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

In adverse listening conditions, human talkers spontaneously adapt the way they speak to maximize the success of communication, a reflex called *Lombard effect*.

This study investigates the Lombard effect of different noise levels on F0 contours of tones in Northern Vietnamese. We hypothesize that due to the human talkers' limited capacity for hyper-articulation, the dynamic range of F0 contour may be decreased, potentially resulting in reduced distinction between lexical tones. Furthermore, for rising tones, the slopes of the overall F0 contours might be narrowed at higher noise levels if the demand for hyper-articulation forces elevation of F0 contours to an unsustainable level. Consistent with previous research [1], nevertheless, we expect some exaggeration of F0 contours under the Lombard effect.

An acoustic analysis of speech produced in quiet and two noise levels confirms raised F0 contours across all lexical tones and talkers. While the present result is inconclusive about the decrease of F0 range in tones produced in 90 dB noise, changes in the F0 contours support 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: acoustics, acoustic phonetics, psychoacoustics, lombard speech, northern vietnamese, hyper-articulation, lexical tones, speech in noise, speech production

1. Introduction

Speech produced in noise has been investigated in various research studies for its characteristics, typically involving compensatory strategies by the speaker to overcome an adverse listening environment. Noise is loosely defined as sound that carries no useful information and at the same time intervenes with sounds that carry useful information. We know from previous research that the presence of background noise affects communication in various ways: from the transmitting side, background noise often prompts the talker to alter their speaking style in order to increase communication efficiency; from the receiving side, noise has a direct impact on the speech signal and thus complicates comprehension and perception of the listener. As a whole, noise and its effects present themselves as interesting objects of inquiry, for they present challenges to communication and a good understanding of the strategies used by communicators to overcome these challenges have implications for speech intelligibility and enhancement techniques.

The Lombard effect of speech production in noise has been extensively studied but not much research has been conducted for tonal languages. Among features cited in the literature about Lombard speech are increased F0 ([2], [3], [4], [5], [1], [6]), increased F1 ([2], [4], [5]), increased intensity ([2], [3], [7], [4], [8]), increased duration of more sonorous segments such as vowels ([2], [7], [4]), reduced rate of speaking [7], flattened spectral tilt [4], and decreased open quotients in the EGG signal [4].

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 [2]. 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 [7]. Given this logarithmic model of the relationship between noise level and response, we hypothesize that although exaggeration of F0 contours in noise would be expected, the Lombard effect 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 classified as a tone language where a syllable could carry different F0 (pitch) patterns, signifying semantic contrasts. In the standard Northern variety, a syllable could theoretically bear six or eight tones [9]. Traditional analyses consider six tones to be phonemic while the remaining two tones are allophonic, with a limited distribution. They are checked tones that 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 ([10], [11]). 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 ('ngang') is a level tone spoken with a modal voice.

Tone A2 ('huyen') is a low to mid-falling tone usually spoken in a modal voice but could also be spoken with a lax or breathy voice [10].

Tone B1 ('sac') is a mid-rising tone spoken with a modal voice.

Tone B2 ('nang') is a mid-falling tone with strong glottalization at the end, or mid-falling with creakiness.

Tone C1 ('hoi') is also a falling tone, with a similar F0

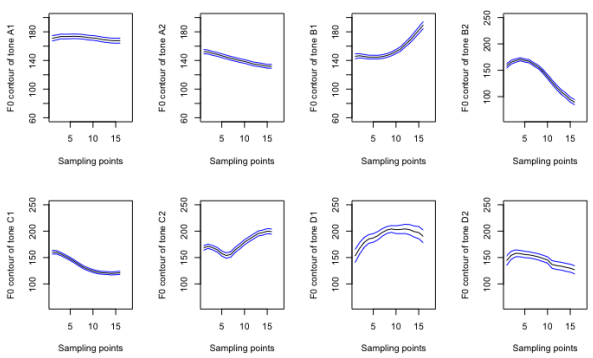

Figure 1: *Averaged F0 contours of Northern Vietnamese phonetic tones*

contour as tone A2, but with slight laryngealization at the end [10]. Some speakers realize this tone with a mid-falling-rising contour, similar to the contour of C2.

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

Tone D1 (checked 'sac') is a rising tone with a much higher F0 than tone B1.

Tone D2 (checked 'nang') has a low-falling F0 contour, but is not glottalized.

Average F0 contours of the tones just described across five Northern Vietnamese talkers and 95% confidence intervals are presented in Figure 1.

3. Methods

3.1. Participants

All study procedures were conducted with the approval of the Institutional Review Board at our institution. Five participants (three males, two females) were recruited to participate in this study. All of them are native speakers of the Northern Vietnamese variety and are between 19 to 26 years old. Each participant signed a consent form and completed a background questionnaire prior to taking part in the study and all of them received a compensation of 10 dollars per hour for their participation. All of them reported being from cities and provinces in the North of Vietnam (Hanoi (4), Phu Tho (1)). None reported any speech or hearing impairment. A hearing test was conducted at the beginning of the first recording session to verify that the participants had a normal hearing range. All participants were able to detect the pure tones at or lower than 10 Hz for the left ear and 15 Hz for the right ear, well within the normal threshold of 20 Hz.

3.2. Materials

Seventy-eight (78) stimuli were created for this experiment. The stimuli were words of V, CV, and CVC syllable shapes. The same set of stimuli was embedded in a carrier sentence *Toi noi cho ban nghe bay gio* ('I say X to you now') and presented to the participants, where X was one of the original stimuli. All the 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 stimuli are possible words in Vietnamese but the combination of the tones on them is phonotactically possible.

3.3. Procedure

Our general setup presented noise at 78 dB SPL and 90 dB SPL to participants over open back headphones in a noise-attenuating audiometric booth. The acoustic and laryngeal signals were recorded at the same time using Praat [12], however, the laryngeal data are not presented in this paper.

The participants were instructed to say aloud the stimuli displayed on the screen, two times each. The stimuli were displayed on a computer screen, hosted by the Javascript-based psycholinguistics experimental platform Ibex Farm. 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 move through the stimuli via a mouse click. They had two practice items before the experiment started and nine self-monitored breaks during the experiment.

Three recording sessions were conducted: one recording session done in a quiet environment; in follow-up recording session 1, the participants produced the stimuli while listening to a speech-shaped noise with the noise level calibrated at 78 dB SPL; in follow-up recording session 2, the participants produced the stimuli while listening to a speech-shaped noise with the noise level calibrated at 90 dB SPL. The noise maskers were generated with a representative spectrum of the Northern Vietnamese variety based on sample speech recordings collected from publicly available corpora. The generated speech-shaped white noise has a long-term average spectrum similar to the long-term average spectrum of the target language's speech corpus and therefore can provide equal masking for all frequencies in the language. This is important because unlike white noise, speech has higher energy at low frequencies and lower energy at high frequencies. Using pure white noise would not have a constant masking effect across all frequencies. The noise maskers were introduced to the participants via the application Audacity. Screen recordings were obtained for all sessions to cross check and confirm the accuracy of the annotated data, if necessary.

3.4. Data processing

As the tone-bearing unit in Northern Vietnamese is the vowel, vowels bearing tones were manually segmented and labelled in Praat [12] for all recorded files. All audio files were double checked semi-automatically against a correct tokens list to ensure that all tokens were properly segmented and labelled. The total number of tokens labelled and analyzed was 4,732 tokens.

The labelled intervals were extracted to individual .wav files and then the F0 contours of these vowels were extracted using a Matlab implementation of the algorithm used in the paper [13], based on glottal inverse filtering and autocorrelation. After that, the extracted vectors were downsampled to 22 points with linear interpolation.

The F0 range was calculated by simply taking the difference between the highest and lowest F0 values of each F0 vector, while the slope of the F0 contours was estimated by fitting a regression line through the extracted F0 values.

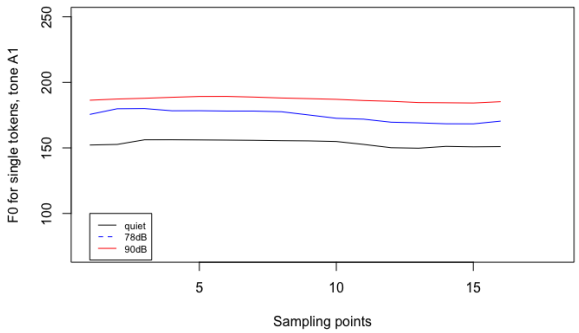

Figure 2: *F0 contours for single tokens on tone A1*

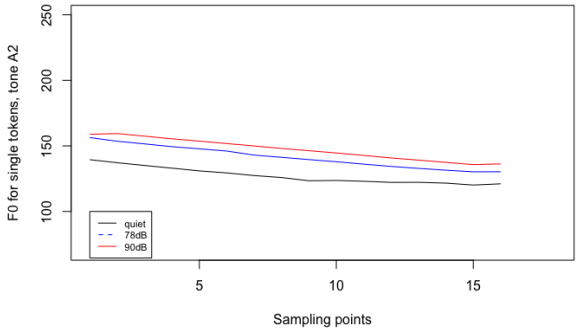

Figure 3: *F0 contours for single tokens on tone A2*

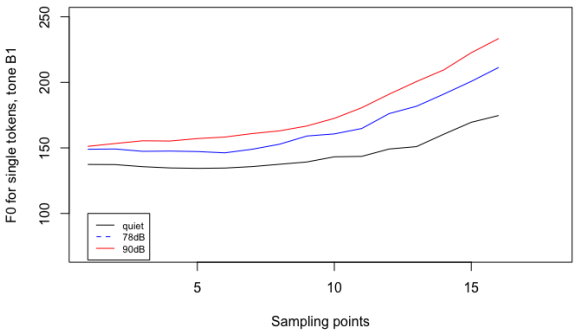

Figure 4: *F0 contours for single tokens on tone B1*

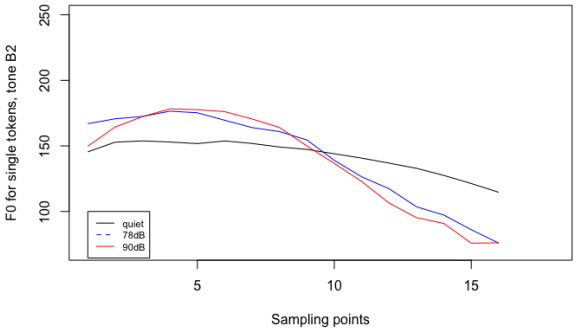

Figure 5: *F0 contours for single tokens on tone B2*

4. Results

4.1. Graphical Results

Looking at tokens uttered in isolation, F0 contours show elevation in noise generally, as demonstrated in the attached plots. For single tokens with tone A1, A2, and B1, F0 is elevated in 78 dB SPL noise level compared to the controlled condition, and again in 90 dB SPL noise level compared to 78 dB SPL noise. Figure 2, 3, and 4 illustrate this pattern.

For tone A2, when regression lines are fitted to F0 values, the absolute value of the slope of the regression lines for 90 dB SPL noise level (1.71) is slightly lower than that of the 78 dB SPL noise level (1.80). This suggests potential narrowing of the F0 range going from 78 dB SPL to 90 dB SPL, especially because this tone's contour is fairly linear.

The extent to which the contour rises is more exaggerated in noise level 78 and 90 for tone B1. When regression lines are fitted to F0 values, the slope of the regression lines for 90 dB SPL noise level (5.18) is higher than that of the 78 dB SPL noise level (4.02) and the slope of the regression lines for 78 dB SPL noise level is higher than that of the controlled condition (2.31).

For single tokens with tone B2 (mid-falling with creakiness), F0 starts off at higher levels in 78 dB SPL and 90 dB SPL noise levels compared to the controlled condition (see Figure 5) and fall off to lower frequency levels. Interestingly, the extent to which the contour falls is more exaggerated in noise level 78 and 90. When regression lines are fitted to F0 values, the absolute value of the slope of the regression lines for 90 dB SPL noise level (7.03) is higher than that of the 78 dB SPL noise level (6.84) and the absolute value of the slope of the regression lines for 78 dB SPL noise level is much higher than that of the controlled condition (2.26). This exaggeration of the contour suggests agreement with hyper-articulation in noise conditions.

For single tokens with tone C1 (falling tone with slight laryngealization), F0 starts off at higher levels in 78 dB SPL and 90 dB SPL noise levels compared to the controlled condition (see Figure 6) but fall off to similar levels. Consequently, the extent to which the contour falls is more exaggerated in noise level 78 and 90. When regression lines are fitted to F0 values, the absolute value of the slope of the regression lines for 90 dB SPL noise level (2.56) is higher than that of the 78 dB SPL noise level (2.27) and the absolute value of the slope of the regression lines for 78 dB SPL noise level is higher than that of the controlled condition (1.44). Similar observations about contour exaggeration in noise can be made for tone C2 (see Figure 7). This exaggeration of the contour suggests agreement with hyper-articulation in noise conditions, although the difference in the contours' slopes are not as pronounced as that in tone B2.

Single tokens with tone D1 show more jagged contour patterns (see Figure 8), but in general, F0 contours are more elevated in 78 and 90 dB SPL noise levels. The regression lines show different patterns from the other tones, however. Although the slope of the regression line for 78 dB SPL (4.27) is higher than that of the controlled condition, (1.01) the slope of the regression line for 90 dB SPL (2.51) is lower than that of 78 dB SPL.

Single tokens with tone D2 also show the same raised F0 patterns in noise conditions (see Figure 9). The slope of the

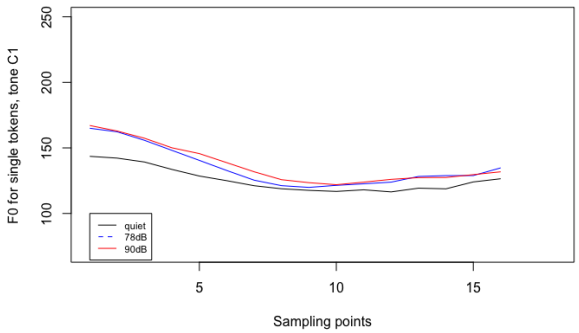

Figure 6: *F0 contours for single tokens on tone C1*

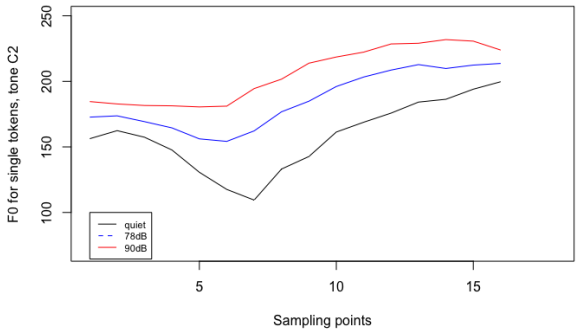

Figure 7: *F0 contours for single tokens on tone C2*

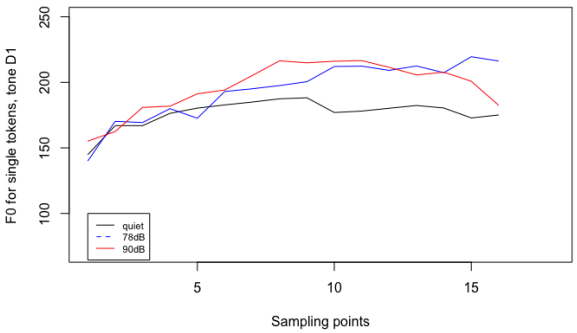

Figure 8: *F0 contours for single tokens on tone D1*

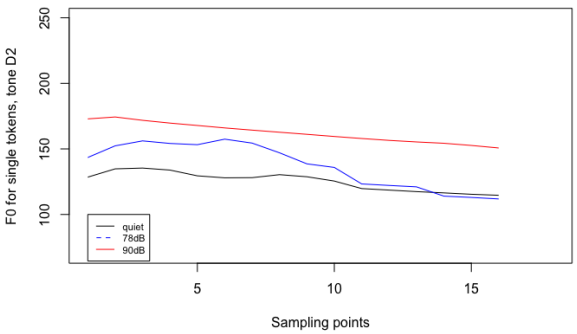

Figure 9: *F0 contours for single tokens on tone D2*

regression lines shows the same pattern as tone D1: narrower range of F0 contour in 90 dB SPL compared to 78 dB SPL even though the contour in 78 dB SPL shows an expanded range.

These figures have been drawn for 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.

4.2. Numerical Results

Slopes of F0 contours of single tokens and tokens in carriers can be seen in Table 1, in all experimental conditions. What comes to the fore is that in carrier context, which is arguably a more natural speaking condition, narrowing of F0 contours' slope in absolute value of the 90 dB SPL condition occurs in tone B1, C2, D1, and D2 whereas in the single context, this happens in tone A1, A2, D1, and D2. It seems that the checked tones are the most likely to experience narrowing of F0 slopes and all tones experiencing this effect are non-creaky tones, except for C2.

The range of F0 contour for all tones (single and carrier tokens) in all experimental conditions can be seen in Table 2. We can see that the result is mixed here. In single tokens, the tones experiencing a narrower F0 range in higher noise levels are A1, A2, C2, D1, and D2. In carrier contexts, the tokens experiencing this are C2 and D2. Similar to the effect on F0 slopes, the checked tones are the most likely to experience narrowing F0 range at 90 dB SPL and all tones experiencing this effect are non-creaky tones, except for C2.

5. Discussion

Examining F0 contours, their regression lines' coefficients and range confirms previous findings about raised F0 contours as the noise levels increase. However, the narrowing Lombard effect in the F0 slopes and range only partially borne out, and mostly we see this in non-creaky tones. With still a modest number of talkers who participated in this study, we can only say that reduced Lombard effect at higher noise levels is not always present but and it seems to be dependent on the individual tones. However, given that reduced effect did happen, we can expect certain tones to be more similar to others. For instance, the narrowing of F0 contour at 90 dB SPL for carrier B1 tone could potentially make it more similar to the level tone A1. Tone C2 in carrier contexts might be mistaken for tone C1 or A1 if the narrowing effect is strong enough and if there is increased regularity or reduced creakiness in the tone's articulation. This initial evidence for potentially tone merging has implications for follow-up perception studies that investigate the extent to which tone confusion happens.

We also note that the checked tones experience more narrowing Lombard effect at high noise levels. However, this result should be interpreted with care, for there are less instances of tones D1 and D2 in the collected data to begin with, and there could be co-articulation effect due to following coda consonant that we are not aware of. It is also interesting that even though B1 and D1 are both rising tones where D1 is B1's allophonic counterpart, plateauing of the Lombard effect is much visible in tone D1. What could explain this is because the F0 contour of tone D1 starts at a much higher frequency than tone B1 and

Table 1: *Slopes of F0 contours in different noise levels and contexts*

Tone	Context	Quiet	78 dB noise	90 dB noise
A1	single	-0.32	-0.81	-0.26
A1	carrier	-0.45	-0.22	-0.31
A2	single	-1.25	-1.80	-1.71
A2	carrier	-1.23	-1.36	-1.56
B1	single	2.31	4.02	5.18
B1	carrier	1.50	1.66	1.62
B2	single	-2.26	-6.84	-7.03
B2	carrier	-4.03	-5.99	-7.94
C1	single	-1.44	-2.27	-2.56
C1	carrier	-3.15	-4.02	-5.14
C2	single	3.60	4.00	4.15
C2	carrier	1.34	1.87	1.66
D1	single	1.01	4.27	2.51
D1	carrier	2.35	1.72	1.65
D2	single	-1.39	-3.24	-1.58
D2	carrier	-2.34	-1.94	-1.38

Table 2: *R0 range in different noise levels and contexts*

Tone	Context	Quiet	78 dB noise	90 dB noise
A1	single	6.45	11.66	5.01
A1	carrier	6.81	5.68	8.92
A2	single	19.34	26.09	23.69
A2	carrier	16.43	18.55	22.17
B1	single	40.30	65.13	82.24
B1	carrier	29.35	31.19	31.76
B2	single	39.16	100.65	102.43
B2	carrier	62.82	81.83	103.04
C1	single	27.13	45.11	45.15
C1	carrier	42.48	51.10	64.17
C2	single	90.27	59.46	51.35
C2	carrier	41.21	32.40	26.14
D1	single	43.39	79.55	61.42
D1	carrier	61.17	38.81	52.86
D2	single	20.77	45.60	23.66
D2	carrier	35.97	42.15	28.86

therefore it is much easier for tone D1 to reach a plateau in the Lombard effect. If this is the case, the starting point of F0 contours and their trajectories are both important factors to determine if a tone experiences a plateau in the Lombard effect or not.

It is also noteworthy that lexical frequency could have an impact in how F0 contours vary in noise. [6] studied how word frequency and noise affect the acoustic realizations of tones in Cantonese and remarked that low-frequency words with mid-range tones are produced with higher F0 than frequent words. In my production experiment, all tones are imposed on the syllable templates, and although they are phonotactically possible, not all of them are real words in the language variety under consideration. In the future, taking into account the word frequency might help elucidate our current findings further.

6. Conclusion

An acoustic analysis of speech produced in quiet, 78 dB and 90 dB noise presents exaggeration of F0 contours and provides additional evidence supporting hyper-articulation due to the Lombard effect. Reduced Lombard effect for 90 dB SPL was detected in non-creaky tones. Tone distortion suggests potential perception research into whether tone confusion is experienced by human listeners.

7. Acknowledgements

The authors would like to thank the reviewers of the First International Conference on Tone and Intonation in providing feedback to the abstract and consequently helping us improve this paper.

8. References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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

[6] 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

[7] 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.

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

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

[10] 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.

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

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

[13] 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.