Distortion in Tone Production due to the Lombard Effect

Giang Le[1], *Chilin Shih*[1,2], *Yan Tang*[1,3]

[1]Department of Linguistics, University of Illinois Urbana-Champaign, USA
[2]Department of East Asian Languages and Cultures, University of Illinois Urbana-Champaign, USA
[3]Beckman Institute for Advanced Science and Technology, University of Illinois Urbana-Champaign, USA

gianghl2@illinois.edu, cls@illinois.edu, yty@illinois.edu

Abstract

This study investigates the Lombard effect triggered at different noise levels on the F0 contours of tones in Northern Vietnamese. We hypothesize that due to the human talkers' limited capacity for hyper-articulation, the dynamic range of the F0 contours may be decreased, potentially resulting in a reduced distinction between lexical tones. For example, for rising tones, the slopes of the overall F0 contours might be flattened at higher noise levels if the demand for hyper-articulation forces elevation of the F0 contours to an unsustainable level. Acoustic analyses of speech produced in quiet and two noise levels confirmed that the F0 contours are raised across all lexical tones and talkers in noise and the dynamic range of the F0 contours decreased for tone C2. The results broadly supported the hypotheses that hyper-articulation due to the Lombard effect may cause tone distortion. The present findings are promising and invite further research into whether tone confusion is experienced by human listeners.

Index Terms: Lombard speech, Northern Vietnamese, hyper-articulation, lexical tones, speech in noise, speech production

1. Introduction

In adverse listening conditions, human talkers spontaneously adapt the way they speak to maximize the success of communication due to the *Lombard effect*. Speech produced under the Lombard effect, known as Lombard speech, has been observed to have characteristics such as increased F0 ([1], [2], [3], [4],), increased F1 ([1], [3], [4]), increased intensity ([1], [2], [5], [3], [6]), increased duration of more sonorous segments such as vowels ([1], [5], [3]), reduced rate of speaking [5], flattened spectral tilt [3], and decreased open quotients in the electroglottographical signal [3]. Different noise types and noise levels have been found to induce different changes to speech in noise. For instance, speech-shaped noise and babble noise induced less energy increase than other types of noise [1]. At lower noise levels, an increase in the noise barrier would give rise to a stronger response and at higher noise levels, the same increase in the noise barrier would induce a less drastic change in the response [5]. Some of the changes lead to better speech intelligibility even for non-native speakers [7].

While Lombard speech has been extensively studied, not much research has been conducted for tonal languages. At a suprasegmental level, tone production in Cantonese has also been found to be influenced by the Lombard effect and produced with a higher F0 [8]. F0 and F0 contours in Thai lexical tones are also seen to be emphasized in noise [9]. With the elevation of the overall F0 floor, one aspect less clear is whether the dynamic range of F0 contour would be reduced when the Lombard effect and hyper-articulation reach the ceiling, especially for tones realized by F0 fluctuations, such as rising and falling-rising tones. If this indeed is the case, certain tone pairs' F0 contours might become similar, leading to confusion between tones under a strong Lombard effect. This study investigates the Lombard effect of different noise levels on the tone articulations in Northern Vietnamese. Given a logarithmic model of the relationship between noise level and response [5], we hypothesize that although exaggeration of F0 contours in noise would be expected, the Lombard effect as manifested in raised F0 contours would be narrowed at higher noise levels. We present a brief overview of Northern Vietnamese tones below before reporting our speech production experiment's results and discussion thereof.

2. Background

Vietnamese is a tone language where a syllable could carry different F0 patterns, signifying semantic contrasts. In the standard Northern variety, a syllable could theoretically bear six or eight tones [10]. Traditional analyses consider six tones to be phonemic while the remaining two tones are allophonic, which are checked tones and only occur in closed syllables ending in voiceless stops.

The acoustic correlates of Vietnamese tones are F0 indicating pitch movement and pitch height, duration, intensity, and voice quality ([11], [12]). Other tone correlates include pitch range and the beginning and ending points of pitch movement. Lexical tones in the Northern variety are characterized by both varying F0 contours and changes in voice quality, whereby breathiness has been found to accompany the low-falling tone and creakiness has been found to accompany the rising and mid-falling tones. All eight phonetic tones in Vietnamese can be described as follows.

Tone A1 is a *level* tone spoken with a modal voice.

Tone A2 is a low to *mid-falling* tone usually spoken in a modal voice but could also be spoken with a lax or breathy voice [11].

Tone B1 is a *mid-rising* tone spoken with a modal voice.

Tone B2 is a *mid-falling* tone with strong glottalization at the end, or mid-falling with creakiness.

Tone C1 is also a *falling* tone, with a similar F0 contour as tone A2, but with slight laryngealization at the end [11]. Some speakers realize this tone with a mid-falling-rising contour, similar to the contour of C2.

Tone C2 is a *rising* tone with a glottal interrupt in its first half, also known as mid-rising with creakiness.

Tone D1 is a *rising* tone with a much higher F0 than tone B1.

Tone D2 has a *low-falling* F0 contour, but is not glottalized.

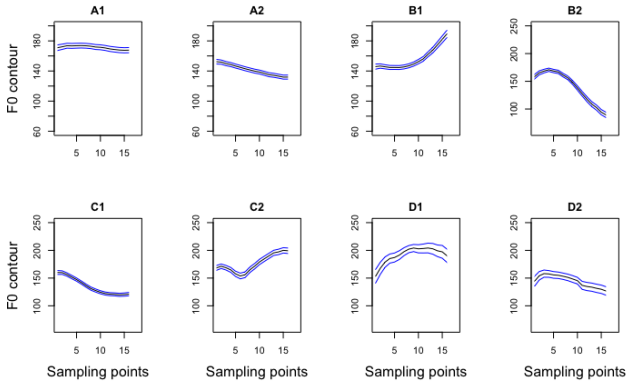

Figure 1: *Averaged F0 contours of Northern Vietnamese phonetic tones with 95% confidence intervals*

Figure 1 illustrates the average F0 contours and 95% confidence intervals of the six tones across five native Northern Vietnamese talkers.

3. Methods

3.1. Participants

Five native speakers of the Northern Vietnamese variety (three males, two females) between ages 19 and 26 were recruited to participate in this study. A hearing screening showed that all the participants had a normal hearing level.

3.2. Materials

Seventy-eight (78) tokens in vowel, consonant-vowel and consonant-vowel-consonant formats were created for this experiment. The same set of stimuli was embedded in a carrier sentence "*Toi noi cho ban nghe X bay gio* ('I say X to you now')" and presented to the participants, where X was one of the original tokens. All stimuli contained one of the three corner vowels /e/, /a:/ or /u/. The reason why /i/ was not chosen is because of higher incidences of vulgar words when combined with certain consonants in the stimuli, which could trigger the participants to react in an unexpected manner. The onset consonants belonged to the alveolar stop series. The alveolar stop series was chosen because it has three-way contrasts, the maximum number of contrasts in the language, as opposed to the limited contrasts in the bilabial and velar stop series. All the syllables were combined with all possible tones. Not all the tokens are possible meaningful words in Vietnamese but the combination of the tones on them is phonotactically possible.

3.3. Procedure

The speech production experiment took place in a noise-attenuating audio booth. To elicit Lombard speech due to different levels of the Lombard effect, three recording sessions were conducted. As a control condition, the first session was done in a quiet environment. During the other two sessions, participants were presented the speech-shaped noise at 78 and 90 dB SPL over a pair of open-back headphones while speaking. The speech-shaped noise was generated to have the long-term average spectrum of the Northern Vietnamese variety, in order to provide a comprehensive masking effect across frequencies.

During the experiment, the tokens were displayed on a computer screen, managed by a Javascript-based web platform. The stimuli were displayed to the participants in groups of the same syllable base, but the tone ordering was randomized and the order by which the syllable bases were presented was also randomized. The participants were instructed to utter the token displayed on the screen. The participants could control how fast they move through the stimuli via a mouse click. They had a practice session before the experiment started and nine self-monitored breaks during the experiment.

3.4. Data processing

As the tone-bearing units in Northern Vietnamese is vowels, vowels were manually segmented and labelled in Praat [13] for all recorded files. All audio files were double-checked semi-automatically against a correct tokens list, leading to a total of 4,732 tokens being labelled and analyzed. The F0 contours of these vowels were then extracted using a Matlab implementation of an F0 contour extraction based on glottal inverse filtering and autocorrelation method [14], followed by downsampling to 22 points by linear interpolation. After excluding six F0 values at the two ends of the contours, the F0 range was calculated by taking the difference between the highest and lowest F0 values of each F0 contour, while the slope of the F0 contours was estimated by fitting a regression line through the F0 values. The mean F0 was calculated by taking the average of the extracted F0 values after downsampling and exclusion.

4. Results

4.1. F0 range

A three-way ANOVA found significant main effects of noise, context, and tone on the F0 range [$\forall p < .001$]. Two- and three-way interaction effects were also found. The vowel predictor was initially included in the model, but because of its insignificant main effect, this predictor was excluded in the analysis. Post-hoc Tukey multiple comparisons with Bonferroni correction for the three-way ANOVA confirmed significant increases in the mean F0 range across tones and contexts from "quiet" to "78-dB SPL" by 2.50 Hz [$p < .05$], and further from "78-dB SPL" to "90-dB SPL" by 2.86 Hz [$p < .05$]. Tokens elicited in isolation had a higher mean F0 range compared to tokens elicited in carrier sentences [$p < .001$], when controlling for tone and noise.

Controlling for context, the noise and tone predictors interacted for tone B1, C1, and C2 and no significant interactions were found for tone B2, modal tones A1, A2 and checked tones D1, D2. Specifically, for tone B1, the mean F0 range increased from "quiet" to "78-dB SPL" by 15.23 Hz [$p < .001$] and from "quiet" to "90-dB SPL" by 17.08 Hz [$p < .001$] while no difference was found between "78-dB SPL" and "90-dB SPL". A similar pattern was found for tone C1, where the mean F0 range increased from "quiet" to "78-dB SPL" by 10.49 Hz [$p < .05$] and from "quiet" to "90-dB SPL" by 19.29 Hz [$p < .001$] while no difference was found between "78-dB SPL" and "90-dB SPL". For tone C2, a very different pattern emerged. The mean F0 range *decreased* from "quiet" to "78-dB SPL" by 24.53 Hz [$p < .001$] and also *decreased* from "quiet" to "90-dB SPL" by 26.51 Hz [$p < .001$] while no difference was found between "78-dB SPL" and "90-dB SPL" conditions. Fig. 2 illustrates the described interaction effects.

4.2. F0 slope

A three-way ANOVA found significant main effects of noise, context, and tone on the F0 range [$\forall p < .001$]. Two- and three-

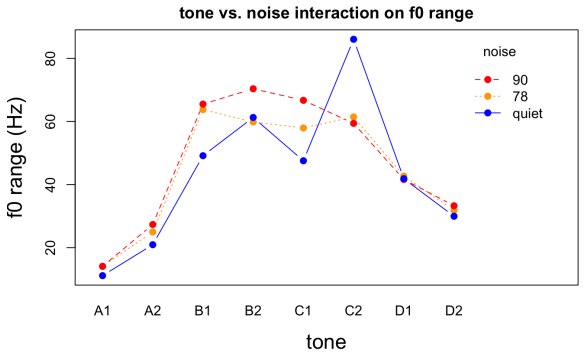

Figure 2: *Interaction between tone and noise on the F0 range*

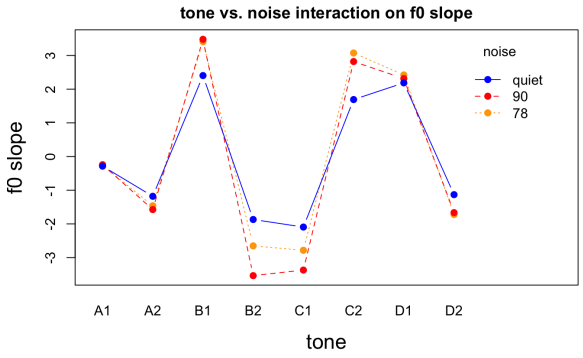

Figure 3: *Interaction between tone and noise on the F0 slope*

way interaction effects were also found [$\forall p < .001$]. The vowel predictor was initially included in the model, but because of its insignificant main effect, this predictor was excluded in the analysis.

Post-hoc Tukey multiple comparisons with Bonferroni correction showed that the absolute value of the F0 slope across tones and contexts increased significantly going from "quiet" to "78-dB SPL" by 0.43 [$p < .01$] and further from "78-dB SPL" to "90-dB SPL" by 0.60 [$p < .01$]. Tokens elicited in isolation had a higher F0 slope compared to tokens elicited in carrier sentences [$p < .001$], when controlling for tone and noise.

Controlling for context, the noise and tone predictors interacted for B1, B2, and C1 but not for the modal tones A1 and A2, the checked tones D1 and D2, and the creaky tone C2. Specifically, for tone B1, the absolute value of the mean F0 slope increased from "quiet" to "78-dB SPL" by 1.10 [$p < .001$] and from "quiet" to "90-dB SPL" by 1.31 Hz [$p < .001$] while no difference was found between "78-dB SPL" and "90-dB SPL". A similar pattern was found for tone C1, the absolute value of the mean F0 slope increased from "quiet" to "78-dB SPL" by 0.76 [$p < .001$] and from "quiet" to "90-dB SPL" by 1.28 [$p < .001$] while no difference was found between "78-dB SPL" and "90-dB SPL". For tone B2, the absolute value of the mean F0 slope increased from "quiet" to "90-dB SPL" by 1.04 [$p < .001$]. No significant difference was found between the other conditions.

4.3. Mean F0

A four-way ANOVA found significant main effects of noise, vowel, tone, and context on the mean F0 [$\forall p < .001$]. Two-way interaction effects were also found for noise and tone [$p < .001$], noise and vowel [$p < .05$], noise and context [$p < .05$], and tone and context, [$p < .001$].

Post-hoc Tukey multiple comparisons with Bonferroni correction confirmed previous findings that the mean F0 increased in noisy conditions; specifically, the mean F0 increased from "quiet" to "78-dB SPL" by 17.17 Hz [$p < .001$] and further from "78-dB SPL' to "90-dB SPL" by 27.00 Hz [$p < .001$]. Tokens elicited in isolation had a higher mean F0 compared to tokens elicited in carrier sentences [$p < .001$], when controlling for tone, noise, and vowel.

Controlling for vowel and context, it was found that the mean F0 increased significantly in noise compared to "quiet" for tones A1, A2, and B1 but no significant difference was detected between the two noisy conditions. Tone B2 did not show a significant mean F0 change from "quiet" to noisy conditions. Tone C1 showed an increased in mean F0 from "quiet" to "90-dB SPL" by 17.46 Hz [$p < .001$] but no difference between the other conditions. Tone C2 patterned the same as the general pattern, the mean F0 increased from "quiet" to "78-dB SPL" by 35.59 Hz [$p < .001$] and further from "78-dB SPL' to "90-dB SPL" by 17.90 Hz [$p < .001$]. Both tones D1 and D2 showed a significant increase in the mean F0 from "quiet" to "90-dB SPL" [$p < .001$].

Controlling for tone and context, the vowel pairwise comparison for the interaction effects showed the same pattern as the general pattern, except for vowel /o/. This is likely due to the smaller number of data points available for vowel /o/ in the elicited speech.

4.4. F0 contours

The F0 contours in three recording conditions for tones A2, B1, B2, C1, C2 were plotted in figures 4, 5, 6, 7, 8, respectively. These figures have been drawn for the average F0 across single tokens for all speakers at 16 sampling points. A more robust approach would be drawing a 95% confidence interval bands for each contour; for clarity of illustration, the average F0 contours are likely to suffice, and so they are adopted here. Due to the lack of statistically significant results in tones D1 and D2, the F0 contours of those two tones were not plotted. Because of similar result patterns between the modal tones A1 and A2, only the F0 contour for tone A2 is presented here.

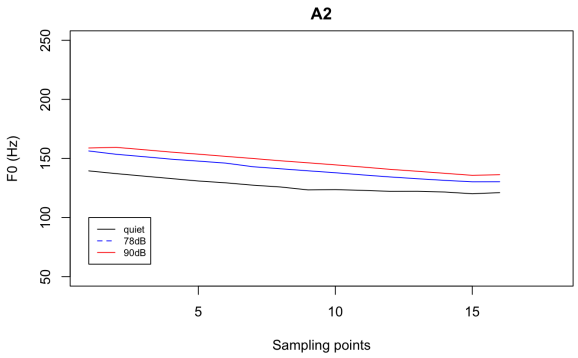

Figure 4: *F0 contours for single tokens on tone A2*

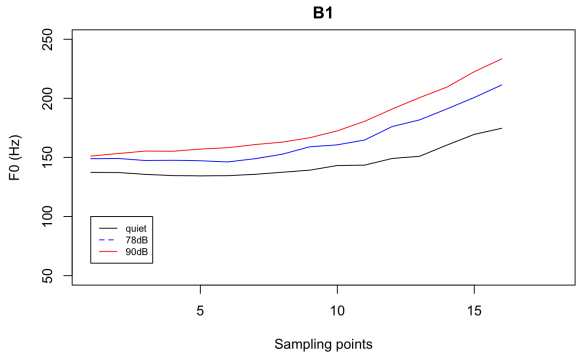

Figure 5: *F0 contours for single tokens on tone B1*

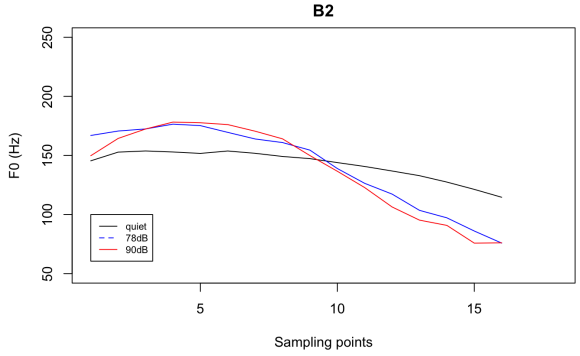

Figure 6: *F0 contours for single tokens on tone B2*

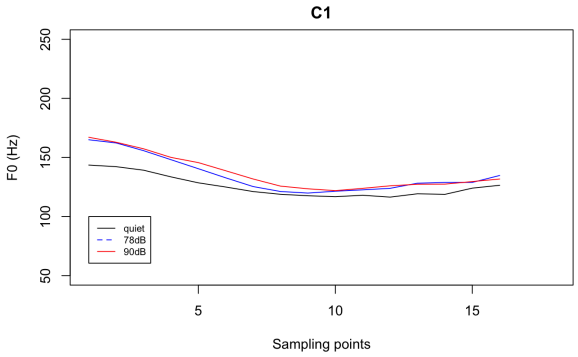

Figure 7: *F0 contours for single tokens on tone C1*

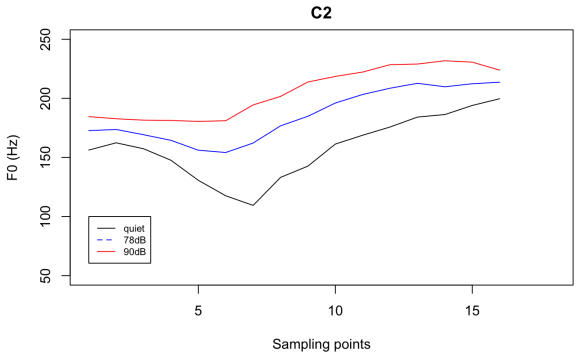

Figure 8: *F0 contours for single tokens on tone C2*

5. Discussion

The F0 range result showed that while in general, the F0 range increased in noise, and this increase became greater at higher noise level ("78-dB SPL" vs. "90-dB SPL"), this pattern was reversed for tone C2. The increase in the F0 range in noise, in particular for a rising tone such as tone B1, seemed to be consistent with general observations in the literature that F0 contours became exaggerated in noise [9]. While tone C2 still experienced an elevated mean F0 in noisy conditions, its F0 range narrowed in noise. Tone C2 is accompanied with creaky phonation in the middle of the segment. The finding that the F0 range of this tone actually decreased in noise provided evidence that hyper-articulation due to the Lombard effect could exhibit different effects in tones with different qualities, supporting our hypothesis that the Lombard effect could be lessened at high noise levels due to limits of human capacity for hyper-articulation. The fact that the F0 range of tone C2 decreased from "quiet" to "78-dB SPL" suggested that the threshold for the reduction in the Lombard effect for tone C2 might be even lower than the tested noise level, or that the mechanism of hyper-articulation for this tone is very different from that of the other tones.

The F0 slope finding, together with the F0 range finding confirmed that F0 contour became exaggerated for a modal, rising tone such as tone B1. However, it is not only in the rising tone where the F0 contour became exaggerated, for tone B2, a falling tone, also exhibited a higher absolute value of its slope in "90-dB SPL" compared to "quiet". Tone C1, also a falling tone, exhibited an increase in the absolute of the F0 slope in noisy condition also. These two findings showed that hyper-articulation also affected falling tones in Northern Vietnamese. It is worth noting that an estimated F0 slope value cannot capture a complex F0 contour such as that of tone C2. The lack of a significant difference in tone C2's F0 slope across conditions could be a result of this.

6. Conclusion

An acoustic analysis of speech produced in quiet, 78 dB and 90 dB noise presents evidence for exaggeration of F0 contours and provides additional evidence supporting hyper-articulation due to the Lombard effect. A reduced Lombard effect at two noise levels, as measured by the range of F0 values, was detected in tone C2. A laryngographic study examining how glottal vibration patterns change in noise for tones with different phonation types would help provide more evidence for articulatory differences in hyper-articulation of tones in Northern Vietnamese. Overall, tone distortion induced by the Lombard effect suggests potential perception research into whether tone confusion is experienced by human listeners.

7. References

[1] J.-C. Junqua, "The lombard reflex and its role on human listeners and automatic speech recognizers." *Journal of the Acoustical Society of America*, vol. 93, no. 1, pp. 510 – 524, 1993.

[2] W. Van Summers, D. Pisoni, R. Bernacki, R. Pedlow, and M. Stokes, "Effects of noise on speech production: Acoustic and perceptual analyses." *Journal of the Acoustical Society of America*, vol. 84, no. 3, pp. 917–928, 1988.

[3] M. Garnier, L. Bailly, M. Dohen, P. Welby, and H. Loevenbruck, "An acoustic and articulatory study of lombard speech: global effects on the utterance," in *INTERSPEECH*, 2006.

[4] Y. Lu and M. Cooke, "Speech production modifications produced in the presence of low-pass and high-pass filtered noise," *The Journal of the Acoustical Society of America*,

vol. 126, no. 3, pp. 1495–1499, 2009. [Online]. Available: https://doi.org/10.1121/1.3179668

[5] T. D. Hanley and M. D. Steer, "Effect of level of distracting noise upon speaking rate, duration and intensity." *Journal of Speech & Hearing Disorders*, vol. 14, pp. 363 – 368, 1949.

[6] S. Kim, "Durational characteristics of korean lombard speech," in *INTERSPEECH*, 2005.

[7] M. Cooke and M. L. G. Lecumberri, "The intelligibility of lombard speech for non-native listeners." *The Journal of the Acoustical Society of America*, vol. 132 2, pp. 1120–9, 2012.

[8] Y. Zhao and D. Jurafsky, "The effect of lexical frequency and lombard reflex on tone hyperarticulation," *J. Phonetics*, vol. 37, no. 2, pp. 231–247, 2009. [Online]. Available: https://doi.org/10.1016/j.wocn.2009.03.002

[9] B. Kasisopa, V. Attina, and D. Burnham, "The lombard effect with thai lexical tones: an acoustic analysis of articulatory modifications in noise," in *INTERSPEECH*, 2014.

[10] J. Kirby, "Vietnamese (hanoi vietnamese)," *Journal of the International Phonetic Association*, vol. 41, no. 3, pp. 381–392, 2011.

[11] V. L. Nguyen and J. Edmondson, "Tones and voice quality in modern northern vietnamese: Instrumental case studies," *Mon-Khmer Studies*, vol. 28, pp. 1–18, 1997.

[12] H. Pham, "Vietnamese tone: Tone is not pitch," Ph.D. dissertation, University of Toronto, Ontario, 2000.

[13] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 5.1.13)," 2009. [Online]. Available: http://www.praat.org

[14] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku, "Hmm-based speech synthesis utilizing glottal inverse filtering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 153–165, 2011.