The Effect of Paralinguistic Emphasis on F_0 Contours of Cantonese Speech

Wentao Gu, Keikichi Hirose and Hiroya Fujisaki

University of Tokyo, Japan
{wtgu; hirose}@gavo.t.u-tokyo.ac.jp fujisaki@alum.mit.edu

Abstract

Emphasis has significant effect on F_0 contours in various languages, among which tone languages require more careful study because their F_0 contours show complex interaction between lexical tones and phrase intonation. Here we employ the command-response model to investigate the effect of paralinguistic emphasis in Cantonese, a typical tone language with nine lexical tones. Following our previous study on target syllables in a fixed carrier frame, the current study continues to investigate the utterances with natural context, in which the effects of emphasis with different scopes and on different parts of utterance are compared. It is shown that the major effect of emphasis is not on tone commands but on phrase commands. The narrowness/broadness of emphasis can be distinguished by the number of phrase commands being affected in the phonetic realization. By use of the command-response model, F_0 contours for expressive speech conveying the information of emphasis can be generated efficiently.

1. Introduction

An accurate and quantitative representation of the essential characteristics of F_0 contours of speech is necessary for synthesis of expressive speech. The F_0 contour of speech is affected by a complex combination of various linguistic, para- and non-linguistic factors [1]. Among them, emphasis, as a quite fundamental factor in information transmission, is well known to have significant effect on F_0 contours of speech in various languages, though the effect may vary with languages.

Emphasis in speech can be classified into different categories. In the current study, we only investigate the effect of *informative emphasis*, which is not directly correlated with morphology or syntax of the text but is intentionally imposed by the speaker on a particular part of utterance to highlight the intended new information. In other words, it is *paralinguistic emphasis*. The scope of emphasis can be as narrow as a single syllable, but can also be broader to encompass polysyllabic words, compound words or even phrases.

Many works have been done on emphasis in various languages, among which tone languages require more careful study because of the complex interaction between phrase intonation and lexical tones which have an important role to distinguish the meaning of words. In this study, we investigate specifically the effect of emphasis in Cantonese, a typical tone language well-known for its complex tone system.

Without introducing a quantitative model, most previous works (e.g. [2] for Cantonese) inspect direct F_0 measurements such as onset/offset/peak/valley/mean F_0 values, slope of F_0 curve, or range of F_0 values in a target syllable. However, such kind of analysis has two drawbacks. First, it does not separate global intonation and local tone patterns explicitly, and hence the analysis can only give a confounded result. Second, the phenomenological measurements cannot capture the essential characteristics of F_0 movement efficiently.

By introducing the command-response model [1] to Cantonese [3, 4], we use a fully quantitative approach to study the effect of emphasis on F_0 contours of Cantonese. Our recent study [5] on target syllables embedded in a fixed carrier frame has shown that the major effect of emphasis is on phrase commands, whereas both the polarity and the amplitude of tone commands in the target syllables are hardly affected so that the inherent tone patterns are maintained. The current study will continue to investigate the effect of emphasis in utterances of natural context, and the emphasis with different scopes and on different parts of utterance will be compared.

2. The command-response model for F_0 contour generation for Cantonese

The command-response model for tone languages describes F_0 contours in the logarithmic scale as the sum of phrase components, tone components and a baseline level $\ln F_b$. The phrase commands produce phrase components through the phrase control mechanism, giving the global shape of F_0 contour, while the tone commands of both positive and negative polarities generate tone components through the tone control mechanism, characterizing the local F_0 changes. Both mechanisms are assumed to be critically-damped second-order linear systems. For a specific tone language, a set of tone command patterns needs to be specified in the model.

As one of the major dialects of Chinese, Cantonese has a complex system of nine lexical tones, as described in Table 1. The syllables of entering tones end with an unreleased stop coda /p/, /t/ or /k/, and are comparatively shorter in duration than those of non-entering tones. Each entering tone has its counterpart of non-entering tone, showing a similar F_0 pattern – T7, T8 and T9 correspond to T1, T3 and T6 respectively.

We have already proposed a set of tone command patterns for the nine lexical tones in Cantonese [3, 4]. In terms of the command polarities corresponding respectively to the earlier and the later parts of a syllable, the command pattern for each tone can be represented phonologically as in the rightmost column of Table 1, where +, - and 0 denote positive, negative and null commands respectively, and the brackets indicate entering tones.

Among the nine tones, T3 and T8 have no tone commands,

Table 1: *Descriptions of Cantonese tone system.*

Tone name in Middle Chinese system		Tone number	Pitch feature	Tone code	Command pattern
Non-entering tones	Upper-level	T1	high level	55	+ +
	Upper-elevating	T2	high rising	35	- +
	Upper-departing	T3	mid level	33	0 0
	Lower-level	T4	low falling	21	- -
	Lower-elevating	T5	low rising	13	- 0
	Lower-departing	T6	low level	22	- -
Entering tones	Upper-entering	T7	high level	5	[+ +]
	Middle-entering	T8	mid level	3	[0 0]
	Lower-entering	T9	low level	2	[- -]

and T2 has a pair of tone commands, while all the other tones actually possess a single tone command. It is to be noted that T4 and T6 show the same command polarities but T4 gives more negative amplitude than T6 [3, 4]. The command patterns of entering tones are similar to those of their respective counterparts of non-entering tones, except the shorter command duration caused by the unreleased stop coda.

3. Speech data

Unlike the speech material in our previous study [5] where the target syllables embedded in a fixed carrier frame are *mentioned*, the sentences in the current speech material consist of words that are *used* in natural context. The speech material consists of eight sets of declarative sentences, each composed of 6~10 syllables. There are no constraints on tone distribution, word constitution or syntactic structure for these sentences.

The sentences within each set share the same text but have emphasis on different parts of the text, either on a syllable/word or on a compound word. Each sentence with a particular target for emphasis was uttered in four versions, with four different degrees of emphasis on the target part: no emphasis (neutral statement), slight emphasis, moderate emphasis and strong emphasis, respectively. The utterance of neutral statement was not prompted by any question, while for the other three versions of utterances, questions were provided together with an indication of the expected degree of emphasis to prompt the emphasis on the expected targets.

The informant is a male native speaker of Cantonese from Guangzhou. Each utterance was recorded twice at the normal speech rate of the speaker. There are altogether 240 utterances.

The F_0 contours were analyzed by the method of Analysis-by-Synthesis with the aid of manual initialization. On the basis of the information of tone identity and syntactic structure, an initial analysis was conducted manually to deconvolve an F_0 contour into underlying commands and a baseline frequency to give a solution in line with the linguistic constraints. For a given speaker, the baseline frequency F_b can be initialized at a constant. Tone commands in each syllable should mostly comply with the inherent command patterns for the particular tone type, whereas the occurrences of phrase commands are largely aligned with major syntactic boundaries and can be determined with the aid of both prosodic perception and comparison of local F_0 values between neighboring tones.

After manual initialization, successive approximation was conducted through a hill-climbing search in the space of model parameters to obtain an optimal solution giving the least mean square error between the observed and the approximated F_0 contours in the logarithmic domain.

4. Analysis results

Since the observed phenomena are quite consistent across the eight sets of sentences, in this section we only take the following sentence as an example for detailed discussion:

(a) “Gaau3 gung1 sei3 dim2 bun3 tai2 din6 jing2.” (The teaching staff will see a movie at half past four.)

This sentence consists of two syntactic constituents: the subject “gaau3 gung1” (the teaching staff) and the predicate composed of an adverbial phrase (ADP) “sei3 dim2 bun3” (half past four) and a verb phrase (VP) “tai2 din6 jing2” (see a movie). The syntactic structure is illustrated in Fig. 1.

Five types of sentences are designed for this sentence text. The first is a neutral statement without any emphasis, whereas the other four, as listed below, place emphases on the different

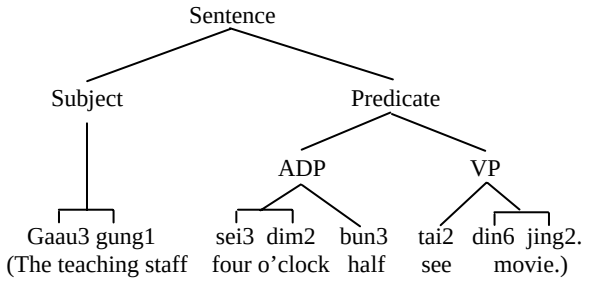


Figure 1: *The syntactic structure of the example sentence*.

parts of the sentence as indicated by the underlines.

(b) “Gaau3 gung1 sei3 dim2 bun3 tai2 din6 jing2.” (The teaching staff will see a movie at half past four.)

(c) “Gaau3 gung1 sei3 dim2 bun3 tai2 din6 jing2.” (The teaching staff will see a movie at half past four.)

(d) “Gaau3 gung1 sei3 dim2 bun3 tai2 din6 jing2.” (The teaching staff will see a movie at half past four.)

(e) “Gaau3 gung1 sei3 dim2 bun3 tai2 din6 jing2.” (The teaching staff will see a movie at half past four.)

Among them, (b) and (c) place emphasis on a syllable or a word, whereas (d) and (e) place emphasis on a compound word. Especially, the target syllable “sei3” (four) in (c) is a part of the target compound word “sei3 dim2 bun3” (half past four) in (d).

These four types of sentences can be regarded as the replies to the following questions respectively:

(Q.b) Who will see a movie at half past four?

(Q.c) Will the teaching staff see a movie at half past five?

(Q.d) When will the teaching staff see a movie?

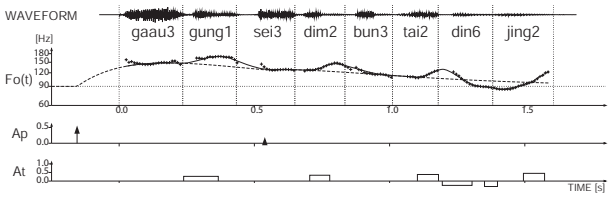
(Q.e) What will the teaching staff do at half past four?

Each of these questions was provided to prompt the speaker during the recording session. Then each of the four types of sentences was uttered in three versions, with slight, moderate and strong emphases on the underlined target part respectively. For the sake of comparison, utterance (a) of neutral statement was uttered together with the three versions, and hence four versions altogether. Each version was recorded twice. Therefore there are altogether $4 \times 4 \times 2 = 32$ utterances for this example sentence.

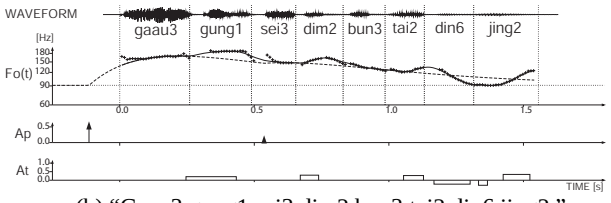
Figure 2 shows the results of Analysis-by-Synthesis of the F_0 contours of five example utterances, the first corresponding to neutral statement (a), whereas the other four corresponding to (b) ~ (e) with moderate emphasis on the respective targets.

Among all these utterances, the tone commands for each syllable mostly coincide with their inherent patterns regardless of the state of emphasis, except that the T2 syllables “dim2” and “tai2” show certain variations. Namely, in utterance (e) a pair of tone commands is clearly shown in the syllable “tai2” which is the initial syllable of the emphasized compound word, but in other occurrences of these two T2 syllables, the first negative commands almost disappear. It indicates that the inherent tone command pattern of T2 may occasionally be reduced in continuous speech by an attenuation of the first negative tone command but it tends to be well maintained in an emphasized context (*e.g.*, in the first syllable of an emphasized target). This is the only effect of emphasis on tone command patterns we have observed in the speech data.

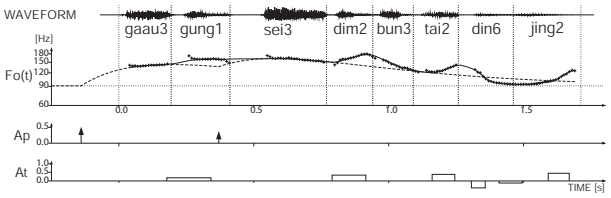
The major effect of emphasis, however, is on phrase commands. This is evident from the example sentence because the three T3 syllables “gaau3”, “sei3” and “bun3” even do not possess tone commands. In utterance (a) of neutral statement,



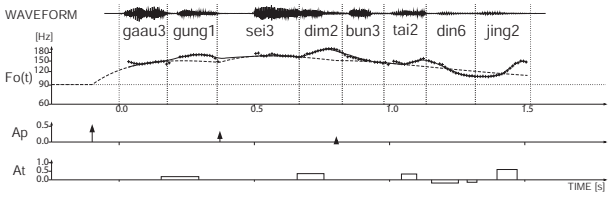
(a) "Gaau3 gung1 sei3 dim2 bun3 tai2 din6 jing2."



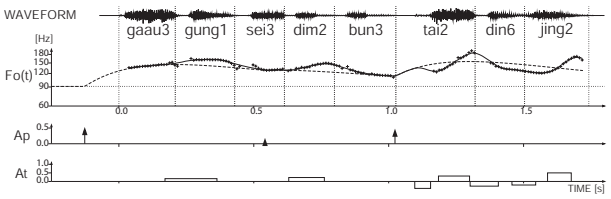
(b) "Gaau3 gung1 sei3 dim2 bun3 tai2 din6 jing2."



(c) "Gaau3 gung1 sei3 dim2 bun3 tai2 din6 jing2."



(d) "Gaau3 gung1 sei3 dim2 bun3 tai2 din6 jing2."



(e) "Gaau3 gung1 sei3 dim2 bun3 tai2 din6 jing2."

Figure 2: *Analysis-by-Synthesis of the F_0 contours of a set of utterances. The underlined parts of sentence are emphasized.*

the occurrences of phrase commands coincide with the syntactic structure: two phrase commands, one before the subject and the second before the predicate (the second is much smaller, and can be absent in some samples).

In utterances (b) and (c), the phrase command immediately before the respective target for emphasis shows larger magnitude than the corresponding one in utterance (a). In utterance (e), on the other hand, an additional phrase command is introduced immediately before the emphasis target "tai2 din6 jing2".

Especially, in utterance (d), not only the magnitude of the second phrase command is increased, but also an additional (third) phrase command is introduced immediately before the syllable "bun3". This is because the target compound word here, "sei3 dim2 bun3", is a combination of two element words "sei3 dim2" (four o'clock) and "bun3" (half), which need to be conveyed with equal importance. In order to put emphasis on the entire compound word instead of only on the first element word as in utterance (c), an additional phrase command is employed.

The results of analysis coincide very well with the intuitive observation. In utterances (b) ~ (e), the increases of F_0 values by emphasis compared with utterance (a) of neutral statement are not localized in the respective target part for emphasis but extended over the entire interval from the target part to the end of the utterance. It is observed that the F_0 values in the sentence-final syllable "jing2" in all the utterances of (b) ~ (e) are higher than those in utterance (a). Also, the nearer the target part for emphasis is to the end of the utterance, the higher the F_0 values in the final syllable "jing2" tend to be. All these observations indicate that the effect of emphasis on F_0 contours has a wide range and coincides exactly with the effect of phrase commands.

It is to be noted that the number of phrase commands being affected in utterance (d) is more than that in utterance (e), though these two utterances both place emphasis on the compound words of the same length (three syllables). This can be explained by the fact that in response to (Q.e) the compound word "tai2 din6 jing2" (see a movie) in utterance (e) is regarded as a single semantic unit that need not be further divided to receive emphasis separately. The comparison between (d) and (e) indicates that the number of phrase commands affected by emphasis is not determined by the length of the linguistic target for emphasis; it also depends on the related syntax, semantics and pragmatics.

In terms of the number of phrase commands affected by emphasis, we can introduce a novel way to distinguish *narrow emphasis* and *broad emphasis*, at least for Cantonese in which the major effect of emphasis is on phrase commands. Namely, emphasis that brings significant change only to a single phrase command is regarded to be narrow, whereas that brings significant changes to more than one phrase command is broad. Hence, emphasis on a syllable or a word is always narrow, whereas that on a compound word can be either narrow or broad in the phonetic realization, depending on the syntactic, semantic and pragmatic natures of the compound word. In our definition, narrowness or broadness of emphasis is not directly relevant to the length of the linguistic units on which an emphasis is intended, but is entirely based on the phonetic realization.

Table 2 gives the mean magnitudes of phrase commands at each possible position for all the utterances of this example sentence. The first and the second columns indicate the type of utterance and the degree of emphasis respectively (0: no emphasis; 1: slight emphasis; 2: moderate emphasis; 3: strong emphasis). It should be noted that the utterances with no emphasis in (b) ~ (e) are actually utterance (a) and hence they should share the same characteristics. The differences between them are due to natural variations in speech.

The magnitudes of phrase commands are listed according to their positions in the utterances, *viz*., the phrase command given in each column occurs immediately before the part of utterance indicated by the column. The cells in gray indicate the target parts for emphasis in each utterance.

It is shown that the stronger the emphasis is, the larger the phrase command tends to be. Also, the increment in magnitude of phrase command is approximately in reverse proportion to the original magnitude in the utterance of neutral statement. For instance, in utterance (a), the utterance-initial phrase command has the largest magnitude in the utterance (this is the usual case for most utterances), and the phrase command immediately before the predicate is much smaller, whereas no

Table 2: *Mean magnitudes of phrase commands for the example sentence.*

Utt	Emp	Magnitude of phrase command			
		gaau3 gung1	sei3 din2	bun3	tai2 din6 jing2
(b)	0	0.43	0.05		
	1	**0.45**	0.07		
	2	**0.53**	0.11		
	3	**0.63**	0.15		
(c)	0	0.48	0.07		
	1	0.45	**0.10**		
	2	0.44	**0.27**		
	3	0.40	**0.41**		
(d)	0	0.50	0.12		
	1	0.50	**0.22**	**0.06**	
	2	0.48	**0.28**	**0.09**	
	3	0.52	**0.39**	**0.12**	
(e)	0	0.41	0.02		
	1	0.40	0.12		**0.15**
	2	0.40	0.09		**0.30**
	3	0.43	0.08		**0.48**

phrase command (zero magnitude) occurs immediately before the verb phrase. As a result, the increment in magnitude of phrase command is in the reverse order: largest before the verb phrase as shown in (e), medium before the predicate as shown in (c), and smallest at the beginning of the utterance as shown in (b). This may be due to the physiological constraint on the magnitude of phrase command.

It is also observed that when a broad emphasis is placed on a compound word and hence increases the magnitudes of more than one phrase command, the increment of magnitude tends to be smaller on the later phrase command than on the earlier one. For instance, in utterance (d) as shown in Table 2, the increment on the phrase command preceding "bun3" is about half of that on the phrase command preceding "sei3". This can be explained by the fact that the effect of F_0 increase caused by the earlier phrase command is still continuing and hence less effort is needed to raise F_0 again for the later part.

Finally, as we can observe from Fig. 2, emphasis has significant effect not only on F_0 contour but also on syllable duration, source intensity and even pauses. Namely, duration and intensity of the emphasized part, together with duration of the short pause inserted before the part may increase steadily with the degree of emphasis, though these effects have been shown to be less consistent than that on phrase commands for F_0 contours [5]. In construction of speech synthesis systems, these effects of emphasis need also to be incorporated.

5. Discussion and conclusion

On the basis of the command-response model for the process of F_0 contour generation, we have shown that paralinguistic emphasis in Cantonese utterances hardly affects the patterns of tone commands, but it has a major effect on the relevant phrase commands. From the phonological point of view, this is a good approach for dealing with the interaction between lexical tones and phrase intonation. Since Cantonese has two tones that possess no tone commands (T3/T8), only by controlling phrase command (if leaving aside syllable duration, source intensity and pause) can we realize the emphasis without distorting lexical tone identities.

Our approach gives better insights into the effect of emphasis by use of a quantitative model separating the local tone patterns from the wide-range phrase intonation. The conclusions here can also be used to interpret the surface observations in previous work [2], where the increase of F_0 peak can be due to the phrase command, whereas the expansion of F_0 range can be due to the lengthened duration of tone command instead of any change in the amplitude.

By comparison, emphasis on a syllable/word affects only a single phrase command, whereas emphasis on a compound word may affect one or more phrase commands, depending on the prosodic structure of the compound word. Hence, without abandoning the concept of linguistic scope of emphasis in terms of speaker's intention (subjectively), a phonetic description of the narrowness/broadness of emphasis has also been introduced in terms of the number of phrase commands being affected, *viz.*, by the phonetic realization (objectively).

In text-to-speech synthesis, the pure data-driven approaches, although successfully applied in many systems, are not efficient in synthesizing expressive speech of varied speaking styles, because they cannot capture the effects of para- or non-linguistic factors very well. In order to construct a system for synthesizing speech that can express different degrees of emphasis on different parts of an utterance, the required speech corpus will be many times larger than that for synthesizing non-expressive read-style speech, and a rich transcription of speech needs to be constructed indicating the degrees of emphasis. This is not only time and storage consuming, but also makes it difficult to control the degrees of emphasis consistently during a long recording session.

The rule/model-based approaches, on the other hand, will be more suitable for expressive speech synthesis, because a set of rules can be derived explicitly from a small set of well controlled utterances. Moreover, a quantitative model will be more accurate than a set of individual heuristic rules.

The advantage of the command-response model lies in that it can capture the essential effect of a specific factor accurately by a very small number of parameters. For example, the F_0 contour of an utterance with emphasis on a particular word can be easily reproduced from that of an utterance of neutral statement by increasing the magnitude of the phrase command immediately before the target word. Since in Cantonese the increase of F_0 values by emphasis is not localized in the target constituent and not distributed evenly in the utterance, it will be difficult to capture this effect by adjusting F_0 values in each syllable separately. However, by use of the command-response model, only changing one or two parameters can be adequate.

6. References

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

[2] Man, C.-H. V., 1999. An acoustic study of the effects of sentential focus on Cantonese tones. Master thesis, University of Victoria, British Columbia, Canada.

[3] Gu, W.-T.; Hirose, K.; Fujisaki, H., 2004. Analysis of F_0 contours of Cantonese utterances based on the command-response model. *Proc. ICSLP'04*, Jeju, Korea, 781-784.

[4] Gu, W.-T.; Hirose, K.; Fujisaki, H., 2005. Identification and synthesis of Cantonese tones based on the command-response model for F_0 contour generation. *Proc. ICASSP'05*, Philadelphia, USA, 289-292.

[5] Gu, W.-T.; Hirose, K.; Fujisaki, H., 2005. Analysis of the effects of word emphasis and echo question on F_0 contours of Cantonese utterances. *Proc. Eurospeech'05*, Lisbon, Portugal, 1825-1828.