WORD-LEVEL AND SYLLABLE-LEVEL PREDICTABILITY EFFECTS ON SYLLABLE DURATION IN TAIWAN MANDARIN

Sheng-Fu Wang, Shu-Chuan Tseng

Institute of Linguistics, Academia Sinica, Taiwan
sftwang@gate.sinica.edu.tw, tsengsc@gate.sinica.edu.tw

ABSTRACT

This study examined how word-level and syllable-level predictability affects the variability of syllable duration in disyllabic words in Mandarin conversational speech. Specifically, we examined how the correlation between predictability measurements (unigram surprisal, bigram surprisal, and informativity) and syllable duration was affected by final lengthening, and whether word-level predictability effects were uniform across syllables. Our results showed that incorporating both levels of predictability variables improved model fit significantly compared to using only one level. We also found that predictability effects weakened at prosodic boundaries, and that word-level bigram surprisal had a stronger impact on syllables closer to relevant word boundaries, while word-level informativity and unigram surprisal were more predictive of duration for a word's first syllable. Overall, our findings highlight the relevance of different types and levels of statistical information for various aspects of durational variability in Mandarin.

Keywords: predictability, Mandarin, syllable, duration, word

1. INTRODUCTION

Linguistic units that occur more frequently or are more likely to occur in a given context are often reduced or shortened (e.g., [15, 2, 21, 12]). Typically, frequency and probability estimation is conducted at the word level, focusing on how word-level frequency and contextual probabilities affect word duration. Although some recent studies have shown this correlation at the syllable level [18, 22], the relationship between word and sub-word predictability measurements in shaping speech production remains underexplored. To complement these studies, we investigated how word-level and syllable-level predictability affect syllable duration in disyllabic words in Taiwan Mandarin.

This study aimed to provide a more nuanced understanding of how different levels of predictability contribute to predicting syllable duration in spontaneous conversations in Taiwan Mandarin. We investigate whether word-level and syllable-level predictability measurements both contribute to modeling durational variability at the syllable level. Answers to this question may further our understanding of the level(s) of representation relevant in speech production planning, which has been discussed as a potential driving force behind the observed predictability effect [2, 12].

This question is especially relevant for Mandarin, as it is a language with an abundance of standalone monosyllabic morphemes that are strongly associated with specific semantic content. In the context of research on the effects of frequency and contextual probability on phonetic realizations, this characteristic makes it more likely that syllable-level predictability plays an important role in speech planning.

Since the target acoustic variable was syllable duration, we were essentially testing to what extent frequency and probability estimates from a higher level percolate to the durational pattern of a lower-level unit. In addition, we investigated whether word-level predictability effects are uniform across syllables within a word. For example, in the word /tjan^{51} iŋ^{214}/ movie' in the context of the phrase /k^{h}an^{51} # tjan^{51} iŋ^{214}/ watch movies', we examined whether the word-level bigram probability P(movies|watch) predicts the duration of both syllables tjan^{51} and iŋ^{214} in the same way. If word-level predictability influences component syllables equally, it suggests that syllable-level speech planning is closely linked to word-level representations.

Finally, we examined these predictability effects' interactions with final lengthening. This continues a recent thread of studies that look at whether predictability effects are modulated by prosodic effects, which boundary marking is supposedly a part of. The rationale was that the highlighting of less predictable units has been served by boundary marking as part of the overall prosodic structure that mediates the relationship between predictability and acoustic cues [1, 20, 11]. This research question also motivated the selection of syllable duration as the target, since previous studies on final lengthening mostly focus on syllable position [7, 6].

Three types of predictability measurements were examined: unigram surprisal (i.e., lexical frequency), bigram surprisal (i.e., contextual predictability at specific local contexts), and bigram informativity. Following [13, 14], we define bigram informativity as a word/syllable type's average bigram surprisal as estimated from a corpus.

2. METHOD

2.1. Speech corpus

We extracted durational data of Taiwan Mandarin from the Sinica Phone-aligned Chinese Conversational Speech Database (SPCCSD, Sinica File No. 24T-1031221; [19]). The corpus contains 3.5 hours of recordings with boundary annotation at the phone and syllable levels. Additionally, the corpus has word segmentation and part-of-speech tagging based on the CKIP system [4], which maps to a simplified tag set [10]. For research on discourse and prosody, the corpus also has annotations of prosodic units (PU), which are based on paralinguistic signals such as pause and inhalation, and discourse units (DU), which refer to units containing a predicate and its key arguments.

As previous studies have shown that PU boundaries with a match and mismatch to a DU boundary exhibit different acoustic cues (e.g., [3]), we controlled for such potential effects by only looking at PU boundaries that coincided with a DU boundary. We also limited our inquiry to syllables from disyllabic words so that we could see a clear picture of how word- and syllable-level predictability measurements compare in modeling phonetic variability. Overall, the analysis contained 7420 disyllabic words.

2.2. Written corpus and language models

For language modeling, we used the Academia Sinica Balanced Corpus of Mandarin Chinese [8, 4], which has 11M word tokens and 17.5M syllable/character tokens. Trigram language models were trained with modified Kesner-Ney smoothing [5] at the levels of words and syllables using the SRILM toolkit [16, 17]. In addition to training the model to obtain probabilities given the previous context, models that read sentences from the backward direction are also trained to obtain surprisal and informativity given the following context. This was motivated by research using the same method showing that informativity and surprisal given the following context account for more variances in word duration than informativity and surprisal given the previous context [15].

2.3. Variables in analyses

The analyses included five predictability variables: unigram surprisal ($-\log P(x)$), bigram forward/backward surprisal ($-\log P(x|context)$), and bigram forward/backward informativity ($-\sum_{\text{context}} P(\text{context}|x)\log P(x|\text{context})$), which were calculated for two unit types (word, syllable). All these variables were log-transformed (base 10) and normalized.

To examine the effect of syllable positions within a prosodic unit (PU), we included a predictor with four categories: initial, medial, penultimate, and final. We highlighted the penultimate position on top of the final position since it has been reported that Taiwan Mandarin has a robust disyllabic domain for final lengthening [7, 6].

Three other variables are included: Speech rate (a prosodic unit's syllable count divided by its length in milliseconds), syllable position (a syllable's position within a word), and each syllable's baseline duration. Following [18], baseline duration was the prediction of a linear regression model that predicted syllable duration based on the syllable's tone and segments.

3. RESULTS

3.1. Word- vs. syllable-level predictability

Three multivariate mixed-effects models were fit: A model with only word-level predictability measurements (*Word* model), a model with only syllable-level predictability measurements (*Syllable* model), and a model with both types of variables (*Full* model). Two likelihood ratio tests showed that the *Full* model provided a significantly better fit than the *Word* model [$\chi^2(5) = 197.95$, $p < .00001$] and the *Syllable* model [$\chi^2(5) = 193.45$, $p < .00001$].

Table 1 summarizes the *Full* model. The effect of within-PU position shows a penultimate and final lengthening, with the initial position not significantly different from the medial position ($p < .0001$ for all pairs of comparisons except for initial vs. medial). Most of the predictability variables had a positive estimate, i.e., higher surprisal/informativity was associated with longer syllable duration. It is worth noting that even though a few predictability variables had a negative estimate, it was likely a result of suppression, i.e., these variables had strong correlations with variables that predicted the dependent variable better [23]. Additional analyses showed that these variables had positive estimates when they were the only predictability variable in a model (Forward bigram word surp.: 0.15, unigram word surp.: 0.09, backward syll. inf.: 0.12, forward syll. surp.: 0.04).

Table 1: Summary of fixed effects in the mixed-effects model; **W**: word-level; **S**: syllable-level

	β	SE	t	$p(\chi^2)$
(Intercept)	2.02	0.04	49.88	
Baseline dur.	0.24	0.01	21.56	< .0001
Speech Rate	-0.31	0.01	-52.11	< .0001
N. density	-0.05	0.01	-4.10	< .001
W Forward Inf.	0.14	0.01	9.82	< .0001
W Backward Inf.	0.03	0.01	1.90	= .05
W Forward Sur.	-0.05	0.01	-6.55	< .0001
W Backward Sur.	0.07	0.01	5.95	< .0001
W Unigram Sur.	-0.05	0.01	-4.10	< .0001
S Forward Inf.	0.05	0.02	2.80	< .01
S Backward Inf.	-0.03	0.02	-1.91	= .06
S Forward Sur.	-0.04	0.01	-3.89	< .001
S Backward Sur.	0.11	0.01	9.18	< .0001
S Unigram Sur.	0.07	0.02	3.27	< .01
Initial	0.00	0.04	0.10	(< .0001)
Penult.	0.49	0.02	24.35	
Final	0.85	0.03	30.31	
2nd Syll in Word	-0.25	0.02	-11.16	< .0001

3.2. Directionality and boundary effects

For analyses in this section, we ran additional regression models where each model only had a predictability variable [23], and the three-way interaction between the predictability variable and syllable position in a prosodic unit and a word. The *lstrends()* function in the lsmeans [9] package was used for post hoc analyses. We use 'S1' and 'S2' to refer to the 1st and 2nd syllables in a word.

The upper panels in Figure 1 show the effect sizes of word-level bigram surprisal in different positions. Surprisal was only a positive predictor of duration in the medial position in both the forward (S1: β = 0.04, $p < .0001$; S2: $\beta = 0.02$, $p < .05$) and backward direction (S1: $\beta = 0.07$, $p < .0001$; S2: $\beta = 0.21$, $p <$ .0001).

Comparisons of the effect size further show that forward surprisal had a stronger effect in S2 than S1 (β = 0.14, p < .0001), which follows the directionality hypothesis. The opposite trend was observed for backward surprisal, which had a stronger effect on S1, even though the difference did not reach statistical significance.

The lower panels in 1 show the effects of global word-level measurements of predictability. The effect of unigram surprisal was significant for both S1 & S2 within a word in the medial position (S1: β = 0.13, p < .0001; S2: β = 0.07, p < .0001) and for S1 in the penultimate position (β = 0.07, p < .01). The effect of forward informativity was significant at all positions (p < .05 for S1 in the final position and p < .0001 for other positions) except for S2 in the penultimate position. The effect of backward informativity was significant at all positions (p < .0001) except for S2 in the penultimate position and S1 in the final position. In other words, the effect of word-level unigram surprisal was neutralized at the final position similar to bigram surprisal, while the effect of informativity was not.

Comparisons of the effects in S1 and S2 show that unigram surprisal was a significantly or nearly significantly stronger predictor at S1 than S2 at the medial ($\beta = 0.06$, $p < .0001$) and the final position (β = 0.07, p = .05). For forward informativity, the same comparisons were significant in the same direction (medial: $\beta = 0.05$, $p < .0001$; penultimate: $\beta = 0.08$, p < .05). Finally, for backward informativity, the same comparison was only significant in the same direction for the penult position (β = 0.11, p < .01). Overall, these comparisons showed that unigram and informativity generally predicted the duration of S1 better.

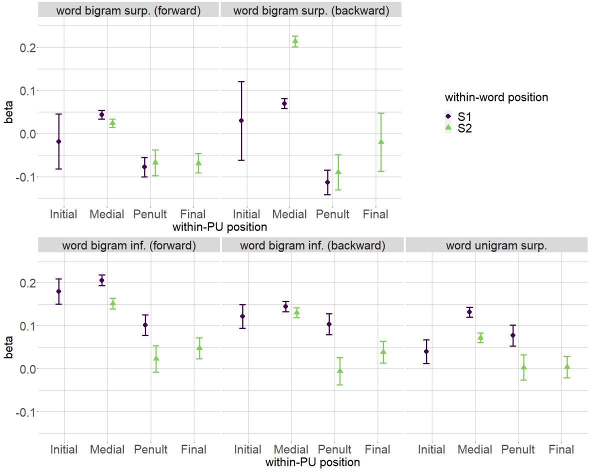

Figure 1: Effect sizes of word-level bigram surprisal (upper panels), informativity, and unigram surprisal (lower panels) at syllable positions within an PU (X axis) and word (color/shape).

Syllable-level bigram surprisal also exhibited a strong directionality effect in terms of its effect size. As shown in the upper panels in Figure 2, at medial and penultimate positions, forward bigram surprisal is a stronger predictor of duration for S2 (medial: p < .0001), whereas backward bigram surprisal was a stronger predictor for S1 (medial: p <

.0001; penult: $p < .001$). In other words, contextual probability conditioned on a syllable within a word had a stronger effect in predicting syllable duration.

On the other hand, as shown in the lower panels in Figure 2, the effect of syllable-level unigram and informativity was either stronger at S1 (backward surprisal at penult: $p < .01$; medial: $p < .0001$) or did not differ between S1 and S2. These two measurements were also a positive predictor of syllable duration at more within-word and within-PU positions (positive estimates with at least $p < .05$ for all except for backward informativity for S2 at the penultimate position) than syllable-level bigram surprisal.

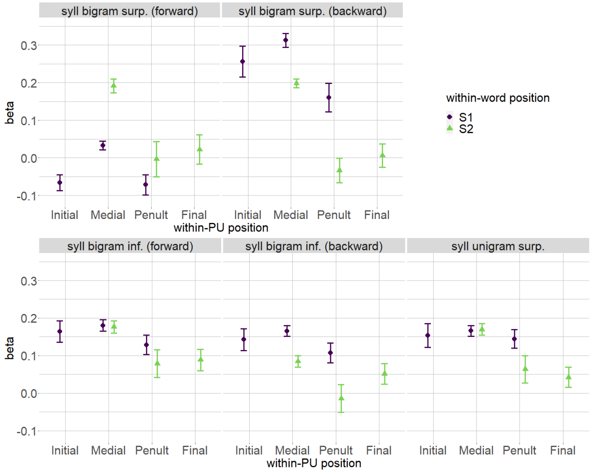

Figure 2: Effect sizes of syllable-level bigram surprisal (upper panels), informativity, and unigram surprisal (lower panels) at syllable positions within an PU (X axis) and word (color/shape).

4. DISCUSSION & CONCLUSION

The results show that having both word-level and syllable-level predictability variables significantly improved the modeling of syllable duration. If the correlation between predictability and durational variability reflects the planning of speech production, this finding suggests that such planning is likely made at both levels of representation. The effect of word-level predictability is especially interesting since it suggests that the phonetic variability at a lower-level unit (syllable) is also affected by word-level statistics.

We further inspected the percolation effect of word-level predictability by examining whether, within a disyllabic word, word-level predictability effects applied uniformly to both syllables. The results show a split between bigram surprisal on one hand and bigram informativity and unigram surprisal on the other. For bigram surprisal, the expected directional effect was found: forward surprisal predicted the duration of the initial syllable better, while backward surprisal predicted the duration of the second syllable better. In other words, word-level bigram surprisal mostly had local effects.

However, word-level informativity and unigram surprisal, which are word-specific measurements that do not vary across a word's occurrences in different local contexts, were more predictive of syllable duration in the initial syllable. It suggests that phonetic variability of a word's first syllable might have been used to signal predictability at the level of word types, i.e., whether the word type is frequent or informative (i.e., often occurring in more predictable contexts).

In addition, we found that syllable-level bigram surprisal is more predictive of syllable duration when conditioned within a word. In other words, even if we focus on how syllable duration and syllable-level predictability correlate, word boundaries are still making an impact, suggesting the crucial role of word-level information (and in terms of data processing for a language like Mandarin Chinese, the role of word segmentation) in the modeling of syllable-level phonetic variability.

Finally, predictability effects (especially for surprisal) were weakened or neutralized at PU boundaries: At the final position, where pre-boundary lengthening was the strongest, and at the initial position, where pitch reset has been found [3]. In other words, there was likely a trade-off between the presence of boundary cues and predictability effects. This finding is consistent with the view that the prosodic structure modulates the relationship between predictability and phonetic variability [1, 20, 11].

To conclude, this study presents a clearer picture of how predictability measurements at different levels affect phonetic variability under different conditions. The results suggest that speakers may be simultaneously tracking two levels of statistical information in speech production. We have further shown that local predictability measurements like surprisal exhibit local effects and are more prone to neutralization in their interaction with prosodic positions, which differs from the behavior of global (i.e., word-type level) measurements such as lexical frequency and informativity.

5. REFERENCES

[1] M. Aylett and A. Turk, "The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic

prominence, and duration in spontaneous speech," *Language and speech*, vol. 47, no. 1, pp. 31–56, 2004.

[2] A. Bell, J. M. Brenier, M. Gregory, C. Girand, and D. Jurafsky, "Predictability effects on durations of content and function words in conversational english," *Journal of Memory and Language*, vol. 60, no. 1, pp. 92–111, 2009.

[3] A. C.-H. Chen and S.-C. Tseng, "Prosodic encoding in Mandarin spontaneous speech: Evidence for clause-based advanced planning in language production," *Journal of Phonetics*, vol. 76, p. 100912, 2019.

[4] K.-J. Chen, C.-R. Huang, L.-P. Chang, and H.-L. Hsu, "Sinica corpus: Design methodology for balanced corpora," in *Proceedings of the 11th Pacific Asia Conference on Language, Information and Computation*, 1996, pp. 167–176.

[5] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Computer Speech & Language*, vol. 13, no. 4, pp. 359–394, 1999.

[6] Fon, J, K. Johnson, and S. Chen, "Durational patterning at syntactic and discourse boundaries in Mandarin spontaneous speech," *Language and speech*, vol. 54, no. 1, pp. 5–32, 2011.

[7] Fon, Y.-J. J, "A cross-linguistic study on syntactic and discourse boundary cues in spontaneous speech," Ph.D. dissertation, The Ohio State University, 2002.

[8] C.-R. Huang and K.-j. Chen, "Academia sinica balanced corpus of modern chinese 4.0," *Academia Sinica*, 2010.

[9] R. V. Lenth, "Least-squares means: The R package lsmeans," *Journal of Statistical Software*, vol. 69, no. 1, pp. 1–33, 2016.

[10] Y.-F. Liu and S.-C. Tseng, "Word use and word-level reduction in story-telling speech of chinese-speaking hearing and hard of hearing children," in *(Dis)fluencies in children's speech*, B. Judit, Ed. Akadémiai Kiadó, 2020.

[11] Z. Malisz, E. Brandt, B. Möbius, Y. M. Oh, and B. Andreeva, "Dimensions of segmental variability: Interaction of prosody and surprisal in six languages," *Frontiers in Communication*, vol. 3, p. 25, 2018.

[12] M. Pluymaekers, M. Ernestus, and R. Baayen, "Articulatory planning is continuous and sensitive to informational redundancy," *Phonetica*, vol. 62, no. 2-4, pp. 146–159, 2005.

[13] U. C. Priva, "Sign and signal: Deriving linguistic generalizations from information utility," Ph.D. dissertation, 2012.

[14] ——, "Informativity affects consonant duration and deletion rates," *Laboratory Phonology*, vol. 6, no. 2, pp. 243–278, 2015.

[15] S. Seyfarth, "Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation," *Cognition*, vol. 133, no. 1, pp. 140–155, 2014.

[16] A. Stolcke, "Srilm-an extensible language modeling toolkit," in *Seventh international conference on spoken language processing*, 2002.

[17] A. Stolcke, J. Zheng, W. Wang, and V. Abrash, "Srilm at sixteen: Update and outlook," in *Proceedings of IEEE automatic speech recognition and understanding workshop*, vol. 5, 2011.

[18] K. Tang and R. Bennett, "Contextual predictability influences word and morpheme duration in a morphologically complex language (kaqchikel mayan)," *The Journal of the Acoustical Society of America*, vol. 144, no. 2, pp. 997–1017, 2018.

[19] S.-C. Tseng, "ILAS Chinese spoken language resources," in *Proc. LPSS*, 2019, pp. 13–20.

[20] A. Turk, "Does prosodic constituency signal relative predictability? a smooth signal redundancy hypothesis," *Laboratory phonology*, vol. 1, no. 2, pp. 227–262, 2010.

[21] R. J. Van Son and J. P. Van Santen, "Duration and spectral balance of intervocalic consonants: A case for efficient communication," *Speech Communication*, vol. 47, no. 1-2, pp. 100–123, 2005.

[22] S.-F. Wang, "The interaction between predictability and pre-boundary lengthening on syllable duration in taiwan southern min," *Phonetica*, vol. 79, no. 4, pp. 315–352, 2022. [Online]. Available: https://doi.org/10.1515/phon-2022-0009

[23] L. H. Wurm and S. A. Fisicaro, "What residualizing predictors in regression analyses does (and what it does not do)," *Journal of memory and language*, vol. 72, pp. 37–48, 2014.