A Laryngographic Study on the Voice Quality of Northern Vietnamese Tones under the Lombard Effect

Giang Le\textsuperscript{1}, Chilin Shih\textsuperscript{1,2}, Yan Tang\textsuperscript{1,3}

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

\{gianghl2, cls, yty\}@illinois.edu

Abstract

While acoustic changes in Lombard speech are well-documented, articulatory changes are less well-studied. This study presents recent findings of Northern Vietnamese tone production in noise from an articulatory perspective, based on laryngographic evidence of voice quality. From the time domain, Lx signals, jitter and shimmer were found to decrease at higher noise levels while the mean F0 and harmonics-to-noise ratio increased. Prior hypotheses predicted that creaky tones might demonstrate different degrees of glottal vibratory changes compared to the modal tones. Interaction effects were detected between tone and noise levels, and a detailed examination showed that differences in Lombard effects were even found within creaky tones themselves. Furthermore, the laryngographic spectral tilt flattened at increasingly higher noise levels in the frequency domain. These findings point towards reduced creakiness in the voice quality of speech produced in noise, and provide evidence that hyper-articulation starts at the sound source.

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

1. Introduction

Lombard speech has been investigated in various research studies for its characteristics, typically involving compensatory strategies by the talker to overcome an adverse listening environment. While acoustic changes 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], and flattened spectral tilt [3] are widely reported, findings of articulatory changes are fewer in number, especially for tonal languages such as Vietnamese and Mandarin Chinese. Garnier et al. examined labial parameters from video data and reported that hyper-articulation could be seen in lip aperture and spreading rather than lip pinching from French speakers [3]. The same study also reported decreased open quotients in laryngographic (Lx) signals of speech in noise.

This study presents an analysis of the Lx signals of Northern Vietnamese. We compared the voice quality of sonorous segments bearing modal and creaky tones in Northern Vietnamese when speech was produced under two noise levels, which were set to trigger the Lombard effect in different degrees. In the time domain of the Lx signals, jitter, shimmer, F0 and harmonics-to-noise ratio (HNR) were measured; in the frequency domain, we further investigated the spectral tilt, which is suggested to be strongly correlated with speech intelligibility in noise by previous studies on acoustic signals (e.g. [7]). We hypothesized that hyper-articulation due to the Lombard effect might reduce the discrepancies between lexical tones essentially distinguished by phonation types (e.g. modal vs creaky), and that hyper-articulation could start at the stage of sound source generation, i.e. vibration of the vocal folds, before a specific sound was formed in the vocal tract.

2. Background

2.1. Northern Vietnamese tones

Vietnamese is a tone language where a syllable can carry different pitch patterns, signifying semantic contrasts. In the standard Northern variety, a syllable can theoretically bear one of six to eight tones [8]. Traditional analyses consider six tones to be phonemic, while the remaining two tones are allophonic with limited distribution, which are checked tones and only occur in closed syllables ending in voiceless stops. All eight phonetic tones in Northern Vietnamese can be described as follows.

- **Tone A1**: 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 [9].
- **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 [9]. Some speakers realize this tone with a mid-falling-rising contour, similar to the contour of C2.
- **Tone C2**: is a rising tone with 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.

The average F0 contours of the tones described above across six native Northern Vietnamese talkers participated in this study and 95% confidence intervals are presented in Fig. 1.

2.2. Creaky phonation versus modal phonation

Creaky phonation occurs when the vocal folds are closely drawn together or tightly adducted, leaving only a narrow gap for voicing. Creaky phonation is associated with irregular vocal pulses. In Northern Vietnamese, tones B2, C1, and C2 are accompanied by glottalization in varying degrees either in the middle of the segment (tone C2) or at the end of the segment (tones B2...
3. Methods

3.1. Participants

Six native speakers of the Northern Vietnamese variety (three males, three females) between the ages of 19 and 34 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 tokens in vowel, consonant-vowel and consonant-vowel-consonant formats were created for this study. These stimuli were embedded in a carrier sentence “Tôi nói cho bạn nghe X bay gió” (I say X to you now), where X was one of the target tokens. All the stimuli contained one of the three corner vowels /e/, /a:/ or /u/. The vowel /i/ was not chosen because of the higher incidences of vulgar words when combined with certain consonants in the stimuli, which could trigger the participants to react in an unexpected manner. /o/ was also elicited as a back-up for /u/ for one syllable template due to the same concern. 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

Three recording sessions took place in a sound-treated audio booth. Apart from the control condition, where participants uttered the stimuli in quiet (“quiet”), speech-shaped noise was presented to participants at 78 dB SPL (“78-dB SPL”) and 90 dB SPL (“90-dB SPL”) over a pair of open-back headphones in the other two sessions to elicit Lombard effect of different degrees. The speech-shaped noise was generated to have a representative long-term average spectrum of the Northern Vietnamese variety; it resulted in similar energetic masking on the target speech across frequencies. In all three sessions, the Lx signals were recorded using a laryngograph. Both acoustic and Lx signals were saved as WAV files sampled at a rate of 44.1 kHz.

The participants were instructed to speak the stimuli displayed on a computer screen aloud twice each. 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 could control how fast they moved through the stimuli via a mouse click.

3.4. Data processing and analysis

Vowels, the tone-bearing unit in Northern Vietnamese, were manually labelled and segmented for the recorded files, yielding 4,820 tokens. For the labelled intervals, the laryngographic signals and features related to voice quality were extracted using Praat [12].

4. Results

4.1. Time-domain analysis

4.1.1. Jitter

A three-way ANOVA found significant main effects of noise, vowel and tone on jitter [\(\alpha < .001\)]. Post-hoc Tukey multiple comparisons with Bonferroni correction confirmed significant decreases in mean jitter across vowels and tones from “quiet” to “78-dB SPL” by 0.73%, and further from “78-dB SPL” to “90-
dB SPL” by -1.57% [∀p < .001]. Overall, jitter was lower for the higher vowels /e/ and /u/ compared to /a:/ [p < 0.05], and was lower for /u/ compared to /e/. This finding is in line with [13] – both jitter and shimmer were the smallest for /u/, intermediate for [i] and greatest for [a]. For tones, A1 and A2 had the lowest jitter compared to other vowels [p < 0.05]. Among the glottalized tones, B2 had higher jitter than C2 [p < 0.05] while no difference was detected between C1 and C2.

In noise, jitter decreased more for the two glottalized tones C1 and C2 than the other tones, but it was not the case for the creaky tone B2. As expected, jitter values were higher for creaky tones, but the jitter difference between C1/C2 and modal tones was the greatest in “quiet” and smallest in “90-dB SPL”. This interaction is seen most clearly in Fig. 3 where tones C1 and C2 experience a greater drop in jitter. B2 does not follow this pattern and retains a relatively strong level of creaky phonation in noise, especially in “78-dB SPL”.

4.1.2. Shimmer

Similar to jitter, a three-way ANOVA found that the same main effects (noise, vowel and tone) on shimmer were significant [∀p < .001]. No two- or three-way interactions were found except between noise and tone [p < 0.05]. Post-hoc pairwise comparisons showed that shimmer decreased significantly from “quiet” to “78-dB SPL” by 2.74%, and further from “78 dB SPL” to “90-dB SPL” by 4.90% [∀p < .001]. In general, shimmer was lower for the higher vowels /e/ and /u/ compared to /a:/ [p < 0.05], and was lower for /u/ compared to /e/ [p < 0.05]. Creaky tones B2, C1 and C2 had higher shimmer values than those of other tones [p < 0.01] but no difference in shimmer was found among them.

Fig. 4 shows that unlike in the case of jitter, compared to modal tones, creaky tones B2 and C1 experienced a lesser degree of shimmer decrease in “90-dB SPL” compared to in “quiet”. However, tones C2 and D1 showed a more significant shimmer decrease in “90-dB SPL” than other tones.

4.1.3. Mean HNR

For HNR, a three-way ANOVA with the same main effects reveals that HNR significantly varied across different noises, vowels and tones [p < 0.001]. No interaction effects were detected except between noise and tone [p < 0.05]. Post-hoc pairwise comparisons confirmed that the mean HNR across vowels increased as the noise level increased: 2.1 dB from “quiet” to “78-dB SPL”, and further 4.4 dB from “78-dB SPL” to “90-dB SPL” [∀p < .001]. High vowels /e/ and /u/ had higher mean HNR than /a:/ [p < 0.001], and /u/ had a higher mean HNR than /e/ [p < 0.05]. The main effect in tone was as expected, modal tones A1 and A2 were found to have a significantly higher mean HNR compared to those in other tones [p < 0.01].

Fig. 5 further illustrates that mean HNR increased across tones from “quiet” to higher noise levels. However, all creaky tones (B2, C1, C2) and allophonic tones D1 and D2 experienced a lesser degree of mean HNR increase than the modal tones A1 and A2. This was also the case for tone B2 when its mean HNR change from “quiet” to “78-dB SPL” was less than that of the modal tones, but this was not the case for C1 and C2, which had about the same amount of mean HNR increase in “78-dB SPL” compared to the modal tones A1 and A2.

4.2. Spectral tilt

Discrete Fourier Transform was performed on the Lx waveforms of all sonorous segments to obtain the spectrum of each
segment. Due to the periodicity of the Lx signal on vowels being regular in general, the long-term spectrum of a segment was calculated from its entire Lx signal. The spectral tilt was then measured as the slope of the decrease in magnitude from 0.1 kHz to 1.4 kHz. Fig. 7 shows the spectral tilt of the Lx signals in the three noise conditions. A three-way ANOVA confirmed the visual impression that the main effects of noise, tone and vowel \( \forall p < 0.01 \) were significant. Post-hoc pairwise comparisons suggested that spectral tilts flattened significantly when the noise level increased from “quiet” to “78-dB SPL” \( [p < 0.001] \), and then to “90-dB SPL” \( [p < 0.001] \).

In conclusion, it was observed that in the time domain, periodicity of the Lx signals increased via decreased jitter and shimmer as well as increased mean HNR in noise. While tone B2 showed resistance to increased glottal regularity, tone C2 showed a greater degree of shimmer decrease in “90-dB SPL”. This could be explained by the relatively more challenging task of strong glottalization in the middle of a sonorous segment in order to articulate tone C2 in noise. Due to the location of the creaky phonation in C2, differentiating the creaky part and neighboring modal parts requires a distinctive switch between on and off of glottalization within a short period. In noise, hyper-articulation induced by the Lombard effect could limit human speakers’ capacity to maintain such quick phonation alternation, hence a greater degree of reduced glottal vibratory irregularity of C2 in noise. The relatively more challenging phonation setting of C2 also explains why in the Southern Vietnamese variety, this tone has merged with tone C1. Its distribution is also much more limited compared to other tones (313 types in single syllablemene compared to 1365 types of tone B1, the tone with the highest distribution [14]).

The mean HNR findings confirmed the different behavior of B2 compared to other tones. B2’s increase in mean HNR in noise was less than that of the modal tones, but for C1 and C2, the mean HNR increase was about the same as that in the modal tones. This was consistent with the observation that for B2 it was relatively easier to sustain creaky phonation at the end of segments, while for C2, it was harder to sustain creakiness because this phonation change occurred in the middle of the segment. Overall, a reduction in glottalization was found for all creaky tones. Consequently, the distinction between tones realized by different phonation types could be minimized, potentially leading to misperception of these tones, which needs to be confirmed by further perceptual experiments.

In the frequency domain of the Lx signals, significant flattening of the spectral tilt at increasingly higher noise levels was observed. When hyper-articulating in noise, the vocal folds were brought together at higher frequencies, as evident in raised F0 in Lombard speech. In the meantime, the increased activities appeared to be taking place not only between the two interactive vocal folds, but also in each individual organ. The boosted high-frequency components in the Lx signals reflected as the reduced spectral tilt suggested that hyper-articulation may also lead to increased oscillations by the vocal folds themselves. Previously, other studies discovered the flattening of the spectral tilt of the acoustic signals as a combined consequence of source generation and vocal tract filtering [3]. What was found in this study can provide evidence that the effect of hyper-articulation can be at as early as the stage of sound source generation. However, further investigation is required to isolate the effect of vocal tract filtering on boosting the high-frequency components of the produced sounds.

In conclusion, it was observed that in the time domain, periodicity of the Lx signals increased via decreased jitter and shimmer as well as increased mean HNR in noise. While tone B2 showed resistance to increased glottal regularity, tone C2 showed a greater degree of shimmer decrease in “90-dB SPL”. This could be explained by the relatively more challenging task of strong glottalization in the middle of a sonorous segment in order to articulate tone C2 in noise. Due to the location of the creaky phonation in C2, differentiating the creaky part and neighboring modal parts requires a distinctive switch between on and off of glottalization within a short period. In noise, hyper-articulation induced by the Lombard effect could limit human speakers’ capacity to maintain such quick phonation alternation, hence a greater degree of reduced glottal vibratory irregularity of C2 in noise. The relatively more challenging phonation setting of C2 also explains why in the Southern Vietnamese variety, this tone has merged with tone C1. Its distribution is also much more limited compared to other tones (313 types in single syllablemene compared to 1365 types of tone B1, the tone with the highest distribution [14]).
6. References


