Tentative planning of prosodic rules for the naturalness of synthetic spoken Chinese

Wu Zongji

试论合成普通话口语自然度所需的韵律特征规则

吴宗济

[摘要] 普通话语音合成的自然度，主要取决于韵律特征的处理是不是理想。近年来各语言信息处理方面，对普通话的单字和短词的合成质量，在清析度、可懂度方面，都有不小的进展；但对连读的韵律特征及成句的语调的处理，由于口语中音变的复杂性，以及语音知识与言语工程的不配套，以致难有较大的进展。

语言学大师赵元任先生早在 1928 年，根据吴语方言调查中的研究，对于口语变调的复杂性，指出分析语调的两个问题：一是说话人的“音高”（调形），因语境、情感而变；二是说话时的“音程”（调域），因生理、情绪而异；他认为“暂时还没有想到好法子把材料改成一致的。”他此论发表了四十年之后，在 1968 年出版的《中国话的文法》中指出，汉语的语调“只是基调的差别，而不是象英语那样上升或下降的曲线，”在今天看来，他对当年的难题已有了解决的曙光了。因为语调既是基调的变化，那就比曲线的升降简单得多，可以从而探索规律了。我们近十年来对普通话声调的实验，证明了两点：(1) 连词或短语单元的调型，在不同的语调中总是稳定的。(2) 语调的变化，是由于连词或短语基调的变化；而不是调形拱度变化。韵律的三特征：音高、音长和音强，都单独或共同对口语的自然度起作用，尤其以音高的变化（包括字调、词调、句调）为主，在这方面，已经发表了一系列的短语连读变调的守恒规则。近来的语调分析、合成实验，又对语调中的基调作移调处理，以及对语调高低，和调域宽度，改用半音音程计量，都取得了一致性的规正结果。可以说对赵先生关心的难题，大致有了解决的希望。

本文就是在前述的基础上，对普通话合成中与自然度有关的短语单元和语句的韵律特征的处理规则，作扼要的叙述。试图对合成自然度质量的改进从语音学的范围提出一些建议。文内对短语中及短语间的韵律特征协同发音处理规则，分别列表，并对目前通行的两种合成系统的优缺点，略予评述：一种是参数合成系统，音色和韵律都能自由处理。但因语音的每一细节都需由程序生成，规则非常繁复；设计稍有粗疏，自然度就差，而且成本也高。另一种是音素编辑系统。由于是调用人声的录音材料，清晰度就高，也比较经济。但因其是音波拼接，协同发音中由于同化作用，需要改动音色时，就很困难，因此也有个提高自然度的问题。据闻此点在最近已有较好的改良方法，如能成功，将可望进一步提高这种合成系统的质量。

Abstract

The naturalness of the synthetic Spoken Chinese (SC) is much more relied on promising processing of the prosodic features. Our previous experiments have proved that (1) the patterns

of phrasal contours (PC) units in different sentential intonations are rather consistent; for which, tone-sandhi rules of di-, tri- and quadro-syllabic combinations were presented, and (2) the variety of sentential intonations are mainly relevant to the modification of phrasal registers (PR) instead of that of PC. Transposition of PRs in different keys may generate sentence intonations of various moods. In this paper, selected prosodic rules and plans in favour of the both processings are given, in an attempt to suggest a way of improving the naturalness of synthetic spoken Chinese from phonetic aspect.

1. Introduction

The quality of the articulatory index and the intellibility of synthetized Standard Chinese (SC) words and short phrases are quite acceptable now in several labs of speech processing. However,as far as we know, the level of naturalness in synthetized spoken Chinese is still in its infancy. Apart from the fact that the analysis of transitional cues in syllables had been an old story in speech processing, the co-production effects between poly-syllabic words and phrases are now playing an important role for perceptual naturalness in spontaneous speech under various attitudinal intonations. Since the inter-syllabic tone-sandhi phenomena within the phrasal contours (PC) related to the lexical, as well as sentential poly-syllabic structures in sentences of different moods are rather consistent in pattern, thus a series of PC patterns based upon poly-syllabic tone-sandhi rules were presented (Wu 1982;1985;1988a;1988b;1990), and had been practiced in certain synthetic processings (Yang 1994). As for naturalness, the prosodic features of duration and stress other than tone also shared the co-production effects on inter-syllabic as well as inter-PC's boundaries (Wu 1995). Thus exploring the field of sentential naturalness based on global prosodic or suprasegmental co-productions becomes a main task in qualified synthetic speech nowadays.

2. Prosodic rules for phrasal contours

The three prosodic features, namely: pitch, tempo and stress have been defined as suprasegmentals by phoneticians over decades. However, as a matter of fact, the tone plays double roles in speech: i.e. lexical identity in words and clauses as the toneme and attitudinal diversity in sentences as intonation. In addition to the four mono-tonemes in SC, there are a lot of tone-sandhi patterns for poly-syllabic phrasal contours being the building units of sentential intonation, in which the other two prosodic features may not be the key factors. However, the syllable length might be shortened in certain words in poly-syllabic phrases, i.e., the second syllable in di- and tri-syllabic phrases and the second and third in qudro-syllabic ones when in sentences. And a series of phonological rules representing how the surface tone-sandhi contours derive from underlying forms had been given (Wu 1990). In a sentencial intonation contour, there are two types of prominent PCs that might have occured. (1) " A phrase or a first clause in a composite sentence is in a slightly higher key than a concluding phrase or clause", as Chao stated (Chao, 1968). It is usually defined as declination of a sentence contour. (2) A phrase or clause which might carry the focus information of the sentence topics becomes prominent. It bears a

name of "logical stress" in Chinese grammar. Thus they have the keys higher than the other prosodic units with different extents according to the various attitudes.

In a sentence, for the sake of more economical processing, a PC may group more than one units of PCs with a common key. Thanks to the consistency of the PCs regardless of the variable intonations, the processing of the shifted PC's keys can be realized just by a simple method of "frequency transposition" of the key like what a musician does when composing scores. Furthermore, as has been stated by I.Lehiste: "the problem of relating contour-like movements to musical intervals seems to be less relevent for a study of English than for a study of tone languages" (Lehiste, 1976). It means that Chinese tonal contours had better relate to musical intervals instead of to the F0 scale. This had been discussed fully in a paper based on the Chinese traditional concept of tonal transcription found in historical works (Wu 1995a). Also, some experiments had been arranged on transposition of the key relating to the prominent PCs in collected spoken SC sentences in different moods (Wu 1993), and also to the range-width in lab-built 4-tonemic tokens under different given keys (Wu 1994). Both the results showed that the range-width as scaled by musical semitone are always constant even under different keys but are always differed by frequency scale.

It is not the whole story after the intra-PC's prosodic processing had been well done. Attention should be paid to that the boundaries between any two tone-sandhi contours or two PC units that would have to shift the prosodic features of the ending of the former syllable and the starting of the latter one due to the CP effect. A gliding contour may take the place of the boundary being a prosodical juncture following a "gang-board-like" gliding rule (Wu 1985). It is not neccessary to give all the prosodic rules required for the naturalness improvement at the moment. A brief list of the planning rules for prosodic CP are given in the following section.

3. List of co-productions for synthetic naturalness in brief

The co-production effects in sentences can be divided into two groups: segmental and supra-segmental. The former includes the CP among the phonetic segments that usually appear as allophones or under-specifications of the consonants and vowels thus giving the naturalness perception. It will not be discussed here. The latter, which is more complex, includes the distribution of any or all of the three prosodic features in certain proportion. Here is a brief list of what rules are required to satisfy the co-production effect, by one or more modifications.

(1) Two types of CP:

A. Opened CP acted on:
   1.initial of a sentence or of a clause.
   2.ending of a sentence or a clause.

B. Closed CP acted on:
   1.inter-syllables
   2.inter-PCs

(2) Representation of CP:

A. CP in initials:

1.voiceless → voiced.

2.fricative → shortened

3.nasal → nasalized.

4.semi-vowel → voiced fricative or glottal stop.

5.aspirated → shortened or -V deleted.

B. CP in finals:

1.mono-vowel → centralized.

2.diphthong:1st vowel dominant → last vowel undershot

3.diphthong:last vowel dominant → last vowel centralized.

4.nasal-ending → nasalized.

5."-n" + back C or V → "-ng".

6."-ng" + front C or V → "-n̂".

C. CP in tones:

1.high → lowered or falling.

2.low → upward or rising.

3.rising → leveling or falling.

4.falling → leveling or rising.

5.citation tone → underspecified or neutralized.

6.tone group → transposition of key.

D. CP in duration:

1.long → shortened.

2.short → lengthened.

E. CP in loudness:

1.stress → weakened.

2.weak → enhanced.

(3) Occuring of CP:

A. CP in di-syllabic PC:

1.2nd syllable weakened (neutralized,shortened,under-specified).

2.1st syllable weakened (ditto).

3.well-formed → under-shot.

4.1st tone of Tone3 + tone3 → 2nd tone.

B. CP in tri-syllabic PC:

1.Structure: (1+[1+1]) → [1+1] = CP rule: A.2.

2.Structure:([1+1]+1)→[1+1] = CP rule: A.1.

3.In ([1+1]+1), [1+1]=[tone3+tone3]: 1st tone3 → tone2.

4.In any structure, 2nd tone → under-shot or weakened.

5.2nd tone → gliding contour (GC) after "gang-board" rule.

6.Last tone → ending drift.

7.1st and 2nd tone are high level or high rising → "gang-board" rule.

C. CP in qudro-syllabic PC:

1.Structure: ([1+1]+[1+1]) → CP rule: (A.1 + GC + A.2).

2.Structure: ([1+1+1]+1) or (1+[1+1+1]) → CP rule: C.1

3.2nd and 3rd tone → under - shot,weakened

4.last syllable → ending drift.

(4) Transposition key of PC in sentence:

1.Each value of the transposition of one semitone interval = every next frequency × 1.0946.

2.Extent of range-width in PCs in a sentence ≤ 12 semitone (according to speaker's voice register)

3.Extent of range-width in a sentence contour ≤ 24 semitone intervals. (according to speaker's expression)

4.The musical key corresponding to measured frequency of PC can be found in any "Pitch Names and F0" converting table, and the measured frequency may correspond to the nearest value.

5.The declination of sentence contours also exist in SC

(5) Distribution and proportion of the prosodic features in PCs

1.The tone pitch and the duration can either bothly or solely share the prominence of PC.

2.The stress should be always accompanied by tone pitch to give the prominence of PC.

4. Discussion

The naturalness of spoken Chinese means the fluency of the speech that is based on the successive movements of articulatory mechanism. A speech sound can be segmented into discrete symbols to represent the words as letters or Chinese characters rather than the actual sound. An educated speaker can read the phrases and sentences in a text of discrete linguistic symbols, and go through fluently every "filling" of coarticulation or transition, gliding correctly and smoothly those symbols, according to the tongue's and muscles' displacements without hesitation. However, those "fillings" are not easy to be presented accurately in synthesis automatically, and only a programmer of higher knowledge and technique might have achieved this. Therefore, those prosodic rules in the list cannot give comprehensive CP items satisfying the required naturalness. This shortcoming can only be improved by more and more experiments.

The declination of a sentence contour made the difference of pitch between the beginning and ending of the sentence, As Chao (1968) also stated: "[s]uch a difference in pitch differs from a similar difference in English in two respects: first, the difference in Chinese is slight, and, secondly, the difference is one of key and not an upward and downward sweep, as is often the case with English." That this declination is not a sweep as said by Chao is quite true. In our analysis and

synthesis, due to the consistency of PC's pattern, this declination is going through the PC units by cascading. So it is easy to be done by negative transposition of each PC's key step by step.

In the case of there being many synthesis tools designed by different systems, certainly the present planning cannot satisfy their total demands. For example, the synthesis of parametric system may work very well for the processing of phonetic quality at will, but for the naturalnes it would have spend much work on the programming. While in the syllable editing system, the sound quality of simple words may reach the higher fidelity, but the segmental data are hard to be modified on the waveforms. So the prosodic features in this system may be easy to arrange, but it is difficult to smooth the "break-off" in perception, especially when having a vocalic ending immediatly followed by a syllable of voiced or "zero" consonant due to the zig-zag of diffrent articulatory positions.

Last, but not the least, it is interesting to learn from a recent paper by Lehiste who reported an experimental result on the percieved prosodic prominence of different language background. "Increased amplitude is a stronger stress cue in English." and "increased duration is a relative stronger cue in Estonian and Swedish" (Lehiste 1994). Now, what will be the prominent cue in a Chinese dialect as percieved by a listener with the same language background? Probably most of the phoneticians would think that the increased tone pitch is the answer. Nevertheless, a few of prosodical analyses of recorded spoken SC sentences were made in our laboratory, in which most of them show that the prominence is related to the increasing of tonal frequency in ordinary declarative sentences. However, in some emotional sentences, other results were obtained. "A number of tentative experiments on speech synyhesis had shown that in a synthesized exclamatory sentence the absolute register of a prominent PC raised to a certain extent is not significant in loadness unless a certain degree of amplitude and/or duration are added to an appropriate proportion (Wu 1995b).

References

Chao, Yuenren (1933) A Preliminary study of English intonation (with American varients) and its Chinese equivalent. In Studies presented to Ts'ai Yuan P'ei on his sixty-fifth birthday, part one, NIHP, Academia Sinica.

Chao, Yuenren (1968) *A grammar of spoken Chinese*. Berkeley: University of California Press.

Lehiste, I. and G. E. Peterson (1976) Some basic considerations in the analysis of intonation. *Acoustic phonetics: a Course of Basic Readers* (D. B. Fry, editor), part 4, Investigation of prosodic features. Cambridge: Cambridge University Press.

Lehiste,I. (1994) Language background and the perception of prosody. *Proceedings of the International Symposium on Prosody*, pp. 65-74. Yokohama: The Japan Society for the Promotion of Sciety.

Wu, Zongji (1982) Rules of intonation in Standard Chinese. In Preprints of papers for the working group on Intonation, pp. 95-119, the 13th international congress of Linguists, Tokyo.

Wu, Zongji (1985) Tone-sandhi rules of tri-syllabic combinations in standard Chinese. In *Bulletin of Chinese Linguistic Society* (Zhou Zumo, Xing Gongwan & Wang Jun, editors),

pp. 70-92. Beijing: The Comercial Press.

Wu, Zongji (1988a) The basic tone-sandhi patterns in standard Chinese intonation. In *Essays on Linguistics, Feschrift for Prof. Wang Li*, pp.54-73. Beijing: The Commercial Press.

Wu, Zongji (1988b) Tone-sandhi patterns of quadro-syllabic combinations in standard Chinese. *Report of Phonetic Research*, Institute of Linguistics, Chinese Academy of Social Sciences.

Wu,Zongji (1990) Can poly-syllabic tone-sandhi patterns be the invarient units of intonation in spoken standard Chinese? *Proceedings of the 1st International Conference on Spoken Language Processing*, section 12.10.1. Kobe: The Acoustical Sciety of Japan.

Wu,Zongji (1993) A new method of intonation analysis for Standard Chinese: Frequency transposition processing of phrasal contours in a sentence, *Report of Phonetic Research*,1-18.

Wu,Zongji (1994) Further experiments on spatial distribution of phrasal contours under different range registers in Chinese intonation, *Proceedings of International Symposium on Prosody*, pp. 65-74. Yokohama: The Japan Society for the Promotion of Science.

Wu, Zongji (1995a) Predictability of different attitudinal intonation in standard Chinese. *Proceedings of the 13th International Congress of Phonetic Sciences*(Kjell Elenius & Peter Branderud, editors), **3**, 726-29. Stockholm: KTH & Stockholm University.

Wu, Zongji (1995b) A new method forintonation processing of standard Chinese based on the relation between tonology and musicology (in press).

Yang, Shun'an (1994) *The Chinese Synthesis Technique Orienting Acoustic-Phonetics*. Beijing: The Documental Press of Social Sciences.