F0-based Pairwise Variability Index: A Prosodic Metric for Holistic Language Processing

Alvin Cheng-Hsien Chen

Department of English, National Taiwan Normal University, Taiwan
alvinchen@ntnu.edu.tw

ABSTRACT

This study examined the effectiveness of a pitch variability metric—f0-based Pairwise Variability Index (f0PVI), which measures the pairwise f0 variability of a speech fragment. Analyzing spontaneous speech production, we asked (a) whether multisyllabic words' f0PVI values are smaller than nonwords'; (b) how multisyllabic words' f0PVI values are connected to lexical frequencies; and (c) to what extent the f0PVI values are conditioned on different representative f0 values of syllables. Multisyllabic words above a frequency cut-off in a native corpus were defined, and a subset of them used in a spontaneous speech corpus were particularly identified for the analyses of the relationship between their f0PVI in the speech corpus and their lexical frequencies in the native corpus. Results show f0PVI effectively differentiated words from nonwords and significantly correlated with lexical frequencies. We discuss the role of f0PVI in uncovering speakers' statistical knowledge in language processing.

Keywords: pitch variability, PVI, frequency effects, usage-based, spontaneous speech

1. INTRODUCTION

Human speech, although rich in forms and combinations, is always subject to the limitation of the biomechanism system. Studies have shown that there seems to be an upper limit (i.e., maximum speed) for human speakers to change their pitch voluntarily [1-3]. Along with the economy of efforts suggested by [4], pitch variation should probably be produced at the time when its pragmatic function is the most readily needed in a communicative setting. For example, [5] observed that pitch change in spontaneous speech production is closely connected to the semantic coherence and completeness of the proposition to be articulated. This suggests that significant pitch change often occurs at the semantic completion of the speaker's intended message. One corollary is that the pitch variation of a holistic linguistic unit is kept to the minimum compared to the pitch variation of a complex structure spanning several fundamental units.

In this study, we take on this connection between pitch variability and holistic language processing, and further investigate whether the pitch variability of a lexical unit is connected to its lexical frequency in use. We present a case study from spontaneous speech production in Taiwan Mandarin.

There are three main goals in this study. First, we experiment with a pitch encoding metric, which is based on the idea of duration-based variability, and validate its effectiveness in measuring the pitch-related variability of a speech segment. Specifically, we extend the computation of the Pairwise Variability Index (PVI), which is used to measure cross-linguistic durational variability, to pitch variability by creating the f0-based Pairwise Variability Index (f0PVI). The duration-based PVI was first proposed by [6] to capture the rhythmic pattern of a speech segment (e.g., a phrase) by analyzing every local durational variation of the composing sub-units (e.g., syllables of a phrase). In this study, we build upon the PVI's ability to capture local variability and explore its potential to measure local pitch variability of a speech segment.

Second, we aim to determine the effectiveness of the f0PVI by comparing the f0PVI values of disyllabic words to those of nonwords' (i.e., any two-syllable combinations that span the word boundary). Our hypothesis is that a holistic unit such as a word should consistently exhibit smaller degrees of pitch variability, reflected in lower f0PVI values, compared to a nonword.

Finally, we will investigate the extent to which the f0PVI values of disyllabic words are associated with their frequency of use. This investigation is particularly significant in language processing because the systematic prosodic encodings contributed by distributional/statistical properties of language can provide strong empirical evidence of the development of usage-based native intuition [5, 7, 8].

This study addresses three questions: (a) whether disyllabic words consistently demonstrate lower degrees of pitch variability than disyllabic nonwords; (b) to what extent the pitch variability of disyllabic words are connected to lexical frequencies; (c) to what extent the analysis of the pitch variability metric (f0PVI values) is dependent on the selection of different representative f0 values. To achieve these objectives, we selected disyllabic words above a frequency threshold in a native corpus of Taiwan Mandarin and identified a subset of these words for analysis of the relationship between their f0PVI in a spontaneous speech corpus and their lexical frequencies in the native corpus.

2. METHOD

2.1. Data

The native corpus used in this study was the Academia Sinica Balanced Corpus of Mandarin Chinese (Sinica Corpus) [9], an eleven-million-word text collection tht covers a wide range of topics and genres. The speech data for pitch analysis came from the Sinica Phone-Aligned Chinese Conversational Speech Database (SPCCSD) [10], which includes approximately 3.5 hours of face-to-face free conversations from 16 speakers (7 males and 9 females; age range: 16–46). The SPCCSD has been time-aligned at the segmental level by the corpus provider, which allows for acoustic and prosodic analysis.

2.2. Disyllabic words

Due to the prevalence of disyllabic words in Mandarin (i.e., words consisting of two Chinese characters, e.g., *tongcai* 同儕 'peer', *shiyie* 事業 'career'), they are ideal candidates for pitch variability analysis. We selected all disyllabic words with frequency values above 5 in the Sinica Corpus as potential units for pitch analysis. Subsequently, a subset of these words was identified in the speech corpus of the SPCCSD for the purpose of examining the relationship between their f0PVI values in the speech corpus and their lexical frequencies in the native corpus.

2.3. Pitch variability metric: f0PVI

In this study, we are concerned with pitch variability of a speech fragment, which refers to the f0 variations of the base units within the speech fragment. In this section, we outline the method used to compute the pitch-related variability of disyllabic words observed in the SPCCSD.

First, we extracted the raw f0 values of each syllable using Praat's autocorrelation-based pitch tracking algorithm, with gender-dependent pitch ranges (75-300 Hz for males, and 100-500 Hz for females). The hertz values were converted into semitones using a 50 Hz base. Second, we computed the pairwise local f0 variability by selecting a representative f0 value from each syllable. To determine a representative f0 value, we explored several alternatives, including (a) the f0 value at the maximum energy (f0dbmax), (b) the maximum f0 value (f0max), (c) the minimum f0 value (f0min), and (d) the mean f0 (f0mean) value of the syllable. The f0dbmax may deserve more explanation. It is important to note that the f0dbmax represents the f0 value at the time point of the maximal decibel value in the syllable. This energy-based extraction of f0 values has been shown to produce more reliable f0 values and reduce the likelihood of tracking errors [5, 11, 12]. For our analysis of the connection between pitch variability and lexical frequencies, we will use the f0dbmax as the basis. However, the other f0 values (i.e., f0max, f0min, and f0mean) will be used for critical comparison of their impact on the f0PVI computation.

The computation of the words' pitch variability was based on the PVI proposed by [6] for the study of the typological rhythmic patterns. This duration-based PVI was first proposed as a quantitative metric to measure the isochrony and/or rhythm of syllables in languages and help determine the timing patterns (e.g., syllable-timing or stress-timing) of languages [13]. The duration-based PVI in [6] calculates the differences of durations between successive pairs of phonological units. For example, given a speech segment of *n* syllables, we first need to obtain the duration values of all the syllables (i.e., D_k), sum all the durational differences between each successive pair of syllables (i.e., $\sum_{k=2}^{n}|D_\mathrm{k} - D_{k-1}|$), and compute the mean difference, as formulated in (1). Because the PVI based on (1) can potentially be scaled up due to the slow speech rates, [13] introduced a normalization mechanism to mitigate the influence of speech rates, as formulated in (2).

(1) Duration-based PVI $= \frac{[\sum_{k=2}^{n}|D_k - D_{k-1}|]}{n-1}$

(2) Normalized Duration-based PVI

$$= 100 \times \left[\frac{\sum_{k=2}^{n}\left|\frac{D_k - D_{k-1}}{(D_k + D_{k-1})/2}\right|}{n-1}\right]$$

In this study, we operationalized an f0-based normalized PVI (f0PVI) to quantify the pitch variability of a speech segment. The f0PVI in (3) captures the variability of f0 differences between each successive pair of phonological units (i.e., syllables) in a speech fragment:

(3) f0PVI $= 100 \times \left[\frac{\sum_{k=2}^{n}\left|\frac{f0_k - f0_{k-1}}{(f0_k + f0_{k-1})/2}\right|}{n-1}\right]$

In contrast, the earlier duration-based PVI in (2) measures the variability of durational differences between syllables. The key trick was to determine a representative f0 value for each syllable within the speech segment. In this study, we computed the f0PVI values based on four different f0 values extracted from each syllable of the disyllabic word: f0dbmax, f0max, f0min, and f0mean, which are represented by $f0_k$ in (3). The f0PVI is an intuitive and informative prosodic metric that quantifies the degree of pitch variability in a speech fragment. Higher f0PVI values indicate greater pitch variability (less regularity), while lower f0PVI values suggest a more monotonous f0 variation within the fragment.

2.4. Research design

The working hypothesis is that a holistic unit should be processed more cohesively, thus demonstrating smaller pitch variability (i.e., smaller f0PVI values) when compared to non-holistic units. To test this hypothesis, we carried out three analyses to determine the effectiveness of f0PVI and its connection to holistic language processing. The first two experiments used the f0PVI values computed based on the f0dbmax.

In Analysis 1, we examined the effectiveness of f0PVI in discriminating between disyllabic words and nonwords. Specifically, we compared the f0PVI values of all the disyllabic words with those of disyllabic nonwords (i.e., all nonword character bigrams) observed in the SPCCSD. If f0PVI is indicative of the extent of holistic processing by speakers, we would expect the f0PVI values of disyllabic words to be significantly lower than those of nonword character bigrams.

In Analysis 2, we investigated the relationship between the disyllabic words' f0PVI values and their lexical frequencies. We posited that as the lexical frequency of a disyllabic word increases, it is more likely to be processed holistically, resulting in a smaller f0PVI value.

In Analysis 3, we assessed the potential impact of different representative f0 values used in f0PVI computation on the results of the first two analyses. In particular, we investigated whether using f0max, f0min, or f0mean instead of f0dbmax for computing f0PVI would lead to different generalizations in distinguishing disyllabic words from nonwords and the relationship between f0PVI values and lexical frequencies of disyllabic words.

2.5. Statistical analysis

A linear mixed-effect analysis was used in this study with the f0PVI value as the dependent variable and the disyllabic wordness (in Analysis 1) and (log-transformed) lexical frequency (in Analysis 2) as the experimental fixed effects. Speaker-based intercepts were included as the random variable. In a spontaneous speech setting, pitch variability can arise from several factors, such as segmental structures, syllabic structures, relative position in utterance, speech rates, prosodic phrasing, and information structure. Therefore, we included several essential control factors in the mixed-effect model: (a) the relative position of the word in a speaker turn; (b) whether the word is at the onset/end of an intonation unit or a pause; and (c) the phonemic length of the word. We compared models using likelihood ratio tests to examine the significance of experimental fixed effects.

3. RESULTS

Disyllabic words and nonwords that could not be extracted with reliable f0 values due to speech reduction were removed from analysis. After filtering, we identified 12,824 tokens of disyllabic words in use from the speech corpus of SPCCSD, encompassing 2,246 word types. In SPCCSD, every character bigram spanning the word boundary formed a potential disyllabic nonword, amounting to 12,109 tokens (7,151 types).

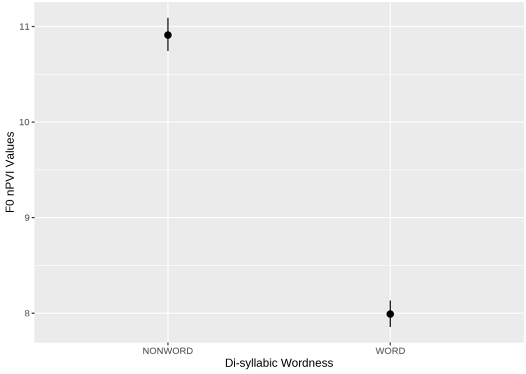

Figure 1: F0PVI means and confidence intervals of disyllabic words and nonwords

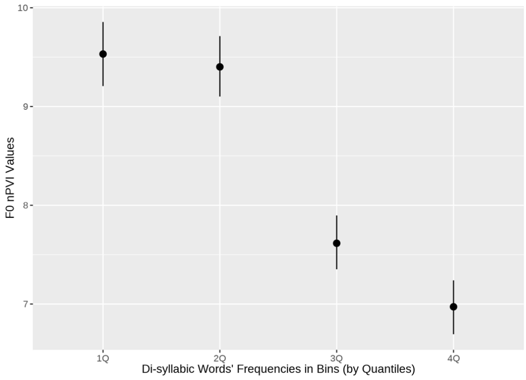

Figure 2: F0PVI means and confidence intervals of disyllabic words by lexical frequency bins

The results of Analysis 1 show that disyllabic words' f0PVI values were significantly lower than nonwords ($\chi^2 = 651.15$, $p < .01$), as shown in Figure 1.

This finding provides evidence that f0PVI can serve as an indicator of segmental integrity and a holistic unit (e.g., a disyllabic word) exhibits consistently lower pitch variability.

In Analysis 2, a significant correlation was observed between di-syllabic words' f0PVI values and their frequency values ($\chi^2 = 215.53, p < .01$). Figure 2 depicts the frequency effects, displaying the f0PVI mean values for disyllabic words divided into four frequency bins based on quantiles. The findings suggest that there is a negative relationship between word frequency and pitch variability, with higher frequency words showing smaller pitch variability.

In the final analysis, we examined whether the choice of representative f0 values for f0PVI computation could impact the results obtained in the previous analyses. Specifically, we assessed whether using f0max, f0min, and f0mean instead of f0dbmax to compute f0PVI values would lead to different generalizations in Analyses 1 and 2. However, the results showed no significant differences, reinforcing the robustness of the findings from the previous two analyses.

4. DISCUSSIONS

Holistic processing is a crucial area of study in psycholinguistics because it has significant implications for the psycholinguistic validity of a linguistic structure. To investigate holistic processing, researchers often examine the durational patterns of linguistic units. Reduction, in particular, is a frequently observed prosodic cue that reflects speakers' holistic processing of entrenched linguistic units [7, 14]. High-frequency words are produced and/or processed faster than low-frequency words, providing evidence for frequency-based effects on duration. This effect has also been observed with multiword units, including idioms [15], binomial phrases [16], collocations [17, 18], and even compositional multiword sequences [19-22].

On the other hand, pitch variation has more often been analyzed as a prosodic cue for discourse disjuncture at various levels [5], and its connection to the idea of holistic processing is relatively underexplored. Although human speech is rich in forms and combinations, it is always subject to the constraints of the biomechanism system.

Particularly relevant to the present study is an upper limit (i.e., maximum speed) for human speakers to voluntarily modulate their pitch [1-3]. Additionally, in accordance with the economy of effort principle proposed by [4], stronger pitch variation should be more likely to be produced only when its pragmatic function is most readily needed in a communicative setting. For example, [5] has shown that pitch change in speech production is closely connected to the semantic coherence and completeness of the proposition to be articulated. Significant pitch change occurs at important semantic boundaries of the speaker's intended message. A corollary of this is that the pitch variability of non-compositional basic linguistic units such as words should be minimized, optimizing communicative needs and the ease of articulatory efforts.

This study has developed an intuitive pitch variability index—f0PVI, which effectively distinguishes between (a) words and nonwords and (b) words at different frequency ranges. Although f0PVI calculation relies on a representative f0 value from the base units (i.e., syllables) of the speech segment (i.e., disyllabic word), our results demonstrate that f0 selection has no significant effect on the strong relationship between f0PVI and lexical frequency.

Our findings are consistent with and extend previous usage-based research by highlighting pitch variability as a strategic cue that reflects speakers' sensitivity to distributional properties in language use [5]. The pitch variability index can serve as a prosodic indicator of the degree of entrenchment of a linguistic unit in language processing. As a lexical unit is repeatedly used, it becomes progressively entrenched to the point of being processed holistically. The inverse relationship between a word's pitch variability in production and its frequency of use provides clear empirical support for this usage-based grammatical competence. In addition, the straightforward and intuitive nature of f0PVI suggests its potential for application in analyzing pitch variation with other more complex linguistic structures.

5. CONCLUSIONS

This study makes significant contributions in three key areas. First, we have introduced and validated a pitch variability metric for characterizing the holistic processing of linguistic units. Second, our focus on pitch-related encodings provides a novel and complementary perspective on the frequency effects observed in usage-based research. Finally, our analysis of spontaneous speech highlights the critical role of speakers' usage-based competence in online speech production and has important implications for psycholinguistic research.

6. ACKNOWLEDGEMENTS

This work was partly supported by National Science and Technology Council, Taiwan (110-2410-H-003-014 and 111-2410-H-003-030).

7. REFERENCES

[1] Y. Xu and X. Sun, "Maximum speed of pitch change and how it may relate to speech," *The Journal of the Acoustical Society of America,* vol. 111, pp. 1399-1413, 2002.

[2] J. Sundberg, "Maximum speed of pitch changes in singers and untrained subjects," *Journal of Phonetics,* vol. 7, no. 2, pp. 71-79, 1979.

[3] J. J. Ohala and W. G. Ewan, "Speed of pitch change," *The Journal of the Acoustical Society of America,* vol. 53, no. 1, pp. 345-345, 1973.

[4] B. Lindblom, "Economy of speech gestures," in *The production of speech*, P. F. MacNeilage Ed. New York, NY: Springer, 1983, pp. 217-245.

[5] 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, pp. 1-22, 2019, doi: 10.1016/j.wocn.2019.100912.

[6] E. L. Low, E. Grabe, and F. Nolan, "Quantitative characterizations of speech rhythm: Syllable-timing in Singapore English," *Language and Speech,* vol. 43, no. 4, pp. 377-401, 2000.

[7] A. C.-H. Chen, "Durational patterns of recurrent multiword combinations in Mandarin spontaneous speech production," *Language and Speech,* vol. 64, no. 3, pp. 742-767, 2021, doi: 10.1177/0023830920966010.

[8] P. M. S. Lin, "The prosody of formulaic expression in the IBM/Lancaster Spoken English Corpus," *International Journal of Corpus Linguistics,* Article vol. 18, pp. 561-588, 2013, doi: 10.1075/ijcl.18.4.05lin.

[9] C.-R. Huang and K.-j. Chen, *Academia Sinica Balanced Corpus of Modern Chinese 4.0.* Taipei, Taiwan: Academia Sinica, 2010.

[10] S.-C. Tseng, "Lexical Coverage in Taiwan Mandarin Conversation," *International Journal of Computational Linguistics and Chinese Language Processing,* vol. 18, no. 1, pp. 1-18, 2013.

[11] D. R. Ladd and C. Johnson, "Metrical factors in the scaling of sentence-initial accent peaks," *Phonetica,* vol. 44, pp. 238-245, 1987, doi: 10.1159/000261801.

[12] E. Grabe, G. Kochanski, and J. Coleman, "Connecting intonation labels to mathematical descriptions of fundamental frequency," *Language and Speech,* vol. 50, pp. 281-310, 2007, doi: 10.1177/00238309070500030101.

[13] E. Grabe and E. L. Low, "Durational variability in speech and the rhythm class hypothesis," in *Papers in Laboratory Phonology 7*, C. Gussenhoven and N. Warner Eds. Hague: Mouton de Gruyter, 2002, pp. 515-546.

[14] N. C. Ellis, "Frequency effects in language processing," *Studies in Second Language Acquisition,* vol. 24, pp. 143-188, 2002, doi: 10.1017.S0272263102002024.

[15] A. Siyanova-Chanturia, K. Conklin, and N. Schmitt, "Adding more fuel to the fire: An eye-tracking study of idiom processing by native and non-native speakers," *Second Language Research,* vol. 27, pp. 251-272, 2011.

[16] A. Siyanova-Chanturia, K. Conklin, and W. J. B. van Heuven, "Seeing a phrase 'time and again' matters: The role of phrasal frequency in the processing of multiword sequences," *Journal of Experimental Psychology: Human Learning and Memory,* vol. 37, pp. 776-784, 2011, doi: 10.1037/a0022531.

[17] P. Durrant and A. Doherty, "Are high-frequency collocations psychologically real? Investigating the thesis of collocational priming," *Corpus Linguistics and Linguistic Theory,* vol. 6, pp. 125-155, 2010, doi: 10.1515/cllt.2010.006

[18] S. Sonbul, "Fatal mistake, awful mistake, or extreme mistake? Frequency effects on off-line/on-line collocational processing," *Bilingualism: Language and Cognition,* vol. 18, pp. 419-437, 2015, doi: 10.1017/S1366728914000674.

[19] I. Arnon and N. Snider, "More than words: Frequency effects for multi-word phrases," *Journal of Memory and Language,* vol. 62, pp. 67-82, 2010, doi: 10.1016/j.jml.2009.09.005.

[20] C. Bannard and D. Matthews, "Stored word sequences in language learning: The effect of familiarity on children's repetition of four-word combinations," *Psychological Science,* vol. 19, pp. 241-248, 2008, doi: 10.1111/j.1467-9280.2008.02075.x.

[21] A. Tremblay and H. Baayen, "Holistic processing of regular four-word sequences: A behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall," in *Perspectives on formulaic language: Acquisition and communication*, D. Wood Ed. London: The Continuum International Publishing Group, 2010, pp. 151-173.

[22] A. Tremblay, B. Derwing, G. Libben, and C. Westbury, "Processing advantages of lexical bundles: Evidence from self-paced reading and sentence recall tasks," *Language Learning,* vol. 61, pp. 569-613, 2011, doi: 10.1111/j.1467-9922.2010.00622.x