RC22389 (C0204-002) April 1, 2002
Other

IBM Research Report

Issues in Speech Synthesis for Tonal Languages

Fangxin Chen
IBM Research Division
China Research Laboratory

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

Issues in Speech Synthesis for Tonal Languages

Fangxin Chen

IBM China Research Laboratory
e-mail: chenfx@cn.ibm.cn

Abstract

This paper discusses linguistic and acoustic issues related to tonal language speech synthesis. Topics covered include physiological constrains on lexical tone production and perception, lexical tone domain, lexical tone rules, lexical tone and intonation interaction, and intonation generation models for tonal languages.

1 Introduction

To get more natural synthesized speech, prosody needs to be predicted, based on linguistic, acoustic and statistical analyses at the syllable/word/sentence level. For tonal languages, there is another dimension of complication for prosody prediction, as compared with non-tonal languages. That is the lexical tone, which is part of the syllable structure for conveying linguistic meaning, and is superimposed on the general intonation contour.

In general linguistic description, lexical tones are usually represented by simple high (H) or low (L) pitch level, or falling/rising pitch contours. In real speech production, lexical tone generation is a more complicated phenomenon, which is decided by various factors, such as the speaker's physiological characteristics, the syntactic and semantic context effects, and the interaction between the lexical tone and the general intonation contour. This paper discusses those aspects that are particularly relevant to tonal languages in speech synthesis.

2 Articulatory Constrains on Lexical Tone Production

First, there are articulartory constrains on lexical tone generation. As it is known, pitch is the primary acoustic cue for lexical tone perception, and it is associated with the vibration of vocal folds. According to the Myoelastic Aerodynamic Theory of Phonation, there are several factors affecting the vibration of vocal folds.

Myoelastic refers to the ways in which the muscles(myo-)change the vocal folds elasticity and tension to effect changes in frequency of vibration. Since the muscles control the vocal fold vibration in a linear way, we expect that the pitch will not change abruptly. The implication for lexical tone production is that tone change is gradual and there will be tone coarticulation if the two tone targets are distant in pitch level. In addition, muscle control of vocal fold vibration is constrained by the speaker's own physiological construct, such as the vocal fold length, elasticity etc. We expect that the pitch will change only within certain range for a particluar speaker in normal speech conditions. For intonation generation, then we need to find out the baseline and the topline F_0s of a particular speaker in his or her pitch fluctuation.

Aerodynamic refers to the fact that vocal folds vibration is activated mainly by the Bernoulli Effect of subglottal air pressure rather than by nerve impulses. Because of the gradual dropping of the subglottal air pressure during the phonation in normal speech, it is expected that the pitch gradually drops too. The aerodynamc theory interprets the pitch declanation observed in natural speech. A speaker tends to pre-plan his/her speech before real phonation. A clause or sentence is normally a linguistic unit associated with a breath. Consequently, there will be F_0 declenation at the clause or sentence level, and

F_0 resets at the starting of a new clause or sentence.

3 Lexical Tone Perception

There are also auditory constrains on lexical tone perception. According to the perceptual experiment [1], there is a minimum tone length requirement for perceiving a contour tone, which is about 40 ms. The implication of this auditory constrain on contour tone perception for speech synthesis is that for a normal male voice, a synthesized syllable needs to be about four periods or more to have a falling or rising tonal quality. For a female voice, the required periods could be doubled due to her higher F_0.

Besides auditory constrains, we need to investigate what are the essential acoustic cues for lexical tone perception. In other words, what are the acoustic dimensions of the lexical tone?

There are language universal dimensions for lexical tone perception. For contour tone languages such as Standard Chinese (SC), Thai, Cantonese, etc, pitch height and contour are the primary acoustic cues; while for register tone languages, pitch height is the primary acoustic cue.

On the other hand, there could also be language specific dimension(s) in lexical tone perception. In other words, those dimensions are only relevant to a particular tonal language. For example, energy level is an acoustic cue for the Falling-rising tone (Tone 3) perception in SC; and tone length is an important cue for distinguishing the level tones in checked and non-checked syllables in Cantonese. However, either energy or tone length are not necessarily essential for tone perception in other tonal languages. In case of tone length, for example, it in fact affects not much on the tone perception in SC, if the minimum tone duration condition for each tone is satisfied.

With the knowledge of lexical tone dimensions, we are possible to manipulate the acoustic value in speech synthesis. For SC, for example, we need to reduce the energy level for Tone 3 syllable; while in Cantonese, we need to shorten the syllable length for checked tones.

In addition, there are interrelationships among the acoustic cues for contour tone perception due to auditory properties. To perceive a rising or falling tone, the required F_0 slope is inversely proportional to the syllable length and directly proportional to the logarithm of the pitch height. In speech synthesis, then, to generate a rising/falling tone, the F0 slope needs to be steeper either the syllable length is short, or the tone starts at a high pitch level.

For auditory constrains on tone perception, there is another important implication for lexical tone synthesis. In real speech signal, we find there are various tone variations due to phonetic context. People argued about weather it was necessary to model those variations in speech synthesis. In general, we can say that if the tone variations have durations that exceed the perceptual threshold for perceiving a steady tone, then the pitch contour need to be modeled. Otherwise, we could lose tone quality. If the micro variations are only transient, then there is no need to model them meticulously, because those pitch variations are perceptually negligible.

4 Lexical Tone Domain

In a tonal language, lexical tone is part of the syllable structure. For lexical tone assignment in speech synthesis, it is important to know where to place the F0 contour in the syllable. Since SC is used for tone domain investigation, a brief description of SC syllable structure is given here to facilitate the ensuing discussion. The SC syllable can be decomposed as follows:

Syllable ⟶ Initial+Medial + Rhyme:
Initial ⟶ Consonant;
Medial ⟶ Glide (j, w,);
Rhyme ⟶ Nucleus + Coda;
Nucleus ⟶ Vowels;
Coda ⟶ Nasal(n, N) or Retroflex r.

According to the perceptual experiment conducted on SC lexical tone domain [1], removal of the Initial or the Medial, or both the Initial and Medial from the syllable did not affect lexical tone perception; while elimination of the Coda part seriously affected listeners' tone judgment. This result indicates that, in SC, the pitch relevant to lexical tone perception is contained in the Rhyme part of a syllable. The

pitch perturbation in the initial part of the esyllable is irrelevent to lexical tone perception.

Lexical tone domain is important in both the parametric and concatenative speech synthesis. For parametric TTS engines, such as the formant synthesizer, lexical tone implementation involves the alignment of tone targets with the correnponding syllable structure. For SC, it should be the Rhyme part of the syllable, which is the domain of lexical tones. For concatenative TTS, the implication of lexical tone domain is that the Rhyme part should not be broken into smaller parts in considering the concatenative units. Otherwise, it could possibly damage the tone integrity. In concatenative TTS systems for non-tonal languages, the commonly used synthesis units are phone, biphone or demi-syllable. For tonal languages like SC, maybe those units are not appropriate because the lexical tone could be disrupted.

5 Lexical Tone

According to Chao's [2] five-level tone scheme, SC tones can be defined as

Tone 1 : High level : 55
Tone 2 : High rising : 35
Tone 3 : Falling rising : 214;
Tone 4 : High falling : 51;

While the six Cantonese tones can be defined as :

Tone 1: High level: 55
Tone 2: High rising: 35/25
Tone 3: Mid level: 33
Tone 4: Low falling: 21/11
Tone 5: Low rising: 23/13
Tone 6: Low level: 22

Chao's five-level tone scheme is an effective method for lexical tone description. However, it is not a precise acoustic representation of lexical tones in real speech. Figure 1 and Figure 2 are the pitch contours extracted directly from real speech of SC and Cantonese syllables read in citation form. The pitch contours of different tones change in contiuned speech. Even read in citation form, it can be seen that the tone levels are not correspondent to what are defined. For example, Tone 4 starting point in SC for that speaker is much higher than the Tone 1 starting point. They obviously do not start at the same pitch level. For tone contours, the acoustic realization is also in discrepency with the linguistic description. For example, Tone 2 in Cantonese is defined as High rising with pitch scale 35 or 25. However, in real speech, Tone 2 has rather a falling-rising pitch contour.

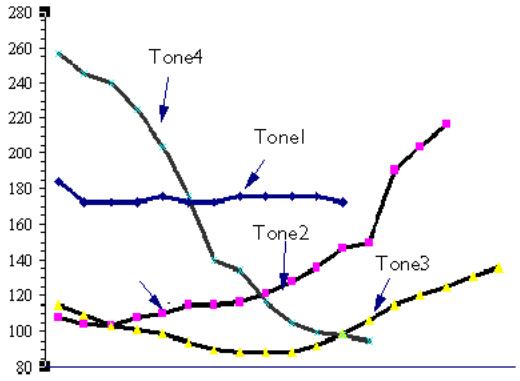

Figure 1 : Four Chinese lexical tones extracted from their respective tone-carrying syllables [TANG]

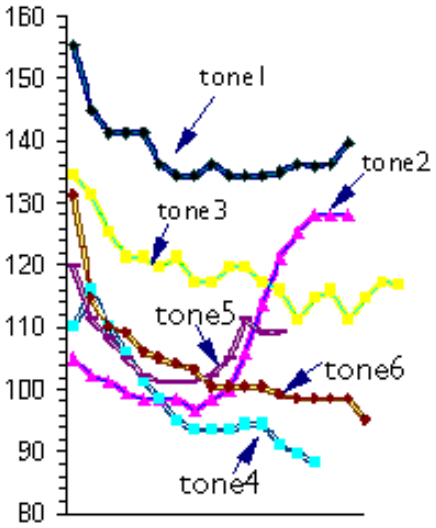

Figure 2 : Six Cantonese lexical tones extracted from their respective tone-carrying syllables [HAAM]

To make the five-level scaling tone patterns closer to real speech tone patterns, Shi [3] proposed an algorithm to approximate the tone levels from the F0 values of the pitch.

$$T=5*\frac{\log_{10} X-\log_{10} b}{\log_{10} a-\log_{10} b}$$

Where a is the topline F0 value in the pitch contour; b is the baseline F0 value; X is the F0 value for the target point. T is the approximated pitch level. T level can be got from following approximation:

0-1.0 → level 1;
1.0 –2.0 → level 2;
2.0-3.0 → level 3;
3.0-4.0 → level 4;
4.0-5.0 → level 5.

This method is useful in mapping a tonal language's real F0 contours into the five-level tone scheme. In TTS, it is also a useful way to map individual speaker's tone patterns into his/her unique five-level representation, which could contribute to the personalized TTS system research work.

6 Lexical Tone Rules

Lexical tone realization in contiuned speech undergoes various linguistic rules. In speech synthesis, it is necessary to sum up and implement those linguistic rules into the lexical tone generation module.

Tone Sandhi rule is defined as the changes in the tone brought about regularly by the effects of adjacent tones. A typical Tone Sandhi rule in SC is the Tone 3 variations. The full Tone 3 pattern is 214, according to the 5-level tone scheme, when the syllable carrying the tone is in citation form, or is stressed and at non-falling phrase intonation end. In all the other phonetic contexts, Tone 3 is read either as 211, or 24 patterns. Following are the Tone 3 Sandhi rules[4]:

In mono-syllabic words :
Tone 3 (unstressed) → 211;
Tone 3 + [%H] → 214;
Tone 3 + [%L] → 211;

In bi-syllabic words :
Tone 3 + [Tone 3] → 24;
Tone 3 + [~Tone 3] → 211;
Tone 3 + [Neutral Tone originated from Tone 3]
→
Optional Neutral Tone → 24;
Obligatory Neutral Tone → 211;
The second char. is concrete word → 24 ;
The second char. is functional word → 211 ;

Tone patterns could also be affected by syllable stress patterns. For example, in SC bi-syllabic words, if two Tone4 syllables are in succession, the tone patterns can be 53+51 if the syllable stress falls on the second syllable; otherwise, the tone patterns are 51+31. Then, the Tone rules could be written as :

In bi-syllabic words :
Tone 4 + Tone 4 [Stressed] → 53 + 51;
Tone 4[Stressed] + Tone 4 → 51 +31;

Tone rules could also come from morphological and semantic factors. In Cantonese [5], for example, the reduplicated adjectives with the [AAB] pattern will change the low tone of the second element into a high rising tone. In that case, the Cantonese tone rule could be written as :

[AA(low tone)B] (Adjective) → [AA(35)B]

eg.
Kam4-kam4-cheng1-> kam4-kam2-cheng1

In SC, another interesting phenomenon for lexical tone realizaiton is the so-called neutral tone(Tone 0]. Neutral tone does not have its own pitch pattern in citation form. The realization of neutral tone depends on its preceding tone context. The neutral Tone rules could be summed up as :

[Tone 1] + Tone 0 → 2;
[Tone 2] + Tone 0 → 3;
[Tone 3] + Tone 0 → 4;
[Tone 4] + Tone 0 → 1;

Lexical Tone and Intonation Interaction

Lexical tone is realized at the syllable level, while intonation at the clause/sentence level. In tonal languages, it is expected that there could be interaction between the lexical tone and intonation, because both rely on pitch contours for its realization. The question is : in what fashion is the interaction between lexical tone and intonation realized in the frequency domain? For example, what will happen to a falling lexial tone in an intonation environment which requires a rising pitch contour, such as at the ending of an interrogative sentence, or a rising lexical tone at a declarative sentence ending, which usually has significant F0 dropping?

Researches for SC show that the basic lexical tone patterns, in the intonation environment, are not affected other than the pitch fluctuation ranges. According to Shen[6]'s interpretation, a speaker's pitch fluctuation range is modulated by the pitch topline and the pitch baseline, which are not constant, but vary with the syntactic/sementic contexts. The topline moves up significantly at the sentence nucleus position. After the sentence nucleus, the topline drops significantly for declarative sentence. The baseline also falls. For interrogative sentence, the topline drops slowly, while baseline rises to certain degree. Figure 3 illustrates the pitch contours of three pairs of SC declarative and interrogative sentences. In the graphs we can see that the basic pitch patterns for lexical tones are kept intact, but the pitch fluctuation range increase, especially at the sentence-ending syllable.

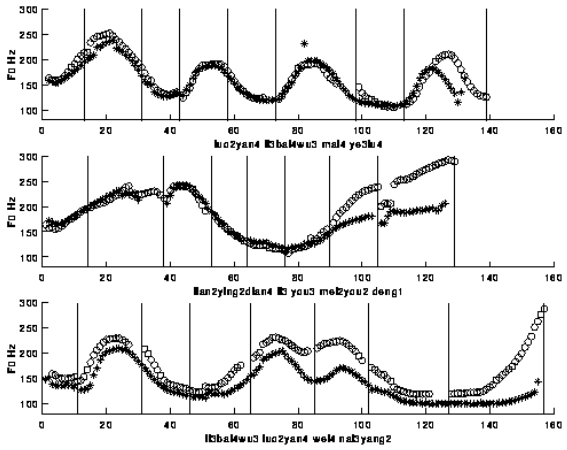

Figure 3 Comparison of the pitch contours for three pairs of SC declarative and interrogative sentences(quoted from Yuan[8]).

One phenomenon worth mentioning is that tonal languages like SC have alternative linguistic resort to express the paralinguistic information other than relying on intonation, so as to avoid the possible confusion caused by lexical tone and intonation interaction effect. SC, for example, has a sentence-final quesiton particle 吗to indicate Yes/No questions. In that case, the sentence intonation is the same as the declarative sentence.

7 Intonation Generation for Tone Languages

Modeling the intonation of a tonal language for pitch contour generation is one of the most challenging tasks in speech synthesis. Here we give a brief description of two popular intonation generation models:

Pierrehumbert's Intonation Model [7] is a structured strings of Low and High tones generated by a finite-state grammar. The strings consist of one or more pitch accents, and a phrase tone following the pitch accent on the main phrase stress, and the boundary tone. Since Pierrehumbert's model was developed from English intonation analysis, it is not clear how this H/L tone scheme could cope with far more complicated pitch patterns in contour tone languages. In Cantonese, for example, there are three level tones: High, Mid and Low; two rising tones: High rising and Low rising, as well as a Low falling tone. The two-tone scheme seems not sufficient for this kind of complicated tone situations.

Fujisaki's Intonation model is based on speech generation mechanisms. In his model, Intonation can be decomposed into two parts: the phrase component and accent component. These two components can be generated separately in the frequency domain by phrase command and accent command, and then superimposed to each other to get the final intonation F0 contour. For tonal languages, Fujisaki simply replaced the accent component with the tone component. Another difference between tonal and non-tonal language implementation in Fujisaki's model is that the

polarity of the accent command can only be positive, while tone command can be both positive and negative. For SC, Fujisaki used a positive command for Tone 1, a negative one followed by a positive one for Tone 2, a negative one for T3, and a positive one followed by a negative one for Tone 4.

Figure 4 is a Chinese Intonation contour generation example illustrated by Fujisaki [9].

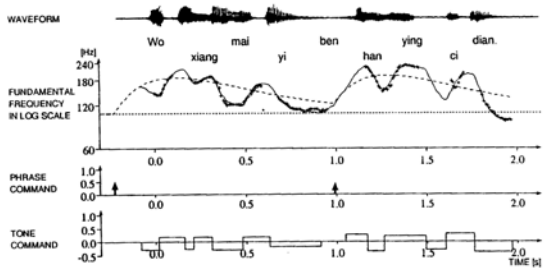

Figure 4 An example of Standard Chinese F0 contour generation

Fujisaki's approach is functional for generating tonal language intonation, because lexical tone can be seen as a separate component of the overall intonation contour, and it can be processed with various linguistic rules before it is superimposed onto the phrase component intonation contour. On the other hand, the phrase component contour can also be handled independently for various sentence types. Like non-tonal languages, tonal languages also have word accent and nucleus at sentence level, The Fujisaki's model for tonal languages may also need the accent command, in addition to the phrase command and tone command. The accent command will handle the pitch change due to word accent and nucleus at the sentence level. As mentioned in the previous analysis, in tonal languages, word accent or nucleus may increase the pitch fluctuation range of the lexical tone, rather than the pitch register. There are also interaction between phrase component and the tone component, which may also affect the pitch fluctuation range. Therefore, how to coordinate those three commands in Fujisaki's model still needs further experimental work.

References

[1] Chen, FX and Rozsypal, A, 1992, Computer Modeling of Mandarin Chinese Tone Perception, ICA 14, Beijing, 1992, G2-6.

[2] Chao, Y-R. 1968, A Grammar of Spoken Chinese, Berkeley: University of California Press.

[3] Shi, Feng, 1994, A few questions on Intonation Analysis, Phonetic Manuscript, Beijing Language Institute Press.

[4] Zhang, B.N, Yang R.H, 2000, Tone Co-articulation in Standard Chinese, Business Press.

[5] Shen, Jiong, 1992, A Tentative Chinese Intonation Model, Chinese Research, Vol. 45, 1992, p16-24.

[6] Mattews, S, and Yip, V, 1994, Cantonese, A Comprehensive Grammar, Routedge.

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

[8] Ladd, D. R. 1996, Intonational Phonology, Cambridge University Press.

[9] Fujisaki, H, 1997, Modeling the Process of Fundamental Frequency Control of Speech for Synthesis of Tonal Features of Various Languages, China-Japan Symposium on Advanced Information Technology, 1997.