Prosodic Markings of Semantic Predictability in Taiwan Mandarin

Po-jen Hsieh
Graduate Institute of Linguistics, National Tsing Hua University, Taiwan
pj.dan.hsieh@gmail.com

Abstract
This paper investigated the effect of semantic predictability on acoustic realizations in Taiwan Mandarin. Most previous studies have been guided by the assumption that in high predictability conditions, speech units tend to be reduced in the temporal and spectral qualities. This paper focuses on a tone language in an attempt to not only examine previous findings about vowels but also look into possible effect of semantic predictability on F0. The results showed that words were shorter in vowel duration, lower in the intensity, and more centralized within the vowel space when they were embedded in high predictability contexts. Male and female speakers both performed the same reduction process systematically. F0 excursion was also found strongly affected by semantic predictability; specifically, it was much smaller in high predictability conditions. Moreover, gender difference in the extent of F0 excursion was mediated in high predictability conditions. This paper provides additional evidence for the acoustic correlates of semantic predictability in speech production and sheds light on the prosodic encoding of semantic structures in tone languages.

Index Terms: semantic predictability, speech production, Taiwan Mandarin, tone

1. Introduction
In daily life, we listen to understand and speak to be understood. Individuals participate in communicative activity in order to both express their ideas and thoughts as well as to receive them from others. Moreover, it is well-known that human speech production is dominated by various linguistic and paralinguistic dimensions [1-6]. With regard to the paralinguistic facets such as, speaking rate, age, and gender, they all can contribute to different phonetic realizations. It has been demonstrated that in fast speech, words are likely to be reduced in both temporal and spectral quality [2, 6]. Female speakers tend to have more standard pronunciation than male speakers [3, 4]. These phonetic variations, however, do not degrade the intelligibility in our daily conversation. With respect to linguistic perspectives, speech production can be influenced by neighboring segments or by larger, higher-level information, such as lexical frequency. It has been demonstrated that vowels in low frequency words tend to be longer in duration and are more dispersed in vowel space than vowels in high frequency words [7]. In tonal languages, the tonal space has been found to be more dispersed for low frequency words [8]. Altogether the versatile phonetic variations yield an interest of the link between language processing and articulatory planning.

The difficulty in giving each linguistic category a precise acoustic definition has been known as the invariance problem. [6] proposed that speakers may vary their articulatory effort along a continuum of hyper- and hypo-speech. Human speech production is controlled by two primary driven mechanisms. The output-oriented force demands that speakers manipulate the displacement of speech motors to yield sufficient discriminability for listeners, namely ‘hyper-speech’. The system-oriented force requires that the low-cost articulator displacement (‘hypo-speech’) be more preferable. [6] further posited that the adaptive processing was motivated by the lack of phonetic invariance.

[1] and [2] also made attempts to account for the varying phonetic realizations of single phonological category. For listener-motivated constraints, the accessibility of information for listeners is taken into consideration according to the need of speaking situations. Speakers performed greater displacement of speech motors to enhance perceptual advantage in adverse listening conditions. Speaker-motivated constraints, on the other hand, demands that speakers reorganize their motor gestures to low-cost speech outputs. Given contexts of favorable listening, speakers spare the least effort to complete the articulation process.

Furthermore, it is well-known that speakers make use of permissible phonetic variations to communicate more efficiently without sacrificing intelligibility [6, 9]. In particular, speakers are able to integrate multi-faceted linguistic information, such as semantic predictability and syntactic structure, into a coherently intelligible discourse. Therefore, the ultimate speech outputs represent not only sequences of speech sounds but also more complex syntactic and semantic information that speakers wish to convey.

Abundant studies have provided extensive discussions about the acoustic reduction in spoken language processing. The shortening effect has been found in the duration of words in Dutch when the target words are contextually predictable either from the preceding words or from the following words [9]. In English words tend to have shorter segments when they are contextually predictable from the neighboring words [10]. [11], based on a monologue corpus, argued that words are also reduced in duration when they are second mentioned in a discourse. Furthermore, a high frequency word tends to be reduced in its temporal and spectral quality than a low frequency word [12, 13]. In terms of semantic predictability, words in high predictability conditions are found shorter in duration, lower in intensity, and were more limited in spectral quality than words in low predictability conditions [14, 15].

Speakers, consciously or subconsciously, spend less articulatory effort in production when the words are either second mentioned in a discourse or higher in contextual predictability, lexical frequency, or semantic predictability. Furthermore, it has been proposed [14] that the reduced items are relatively easy, fast, and effortless to retrieve from the speakers’ mental lexicon.

To date, few attempts have been done to examine whether such semantic effects still hold at the suprasegmental level. Given that tone is contrastive in tone languages, how the predictability effects interact with tone merits investigation. Previous studies [16] on tonal variations focused primarily on how tonal sequences affect each other. The present study is the first to examine the effect of the semantic predictability on elicited speech in Mandarin Chinese.
2. Method

2.1. Participants

Twenty native speakers of both Taiwan Mandarin and Taiwan Southern Min were recruited from National Chiao Tung University. They were 10 females (M = 21.9, SD = 1.6) and 10 males (M = 23.2, SD = 2.1). None of them ever resided in foreign countries for more than two months. They didn’t receive any linguistic training, and claimed to be without any speech, language, and hearing difficulties. All were paid for their participation in the present study.

2.2. Corpus

There was a pretest conducted to evaluate the semantic predictability of the corpus used in the later production experiment. Additional twelve native Taiwan Mandarin speakers were recruited. The materials consisted of 12 disyllabic target Mandarin words. The first syllables had a fixed CV structure, with three-vowel (i, u, a) and four-tone options. The second syllable was controlled as the semantically neutral morpheme ‘zi’. The 12 (3 × 4) target words were situated in two semantic contexts: Low Predictability (LP), i.e., the sentence-final target words were unpredictable from the preceding semantic contexts, and High Predictability (HP), i.e., the target words were embedded in semantically predictable contexts. The order of the 24 sentences (12 target words × 2 semantic contexts) was randomized for each subject. During the pretest, the subjects first saw a sentence on the screen with the clause-final target word missing. They were instructed to fill in the word that they thought was most likely to occur in the position. After they wrote down their answers, the same but complete sentence appeared on the screen again along with predictability judgment criteria. They were then instructed to rate semantic predictability of the sentences at the scale of 1-7 by their participation in the present study.

2.3. Procedure

The recording was conducted in a sound-attenuated room at National Chiao Tung University. Speakers sat before a 24-inch computer monitor wearing headphones (audio-technica AT-HANC1). A Shure SM57 microphone was placed 15 cm from the mouth of the speakers to collect their speech which was then recorded with KORG MR-1000 recorder at a sampling rate of 44.1 kHz. During each trial, the target sentence was shown on the screen, and an eliciting question Ni yao shuo she me meaning ‘What would you say?’ was presented over the headphones after 10 seconds. Speakers were instructed to reply using the displayed sentence at a normal speed and in a natural manner. The displayed sentences were not removed while the subjects were making their replies, and then they pressed a key on the keyboard to proceed to the next trial.

Prior to the recording, the subjects underwent a six-trial practice session which contains different sentences from those in the later production experiment. Only one subject at a time participated in the study. No time limit was given to complete either each trial or the entire session. The whole procedure took each subject about 30 minutes.

2.4. Data analysis

Acoustic analyses were made using Praat [17] along with custom written scripts. The speech data was first hand-labeled by the experimenter. Duration, intensity, pitch contours, F0 excursion, first formant (F1), and second formant (F2) were calculated from the steady portion of the vowels. Three scripts were written: one to extract duration, intensity, and time-normalized F0 contours [18], one to extract the maximal and minimal F0 values [19], and the other to extract the F1 and F2 values for individual vowels [19]. The semantic predictability effect on vowel space was calculated by using the Formant Centralization Ratio (FCR) formula: (F2u + F2a + F1i + F1u) / (F2i + F1a). The greater FCR values represented the smaller vowel space [20]. For each token, F0 excursion was further obtained from the difference between maximal and minimal F0 values.

2.5. Results

2.5.1. Vowel duration

A Semantic Predictability (2) × Gender (2) two-way mixed ANOVA was done on the vowel duration. The results showed that the two-way interaction was not significant. The main effects of Semantic Predictability was significant (F (1, 18) = 138.1, p < .001]. Vowel duration was significantly longer in low predictability conditions (M = 143.7, SD = 29.9) than in high predictability conditions (M = 147.9, SD = 29.9). As can be seen from Figure 1, both male and female subjects produced the words with longer vowel duration in low predictability conditions (Male: = 152.2, SD = 35.9; Female: M = 143.7, SD = 29.9) than in high predictability conditions (Male: M = 113.5, SD = 25.3; Female: M = 105.2, SD = 23.2). However, the differences between male and female speakers in two semantic conditions were minor.

2.5.2. Vowel quality

A Semantic Predictability (2) × Gender (2) two-way mixed ANOVA was performed on the formant centralization ratio.
The results showed that both the two main effects were significant [Semantic Predictability: $F(1, 18) = 22.6, p < .001$; Gender: $F(1, 18) = 4.3, p < .05$]. However, the two-way interaction was not significant [$F(1, 18) = .8, p = .391$].

Speakers had significantly smaller formant centralization ratio in low predictability contexts ($M = 1.2, SD = 0.02$) than in high predictability contexts ($M = 1.3, SD = 0.04$) (see Figure 2). Specifically, vowel space was more centralized in high predictability conditions.

### 2.5.3. Intensity

A Semantic Predictability (2) × Gender (2) two-way mixed ANOVA was performed on the intensity. The results showed that both of the main effects were significant [Semantic Predictability: $F(1, 18) = 29.2, p < .001$; Gender: $F(1, 18) = 7.2, p < .05$]. The interaction effect between Semantic Predictability and Gender was marginally significant [$F(1, 18) = 2.5, p = .068$].

As shown in Figure 3, vowels had significantly greater intensity in low predictability contexts ($M = 61.2, SD = 6.1$) than in high predictability contexts ($M = 59.1, SD = 6.4$).

### 2.5.4. F0 excursion

A Semantic Predictability (2) × Tone (4) × Gender (2) three-way mixed ANOVA was performed on F0 excursion. The results indicated that the three-way interaction effect was significant [$F(3, 54) = 5.9, p < .001$].

A comparison across the four tones (see Figure 4) indicated that words in low predictability conditions (Tone 1: $M = 20.5, SD = 9.9$; Tone 2: $M = 33.9, SD = 17.1$; Tone 3: $M = 40.9, SD = 17.5$; Tone 4: $M = 58.2, SD = 22.7$) had greater F0 excursion than those in high predictability conditions (Tone 1: $M = 10.5, SD = 5.8$; Tone 2: $M = 12.5, SD = 6.5$; Tone 3: $M = 16.9, SD = 7.7$; Tone 4: $M = 29.6, SD = 15.2$).

A series of post-hoc tests were done to further determine the three-way interaction effect on F0 excursion. Results revealed that F0 excursion was significantly affected by semantic predictability conditions. Male and female speakers both produced words with much smaller F0 excursion in high predictability contexts than in low predictability contexts. Furthermore, the reduction effect was qualified by the interaction effect among predictability, tones, and gender. In particular, F0 excursion differences between high predictability and low predictability contexts were even greater for Tone2, Tone 3, and Tone 4 by female speakers.

Figure 2: Interaction effect between SP and gender on formant centralization ratio.

Figure 3: Interaction effect between SP and gender on Intensity.

Figure 4: Interaction effect between SP and Tone on F0 excursion.

Figure 5: Interaction effect between SP and gender on F0 excursion.
3. Discussion

3.1. Probabilistic Reduction Hypothesis

The present study has demonstrated that words in high predictability contexts tended to be reduced in their acoustic properties. Furthermore, this study provides further evidence that spoken language processing is dominated by multidimensional linguistic factors [1, 2, 3, 5, 6, 9, 21]. This study’s findings are in line with previous studies on probabilistic reduction [9, 10, 11, 12, 13, 14, 15]. Speech output from high probability conditions tended to be reduced in its acoustic properties such as duration, vowel quality, and intensity. Human speech production is shaped and modulated by high-level linguistic structure such as lexical frequency and semantic predictability. The fine linguistic details embedded in the acoustic speech signals are far more than just the sum of the lexical and sentential meanings. Consequently, finer-grained phonetic variations can serve to illustrate the semantic predictability contrasts.

This study’s findings confirm the probabilistic reduction hypothesis [10, 22] that the semantic predictability has a strong reduction effect on the acoustic realizations in Taiwan Mandarin. The twenty speakers in the present study employed temporal and spectral quality, intensity, and F0 excursion to mark the semantic predictability contrasts. Specifically, words in high predictability contexts were produced with shorter duration, lower intensity, and smaller F0 excursion, and the individual vowel targets were more centralized in the vowel space. On the other hand, speech outputs from low predictability contexts tended to be longer in duration, higher in intensity, and greater in F0 excursion, and the vowel targets were located at the more peripheral positions.

The lenition effect on speech processing can be explained from two perspectives. First, the listener-oriented constraints demand that the speech signals should be as clear and intelligible as possible [1, 2, 6]. In low predictability contexts, speakers assume that listeners have no idea about what they are going to perceive, and thus the precise speech should be given. In high predictability contexts, however, speakers assume that listeners can comprehend the whole sentences even when parts of the information are missing. The preceding semantic contexts may compensate for the missing or unclear messages. Therefore, speakers will not have to spend so much articulatory effort on the information transfer.

Secondly, the speaker-oriented constraints demand that certain “easy” words be much easier to retrieve from speakers’ mental lexicon and thus less time is spent on it [1, 2, 6]. On the other hand, it takes more time to retrieve “hard” words. The “easy” words are defined as words with high lexical frequency, in high predictability contexts or being second mentioned in a discourse. The “hard” words are just the opposite. Speech production is well-known a continuous process in which the articulators move rapidly from one target configuration to the next [23]. Shorter retrieval implies that less time was given to complete the whole articulatory process; accordingly, a reduced form is more likely to be acoustically realized [14] hypothesized that words in high predictability contexts are much easier to retrieve from speakers’ mental lexicon and thus they are more likely to be reduced in their acoustic properties.

It is obvious that speakers are aware of the high-level linguistic organization during their articulatory planning, and are capable of encoding the information into speech utterances [1, 2, 6]. Consequently, speakers tend to reduce the acoustic properties of words in high predictability contexts but not in low predictability contexts.

One point to be raised in the present study is that most of the literature has focused only on the acoustic reduction in high probability conditions. Few studies concern the acoustic enhancement in low probability conditions. Will the processing of speech production be unidirectional? As far as the hyper- and hypo-speech theory [6] is concerned, the modulation of speech production should be bidirectional, in which speakers spend more articulatory effort enhancing the acoustic quality of their utterances in adverse perceptual conditions (hyper-speech) and less effort reducing the acoustic quality in favorable perceptual conditions (hypo-speech). Taken together, speakers tune their articulation force along the hyper- and hypo-continuum in accordance with the redundancy of the information they wish to convey. The norm of speech production, however, must be well-defined prior to the establishment of the concept that both acoustic reduction and enhancement processes are involved in the course of speech motor planning.

4. Concluding remarks

The present study has demonstrated that speech production is strongly affected by semantic predictability in Taiwan Mandarin. Congruent with previous studies, vowels were reduced in the temporal quality and intensity to a large extent in high predictability contexts. In terms of spectral quality, vowels tended to be centralized in high predictability contexts. Furthermore, male and female speakers both made use of F0 excursion variations to mark the contrasts of semantic predictability. In particular, the asymmetries in utilizing F0 excursion between males and females were neutralized in high predictability contexts. In sum, these findings not only further support previous studies in the temporal and spectral reduction in high predictability contexts, but also first indicate the semantic predictability effect on F0 excursion in tone languages.

5. Acknowledgements

Special thanks go to my thesis advisor, Dr. Hui-chuan Hsu, whose influence on this work is more than obvious. I have also benefited from comments, criticism by Dr. Feng-fan Hsieh and Dr. Yuwen Lai. All errors are my own.

6. References


Xu, Y. 2011. http://www.phon.ucl.ac.uk/home/yi/ProsodyPro/


