Quantitative Analysis and Synthesis of Focus in Mandarin

Gao Peng Chen, Yu Hu*, Ren Hua Wang* and Hansjoerg Mixdorff***

*iFlytek Speech Lab, University of Science and Technology of China
**Faculty of Computer Science, Berlin University of Applied Sciences, Germany
{gpchen@ustc.edu, jadefox@ustc.edu, rhw@ustc.edu.cn, mixdorff@tfh-berlin.de}

Abstract

This paper analyzes the influence of narrow and broad focus on the F0 contours in Mandarin utterances. We employ the Fujisaki model as a means for parameterizing F0 contours. We can regard the F0 contour in Mandarin as a superposition of slowly and fast changing components. Whereas the slow phrase component is associated with declination, the fast tone component is the manifestation of the syllabic tone shape. If there is narrow focus on a word, the amplitude of tone commands (of positive and/or negative polarity) aligned with it increases. The amplitudes of commands on the preceding word increase likewise whereas the amplitudes of commands on all following words are reduced. When focus broadens, that is, when it is placed on more than one word of an utterance, command amplitudes are reduced as the length of the focus domain increases. As a consequence, the influence of focus on words at the tail of an utterance is relatively weak. Focus also influences duration and intensity.

1. Introduction

There are two prevailing TTS technologies today: parametric synthesis and concatenative synthesis. As concatenative synthesis (corpus-based) can produce more natural-sounding speech, it has become the most widely adopted technology, especially in commercial TTS systems. Most of current corpus-based synthesis systems, however, are based on neutral reading-style sentences. Therefore they can only output utterances in this style and lack the capability of varying prominences of words in the utterance. When TTS is used in a broadcast environment, or in dialogs requiring contrastive meanings, we sometimes need to highlight certain words or phrases, so these words or phrases become the focus of the whole sentence. For example, in a weather report the number "20" in the sentence "Today's high will be of 20 degrees centigrade" is essential for the listeners and will therefore be the focus. In the reply to a question the key point of the answer is under focus as well. "Where did you go?", "I went to school." When speakers want to emphasize certain words they will usually focus on these. This kind of utterance is still of declarative type, but they exhibit distinct intonation patterns, since focus is typically marked by increased prominence. If we wish to realize focus in corpus-based synthesis directly, we must record a special inventory which could be very costly. Alternatively we can adjust the prosody of neutral, broadly focused utterances in order to convey the connotation of the desired focus condition.

Since we want to establish the quantitative relationship between focus and the F0 contour we use the Fujisaki Model for parameterizing F0 contours of Mandarin. Based on the results of analysis we then develop a method for synthesizing the effect of varying focus on the F0 contour.

2. Method of Analysis

In the concept underlying the Fujisaki Model, the log F0 contour is composed of a base frequency Fb, a phrase component and an accent component (tone component in tone languages); it can be represented by the following formula:

$$\ln F0(t) = \ln Fb + \sum_{i=1}^{I} Ap_i Gp(t - T_{0i}) + \sum_{j=1}^{J} Aa_j [Ga(t - T_{1j}) - Ga(t - T_{2j})]$$

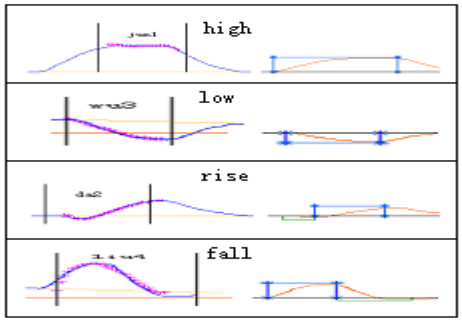

Figure 1: *Examples of tone command configurations for producing F0 contour of syllabic tones. The four contours are corresponding to high, rising, low, and falling tones, respectively.*

In utterances of Mandarin, F0 contours are considered as a combination of a slow decaying phrase contour and the fast changing syllabic contour associated with individual tones. This analysis is similar to that of the Fujisaki model. We can decompose the F0 contours of Mandarin into phrase component and syllable component by automatic or manual method. Since Mandarin is a tone language, we combine tone commands to model the syllabic F0 contour (Figure 1). Distinct tone command combinations are used to produce the F0 contour of Mandarin high, rising, low, falling and light tones, respectively. Tones 2, 3, and 4 or their combination, have the adjacent negative-positive or positive-negative command which is called tone switch, and the point it switches is tone switch time. The switch time corresponds with the pole position of the F0 contour. In the following experiment we mainly use tone commands to establish rules of F0 changes due to varying focus.

Using the FujiParaEditor[2] we applied a method of automatic analysis and manual correction to get the accurate Fujisaki model parameters.

3. Database Design

In this experiment we designed two databases. Database 1 consists of 24 sentences. Each sentence is composed of a carrier sentence combined with a number of words that are varied. The carrier sentence is "Zhang1 jun1 zheng1 jia1 × × fa1 yin1" into which a two-syllable word is inserted. The reason for using high tone syllables is that from the sequence of high tones we can easily observe and extract the declining phrase command curve. "××" covers all tone combinations of Mandarin, like "da1 ma1", "da1 ma2", etc. A broadcast announcer read these sentences at a normal speech rate and each sentence was read five times. In the first four utterances, narrow focus was placed on "zhang jun", "zheng jia", "× ×" or "fa yin", respectively, the fifth has a broad focus on the whole predicate phrase of the sentence. Therefore we yielded five times 24 utterances, a total of 120 recorded sentences.

In Database 1 the recorded sentences were extremely artificial since they were specially constructed. Since focus is strongly linked with the semantics of an utterance we designed another database of more meaningful utterances to conduct a perceptual experiment. Database 2 contains 10 meaningful sentences. Each sentence was read by the same announcer without focus first yielding 10 utterances, and then with narrow focus on some parts of the sentence yielding 27 utterances. Finally we got a total of 37 utterances. Subsequently Database 2 was used to verify the rules yielded by the analysis of Database 1.

Recordings were made at 16kHz/16bit. We used Praat[5] to calculate F0 curves at a step of 10 ms.

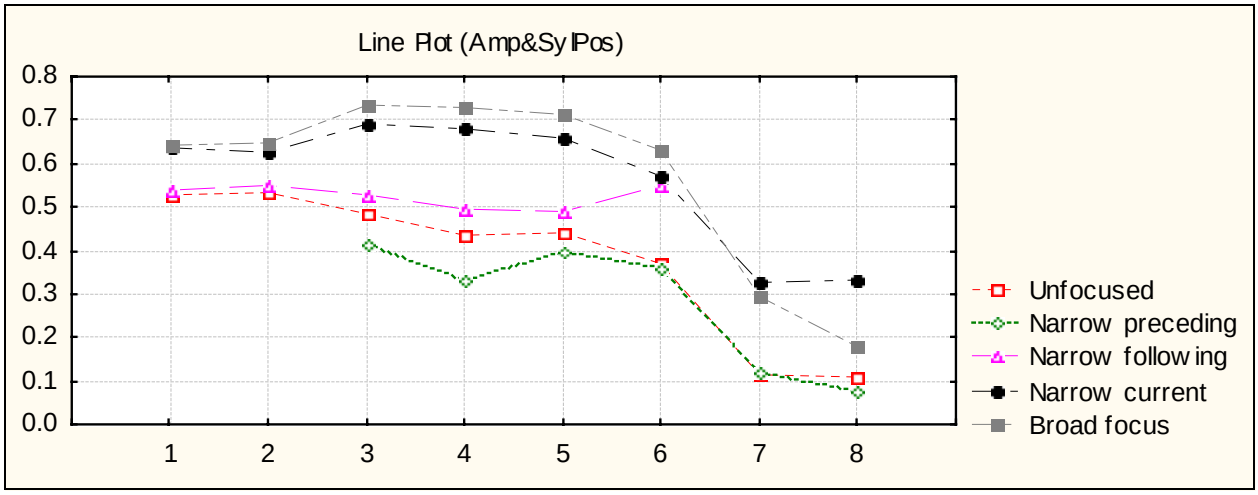

Figure 2: *the horizontal axis presents 8 syllables, vertical axis presents difference amplitude of tome command under different focus conditions.*

4. Results of Analysis

4.1. Narrow focus analysis

In Database 1, all of the sentences had the same carrier sentence, so the phrase component was essentially the same for all utterances. The tone command will be the main parameter and the phrase command will be ignored in the following analysis.

By estimating the tone commands of each of the eight syllables in the utterance, we yielded their respective amplitudes. For the syllables which exhibited two tone commands (mainly rising, falling tones, and low), we defined its amplitude to be the sum of absolute amplitude values of each of the two tone commands, which reflects both height and range. By performing statistic analysis on the tone command amplitudes of all sentences we yielded the results displayed in Figure 2.

From Figure 2 it can be seen clearly that when there is a narrow focus on a word, its command amplitude increases higher than otherwise; if the focus lies on the following word, the amplitude of commands will rise; if the focus lies on the preceding word, the amplitude will decrease. That is to say, the focus on a word in an utterance is clearly reflected by F0. The amplitude of the tone command starts to rise when coming to the emphasized word to make a preparation, then rises to its highest on this word, but falls to an even lower level instead after the word.

Here the broad focus is considered as the same with the narrow focus. Figure 1 shows that broad focus has the highest tone command amplitude in the initial part. In next subsection we'll discuss and analyse why broad focus has higher amplitude than narrow focus.

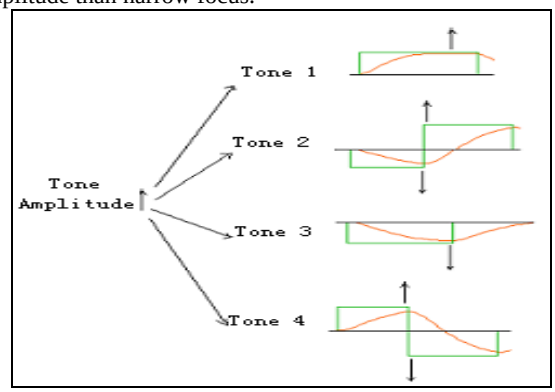

Figure 3: *The tone commands vary with tone amplitude.*

Narrow focus emphasizes a single word and has an influence on tone commands' amplitude in the way that the F0 of the preceding syllable increases and of the following decreases. Amplitudes' increase means that pitch is higher for tone 1. For tone 2 and 4, it means a wider pitch range except being higher. As for tone 3, it causes a lower and full pitch (showed in Figure 3). And vice versa. Because of lack of enough data, the amplification ratio of the positive command to the negative cannot be ruled well.

Table 1 displays mean syllable durations for the four tones under broad and narrow focus conditions. Obviously focus has little influence on tone 1, and its prosodic effect is mainly on F0. However with respect to tones 2, 3 and 4, focus not only increases the amplitude of tone commands causing raises in F0 as well as F0 range expansions, but also increases the syllable duration. However, when a syllable is aligned with a pair of positive and negative tone commands, the relative position change of tone switch time is very small. That is, the position of maximum or minimum of F0 contour remains generally stable relative to the syllable.

Table 1: *mean duration of each tone*

	Tone 1	Tone 2	Tone 3	Tone 4
Focused	227ms	235ms	247ms	239ms
Unfocused	221ms	218ms	218ms	214ms

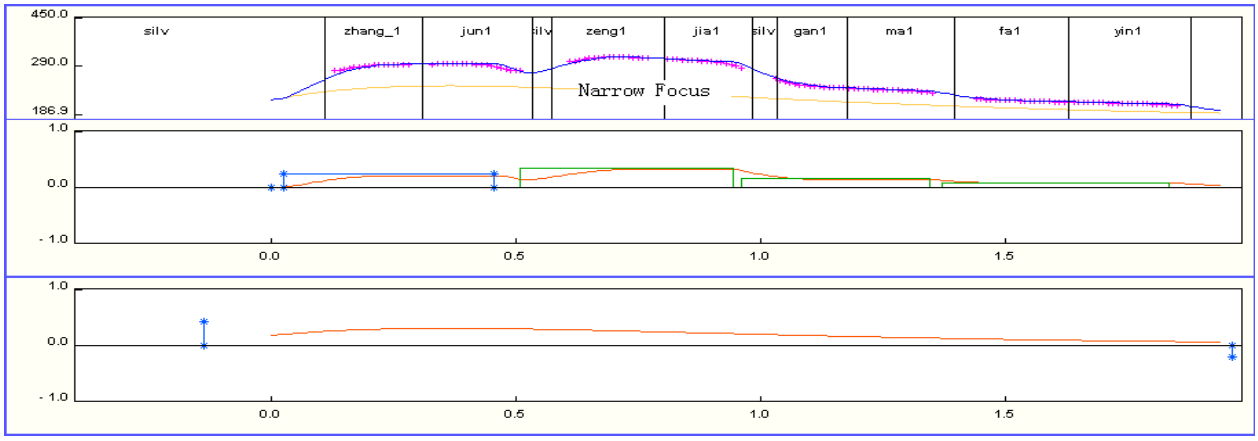

Figure 4: *Narrow focus on " zeng1 jia1"*

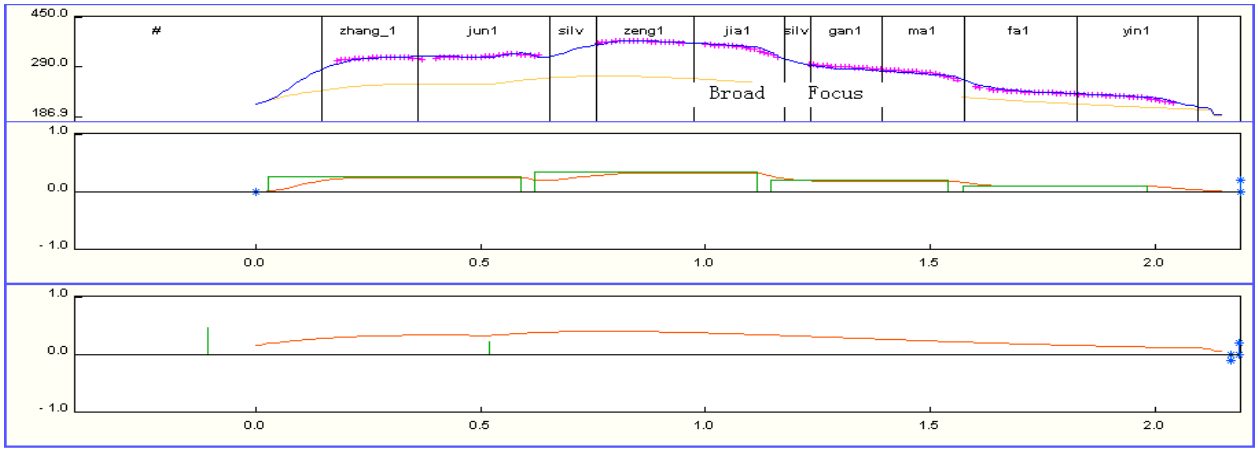

Figure 5: *Broad focus on "zeng1 jia1 gan1 ma1 fa1 yin1"*

4.2. Broad focus analysis

Comparing the broad focus contour with unfocused F0 contours (Figure 2), such a conclusion is draw that F0 of the syllables before the focus rises along with the focus, and in the range of broad focus (the 6 syllables as predicate), the rising step of F0 declines gradually. In the result of Figure 2, the phrase commands are considered as the same, so the command amplitude of broad focus is larger than that of narrow. However, broad focus influences phrase too. In Figure 5, broad focus falling is expressed by adding a phrase command compared with the narrow focus in Figure 4. We can see clearly that broad focus creates a low phrase command in the start of the broad focus. Its amplitude of new tone commands is nearly equivalent to that of the narrow focus. Furthermore, a longer break occurs in the front of the broad focus. It is a remarkable indication. It is considered that there is a "voice reset" in the front of the broad focus.

4.3. Perception experiment

From the above analysis we can only draw tentative conclusions, since the data we have is too limited. In the

following, we designed a perception experiment to verify our findings.

Toward the 10 unfocused sentences in Database 2, respectively we adjust the tone command amplitude and syllable duration on the word that we want to focus according to the results discussed in this section previously. The commands should be extracted automatically from the F0 contour firstly. Increase the amplitude directly for tone-1 syllable. Increase the amplitude of tone commands meanwhile lengthen the duration for tone 2, 3 and 4. As for a broad focus we want, test to add a phrase command and the internal words' commands should be adjusted little. Then we got a new F0 contour rebuilt by modified commands. Synthesize it on the basis of unfocused utterances to get focused utterances. We applied PSOLA for resynthesis of 27 utterances and presented these to three subjects.

They are required to score the naturalness of the synthesized wave, pointed out which word are focused, and compared it with the corresponding focused utterance recorded. On the average 80% of the sentences are perceived as focused and the focus position is clear. This suggests that the above results are to some extent generalizable. It must be stated, however, that when a part of an utterance is emphasized, not only F0 and duration change, but also the energy and spectral characteristics. PSOLA synthesizer cannot process energy upgrading or spectrum smoothing. The synthesized speech is reduced in quality and the speech is easily perceived as synthesized. From the adjusted sentences, the performance of words or phrases in the front of a sentence is better than that in the back. That is because front words are normally louder, F0 and duration enlargement acts the same as speaking with a higher pitch. However, the amplitude of the speech signal decreases towards the end of an utterance, so that after amplification the distortion is severe. To improve on this an algorithm with stronger adjusting ability may be used or picking out the more proper syllable unit while synthesizing.

5. Conclusions

From the above analysis, we can get the rules of tone commands affected by focus, the difference and relations between narrow and broad focus. Nevertheless we have not derived a satisfactory result about the amplitude ratio of positive and negative accent, the ratio of duration and the energy change under the effect of focus when there is tone switch in tone command. The sparsity of our data greatly effects our observation, and we will continue this work. Moreover, shedding light on how the segmental characteristics are influenced by focus, how they influence synthesis performance as well as further improvement of our synthesizer are subjects of future work.

6. References

[1] H. Fujisaki and K. Hirose, 1984. Analysis of voice fundamental frequency contours for declarative sentences of Japanese. *J. Acoustic. Soc.* Jpn , 233-242.

[2] Hansjoerg Mixdorff, Hiroya Fujisaki, GaoPeng Chen, Yu Hu, 2003. Towards the Automatic Extraction of Fujisaki model Parameters for Mandarin. *EuroSpeech 2003*.

[3] Hansjoerg Mixdorff, 1997. Production of Broad and Narrow Focus in German - A Study Applying a Quantitative Model. In *Proceedings of the '97 ESCA Workshop on Intonation*, 239-242.

[4] Tan Lee, Greg Kochanski, Chilin Shih and Yujia Li, 2002. Modeling Tones In Continuous Cantonese Speech. *ICSLP 2002*.

[5] Fujisaki H., Hallé P. and Lei H., 1987. Application of F0 contour command-response model to Chinese tones. *Reports of Autumn Meeting, Acoustical Society of Japan*, 197-198, 1987.

[6] http://www.fon.hum.uva.nl/praat/.