THE EFFECT OF DOWNDRIFT IN THE PRODUCTION AND PERCEPTION OF CANTONESE LEVEL TONES

Patrick C.M. Wong
University of Texas at Austin, Austin, Texas, USA

ABSTRACT

This study investigates downdrift in Cantonese. Analysis of Cantonese sentences spoken by seven speakers found that downdrift exists in Cantonese production. In a subsequence perception experiment, the same sentences were resynthesized so that the context was divided into two halves and their pitch was raised or lowered independently while the pitch of the target remained constant. Listeners identify the pitch of the target syllable based on the latter half of the context, i.e. a recency effect. Furthermore, when the pitch pattern of the initial half of the context was lower than the latter half (i.e. a violation of downdrift), listeners were less likely to make tonal judgement based on the latter half. These results are explained by listeners' expectation for downdrift when perceiving tones.

1. INTRODUCTION

Most, and maybe all, languages in the world exhibit a gradual fall in pitch from the beginning to the end of an utterance [1]. This phenomenon, known as "downdrift," has been observed in both tone languages and nontone languages. In tone languages, downdrift was observed primarily in African tone languages (e.g. Akan [2]). In tone production, the result of downdrift can be that successive tones become lower and lower in pitch, resulting in a situation where at the end of an utterance, a high tone can have a fundamental frequency (F0) which is as low as the low tone at the beginning of the utterance.

Downdrift also affects speech perception. The auditory reality of downdrift was first investigated by Breckenridge [3]. Listeners in her study were asked to judge the peaks of two pitch prominences in a synthesized English sentence. It was found that listeners judged a syllable occurring later in an utterance to have the same pitch as a syllable occurring earlier even when it was a few hertz higher. Similar results were found by Pierrehumbert [4] in which English listeners judged two stressed syllables to be equal in pitch when the second syllable was actually lower. Pierrehumbert argued that listeners' judgements reflected normalization for expected declination in pitch toward the end of an utterance.

From the previous studies, we know that downdrift is both a production and perception phenomena, and it is present in both tone and nontone languages. However, all the perceptual experiments of downdrift have focused on nontone languages. The present study, in addition to investigating downdrift in the production of a tone language other than African tone languages, investigates the perceptual consequences of downdrift in a tone language.

2. CANTONESE TONES

Cantonese is the language of investigation in this study. Cantonese has six tones: Tone 1, high level; Tone 2, high rising; Tone 3, mid level; Tone 4, low falling; Tone 5, low rising; and Tone 6, low level. The three level tones, i.e. Tone 1, Tone 3, and Tone 6, were used in this study. Since downdrift, and related phenomena like downstep, has been reported to occur in tone languages which have primarily level tones [5], Cantonese is a good candidate for the investigation of downdrift.

3. TONE NORMALIZATION

3.1 Introduction

In their study, Wong and Diehl [6] investigated tone normalization by examining how Cantonese level tones were perceived when preceded by a five-syllable context. They raised and lowered the F0 pattern of the context as a whole and found that a stimulus that is perceived as a low tone when preceded by a high-F0 context can be perceived as a high tone when preceded by a low-F0 context. Because the present study employed a similar methodology, the next section will discuss the methods used in the Wong and Diehl study.

3.2 Methods

The following sentence spoken by seven native Cantonese speakers was used: /ha6 yat1 go3 ji6 hai6 si3/ ('The next word is try"). This sentence will be called the "Original Sentence." The context (i.e. the whole sentence except for the last word) was re-synthesized using the Kay Analysis/Synthesis Laboratory (ASL) so that the F0 of each word in the context could be raised or lowered. According to Chao [7], Tone 3 is approximately three musical semitones lower than Tone 1 and two musical semitones higher than Tone 6[1]. Accordingly, two new sentences were generated, one being two musical semitones higher than the Original Sentence (this will be called the "Tone 6 Sentence" because the last word is predicted to be perceived as the Tone 6 'yes'), and one being three semitones lower (this will be called the "Tone 1 Sentence"). As an increase of one musical semitone reflects approximately a 6% increase in F0, the contexts were resynthesized as follows: the F0 of the context of the Tone 6 Sentence was 1.13 (1. 06^2) times the Original Sentence, while the F0 of the context of the Tone 1 Sentence was 0.82 ($1/1.06^3$) times the Original Sentence. Instead of using the Original Sentence as a stimulus token, a "Tone 3 Sentence" was generated by raising the F0 of the context of the Original Sentence by 1% (raising the F0 by 1% should not be large enough to affect tone perception). Because the quality of the resynthesized sentences was slightly different from the original sentences, generating a re-synthesized Tone 3 sentence helped to ensure that the quality of all the stimulus sentences was the same.

Therefore, for each speaker, there were three resynthesized sentences generated. Listeners listened to the 21 stimulus sentences five times each.

3.3 Results and Discussion

Wong and Diehl [6] found that in 94.34% of the time listeners identified the target word (the last word in an stimulus sentence) based on the context as predicted. In other words, they identified the target syllable as Tone 6 in the Tone 6 stimulus sentences (i.e. when the context was raised by two musical semitones); they identified the target as Tone 1 in the Tone 1 stimulus sentences (i.e. when the context was lowered by three musical semitones), and they identified the target as Tone 3 in the Tone 3 stimulus sentences (i.e. when the context was raised by 1%).

The results showed that listeners' tonal judgement was made based on the preceding context. The same syllable (having the same F0) could be perceived as different tones depending on the pitch pattern of the preceding context. Based on the Wong and Diehl [6] study, the present study was designed. However, in this study, the context was divided into two halves and their pitch patterns were raised and lowered independently. The results tell us on which half the listeners based their tonal judgment. The results are explained by downdrift.

4. METHODS

4.1 Subjects

4.1.1. Speakers. Speakers were seven male native Cantonese speakers (same speakers used in Wong and Diehl [6]). They were attending the University of Texas at Austin and were paid for their participation in the study.

4.1.2. Listeners. Listeners were eight native Cantonese speakers, four females and four males. They were also attending the University of Texas at Austin and were paid for their participation in the study.

4.2. Stimuli

4.2.1. Original Sentence. They same Original Sentences from the Wong and Diehl [6] study were used. Speakers were asked to produce the following Cantonese sentence three times: /ha6 yat1 go3 zi6 hai6 si3/ 'The next word is teacher.' Based on his judgement as a native Cantonese speaker, the author chose the best exemplar of the three instances to be the sentence for resynthesis.

4.2.2. Resynthesis. Resynthesis procedures were similar to what was employed in Wong and Diehl [6]. However, in this study, the context was divided into two halves. The first half consisted the first three syllables /ha6 yat1 go3/ and the second half consisted the last two syllables /zi6 hai6/. On the average, the first half (375.13 msec long) was 26.4 msec longer than the second half (348.73 msec long). The pitch of the two halves of the context were raised or lowered independently.

In Wong and Diehl [6], the context as a whole was raised or lowered and stimulus sentences were named depending on how the target was predicted to be identified. For example, a stimulus sentence was called a "Tone 6 Sentence" when the context was raised by two musical semitones because the authors predicted that the target syllable in that context would be identified as Tone 6. In the case of this study because the two halves of the context were raised and lowered independently, stimulus sentences were named differently. In all, six resynthesized sentences were generated for each speaker and were named as follow: F1S3 (first half Tone 1, second half Tone 3), F1S6, F3S1, F3S6, F6S1, and F6S3. In F1S6, i.e. first half Tone 1, second half Tone 6, the first half was lowered by three semitones and the second half was raised by two semitones. Accordingly, if listeners made tonal judgements on the basis of the first half of the sentence, they would identify the target tone as Tone 1; if they based their perception on the second half, they would identify the target tone as Tone 6. The other stimulus sentences can be explained in a similar fashion.

4.3. Procedures

Listeners listened to each of the six resynthesized sentences for each of the seven speakers five times; therefore, they listened to 210 sentences (6 sentences x 7 speakers x 5 times). A practice session consisting of all the stimulus sentences was administered before the actual experiment. All listening sessions were conducted in a double-walled sound-attenuated chamber.

Listeners were asked to identify the last word (the target) of the stimulus sentences they heard. Responses were made by pressing one of the three response keys labeled with the Chinese character of the following three words: 'teacher' (for /si1/), 'to try' (for /si3/), and 'yes' (for /si6/).

5. RESULTS

5.1 Production

The F0 for each syllable of the original, non-resynthesized sentences, spoken by the seven speakers was obtained and is summarized in Table 1. A paired t-test was performed and showed that the last mid tone (Tone 3) syllable /si3/ in the Original Sentences did not differ from the first low tone (Tone 6) syllable /ha6/ in F0 [$t(6) = -2.194$, $p > .07$]. Thus, a mid tone at the end of a sentence was as low as a low tone at the beginning of the sentence. Further, the syllable /go3/ had significantly higher F0 than /si3/ [$t(6) = 3.017$, $p < .025$]. Thus, the same tone, when it occurred earlier in an utterance, had a higher F0.

5.2 Perception

Table 2 summarizes the results of the perceptual experiment. It was found that for all the stimulus sentences, listeners mostly based their tonal judgement on the second half of the sentence (the highest response rate for each stimulus sentence type is underlined). For example, for stimulus F3S1, listeners perceived the target syllable to be Tone 1 over 82% of the time. When all the stimulus sentences were taken into consideration, listeners identify the target syllable based on the second half of the context 77.69% of the time. Thus, we found that there was a recency effect in tone perception meaning that listeners identified tones mostly on the basis of the context that was closer to the target.

6. DISCUSSION

6.1 Production

The results clearly show that downdrift exists in Cantonese. The final syllable /si3/ which has the mid tone was lower than the mid tone /go3/ that occurred before it, and had compatible F0 with the low tone /ha6/ that occurred at the beginning of the utterance. Given what is known from research on African tone languages,

this is not surprising. The production results, however, give us more information on downdrift in non-African tone languages.

6.2 Perception

The recency effect observed in perception can be explained by the fact that listeners "expected" that the tones at the beginning of an utterance would have different F0 patterns than tones toward the end (even when the two tones were the same). In this case, listeners expected downdrift in which tones at the beginning had higher F0 than tones at the end. For more accurate tonal judgement, it was more reliable for listeners to rely on the latter half of the context which was closer and had a more similar F0 pattern to the target syllable.

Table 2 shows that listeners made tonal judgement mostly based on the latter half of the context; however, they did it more so in some incidences (e.g. F6S3) than others (e.g. F1S3). This can also be explained by downdrift. In cases like F1S3 and F1S6 when listeners relied on the latter half of the context less (68.11% and 68.03% respectively), the initial half of the context (i.e. F1) had a lower pitch pattern than the latter half. Recall that in F1, the first half of the sentence was lowered by three musical semitones. Because in downdrift, the initial half of an utterance is higher in pitch than the latter half, both F1S3 and F1S6 violated the overall downdrift pattern expected by listeners. Listeners, therefore, could not use the usual "recency strategy" they used to identify the target syllable reliably compared to instances when the overall downdrift pattern was preserved (as in F3S1 and F6S1).

7. CONCLUSION

This study, in addition to providing production data for downdrift in a tone language other than African tone languages, provided data on the auditory reality of downdrift in a tone language. It showed that when identifying tones, listeners used a recency strategy. Listeners depended more on the tones that were closer to the target syllable because of downdrift. It is because as a result of downdrift, the initial and latter halves of the context had different pitch pattern, and the latter half had more similar pitch pattern to the target syllable. If the overall downdrift pattern was violated, listeners were more confused and used the recency strategy less.

ACKNOWLEDGMENTS

The author wishes to thank Dr. Randy Diehl for his support and insightful comments. This work is supported by an NIDCD grant given to Dr. Diehl.

NOTES

1. In Chao [7], Tone 1 produced by a male speaker was claimed to be four musical semitones higher than Tone 3, and Tone 3 was two musical semitones higher than Tone 6. However, in the musical score he showed for all the tones, Tone 1 was only three musical semitones higher (instead of four). Based on the author's judgment as a native speaker, the author found that three musical semitones sounded more natural than four, and therefore was the value used for re-synthesis in this study.

REFERENCES

[1] Ohala, J. 1978. Production of Tone. In Fromkin, V. (ed.), *Tone: A Linguistic Survey*. New York: Academic Press.

[2] Schachter, P. and Fromkin, V. 1968. A phonology of Akan. *UCLA Working Papers in Phonetics*, 9.

[3] Breckenridge, J. 1977. The declination effect. *Journal of the Acoustical Society of America*, 60, S90.

[4] Pierrehumbert, J. 1979. The perception of fundamental frequency declination. *Journal of the Acoustical Society of America*, 66(2), 363-369.

[5] Woo, N. 1969. Prosody and phonology. Ph.D. dissertation, MIT.

[6] Wong, P.C.M. and Diehl, R.L. 1998. Effects of speaking fundamental frequency on the normalization of Cantonese level tones. *Journal of the Acoustical Society of America*, 104(3), 1834.

[7] Chao, Y.-R. 1947. *Cantonese Primer*. Cambridge: Harvard University Press.

	/ha6/	/yat1/	/go3/	/zi6/	/hai6/	/si3/
Speaker						
1	117	133	118	109	97	106
2	109	147	115	102	102	110
3	120	156	119	105	101	109
4	112	127	109	99	100	104
5	169	223	189	159	154	159
6	159	200	169	149	141	159
7	130	172	135	118	116	133

Table 1. F0 of each syllable in the original sentences spoken by the speakers in hertz (Hz)

	Tone 1	Tone 3	Tone 6
Stimulus Sentence			
F1S3	18.48	***68.11***	11.76
F1S6	16.31	14.26	***68.03***
F3S1	***82.65***	13.12	3.8
F3S6	7.29	11.53	***79.06***
F6S1	***77.72***	18.25	4.03
F6S3	.51	***83.59***	14.88

Table 2. Listeners' identification (%). The highest response rate for each stimulus sentence type is underlined.