# THE EFFECT OF TONAL CONTEXT ON CANTONESE CONCATENATIVE SPEECH SYNTHESIS

*Tien-Ying FUNG* and *Helen M. MENG*

Human-Computer Communications Laboratory,
Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong, Shatin, N.T,
Hong Kong SAR
tyfung@se.cuhk.edu.hk, hmmeng@se.cuhk.edu.hk

## ABSTRACT

This paper describes our study of the effect of tonal context on Cantonese concatenative speech synthesis. We have previously developed a speech synthesizer, CU VOCAL, that concatenates syllables to generate Cantonese and Mandarin speech [1, 2]. The preliminary version of CU VOCAL captures only the place of articulation as coarticulatory context by the use of distinctive features in unit selection [3]. However, we noticed discrepancies between the perceived tone and the desired tone for some Cantonese syllables in the synthesized speech, which affected the perceived quality of the synthesis outputs. This suggests the need to extend our unit selection strategy to incorporate tonal context as well. In order to devise such a strategy, we studied the comparative importance between the left and right tonal contexts in terms of their influence on the perceived tone of the current syllable. We also defined a scheme by which we can measure the difference between a desired syllable token and its tonal variant, in terms of attributes such as tone shape, tone height and tone trajectory. Hence, if a desired syllable token is unavailable during concatenative synthesis, we can substitute with its "closest" tonal variant as suggested by our unit selection scheme.

## 1. INTRODUCTION

Cantonese is a major dialect of Chinese, spoken by over 64 million people in South China, Hong Kong, Macau and many overseas Chinese communities [4]. The dialect has nine tones, hence Cantonese presents a rich tonal structure for the study of tones. The acoustic correlate of tone is fundamental frequency (f0), and it has been observed that the actual realization of f0 in a tonal syllable is heavily influenced by its neighboring syllables in continuous speech. We aim to investigate how the perceived tone of a syllable may be affected by the tones of the surrounding (left and right neighboring) syllables. The findings of this study are applied to Cantonese synthesis by our text-to-speech engine, CU VOCAL [2]. CU VOCAL is a corpus-based concatenative synthesizer for both Cantonese and Mandarin that uses the tonal syllable as the basic unit for concatenation. Coarticulation is modeled only in terms of the place of articulation by the use of distinctive features. The investigation presented in this paper aims to extend the unit selection strategy in CU VOCAL to incorporate considerations in tonal context, thereby improving the perceived quality of the synthesis outputs. This paper focuses on Cantonese synthesis.

## 2. PROPERTIES OF CANTONESE TONES

### 2.1 Cantonese Tone System

The pronunciation of a Chinese character can be represented with a syllable and a lexical tone. Cantonese has nine tones as shown in Figure 1.

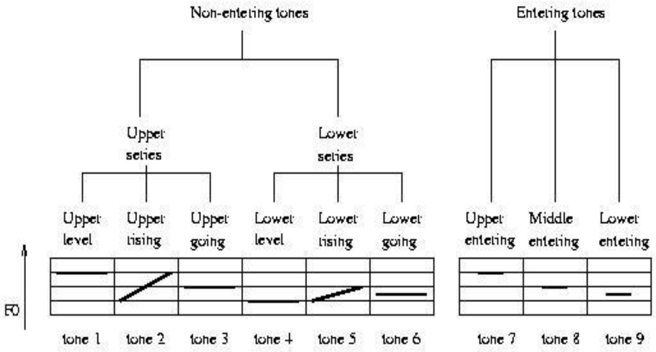

Figure 1. The Cantonese nine-tone system [5].

Based on duration, the nine tones can be divided into two groups: *non-entering tones* (longer in duration) and *entering tones* (shorter in duration). Within each group the tones can be further distinguished by their heights and shapes. Tone height refers to the level of f0 (fundamental frequency) and tone shape describes the trajectory of the f0 within the syllable. Since our investigation focuses only on the tone height and tone shape, we can consider only the non-entering tones and reduce the number of tones to six. There are two different shapes in the six tones: the *level* tones (tones 1, 3, 4 and 6) and the *rising* tones (tones 2 and 5). Tones with the same shape are further distinguished by tone heights.

### 2.2 Role of Tone in Corpus-based Speech Synthesis

The preliminary version of CU VOCAL captures only the place of articulation as coarticulatory context by the use of distinctive features in unit selection [3]. However, we noticed discrepancies between the perceived tone and the desired tone for some Cantonese syllables in the synthesized speech. Such discrepancies affect the perceived quality of the synthesis outputs. The acoustic realization of a syllable's tone (in terms of f0) is affected by the tones of the neighboring syllables, i.e. the tonal context. Thus the tonal context can distort the tone height and shape of a syllable, which may affect the perceived tone of the syllable.

Consider an example in continuous speech “零三六三上海實業”(translation: 0363 Shanghai Industry Holdings) as shown in Figure 2:

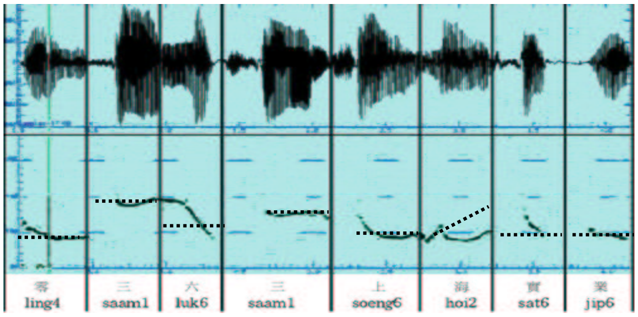

Figure 2. Example of the pitch contour in the real recorded phrase of “零三六三上海實業”. The solid line and the dotted lines show the real tone contours and the canonical tone contours respectively.

The character sequence is pronounced as the tonal syllable sequence /ling4 saam1 luk6 saam1 soeng6 hoi2 sat6 jip6/. Tonal distortion can be found in the syllables /luk6/, /soeng6/ and /sat6/. In each case, the syllable should have a level canonical tone shape, but since it is preceded by a left syllable of high f0 level, the f0 trajectory has to dive rapidly towards the target f0, leading to a rapidly falling tone shape instead. In contrast, some syllables such as /ling4/, /saam1/ and /jip6/ largely maintain their level tone shapes in different tonal contexts. If a syllable segment with distorted tone shape is used in concatenative synthesis for an alternate tonal context, the distortion may lead to the perception that an incorrect tone has been synthesized.

## 3. COMPARATIVE IMPORTANCE BETWEEN THE LEFT AND RIGHT TONAL CONTEXTS

### 3.1 Experiments with Digit Triplets

We investigate the relative importance between the left and right tonal contexts based on Chinese numeric characters. This is because the nine numeric characters ranging from zero(零) to nine(九) fully cover the six Cantonese tones as illustrated in Figure 3.

| Tone 1 | Tone 2 | Tone 3 | Tone 4 | Tone 5 | Tone 6 |
|---|---|---|---|---|---|
| 一(1)/jat1/<br>三(3)/saam1/<br>七(7)/caat1/ | 九(9)/gau2/ | 四(4)/sei3/<br>八(8)/baat3/ | 零(0)/ling4/ | 五(5)/ng5/ | 二(2)/ji6/<br>六(6)/luk6/ |

Figure 3. The Chinese numeric characters ranging from zero(零) to nine(九) have syllable pronunciations that fully cover the six Cantonese tones.

A given Chinese numeric character (and its corresponding tonal syllable) at the center of a triplet may be preceded by any one of the six tones due to its left neighboring syllable as well as followed by any one of the six tones due to its right neighboring syllable. Hence the tonal syllable at the center of a triplet has different instances due to different tonal contexts. We refer to these different instances as *tonal variants* of the syllable. We recorded a corpus of digit triplets, which attains a 60% coverage of the possible tonal variants for each tonal syllable.

To investigate the comparative importance of the left and right tonal contexts, we performed syllable-based concatenative synthesis using our corpus of Chinese digit triplets e.g. “零一三” (translation: “zero one three”, pronunciation: /ling4 jat1 saam1/). We synthesized a pair of waveforms for each digit triplet by using a different tonal variant for the central syllable (e.g. **/jat1/** in /ling4 **jat1** saam1/). One waveform in the pair incorporates a syllable with matching tonal contexts (i.e. the MATCHED condition). The other waveform corresponds to a one-sided MISMATCHED condition. We synthesized 16 waveform pairs in total – 8 pairs (i.e. MATCHED versus MISMATCHED) with respect to the left tonal context (denoted as category L); and the other 8 pairs with respect to the right context (denoted as category R). Examples of the waveform pairs in the two categories are shown in Figure 4.

| Category | Triplet | MATCHED condition | MISMATCHED condition |
|---|---|---|---|
| L | “013” | /ling**4** (**4**)jat1(1) saam1/ | /ling**4** (**0**)jat1(1) saam1/ |
| R | “244” | /ji6 (6)sei3(**3**) sei**3**/ | /ji6 (6)sei3(**6**) sei**3**/ |

Figure 4. Example of the waveform pairs in categories L and R. For category L, we have one synthesized waveform for a digit triplet, in which the central syllable segment /jat1/ is chosen such that its left tonal context (tone 4) matches that of the preceding syllable /ling4/. The other synthesized waveform presents a mismatched condition, where the left tonal context (tone 0) is used/. Similarly, we have matched and mismatched conditions for the right tonal context in the pair of waveforms under category R.

### 3.2 Listening Tests

A listening test was set up as a within-group experiment. 72 university students aged between 20 to 25 were invited to be our subjects. Precautions were taken to ensure an unbiased environment for the listening test. For example, for each pair of waveforms synthesized (under MATCHED and MISMATCHED conditions), we randomized the order of the pair. We also randomly choose between a mismatch in the left or right tonal context. Each pair of waveforms was played three times before each listener was asked to record his/her judgment regarding *one* of the following:

- the former waveform sounded better than the latter
- the former waveform sounded equally good as the latter
- the former waveform sounded worse than the latter

These judgments were re-organized into the following categories:

- the MATCHED condition sounded better than the MISMATCHED condition (denoted by MATCHED> MISMATCHED)
- the MATCHED condition sounded equally good as the MISMATCHED condition (denoted by MATCHED=MISMATCHED)
- the MATCHED condition sounded worse than the MISMATCHED condition (denoted by MATCHED< MISMATCHED)

Since we have 576 waveform pairs, we collected 576 ratings (judgments). Results for comparative experiments in category L (MATCHED versus MISMATCHED left tonal context) and category R (MATCHED versus MISMATCHED right tonal context) are shown in Figures 5 and 6 respectively.

Based on the experimental results in category L, we conducted two statistical tests. The first is a two-tailed test to establish that there are perceivable differences between the MATCHED condition and the MISMATCHED condition in the left tonal context. The second is a one-tailed test that focuses only on the subset of waveforms with perceived differences in the subjects' ratings. This test shows that the listeners rate the MATCHED condition superior to the MISMATCHED condition in terms of synthesis quality. Both tests gave statistically significant results at $\alpha=0.01$. Similarly, we conducted two statistical tests based on the experimental results in category R. Results from both tests (with $\alpha$=0.01) indicate that there is no perceivable differences between the MATCHED and MISMATCHED conditions in the right tonal context.

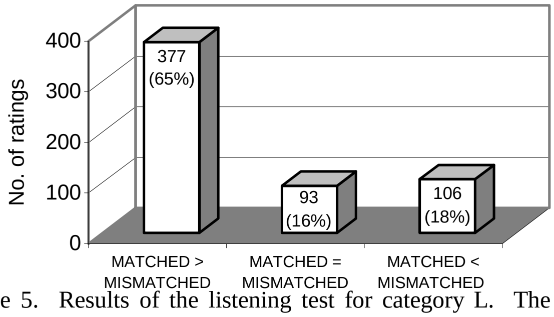

Figure 5. Results of the listening test for category L. The figure shows ratings regarding three possible judgments. The ratings indicate that a matched left tonal context sounds significantly better than a mismatched case for concatenative synthesis.

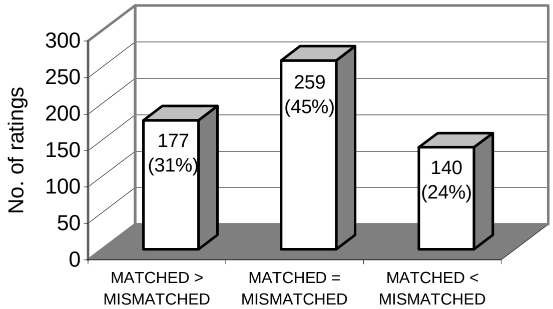

Figure 6. Results of the listening test for category R. The results show that no significant difference is perceived by our subjects between a matched right tonal context and a mismatched case for concatenative synthesis.

## 4. UNIT SELECTION SCHEME FOR TONAL VARIANTS

Our investigation thus far has indicated the importance of the left tonal context for syllable-based concatenative synthesis. We proceed to incorporate this finding in our unit selection strategy. We have devised the following scheme for unit selection based on tonal features. As we concatenate syllable units from left to right during synthesis, we need to analyze the *desired* tonal context based on the original character sequence. The ideal scenario is when our synthesis corpus can provide a syllable unit with matching left and right tonal contexts. If such an instance cannot be found in our corpus, we attempt to enforce a match in the left tonal context. Otherwise, we follow the incremental matching rules as follows:

**Rule 1:** We favor a syllable instance that maintains the slope in the tone trajectory going from the preceding syllable unit to the current syllable unit.

**Rule 2:** If the condition as specified in rule (1) is met, we attempt to find the syllable unit whose left tonal context has the same tone shape as that of the desired syllable unit.

**Rule 3:** If the conditions as specifed in rules (1) and (2) are met, we attempt to find the syllable instance that minimzes transitional movements in the tone trajectory going from the preceding to the current syllable unit.

**Rule 4:** We avoid the use of a syllable instance that has tone 2 as its left context. Such a tonal context tends to produce a dynamic and transitional trajectory in the syllable instance.

In the following we explain each rule in detail.

**Rule 1: Maintain slope in tone trajectory between connected syllables**

Assume that the target syllable unit for concatenation is one with tone $T$ and left tonal context $L_D$, denoted by ($L_D$)SYL$T$. As we proceed from the (left) preceding syllable to the current syllable, the observed difference in tone height is $d$=$T$-$L_D$. We denote the substituting tonal variant used for concatenation as ($L_S$)SYL$T$, where $L_S$ is the substituted left tonal context. The difference in tone height observed in this tonal variant substitute is $d'$=$T$-$L_S$. Then the substitute is chosen such that $d'$ and $d$ have the same sign (both positive or negative). This maintains the slope in the tone trajectory as we move from the preceding syllable to the current one. This is illustrated in Figure 7.

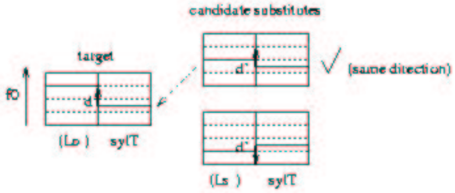

Figure 7. To find a tonal variant ($L_S$)SYL$T$ to substitute for ($L_D$)SYL$T$, we compare the signs of $d$=($T$-$L_D$) and $d'$=($T$-$L_S$), and select ($L_S$)SYL$T$ such that $d$ and $d'$ have the same sign.

**Rule 2: Preserve the tone shape**

Given that above condition is met, we favor substitutions with tonal variants whose left tonal context $L_S$ has the same tone shape as $L_D$. Recall that there are two tone shapes – *rising* (as in tones 2 and 5) and *level* (as in tones 1, 3, 4 and 6). This is illustrated in Figure 8.

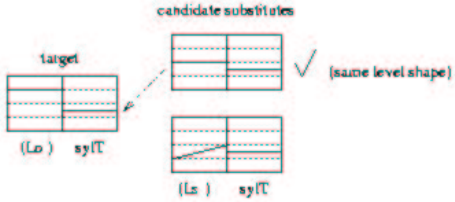

Figure 8. To find a tonal variant ($L_S$)SYL$T$ to substitute for ($L_D$)SYL$T$, we compare the tone shapes of $L_D$ and $L_S$, and favor ($L_S$)SYL$T$ whose $Ls$ has the same tone shape as $L_D$.

**Rule 3: Minimize transitional movements in the tone trajectory**

We attempt to apply rule (3) given that conditions in rules (1) and (2) have been met. We favor substituting with the tonal variant that gives a $d'$ value whose magnitude is smaller than but closest to $d$. This principle avoids large transitional movements in the tone trajectory going from the preceding syllable to the current syllable. This is illustrated in Figure 9.

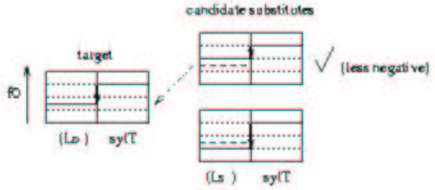

Figure 9. To find a tonal variant ($L_S$)SYL$T$ to substitute for ($L_D$)SYL$T$, we compare the magnitudes of $d$=($T$-$L_D$) and $d'$=($T$-$L_S$). We choose the tonal variant that gives a $d'$ value whose magnitute is smaller than but closest to $d$.

**Rule 4: Avoid using syllables with tone 2 as its left context**

We try to avoid substituting with tonal variants whose left context $L_S$ is tone 2. This is because tone 2 has one of the most dynamic tone shapes and often leads to overshooting and undershooting in tone trajectories, which distorts the tone shape and height of the current syllable. This is illustrated in Figure 10.

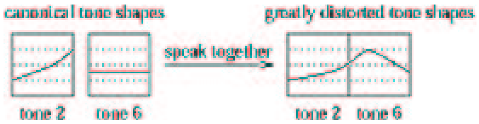

Figure 10. We avoid using tone variants ($L_S$)SYL$T$ that have $L_S$ being tone 2 since it tends to introduce overshooting and undershooting in the tone trajectory.

## 5. LISTENING TEST TO VALIDATE SELECTION SCHEME FOR TONAL VARIANTS

We designed a listening test to assess the validity of our unit selection scheme for tonal variants. This is again based on synthesized digit triplets in Cantonese, similar to the listening test in the previous section. The digit triplet calls for a tonal variant for the central syllable with tone $T$ and *desired* left context $L_D$. Instead, we concatenate with the tonal variant (with tone $T$) and *substituted* left context $L_S$ according to our selection scheme. We have used 15 digit triplets, covering six tones for $T$ and five tones for $L_D$, but not all combinations are included due to the limited size of our syllable corpus. For a given $T$ and $L_D$, our selection scheme provides a rank order of up to five alternatives for $L_S$. Of these five alternatives, we include three in our generated waveforms. Hence, for each digit string, we generate a group of four waveforms.

As an example, consider the digit string “九一三” (i.e. nine one three) pronounced as /gau2 jat1 saam1/:

- the first generated waveform is /gau2(1) (2)jat1(1) (1)saam1/ (ideal case with matching tonal variants, denoted as REF)
- the second generated waveform is /gau2(1) (5)jat1(1) (1)saam1/ (replaced the $L_D$ = 2 with $L_S$ = 5) (denoted as SUBSTITUTE1)
- the third generated waveform is /gau2(1) (4)jat1(1) (1)saam1/ (replaced the $L_D$ = 2 with $L_S$ = 4) (denoted as SUBSTITUTE2)
- the fourth generated waveform is /gau2(1) (1)jat1(1) (1)saam1/ (replaced the $L_D$ = 2 with $L_S$ = 1) (denoted as SUBSTITUTE3)

During the listening test, we began by playing the REF waveform, followed by the other three waveforms in randomized order. Subjects were asked to rank SUBSTITUTE1, SUBSTITUTE2, SUBSTITUTE3 in the descending order of synthesis quality. We then compare the subjects' rankings with those suggested by our selection scheme. For example, if the listener's ranking is: SUBSTITUTE1 > SUBSTITUTE2 = SUBSTITUTE3 then the pairs are (1>2), (1>3) and (2=3). Furthermore, if the selection scheme suggests the ranking: SUBSTITUTE1 > SUBSTITUTE3 > SUBSTITUTE2 and then the pairs are (1>2), (1>3) and (3>2). Comparison between these two sets of rankings shows that two pairs out of the three, i.e. (1>2) and (1>3), are in agreement ($N_{agree}$=2), and the remaining pair has no perceivable difference ($N_{no\_difference}$=1). Hence we also denote the number of pairs in disagreement as ($N_{disagree}$=0). Recall that we have 15 digit triplets from which we generate 45 waveforms. Each waveform is rated by 55 listeners. Hence we have 2475 pairs from which we can derive $N_{agree}$, $N_{disagree}$, and $N_{no\_difference}$, as illustrated in Figure 11. There is significant agreement between human judgement and our selection scheme.

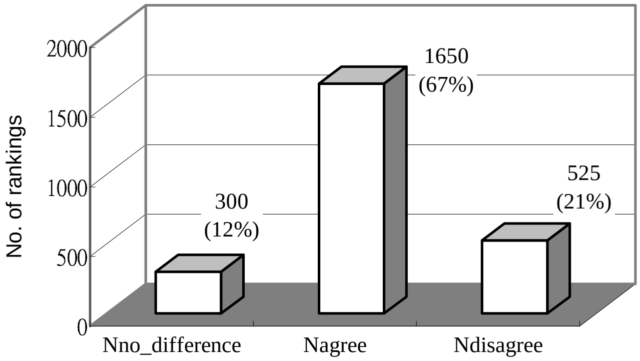

Figure 11. Results of the listening test to assess the validity of our selection scheme for tonal variants. $N_{no\_difference}$ is the number (or fraction) of rankings through which the listener indicates no perceivable difference between different synthesized outputs. $N_{agree}$ is the number (or fraction) of rankings with agreement between the listener's judgement and our selection scheme. $N_{disagree}$ is the corresponding number with disagreement.

## 6. CONCLUSIONS

This paper investigates the effect of tonal contexts in Cantonese syllable concatenation. We began with a study of the relative importance between the left and right tonal contexts in concatenative synthesis. Our listening tests indicate that the left tonal context exerts significant effect in the quality of the synthesized outputs. We incorporated this finding in the development of a syllable unit selection scheme for tonal variants in place of a desired syllable. The scheme is applied in our concatenative synthesizer under the condition that a desired syllable unit (i.e. one with matching left and right tonal contexts) is non-existent in our corpus. The selection scheme rank orders the available tonal variants to be used as substitutes. We assessed the validity of our selection scheme through comparison with human judgments based on a listening test. We found significant agreement between our selection scheme and human perception 67% of the time.

## 7. ACKNOWLEDGMENT

This research is funded by a Direct Grant 2001-2002 from The Chinese University of Hong Kong. We thank Li Yuk Chi and Keung Chi Kin for their comments.

## 8. REFERENCES

[1] Fung, T. Y. and H. Meng, “Concatenating Syllables for Response Generation in Spoken Language Applications,” *Proceedings of ICASSP 2000.*

[2] Meng, H. et al, “CU VOCAL: Corpus-based Syllable Concatenation for Chinese Speech Synthesis across Domains and Dialects,” *Proceedings of the International Conference on Spoken Language, 2002.*

[3] Rabiner, L. R. and Schafer, R. W. “Digital Processing of Speech Signals” *pages 39-41, Prentice-Hall, 1978.*

[4] Grimes, B. F., ed. 1992. Ethnologue: Languages of the World. Dallas, Texas: Summer Institute of Linguistics

[5] Lee, T., H. Meng, W. Lau, W. K. Lo and P. C. Ching, "Micro-prosodic control in Cantonese text-to-speech synthesis", *Proceedings of the Sixth European Conference on Speech Communication and Technology*, vol. 4, pp. 1855-8, Budapest, 1999.