# Quantitative Analysis of $F_0$ Contours of Emotional Speech of Mandarin

*Wentao Gu and Tan Lee*

Department of Electronic Engineering, the Chinese University of Hong Kong, Hong Kong

{wtgu, tanlee}@ee.cuhk.edu.hk

## Abstract

The $F_0$ characteristics of Mandarin speech in four basic emotions (anger, fear, joy, and sadness) as well as in neutral reading are compared quantitatively. Two approaches are employed: analysis of surface features from time-normalized $F_0$ contours, and analysis-by-synthesis of time-intact $F_0$ contours based on the command-response model, which turns out to be also applicable to emotional speech. For surface $F_0$ features, the height and range of $F_0$, the local tonal variation, and the sentential $F_0$ declination are all investigated. In model-based analysis, the parameters of both phrase and tone commands are compared systematically. The study shows that those surface $F_0$ phenomena can be explained better by the model-based approach, which can later be used in $F_0$ generation for emotional speech synthesis.

## 1. Introduction

During the last two decades, there has been an inspiring growth in the works on emotional speech. Especially, the technologies on emotional speech synthesis and automatic recognition of emotional speech have progressed steadily by the aid of data-driven statistical methods, without having attained a really clear picture of the acoustic characteristics of various vocal emotions. However, such basic questions as to how a given emotion is expressed in speech still need to be answered, not only from scientific considerations but also for a further improvement of the related practical technologies.

Although it has been well known that both segmental and suprasegmental (prosodic) features play important roles in conveying vocal emotions [1], the latter is usually regarded to be primary. In the present study, we shall only investigate the characteristics of $F_0$ contours in vocal emotion expression.

In contrast to a great number of analysis works for non-tone languages like English (e.g. [1]), rather few studies on prosodic features of vocal emotions in tone languages like Mandarin Chinese have been reported in literature. The reason may partly lies in that the presence of lexical tones significantly constrains the manipulation of $F_0$ in emotional speech, as discussed in [2] where three $F_0$ features ($F_0$ slope, $F_0$ variation, and $\Delta F_0$) are measured. Therefore, although the acoustic realizations of vocal emotions share many common properties across languages, there may be still some attributes specific to tone languages that need more investigation.

Among those very few studies on Mandarin, Yuan et al. [3] claimed that anger and fear are mainly realized by phonation; joy is mainly realized by $F_0$; and sadness is realized by both. Zhang et al. [4] investigated $F_0$, duration as well as short-time amplitude, not only at the sentential layer but also on the syllable-by-syllable base; their study also showed that stressed words carry more identifiable acoustic features for vocal emotions than unstressed words. Table 1 summarizes the qualitative results on $F_0$ features of Mandarin speech in the four basic emotions obtained in these studies, in company with those for English as obtained in [1]. It should be noted that the features in [2] and [3] are defined differently. As shown, the tendencies on both height and range of $F_0$ differ only slightly between the two languages.

*Table 1*: Summary of $F_0$ features for emotional speech in literature, [1] for English and [2, 3] for Mandarin

| Lit. | $F_0$ feature | Anger | Fear | Joy | Sadness |
|---|---|---|---|---|---|
| Murray et al. [1] | Average | very much higher | very much higher | much higher | slightly lower |
| | Range | much wider | much wider | much wider | slightly narrower |
| | Inflection | abrupt on stressed | normal | smooth upward | downward |
| Yuan et al. [2] | Height | high | high | high | low |
| | Fluctuation of top-line | large | small | large | small |
| Zhang et al. [3] | Average | highest | higher | higher | slightly lower |
| | Range | widest | slightly wider | wider | slightly narrower |
| | Stressed in contrast to unstressed words | higher | higher | higher | wider range |

However, many important phenomena in $F_0$ variation have not been studied yet. For instance, how lexical tone patterns and sentential $F_0$ declination vary with those emotions needs a systematic investigation. In addition, all the aforementioned works inspect surface $F_0$ features such as max/min/mean $F_0$ values, slope of $F_0$ curve, or range of $F_0$ values in a target syllable or in a larger domain. This kind of analysis, however, has the following drawbacks. First, it does not separate global intonation and local tone patterns explicitly, and hence only gives a confounded result. Second, the surface measurements are phenomenological and cannot capture the essential characteristics of $F_0$ movements efficiently. Third, the surface measurements are vulnerable to microprosody and noises in $F_0$ extraction.

For emotional speech synthesis, the construction of prosodic rules is necessary. Hence, a quantitative model giving a parametric representation of $F_0$ contours needs to be introduced. In this sense, the command-response model for the process of $F_0$ contour generation [5] is quite efficient, which was originally proposed for Japanese but later also applied to many other languages including Mandarin [6]. The method has been employed to analyze quantitatively $F_0$ contours of emotional speech of Japanese [7]. In the present study, we will investigate whether the model can be applied successfully to emotional speech of Mandarin, and if so, what parametric differences can be captured between those vocal emotions as a result of a fully quantitative analysis. The results will also be compared with the analysis of surface $F_0$ features observed directly from time-normalized $F_0$ contours.

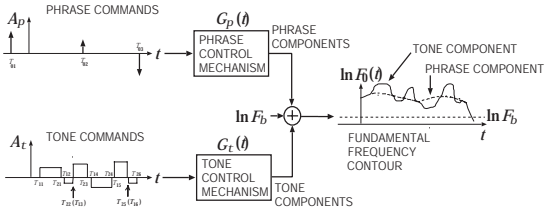



*Figure 1*: The command-response model for the process of $F_0$ contour generation.

## 2. The command-response model for $F_0$ contours of Mandarin

Figure 1 shows the diagram of the command-response model. It describes $F_0$ contours in the logarithmic scale as the sum of phrase components, accent/tone components, and a baseline level $\ln F_b$. The phrase commands (pulses) produce phrase components through the phrase control mechanism, giving the global shape of $F_0$ contours, while the accent/tone commands (pedestals) generate accent/tone components through the accent/tone control mechanism, characterizing the local $F_0$ changes. Both mechanisms are assumed to be critically-damped second-order linear systems. The model can give highly accurate approximations to $F_0$ contours from a small number of linguistically meaningful parameters, and has been applied to many languages [5]. The details of model formulation are described in [5]. In the present study, the constants $\alpha$, $\beta$, and $\gamma$ in the model are fixed at 3.0 (1/s), 20.0 (1/s), and 0.9, respectively, following the previous studies.

Unlike Japanese, tone languages usually require both positive and negative tone commands due to faster local $F_0$ changes. For a specific tone language, a set of tone command patterns needs to be specified in the model. As listed in Table 2, Mandarin has four lexical tones, as well as a neutral tone – any lexical tones can be neutralized in an unstressed syllable. The rightmost column of the table gives the tone command patterns for each tone [6]. The neutral tone does not have a stable tonal shape and hence it has no intrinsic tone command pattern; instead, it varies largely with the preceding tone.

*Table 2*: Mandarin tone system

| Tone type | Pitch feature | Tone code | Command pattern |
|---|---|---|---|
| T1 | high | 55 | positive |
| T2 | rising | 25 | negative to positive |
| T3 | low | 21(4) | negative |
| T4 | falling | 51 | positive to negative |
| T0 | neutral | variable | context-dependent |

## 3. Speech data

We designed ten short sentences (part of them are from [4]), each consisting of 4 to 9 characters. They are declarative sentences or wh-questions, hence inherently with a declining intonation in neutral reading. The sentence texts are neutral (i.e. not literally associated with any specific emotion) but can be placed in different contexts to induce various emotions. Each sentence was uttered in five styles: four basic emotions (anger, fear, joy, sadness) and a neutral reading at normal speech rate (i.e. neutral emotion). Here, anger and joy are active emotions, while fear and sadness are passive emotions. Each utterance was recorded with three repetitions at consistent degrees of emotion expression.

Two speakers, one male and one female, who were both graduate students, were asked to record the speech. Before the recording of each utterance, a designed context was prompted by the instructor to help the speaker induce the required emotion. For instance, for the following sentence, of which the text is not inherently associated with any of the four basic emotions: "Ju1 ran2 hui4 fa1 sheng1 zhe4 zhong3 shi4" (Unexpectedly this thing happened), a prompt story that apparently causes anger, fear, joy, and sadness was described to the speaker respectively before his/her recording.

Although there are always speaker differences in vocal emotion expression, our preliminary study shows that these two speakers largely share the common strategy of expression (though differ in quantity). Hence, in the present work, only the analysis of the female speaker's data will be presented.

## 4. Method of data analysis

$F_0$ values were extracted by a modified autocorrelation analysis, while syllables were segmented manually by visual inspection of waveform and spectrogram.

One difficulty in comparing $F_0$ contours lies in that they are not aligned in time. The $F_0$ contour is not a unit-based measurement like syllable duration; instead, it is a time-varying sequence which implicitly involves the timing information. For a direct comparison, $F_0$ contours need to be time-normalized. Hence, the measured $F_0$ values were first smoothed and interpolated for voiceless intervals to produce a continuous $F_0$ contour. Then, ignoring durational differences, a time-normalized $F_0$ contour was obtained by extracting a 10-point (equally spaced) sequence of $F_0$ values in each syllable from the continuous $F_0$ contour.

Unlike surface feature analysis, model-based analysis tries to give an optimal approximation to the entire $F_0$ contour through a set of parameters. This procedure, named analysis-by-synthesis, was first done manually with the aid of syllable timing and linguistic information such as tone identity and syntactic structure, and later the parameters were optimized by successive approximation, as discussed in more details in [6]. It should be noted that this is not merely a mathematical procedure of curve fitting; instead, the minimum error criterion is only effective under the linguistic constraints to ensure the linguistic meaningfulness of the analysis.

For the utterances of a fixed speaking style, the baseline frequency $F_b$ can be considered to be constant for the sake of simplicity of modeling, and it is usually initialized by visual inspection of $F_0$ contours of many utterances in the same style.

In read speech of neutral emotion, tone commands in each syllable should basically comply with the inherent command patterns for the particular tone type, though closely neighboring tone commands with the same polarity are allowed to be merged. In emotional speech, however, the situation becomes complicated due to frequent reduction, neutralization, or change of lexical tones. Hence, tone identities should be based on acoustic realization instead of linguistic form. Also, the following two heuristic rules can be adopted. First, some tone commands may disappear, but it rarely occurs that the polarity of a tone command is reversed. Second, the stressed syllables tend to preserve the canonical form of tones better and hence a better coincidence with the inherent tone command patterns should be given there.

The occurrences of phrase commands are largely aligned with major syntactic boundaries and can be determined by comparison of syllabic $F_0$ pattern and the canonical form of tones, by comparison of $F_0$ patterns in adjacent tones, and sometimes also with the aid of prosodic perception. Phrase

commands are only assigned when necessary and linguistically meaningful. At many places, whether to add a very small phrase command or not usually has little effect on the accuracy of approximation; in this case, we do not add it.

## 5. Results

### 5.1. Analysis of time-normalized $F_0$ contours

Figure 2 shows the average time-normalized $F_0$ contours of the utterances of five sentences (one sentence for each panel). For each sentence, the $F_0$ contours (averaged over the three repetitions) for five emotions (including neutral) are plotted. From the figure the following characteristics are observed:

(1) The five emotions are distinctly clustered into two groups: one is anger/fear/joy and the other is sadness/neutral, the former showing significantly higher $F_0$ than the latter.

(2) Anger shows the widest $F_0$ range, and usually gives the highest sentential maximum $F_0$. Within the higher group, anger also gives the lower sentential minimum $F_0$ than fear and joy. Especially, compared with other emotions, anger always raises $F_0$ in a certain syllable to produce an $F_0$ peak, as shown in /zhe4/, /jin4/, /zhe4/, /yue4/, and /zhe4/ in the five sentences, respectively (they happen to be all of T4 here). The $F_0$ values immediately before the syllable of peak $F_0$ are usually also raised, while those after the peak syllable are conspicuously lowered, hence resulting in a larger $F_0$ declination than other emotions.

(3) Within the higher group, fear shows the narrowest $F_0$ range. Especially, at the end of an utterance, fear gives higher $F_0$ than anger and joy (in fact, the highest among the five emotions). In other words, fear shows a weaker sentential $F_0$ declination than anger and joy.

(4) Sadness shows both the lowest $F_0$ value and the narrowest $F_0$ range. Especially, $F_0$ contours of sadness and neutral often coincide at the major $F_0$ valleys of the latter, but at other positions $F_0$ contours of sadness are significantly lower and flatter than those of neutral.

(5) Among the five emotions, fear and sadness (both are passive) show similar $F_0$ trajectories (nearly parallel) and differ mainly in the height – fear is higher (and as we will show later, they also differ in tempo – fear is faster). These two give flatter $F_0$ contours than others. This is especially distinct in panels (c) ~ (e).

(6) Among the five emotions, fear and joy compose a pair that seems the most difficult to be distinguished. The major distinction is that fear gives a flatter $F_0$ contour, i.e. with a narrower $F_0$ range. In the earlier part of an utterance fear has comparable or even lower $F_0$, but in the later part (especially at the end) it keeps significantly higher $F_0$ than joy.

Besides the above findings, we further look separately into sentential intonation and syllabic tones, though at this stage these two components cannot be separated effectively.

On the one hand, it is well known that $F_0$ declines gradually in an utterance of neutral reading (except some yes/no questions). This can roughly be observed from Figure 2, especially in those relatively longer sentences. For instance, among the five T4 syllables in the neutral reading of sentence (d), the utterance-initial /zhe4/ is higher than the utterance-medial /bian4/ and /yue4/, which are again higher than /yue4/ and /re4/ near the end of the utterance.

In emotional speech, the magnitudes of sentential $F_0$ declination rank as follows: anger > joy > neutral > fear > sadness, which also indicates the order of active emotions > neutral > passive emotions. This can be seen from Table 3, where the ratios of utterance-final $F_0$ to utterance-initial $F_0$

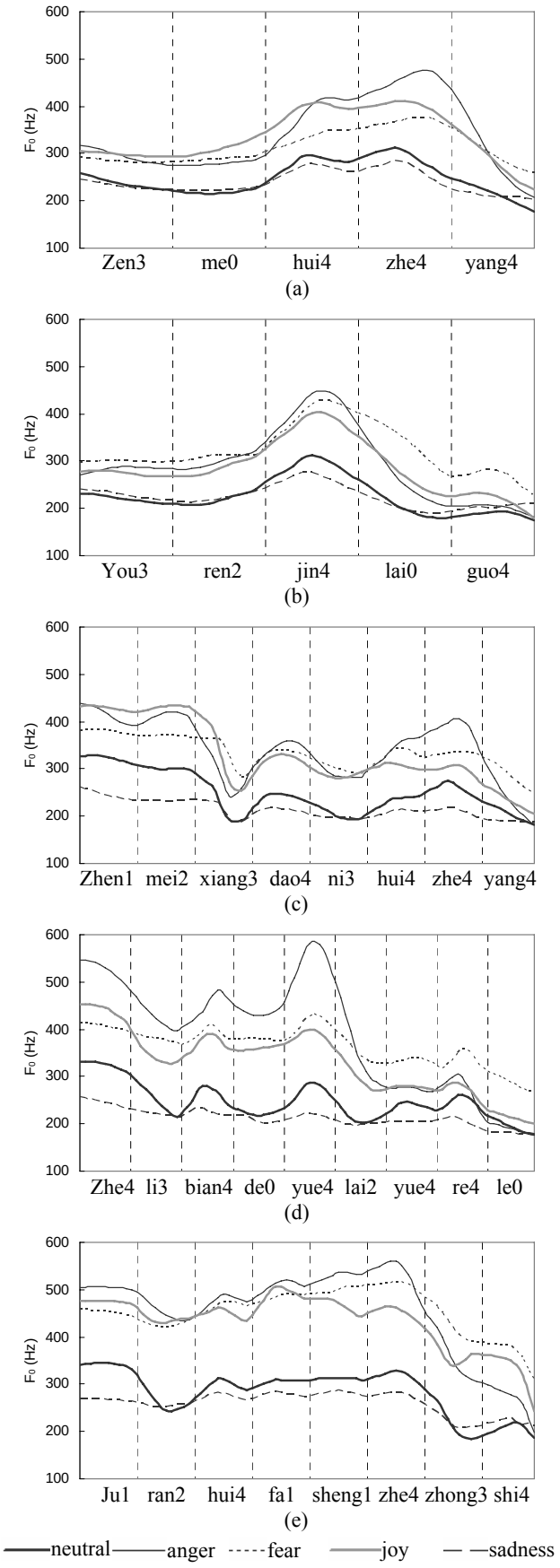



*Figure 2*: Average time-normalized continuous $F_0$ contours of the utterances in five emotions (including neutral).

*Table 3*: The ratios of utterance-final $F_0$ to utterance-initial $F_0$, approximately indicating sentential declination

| | Neutral | Anger | Fear | Joy | Sadness |
|---|---|---|---|---|---|
| (a) | 0.69 | 0.65 | 0.89 | 0.73 | 0.83 |
| (b) | 0.76 | 0.67 | 0.76 | 0.65 | 0.87 |
| (c) | 0.56 | 0.42 | 0.65 | 0.47 | 0.71 |
| (d) | 0.53 | 0.33 | 0.64 | 0.44 | 0.69 |
| (e) | 0.55 | 0.39 | 0.67 | 0.51 | 0.78 |
| Average | 0.61 | 0.47 | 0.72 | 0.55 | 0.78 |

*Table 4*: The percentages of correctly identified lexical tones when lifted out of continuous speech (%)

| | Neutral | Anger | Fear | Joy | Sadness | Average |
|---|---|---|---|---|---|---|
| T1 | 100 | 100 | 100 | 100 | 75 | 95 |
| T2 | 33 | 33 | 8 | 42 | 33 | 30 |
| T3 | 100 | 67 | 44 | 94 | 50 | 81 |
| T4 | 71 | 59 | 37 | 57 | 35 | 52 |
| All | 75 | 62 | 43 | 68 | 43 | 58 |

are given. Although the difference between the two ends is not entirely ascribed to sentential declination, a comparison of such differences can roughly show the effects of different emotions on sentential declination. A more accurate analysis will be given in the model-based approach in the next section.

On the other hand, it is also well known that tones in continuous speech deviate from their canonical form in isolated syllables, and tone identities of some syllables may be completely lost, especially in spontaneous speech. For example, Tseng's study [8] showed that in fluent spontaneous speech of Mandarin only 36% of syllables preserve their lexical tones. The reason lies in that speech production is a compromise between maximizing communicative function and minimizing articulatory effort, as suggested by the hypo- and hyper-articulation theory [9].

In order to investigate the acoustic realization of lexical tones in emotional speech, we conducted a perceptual test, in which each syllable (except those that can be predicted from the text to be in neutral tone) was lifted out of the utterance and played back to a native subject for a perceptual identification of tone ('unidentifiable' is also set as an option).

Table 4 lists the percentages of correctly identified lexical tones. Compared with neutral reading, the proportions of identifiable tones decrease significantly in emotional speech, especially in passive emotions (fear and sadness), indicating that the use of hypo-speech increases in expressing emotions (especially for passive ones). This is consistent with the observation that $F_0$ range is narrowed in fear and sadness. Besides, the contour tones (T2 and T4) are found to be less preserved than the level tones (T1 and T3). It coincides with our model-based analysis, in which a pair of tone commands inherently associated with a contour tone is frequently reduced to a single tone command.

## 5.2. Model-based analysis of $F_0$ contours

Table 5 gives the average model parameters as a result of model-based analysis-by-synthesis of $F_0$ contours of all the utterances. Since $F_0$ contours of the five emotions are clearly clustered into two groups according to the overall $F_0$ level, two different $F_b$ values are set heuristically for the two groups, respectively. Because utterance-medial phrase commands may occur at different positions and some utterances even do not have medial phrase commands, only the magnitude ($A_p$) of utterance-initial phrase command is listed for comparison.

*Table 5*: Average model parameters for $F_0$ contours of emotional speech

| Parameters | Neutral | Anger | Fear | Joy | Sadness |
|---|---|---|---|---|---|
| $F_b$ [Hz] | 160 | 220 | 220 | 220 | 160 |
| Num of phrase cmd. | 2.33 | 1.08 | 1.75 | 1.42 | 2.25 |
| Utterance-initial $A_p$ | 0.49 | 0.52 | 0.49 | 0.49 | 0.43 |
| Num of tone cmd. | 6.58 | 7.42 | 6.83 | 7.17 | 5.50 |
| Abs amp. tone cmd. | 0.35 | 0.54 | 0.37 | 0.39 | 0.23 |
| Dur of tone cmd. [s] | 0.10 | 0.08 | 0.06 | 0.07 | 0.10 |

For tone commands, both absolute amplitude and duration (from onset to offset) are given.

From the statistics the following tendencies are observed:

(1) Baseline $F_b$: (anger, fear, joy) > (neutral, sadness).

(2) Number of phrase commands: (neutral, sadness) > fear > joy > anger.

(3) Magnitudes of phrase commands: sadness < others.

(4) Number of tone commands: anger > joy > fear > neutral > sadness. Especially, sadness gives an obviously smaller number than others.

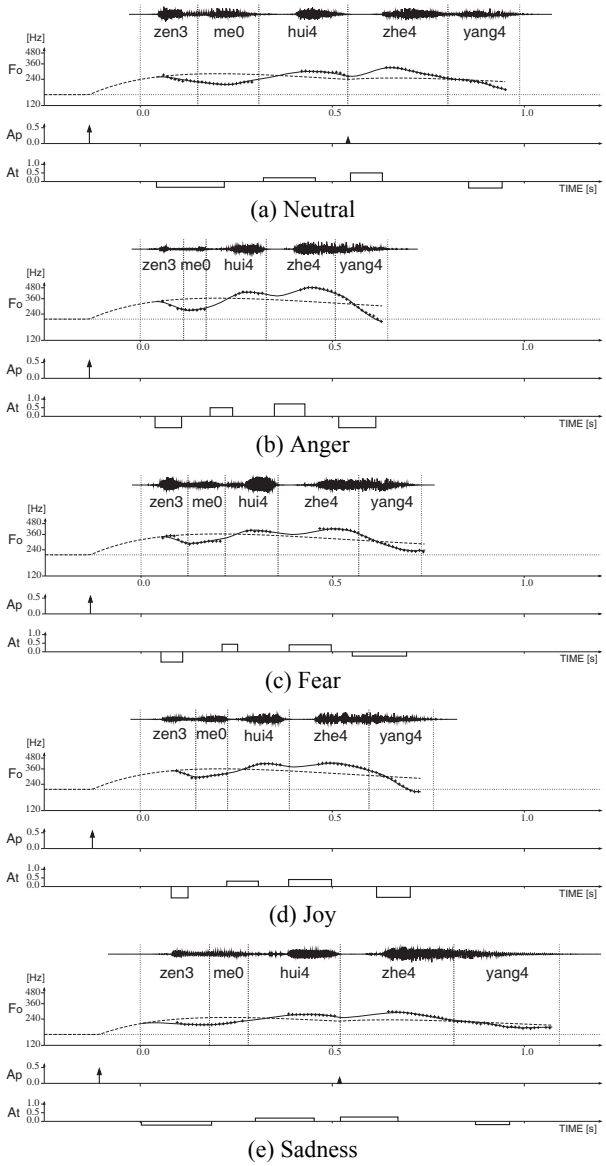
(5) Absolute amplitudes of tone commands: anger > joy > fear > neutral > sadness. This rank is the same as that for the number of tone commands, but the differences between joy, fear, and neutral are not as significant as between others. Both (4) and (5) suggest how big local $F_0$ variation is: fewer and smaller tone commands lead to smaller local $F_0$ variation.

(6) Duration of tone commands: (sadness, neutral) > (anger, joy, fear). Although duration of tone commands is not inherently correlated with syllable duration (as we will show in a later example), in many cases they are approximately in proportion. The result indicates that the speech in neutral or sadness is slower than in other emotions, which is basically consistent with the report on syllable/sentence duration in [4].

The different magnitudes of sentential $F_0$ declination as discussed in Section 5.1 can be explained from the above analyses. Among the emotions giving comparable speaking rates, a larger number of phrase commands indicates more $F_0$ resets and thus results in a weaker sentential $F_0$ declination – hence the declination ranks as anger > joy > fear. Sadness gives a weaker declination than neutral, mainly due to the smaller phrase command. However, a comparison between the emotions giving different speaking rates is a little complex. Although $F_0$ of slower speech may decline more due to the longer stretch of phrase components, the difference in the number of phrase commands should also be considered because slower speech usually needs more phrase commands for $F_0$ resetting – in this case the result is a combined effect of these two factors.

In comparison, this model-based analysis captures sentential $F_0$ declination more accurately than the surface feature analysis in Section 5.1, because phrase and tone components are separated explicitly in the framework of the model and hence the differences in local tonal variation due to different amplitudes of tone commands can be excluded.

The above simple statistics, however, are still not sufficient to give a full view of the parametric differences between vocal emotions. The detailed distributions of model parameters need also to be investigated. For simplicity of illustration and comparison, we select the short sentence shown in Figure 2(a) as an example, which happens to give the same number of tone commands in all the five emotions. Figure 3 shows the results of analysis-by-synthesis of $F_0$ contours of five utterances in the respective emotions. The crossed symbols indicate the measured $F_0$ values, while the solid, dotted, and dashed lines indicate the approximated $F_0$




(a) Neutral

(b) Anger

(c) Fear

(d) Joy

(e) Sadness

*Figure 3*: Analysis-by-synthesis of $F_0$ contours of the utterances in five different emotions.

contours, the baseline frequencies, and the contributions of phrase components, respectively. The differences between the approximated $F_0$ contours and the phrase components correspond to the tone components.

As shown, $F_0$ contours of speech in all these emotions can be approximated at a very high accuracy with a small number of commands in the framework of the model. Among the five syllables, /zen3 me0/ are merged into a single negative tone command (because /me0/ is in neutral tone); /hui4/ almost loses the falling feature and is reduced to a high level tone – hence a positive tone command; and /zhe4 yang4/, as a prosodic word composed of two consecutive T4 syllables, shows the well-known tone coarticulation (viz., both T4's are reduced to half-falling), and hence can be modeled like a single T4 syllable – a positive tone command followed by a negative one. Besides, anger, fear, and joy only give an utterance-initial phrase command, while neutral and sadness also give an additional phrase command (with a lower magnitude than the utterance-initial one) before the utterance-final prosodic word /zhe4 yang4/, because neutral or sad speech is slower than others and hence needs more $F_0$ resets.

*Table 6*: Model parameters for $F_0$ contours of the example sentence in five emotions

| | | Neutral | Anger | Fear | Joy | Sadness |
|---|---|---|---|---|---|---|
| $F_b$ [Hz] | | 160 | 210 | 210 | 210 | 160 |
| 1st $A_p$ | | 0.48 | 0.50 | 0.44 | 0.49 | 0.44 |
| 2nd $A_p$ | | 0.15 | -- | -- | -- | 0.14 |
| Amp. of tone cmd. | zen3-me0 | -0.32 | -0.70 | -0.55 | -0.53 | -0.25 |
| | hui4 | 0.26 | 0.50 | 0.44 | 0.42 | 0.22 |
| | zhe4 | 0.45 | 0.67 | 0.45 | 0.41 | 0.27 |
| | yang4 | -0.35 | -0.59 | -0.34 | -0.58 | -0.14 |
| Dur. of tone cmd. [s] | zen3-me0 | 0.16 | 0.06 | 0.05 | 0.06 | 0.15 |
| | hui4 | 0.11 | 0.06 | 0.04 | 0.07 | 0.10 |
| | zhe4 | 0.08 | 0.10 | 0.11 | 0.12 | 0.10 |
| | yang4 | 0.09 | 0.11 | 0.10 | 0.10 | 0.14 |

Table 6 lists the relevant model parameters averaged over three repetitions of utterances for this example sentence in all the five emotions. Although the overall tendencies are consistent with the statistics given in Table 5, the differences in tone command parameters are not homogenous in the utterance; instead, they vary with word position, and some local features may be more important in characterizing the particular emotions, for instance:

(1) In the earlier syllables /zen3 me0 hui4/, fear and joy give comparable amplitudes of tone commands, as is consistent with the average result shown in Table 5, but in the later part, especially in the utterance-final syllable /yang4/, fear gives significantly higher tone commands than joy. This explains the surface observation that fear ends with a higher $F_0$ and shows a weaker sentential $F_0$ declination than joy.

(2) In the earlier syllables /zen3 me0 hui4/, anger, fear, and joy give much shorter tone commands than neutral speech, as is consistent with the average result shown in Table 5, but in the final two syllables /zhe4 yang4/ the case turns opposite, viz., they give even longer tone commands than neutral, though syllable duration is still shorter – in fact the longer tone commands here are not due to syllable duration but due to the larger local $F_0$ variation than in neutral speech. For sadness, on the other hand, the durations of tone commands in the earlier syllables /zen3 me0 hui4/ are comparable with those in neutral speech, while the tone commands in the final two syllables /zhe4 yang4/ are longer, which is consistent with the longer syllable durations than in neutral speech. It should be noted that the final two syllables /zhe4 yang4/, as a prosodic word, is perceptually the most prominent part in the emotional utterances, as is consistent with Li's finding that in emotional speech sentence stress tends to be placed on the utterance-final prosodic word [10].

Hence, the model-based parametric analysis shows that the prosodic characteristics of emotional speech vary systematically with word position, or with the status of sentence stress. The above results of analysis for this short sentence are similarly observed in the utterances of other sentences. The major difference lies in that for those longer sentences the numbers of tone commands also vary with vocal emotions, as indicated by the statistics shown in Table 5.

## 6. Discussion and conclusion

The $F_0$ characteristics of emotional speech of Mandarin are investigated, both by surface feature analysis and by model-based analysis. Although the present study is still preliminary and does not involve a large amount of data, many valuable results have been obtained.

The quantitative differences in the time-normalized $F_0$ contours for different vocal emotions can be summarized qualitatively as follows. The five emotions are first clustered into two groups, i.e., anger/fear/joy vs. neutral/sadness, the former giving globally higher $F_0$ contours than the latter. Among the higher group, anger shows the largest sentential declination and the widest $F_0$ range, and usually an $F_0$ peak is raised to give a sentence stress, which is often placed on the utterance-final prosodic word; in the remaining, joy shows larger sentential declination and wider $F_0$ range than fear. Among the lower group, sadness gives smaller sentential declination and narrower $F_0$ range than neutral speech. Besides, in emotional speech the lexical tones are less preserved than in neutral speech, and the preservation of lexical tones is even less in passive emotions than in active emotions. Lastly, although syllable duration is not discussed in depth here, we found a rather similar result as that reported in [4], namely, neutral and sadness are slower than others.

The command-response model is applied successfully to $F_0$ contours of emotional speech of Mandarin, though tone command patterns show larger variation due to the increase in the use of hypo-speech in vocal emotion expression. The observations on surface $F_0$ features can be explained better by the model-based analysis. The global $F_0$ level is represented by the baseline frequency; the sentential $F_0$ declination is characterized by the number as well as the magnitudes of phrase commands; and the local $F_0$ variation is described by the amplitude and duration of tone commands, which also vary systematically with word position in the utterance.

In comparison, analysis of surface $F_0$ features is straightforward, but durational information has to be discarded in comparing time-normalized $F_0$ contours. More importantly, such surface analysis cannot be used directly in emotional speech synthesis. Model-based analysis-by-synthesis of $F_0$ contours, on the contrary, is more efficient in capturing the essential characteristics of $F_0$ movements if a good model is introduced, though we believe that the two approaches can be employed together to give a better validation. With a set of quantitative parameters characterizing different vocal emotions, the model-based approach can be used for $F_0$ generation in emotional speech synthesis [11].

## 7. References


[1] Murray, I.R. and Arnott, J.L, "Toward the stimulation of emotion in synthetic speech: A review of the literature on human vocal emotion", *JASA*, 93: 1097-1108, 1993.

[2] Ross, E.D., Edmondson, J.A., and Seibert, G.B., "The effect of affect on various acoustic measures of prosody in tone and non-tone language: A comparison based on computer analysis of voice," *Journal of Phonetics*, 14: 283-302, 1986.

[3] Yuan J., Shen, L., and Chen, F., "The acoustic realization of anger, fear, joy and sadness in Chinese," *Proc. ICSLP*, pp.2025-2028, Denver, USA, 2002.

[4] Zhang, S., Ching, P.C., and Kong, F., "Acoustic analysis of emotional speech in Mandarin Chinese," *Proc. ISCSLP*, pp.57-66, Singapore, 2006.

[5] Fujisaki, H. "Information, prosody, and modeling – with emphasis on tonal features of speech," *Proc. Speech Prosody*, pp.1-10, Nara, Japan, 2004.

[6] Fujisaki, H., Wang, C., Ohno, S., and Gu, W., "Analysis and synthesis of fundamental frequency contours of Standard Chinese using the command-response model," *Speech Communication*, 47: 59-70, 2005.

[7] Hirose, K., Kawanami, H., and Ihara, N., "Analysis of intonation in emotional speech," *Proc. ESCA Workshop on Intonation*, pp.185-188, 1997.

[8] Tseng, C., *An Acoustic Phonetic Study on Tones in Mandarin Chinese (2nd ed.)*. Institute of Linguistics, Academia Sinica, Taiwan, 2006.

[9] Lindblom, B., "Explaining phonetic variation: a sketch of the H&H theory," In *Speech Production and Speech Modeling*, pp. 403-439, Kluwer Academic Publishers, 1990.

[10] Li, A., 情感句重音模式, *Proc. 7th Phonetic Conference of China*, Beijing, 2006.

[11] Hirose, K., Sato, K., Asano, Y., and Minematsu, N., "Synthesis of $F_0$ contours using generation process model parameters predicted from unlabeled corpora: application to emotional speech synthesis," *Speech Communication*, 46: 385-404, 2005.