Towards the Organization of Mandarin Speech Prosody: Units, Boundaries and Their Characteristics

Chiu-yu Tseng[†]
Institute of Linguistics, Academia Sinica
Taipei, Taiwan 115
E-mail: cytling@sinica.edu.tw

ABSTRACT

This study attempts to establish a possible proto organization of speech prosody for Mandarin Chinese on the basis of perceptual as well as acoustic analyses of units and boundaries in production data. These analyzed boundaries and units were used in applying Fujisaki's model of sentence intonation to generate sequences of prosodic units. Though at this stage the approach appears more data driven, our purpose is to address (1.) how phrases are grouped into larger prosodic units in speech output, (2.) how an existing phrasal/sentential model can be elaborated to include such grouping in order to form larger prosodic units, (3.) how such grouping is manifested in prosody, and (4) how a possible structure of prosody organization can be achieved by incorporating these features while leaving room for future enhancement.

1. INTRODUCTION

We note that in spoken Mandarin Chinese, instead of complete or complicated sentences, native speakers tend to speak in a sequence of phrases. These phrases, somewhat loosely governed by semantics and are currently under investigation, are grouped into perceptually identifiable larger units. So far such identification, largely consistent across listeners, can be characterized on the basis of perceived boundaries as units in prosody. These results have been analyzed and reported in our earlier investigations [1, 2]. The fundamental frequency (F0) patterns can also be characterized with respect to their positions in a prosodic group. Therefore, we are interested in seeing how we could incorporate these findings into a working model of intonation generation for the above mentioned phrase grouping. In this paper, we will present preliminary results of applying our results to the Fujisaki model [3 to 9] to show that it is possible to elaborate a phrase intonation model into a much larger prosodic unit that may reflect the overall planning of speech prosody better.

2. Experiments

The aim of the reported experiments was to test if a phrasal F0 contour model can be used to generate sequences of phrases, and capture the overall prosody of the phrase/utterance group, called the prosodic group in prosody organization. The materials under investigation were transcribed speech data of one male speaker. Read speech of 582 relatively long paragraphs at up to 180 characters/syllables in length were used. All of the speech data were transcribed for phonetic segments, perceived boundaries and breaks/pauses, perceived focus/prominence, and F0 measurements. Segmental transcription was first performed by the HTK software and then manually checked by trained transcribers. Perceived boundaries and breaks were manually labeled by 3 transcribers and checked for both intra- and inter-transcriber consistency. Perceived focus/prominence followed the same pattern of break transcription. F0 patterns were calculated automatically.

Experiment 1 aimed to see how we used perceptually based labeling results in applying the Fujisaki phrase intonation model [3-9], and whether the modification of parameters was ordered. This is the first step towards using an existing model to generate a sequence of phrases. Instead of using punctuation marks in the text as indictors of boundaries only or incorporating results from parsing, we used our labeling results of perceived breaks in collected speech data to denote boundaries. Three levels of our labeled breaks in our system, namely, B3, B4 and B5, were used to denote boundaries corresponding to prosodic phrase, utterance and prosodic group. The model of phrase command applied at each boundary, treating each boundary as T0 and each unit before T0 as a phrase. These applications continued until it reached B5 that in our system denotes the end of the grouping of phrases. Figure 1 shows how the generation of a sequence of phrases forming a prosodic group was achieved. The upper column shows the F0 patterns, the blue spikes are the generated F0 peaks, the pink contour the actual speech. The lower column of spikes shows the location of boundaries in the prosodic group.

Subsequently, the remaining two parameters of the model, namely, Ap and α, were adjusted to optimize the generated

ISBN 1-876346-48-5

output towards actual speech output. During the adjustment stage we set out to look into (1.) whether there exists an ordering in the parameters for phrase command and (2.) whether we could group a sequence of phrases into a prosodic group by utilizing some physical characteristics.

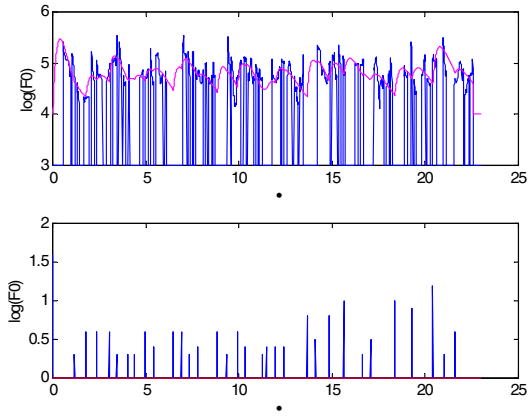

Figure 1.Example of synthesized sequences of intonation contour

Experiment 2 aimed to see if adjustment of parameters involved any ordering. Given the two available parameters, two possible orders of adjustments were applied. The first order was to adjust Ap first, then α for phrase command; the second adjustment was on the results of the first adjustment. A reversed order was applied a second round. Comparisons were then made of these two orders to see if ordering of application would affect the generated output in the end. Figure 2 shows the comparison of two application orders. Results showed no significant difference between these two orders.

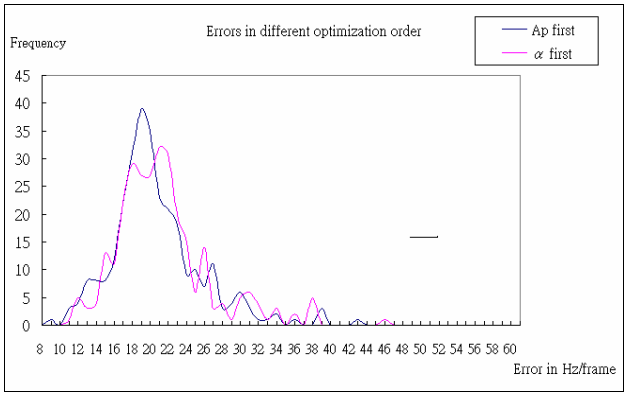

Figure 2. Comparison of two orders of parameter adjustments.

Experiment 3 aimed to see whether prosodic characteristics could be found with respect to the beginning and ending portions a prosodic group that contained a sequence of phrases. We took the fist and last Ap within each prosodic group, in other words, the first Ap of the first prosodic unit and the last Ap of the last prosodic unit, to see if significant difference can be found. Since we treated all of these prosodic units as phrases in Fujisaki's term for the time being, we are well aware that future refinements may be more than necessary.

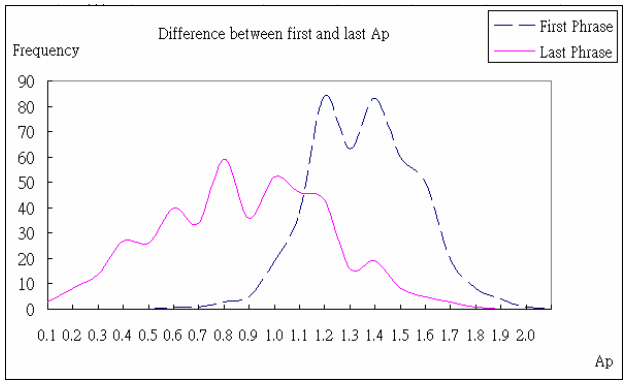

Figure 3. Frequency of Ap in prosodic group's head and tail

However, note that our previous analyses of breaks [1] showed a sharp duration difference between mean duration of B2 (12ms) and B4, B5 (over 600ms) but a wide distribution of duration for B3 at the mean duration of 346ms, we used 150 ms as a reference duration to denote phrase end within a prosodic group. A t-test is performed and showed that significant difference was found and 150ms is a feasible choice. Figure 3 shows the results while Table 1 shows the statistical analysis. Note that significant difference is found between the compared two sets.

Mean	1.322	0.848
Variance	0.044	0.107
Data number	440	440
Pooled variance	0.0757	
df	878	
T-value	25.56	
P(T<=t) one-tail	1.8E-108	
threshold： one-tail	1.646	

Table 1. t-test for Ap in prosodic group's head and tail

3. Discussion

One major challenge faced by linguists and language engineers working on developing unlimited text-to speech (TTS) of Mandarin Chinese has been the naturalness issue, naturalness of speech output as well as a set of commonly agreed evaluation or measurement criteria for speech naturalness. In essence, naturalness boils down to our insufficient knowledge of how prosody is organized to generate the overall rhythmic and melodic aspects of speech flow. Or simply, what speakers do when NOT speaking in isolated phrases or sentences. This problem was best exemplified when the goal of speech synthesis is to convert paragraphs of written text, generating isolated

sentences and sequencing them one after another appear insufficient. The problem is further complicated due to the nature of written Chinese. Chinese text is composed of characters that correspond to syllables but not necessarily to words; a Chinese word can be polysyllabic, but unfortunate for linguists and speech engineers, no morphological cues exist. Linguistically native speakers' intuition appears rather fuzzy and opaque, largely due to the fact that no spacing of words is required in writing Consequently, written Chinese is quite often punctuated with a string of commas separating characters that correspond phrases whereas a period marking the end could be far ahead. For example, our database of read speech contained text of paragraphs that were easily over 400 characters/syllables, with quite a few commas in between, but usually nothing else to break down the paragraphs into some units in between or above until it reaches the lone period at the very end. At an average of about to 250ms mean duration per syllable, this could mean around 90 seconds of speech data. It is quite obvious that any speaker would have to break these 90 seconds into smaller units in actual speech, catching breath for one thing, planning where to end for another. The question then is: how can we predict where the boundaries could and should fall? Further, how can we predict the overall rhythmic as well as melodic structure from text?

In spoken Chinese, this is also the case. A series of phrases are spoken to denote a relatively comprehensive semantic, pragmatic or simply prosodic units rather than a series of sentences that are separated by the punctuation mark period in written form. That is, Mandarin Chinese speakers tend to speak in groups of phrases instead of separated sentences. In synthesizing Mandarin Chinese, the most practiced approach so far has been to treat these phrases as sentences. Each of them would be given a sentential intonation patterns, most likely the declarative pattern, and what we ended up with would be a series of short falling intonations, resulting the impression of abruptness and choppiness that are far from natural We are also lacking in our knowledge as to how many breaks/pauses to insert and where to insert them, especially with respect to sequencing phrases that do not constitute sentences and therefore are far from coming up with the overall rhythm close to that of actual speech output.

In short, we are faced with two major difficulties that are pretty much two sides of the same issue. One is how to predict prosody from text; locating boundaries and units in prosody; the other is how to group phrases into larger prosodic units to avoid short and choppy utterances. This is quite similar to generate prosody for complex sentences from text of English where a number of phrases are punctuated properly by syntactic requirements to mark the boundaries, and correctly end in a comma when punctuation is only a very loose reference in Chinese and the grouping of phrases could not be obtained from syntactic information.

A series of our previous studies analyzed speech data of read paragraphs by locating perceptually identifiable breaks and boundaries focusing on how such breaks and boundaries corresponded to prosodic properties [summarized in1]. Our experiments reported here showed the following: (1.) 3 units of prosody can be treated as phrases in a phrase intonation generation model, marked by boundaries that correspond to breaks (pauses, silence) in speech output. (2.) These breaks (pauses, silence) in running speech are systematic, as we have found in previous investigations. Note that our break analyses focused on production data based speech units and the grouping of these units, and are best exemplified in our proposal of the prosodic group as the largest prosody unit instead of phrase or sentence [2]. In this study, we have shown that these notions have been and can be applied notion to the generation of speech output. We believe that this is a significant step towards the naturalness issue. (3.) Prosodic boundaries and units are hierarchical, coupled with following breaks to make up the organization of speech prosody. These findings paved a foundation of a possible organization of speech prosody.

4. CONCLUSIONS

In this paper, we showed, briefly though, that the organization utterance grouping could be achieved to a reliable extent by incorporating our analyses of breaks and boundaries into the existing phrase/sentence models by Fujisaki et al [3-9]. In particular, how continuous running speech is broken into perceivable units larger than phrases and/or sentences and how grouping of phrases can be addressed in prosody organization on the basis of an existing phrasal model. But we are aware that further investigation on how these perceived units are related to speech planning is necessary to refine the proposed organization. Note that by no means we imply from the above results that the planning of such units operating at the phrase level, and will need to look into prosodic properties in more detail [10, 11]. At this stage, we can only speculate that utterance grouping is related to semantic components of subject and theme, and intend to pursue it in future investigations. Further, for a comprehensive organization of prosody, we will need to address issues of temporal organization as well [12] and incorporate it into any model or organization of speech prosody.

REFERENCES

[1] C. Tseng, "The prosodic status of breaks in running speech: Examination and evaluation", in *Speech Prosody 2002*, 11-13 April, Aix-en-Provence, France, pp. 667-670, 2002

 ISBN 1-876346-48-5

[2] C. Tseng and F. Chou, "A prosodic labeling system for Mandarin speech database", *XIV International Congress of Phonetic Science*, Aug.1-9 San Francisco, USA, pp2379-2382, 1999

[3] H. Fujisaki, S. Ohno and O. Tomita "On the levels of accentuation in spoken Japanese." *Proceedings of 1996 International Conference on Spoken Language Processing*, vol. 2, pp. 634-637, Philadephia, USA., 1996

[4] H. Fujisaki, S. Ohno and C. Wang "A command-response model for F0 contour generation in multilingual speech synthesis." *Proceedings of the 3rd ESCA/COCOSDA International Workshop on Speech Synthesis*, pp. 299-304, Blue Mountain, Australia, 1998

[5] H. Mixdorff "A novel approach to the fully automatic extraction of Fujisaki model parameters." *Proceedings of ICASSP 2000*, vol. 3, pp.1281-1284, Istanbul, Turkey, 2000

[6] O. Jokisch, H. Mixdorff and U. Kordon "Learning the parameters of quantitative prosody models." *Proceedings of 2000 International Conference on Spoken Language Processing*,, vol. 1, pp. 645-648. Beijing, China, 2000

[7] H. Mixdorff "MFGI, a linguistically motivated quantitative model of German prosody." *Improvements in Speech Synthesis*, E. Keller, G. Bailly, A. Monaghan, J. Terken and M. Huckvale (Ed.), Wiley Publishers, pp.134-143, 2001

[8] H. Fujisaki "Modeling in the study of tonal feature of speech with application to multilingual speech synthesis." *Joint International Conference of SNLP-Oriental COCOSDA 2002*, pp.D1-D9, Prachuapkirikhan, Thailand, 2002 (Invited papers)

[9] C. Wang, H. Fujisaki, R. Tomana and S. Ohno "Analysis of fundamental frequency contours of standard Chinese in terms of the command-response model and its application to synthesis by rule of intonation." *Proceedings of 2000 International Conference on Spoken Language Processing*, vol. 3, pp. 326-329. Beijing, China, 2000

[10] J. Hirschberg and G. Ward "The influence of pitch range, duration, amplitude, and spectral features on the interpretation of L*+H L H%." *Journal of Phonetics* **20**(2): 241-251, 1992

[11] A. Cutler and D. R. Ladd, Eds. *Prosody: Models and Measurements,* Springer, Berlin, 1983.

[12] K. Hirose and H. Kawanami "Temporal rate change of dialogue speech in prosodic units as compared to read speech." *Speech Communication* **36**(1-2): 97-111, 2002
H. H. Clark "Speaking in time." *Speech Communication* **36**(1-2): 5-13, 2002