Mandarin Speech Prosody: Issues, Pitfalls and Directions

Chiu-yu Tseng

Institute of Linguistics
Academia Sinica, Taipei, Taiwan
cytling@sinica.edu.tw

Abstract

From the perspective of speech technology development for unlimited Mandarin Chinese TTS, two issues appear most impedimental: (1.) how to predict prosody from text, and (2.) how to achieve better naturalness for speech output. These impediments somewhat brought out the major pitfalls in related research, i.e., characteristics of Chinese connected speech and the overall rhythmic structure of speech flow. This paper discusses where the problems stem from and how some solutions could be found. We propose that for Mandarin, prosody research needs to include the following: (1.) characteristics of Mandarin connected speech that constitute the prosodic properties in speech flow, i.e., units and boundaries, (2.) scope and type of speech data collected, i.e., text other than isolated sentences, (3.) prosody in relation to speech planning, i.e., information other than lexical, syntactic and semantic, and (4.) an overall organization of prosody for speech flow, i.e., a framework that accommodate the above mentioned features.

1. Introduction

One major bottleneck in Mandarin Chinese speech synthesis for unlimited TTS is to generate prosody from text, especially when the text comes in paragraphs. This bottleneck also brings forth the question of what constitutes the overall prosody of connected speech flow. It is obvious that producing the right sentence or phrase intonations, as many of our colleagues already had successfully achieved, did not guarantee the right speech prosody. And this also brings forth the naturalness issues, which is very much the other side of the very same coin. Two sets of analysis will be presented to illustrate this point. The next question is: What then are some concrete measures that would facilitate the generation of prosody of connected speech flow and adequately improve output naturalness for Mandarin Chinese? Some discussion will be devoted to this issue, as well as propose possible future directions to improve our understanding of speech flow.

This paper will present 2 kinds of analysis. The first experiment of punctuating text shows grouping of short phrases into complicated sentences is the preferred choice; the choice. However, the grouping varies considerably from one speaker to another. The second analysis of perceived boundaries in a corpus of read speech of well punctuated text showed that speakers' choice of boundaries also varies considerably. These phenomena illustrated perhaps the most inherent properties of Mandarin speech and prosody.

2. Experiments

2.1. Experiment 1

The aim of Experiment 1 was to show that for native speakers of Mandarin Chinese, (1.) punctuation marks serve as a loose reference of syntactic structures as well as semantic domains, (2.) Mandarin speech is characterized by a string of short phrases; simple and short sentences are hard to come by in text. Taken the above with reference to prosody structure and manifestation, how short phrases are grouped should be the key to solve the major bottleneck of Mandarin speech prosody. However, between-speaker variations also showed that our mission was still far from accomplished for attempts made to predict prosody from text.

2.1.1. Methodology and Procedures

Two native speakers, one male and one female, were asked to punctuate two pieces of text whose original punctuation marks had been removed. Both subjects received master's degrees and are fluent readers of Chinese. They were each given two pieces of text composed of frequently used words at 1118 (Text 1) and 1075 (Text 2) characters respectively without any punctuation marks. Each subject punctuated the text independently. Their results were compared. The task may appear quite simple and simple-minded. However, it is worth noting that unlike alphabetic and inflectional writing systems such as English, a logographic writing system such as Chinese does not provide morphological indications to syntactic structures, thereby making punctuation relatively freer and prosody prediction from text all the more difficult.

2.1.2. Result.

Table 1 is the results punctuations used by type and by Ss, and the number of commas within period by subjects. In this experiment periods denote sentence final punctuations and therefore include question and exclamation marks.

	S1			S2		
	# of periods	M of commas	Var	# of periods	M of comma	Var
Text1	27	2.96	8.83	18	3.21	5.39
Text2	28	2.60	1.73	24	2.88	4.11

Table1. Results of the number periods used by each S and the mean number commas within each period. There are 1118 characters in Text 1, 1075 in Text 2

The performance of the two subjects differs both in the total number of periods used and the number of commas within each period. S1 used a total of 107 punctuations for Text 1, 104 for Text 2 whereas S2 used 77 punctuations for Text 1, 93 for Text 2. Note that S1 used more periods than S2 for both pieces of text. The mean numbers of within-period commas also varied. The more periods used in, the number of complete sentences generated increased, and the less numbers of commas within sentences. The simple comparison results indicate the following: (1.) Ss varied considerably in their grouping of syntactic as well as semantic units. (2.) Both subjects showed no inclination for choosing simple sentences, always grouping phrases into sentences. (3.) The grouping of phrases into complicated sentences varies between subjects. In fact, the result implied clearly that in unlimited TTS, intonation types in accordance with sentence types are definitely insufficient. This preliminary analysis implies that the preferred sentences consist of at least 3 phrases. Identifying intonation groups is necessary; coming up with some kind of prosody framework to accommodate complicated sentences of 3 phrases and above, and represent the grouping effect with some prosody devices is more than desirable.

Table 2 shows further analysis of matched punctuation marks across Ss. By matched I mean identical punctuations at identical locations between Ss. S1 used 107 punctuations for Text 1 and 104 for Text 2. S2 used 77 punctuations for Text 1 and 93 for Text 2. Of the 107 punctuations used by S1 for Text 1, only 53 (49%) matched with S2's choices. Of the 104 punctuations used for Text 2, only 64 (61%) matched those of S2's. Identical punctuations at identical locations for each piece of text, indicating the likelihood of identical boundaries, turned out to be low.

Text 1	Match (53 in total)		Mismatch (59 in total)		
	Commas	Periods	Same location different mark	S1-only Marks	S2-only Marks
# %	40 75.4%	13 24.6%	19 32.2%	35 59.3%	5 8.5%
Text 2	Match (64 in total)		Mismatch (54 in total)		
# %	48 75%	16 25%	15 27.8%	25 46.3%	14 25.9%

Table2. Comparison of matched vs. mismatched punctuations used. S1 used 107 punctuations for Text 1, 104 for Text 2. S2 used 77 punctuations for Text 1, 93 for Text 2.

2.2. Experiment 2

The aim of Experiment 2 was to show that for native speakers of Mandarin Chinese, when reading punctuated text, they also tend to group phrases into larger units that are identifiable by prosody. Earlier investigations [1, 2] showed the grouping effect in speech production is perceptually consistent across listeners, and the perception of boundaries is also consistent across listeners. Furthermore, the boundaries and the respective following silences are systematic. Experiment 2 gives more comparisons across speakers to bring out the inter-speaker variation issue, and hopefully illustrates the importance of the problem.

2.2.1. Methodology and Procedures

Four native speakers of Mandarin Chinese, 2 males and 2 females, produced speech data for a speech database. Each speaker read a total of 599 paragraphs of text. The text was hand-tailored text consisted of most frequently used words from the CKIP database (http://godel.iis.sinica.edu.tw/CKIP/), and balanced for phonetic and tonal distributions. The paragraphs consisted of simple 2-character sentences up to 181-character complicated sentences. The focus was on longer complicated sentences. The speech data was microphone speech recorded in sound proof chambers. Table 3 summarizes the speech data used for Experiment 2. We developed a break labeling system in TOBI spirit [1], and the speech data used in Experiment 2 was labeled independently by 3 transcribers for perceived boundaries and breaks (pauses). We have also shown that both intra- and inter-transcriber consistencies could be achieved [2], proving that the perception of the silent portions in running speech is systematic.

Speaker.	Total Characters/Syllables.	Analogue Data	Labeled Digitized Data
F01	24,803	176m	322MB
F03		151m	277MB
M01		151m	276MB
M02		123m	226MB

Table3. The 599 paragraphs of text consist of 24,803 characters in total, corresponding to the same number of syllables. The total duration of speech data and corresponding size of digitized and labeled speech is also shown.

2.2.2. Result

We propose the canonical prosody framework as a hierarchical organization both in terms of prosodic units and their following breaks. In other words, the higher the prosodic unit is, the longer the following break should be. This framework would specify a PW (prosodic word) followed by B2 (Break 2); PPh (prosodic phrase) by B3 (Break 3); UTR (utterance) by B4 (Break 4); and PG (prosodic group) by B5 (Break 5). At the current state, we have not yet obtained analysis on each prosodic unit. However, based on the result of break labeling, we aligned the location of perceived breaks and calculated the overlaps across speakers. Table 4 shows the results of occurrence of aligned breaks across 4 speakers.

	Number of Cross-Speaker Overlap
PW/B2	1643
PPh/B3	1164
UTR/B4	2
PG/B5	220

Table4. Overlap of perceived breaks across 4 speakers.

Table 5 shows the percentage of overlap in each speaker's speech data. The left panel under each speaker is the actual number of labeled breaks, the number in the left panel under each speaker in black is the actual occurrence of labeled breaks; the number in the right pane in purple is the percentage of overlapped breaks, as shown in Table 4, in proportion to the actual occurrence.

	F01		F03		M01		M02	
	Occur rence	Ovlp %	Occur rence	Ovlp %	Occur rence	Ovlp %	Occur rence	Ovlp %
PW/B 2	4555	36	4221	39	5365	31	4997	33
PPh/B 3	3526	33	3656	32	2560	45	4267	27
UTR/ B4	182	1	110	2	192	1	247	1
PG/B 5	460	48	516	43	447	49	543	40

Table5. Total count of each labeled break for each speaker and percentage of the cross-speaker overlap in relation to total occurrence.

The low overlap across speakers showed that speakers varied considerably in speech production.

3. Discussion

Over two decades of speech synthesis research for Mandarin Chinese has made quantal leaps in advancement. But some major bottlenecks remain. At the prosody level, two issues remain unsolved and hence deserve proper attention. The first issue, best exemplified in previous collective efforts towards unlimited TTS, looks for tools that could predict prosody from text in spite of punctuation denotation. The second issue, best exemplified in synthesized speech, looks for ways to improve the overall flow output naturalness. These two issues, in fact, bring us to face the next questions: why is it insufficient after generating sentence intonation? What really work for Chinese? If both Chinese declarative sentences and yes-no questions also exhibits similar intonation patterns as their counterparts in English [3], we are still unable to come up with the right prosody for punctuated text after we tackled the segmental, tonal and intonation aspects of Mandarin? Why it is that connecting sentence and/or phrase intonations doesn't produce the right kind of flow of speech output?

Experiments 1 of text punctuation showed clearly that phrases and simple sentences are not the preferred units for Chinese. The simple statistics from Experiment 1 suggests that 1.) Punctuations serve only as a loose reference to syntactic structures as well as semantic domains. Nevertheless, a unit that is marked by the final period, question mark or exclamation mark, is a unit that consists of a string of commas within. This phenomenon is often seen and Chinese and reflects in spoken Chinese as well. (2.) Simple and short sentences, as often seen in intonation studies, are actually hard to come by in text. Instead of speaking in sentences, Chinese speech is characterized by a string of phrases. (3.) The high non-overlap of punctuations is evidence of variation that also deserves attention.

The results of Experiment 1 suggest that a possible and somewhat optimal Chinese sentence may consist of 3 to 4 short phrases and a larger unit emerges. Taken this result with reference to prosody structure and manifestation, how short phrases are grouped may very well be the key to solve one of the major bottlenecks of Mandarin speech prosody. This string-of-phrases phenomenon may appear language specific on the surface, but is fact comparable to complicated sentences in languages like English. However, since previous collective efforts of the field dwelled on sentence intonation generations, focusing on short and simple sentences, the canonical form of generation stopped short at this level. A major pitfall in Mandarin prosody research is the lack of a prosody framework that addresses complicated sentences, especially emphasizing the grouping of phrases In other words, the inadequacy that stemmed from concentrating on sentential and/or phrasal intonations without a larger and higher prosody. This short-phrase oriented approach would invariably result in the outcome of short and somewhat choppy prosody that lacks the flow in running speech.

Predicting prosody from text would still need more understanding of the interrelation among prosody, syntax and semantics. However, characterizing PG related characteristics is relatively less difficult. Our recent attempt [4] showed that a sentence intonation model such as Fujisaki's [5,6,7,8] could be elaborated to accommodate PG related phenomenon. By including more than one phrase into a PG while specifying the prosodic characteristics of PG-initial, PG-middle and PG-final acoustic features, and the correspondence between prosodic units and following breaks, a rough approximation of the canonical PG could be achieved.. The advantage of this approach is several folds. It saves the effort to propose a separate model, thereby preserving the physiological aspects of the model that not only addresses articulation and speech production in general, but also distinguish the model from other attempts of tonal or prosody modeling [9,10]. It also leaves the model free to produce a single phrase or a complicated sentence like a PG. Moreover, there is still room for speech planning related features in the future.

As for the role of intonation in Mandarin Chinese, we postulate that phrasal intonation is a component of a larger prosody unit and is only significant within the overall organization of prosody for connected speech. In our corpus analysis, as required for the kind of statistics performed in Experiment 2, collecting and analyzing sentences of different types and more importantly longer duration proved to be more enlightening. How would a speaker break a 181-character complicated sentence, marked by commons in between and one period at the very end? The function of this larger unit surfaced. Speakers grouped phrases into prosodically and semantically identifiable units characterized roughly by different degrees of F0 reset at the PG-initial and PG-final positions, different degrees of final tapering of F0 and weakening of amplitude, and various degrees of breaks between prosodic units. The significance of phrasal intonation is better manifested in the organization of PG-governed speech prosody for connected speech. For development in speech technology, the flow of speech output can also be better described in light of PG from two perspectives. For speech synthesis, how phrasal intonation can be manipulated and modified as a component within a larger prosody unit. For speech recognition, how intonation and prosody organization may function perceptually in relation to semantics and speaker's intentions.

However, acknowledging and accepting a PG-oriented framework is but the first step. The mismatches in Experiment 1 and the non-overlap in Experiment 2 can be discussed together. Both results point to the direction of issues of speech variation, reflecting different planning and intention of the speakers. So obvious is the fact that speakers tend to plan speech output differently. The tasks in speech technology development would be enhanced if more understanding in this respect is available. In other words, these variations deserve proper attention. At this fairly preliminary stage, these results already indicate little agreement among speakers in where boundaries were assigned and the kind of breaks followed in actual production, although each speaker operated with the same kind of framework. It is quite obvious that these variations were significant in the make-up of overall rhythm of speech output, constituting in part output naturalness as well. For technology development, it may very well be the case that the canonical form of PG may serve as a possible base form for prosody generation. But further and somewhat sufficient simulation of variation may very likely render better generation of more natural speech flow.

4. Conclusions

Much of the attention on Mandarin prosody investigation and generation had been given to intonation phases or sentences in isolation. For example, how yes-no questions possess overall higher register in declarative sentence [3] whereas relatively little attention on characteristics of Mandarin prosody. This usually resulted in intonation of sentences under 10 syllables and almost all the time implied, perhaps inexplicitly that connected speech could be seen as connecting isolated sentences in succession. Researches in TTS have demonstrated that this kind of approach is hardly sufficient for the generation when the text is unlimited. This paper shows one of the major prosody related characteristics of Mandarin Chinese. Namely, a larger prosody unit PG consisting of more than one phrase or simple sentence appears to be the operating unit in connected speech. The organization of PG, stressing the grouping of utterances into identifiable unit and reflecting global planning of speech unit, is essential to characterize Mandarin prosody. Consequently, phrase- or sentence-intonation contours specification is less significant. These intonations, in turn, should be seen as units within a canonical prosody form; their characteristics and manifestation vary in relation to their relative positions within the PG. The present study demonstrates that (1.) larger units that imply overall planning of speech output must be taken into consideration. (2.) Phrasal intonations of tone languages are not as significant as they are in intonation languages unless specified within PG. Future directions should definitely include correlations and mismatches between prosody units to syntactic and semantic analyses, and address characteristics of the referred global planning. More understanding of speech rate in relation to global planning and PG should be desirable. Furthermore, how to fit variations into a prosody generation framework, via what kind of mechanism, and how much variations should be considered, in speech synthesis, even minimally, should collectively enhance better speech output in the future.

5. References

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

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

[3] Lin, M-C, “Hanyu yunlyu jiegou han gongneng yudiao (Mandarin prosody organization and functional intonations, in Chinese)”, *Report of Phonetic Research* 2002, Phonetics Laboratory, Institute of Linguistics, Chinese Academy of Social Sciences pp. 7-23

[4] C. Tseng, “Towards the organization of Mandarin speech prosody: Units, boundaries and their characteristics”, *XIV International Congress of Phonetics Science,* Aug.1-9, 2003, Barcelona, Spain.

[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] Y. Xu “Articulatory constraints and tonal alignment” *Speech Prosody 2002*, 11-13 April, Aix-en-Provence, France, pp. 91-100, 2002.

[10] J. van Santen, C. Shih and B. Mobius “Intonation” *Multilingual Text-to-Speech Synthesis: The Bell Labs Apporach,* R. Sporat Ed. Kluwer Academic Publishers, Ch. 6, pp.141-1901998