Towards Automatic Parameter Extraction of Command-Response Model for Cantonese

Raymond W.M. Ng[1], Tan Lee[1] and Wentao Gu[2]

[1]Dept. of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong
[2]Dept. of Information and Communication Engineering, University of Tokyo, Japan

{wmng; tanlee}@ee.cuhk.edu.hk wtgu@gavo.t.u-tokyo.ac.jp

Abstract

Command-response model is one of the quantitative models capable of illustrating the voice fundament frequency (F_0) characteristics of different languages. In recent years, the model has been adapted to tone languages including Cantonese. This paper presents an automatic optimization procedure to enhance parameter extraction of command-response model for short phrases of Cantonese. We conduct a parameter extraction test on 128 speech segments produced by three speakers, all of which are modeled by single phrase command in the command-response model. The extracted parameters attained accuracy compatible with that obtained from the conventional manual estimation.

Index Terms: F_0, Cantonese, command-response model, automatic parameter extraction

1. Introduction

1.1. Cantonese tone system

Cantonese is a tone language spoken in southern China, Hong Kong, Macau, and overseas Chinese communities. In Cantonese, a tone is realized over a syllable. The contrastive lexical tones in the language are associated with differences in meanings. It is commonly said that there are nine citation lexical tones in Cantonese. Figure 1 illustrates the pitch patterns of these nine lexical tones [1].

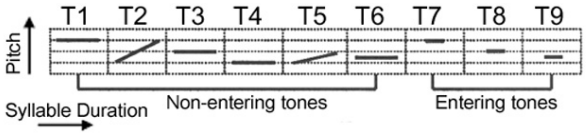

Figure 1 *Tones in Cantonese (Schematic description) [1]*

In Figure 1, T7, T8 and T9 are entering tones and they are of short durations. In terms of pitch height, T7, T8 and T9 resemble T1, T3 and T6 respectively. Therefore, the nine lexical tone categories are sometimes collapsed to six, with the difference in syllable duration neglected.

1.2. Command-response model

Voice fundamental frequency (F_0) is the direct acoustic correlate of the perceived pitch, and carries the information of both local syllable tone or word accent and global sentence intonation.

The command-response model is a quantitative model for voice fundamental frequency (F_0) proposed by Fujisaki and his coworkers. It was originally proposed for Japanese [2], [3], but later was also applied to other languages such as German [4], Mandarin [5], and Cantonese [6]. This model describes F_0 contour as an addition of three components (i.e., baseline frequency, phrase component and tone component) in the logarithmic scale. Detailed mathematical representations are given in [2]. Figure 2 shows an example, where the F_0 contour of a 5-syllable fragment of Cantonese utterance is modeled by one impulse (phrase command) and six pedestals (tone commands).

1.3. Parameter extraction

In most previous work, the underlying parameters of the model are derived by the method of Analysis-by-Synthesis [4], [6], [7]. The standard paradigm comprises two steps: manual initialization of parameters with the aid of linguistic information, and then a hill-climbing procedure of automatic optimization. Recently, fully automatic algorithms for parameter extraction were also proposed [8], [9].

For manual estimation, we can take care of every minutia and thus obtain a comprehensive and realistic picture of the intonation. However, subjectivity sets in during the course of modeling, affecting the results of Analysis-by-Synthesis. Also, the manual approach is time-consuming and labour-intensive. In automatic extraction, we face the problem of undesired optimal solutions. For instance, if the linguistic and paralinguistic factors are neglected, the extracted parameter will be meaningless. To prevent from undesired solutions, we can introduce explicit constraints.

Gu et al. [6] proposed a system of tone command patterns for each lexical tone in Cantonese, as depicted in Table 1. In the table, '– +' stands for a negative tone command followed by a positive tone command. Tone commands for T3, T8 and the second half of a T5 syllable are zero. T7, T9 are entering tones and commands are bracketed, denoting shorter durations. With explicit constraints derived from relevant information such as lexical tone identities and syllable boundaries, automatic parameter extraction of tone commands is feasible. In this study we focus on the parameter optimization for tone command estimation. Our objective is to enhance the procedure of parameter extraction such that higher efficiency can be attained and less supervision is required.

Tone Type	T1	T2	T3	T4	T5	T6	T7	T8	T9
Magnitude	+ +	– +	0 0	– –	– 0	– –	[+ +]	[0 0]	[– –]

Table 1 *Proposed System of Tone Command Polarity*

2. Methodology

2.1. Speech data

Three speakers, including one male (Speaker R) and two females (Speakers E & M), were recruited to provide the speech data in the study. Their ages vary from 22 to 35. The male speaker R and female speaker E are radio announcers. The female speaker M is a university undergraduate student. None of them has knowledge about intonation models.

12 test sentences are mixed with 2 fillers and some distracter sentences in the corpus. Each sentence comprises 4-8 short segments delimited by punctuation marks. Each short segment comprises 2 to 17 syllables. Totally there are 70 segments and 504 syllables. These sentences are transcribed scripts of radio news, newspaper, or text from primary school Chinese language textbooks.

On average each sentence consists of 42 syllables. We use breaks to delimit a recorded sentence into shorter speech segments. In [3], these break-delimited speech segments are termed *spoken sentences* or *prosodic clauses*, depending on the length of the delimiting pause, and whether or not the preceding phrase component is completely reset. Tseng [10] studied the prosodic behaviour in Mandarin Chinese and made a comparison between pauses and F_0 reset. She confirmed the validity of using pauses as the main cue of distinction of prosodic hierarchy.

Although we cannot explicitly control the pause pattern of the speakers, we can manipulate the punctuations in the corpus. During the recording sessions, we ask the speakers to try to make their rhythm of speech follow the punctuation marks.

2.2. Data processing

We first extract F_0 with *Praat* and make necessary adjustment and removal manually. Then we annotate the recorded speech, which includes LSHK standard Jyut-pin transcriptions and syllable boundaries.

After aural evaluation, perceived breaks are included to delimit the recorded speech. There are different manners of pauses and F_0 reset according to Fujisaki [3] and Tseng [10]. As a preliminary study, we assume F_0 reset occurs whenever there is a perceived break. Thus, each of these break-delimited segments is inputted to the optimization program in isolation.

For the sake of illustration in this paper, we will refer to the delimited segments mentioned above as *speech segments.*

With the measured F_0 and annotation of a *speech segment*, we follow a rule-based method to initialize the parameters. Then, a constrained optimization for the parameters is carried out. The F_0 synthesis error is defined as the *objective function*, model parameters are specified as *optimization variables*. We use the quadratic programming method implemented in *Matlab* to do error minimization and extract the optimal parameters.

In this study, we limit our interests to *speech segments* which are modeled by one and only one phrase command. We carry out manual extraction and automatic extraction in parallel, and use the manual result to sort out the target *speech segments* which are modeled by only one phrase command.

2.3. Objective function and optimization variables

The objective function of optimization is the root-mean-square of the difference between the measured F_0 contour $F_0(t)$ and the model-generated contour $F_S(t)$ as shown below:

$$\text{RMS-error} = \sqrt{\frac{1}{T_v}\sum_{t\in\Omega}(F_S(t)-F_0(t))^2} \tag{1}$$

where Ω is the set of temporal locations with valid measured F_0 data. T_v is the number of elements in Ω.

The model-generated contour $F_S(t)$ is calculated by the equation of the command-response model:-

$$\begin{aligned}\ln F_S(t) = \quad & \ln F_b + \sum_{i=1}^{I} A_{pi} G_p(t-T_{0i}) \\ & + \sum_{j=1}^{J} A_{aj}\left\{G_a(t-T_{1j}) - G_a(t-T_{2j})\right\}\end{aligned} \tag{2}$$

G_p and G_a are impulse and step functions. Details of equation (2) can be referred to in [2]. A_{pi} and T_{0i} indicate the amplitude and onset time of the *i*th phrase command respectively. A_{aj}, T_{1j} and T_{2j} indicate the amplitude, onset time and offset time of the *j*th tone command respectively. They are the optimization variables which are to be adjusted to give minimum RMS-error in equation (1).

Optimization variables are initialized by a rule-based method. Explicit constraints are also assigned. For a phrase command, A_{pi} is initialized at 0.5 and constrained between 0 and 1. T_{0i} is initialized at the first syllable onset in a *speech segment*. The lower bound of T_{0i} is at 0.2 sec before the initial value. Its upper bound is at the midpoint of the first syllable.

For tone commands, we first refer to the annotation for transcription, onset and offset boundaries of the syllable. Table 2 describes the initial and bound values assigned to A_{aj}, according to the tone category. Lower bounds of T_{1j} and T_{2j} are set respectively to 0.025 sec and 0 sec before the onset boundary. Upper bounds of T_{1j} and T_{2j} are set to exactly on the offset boundary. T_{1j} is initialized at the temporal position just after the syllable's first quadrisection, while T_{2j} just after the third quadrisection. For Tone 2 and Tone 5 syllables, there are two tone commands in one syllable, and the above temporal positions are scaled accordingly. An extra constraint forbids the overlap or inversion of orders for successive tone commands.

At most 3000 function evaluations are allowed in the iterative optimization process. Termination tolerance on the optimization variable is 1e-12.

Tone	1/7	2		3/8	4	5		6/9
LB	0.2	–0.8	0.15	0	–0.8	–0.45	0	–0.4
Initial Value	0.2	–0.15	0.15	0	–0.2	–0.1	0	–0.1
UB	0.8	–0.15	0.8	0	–0.2	–0.1	0.15	–0.1

Table 2 *A_{aj} constraints (LB: lowerbound; UB: upperbound)*

3. Results

The corpus contains 70 punctuation-delimited segments. The total number of break-delimited segments (i.e. *speech segments* as defined in Section 2.2) for speaker E, M and R are 77, 70 and 74 respectively. These patterns conform to the punctuations, with slight variations. The number of single-phrase-command *speech segments* for the three speakers are 38, 45 and 45, adding up to a total of 128 target *speech segments*.

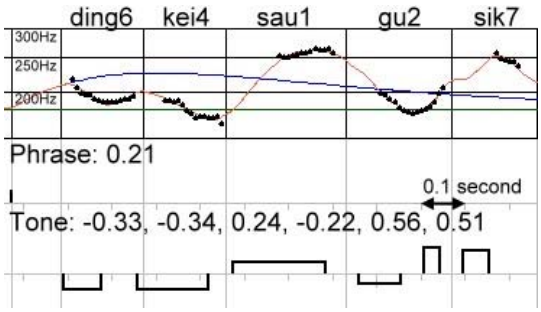

Figure 2 *Parameter extraction: [ding6 kei4 sau1 gu2 sik7]*

Figure 2 shows the reconstructed F_0 using the automatically extracted parameters for the 5-syllable *speech segment* [ding6 kei4 sau1 gu2 sik7], which means "regularly receive dividend". Measured F_0 are denoted by the dots on the top. 1 phrase impulse with amplitude 0.21 and 6 tone pedestals with amplitude -0.33, -0.34, 0.24, -0.22, 0.56 and 0.51 give the smooth interpolated of F_0 contour.

Spk	Number of Syllables	Syllable RMSE (sd)		Number of Phrases	Segment RMSE (sd)	
		Manual	Auto		Manual	Auto
E	163	10.31(6.1)	5.96(3.85)	38	10.58(3.9)	6.00(2.36)
M	228	6.73(3.93)	3.42(2.42)	45	7.30(2.53)	3.62(1.53)
R	222	6.65(4.55)	4.18(3.29)	45	7.48(3.33)	4.44(1.82)
Total	613	7.60(5.10)	4.37(3.32)	128	8.34(3.56)	4.61(2.13)

Table 3 *RMS-error wrt speaker (unit: Hz)*

Tone	Number of Syllables	Syllable RMSE (sd) / Hz	
		Manual	Auto
1	140	10.26 (5.64)	3.96 (2.20)
2	78	6.64 (4.58)	2.96 (1.67)
3	67	8.14 (5.51)	6.89 (3.45)
4	60	5.16 (3.17)	3.65 (3.13)
5	69	6.94 (4.12)	3.74 (2.52)
6	73	6.13 (3.96)	3.02 (1.84)
7	50	7.31 (3.93)	4.76 (2.94)
8	20	10.84 (7.25)	11.21 (6.71)
9	56	6.08 (4.84)	4.89 (3.85)

Table 4 *RMS-error wrt lexical tone of syllable (unit: Hz)*

The root-mean-square synthesis errors for 128 target *speech segments* are evaluated. The synthesis errors are in the unit of Hertz, indicating the average discrepancy between every point of measured F_0 and its corresponding point on the model-generated contour. Syllable RMS-error is calculated by equation (1) with Ω covering one syllable. Segment RMS-error is calculated the same way with Ω covering one *speech segment*. Table 3 and Table 4 show the RMS-errors with respect to speaker and lexical tone identity respectively. The automatic extraction set and the manual estimation set are compared.

4. Discussion

4.1. Factors affecting synthesis RMS-error

There are three factors which affect the synthesis RMS-error, namely the F_0 measurement error, micro-prosodic effects and individual factors.

Artifacts of F_0 estimation or speech with creaky voice will give incorrect F_0 measurement with extreme values. The intonation model is not capable of modeling, and should not model, these perturbations. Validation of the extracted F_0 has to be carried out before parameter extraction.

We refer to the influence of speech sounds on the F_0 contour as 'micro-prosody'. These influences are on the sub-syllabic or segmental level. They are often treated as unperceivable minutiae and thus neglected in F_0 analysis [3], [4].

In Cantonese, however, segmental effects are not totally trivial. Comparing with stress languages and pitch-accent languages, the F_0 contour of a tone language varies much rapidly and a tone command models a temporally shorter F_0 trajectory. The onset and offset of a tone command are often situated around segmental boundaries. Sometimes it even seems to be crucial that we shall pay attention to the segmental effects.

For example, an entering tone is distinct from other tones because of the segmental feature of stop codas, which gives a shorter and falling contour. In Table 4, we can see that the negligence of this segmental feature gives a generally higher RMS-error.

Table 3 shows that the RMS-errors of the three speakers vary. The primary reason for such a discrepancy is their difference in pitch excursion within a syllable and undulations across different syllables in a *speech segment*. Because the optimization variables are constrained, the pedestals of tone commands are not 'big' and 'fast' enough to reach the high peaks, low troughs, steep rising or falling.

Table 5 makes use of the pitch excursion (highest F_0 – lowest F_0) of a syllable and a *speech segment* respectively, and calculates *within-syllable excursion*, by averaging the pitch excursion of all syllables; *across-syllable excursion*, by taking the standard deviation to the pitch excursion of all syllables; and *across-segment excursion*, by taking the standard deviation to the pitch excursion of all *speech segments*.

Speaker	Idiosyncratic F_0 mean	Average Excursion		
		Within syllable	Across syllable	Across segment
E	196.4 Hz	45.96 Hz	50.11 Hz	28.63 Hz
M	221.6 Hz	30.83 Hz	38.77 Hz	26.26 Hz
R	132.4 Hz	24.80 Hz	37.08 Hz	22.68 Hz

Table 5 *Syllable F_0 excursion for 3 speakers*

Speaker E is having the biggest within-syllable excursion and this can be reflected in the highest synthesis RMS-error of her speech data in Table 3. For subject R, although his absolute excursion across phrase is not the largest, its effect stands out when we normalize the excursion by the idiosyncratic F_0 mean. This coincides with the instability of his baseline value F_b.

4.2. An efficient optimization towards the underlying factors of F_0

Table 3 shows that the automatic extraction attains an RMS-error comparable to that of manual estimation in the case of single-phrase-command *speech segments*. By the automatic extraction, the effort in manual estimation can be reduced.

Another goal of our study is to make room for further investigations in different biasing factors in the course of parameter extraction. F_0 is the result of interaction of different factors, such as linguistic formatives, physiological and

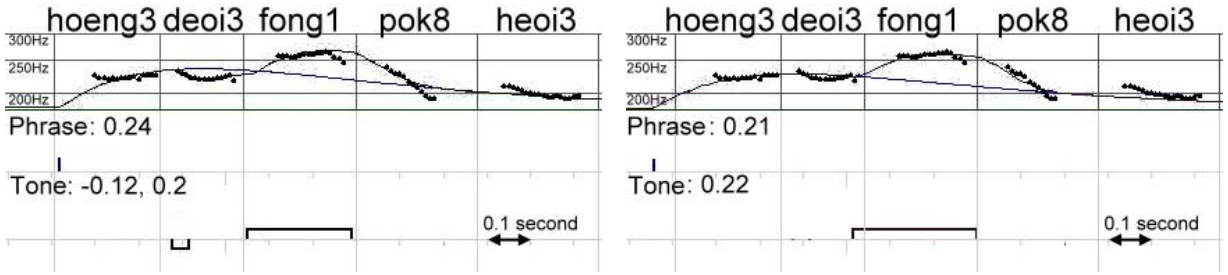

Figure 3 *Parameter extraction result for [hoeng3 deoi3 fong3 pok8 heoi3] with manual (left) and auto (right) extraction methods*

acoustical factors, etc. [11]. This inverse problem of inferring underlying commands is analytically unsolvable [7].

For example, canonical tone command amplitudes for T3 and T8 are zero (Table 2). Nevertheless, a small tone command would be allowed in special circumstances. Figure 3 above shows the analysis for [hoeng3 deoi3 fong1 pok8 heoi3], which means "(enemies) thrust into each other". In the manual case, an extra tone command is added at the second syllable [deoi3]. Meanwhile, the automatic case shows the fitting result without the extra command. Currently we are not able to tell which parameter extraction method gives a 'better' parameter set for this particular *speech segment*. We are not even sure about the existence of 'better' parameter sets. More studies would be necessary before we can draw further conclusions.

4.3. About multi-phrase-command speech segments

Other than the 128 target *speech segments*, we did not deal with other multi-phrase-command *speech segments*. These segments involve phrase command modeling. In other studies, Mixdorff [8] looked for modeling hints from low-frequency components of F_0 contour. Gu [6] took linguistic factors into consideration. For a given speech segment, it is not always easy to tell the number of phrase commands necessary in modeling.

Although it is technically possible to give rules accommodating extra phrase commands such that modeling can be done, explicit constraints are yet to work out. We evaluate the RMS-error of these long segments under the current automatic extraction strategy. In these partially optimized long segments, we leave out all phrase commands but the first. Their average segment RMS-error is 7.15Hz (*sd* 3.63Hz). This shows a contrast to the average segment RMS-error of short segments (4.61Hz, *sd* 2.13Hz). Mixdorff claims "the omission of phrase commands would reduce the modeling accuracy for German sentences" [4]. It is also valid for Cantonese.

Nevertheless, how and where to assign extra phrase commands remains a non-deterministic question. Prosodic phrase duration, syllable count, and synthesis error in sub-syllabic level have been investigated. We found that these mentioned factors are necessary, but not sufficient cues. More acoustic comparison and linguistic analysis may be necessary.

5. Conclusion

This study demonstrates automatic parameter extraction of command-response model for single-phrase-command *speech segments* in Cantonese. Methodology is explained and the fitting error is used as the general indicator for evaluation. We look for factors affecting the fitting accuracy, and we argue that automatic parameter prediction makes room for future enhancement in efficiency and objectivity. Finally, we study briefly the long *speech segments* as a suggestion towards a comprehensive parameter extraction system.

6. Acknowledgments

This research is partially supported by an Earmarked Research Grant (Ref: CUHK 4227/04E) from the Hong Kong Research Grants Council.

7. References

[1] Li, Y.J., Lee, T., Qian, Y., "Acoustical F0 analysis of continuous Cantonese speech," in *Proc. ISCSLP 2002*, pp. 127-130.

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

[3] Fujisaki, H. et al., "Manifestation of linguistic information in the voice fundamental frequency contours of spoken Japanese," *IEICE Trans. Fundamentals of Electronics, Communications and Computer Sciences*, vol. E76-A, no. 11, pp. 1919-1926, November 1993.

[4] Mixdorff, H., "Intonation patterns of German – model-based quantitative analysis and synthesis of F_0-contours," Ph.D. dissertation, TU Dresden, Germany, 1998.

[5] Fujisaki, H., Wang, C., Ohno, S., Gu, W., "Analysis and synthesis of fundamental frequency contours of Standard Chinese using the command-response model," *Speech Communication*, vol. 47, no. 1-2, pp. 59-70, 2005.

[6] Gu, W., Hirose, K., Fujisaki, H., "Analysis of F0 contours of Cantonese utterances based on the command-response model," in *Proc. INTERSPEECH 2004*, pp. 781-784.

[7] Fujisaki, H., "Prosody, models, and spontaneous speech," in *Computing Prosody: Computational Models for Processing Spontaneous Speech*, Sagisaka, Y., Eds., NY: Springer, 1997, pp. 27-42.

[8] Mixdorff., H. et al., "Towards the automatic extraction of fujisaki model parameters for Mandarin," in *Proc. Eurospeech 2003*, pp. 873-876.

[9] Gu, W., Hirose, K., Fujisaki, H., "A general approach for automatic extraction of tone commands in the command-response model for tone languages", in *Proc. Speech Prosody 2006*, pp. 153-156.

[10] Tseng, C.-Y. et al., "Speech prosody: issues, approaches and implications," in *From Traditional Phonology to Modern Speech Processing*, Fant, G., Eds., Beijing: FLTR Press, 2004, pp. 417-437.

[11] Wang, W.S.-Y., "The many uses of F0", in *Explorations in Language*, Taipei: Pyramid Press, 1991, pp. 193-204.