Pre-boundary lengthening modulates predictability effects on durational variability in Taiwan Southern Min

Sheng-Fu Wang
Institute of Linguistics, Academia Sinica, Taiwan
sftwang@gate.sinica.edu.tw

Abstract
Linguistic units with lower predictability (i.e., higher surprisal) tend to be realized with stronger acoustic cues, so do units closer to a prosodic break. This study examined how predictability measurements such as lexical frequency, surprisal, and neighborhood density affect the variations of syllable duration and how these effects interact with the durational marking of prosodic phrasing in Taiwan Southern Min. Speech data were extracted from an eight-hour spontaneous speech corpus. Surprisal was estimated with trigram language models trained on a written corpus with 4.13M words. Results show that syllable duration had a positive correlation with surprisal but a negative correlation with neighborhood density. As for positional effects, all predictability effects were neutralized at the pre-boundary syllable, even though there were still predictability effects in cases where penultimate and ante-penultimate lengthening is observed. These findings highlight how predictability effects are modulated by the durational marking of prosodic phrasing and have implications for an information-theoretic view of the balance of signal redundancy in human speech.

Index Terms: predictability, surprisal, neighborhood density, final lengthening, tone language

1. Introduction
This study investigated how local and global measures of contextual predictability affect syllable duration and how they interact with the durational marking of prosodic phrasing in Taiwan Southern Min, commonly known as “Taiwanese.” The findings have implications for an information-theoretic view of the balance of signal redundancy in human speech.

Shortening or segmental reduction of words and syllables that have higher lexical frequencies is well-noted by research in linguistic, and speech sciences [1, 2, 3, 4, 5]. The same observations are also found for words and syllables in the contexts that they are more likely to occur [6, 7, 8, 3, 4, 9, 10, 11, 12, 1, 5]. The consistent finding is that higher frequencies or higher contextual likelihoods (i.e., lower surprisal) are correlated with phonetic shortening or segmental reduction in the signal. These findings can be referred to as “predictability effects” on phonetic realizations, based on the notion that it is easier to guess, or predict, the occurrence of a unit when it is more probable regardless of context (i.e., high lexical or type frequency) or in a particular context (i.e., contextual probability).

From an information-theoretic [14, 15] point of view, the correlation between higher predictability and acoustic weakening shows that a more predictable unit carries less information and thus does not require strong acoustic cues. On the other hand, if a linguistic unit is infrequent itself or occurs in a surprising context, its acoustic signal needs to be strengthened to increase its chance of being correctly recognized. This account has parallels with proposals on speech production and processings such as the Hyper- & Hypo-articulation (H&H) theory and the [16] Smooth Signal Redundancy Hypothesis [4, 17], the latter of which states that the amount of information is distributed evenly across each element of an utterance so that “linguistic redundancy” (i.e., lexical, syntactic, semantic, and pragmatic information) and “acoustic redundancy” (i.e., acoustic cues) have an inverse relationship. A schematic representation of the relationship is shown in Figure 1 (Figure 5 in [13]).

Phonological neighborhood density is another measurement that has an information-theoretic interpretation. For a word X, its neighborhood density refers to how many words in the lexicon are just one edit distance (i.e., addition, substitution, deletion) away from X [18]. From the view of predictability in word processing, having more neighbors means that among the pool of possible words that are phonologically similar, there are more potential choices in word recognition so that it is harder to identify the target. In other words, having more phonological neighbors may also result in low predictability. Previous studies have shown that words with a higher number of phonological neighbors tend to exhibit greater phonetic salience [19, 20, 21, 22, 23, 24], which is consistent with the findings on low predictability units. However, results in the opposite direction have also been reported: [25] show that words with higher neighborhood density are more often reduced, which is taken to be evidence supporting a production-based view of phonetic variation: words with higher neighborhood density are easier to access thus easier to produce.

Predictability is not the only factor that affects phonetic realizations. Especially for durational variability in acoustic cues, prosodic factors such as phrasing and prominence also play an important role. For example, it has been commonly found that lengthening occurs at the right edge of prosodic units [26, 27]. The interaction between this boundary lengthening effect and predictability has been discussed in the literature. [4] hypothesize that the prosodic structure encodes and thus mediates the relationship between information redundancy and acoustic cues. Under this view, the relationship between predictability
and phonetic cues is not direct. Rather, the placement of prominence and organization of phrasing reflects the predictability profile of an utterance (e.g., unpredictable units are more likely to be focused on in certain prosodic positions), and the seemingly direct correlation between phonetics and predictability follows from the phonetic correlates of focus and phrasing (see [28] for a review of this view). This view is supported by empirical works showing how a prosodic structure can account for a large proportion of what predictability variables could explain [29, 4, 28].

Recently, some studies directly examined the relationship between surprisal and prosodic structure. [11] investigates phonetic variability in multiple cues (syllable duration, vowel dispersion, vocalic spectral emphasis, consonant center of gravity) in six languages and show that while prosodic structure accounts for phonetic variability, there is also some direct effect of predictability on phonetic realizations. [12] compares the effect of predictability on syllables at strong and weak boundaries with or without a phrasal accent. They find stronger predictability effects at stronger boundaries, while accent placement does not interact with surprisal. It suggests that the effect of predictability and prosodic structure do not necessarily have a trade-off relationship.

Building on these previous works, this study aimed to compare the effect of predictability at different prosodic conditions directly. The research target was syllable duration in the spontaneous speech of Taiwan Southern Min. The main research question was whether the effect of predictability on syllable duration was consistent across different positions before a prosodic boundary. In other words, given the known effect of pre-boundary lengthening, this study investigated whether the effect of predictability was neutralized by boundary effects, as predicted by the Smooth Signal Redundancy Hypothesis. This study also asked an additional question on whether the strength of lengthening between two boundary types (with and without final particles) affected the patterning of predictability effects. Finally, targeting Taiwan Southern Min also expands the inquiry of the role of predictability in phonetic variability to tone languages, which is currently under-explored.

2. Method

2.1. Speech Corpus

Durational data were extracted from a corpus of eight hours of Taiwan Southern Min spontaneous speech [30]. Sixteen speakers each contributed around 30 minutes of recordings to the corpus. The speakers were evenly split in gender and two age groups. The “old” and “young” groups were divided by whether the speakers were born before or after 1975. The alignment between transcription and sound files was done at the syllabic level with EasyAlign [31] and manually checked by trained phoneticians. Word segmentation and orthography follow the official dictionary released by Taiwan’s Ministry of Education (MOE). The corpus also contains annotations of discourse and prosodic boundaries. Discourse annotation was conducted independently of the sound files. The transcriptions were segmented into clausal units. The relationship between adjacent units was labeled according to the degree of separation of their topics. Prosodic annotations identified two levels of prosodic boundaries (intonational and intermediate phrases).

Taiwan Southern Min has a tone sandhi process where all syllables switch their tones except for the final one in a “tone sandhi group”, the formation of which is governed by morphosyntactic organization [32] as well as phonological and lexical factors [33, 34]. Since the tone sandhi group has been described as a level in the prosodic hierarchy in this language [35], this study performed additional annotations on whether a syllable was realized in the sandhi tone, so that tone sandhi groups could be identified.

To have a more stringent control of boundary type, this study focused on the right edges of discourse units that were also intonational phrases. Among the 9981 clausal boundaries marked in this corpus, 6567 (65.79%) of them corresponded to an intonational phrase boundary. Since the targets were final, penultimate, ante-penultimate, and initial/medial positions before a boundary, the analysis only included discourse units longer than three syllables, which amounted to 5817 discourse units (58.28% of all boundaries in the data set). The selected data were further filtered to exclude disfluencies, code-switching, and additional prosodic breaks within the discourse unit. The final data set going into the analyses contained 42925 syllables (33368 words).

2.2. Language modeling corpus

This study used a written corpus of Taiwan Southern Min collected by [36] originally for a study on lexical frequencies. The corpus covers various genres, including literary works, journalist reports, conversation transcripts, commentaries, and academic writings. Additional pre-processing was done to transform the orthography to be consistent with the MOE dictionary. Word segmentation was performed following the MOE dictionary with the maximal length matching method. Given these pre-processing measures, the final word count was around 4.13M (from 5.42M syllables/characters). Finally, trigram language models were trained on the corpus with modified Kesner-Ney smoothing [37] at the word level using the SRILM toolkit [38] in forward and backward directions separately.

2.3. Variables in the regression analyses

Surprisal given previous and next word were calculated from language models trained on the corpus described in Section 2.2. For each word token, surprisal given the previous and following word token was calculated and used as separate variables. Bigram surprisal was used even though trigram language models were trained since it has been shown that using trigrams instead of bigrams only produces negligible improvement in predicting reductions in speech [39]. Unigram surprisal from the language models was also included in the analyses as a measurement of lexical frequencies.

Neighborhood Density refers to the number of phonological neighbors that a syllable has, based on the syllable types that occur in MOE dictionary. A phonological neighbor is defined as a syllable that is different from the target syllable by deletion, addition, or substitution of a segment or tone. For example, /pak/’s neighbors would include /bak/, /pat/, /pak/, and /ak/.

Following a similar method reported in [1], a linear regression model was trained on the corpus to predict the duration of each syllable from what its onset, onglide, vowel, offglide, and coda are. Syllable length measured in the number of segments was also added as a variable in this linear model. Predictions made by this model are used as a predictor, Baseline duration, in the analysis. Speech rate refers to the number of syllables per second in each prosodic unit. A larger number indicates a faster speech rate.

All the aforementioned numeric predictors were log-
transformed with base ten and transformed to z scores, except for Neighborhood Density, which was only z-score transformed since the distribution before log-transformation was more normal.

**Syllable position** refers to a syllable’s relative position to the nearest discourse/prosodic boundary that followed it. Four categories were examined in this study: Initial/Medial, Ante-penultimate, Penultimate, and Final. **Final Particle** is a related controlling variable that refers to whether the prosodic/discourse boundary ended with a final particle. The interaction between **Syllable position** and **Final Particle** was included in the analysis to account for the potentially different pre-boundary lengthening patterns conditioned on the presence of final particles. **Word length** describes the length of the word measured in the number of syllables. This predictor aims to model a potential inverse relationship between the length of a word and the length of its components (e.g., [40, 41]).

**Sandhi grouping** described whether a syllable occupied the right edge of the domain of sandhi application. Inclusion of this variable tested and controlled for the effect of tone sandhi groups on phonetic variability.

**Surface tone** refers to the surface tonal category of a syllable, instead of the ‘base’ tonal category before the tone sandhi process in Taiwan Southern Min. In addition to the seven lexical tonal categories, there are also the ‘de-stressed’ category and the ‘particle’ category. The ‘de-stressed’ category refers to cases where a syllable carries neither the base or surface tone and is often accompanied by phonetic reduction. For example, the expression 他家中的小孩 / ‘he jia zhong de xiao huan’ can either be pronounced with the final two syllables de-stressed (他家中的小孩 / ‘tia jia zhong de xiao huan’ ‘buy a few’) or with regular tone sandhi application (vs. 他家中的小孩 / ‘tia jia zhong de xiao huan’ ‘buy exactly two’). ‘Particle’ categorizes expressions such as ‘ah’ and ‘lah’, which do not have inherent tonal targets.

### 2.4. Modeling procedure

The main analysis was conducted with the *lmer* function in the lme4 package [42]. Fixed effects included the variables mentioned in the previous section. We also added random intercepts for Morpheme (monosyllabic), Word, and Speaker.

Additional models were run where only each of the four predictability variables was included at one time, along with the controlling variables and random intercepts. This follows the discussion and recommendation of [43], who argued that in cases where collinearity might be a concern, estimating the effect size of each variable in such ‘simultaneous regression analyses’ is preferable both to multivariate regressions and to using residualized variables. To examine the effectiveness of these predictors at each position given different boundary types, the same set of simultaneous regression analyses was performed with the entire dataset, as well as at each of the eight categories of boundary position × whether the boundary ends with a final particle.

### 3. Results

Fixed effects of the main analysis are summarized in Table 1. Neighborhood density, surprisal conditioned in either position, and unigram surprisal were all significant predictors of syllable duration. Bigram and unigram surprisal had an expected positive relationship with longer syllable duration. On the other hand, Neighborhood density had a negative correlation with syllable duration, i.e., syllables with more phonological neighbors tended to be shorter.

#### Table 1: Summary of fixed effects in the mixed-effects model for the entire dataset

<table>
<thead>
<tr>
<th>Fixed effects</th>
<th>β</th>
<th>SE</th>
<th>t</th>
<th>p(χ²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.21</td>
<td>0.049</td>
<td>4.29</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Baseline dur.</td>
<td>0.26</td>
<td>0.010</td>
<td>26.13</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Speech Rate</td>
<td>-0.16</td>
<td>0.004</td>
<td>-40.99</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>N. density</td>
<td>-0.04</td>
<td>0.012</td>
<td>-3.57</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Sur. given prev.</td>
<td>0.04</td>
<td>0.004</td>
<td>8.95</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Sur. given next</td>
<td>0.05</td>
<td>0.004</td>
<td>12.85</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Unigram sur.</td>
<td>0.04</td>
<td>0.009</td>
<td>4.65</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Ante-penult.</td>
<td>0.29</td>
<td>0.016</td>
<td>18.88</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Penult.</td>
<td>0.59</td>
<td>0.017</td>
<td>35.17</td>
<td></td>
</tr>
<tr>
<td>Final</td>
<td>1.19</td>
<td>0.024</td>
<td>49.55</td>
<td></td>
</tr>
<tr>
<td>No FP</td>
<td>-0.04</td>
<td>0.008</td>
<td>-4.64</td>
<td></td>
</tr>
<tr>
<td>× Ante-penult.</td>
<td>-0.27</td>
<td>0.020</td>
<td>-13.84</td>
<td></td>
</tr>
<tr>
<td>× Penult.</td>
<td>-0.33</td>
<td>0.021</td>
<td>-15.80</td>
<td></td>
</tr>
<tr>
<td>× Final</td>
<td>-0.19</td>
<td>0.028</td>
<td>-6.60</td>
<td></td>
</tr>
<tr>
<td>Base tone</td>
<td>0.13</td>
<td>0.012</td>
<td>11.24</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Word length</td>
<td>-0.15</td>
<td>0.013</td>
<td>-11.44</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Unstressed</td>
<td>-0.02</td>
<td>0.047</td>
<td>-0.47</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Tone2 (53)</td>
<td>0.02</td>
<td>0.017</td>
<td>1.03</td>
<td></td>
</tr>
<tr>
<td>Tone3 (21)</td>
<td>-0.08</td>
<td>0.020</td>
<td>-3.74</td>
<td></td>
</tr>
<tr>
<td>Tone4 (2)</td>
<td>-0.07</td>
<td>0.031</td>
<td>-2.28</td>
<td></td>
</tr>
<tr>
<td>Tone5 (24)</td>
<td>0.05</td>
<td>0.030</td>
<td>1.80</td>
<td></td>
</tr>
<tr>
<td>Tone6 (22)</td>
<td>0.02</td>
<td>0.017</td>
<td>1.33</td>
<td></td>
</tr>
<tr>
<td>Tone8 (4)</td>
<td>-0.06</td>
<td>0.030</td>
<td>-2.13</td>
<td></td>
</tr>
<tr>
<td>Particle</td>
<td>-0.79</td>
<td>0.055</td>
<td>-14.52</td>
<td></td>
</tr>
</tbody>
</table>

Pre-boundary lengthening interacted with whether a boundary had a final particle, as shown in Figure 2. With a final particle, there was a trisyllabic domain of pre-boundary lengthening (p < .0001 for all six pairs of comparisons with lsmeans [44] post hoc analysis), while there was only a bisyllabic domain of lengthening without final particles (p < .0001 for five pairs of comparisons; Ante-penultimate not significantly different from Initial/Medial at p = .68).

![Figure 2: Syllable duration (z-scores) at different prosodic positions as a function of whether the boundary ends with a final particle](image-url)

The summary in Table 1 also shows that controlling variables had expected effects. There was a positive relationship between baseline duration from a syllable’s segmental content and its actual duration. When speech rate was faster, syllable duration was shorter. Syllables in base tone, i.e., next to a tone sandhi group boundary, were also longer. Syllables in longer words tended to be shorter (main effect on word length). Summary of random effects in Table 2 shows that the random intercept of Morpheme accounted for more residual variances than the intercepts of Word and Speaker.

The results of simultaneous regression models for each pre-
dictability measurement are shown in Table 3, which gives a more interpretable picture of the effect size of each variable. We see a stronger effect of unigram surprisal as compared with surprisal, which is not observed in the main analysis in Table 1.

Table 3: Summary of fixed effects in simultaneous regression analysis for each predictability variables

<table>
<thead>
<tr>
<th>Predictability Variables</th>
<th>( \beta )</th>
<th>SE</th>
<th>( t )</th>
<th>( p(\chi^2) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>N. density</td>
<td>-0.042</td>
<td>0.012</td>
<td>-3.71</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Surp. given previous</td>
<td>0.057</td>
<td>0.004</td>
<td>16.66</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>Surp. given next</td>
<td>0.064</td>
<td>0.004</td>
<td>17.86</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>Unigram surpr.</td>
<td>0.087</td>
<td>0.008</td>
<td>10.23</td>
<td>&lt; .0001</td>
</tr>
</tbody>
</table>

Finally, Figure 3 presents the estimate of effect size for each predictability variable in each prosodic position and boundary type (with vs. without final particle). This allows us to see how predictability and pre-boundary lengthening interact with each other. The major trend is that all predictors converge to a smaller effect size moving towards the boundary. The convergence was more dramatic at the final position of a boundary without a final particle, as none of the predictors were significant in that setting. This neutralization only partially applied to the final particles themselves, as they were not longer with higher bigram surprisal, although there was a nearly significant effect of less frequent particles being longer (\( p = .09 \)) under post hoc analysis.

![Figure 3: The effect sizes of predictability variables in simultaneous regression analyses at each prosodic position in Taiwan Southern Min. The effect size for neighborhood density at the final position with final particles (i.e., the effect of neighborhood density on final particles) is omitted for its extreme estimate and standard error as compared with other predictors: \( \beta = -0.6, SE = 0.53 \).](image)

It is also worth noting that predictability effects were still present in cases where penultimate and ante-penultimate lengthening was observed, even though the effect sizes become smaller in most cases. Neutralization of predictability effects is only observed at the penultimate position for surprisal given the previous word. Since the final syllable was always a final particle in this setting, the maintenance of effects for surprisal given the next word might be an anomaly to a real neutralization effect, as probabilities of words given a final particle could be exceedingly predictive of syllable durations.

4. Discussion

Overall, this study replicated the positive correlation between higher surprisal (lower predictability) and syllable duration, and this is true both for unigram surprisal (lexical frequency) and surprisal (contextual predictability). Neighborhood density, however, had a negative relationship with syllable duration. This is similar to the findings in [25] and [45] and is more in line with a production-based view on the relationship between neighborhood density: Words with more neighbors are more easily accessible, and quicker retrieval leads to more phonetic reductions, as commonly assumed in the literature [46, 10, 47]. It also shows that neighborhood density, while also having an information-theoretic interpretation, has a different behavior than measurements based on word frequencies and contextual predictability, suggesting that a different representational and processing mechanism might be involved.

As for the interaction between pre-boundary lengthening and predictability, which is the main research target of this study, the results show that predictability effects were almost completely erased at the final syllable before a prosodic/discourse boundary, where pre-boundary lengthening was the strongest. This neutralization effect was found regardless of whether the boundary had a final particle. However, the presence of any pre-boundary lengthening did not always result in neutralization of predictability effects, as neighborhood density and surprisal still predicted syllable duration in cases where ante-penultimate and penultimate lengthening occurred.

To conclude, these findings are consistent with the Smooth Signal Redundancy Hypothesis, as the strength of prosodic marking inversely correlated with the role of linguistic predictability factors moving towards a discourse and prosodic boundary in Taiwan Southern Min. This relationship was found even for neighborhood density, a variable whose predictability effect went in the opposite direction with that of surprisal. While gradient trade-off between the roles of predictability and prosodic phrasing is found moving toward a boundary, the almost absolute neutralizing effect at the pre-boundary syllable might be viewed as supporting a stronger version of the hypothesis that prosodic structure mediates between information density and phonetic realizations.

5. References


