10th International Conference on Speech Prosody 2020
25-28 May 2020, Tokyo, Japan

Pitch peak and word predictability: Results from CSJ corpus

Daiki Hashimoto[1]

[1]Joetsu University of Education
daiki@phiz.c.u-tokyo.ac.jp

Abstract

It has been widely demonstrated that a word is pronounced with lower phonetic redundancy when it has higher contextual predictability. For example, when a word is predictable given a preceding word and when a word has higher contextual predictability given a following word, it is pronounced with shorter duration [1, 2]. Likewise, words with higher contextual predictability are produced with centralized formant values [3]. This probability-oriented reduction is known as “probabilistic reduction [4].”

This phenomenon can neatly be captured by Message-Oriented Phonology (MOP) [5]. MOP hypothesizes that a speaker balances the efficiency and accuracy of message transmission. When a word is contextually predictable, it can be conveyed successfully to an addressee, the result of which is that the speaker improves the efficiency of the message transmission. On the other hand, when a word is less predictable, the message transmission is more likely to fail, and thus a speaker needs to invest more resource cost in a speech signal, with the result that the phonetic redundancy is increased.

The aim of this study is to explore whether probabilistic reduction can be extended to pitch values. Most previous literature discusses probabilistic reduction in relation to word duration, so therefore, to the best of my knowledge, this study is the first study to investigate the relationship between pitch values and contextual predictability of a word in natural speech. It will be demonstrated that a word is pronounced with a higher pitch value, when it is less predictable given a preceding word. On the other hand, the higher predictability of a word given a following word leads to a higher pitch value of the word.

Index Terms: speech corpus, pitch peak, predictability, phonetic redundancy, Message-Oriented Phonology

1. Introduction

Previous literature has demonstrated that the phonetic realization of a word is affected by the contextual predictability. For example, a word with higher contextual predictability given a preceding word (i.e., $p(\mathrm{w_i}|\mathrm{w_{i-1}})$) and one with higher contextual predictability given a following word (i.e., $p(\mathrm{w_i}|\mathrm{w_{i+1}})$) are produced with shorter duration [1, 2]. Similarly, a word is pronounced with a centralized formant value when it is more predictable in given contexts in comparison with when it has lower contextual predictability [3]. This phonetic reduction of a word as a function of the contextual predictability is known as “probabilistic reduction [4].”

Message-Oriented Phonology (MOP) gives a theoretical account for probabilistic reduction. In this theoretical framework, a speaker is hypothesized to use “the right sort of redundancy” to avoid “inefficient redundancy,” and thereby augment “the likelihood of sufficiently accurate and cost-effective message transmission. [5]” That is, a speaker feels pressure to maximize the accuracy of the message transmission and pressure to minimize resource cost in the message transmission. These two pressures usually conflict each other, and thus a speaker needs to balance the two pressures. For example, a hyper-articulated form such as [wiːt^h] may increase the accuracy of the message transmission because it has clear phonetic cues to each segment in the word and it is more likely to be recognized by an addressee. However, this pronunciation incurs higher redundancy in the articulation, and thus the efficiency of the message transmission may be decreased. On the other hand, a hypo-articulated form such as [wiːʔ] may decrease the accuracy of the message transmission. This is because it has an ambiguous phonetic cue to identify the final segment, and it may be confounded with “week.” At the same time, this relaxed pronunciation may increase the efficiency of the message transmission, because a speaker does not need to make an effort to burst the final consonant.

A message with higher contextual predictability is not so important as that with lower contextual predictability. This is because a listener can reconstruct the message with higher contextual predictability on the basis of the context, even if the listener misses it in speech. For example, the message “pepper” may be inferred given a context such as “salt and” more easily as compared with given a context such as “please pass me.” Hence, the speaker adds more redundancy to a message with lower contextual predictability, so that the speaker can make sure that the important message is successfully conveyed to a listener; the speaker decreases redundancy in a message with higher contextual predictability, the result of which is that the speaker can improve the efficiency in the message transmission. This balancing mechanism accounts for the observation that the duration of a word is a function of the contextual predictability. A speaker improves the efficiency of the message transmission by reducing articulatory effort in the pronunciation of the unimportant message (i.e., message with higher predictability), with the result that the phonetic signal becomes shorter; a speaker improves the accuracy of the message transmission by increasing phonetic redundancy in the pronunciation of the important message (i.e., message with lower predictability), the result of which is that the phonetic signal becomes longer.

Probabilistic reduction at the word level has been discussed mainly in relation to word duration [1, 2], and occasionally with regards to formant values [3]. To the best of my knowledge, this phenomenon has not been discussed with respect to pitch values in natural speech, although it was explored to some extent in an experimental setting [6]. The current paper is intended as a first step in this direction. Filling the gap increases our understanding of how a speaker balances the accuracy and the efficiency in communication. The aim of this paper is to address the following question:

[RQ] Is the pitch value of a word affected by the contextual predictability?

10.21437/SpeechProsody.2020-91

This paper consists of five sections. Section 2 puts forward a theoretical prediction on the basis of MOP. Section 3 explains how the data were collected from the Corpus of Spontaneous Japanese (CSJ) [7], and Section 4 statistically analyses the data. Section 5 concludes the findings and summarizes this study.

2. Theoretical prediction

This section aims to deduce a prediction with regards to the above research question. As reviewed in the preceding section, a speaker is hypothesized to balance a bias to improve the message transmission accuracy and a bias to minimize resource cost in the message transmission. This is why a speaker is likely to add more redundancy to an important message, that is, a word with lower contextual predictability, and thereby the signal becomes more intelligible and the accuracy of message transmission is improved. As demonstrated in previous literature, a speaker may improve the intelligibility of a word by producing the word with longer duration. The current study aims to test whether this probabilistic reduction extends to another phonetic cue, the pitch value, in natural speech.

One of the key hypotheses in relation to the research question is that "the intelligibility of a word may also be improved by raising the pitch value." It is known that human beings can perceive a higher pitched sound easily in comparison with a lower pitched sound as long as the pitch value is around 4,000 Hz or lower. This effect is known as the Fletcher-Munson curve [8]. Hence, a speaker may be able to improve the accuracy of message transmission by producing a word with a higher pitch value in oral speech. Note that the pitch range is usually around 120Hz for men and 210Hz for women. In addition to the hypothesis about high pitch salience, it can be hypothesized that "raising the pitch value incurs higher resource cost due to the articulatory effort." This is because a speaker is required to tighten her vocal cord further and increase the rate of the vibration to raise the pitch value of a word [9]. Taken together, the following general prediction can be made:

[Prediction] A word with lower contextual predictability may be produced with a higher pitch value, so that the message transmission accuracy may be improved.

3. Methodology

The aim of this section is to outline the methodology employed in the current study to test the prediction. Section 3.1 explains the corpus data collected from CSJ-Core [7], Section 3.2 illustrates how the corpus data was converted to statistical variables.

3.1. Corpus data

The current study explores academic presentation and simulated public speech stored in the corpus called CSJ-Core. They are fully annotated with TextGrid files [10]. The annotation includes a variety of information such as word intervals, utterance intervals, and prosodic break points. Using a Praat script [10], the wav files and TextGrid files were converted into a csv file, with the result that 372,631 word tokens were collected along with not only annotated information but also phonetic information such as duration and a pitch value.

The current study explores only word tokens that appear at the beginning of the Intonation Phrase (IP). This is because Japanese has a downward shift from the beginning of IP to the end of IP similar to other languages [11]. Hence, the pitch value of a word highly differs depending on the location within IP. In order to explore the effect of contextual predictability on the pitch value independently of the downward pitch shift, we decided to control the location within IP.[1] After removing the non-IP-initial word tokens, 54,814 tokens remain in our dataset.

3.2. Calculation of variables

As our interest lies in the pitch value of a word, we measured this when the Praat script was running the wav files. The peak value of a pitch within a word interval was retrieved. This numeric variable is called *pitchPeak*, and will be fitted as a dependent variable in the following statistical analyses. The distribution is shown in Figure 1, with the dashed line showing the mean value. The mean value is 216.47Hz, the median is 211.68Hz, and the standard deviation is 68.64.

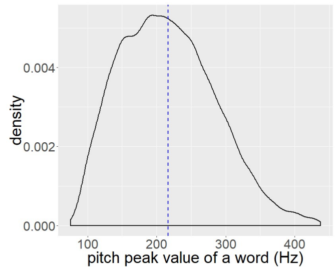

Figure 1: *Pitch values of IP-initial word tokens.*

The variable of interest is contextual predictability, as in our research question and prediction. Following previous literature, we calculated two types of contextual predictability: predictability given a preceding word (i.e., $p(w_i|w_{i-1})$) and predictability given a following word (i.e., $p(w_i|w_{i+1})$). For example, the predictability of a word given a preceding word can be calculated by dividing the frequency of the combination of the word plus the preceding word by the frequency of the preceding word in the corpus (i.e., $N(w_{i-1}\cap w_i) / N(w_{i-1})$). For a statistical reason, they were transformed into information content (IC) by taking $-\log_2$. These probabilistic values were obtained in accordance with the equations (1) and (2):

$$\mathrm{IC}(w_i|w_{i-1}) = -log_2(p(w_i|w_{i-1})) \tag{1}$$

$$\mathrm{IC}(w_i|w_{i+1}) = -log_2(p(w_i|w_{i+1})) \tag{2}$$

The former type of information content is called *precIC*, and the latter type is called *folIC*. Note that the raw predictability

[1] It may be possible to fit a variable that represents the location of a word within IP into a statistical model along with the contextual predictability, so that the effect of the location within IP can be controlled statistically. However, this requires additional annotation to measure the variable. Hence, this preliminary work controls the variable solely by removing non-IP-initial word tokens from the data set. Our future research will measure and fit the variable as a control variable, which enables us to explore the whole data including non-IP-initial tokens as well as IP-initial tokens.

and the information content negatively correlate, that is, lower predictability results in higher information content, as illustrated in Figure 2.

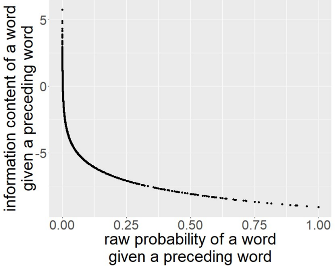

Figure 2: *The relationship between raw probability $p(w_i|w_{i-1})$ and information content $IC(w_i|w_{i-1})$ in our dataset.*

In addition to these variables of interest, some variables that may affect a pitch value were also explored as control variables. First, we explored word frequency, which is known to influence the degree of phonetic reduction. This variable was calculated by counting the number of word tokens within the corpus and transformed with base e log, which is the default in R [12]. This variable is called *logFreq*. If a pitch value also shows a reductive tendency, it is expected that a high-frequency word is realized with a lower pitch value.

Second, *gender* is well-known to affect a pitch value. As noted above, the pitch range in female speech is around 210Hz, while that in male speech is around 120Hz. The pitch value in female speech should be higher than that in male speech.

Table 1: *Fixed effects in statistical analyses.*

Fixed effects	Description
precIC	[numeric] Information content of a word given a preceding word (mean = 9.09; median = 9.47; sd = 3.42)
folIC	[numeric] Information content of a word given a following word (mean = 6.96; median = 6.8; sd = 3.93)
logFreq	[numeric] Log-transformed word frequency (mean = 4.28; median = 4.34; sd = 2.29)
gender	[binary] Speaker is female vs. male (female: 25,596 / male: 28,354)
speech	[binary] Speech is academic presentation vs. simulated public speech (academic: 24,448 / simulated: 39,502)
speechRate	[numeric] Number of morae per second within an utterance (mean = 8.92; median = 8.85; sd = 1.58)
accented	[binary] Word is lexically accented vs. unaccented. (yes: 28,321 / no: 25,629)

Next, we also explored the effect of speech type, since the corpus includes academic presentation and simulated public speech. This binary variable is called *speech*.

The effect of speech rate may also influence a pitch value. This variable is calculated by dividing the number of morae within an utterance by the duration of the utterance. This numeric variable is called *speechRate*. If a pitch has reductive nature, it may be lower in fast speech.

Finally, the effect of a lexical accent is also fitted into a statistical model. This is suggested by an anonymous reviewer. In Japanese, some words are lexically accented, while the others are unaccented. This lexical property is treated as a binary variable called *accented*.

To wrap up this section, the fixed effects explored in the following statistical analysis can be summarized as in Table 1. An anonymous reviewer recommended removing outliers. Following their suggestion, if a value was more than 3 SD away from the mean of a numeric variable, it was removed from the dataset. In all, 53,950 tokens will be explored in what follows. Note that every numeric variable will be centred on the mean.

4. Statistical analyses

The purpose of this section is to describe the statistical analyses used in this research. The 53,950 IP-initial word tokens were hand-fitted into a mixed-effects linear regression model. This analysis is performed using the *lmer* function in the *lme4* library [13]. The seven fixed effects (*precIC*, *folIC*, *logFreq*, *gender*, *speech*, *speechRate*, and *accented*) were fitted into a model alongside with by-speaker and by-word random intercepts and by-word and by-speaker random slopes for the variables of interest (*precIC* and *folIC*). According to the model comparison, every predictor has achieved a significant level ($p > 0.05$). The model is shown in Table 2. The reference level is set as *gender* female, *speech* academic, and *accented* no. The VIF values of the predictors were checked, and they are around 1.0, which suggests that this model has no multicollinearity problem. According to the MuMIn package [14], the marginal R squared value is 0.403, and the conditional R squared value is 0.667.

Table 2: *Best-fitted model to predict a pitch peak value of an IP-initial word.*

	β	***SE***	***t***
(Intercept)	257.38	4.13	62.25
precIC	0.29	0.08	3.44
folIC	-0.26	0.07	-3.65
logFreq	-1.77	0.16	-10.52
gender male	-77.83	3.91	-19.86
speech simulated	-17.52	3.99	-4.38
speechRate	-2.33	0.12	-18.36
accented yes	12.34	0.56	21.71

Now, let us discuss how a pitch value of an IP-initial word is affected by the fixed effects. The intercept tells us that the pitch peak value of a word is 257.38Hz ± 4.13, when the speaker is female, the speech is academic, the word is lexically unaccented, and the three numeric variables are mean values. The variable called *precIC* shows that the pitch peak value is increased by about 0.29Hz, when the information content given a preceding word increases by 1 bit. On the other hand, the variable called *folIC* indicates that the pitch peak value is decreased by 0.26, as the information content given a following word increases by 1 bit. The significance of *logFreq* tells us that a high frequency word is phonetically realized with a lower pitch. The variable *gender* suggests that the pitch value of a word produced by a male speaker is 77.83Hz lower than the pitch value of a word produced by a female speaker. The best-fitted model also shows that the pitch value of a word is lower in simulated public speech than in academic presentation (*speech*), it is lower in fast speech (*speechRate*), and it is higher in a lexically accented word (*accented*). Note that these seven variables significantly affect the pitch value of a word independently of each other, as they are fitted into a single model and reach significance with lower VIF values.

5. Conclusions

This final section aims to discuss the results with regards to the pitch value of a word in our corpus study. First, let us consider the control variables. As expected, a high frequency word is produced with lower pitch (*logFreq*). This may be because a high frequency word is more likely to be subject to phonetic reduction and it may be produced by a less tightened vocal cord. It was also found that the pitch value is lower in fast speech (*speechRate*). Fast speech is known to lead to phonetic reduction, as a speaker cannot spend adequate time to articulate speech sounds. These two observations suggest that phonetic reduction may cause lower pitch sound waves. Our statistical analyses also show that a pitch value is affected by the type of a speaker (*gender*) and the type of speech manner (*speech*). It was also demonstrated that a lexically accented word is phonetically realized with a higher pitch peak as compared with a lexically unaccented word (*accented*).

Now, let us discuss the main findings of this study in relation to our research question and prediction. In Section 2, we deduced the following prediction based on MOP: A word with lower contextual predictability may be produced with a higher pitch value, so that the message transmission accuracy may be improved. In order to test this prediction, we explored two types of contextual predictability, which are converted into information content: $IC(w_i|w_{i-1})$ and $IC(w_i|w_{i+1})$. It was found that the two types of information content both have significant effects on the pitch peak value of a word, and the directions are different from each other, i.e., higher information content given a preceding word (i.e., lower contextual predictability) leads to higher pitch, whereas higher information content given a following word leads to lower pitch. These two opposite effects are illustrated in Figure 3. The y-axis indicates a predicted value of a pitch peak, and the x-axis indicates a centred value of information content. The magenta line represents the coefficient of *precIC* in the best-fitted model (Table 2), and the sky blue line represents the coefficient of *folIC*.

The effect of *precIC* can be neatly captured by MOP. As noted at the beginning of this paper, this theory hypothesizes that a speaker balances a bias to maximize the accuracy of message transmission and a bias to minimize the resource cost in speech. Hence, a speaker adds more redundancy to a speech signal and increases the intelligibility when the message is important, the result of which is that the accuracy of the message transmission can be improved. On the other hand, a speaker decreases redundancy in a speech signal and decreases the articulatory effort when the message is not important, with the result that the efficiency of the message transmission can be improved. A message with higher contextual predictability is considered to be less important because it can be reconstructed from the context. Taken together, a speaker should invest higher resource cost in the production of a word with lower predictability, and lower resource cost in the production of a word with higher predictability. As hypothesized in Section 2, raising the pitch value may increase the intelligibility of a word, because a higher pitch in speech is generally more salient as compared with a lower pitch. Likewise, raising the pitch may lead to more articulatory effort, because a speaker is required to tighten his/her vocal cords further. When coupled with the MOP hypothesis, this hypothesis accounts for why the pitch value of a word is a negative function of contextual predictability given a preceding word. A speaker raises the pitch value of a word with lower contextual predictability in order to increase the intelligibility of a word, because it is important in the message transmission. By doing so, a speaker can make sure that an important message is conveyed successfully. On the other hand, a speaker lowers the pitch value of a word with higher contextual predictability in order to improve the ease of articulation. This results in efficient message transmission.

The effect of *folIC* seems difficult to interpret using MOP, because the effect went in the opposite direction than predicted. An anonymous reviewer pointed out that a following word should be more closely related to a target word in comparison with a preceding word, as the current study explores only the IP-initial words, of which preceding words belong to a different IP and are unlikely to be syntactically related to the target words. Due to the strong connection between a target word and a following word, the message-oriented biases may be cancelled out. There may be some interaction between message predictability and syntactic grouping. This speculation needs to be tested by taking into account pitch values of words appearing at the beginning of IP and those appearing at the middle of IP alongside each other. This effect is left for future study, which will explore the whole data including non-IP-initial tokens as well as IP-initial tokens.

Most previous literature about probabilistic reduction has discussed the duration of a word in general, but our study suggests that other phonetic properties should also be explored in relation to the predictability of a word. In particular, this study has demonstrated that a speaker changes the pitch value of a word depending on the contextual predictability. Some of our findings are amenable to MOP, whereas the others are not. This suggests that the message-oriented biases may be more complicated than previously thought.

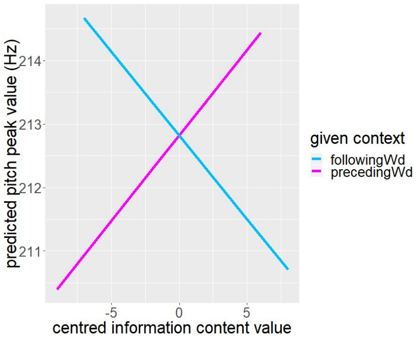

Figure 3: *Effects of precIC (magenta) and folIC (skyblue) on a pitch peak value of a word.*

6. Acknowledgements

This study was conducted at the University of Tokyo, which allows me to use facilities such as CSJ and a desktop computer. This paper benefits from the discussion at Tokyo Circle of Phonologists (TCP). I also thank Professor Shin-ichi Tanaka and his students for inviting me to their seminar. My special thanks go to Michelle van Bokhorst for her reading through this paper and providing insightful suggestions.

7. References

[1] D. Jurafsky, A. Bell, M. Gregory, and W. Raymond, “Probabilistic relations between words: Evidence from reduction in lexical production,” in *Frequency and the emergence of linguistic structure*, pp. 229-254, Amsterdam: Benjamin, 2001.

[2] A. Bell, D. Jurafsky, E. Frosler-Lussier, C. Girand, M. Gregory, and D. Gildea, “Effects of disfluencies, predictability, and utterance position on word form variation in English conversation,” *Journal of Acoustical Society of America*, vol. 113, pp. 1001-1024, 2003.

[3] M. Aylett, and A. Turk, “Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei,” *Journal of Acoustic Society of America*, vol. 119, pp. 3048-3058, 2006.

[4] T. F. Jaeger, and E. Buz, “Signal reduction and linguistic encoding,” in *The handbook of psycholinguistics*, pp. 38-81, New Jersey: Wiley-Blackwell, 2017.

[5] K. Hall, E. Hume, T. F. Jaeger, A. Wedel, “The message shapes phonology,” https://www.researchgate.net/publication/309033386_The_ Message_Shapes_Phonology/download, 2016.

[6] R. Turnbull, “The role of predictability in intonational variability,” *Language and Speech*, vol. 60(1), pp. 123-153, 2017.

[7] K. Maekawa, “Spontaneous speech corpus of Japanese,” *Nihon-go kagaku*, vol. 11, pp. 111-133, 2004.

[8] H. Fletcher, and W. A. Munson, “Loudness, its definition, measurement and calculation,” *Journal of the Acoustical Society of America*, vol. 5, pp. 82-108, 1933.

[9] P. Ladefoged, *A Course in Phonetics (Fifth Edition)*, Boston: Thomson Higher Education, 2007.

[10] P. Boersma, and D. Weenink, “Praat version 6.0.49,” praat.org, 2019.

[11] J. Pierrehumbert, and J. Beckman, *Japanese Tone Structure*, Cambridge: MIT Press, 1988.

[12] R Core Team, R version 3.5.3, https://www.r-project.org/, 2019.

[13] D. Bates, M. Maechler, B. Bolker, and S. Walker, Fitting Linear Mixed-Effects Models Using lme4, *Journal of Statistical Software*, vol. 67(1), pp. 1-48, doi:10.18637/jss.v067.i01, 2015.

[14] B. Kamil, MuMIn: Multi-Model Inference, R package version 1.42.1, 2018.