Implications of Prosody Modeling for Prosody Recognition

Chilin Shih, Greg Kochanski, Eric Fosler-Lussier,
Melody Chan †, Jia-Hong Yuan‡

Bell Laboratories, Lucent Technologies
Yale University †
Cornell University ‡
{cls,gpk,fosler}@research.bell-labs.com

Abstract

This paper introduces Stem-ML, which is a model of the prosody generation process with an associated description language, and suggests how it may help prosody recognition. We applied Stem-ML modeling to three topics: the modeling of prosodic strengths, intonation types, and noun phrase patterns. Stem-ML parameters derived from f_0 contours may have a more consistent relationship with prosodic events than raw f_0 values. This may improve identification of accent classes, accent strengths, and intonation types.

1. Introduction

This paper introduces Stem-ML[1], which is a model of the prosody generation process with an associated description language, and suggests how it may help prosody recognition.

Recognition of prosody is difficult because many linguistically meaningful gestures of intonation are not obvious from the surface intonation contour. For example, f_0 height does not always correspond to linguistic prominence, phrase curves are not directly measurable, and the context influences the shapes of accents in the same way that neighboring phones influence each other. Given a reasonable model, Stem-ML can be used to find the optimal description of prosody within that model, and can uncover meaningful gestures that are not apparent on the surface.

Stem-ML (Soft TEMplate Markup Language) is a physiologically based model of the prosody generation process that is driven by linguistically-defined accents. It can be used as an intonation coding system which combines the linguistic descriptive function of tagging systems such as ToBI [2, 3], and a f_0 generation capability analogous to Fujusaki's intonation model [4]. It defines a set of tags that can be used to describe abstract linguistic attributes of prosody, including accent classes and phrase curves, with numerical attributes that can describe intonation variations. The tags are mathematically defined with an algorithm for translating tags into quantitative prosody.

The tag to surface f_0 mapping is unambiguous. Given tags, Stem-ML generates f_0 deterministically. However, mapping relation in the other direction is ambiguous. Similar to the problem of speech coding by articulatory parameters, there are multiple possibilities to represent a given f_0 contour. One may constrain the occurrence of tags as well as the parameter values of tags by employing intonation models, which allows one to predict the usage of accent types and phrase curves.

In the following sections, we first describe the intonation models, then discuss the unique modeling advantages offered by Stem-ML and their potential impact to ASR in three areas:

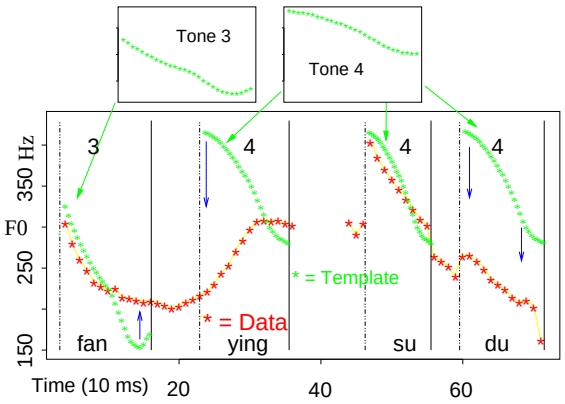

Figure 1: *Tones vs. realization in the phrase fan3 ying4 su4 du4 "reaction time". The upper panels show shapes of tones 3 and 4 taken in a neutral environment and the lower panel shows the realization of the phrase containing those tones. The grey curves show the templates, and the black curve shows the f_0 vs. time data.*

1. The modeling of prosodic strengths, explaining why unstressed words can have high f_0 values.
2. The modeling of intonation types, where a few underlying patterns account for diverse patterns on the surface.
3. Accent shape modeling and the classification of intonation contours over English noun phrases.

Stem-ML parameters derived from f_0 contours may have a more consistent relationship with prosodic events than raw f_0 values. This may improve identification of accent classes, accent strengths, and intonation types. In the paper, we report works in Mandarin Chinese and English.

2. Prosodic Model: Soft Template Markup Language

Stem-ML was initially inspired by tonal distortion data from Mandarin Chinese such as the one shown in Figure 1 [5]. The example shows tone templates vs. the realized pitch track of the phrase *fan3 ying4 su4 du4* "reaction time". The upper panels show shapes of tones 3 and 4 taken in a neutral environment and the lower panel shows the lexical tone templates in grey curves and the actual f_0 vs.time data in black curve. The tone shape of the second syllable is drastically altered to the extent that a lexical falling tone is realized with a surface rising shape.

This kind of distortion occurs in fast speech on a prosodically weak syllable. The direction of the change is predictable: the resulting tone shape conforms to the neighboring tones.

Stem-ML models f_0 by modeling the dynamics of the muscles that control the tension of the vocal folds. Muscles cannot move instantaneously, so it takes time to make the transition from one intended tone or accent target to the next. We represent the surface realization of prosody as an optimization problem, minimizing the sum of two functions: a physiological constraint G, which imposes a smoothness constraint by minimizing effort required to produce the pitch track p, and a communication constraint R, which minimizes the sum of errors r between the realized pitch p and the targets y.

$$G = \sum_t \dot{p}_t^2 + \tau^2 \ddot{p}_t^2$$

$$R = \sum_{i \in \text{tags}} s_i^2 r_i$$

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

(τ, α and β are constants that help define how tones interact, $\bar{p}$ is the average pitch over the scope of a tag, and $\bar{y}$ is the average of y over the tag. $\dot{p}$ and $\ddot{p}$ are first and second time derivatives of p. The above equations are simplified for presentation.)

The errors are weighted by the strength, s_i, of the tag. s_i indicates how important it is to satisfy the specifications of the tag. If a tag is weak, the physiological constraint takes over and in those cases, smoothness becomes more important than accuracy, and the pitch is then dominated by the tag's neighbors. Stronger tags impose their shape on p, and exert more influence on their neighbors.

With this model, the distorted tone shape on the second syllable in Figure 1 is accounted for with a low strength value. A tag set of $\mathbf{3}_{1.1}\ \mathbf{4}_{0.2}\ \mathbf{4}_{1.2}\ \mathbf{4}_{0.8}$, in conjunction with global parameters that define pitch range and lexical tone templates, reproduces the observed f_0 contour. The leading numerals in the tag set represent the lexical tone templates (each implemented as a 5 point representation describing the tone shape), and the subscript represents the strength of the tone template.

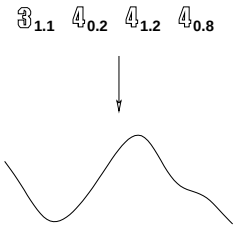

Figure 2: *F_0 curve generated by Stem-ML from the tag set $\mathbf{3}_{1.1}\ \mathbf{4}_{0.2}\ \mathbf{4}_{1.2}\ \mathbf{4}_{0.8}$, and global parameters defining pitch range and lexical tone templates.*

3. Prosodic Strength

Strength in Stem-ML is a measure of how precisely a speaker adheres to the specification of the tone or accent template. This definition has some advantages over a definition of strength that is based on pitch height or pitch range: it links distorted tone shapes to prosodically weak positions and explains the possible outcome. Under this definition, the second syllable in Figure 1 is interpreted as weak while it has a reasonably wide pitch range and high f_0 value.

It is well-known that f_0 height is not always a good indicator of prosodic strength [6]. The relationship between height and strength can be improved by taking into account various sentence effects and discourse factors. Nontheless, such normalization procedures cannot solve the problem of local interpretation of f_0 height relative to nearby words. One frequently finds cases where unimportant function words have higher f_0 than their immediate neighbors. This complicates any algorithm designed to derive information from prosody. Stem-ML offers a model of accent interaction which can account for the high f_0 of these unaccented words.

Figure 3 shows such an example. A natural f_0 curve is plotted from the phrase "I would like to arrive ... " found in the DARPA Communicator database [7]. In this example, *to* has higher f_0 than the surrounding content words which are obviously stressed. The dashed line shows the predicted f_0 values of an unaccented *to* by linear interpolation from the end of the preceding L*+H accent to the next L*+H accent. The predicted f_0 value is too low, and if one assumes that f_0 is locally related to strength, the most natural way to account for the higher f_0 is to assign a unreasonably large strength to *"to"*.

On the other hand, the solid line shows a Stem-ML model of the region, where the height of the word *to* is a natural consequence of its environment. In this model, the three words *I*, *like* and *arrive* are the only accented words, all sharing the same rising accent template. *I* is stressed weakly while *like* and *arrive* are stressed strongly. The function word *to* rides on the slope defined by its more important neighbors. Because *"to"* has little strength, it does not affect the prosody in its vicinity.

This strong tonal coarticulation is physiologically necessary, as the muscles that control f_0 are simply not fast enough to adjust between the end of one syllable and the beginning of the next. Most muscles cannot respond faster than 150 ms, a time which is comparable to the duration of a syllable.

In recent work [8] we are able to replicate Mandarin sentence intonation to within 12 Hz rms error with 0.68 parameters per syllable. The parameters include one strength parameters per word and global settings including lexical tone templates,

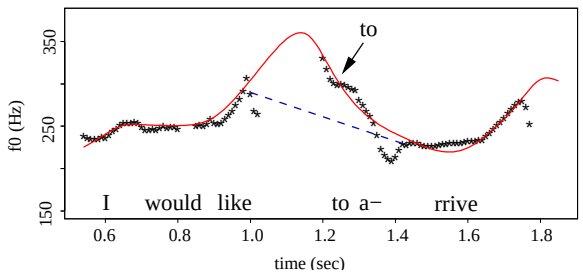

Figure 3: *Example of a high-pitched function word. Data is plotted as "*". Dashed line is predicted from ToBI label interpolation; solid line from Stem-ML constraints.*

pitch range and smoothing window of muscle dynamics. The Stem-ML fitted strengths correlate with linguistic structure better than surface f_0. We expect that this finding will generalize to the interpretation of prosodic strength in English.

4. Question intonation

Mandarin question intonation shows an interesting diversity due to tone and intonation interaction. A sentence ending in a rising tone has a higher rising tail, much like English question intonation. In contrast, a sentence ending in a falling tone shows a higher peak without a rising tail, behavior similar to Greek questions. Consequently, a H% boundary tone aligned with the end of the sentence may account for English as well as Mandarin rising tones, but fail for Mandarin falling tones and Greek.

Previous literature has talked about rising phrase curve [9, 10] or high boundary tones [11, 12] of question intonation. But neither of the accounts can explain all question patterns in Mandarin. While one typically finds regions of high pitch near the end of a question, exactly where they occur depends on the tone sequence. In sentences with final falling tone or final low tone, the pitch may end low.

The optimal models trained by Stem-ML can precisely explain the difference between declarative and interrogative sentences as a combination of two mechanisms: an overall higher phrase curve for the question, and increasing strength values of tones near the end of the sentence [13]. This result is consistent with a perception study of question intonation [14], where listeners are more likely to interpret higher peak and higher ending pitch as questions, independent of their language background.

Furthermore, the optimal phrase curves of the two intonation types are roughly parallel, as shown in Figure 4. The solid line represents the phrase curve of declarative sentences while the dashed line represents that of interrogative sentences. The difference between the two phrase curves corresponds to 8.48 Hz.

The picture shown comes from a model using two points to represent phrase curve. The nearly parallel phrase curves are also found consistently in other models that use three or more points to represent phrase curve.

The higher f_0 at the end of a question intonation is accounted for by higher accent/tone strengths. Figure 5 shows the differences of strength values between interrogative sentences and declarative sentences plotted by syllable positions. The increased strengths at the end imply tighter adherence to the ideal tone shapes and larger pitch excursions.

The Stem-ML models show the correct interaction between tone and intonation. Higher strength accounts for higher ending pitch of rising and high tones, but raises the peak of a falling tone without affecting the final pitch.

We obtained excellent fits for sentences with different tonal combinations using higher phrase curve and increasingly higher strengths on sentence final syllables to model question intonation. Figures 6 and 7 shows the match between the model f_0 and natural f_0 for sentences ending in rising and falling tones, respectively. The filled circles represent natural f_0 and the solid lines represent the calculated f_0. Tones are labeled on top of the figures and the grey dashed lines mark syllable centers.

5. English Noun Phrases

In this section, we report preliminary results of a study on English noun phrases in the DARPA Communicator database [7]. We studied whether consistent prosodic patterns could be found

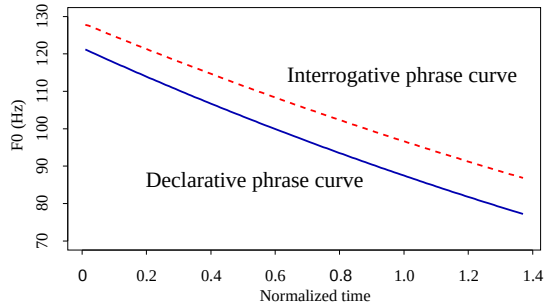

Figure 4: *Phrase curves of question intonation (dotted line) and declarative intonation (solid line). The two lines are roughly parallel: question intonation has higher phrase curve.*

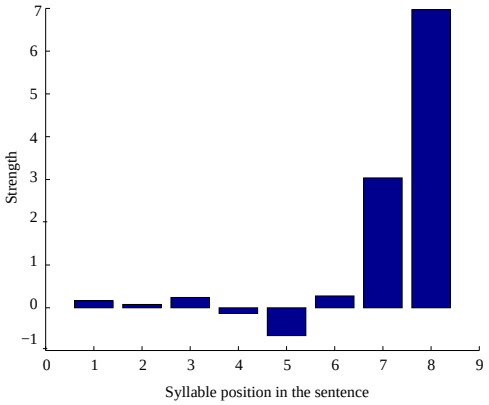

Figure 5: *Difference of syllable strengths between question intonation and declarative intonation, plotted by sentence positions.*

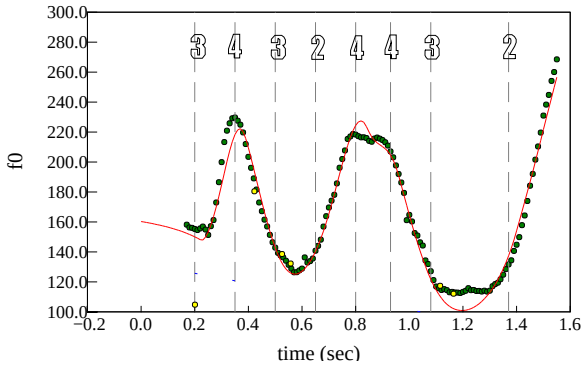

Figure 6: *Natural (filled circles) and model (solid line) intonation curves of a sentence ending in a rising tone: Li3-bai4-wu3 luo2-yan4 yao4 mai3 yang2. "Luo-Yan wants to buy sheep on Friday."*

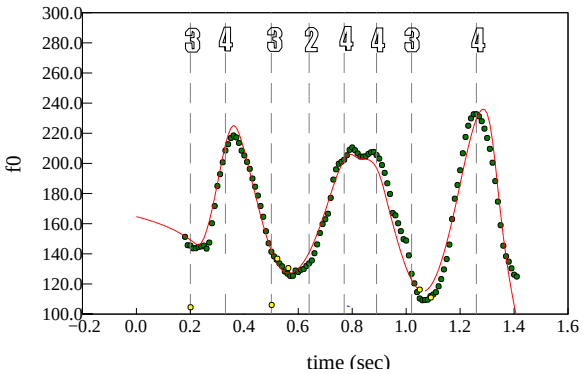

Figure 7: *Natural (filled circles) and modeled (solid line) intonation curves of a sentence ending in a falling tone: Li3-bai4-wu3 luo2-yan4 yao4 mai3 lu4. "Luo-Yan wants to buy a deer on Friday."*

in noun phrases.

We first hand-classified prosodic patterns of noun phrases [15, 16], and then modeled these patterns with Stem-ML. We found that speakers use just a few prosody patterns in long noun phrases; therefore prosody can provide some information for identifying these regions automatically [17].

Our sub-selected database consists of 57 utterances from 26 speakers. These utterances contain 103 noun phrases. Five prosodic patterns are found in these noun phrases, with the following frequency distribution. In addition to the 5 patterns, we mark regions outside of the noun phrases as *OTHERS*. A noun phrase occurring before pause is also marked with a boundary tone at the end.

Pattern	Code	Freq	Description
DROP	δ	40	primarily falling
RISE	ρ	38	primarily rising
LEVEL	λ	9	no movement
HAT	η	9	initial rise, terminal fall
VALLEY	υ	7	initial fall, terminal rise
OTHERS	o	76	
BOUNDARY	β	89	

To prepare database for modeling, we marked the noun phrases with the category of prosodic patterns and assigned boundary tones after long pauses and at the ends of phrases. For example:

from	Atlanta GA	going	to	London England	
o	λ	o	o	ρ	β
from	Phoenix	Arizona	to	Bangkok Thailand.	
o	ρ	υ	ϵ	δ	β
Leaving	San Antonio,	October 17th,		2000.	
o	η β	η	β	δ	β

Using the prosodic marking of the database as input, we fit Stem-ML models to natural f_0 contours by optimizing the shapes of the prosodic templates, the strength of each occurrence of a template, and a set of global parameters. Figure 8 plots the shapes of the five prosodic templates that are learned from the database.

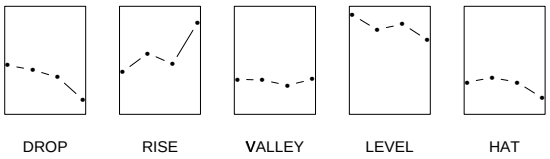

Figure 8: *Stem-ML fitted templates of noun phrases in the Communicator database.*

Each prosodic pattern is represented as a template defined by four points. [1] The templates captured the broad f_0 movement in the noun phrase regions, using one template for each pattern. The model ignores short-term f_0 movements such as segmental effects and even lexical stress. The question is how much of the f_0 movement can be accounted for with a simple model like this one.

Figures 9, 10 and 11 compare f_0 tracks generated from the coded parameters to the original ones. The natural f_0 is plotted by circles and the model f_0 by solid lines.

Figure 9 includes a LEVEL pattern followed by a RISE pattern. This is the first sentence in a dialogue, where speakers often used the RISE pattern to make requests. The model f_0 comes from the template and strength coding shown below, where the Greek letters represent the coded prosodic templates and the subscripts are the fitted strength values. The boundaries of the patterns are marked by dotted lines.

from	Atlanta Georgia	going
$o_{0.09}$	$\lambda_{0.35}$	$o_{0.75}$
to	London England	
$o_{0.57}$	$\rho_{0.62}$	$\beta_{0.85}$

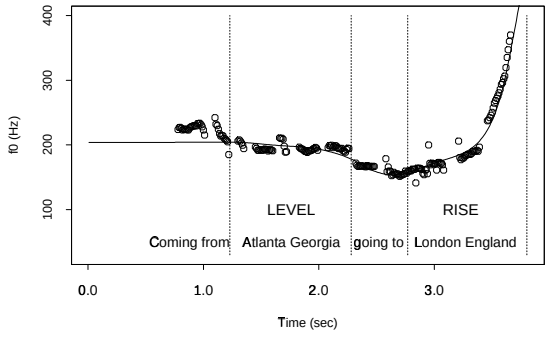

Figure 9: *"from Atlanta Georgia going to London England." A LEVEL pattern followed by a RISE pattern. LEVEL are typically used in non-final positions. This is the first sentence in a dialogue.*

Figure 10 includes a RISE pattern followed by a VALLEY pattern, and terminates in a DROP pattern. This is also the first sentence in the dialogue. The model f_0 is derived from the following coding:

[1] Experiments with large numbers of points showed equally good fits. Four points per template was chosen as the minimal good fit model.

from	Phoenix	Arizona
$o_{6.65}$	$\rho_{2.1}$	$v_{3.07}$

to	Bangkok Thailand.	
$o_{6.92}$	$\delta_{2.19}$	$\beta_{0.85}$

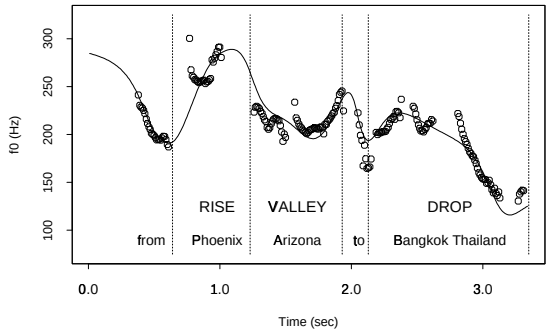

Figure 10: *"from Phoenix Arizona to Bangkok Thailand." This sentence contains three of the prosodic patterns: RISE, VALLEY, and DROP. This is the first sentence in the dialogue.*

Figure 11 includes two HAT patterns followed by a DROP pattern. This is the thirteenth utterance in the dialogue after several rounds of false recognition from the ASR system. The speaker was getting impatient and frustrated, which was expressed by multiple usage of the HAT pattern, terminal DROP, multiple pauses and very slow speaking rate. The model f_0 is derived from the following coding:

Leaving	San Antonio,	October 17th,	2000.			
$o_{11.67}$	$\eta_{5.15}$	$\beta_{0.85}$	$\eta_{3.7}$	$\beta_{0.85}$	$\delta_{1.7}$	$\beta_{0.85}$

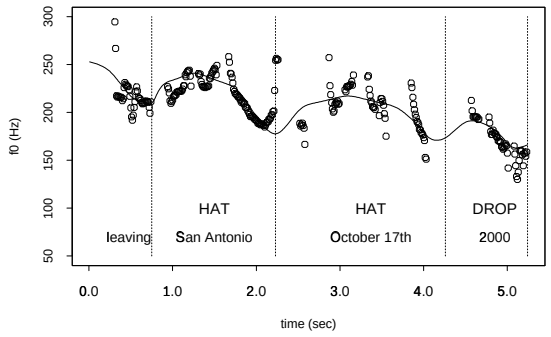

Figure 11: *"Leaving San Antonio, October seventeenth, two thousand." Two HAT patterns followed by a DROP pattern, This is the 13th sentence in the dialogue, after many recognition failures.*

In the context of the communicator dialogue, the speakers tend to be polite initially when they present new information to the system as requests, using rising intonation on noun phrases that contain information such as flight origin, destination, and date and time of travel. As the systems fail to recognize these information, The speakers often slow down, pause more, and switch the prosodic pattern from rising ones such as RISE and VALLEY to falling ones such as HAT and DROP.

There are modest, but real correlations between different f_0 patterns and the information in an utterance that a dialogue system can use. For instance, the pattern was correlated with the frustration level of the speaker. We measured frustration by asking a subject to listen to each dialogue, and to rank at every dialogue turn the user's frustration level on a scale of 1 to 3. Knowledge of the prosodic pattern gives 0.3 bits of information toward selecting among the three marked frustration levels. If we assume that an automated classification of prosodic patterns would yield the same results as the human classification we used, this information could be used to simplify the dialogue, and provide more feedback to the user when he/she starts becoming frustrated.

Likewise, the RISE pattern is associated with new information slightly more often than with other patterns, and the HAT pattern with old information. Overall, knowledge of the pattern yields 0.1 bits of information about the binary choice of whether a person is repeating old information or adding new information into a dialogue.

6. Implications for ASR and Dialogue Systems

There has been significant work to date in integrating prosodic features into detectors of linguistic events, such as errors made by dialogue systems [18, 19, 20], or dialogue acts [21, 22]. We believe that the lessons we have learned in building quantitative Stem-ML models of intonation and prosody can help improve the feature vectors used in these types of classification systems. Our experiments here show that we can accurately describe the prosody of user utterances by characterizing prosodic patterns with a sparse set of template and strength parameters. By finding correlates of the Stem-ML parameters to linguistic phenomena, therefore, we can begin to develop models for detection of these events.

In Mandarin, for example, it is difficult to predict whether a sentence is declarative or interrogative using sentence-final pitch values because of the interference of tones. However, Stem-ML strength values and phrase curves do give a more accurate assessment of the type of sentence.

If the tone sequence is known, we can predict where one can find the biggest difference between declarative and question intonation. By coordinating with initial word hypotheses from an ASR system, we can gather evidence as to the sentence intonation type. In practice, there may not be a unique solution, but there will be evidence favoring the combination of certain tone sequences and intonation types. This can greatly aid spoken dialogue systems by providing confirmation of whether the user is providing information to the system, or is making a request of some type.

Our investigation of English noun phrases in a spoken dialogue system shows that templatic patterns also carry some information for discourse analysis. Certain patterns in our (admittedly small) database are used with different frequencies when the speaker is frustrated, or is repeating information. In future work, we hope to find similar effects in other languages, both in the modeling and recognition of intonation types and emotions.

In the future, we intend to extend our model to find prosodic patterns within ASR recognition hypotheses by searching over the possible templatic patterns. Once this is accomplished, we

can automate the training process further by bootstrapping from the hand-labeled data, automatically labeling larger corpora for further model training.

The current work carries some important implications for spoken language understanding systems – when we are able to detect coherent prosodic patterns corresponding to linguistic structures, we can apply this knowledge to the verification of hypotheses made by various components of a spoken dialogue system, e.g., an ASR system or a pragmatic interpreter that makes inferences about user input. However, only by studying the prosodic patterns that are present within natural speech can we hope to extract information that can be integrated into these dialogue systems.

7. References

[1] Greg P. Kochanski and Chilin Shih, "Stem-ML: Language independent prosody description," in *Proceedings of the 6th International Conference on Spoken Language Processing*, Beijing, China, 2000.

[2] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: A standard for labeling English prosody.," in *International Conf. on Spoken Language Processing*, Banff, 1992, International Conf. on Spoken Language Processing, vol. 2, pp. 867–870.

[3] Mary E. Beckman and Gayle Ayers Elam, *Guidelines for ToBI Labelling*, The Ohio State University Research Foundation, Ohio State University, 1997, http://www.ling.ohio-state.edu/phonetics/E_ToBI/.

[4] Hiroya. Fujisaki, "Dynamic characteristics of voice fundamental frequency in speech and singing.," in *The Production of Speech*, P. F. MacNeilage, Ed., pp. 39–55. Springer-Verlag, 1983.

[5] Chilin Shih and Greg P. Kochanski, "Chinese tone modeling with Stem-ML," in *ICSLP*, Beijing, China, 2000.

[6] Janet Pierrehumbert, "The perception of fundamental frequency declination," *J. Acoustical Soc. Am.*, vol. 66, no. 2, pp. 363–369, 1979.

[7] National Institute of Standards and Technology, "DARPA communicator travel reservation corpus – June 2000 evaluation," Tech. Rep., Gaithersburg, MD, 2000, Speech Data published on CD-ROM.

[8] Greg Kochanski and Chilin Shih, "Hierarchical structure and word strength prediction of Mandarin prosody," in *4th ISCA Tutorial and Research Workshop on Speech Synthesis*, Scotland, August 29th – September 1st 2001.

[9] Eva Gårding, "A generative model of intonation," in *Prosody: Models and Measurements*, Anne Cutler and Robert Ladd, Eds., pp. 11–25. Springer, Heidelberg, 1983.

[10] Xiao-Nan Susan Shen, *The Prosody of Mandarin Chinese*, University of California Press, 1990.

[11] Janet Pierrehumbert, *The Phonology and Phonetics of English Intonation*, Ph.D. thesis, MIT, 1980.

[12] Mark Y. Liberman and Janet B. Pierrehumbert, "Intonational invariance under changes in pitch range and length," in *Language Sound Structure*, M. Aronoff and R. Oehrle, Eds., pp. 157–233. M.I.T. Press, Cambridge, Massachusetts, 1984.

[13] Jia-Hong Yuan, "Comparison of declarative and interrogative intonation in Chinese," Manuscript, Bell Labs, Murray Hill, NJ, 2001.

[14] Carlos Gussenhoven and Aoju Chen, "Universal and language-specific effects in the perception of question intonation.," in *Proceedings of ICSLP 2000*, Beijing, China, 2000.

[15] Douglas O'Shaughnessy , "Linguistic features in fundamental frequency patterns," *Journal of Phonetics*, vol. 7, pp. 119–145, 1979.

[16] J. 't Hart, Collier R., and Cohen A., *A Perceptual Study of Intonation: An Experimental-Phonetic Approach*, Cambridge University Press, 1990.

[17] Melody Chan, "Prosodic modeling and recognition of English noun phrases," Manuscript, Bell Labs, Murray Hill, NJ, 2001.

[18] Julia Hirschberg, Diane Litman, and Marc Swerts, "Generalizing prosodic prediction of speech recognition errors," in *International Conference on Spoken Language Processing (ICSLP)*, Bejing, China, September 2000.

[19] Jun ichi Hirasawa, Noboru Miyazaki, Mikio Nakano, and Kiyoaki Aikawa, "New feature parameters for detecting misunderstandings in a spoken dialogue system," in *International Conference on Spoken Language Processing (ICSLP)*, Bejing, China, September 2000.

[20] Katrin Kirchhoff, "A comparison of classification techniques for the automatic detection of error corrections in human-computer dialogues," in *NAACL Workshop on Adaptation in Dialogue Systems*, Pittsburgh, Pennsylvania, June 2001, pp. 33–40.

[21] Helen Wright, Massimo Poesio, and Stephen Isard, "Using high level dialogue information for dialogue act recognition using prosodic features," in *ESCA Workshop on Prosody and Dialogue*, Eindhoven, Holland, September 1999.

[22] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van Ess-Dykema, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Computational Linguistics*, vol. 26, no. 3, pp. 339–373, 2000.