TOP-DOWN EFFECTS OF LEXICAL PITCH ACCENT ON PHONETIC CATEGORIZATION IN JAPANESE

Terumichi Ariga, Risa Matsubara

Graduate School of Arts and Sciences, The University of Tokyo
ariga@phiz.c.u-tokyo.ac.jp, matsubara-risa761@g.ecc.u-tokyo.ac.jp

ABSTRACT

Although recent studies have shown that spoken word recognition is achieved with reference to not only segmental but also lexical prosodic information, the role of lexical prosody in speech perception is not well established. The present study investigated whether Japanese pitch accent has effects on phonetic categorization of voiceless–voiced continua (e.g., *tansu* LHH – *dansu* LHH, in which only *tansu* LHH "wardrobe" is a real word, and *tansu* HLL – *dansu* HLL, in which only *dansu* HLL "dance" is a real word). The results demonstrated that listeners were inclined to identify the voicing of the initial consonant so that the entire input resulted in a real word with consideration of pitch accent information. This suggests that Japanese pitch accent has a top-down effect on phonetic categorization, providing further evidence that lexical prosody plays a role in spoken word recognition in Japanese.

Keywords: speech perception, lexical prosody, phonetic categorization, top-down effect, Japanese

1. INTRODUCTION

Psycholinguistic studies have proposed various models of spoken word recognition by human listeners [8, 16, 18, 21, 30, 33]. Spoken word recognition is achieved under various levels of linguistic information. For example, the TRACE model of spoken word recognition [18] hypothesizes that there are three levels of processing: acoustic-feature, phoneme, and lexical levels. Each level of processing interacts with each other and gives both positive and negative feedback on other levels of processing.

This allows the redundancy of information in speech perception, making the mechanism of spoken word recognition robust against adverse situations in speech comprehension. For example, even when a phoneme is missed in auditory sentence comprehension, listeners can restore the missing input of a phoneme by referring to the context [31, 32]. When an acoustically equivocal phoneme is presented within a lexical environment, listeners interpret the phoneme so that the entire word becomes a real word [12, 19]. These effects are offered by top-down effects that a higher-level processing has on a lower-level processing. The likelihood of existing real words computed in the lexical-level processing has a bias on the processing of perception of segments.

Recently, ample studies have offered evidence that spoken word recognition is achieved with reference to not only segmental but also suprasegmental information, such as lexical prosody [9]. Words spoken with incorrect lexical prosody impose difficulty in recognizing the words appropriately [7, 20, 27]. When listeners listen to a spoken word, they activate words whose lexical prosody is identical to the perceived prosody but not words whose lexical prosody contradicts the perceived prosody [2, 11, 23, 28]. In languages in which prosody is lexically specified, lexical prosody is a potential cue for identifying spoken words.

The role of lexical prosody is not restricted to constraining lexical access as a bottom-up processing. In English, lexical prosody has a top-down influence on lower-level segmental processing. Connine and others [6] investigated perceptual phonetic categorization of consonants [15] in voice onset time (VOT) continua of a stress-contrastive pair *diGRESS* – *tiGRESS* and *DIgress* – *TIgress* (*diGRESS* and *TIgress* are real words; capital letters indicate stressed syllables). They reported that listeners tended to identify the voicing of initial consonants as voiceless (/t/) if the stress pattern of continua was strong-weak (i.e., *DIgress* – *TIgress*) than weak-strong (i.e., *diGRESS* – *tiGRESS*). This was because the strong-weak continua would become a real word if the initial consonant was /t/ (*TIgress*) whereas the weak-strong continua would become a real word if the initial consonant was /d/ (*diGRESS*), suggesting that the acoustically equivocal consonant was perceptually categorized so that the entire input becomes a real word rather than a nonword.

This replicated the finding that the categorization of phonemes is subject to be influenced by lexical status [12, 19]. Nevertheless, since the lexical status was determined by the stress pattern of the items, these results could be accounted for by the top-down effect of lexical prosody on phoneme-level processing.

However, it is still unclear whether these effects of

lexical prosody obtained in the English study can be generalized to other languages. In the present study, we investigated whether Japanese lexical prosody has top-down effects on the categorization of an equivocal input of phonemes.

Unlike stress languages such as English, Japanese lexical prosody is pitch accent, realized by high-or-low tonal patterns specified for each lexical item [13, 17, 29]. In Japanese, Cutler and Otake [10] reported that listeners can infer the entire input of a word based on pitch accent information in a gating paradigm. This suggests that lexical prosodic information is beneficial for compensating for the upcoming input of segments. However, it remains unknown whether Japanese pitch accent influences the perceptual categorization of an ambiguous input of segments.

In addition, Japanese is suitable for assessing the top-down effects of lexical prosody on the segment-level processing for the following reasons. First, compared to English, Japanese has relatively many lexical items with prosodic contrasts [25]. Thus, it is easy to find experimental materials. Contrary to the previous study [6] which used only a single item (a *TIgress–diGRESS* pair), we used multiple items in Japanese in the present experiment.

Second, Japanese pitch accent is unique in that the produced form of lexical prosody is not constant. In Japanese, the surface pitch pattern of a lexical item is variable due to phonological rules such as compound accent rule and deaccentuation [13]. There is also a dialectical variety of pitch accent [14]. Despite this variability of pitch accent, listeners of Japanese can recognize words presented with different lexical prosody from their lexical knowledge with minimal difficulty [3, 24]. Considering this, there is a possibility that pitch accent in Japanese may not be a reliable cue for categorizing ambiguous inputs.

Therefore, the present study addressed whether Japanese lexical pitch accent plays a role in phonetic categorization. We conducted an identification task of VOT continua, following Connine and others' experiment [6].

2. EXPERIMENT

2.1. Materials

Six pairs of experimental materials were prepared. All pairs were tri-moraic real words whose onset was a stop, segmentally contrastive by the voicing of the initial consonant, and prosodically contrastive by lexical pitch accent (either unaccented LHH or initially accented HLL). The full pairs are shown in Table 1. Pitch accent types of voiceless materials were counterbalanced. By referring to the Japanese lexical database [1], all materials were confirmed that they were not pronounced by any other pitch patterns in standard Tokyo Japanese, and highly familiar words among Japanese lexical items.

voiceless biased	voiced biased
koten LHH "classic"	*goten* HLL "palace"
kazai HLL "valuables"	*gazai* LHH "art supplies"
penki LHH "paint"	*benki* HLL "toilet"
panchi HLL "punch"	*banchi* LHH "address"
tansu LHH "wardrobe"	*dansu* HLL "dance"
teashi HLL "limb"	*deashi* LHH "start"

Table 1: Full sets of the experimental materials.

In these items, the voicing of the initial consonant was dependent on the type of pitch accent. Therefore, lexical status should be determined only by specifying the lexical prosodic pattern. For example, unaccented and voiceless-initial *tansu* LHH "wardrobe" was a real word, whereas *dansu* LHH was a nonword. Similarly, accented and voiced-initial *dansu* HLL "dance" was a real word, whereas *tansu* HLL was a nonword. Thus, if there are any effects of lexical prosody in categorizing the initial consonants, listeners should be inclined to report the voicing of the initial consonants so that the entire word becomes a real word. Items which would be recognized as real words if the initial consonant was voiceless were categorized as a voiceless-biased condition, while items which would be recognized as real words if the initial consonant was voiced were as a voiced-biased condition. Note that this bias was offered by the pitch accent of the items.

Auditory materials were made by the following procedure. Firstly, we recorded natural tokens of both voiced and voiceless items in both unaccented and accented pitch patterns (i.e., we made four tokens for each item: *tansu* LHH, *dansu* LHH, *tansu* HLL, and *dansu* HLL). The tokens were produced by a female native Tokyo Japanese speaker in a sound attenuated room, digitized at a sampling rate of 44.1 kHz with 16-bit resolution using a Samson Q2U microphone. The intensity of all the sound files were normalized using Praat 6.2.12 [5]. Then, using a Praat script [34], we synthesized six-level continua of the words in which the voicing of the initial consonant gradually ranged from a voiced consonant to a voiceless consonant. Following the previous study [6], these continua were made by manipulating VOT by 12 ms increments from 10 to 70 ms (i.e., 10, 22, 34, 46, 58, and 70 ms). We must note that this range of the VOT steps was based on English voicing contrast and the contrast of voicing in Japanese takes slightly different range [26]. However, the continua of a shorter VOT should be categorized as words with voiced consonants, whereas the continua of a longer VOT should be as words with voiceless consonants.

2.2. Procedure

The experiment was conducted online via PCIbex [35]. Participants were instructed to listen to the auditory materials through a headphone at a comfortable volume.

The task was an identification task of the voicing of the initial consonant in a two-alternative forced choice (2AFC) paradigm. At the beginning of each trial, a fixation point (+) was displayed at the center of the display. After 1000 ms, participants listened to either one of the six-level continua. Participants were asked to judge whether the initial consonant of the auditory material was voiced or voiceless. They answered by choosing either one of the buttons with labels of the voiced or voiceless initial mora in *hiragana* (Japanese syllabary characters) presented on the display. While there was no timeout for the response, participants were forced to select one of the choices. After the response, the next trial started automatically with a 1000 ms interval.

We emphasized the following notes in the instruction. First, the aim of the experiment was to investigate the mechanism of human speech perception and not to test the participants' listening ability: participants should answer intuitively without too much serious consideration. Second, therefore, there were no absolute correct answers, and the ratio of voiced and voiceless responses need not to be 1:1.

The task consisted of 72 main trials (2 bias conditions × 6 VOT lengths × 6 items), after twelve practice trials (where a *koban* HLL – *goban* LHH item with six-level VOTs was used). The items were presented in a random order. Main trials were divided into two sessions and participants could take breaks between the sessions.

After the identification task, we conducted a pronunciation task to confirm participants' knowledge on the pitch accent of the words used in the experiment. This confirmation was necessary because lexical knowledge on prosody could slightly vary from listener to listener even if they were standard Tokyo Japanese listeners.

Twelve words used as the auditory materials were presented on the display one by one. Participants pronounced the words naturally. The recording files were submitted to the experimenter after the experiment and the experimenter checked the listeners' pronunciation. The experiment lasted for 10–20 minutes in total.

2.3. Participants

Twenty-five native speakers of Tokyo Japanese participated in the experiment. Their mean age was 24.40 (SD = 6.73). They received monetary compensation for their participation.

2.4. Results

Data of the trials were excluded if the trials used auditory materials which each participant failed to pronounce with a standard Tokyo Japanese pitch pattern in the pronunciation task. This totaled 13.3 % of all the trials. Responses were coded as 1 if they answered that the initial consonant of the auditory material was voiceless and as 0 if they answered that the initial consonant was voiced.

The rates for voiceless response for each condition are shown in Figure 1. Changes in response rates as a function of VOT were regressed with logistic curves individually for each bias condition.

Response rates were analyzed by generalized linear mixed-effects models (GLMM), using lme4 package [4] in R 4.2.1 [22]. Participants' response (either 0 or 1) was set as the dependent variable. For the independent variables, the bias conditions, the VOT conditions, and their interaction were set as fixed effects whereas intercepts of both participants and items were set as random effects. The bias conditions were numerically coded as –0.5 or 0.5 for the voiced-bias condition and the voiceless-biased condition, respectively. The VOT conditions were also numerically coded as 10, 22, 34, 46, 58, or 70 for each six-level length (70 is the most likely to be categorized as voiceless).

The GLMM analysis suggested that there were significant main effects of bias conditions (β = 1.01, SE = 0.50, z = 2.00, p = .045) and VOT (β = 0.17, SE = 0.01, z = 18.03, p < .001), whereas their interaction was not significant (β = 0.02, SE = 0.02, z = 0.95, p = .342). As predicted, the rate for the voiceless response became higher as the VOT became longer. However, listeners tended to report more that the initial consonant was voiceless in the voiceless-

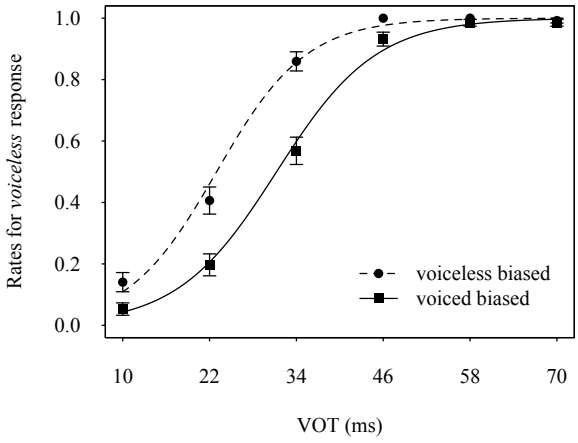

Figure 1: Rates for voiceless response for each bias condition and for each VOT. Error bars indicate the SE. The dashed line is a logistic curve for the voiceless-biased condition and the solid line is for the voiced-biased condition.

biased condition than in the voiced-biased condition.

Then, to investigate the difference in response rates by the bias conditions in detail, we made six post-hoc pairwise comparisons for each VOT condition with the Bonferroni correction (adjusted α = 0.008). When VOTs were 46, 58 and 70 ms, response rates did not differ significantly between the voiced-bias condition and the voiceless-biased condition (46 ms: p = .948; 58 ms: p = .907; 70 ms: p = .976). However, when VOTs were 10, 22, and 34 ms, listeners tended to identify the initial consonant as voiceless in the voiceless bias condition than in the voiced-bias condition (10 ms: p = .002; 22 ms: p < .001; 34 ms: p < .001). In these conditions, a bias occurred in phonetic categorization so that the entire input of auditory materials became a real word rather than a nonword.

3. CONTROL EXPERIMENT

The bias in phonetic categorization obtained in the experiment can be interpreted not only by the effects of lexical prosodic pattern but also by some unexpected VOT properties of the materials. To exclude this confounding interpretation, we conducted a supplemental control experiment.

We extracted the first mora of the original tri-moraic materials (e.g., *da–ta* continua with either H or L tone). These mono-moraic materials (72 trials in total) were presented in the same procedure as an identification task in the main experiment.

Nine native speakers of Tokyo Japanese who did not participate in the main experiment participated in the experiment. Their mean age was 26.40 (SD = 2.07).

The rates for voiceless response for each condition are shown in Figure 2. The GLMM analysis suggested that there was a significant main effect of VOT (β = 0.19, SE = 0.02, z = 11.34, p < .001), whereas there were neither significant main effect of the bias conditions (β = –0.36, SE = 0.84, z = –0.43, p = .194) nor their interaction (β = 0.01, SE = 0.03, z = 0.18, p = .857). The trend of categorical perception of VOT continua did not differ between the bias conditions when the auditory materials were the extractions of the initial mora. This suggested that the inclination of phonetic categorization in the main experiment occurred not due to the bias in VOT lengths of the materials themselves, but due to the bias that offered by the lexical prosodic pattern of the materials.

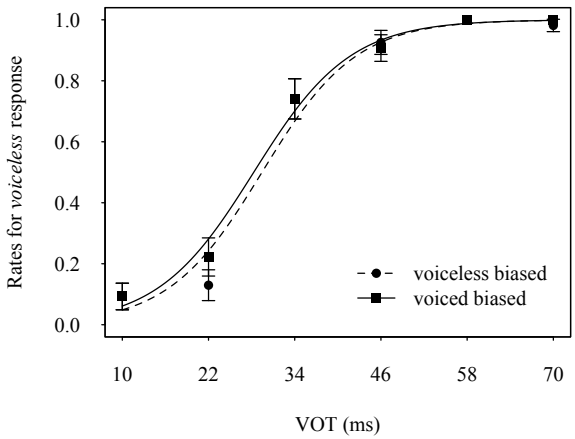

Figure 2: Rates for voiceless response for each bias condition and for each VOT (control experiment). Error bars indicate the SE.

4. DISCUSSION

The present study investigated whether lexical prosody plays a role in phonetic categorization in Japanese speech perception. Results demonstrated that listeners' categorization of the initial consonants in VOT continua was affected by the lexical-status bias offered by the pitch accent information. In the *tansu–dansu* pair, for example, when the pitch pattern was unaccented LHH, listeners tended to perceive /t/ more than /d/ since the continua became a real word if the initial consonant was /t/ (*tansu* LHH "wardrobe"). Similarly, when the pitch pattern was initially accented HLL, listeners tended to perceive /d/ more than /t/ since the continua became a real word if the initial consonant was /d/ (*dansu* HLL "dance").

This replicated the previous study in English [6], suggesting that the previous findings in English can be extended to the case of Japanese. Contrary to the hypothesis that Japanese pitch accent may not be a reliable cue due to the variability of lexical prosody [24], pitch accent may play a role in computing lexical status of perceived spoken words.

These behavioral results can be accounted for within the paradigm of the parallel processing models of spoken word recognition [18], assuming that there is a level for processing lexical prosody in the architecture. In Japanese, the processing of lexical prosody fundamentally plays a role in constraining lexical activation [2, 23]. In addition, this processing also affects the phonetic categorization that should be processed in the lower phoneme-level processing.

Therefore, pitch accent has both bottom-up and top-down effects in Japanese spoken word recognition. The mechanism of spoken word recognition should be robust against variety of acoustic signals by utilizing both segmental and lexical prosodic information. The findings in the present study provide further evidence that pitch accent plays a role in spoken word recognition in Japanese.

5. REFERENCES

[1] Amano, S., Kondo, T. 1999. *NTT Deetabeesu Shiriizu: Nihongo no Goitokusei* [*NTT Database Series: Lexical Properties of Japanese*]. Sanseidoo.

[2] Ariga, T. 2022. Pitch accent constrains lexical activation in Japanese spoken word recognition: A semantic priming study. *Language and Information Sciences* 20, 1–17.

[3] Ariga, T. 2022. Bummyakuka de teiji sareta ayamatta akusento no saikaishaku no katei [Process of reanalysis of mispronounced prosody in sentential context]. *IEICE Technical Report* TL2022-20, 11–16.

[4] Bates, D., Mächler, M., Bolker, B., Walker, S. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1), 1–48.

[5] Boersma, P., Weenink, D. 2022. Praat: Doing phonetics by computer. https://www.praat.org/

[6] Connine, C. M., Clifton, C. E., Cutler, A. 1987. Effects of lexical stress on phonetic categorization. *Phonetica* 44(3), 133–146.

[7] Cutler, A., Clifton, C. E. 1984. The use of prosodic information in word recognition. In: Bouma, H., Bouwhuis, D. G. (eds), *Attention and Performance X: Control of Language Processes*. Lawrence Erlbaum Associates, 183–196.

[8] Cutler, A. 1995. Spoken word recognition and production. In: Miller, J. L., Eimas, P. D. (eds), *Speech, Language, and Communication*. Academic Press, 97–136.

[9] Cutler, A., Dahan, D., van Donselaar, W. 1997. Prosody in the comprehension of spoken language: A literature review. *Language and Speech* 40(2), 141–201.

[10] Cutler, A., Otake, T. 1999. Pitch accent in spoken-word recognition in Japanese. *The Journal of the Acoustical Society of America* 105(3), 1877–1888.

[11] van Donselaar, W., Koster, M., Cutler, A. 2005. Exploring the role of lexical stress in lexical recognition. *Quarterly Journal of Experimental Psychology* 58A(2), 251–273.

[12] Ganong, W. F. 1980. Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, 6(1) 110–125.

[13] Kawahara, S. 2015. The phonology of Japanese accent. In: H. Kubozono (ed), *The Handbook of Japanese Phonetics and Phonology*. De Gruyter Mouton, 445–492.

[14] Kubozono, H. 2021. *Ippan Gengogaku kara Mita Nihongo no Purosodii* [*Japanese Prosody from General Linguistics Perspectives*]. Kuroshio Shuppan.

[15] Liberman, A. M., Harris, K. S., Hoffman, H. S., Griffith, N. C. 1957. The discrimination of speech sounds within and across phonemic boundaries. *Journal of Experimental Psychology* 54(5), 358–368.

[16] Marslen-Wilson, W. D. 1987. Functional parallelism in spoken word-recognition. *Cognition* 25, 71–102.

[17] McCawley, J. D. 1968. *The Phonological Component of a Grammar of Japanese*. Mouton.

[18] McClelland, J. L., Elman, J. L. 1986. The TRACE model of speech perception. *Cognitive Psychology* 18(1), 1–86.

[19] McQueen, J. M. 1991. The influence of the lexicon on phonetic categorization: Stimulus quality in word-final ambiguity. *Journal of Experimental Psychology: Human Perception and Performance* 17(2), 433–443.

[20] Minematsu, N., Hirose, K. 1995. Role of prosodic features in the human process of perceiving spoken words and sentences in Japanese. *Journal of the Acoustical Society of Japan (E)* 16(5), 311–320.

[21] Norris, D. 1994. Shortlist: A connectionist model of Continuous Speech recognition. *Cognition* 52(3), 189–234.

[22] R Core Team 2022. R: A language and environment for statistical computing. https://www.R-project.org/

[23] Sekiguchi, T., Nakajima, Y. 1999. The use of lexical prosody for lexical access of the Japanese language. *Journal of Psycholinguistic Research* 28(4), 439–454.

[24] Shibata, T. 1961. Nihongo no akusento [Japanese accent]. *Gengo Seikatsu* 117, 14–20.

[25] Shibata, T., Shibata, R. 1990. Akusento wa dooongo wo donoteido bembetsu shiuru ka: Nihongo, eigo, chuugokugo no baai [How significant is word accent in differentiating homonyms in Japanese, English, and Chinese?]. *Keiryoo Kokugogaku* 17(7), 317–327.

[26] Shimizu, K. 2018. Heisa shiin no yuusei sei musei sei no onkyoo teki tokutyoo ni kansuru koosatsu [A Study on Phonetic Characteristics of Voicing Contrasts of Stop Consonants]. *Journal of the Phonetic Society of Japan* 22(2), 69–80.

[27] Slowiaczek, L. M. 1990. Effects of lexical stress in auditory word recognition. *Language and Speech* 33(1), 47–68.

[28] Soto-Faraco, S., Sebastián-Gallés, N., Cutler, A. 2001. Segmental and suprasegmental mismatch in lexical access. *Journal of Memory and Language* 45, 412–432.

[29] Vance, T. J. 1987. *An Introduction to Japanese Phonology*. State University of New York Press.

[30] Vitevitch, M. S., Siew, C. S. Q., Castro, N. 2018. Spoken word recognition. In: Rueschemeyer, S., Gaskell, M. G. (eds), *The Oxford Handbook of Psycholinguistics* (2nd edition). Oxford University Press, 31–47.

[31] Warren, R. M. 1970. Perceptual restoration of missing speech sounds. *Science* 167, 392–393.

[32] Warren, R. M., Warren, R. P. 1970. Auditory illusions and confusions. *Scientific American* 223, 30–36.

[33] Weber, A., Scharenborg, O. 2012. Models of spoken-word recognition. *Wiley Interdisciplinary Reviews: Cognitive Science 3*, 387–401.

[34] Winn, M. 2020. Praat script to manipulate VOT in natural speech. https://github.com/ListenLab/VOT

[35] Zehr, J., Schwarz, F. 2018. PennController for Internet Based Experiments (IBEX). https://www.pcibex.net/