Words at the beginning of a sentence, clause, or phrase have greater strengths than words at the final positions

Fig. 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.

The hierarchy of linguistic units is displayed with strengths that increase with the size of the unit. Note that the regression coefficient of words not at a boundary is defined to be zero, and that zero (horizontal line) neatly divides the initial words of units (sent-i, clau-i, phr-i) from the final words of the units (phr-f, clau-f, sent-f). These results are consistent with previous findings that speakers use high pitch to mark discourse initial segments (Hirschberg and Pierrehumbert, 1986).

5.5.2. Nouns and adverbs typically have more strength than words of other part of speech, and particles have the lowest strengths

Fig. 12 shows the regression coefficients for different parts-of-speech (Eq. (10)). As we can see, adverbs on average have a greater strength than words of other parts-of-speech. The strengths for nouns, verbs, and conjunctions are slightly weaker than that for adverbs and their strengths are close to each other. In contrast, the strength for particles

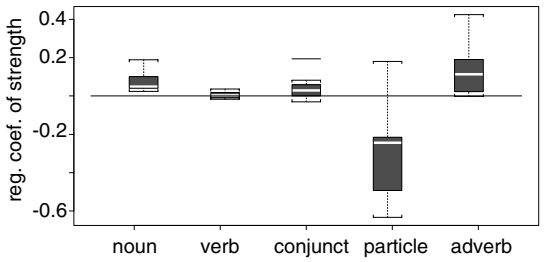

Fig. 12. Relation between part-of-speech and strength. Regression coefficients for Eq. (10) are shown.

(e.g., neutral tones) are much weaker than that for other parts-of-speech. This may be related to the low information content of function words. These results are consistent with previous results which were obtained using human-annotated accents (e.g., Hirschberg, 1993).

5.5.3. Words with more syllables have greater strength than words with smaller number of syllables

Fig. 13 shows the regression coefficients (Eq. (10)) for strengths for words of different lengths. The regression coefficient for three-syllable words is defined as zero, which is shown as the horizontal line in the figure. The plot shows three populations of monosyllabic words, bisyllabic words, and longer words, where word strength increases as a function of word length. The weak status of a monosyllabic word is consistent with previous linguistic observations, where such phenomenon prompted the postulation of the *monosyllabic destressing rule* (Selkirk, 1984).

The correlations between strength in our Stem-ML models and the above linguistic features suggest that the strengths indeed represent the prosodic relations of syllables and words. This has two consequences: First, this knowledge allows us to use features such as position, part-of-speech, and number of syllables in word to predict the strength of a word, and thus improve prediction of f_0 in a Mandarin speech synthesizer. Second, it may be possible to apply it to speech recognition systems, so that the recognizer can detect word boundaries and to deduce whether a word is being emphasized (see Shih et al., 2001 for discussion).

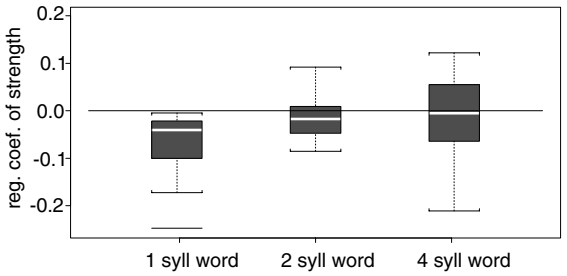

Fig. 13. Relation between strength and the number of syllables in a word. The boxes are plotted relative to three-syllable words, which are shown as the zero line.

5.6. The correlation of strength and duration

We can also calculate the correlation between the fitted strength values with acoustic measurements such as duration. Many duration studies reported a lengthening effect of stressed vowels (Crystal and House, 1988; Klatt, 1973). It is generally expected that, everything else being equal, strong words would have longer duration than weak words.

We calculated the correlation scores between strength and duration in our models, excluding the models using random word segmentation. Outliers are trimmed by excluding the 5% of the population that is farthest from the regression line that defines the correlation, again using a trimmed linear regression. The mean correlation scores of these models are 0.40 in the sentence final position, and 0.27 in the non-final positions.

Fig. 14 show the strength/duration correlation from one of the models. The left panel shows the

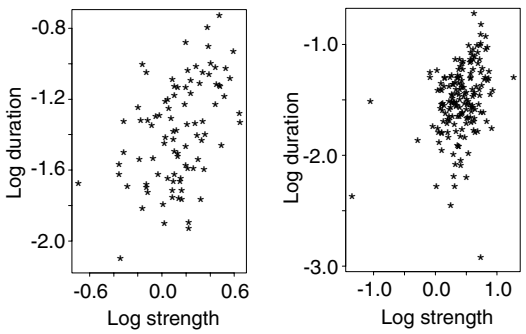

Fig. 14. Correlation of strength values and duration. The left panel shows the sentence final syllables, and the right panel shows the non-final syllables.

population in the sentence final position, and the right panel shows the population in the non-final position. All sample points are used in these plots, where the correlation scores are 0.45 in the sentence final syllables and 0.34 in the non-final syllables.

Phrase final syllables are subject to final lengthening effect (Edwards et al., 1991) and this trend is reflected in the discrepancies between the strength values of final and non-final populations. The phrase final population is characterized by lower strength values and longer duration.

5.7. Mutual information and observed metrical structure

Why might we observe word initial syllables with higher strength than other syllables in the word? We investigate the hypothesis that the speaker is willing to spend more effort to articulate a speech sound clearly when the material is less predictable, but will accept sloppy pronunciation when the material is more predictable. In this section, we use the point-wise mutual information between adjacent syllables to estimate how well a syllable can be predicted from the preceding one, and show that there is a correlation between mutual information scores and prosodic strength.

Point-wise mutual information (Church and Gale, 1991; Fano, 1961) is a measure of how strongly two events are associated, and is defined as

$$I(a;b) = \log_2(P(a,b)/P(a)P(b)), \tag{11}$$

where $P(a)$ is the probability of the event a, $P(b)$ is the probability of the event b, and $P(a,b)$ is the probability of a and b occurring together.

If a and b are independent events, then the probability of them occurring together is the product of the probabilities of a and b: $P(a,b) = P(a)P(b)$ and the mutual information is zero. Applying this measure to text, we can estimate mutual information of two words by using frequency information obtained from a database.

If two words tend to occur together, their mutual information score is positive. Negative mutual information scores suggest some level of avoidance so that the two syllables occur together less often than chance.

In the speech channel, orthographic information is not represented. Therefore, instead of using units like words or Chinese characters (Sproat and Shih, 1990) that apply to written text, we use the syllable, a sound-based unit, to compute mutual information. Syllables with different tones are considered different events.

We used a database with 15 000 sentences (half a million characters). We converted written text into syllable transcriptions using the text analysis component of a text-to-speech system (Shih and Sproat, 1996). The system uses a dictionary together with a homograph disambiguation component to allow context sensitive character-to-sound mapping. We then compute the frequency count of each syllable and each syllable pair from the database, and estimate their probability by dividing the frequency with the total syllable count of the database.

Fig. 15 compares the mutual information scores of the 737 pairs of adjacent syllables in the speech corpus. The figure compares syllable pairs where the second member is word initial (the syllable pair straddles a word boundary) vs. pairs where both syllables are within the same word. The mutual information is high within a word: if you hear the beginning of a word, you have more information about the next syllable. On the other hand, knowing the syllable at the end of one word is not as helpful for predicting the beginning of the next

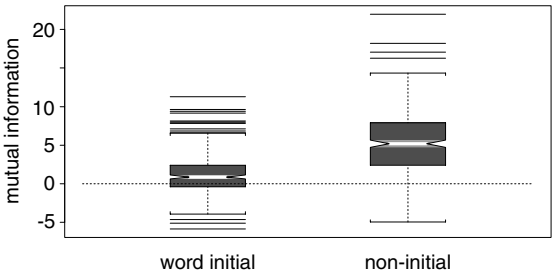

Fig. 15. Mutual information scores of syllables, based on the preceding syllable. The mutual information is lower for initial syllables (i.e., a prediction across a word boundary), thus they are less predictable from the preceding syllable than word internal syllables are.

word. We suggest that where the speech sound is less predictable, speakers spend more effort in pronunciation to make the speech clearer. This may be part of the explanation of the higher prosodic strength we obtained in the word initial positions. Fig. 15 uses word boundaries from the *J* segmentation, but plots from the other two labelers are nearly identical to the one shown. These results are consistent with those obtained by Pan and Hirschberg (2000), using human identification of accent locations.

5.8. Correlation of strength and mutual information

We compute the correlation between mutual information and the prosodic strength of the word initial syllables from three 'wAT' models, one from each word boundary labeler. We do not include word internal syllables in this computation, because the prosodic strength of the word internal syllables is distributed by the metrical structure. The correlation scores of the three models for labelers A, J, and S are −0.20, −0.17, and −0.16, all significant at the 95% confidence level. As expected, there is a negative correlation between mutual information scores and fitted prosodic strengths. Again we see that the less predictable syllable is spoken with higher prosodic strength.

We note that the available database is barely sufficient for calculating mutual information scores across word boundaries: the median syllable occurs only 135 times, thus most possible pairs of syllables simply are not sampled. Consequently, we view these correlations as suggestive, rather than conclusive. However, the observed correlations in Section 5.5.1 are consistent with this hypothesis that strength is at least partially controlled by mutual information. We expect words at the beginning of sentences, clauses, and phrases to be less predictable than words in the middle, because these boundaries can introduce new topics.

As a comparison, we calculated the correlation between mutual information and the high f_0 region in each word. It has been generally expected that a speaker will raise pitch to signal less predictable information. We calculated the f_0 mean of three consecutive voiced samples and took the highest value in each word. The correlation scores of the three segmentations are −0.14, −0.12, and −0.11, smaller than the correlation obtained from fitted prosodic strength, and only the first is significant at the 95% level.

There are several reasons why the fitted strength performs better than surface f_0 values. The raw f_0 values are not corrected for tone class or the effects of the neighboring tones, while the Stem-ML strengths include those basic normalizations. For example, high f_0 may be the result of a preceding rising tone, especially if that tone is emphasized. Not all high f_0 correspond to local intentional emphasis (Shih, 1988; Shih et al., 2001). Furthermore, speakers may use tone-dependent strategies to convey the same prosodic meaning. For example, to express emphasis, people may raise pitch for a high tone but lower pitch for a low tone.

5.9. The scope of prosodic strength

Is the scope of prosodic strength in Mandarin a word or a syllable? We cannot directly answer this question because we assume that Eq. (8) relates the word strengths to the syllable strengths. All of our models in this work assume that one is exactly proportional to the other, therefore the models do not distinguish between the two.

However, we can compare our results here to previous work by (Kochanski and Shih, 2001) where we built models with a separate strength value for each syllable (thus syllable-scope strengths) to fit the same corpus. Since the RMS errors are only marginally worse when we tie the syllable strengths together to make a word strength (13 Hz in this work, vs. 12 Hz in Kochanski and Shih, 2001), we can see that associating strength with words works just as well as associating it with syllables, but leads to a much simpler, more compact model with fewer parameters. Occam's razor thus leads us to associate strengths with words.

However, a comparison of RMS errors has its limitations. It averages over the entire data set, and so cannot exclude the possibility that while most words are spoken in the default word-scope manner, the speaker exercises more detailed syllable-scope control over a few words.

6. Conclusion

We have used Stem-ML to build a model of continuous Mandarin speech that connects the acoustic level up to the results of text analysis (part-of-speech information, and word, phrase, clause, and sentence boundaries). When fit to a corpus, the model shows that prosody is used in a consistent way to mark divisions in the text: sentences, clauses, phrases, and words start strong and end weak. Our prosodic measurements also show a useful correlation with word length, and the part-of-speech of words. We also show that the strength values correlate in expected ways with other acoustic observables such as duration. There is also a correlation between the strength values and mutual information, which suggests that speakers apply a higher prosodic strength to less predictable materials.

The results point to the conclusion that the mathematical models behind Stem-ML provide a quantitative method for measuring prosodic strength. 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., Anderson, V., 2001. Statistical prosodic modeling: from corpus design to parameter estimation. IEEE Trans. Speech Audio Process. 9 (1), 52–66.

Browman, C.P., Goldstein, L., 1990. Tiers in articulatory phonology, with some implications for casual speech. In: Kingston, J., Beckman, M. (Eds.), Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech. Cambridge University Press, pp. 341–376.

Chen, Y., Gao, W., Zhu, T., Ma, J., 2000. Multi-strategy data mining on Mandarin, prosodic patterns. In: Proceedings of the Sixth International Conference on Spoken Language Processing (ICSLP), Beijing, China, October 16–20.

Chen, S.-H., Hwang, S.H., Tsai, C.-Y., 1992. A first study of neural net: based generation of prosodic and spectral information for Mandarin text-to-speech. In: Proceedings of IEEE ICASSP 2, pp. 45–48.

Church, K.W., Gale, W., 1991. A comparison of the enhanced Good-Turing and deleted estimation methods for estimating probabilities of English bigrams. Comput. Speech Lang. 5 (1), 19–54.

Computational Linguistic Society of the Republic of China, 1993. ROCLING Chinese Corpus. Institute of Information Science, Academia Sinica, Nankang, Taipei, Taiwan 11529, ROC, Newspaper texts collected in 1979 and 1980 in Taiwan. Announced Linguist List 4.191.

Crowninshield, R.D., Brand, R.A., 1981. Physiologically based criterion of muscle force prediction in locomotion. J. Biomech. 14 (11), 793–801.

Crystal, T.H., House, A.S., 1988. Segmental durations in connected speech signals: syllabic stress. J. Acoust. Soc. Am. 83, 1574–1585.

Edwards, J., Beckman, M., Fletcher, J., 1991. The articulatory kinematics of final lengthening. J. Acoust. Soc. Am. 89, 369–382.

Fano, R., 1961. Transmission of Information. MIT Press.

Feldman, A.G., Adamovich, S.V., Ostry, D.J., Flanagan, J.R., 1990. The origin of electromyograms—explanations based on the equilibrium point hypothesis. In: Winters and Woo (1990), pp. 195–213, and references therein.

Flash, T., Hogan, N., 1985. The coordination of arm movements: an experimentally confirmed mathematical model. J. Neurosci. 5 (7), 1688–1703.

Flemming, E., 1997. Phonetic optimization: compromise in speech production. University of Mainland Working Papers in Linguistics vol. 5, pp. 72–91. See http://www.stanford.edu/flemming/.

Flemming, E., 2001. Scalar and categorical phenomena in a unified model of phonetics and phonology. Phonology 18, 7–44.

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

Hirschberg, J., 1993. Pitch accent in context: Predicting international prominence from text. Artif. Intell. 63, 305–340.

Hirschberg, J., Pierrehumbert, J., 1986. The international structuring of discourse. In: Proceedings of the 24th Annual Meeting of the Association for Cornputalional Linguistics, vol. 24. pp. 136–144.

Hogan, N., Winters, J.M., 1990. Principles underlying movement organization: upper limb. In: Winters and Woo (1990), pp. 182–194, and references therein.

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

Klatt, D.H., 1973. Interaction between two factors that influence vowel duration. J. Acoust. Soc. Amer. 54, 1102–1104.

Kochanski, G.P., Shih, C., 2000. Stem-ML: language independent prosody description. In: Proceedings of the International Conference on Spoken Language Processing, Beijing, China, vol. 3, pp. 239–242.

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

Kochanski, G., Shih, C., 2003. Prosody modeling with soft templates. Speech Comm. 39 (3–4), 311–352.

Kochanski, G., Shih, C., Jing, H., 2003. Hierarchical structure and word strength prediction of Mandarin prosody. Internat. J. Speech Technol. 6 (1), 33–43.

Laboissière, R., Ostry, D.J., Feldman, A.G., 1996. The control of multi-muscle systems: human jaw and hyoid movements. Biol. Cybernet. 74, 373–384.

Ladd, D.R., 1996. Intonational Phonology. Cambridge University Press.

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

Lee, L.-S., Tseng, C.-Y., Hsieh, C.-J., 1993. Improved tone concatenation rules in a formant-based Chinese text-to-speech system. IEEE Trans. Speech Audio Process. 1 (3), 287–294.

Levenberg, K., 1944. A method for the solution of certain problems in least squares. Quart. Appl. Math. 2, 164–168.

Liberman, M.Y., Pierrehumbert, J.B., 1984. Intonational invariance under changes in pitch range and length. In: Aronoff, M., Oehrle, R. (Eds.), Language Sound Structure. MIT Press, Cambridge Massachusetts, pp. 157–233.

Liberman, M.Y., Prince, A., 1977. On stress and linguistic rhythm. Linguist. Inq. 8, 249–336.

Lin, M.-C., Yan, J., 1983. The stress pattern and its acoustic correlates in Beijing Mandarin. In: Proc. 10th Internat. Congress of Phonetic Sciences. pp. 504–514.

Lindblom, B., 1963. Spectrographic study of vowel reduction. J. Acoust. Soc. Amer. 35 (11), 1773–1781.

MacNeilage, P.F., Sussman, H.M., Westbury, J.R., Powers, R.K., 1979. Mechanical properties of single motor units in speech musculature. J. Acoust. Soc. Amer. 65 (4), 1047–1052.

Marquardt, D., 1963. An algorithm for least-squares estimation of nonlinear parameters. SIAM J. Appl. Math. 11, 431–441.

MathSoft Inc., 1995. S-plus Online Documentation, 3.3 ed., Subroutine *ltsreg*( ), set to exclude the 5 most extreme data points from the objective function.

McFarland, D.H., Smith, A., 1992. Effects of vocal task and respiratory phase on prephonatory chest-wall movements. J. Speech Hearing Res. 35 (5), 971–982.

Monsen, R.B., Engebretson, A.M., Vemula, N.R., 1978. Indirect assessment of the contribution of subglottal air pressure and vocal fold tension to changes in the fundamental frequency in English. J Acoust. Soc. Amer. 64 (1), 65–80.

Moon, S.-J., Lindblom, B., 1994. Interaction between duration, context, and speaking style in English stressed vowels. J. Acoust. Soc. Amer. 96 (1), 40–55.

Nelson, W.L., 1983. Physical principles for economies of skilled movements. Biol. Cybernet. 46, 135–147.

Ohala, J.J., 1992. The segment: primitive or derived? In: Docherty, G.J., Ladd, D.R. (Eds.), Papers in Laboratory Phonology II: Gesture, Segment, Prosody. Cambridge University Press, pp. 166–183 (ISBN 0-521-40127-5).

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

Pan, S., Hirschberg, J., 2000. Modeling local context: for pitch accent prediction. In: Hyman, L. (Ed.), Proc. 38th Ann. Mtg. Association for Computational Linguistics (ACL2000). Association for Computational Linguistics, Hong Kong, pp. 15–70.

Perkell, J.S., Zandipour, M., 2002. Economy of effort in different speaking conditions. II. Kinematic performance spaces for cyclical and speech movements. J. Acoust. Soc. Amer. 112 (4), 1642–1651.

Perkell, J.S., Zandipour, M., Matthies, M.L., Lane, H., 2002. Economy of effort in different, speaking conditions. I. A preliminary study of intersubject differences and modeling issues. J. Acoust. Soc. Amer. 112 (4), 1627–1641.

Pierrehumbert, J.B., Beckman, M.E., 1988. Japanese Tone Structure. The MIT Press.

Prince, A., Smolensky, P., Optimality Theory: Constraint Interaction in Generative Grammar. MIT Press, Blackwell, Oxford, UK, in press (to be published in 2004). Also available as Techinical Report 2 from the Center for Cognitive Science (RuCCS). Rutgers University. Busch Campus. New Brunswick, NJ 08903.

Seif-Naraghi, A.H., Winters, J.M., 1990. Optimized strategies for scaling goal-directed dynamic limb movements. In: Winters and Woo (1990), pp. 312–334, and references therein.

Selkirk, E.O., 1984. Phonology and Syntax: The Relation between Sound and Structure. The MIT Press, Cambridge, MA.

Shih, C., 1986. The prosodic domain of tone sandhi in Chinese. PhD thesis, University of California, San Diego.

Shih, C., 1988. Tone and intonation in Mandarin. Working Papers of the Cornell Phonetics Laboratory. Number 3: Stress. Tone and Intonation, Cornell University, pp. 83–109.

Shih, C., 2000. A declination model of Mandarin Chinese. In: Botinis, A. (Ed.), Intonation: Analysis, Modelling and Technology. Kluwer Academic Publishers, pp. 243–268.

Shih, C., Ao, B., 1997. Duration study for the Bell Laboratories Mandarin text-to-speech system. In: van Santen, J., Sproat, R., Olive, J., Hirschberg, J. (Eds.), Progress in Speech Synthesis. Springer-Verlag, New York, pp. 383–399.

Shih, C., Kochanski, G.P., 2000. Chinese tone modeling with Stem-ML. In: Proc. Internat. Conf. on Spoken Language Processing, Beijing, China, vol. 2, pp. 67–70.

Shih, C., Kochanski, G.P., 2001. Prosody control for speaking and singing styles. In: Proc. Eurospeech 2001. International Speech Communication Association, Aalborg, Denmark, pp. 669–672.

Shih, C., Sproat, R.W., 1992. Variations of the Mandarin rising tone. In: Proc. IRCS Workshop on Prosody in Natural Speech. University of Pennsylvania, pp. 193–200.

Shih, C., Sproat, R.W., 1996. Issues in text-to-speech conversion for Mandarin. Comput. Linguist. Chinese Lang. Process. 1 (1), 37–86.

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

Silverman, K.E., 1987. The Structure and processing of fundamental frequency contours. PhD thesis, University of Cambridge, UK.

Sproat, R.W., Shih, C., 1990. A statistical method for finding word boundaries in Chinese text. Comput. Process. Chinese Oriental Lang. 4 (4), 336–351.

Stevens, K.N., 1998. Acoustic Phonetics. The MIT Press.

Talkin, D., Lin, D., 1996. ESPS/waves online documentation, 5.31 ed., 1996. Program get_f0. ESPS was purchased by Microsoft in 2000. Algorithm is based on: Talkin, D. (1995). A robust algorithm for pitch tracking (RAPT). In: Kleijn, W.B., Paliwal, K.K. (Eds.), Speech Coding and Synthesis. Elsevier, New York.

van Santen, J.P.H., Buchsbaum, A.L., 1997. Methods for optimal text selection. In: EuroSpeech 97, Rhodes, Greece, vol. 2, pp. 553–556.

Whalen, D., Kinsella-Shaw, J.M., 1997. Exploring the relationship of inspiration duration to utterance duration. Phonetica 54, 138–152.

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

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

Winters, J.M., 1990. Hill-based muscle models: a systems engineering perspective. In: Winters and Woo (1990), pp. 69–93, and references therein.

Winters, J., Woo, S. (Eds.), 1990. Multiple Muscle Systems: Biomechanics and Movement Organization. Springer-Verlag, New York.

Xu, Y., 2001. Pitch targets and their realization: evidence from Mandarin Chinese. Speech Comm. 33, 319–337.

Xu, Y., Sun, X.J., 2000. How fast can we really change pitch? maximum speed of pitch change revisited. In: Proc. Sixth Internat. Conf. on Spoken Language Processing (ICSLP), Beijing, China, October, pp. 16–20.

Zahalak, G.I., 1990. Modeling muscle mechanics (and energetics). In: Winters and Woo (1990), pp. 1–23, and references therein.

Zajac, F.E., 1989. Muscle and tendon: properties, models, scaling, and application to biomechanics and motor control. Crit. Rev. Biomed. Eng. 17 (4), 359–411.

Zajac, P.E., Winters, J.M., 1990. Modeling musculoskeletal movement systems: Joint and body segmental dynamics, musculoskeletal actuation, and neuromuscular control. In: Winters and Woo (1990), pp. 139–146, and references therein.