The Lombard Effect with Thai Lexical Tones: An acoustic analysis of articulatory modifications in noise

Benjawan Kasisopa, Virginie Attina, Denis Burnham
MARCS Institute, University of Western Sydney, Australia
b.kasisopa@uws.edu.au, virginie.attina@inserm.fr, denis.burnham@uws.edu.au

Abstract

When in a noisy environment speakers modify their speech production by increasing loudness, vowel duration, and fundamental frequency (F0), a phenomenon known as Lombard speech. Here Lombard speech in Thai was investigated in order to determine the effects of noise on the realisation of F0 in Thai lexical tones. Analysis of the acoustic characteristics of the five Bangkok Thai tones, in both continuous speech and citation form, produced in noise and in quiet showed that F0 is heightened in Lombard compared with clear speech. In addition, generally the contour of the tones also changes in Lombard speech; contours tend to be exaggerated towards the end of the tone in a direction consistent with the contour of the tone.

Index Terms: Speech production, Lombard speech, Tone languages

1. Introduction

Speakers tend to modify the manner in which they speak depending on the audience, e.g., an adult, a foreigner, an infant, or a pet [1, 2], in response to the communicative context, e.g., whispered speech in necessarily quiet circumstances or what is known as Lombard speech [3] in noisy environments. Lombard speech is a complex phenomenon; its main acoustic characteristics compared to clear speech are heightened intensity and heightened fundamental frequency (F0), lengthened vowel duration, as well as flattened spectral tilt and upward shift of mostly the first but also the second formant (F1, F2) frequency [3, 4, 5, 6; for a recent review, see 7]. Lombard speech is studied here with particular regard to Thai lexical tones.

Lexical tones are linguistically contrastive in 70% of the world’s languages [8]. While duration, F2 values, voice quality, and amplitude (perceived as vowel length, vowel height, vowel quality, and loudness respectively) all play some part in lexical tone [9, 10, 11], the phonetic feature of lexical tone is F0 (perceived as pitch). In tone languages, a change in F0 height and/or contour can result in a change in the identity of the lexical tone and the meaning of the word. For example, Bangkok Thai has five tones, three level (Low[21], Mid[33], and High[45]) and two contour (Falling[241] and Rising[315]); the numbers are Chao values [12], in which the 1st number is the relative pitch height at word start, the last number the pitch height at word end, and any numbers in between are the relative pitch height of any internal point of inflection). Accordingly, [maa33] (i)-mid tone means ‘to come’, [maa21] (ii)-low tone ‘brew’, [maa45] (i)-high tone ‘horse’, [maa241] (i)-falling tone ‘grandmother’, and [maa315] (ii)-rising tone ‘dog’.

Speech corpora collected for Lombard speech studies usually consist of citation form words produced while the speakers listen to Gaussian white-noise through headphones [4, 13]. However, it has now been found that listeners adjust their speech output depending on the amplitude modulations in the noise and the type of background noise e.g., white, pink, or babble noise [6, 14]. Lu and Cooke [14] have suggested that speakers monitor background noise when producing speech in noise and adopt the most efficient and effective speech production strategy that most facilitates the listener’s perception in that particular noise environment.

Since one of the key features of Lombard speech is an increase in F0, it is of theoretical and practical importance to investigate Lombard speech with lexical tones, which depend heavily upon F0 modulation for their realisation. How will background noise influence the realisation of the word level F0 fluctuations that are important in lexical tone production and perception? In the only study of Lombard speech in a tone language thus far, Zhao and Jurafsky [15] compared Cantonese monosyllabic words produced in citation form in clear conditions and in 75 dB white noise. They found that all six Cantonese tones were produced with higher F0 in Lombard than in clear conditions. The authors also investigated differential adjustments for more or less frequent words and found that less frequent words were produced with higher pitch especially for mid-low tones. However, as this effect did not interact with the effect of Lombard speech, it can be concluded that the Lombard effect occurs at a relatively low level.

Zhao and Jurafsky [15] only investigated words in isolation. We already know that there are adjustments to Cantonese tones in continuous speech in another form of hyperarticulation, namely infant-directed speech. Xu, Reilly and Burnham [16] found that compared to speech to another adult, Cantonese mothers talking to their 3- to 12-month-old infants expanded their F0-onset/F0-offset tone space [17] and also had a general upward shift in mean pitch, and expansion of pitch range. Given these adjustments, it is reasonable to expect that in sentences adjustments to tone realisation in Lombard speech might be more complex that in citation words. Here the Lombard effect is investigated in Thai and comparisons are also made between the production of isolated words and continuous speech.

2. Method

2.1. Participants

Six native Thai speakers (age range, 27 to 34 years, mean age 31 years; four females) participated in this study. All of them were native speakers of Bangkok Thai, born and raised in Thailand and had lived in Australia for less than four years. They reported no known speech or hearing disorders and presented no speech production problems. They all gave informed consent to participate in the experiment and received AUD$50 compensation for their participation. The study was conducted at MARCS Institute under University of Western Sydney Human Research Ethics Committee approval.

Copyright © 2014 ISCA
2.2. Materials (or Corpus)

The corpus was composed of four different syllables (/ma:/, /lon:/, /kai/, /kai/ = IPA annotation) each produced with each of the five Thai tones giving a total of 20 different Thai syllables, which are all words in Thai. Each was repeated five times in citation form and five times in sentential context. The utterance "ติ่มผักผักนำมาแล้ว...เป็นติดอยู่ [t̚wān/ʔpʰom]kaʔpʰoːt kʰām̥wāː...pʰn̥t̚ājəːŋ]" in Thai meaning "I will say 'TARGET' as an example." was used as the carrier sentence. The F0 contours of the 5 tones are illustrated in Figure 1.

![Figure 1: F0 contours of 5 Bangkok Thai tones](image)

2.3. Data recording

The recordings were conducted in a soundproof booth in the Face and Voice Lab at MARCS Institute, University of Western Sydney. The speakers were asked to produce target syllables in blocks of speech in clear or noise conditions via a Behringer microphone connected to an Optotak Data Acquisition Unit II (Northern Digital Inc.) through a Behringer Eurorack MX 602A mixer. (Note that Optotak equipment was used to collect motion capture data synchronously with the audio but these data are not presented in this paper.) Sound was sampled at a frequency of 44,100Hz. Target syllables were presented stationary within the sentence frame in the centre of a computer screen one at a time; all speakers were asked to remember the sentence frame and to produce each target syllable in the utterance context first, and then after a short pause, in isolation. In the noise condition, participants produced the target syllables while listening through Sennheiser HD25-II headphones to 75 dB SPL white-Gaussian background noise which was calibrated by a sound-level meter prior to the experiment. In addition, self-monitoring feedback of their own voice was presented through the headphones as this does not affect mean F0 [18]. The target syllables were presented pseudo-randomly in each set of clear versus Lombard speech recordings so that the same segments with different tones were not presented contiguously. The participants produced all target syllables in different sequences and in both the clear and the noise condition they provided five productions of each word.

2.4. Data processing and analysis

The sound files for both conditions were segmented and analysed using Praat version 5.3.57 [19]. All the word and tone productions were checked by a native phonetically-trained Thai speaker. Word boundaries and tone contour trajectory were hand-labelled and annotated by a trained phonetician.

The data for each of the 6 speakers (2 male, 4 female) consisted of 2 contexts (words in isolation, words in sentences) x 2 speaking conditions (in noise or clear) x 4 syllables (/ma:, /lon:, /kai:, /kai/) x the 5 Thai tones (low, mid, high, falling, rising) x 5 repetitions. Thus there was a total of 20 syllables repeated 5 times (n=100) by each speaker in each of the 4 Context x Condition sub-sets, a total of 400 productions / speaker, and a total 2,400 productions in all.

For each production, the F0 values of tone trajectory of each syllable were automatically measured in ten equidistant and time-normalised points (10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 100%) across different syllable structures using the ProsodyPro Praat script created by Xu [20]. As there were both male (n=2) and female (n=4) speakers in order to take into account individual differences, F0 value was normalised for each speaker using the formula: (F0 – minF0)/(MaxF0 – MinF0) because the data of each speaker were normally distributed without any outliers.

The data were then analysed separately in five analyses of variance (ANOVAs), one for each of the five Thai tones. The factors in each of the 5 analyses were Context (Words/Sentences) x Condition (Clear/Noise) x Time Points (10 points 10% intervals) with repeated measures on all three factors. In addition to the main effects which would reflect F0 height, for the time points, Linear, Quadratic and Cubic trends over time were tested in each of the five tone ANOVAs in order to determine the contour of each tone in terms of polynomial components, and, in interaction with the Context and Condition factors and their interaction, how the contour of each tone was affected by context of noise. The data were analysed across three random factors – Speakers (F0 was normalised so any differences between the 2 male and 4 female voices were not an issue), Syllables (4 different syllables were included simply to provide a variety of phonetic contexts and increase the generalisability of the results), and Repetitions (5 for each word).

3. Results

Plots of the normalised F0 for each of the five Thai tones are presented in Figures 2L, 2M, 2H, 2F, and 2R for low, mid, high, falling, and rising respectively. In each there are four curves, for speech in clear vs noise (Lombard) x words vs sentences. As can be seen, F0 for words in sentences is generally higher than for the same words produced in isolation, and this is the case across Clear and Lombard speech and across all five tones. With respect to Lombard speech (i) for words in isolation it can be seen that tones are generally produced at a higher F0 in Noise than in Clear conditions; and (ii) for words in sentences there appears to be some heightening of F0 in Noise compared to Clear, but this is a much smaller effect than for words in isolation, and is evident mainly in the low, mid and rising tones.

To investigate the Lombard Effect in detail, as set out above, the results of the five analyses of variance were conducted, one for each tone all with the design: Context (Words/Sentences) x Condition (Clear/Noise) x Time Points (10 10% points) with repeated measures on all these factors. The critical F-value at \( \alpha=0.05 \) is 3.921. The F-values for all factors and their interactions for each of the 5 tones are shown...
in Table 1, with significant F-values, i.e., >3.921 in italics, bold or bold italics.

Words vs Sentences: For all 5 tones, Low, Mid, High, Falling, Rising, words in sentences were produced at a higher normalised F0 than words in isolation, and the tone contours were flatter in sentences than in isolation.

Lombard vs Clear: Most importantly, Lombard speech was generally produced at a higher F0 than Clear speech, and this was significant for 4 of the 5 tones, Low, Mid, High, Rising, but not Falling. Moreover there were differences as a product of the Lombard/Clear x Words/Sentences interaction for Mid, High, and Rising, but not for Low or Falling.

Trend Analyses: The trend analyses showed that there were significant trends over time for all 5 tones as would be expected, but of most interest in this regard, is the fact that trend coefficients interacted with Lombard vs Clear and also with Lombard/Clear x Words/Sentences for 4 of the five tones, Mid, High, Falling, Rising, but not Low.

Table 1: F-values for all factors and their interactions for each of the 5 tones

<table>
<thead>
<tr>
<th></th>
<th>Low</th>
<th>Mid</th>
<th>High</th>
<th>Falling</th>
<th>Rising</th>
</tr>
</thead>
<tbody>
<tr>
<td>C vs L</td>
<td>24.9</td>
<td>29.6</td>
<td>33.1</td>
<td>N/S</td>
<td>40.1</td>
</tr>
<tr>
<td>W vs. S</td>
<td>1162.1</td>
<td>907.1</td>
<td>421.0</td>
<td>859.8</td>
<td>106.4</td>
</tr>
<tr>
<td>C/L x W/S</td>
<td>N/S</td>
<td>12.4</td>
<td>31.9</td>
<td>N/S</td>
<td>6.1</td>
</tr>
<tr>
<td>Linear</td>
<td>1895.6</td>
<td>813.1</td>
<td>66.7</td>
<td>568.3</td>
<td>7.5</td>
</tr>
<tr>
<td>Cubic</td>
<td>N/S</td>
<td>121.4</td>
<td>64.3</td>
<td>153.8</td>
<td>87.4</td>
</tr>
<tr>
<td>C/L x Linear</td>
<td>N/S</td>
<td>N/S</td>
<td>N/S</td>
<td>N/S</td>
<td>N/S</td>
</tr>
<tr>
<td>C/L x cubic</td>
<td>N/S</td>
<td>N/S</td>
<td>9.4</td>
<td>6.2</td>
<td>6.4</td>
</tr>
<tr>
<td>W/S x Linear</td>
<td>179.5</td>
<td>38.9</td>
<td>144.3</td>
<td>253.5</td>
<td>429.6</td>
</tr>
<tr>
<td>W/S x Cubic</td>
<td>N/S</td>
<td>49.9</td>
<td>90.4</td>
<td>31.2</td>
<td>57.4</td>
</tr>
<tr>
<td>C/L x W/S x Linear</td>
<td>N/S</td>
<td>N/S</td>
<td>20.1</td>
<td>43.1</td>
<td>34.6</td>
</tr>
<tr>
<td>C/L x W/S x Cubic</td>
<td>N/S</td>
<td>17.8</td>
<td>13.5</td>
<td>59.3</td>
<td>N/S</td>
</tr>
<tr>
<td>C/L x W/S</td>
<td>N/S</td>
<td>15.7</td>
<td>27.4</td>
<td>N/S</td>
<td>5.9</td>
</tr>
</tbody>
</table>

Note: C stands for Clear speech; L stands for Lombard speech; W stands for Word; S stands for Sentence; and N/S stands for not significant.

Italic figure = p<.05; Bold figure = p<.01; and Bold Italic figure = p<.001.

Figure 2: Normalised F0 across normalised time at 10% intervals for speech in clear vs noise (Lombard) x words vs sentences for low (2L), mid (2M), high (2H), falling (2F), rising (2R).
4. Discussion

The results show that there is a relatively consistent effect of producing speech in noise. As was also found by Zhao and Jurafsky [15] with Cantonese, we found that in noise Thai tones are produced with a higher overall F0. However, when considering only the main effects, this was found for just 4 of the 5 Thai tones, with Falling tone being the odd one out. This difference between this and the only previous study of its kind could be explained by a combination of factors. First, there are methodological differences between our study and Zhao and Jurafsky [15]. Here, self-monitoring feedback was added to the background noise via the headphones, and this is known to reduce Lombard effects although mainly in the intensity, not the frequency (F0) domain [18]. Second, inspection of Figure 1 shows that the Falling tone is the tone with the least competition from the other four tones in the Thai tone space. So it is possible that the distinctiveness of the Falling tone may have assisted in its self-monitoring here, and that modification of this tone in noise may be less necessary for the speaker to be understood by the perceivers.

In addition to the mean F0 effects, the Clear/Lombard x Words/Sentences interactions were also significant (for all but the Falling and the Low tones) showing that the Lombard effect tends to differ in manner over words and sentences. Further information on this is apparent in the interactions of Trends x Clear/Lombard, and Trends x Clear/Lombard x Words/Sentences; for all the tones except Low, at least two of the three possible interactions with the trend coefficients (Linear, Quadratic, or Cubic) were significant.

These results may be summarised as follows. There is a relatively consistent heightening of mean F0 in Thai tones when they are produced in noise. The only exception is the Falling tone, which overall had the same mean F0 in noise and clear. In addition to differences in mean F0, the F0 contour manifested over time was different in noise than clear for all tones except the Low tone. The effect of noise in each of the tones is described below in terms of whether there were effects for mean F0, contour F0, or both.

The Low tone shows a consistent Lombard effect for mean F0, but the tone contour is unaffected by noise.

The Falling tone does not show a consistent Lombard effect for mean F0, but the tone contour is affected by noise; in isolated words the tone contour is more exaggerated with a higher peak and a lower F0-offset. Two reasons for the lack of a mean F0 effect were posited above. It seems that, as the contour of the Falling tone is selectively unaffected on the F0 dimension, it is the characteristics of the Thai tonal space rather than the particular method employed here that elicits these effects. The Falling tone has a unique contour in Thai, and the effect of noise is to emphasise that difference in contour, not the overall pitch height.

The Mid, High, and Rising tones all show a consistent heightening of F0 in noise, i.e., a Lombard effect for mean F0. In addition, the contour F0 of the Mid, High, Rising, and falling tones also changes from Clear speech to Lombard Speech: For the Mid tone there is a higher peak towards the end of the tone productions in Lombard speech; and for High and for Rising tones, there is a higher F0-offset of the tone productions in Lombard speech.

5. Conclusions

Consistent with the results of Zhao and Jurafsky [15] with Cantonese, mean F0 of Thai tones is raised in Lombard speech. Over and above the Zhao and Jurafsky results, however, it was found that the contour of the tones changes in Lombard speech. These changes tend to exaggerate the tones towards the end of the tone production in a direction consistent with the contour of the tone; for the Falling tone the F0-offset is lower, but for the three higher/rising trajectory tones, Mid, High, and Rising, tone offset is at an exaggerated higher frequency at the end of the tone. These results add to the results of Zhao and Jurafsky [15] in two ways. First, there were differences found here as a product of whether the words were in isolation or in sentences. Zhao and Jurafsky did not use words in sentences and so could not test this. Second, both F0 height and F0 contour effects were found here. Zhao and Jurafsky found no such contour effects, possibly because they only analysed raw and not normalised F0, and also they did not look for them, they did not include trend coefficients in their analyses. Another reason for any lack of effect of noise on tone contours in Cantonese may be that the range of individual tone trajectories in Cantonese is not as great as in Thai, and the tone space is more condensed than in Thai, especially in its three level tones. Further studies using words in isolation and words in sentences with normalised F0 analysed for both height and contour (via trend analyses) are required for a range of tone languages to investigate the generality of these findings across tone languages.

6. Acknowledgements

This research was supported by an Australian Research Council (ARC) Discovery grant (DP0988201) to the third author.

7. References


