THE PROCESSING OF WORD PROSODY IN JAPANESE

Anne Cutler* and Takashi Otake†

*Max-Planck-Institute for Psycholinguistics

†Department of English
Dokkyo University

ABSTRACT - The prosodic structure of Japanese polysyllabic words is defined by patterns of high and low pitch accents. The present study investigated whether the accent level of a single syllable extracted from its word context can be reliably identified by listeners. 96 tokens of the same CV sequence, extracted from the utterances of 32 words by three speakers, were presented to 24 listeners; their correct identification rates were high. Scores were higher for word-initial than for word-final syllables, and acoustic correlates of accent level were stronger in word-initial syllables, which is consistent with a role for pitch accent information in lexical access in Japanese.

INTRODUCTION

The prosodic structure of words differs across the languages of the world. In some languages all words have the same prosodic structure - thus in Polish all words are stressed on the penultimate syllable, for instance. In other languages, words consisting of the same segmental sequence can differ solely in prosody. One of the principal areas of cross-linguistic difference in lexical prosodic structure is the domain within which a prosodic contrast is realised. In tone languages such as Mandarin or Thai, for example, this domain may be a single syllable. In such languages one may contrast two patterns (Mandarin Tone 1 versus Tone 2, for example) by comparing isolated syllables. With other kinds of contrast it makes no sense to speak of the prosody of a single syllable; stress in English or Dutch or Finnish, or pitch accent in Japanese, are essentially polysyllabic phenomena, and one can only contrast two stress patterns, or two pitch accent patterns, by comparing polysyllabic utterances.

Spoken word recognition is a process which occurs rapidly and easily. Incoming acoustic information activates compatible candidate words in the listener's mental lexicon. However the role of specifically prosodic information in this process of lexical activation is controversial. Experimental evidence from English, for example, shows that listeners can effectively achieve lexical access solely by reference to segmental contrasts, without needing to take purely prosodic contrasts into account (thus the English noun *FORbear* and verb *forBEAR* are both activated whichever is spoken; Cutler, 1986). Lexical stress information may not be reliably conveyed by the initial syllable of a word (thus Dutch listeners can not reliably tell whether, for instance, a token of the syllable *ka-* comes from *kanarie*, with stress on the second syllable, or *kanon*, with stress on the first; Jongenburger & van Heuven, 1995). The domain over which stress contrasts are realised, in variable stress languages such as English and Dutch, could simply rule out the usefulness of such contrasts in lexical access; processing prosodic cues to stress may simply be too slow to be worthwhile. On the other hand, it could also be the case that stress is hard to use because it is often realised (in languages such as English and Dutch, at least) to a considerable extent via durational contrasts, which are blurred by concurrent effects of phrase position etc. on syllable duration. Or it may simply be the case that (again, at least in English) the prosodic cues to stress are hardly worth processing because segmental structure gives the listener sufficient information about stress.

The case of pitch accent in Japanese is an interesting one for the further elucidation of questions about the role of prosodic information in lexical activation. Pitch accent is in a sense like tone in that contrasts are realised via variation in pitch level - syllables have either high (H) or low (L) accent. But it strongly resembles stress in that it is realised within a polysyllabic domain; monosyllables in isolation cannot exemplify contrasts between pitch accent patterns. (In fact, monosyllabic - or, more accurately, monomoraic - words in Japanese by convention have high pitch accent; Vance, 1987.)

The phonology and the production of pitch accent patterns on Japanese words have been extensively studied (see Haraguchi, 1988, and Vance, 1987 for reviews in English). Relatively little, however, is known about pitch accent perception. Recent studies by Nishinuma and colleagues (e.g. Nishinuma, 1994;

Nishinuma, Arai & Ayusawa, 1996) have demonstrated that discrimination of pitch accent patterns is very difficult for foreign learners of Japanese. Walsh Dickey (1996) conducted a same-different judgement experiment in which Japanese listeners heard pairs of CVCV words or nonwords which were either the same, or differed either in pitch accent or in one of the four segments. "Different" judgements were significantly slower for pairs varying in pitch accent than for pairs which varied segmentally, irrespective of the position of the segmental difference. Thus even a difference in the final vowel (at which time the pitch accent pattern should also be unambiguous) led to significantly faster responses than the pitch accent difference. Otake, Hatano, Cutler and Mehler (1993) found no effects of pitch accent in a syllable-detection task with Japanese listeners: the first CV of a word was perceived equally rapidly and accurately irrespective of whether the word had HLL (e.g. *monaka*) or LHH (e.g. *kinori*) accent pattern. None of these results suggest that pitch accent contributes readily to the lexical access code.

In the present study we ask whether naturally spoken Japanese words present the listener with reliable information about pitch accent patterns of a type which could be useful for narrowing the set of possible word candidates. We pose this question by investigating whether a syllable extracted from a word contains information about its pitch accent level: high versus low. Of particular interest is whether word-initial syllables contain exploitable pitch-accent information. Japanese contains pitch accent minimal pairs such as *ame* (HL *rain*) and *ame* (LH *candy*), which could be unambiguously distinguished if pitch accent could be determined from the initial syllable; unlike stress pairs in English, such minimal pitch accent pairs would then not be effectively homophonous. The restricted phonological inventory of Japanese also means that exploiting the pitch accent information in a word's first syllable could drastically cut the number of potential candidate words.

METHOD

Materials

32 words were chosen, all with the segmental structure CVCV and containing the mora/syllable *ka*. Half of the words had HL accent pattern, half LH. For each pattern, in half of the words the syllable *ka* was word-initial, in half word-final. Each word was paired with another, with the contrasting accent pattern, such that the two members of a pair contained the same phonetic material adjacent to the *ka* (e.g. *kage/kagi*; *baka/gaka*). The full set of words was: HL: *baka, kika, waka, huka, naka, buka, deka, yoka, kage, kako, kame, kare, kagu, kazu, kasa, kaku*; LH: *gaka, shika, taka, nuka, haka, yuka, geka, hoka, kagi, kake, kami, kara, kago, kaze, kase, kaki.*

All words were recorded by three female speakers of Tokyo Japanese, who were naive as to the purpose of the experiment. The 96 resulting productions were digitised, using the ESPS speech editing system with WAVES+, and the *ka* syllables were extracted from each production. The following nine acoustic measures were computed for each of the extracted syllables: Minimum F0; Maximum F0; F0 Range; Mean F0; Standard Deviation of F0; Total Syllable Duration; Vowel Duration; Mean RMS Amplitude; Standard Deviation of RMS Amplitude. F0 and amplitude measures were computed for voiced portions of the signal only. The 96 *ka* tokens were recorded, in random order, onto Digital Audio Tape. Vowel-final syllables produced in isolation are typically closed with a glottal stop, and this was the case in all of the 48 *ka*-final tokens; this glottal stop was included in the *ka* tokens on the tape.

Subjects and Procedure

The subjects were 24 undergraduates of Dokkyo University, all from the Kanto area (Tokyo and environs). They were presented with the tape containing the *ka* tokens and were required to choose for each token between two words from which it might have come (e.g. *kage* HL vs. *kagi* LH; *baka* HL vs. *gaka* LH). These choices were presented in written form, in both *kanji* and *hiragana* orthography, and the subjects circled their choice for each token. Note that subjects were never asked to decide whether a syllable was word-initial or word-final; each choice was between two initial syllables (one H, one L) or between two final syllables (one H, one L). The choice was, further, always between the two members of a phonetically matched pair, so that possible coarticulatory information adjacent to the *ka* boundary could not provide clues to identify the source word. Each pair occurred on the response sheet six times (corresponding to the two source words spoken by each of the three speakers), and it was given three times in each possible order, with neither source word nor speaker always having the same order.

RESULTS

Perceptual judgements

The overall correct response rate was very high (74%). Identification was more accurate for H (87%) than for L syllables (61%; F1 [1,23] = 72.75, $p < .001$; F2 [1,84] = 97.63, $p < .001$), and for initial (80%) than for final syllables (68%; F1 [1,23] = 23.92, $p < .001$; F2 [1,84] = 18.41, $p < .001$).

There was however a significant effect of speaker, with Speaker 1 receiving lower correct-identification scores (64%) than Speakers 2 and 3 (78%, 79%; F1 [2,46] = 17.51, $p < .001$; F2 [2,84] = 13.02, $p < .001$). An analysis of the results excluding Speaker 1's productions revealed that both the main effect of accent level (H 86%, L 72%) and the main effect of position in the word (initial 84%, final 74%) remained statistically significant.

Fifteen of the 96 items received scores below chance; all were L syllables mistakenly judged by the majority of subjects as H. Eleven of those were spoken by Speaker 1. Of the eight items with scores significantly below chance (9/24 or less), six were spoken by Speaker 1, and five of these were final L syllables. Thus this speaker systematically failed to signal L accent on a final syllable (not one of her eight final-L items was identified with accuracy significantly above chance).

Acoustic analyses

Table 1 shows the mean value on each of the nine measures, separately for the four syllable types. Analyses of variance across the tokens were computed for each measure. The main focus of interest here is where acoustic differences between H and L syllables are to be observed, since the H/L categorisation was essentially the listeners' task in this experiment.

Pitch: The five measures which we made of the pitch characteristics of the syllables revealed a simple and consistent pattern. The minimum, maximum and mean F0 values for the syllables tended to pattern together: if one of these measures showed a significant difference between H syllables and L syllables, so did the others. Likewise, the two remaining measures, F0 range and standard deviation of F0 (both of which provide crude estimates of the amount of pitch movement across a syllable) also patterned together, and separately from the other set.

The minimum, maximum and (therefore also the) mean F0 were all significantly higher in H syllables than in L syllables (F0min: F [1,28] = 259.33, $p < .001$; F0max: F [1,28] = 56.43, $p < .001$; F0mean: F [1,28] = 310.78, $p < .001$), and were also significantly higher in initial than in final syllables (F0min: F [1,28] = 107.75, $p < .001$; F0max: F [1,28] = 9.08, $p < .01$; F0mean: F [1,28] = 126.45, $p < .001$). On each measure there was also a significant interaction between syllable position and accent level, whereby the H/L difference was greater in initial than in final syllables (F0min: F [1,28] = 16.28, $p < .001$; F0max: F [1,28] = 64.34, $p < .001$; F0mean: F [1,28] = 58.92, $p < .001$).

All three of these measures also showed a significant effect of speaker (F0min: F [2,56] = 79.23, $p < .001$; F0max: F [2,56] = 48.53, $p < .001$; F0mean: F [2,56] = 104.49, $p < .001$). The source of this effect was that Speaker 1 had a noticeably higher voice, approximately 35 Hz higher on each F0 measure, than the other two. An analysis of the results for only the syllables uttered by Speakers 2 and 3 showed that all the main effects of accent level and of syllable position, and the interactions between these two factors, remained significant as reported above.

Both the F0 range and the standard deviation of F0 were significantly greater for L than for H syllables, and significantly greater in final than in initial syllables. The interaction between accent level and syllable position was also significant in the opposite direction (greater H/L differences in final than in initial syllables). On neither of these two measures was there a significant effect of speaker.

Duration: Neither durational measure showed significant differences between H and L syllables. Final syllables were however significantly longer than initial syllables (overall: F [1,28] = 4.9, $p < .05$; vowel: F [1,28] = 29.8, $p < .001$).

Amplitude: H syllables had significantly greater mean amplitude than L syllables ($F [1,28] = 10.85$, $p < .005$). There was no difference between initial and final syllables, or interaction between accent level and syllable position. The standard deviation of amplitude showed no main effect of either accent level or position. However, there was again an effect of speaker on both measures (mean: $F [2,56] = 64.61$, $p < .001$; sd: $F [2,56] = 31.92$, $p < .001$), and again, this was due to deviance of the productions of Speaker 1, who spoke significantly louder than the other two.

	Initial Syllables		Final Syllables	
	H	L	H	L
minimum F0 (Hz)	242	180	197	160
maximum F0 (Hz)	266	212	227	229
mean F0 (Hz)	258	195	211	186
sd F0 (Hz)	7.8	10.0	8.3	21.6
mean rms amplitude	1087	780	937	790
sd rms amplitude	258	220	248	299
total duration (sec.)	1.30	1.45	1.51	1.43
vowel duration (sec.)	0.82	0.84	1.09	0.99
percent correct responses	90.3	69.1	84.0	52.4

Table 1. Mean values on eight acoustic measures (note: the ninth measure referred to in the text, F0 range, is the difference between minimum and maximum F0), and mean percent correct responses, for H versus L *ka* syllables in initial versus final position.

Correlations

To obtain a uniform measure of listeners' performance, the responses were converted to percentage H judgements - that is, the percentage of correct responses for syllables which actually were H, and the percentage of error responses for those which actually were L.

Over all 96 tokens, there were significant positive correlations between responses and four of the nine acoustic measures: subjects were more likely to decide that a syllable was H when it had high minimum F0 ($r [95] = .66$, $p < .001$), high maximum F0 ($r [95] = .52$, $p < .001$), high mean F0 ($r [95] = .67$, $p < .001$) and high mean amplitude ($r [95] = .38$, $p < .001$). There were significant negative correlations with two other measures: subjects were more likely to decide that a syllable was H when it had low F0 range ($r [95] = -.32$, $p < .002$) and low F0 standard deviation ($r [95] = -.38$, $p < .001$). Thus high absolute F0 and high amplitude signalled a H syllable; pitch movement signalled a L syllable.

Responses to initial syllables showed the same pattern of relationship to F0 and amplitude as displayed in the overall correlations, while only four of the six significant correlations in the overall analysis were significant for L syllables (minimum F0 and mean F0, and F0 range and F0 standard deviation). The pattern of correlation was furthermore not the same for each speaker. Responses to all three speakers' productions correlated in the same way with the F0 measures, but only the responses to the productions of Speaker 2 showed a statistically significant relationship to amplitude.

Nor was the pattern the same for H versus L syllables separately. The likelihood of H responses to syllables which actually were H correlated only with the maximum and the mean F0, and only relatively weakly: F0max: $r [47] = .29$, $p < .05$; F0mean: $r [47] = .32$, $p < .05$. In contrast, the likelihood of H responses to syllables which actually were L correlated with minimum F0 ($r [47] = .37$, $p < .01$), with maximum F0 ($r [47] = .43$, $p < .002$) and with mean F0 ($r [47] = .44$, $p < .002$) as well as a marginal correlation on amplitude.

DISCUSSION

These analyses motivate a number of conclusions. First, there is in this task a bias towards responses which observe the conventional H marking for isolated syllables. This can be seen in the overall higher percentage of correct responses for H than for L syllables, and in the somewhat lower correlations of responses to H syllables with acoustic factors.

Second, listeners' judgements are principally based, as of course was expected, and as the pattern of correlations certainly showed, on F0 values: high absolute F0 signals a H accent level, F0 movement is more likely to cue a L accent level. Listeners can also make some use of the amplitude. Durational factors apparently play little role in signalling whether a syllable is H or L.

Third, not all speakers are equally successful at conveying the H/L difference. Our Speaker 1 produced these two syllable types in a less differentiated way than Speakers 2 and 3, and correspondingly she received a lower mean percentage of correct responses from the listeners.

Fourth, and most interestingly, cues to the H/L accent level distinction are conveyed most sharply, and most usefully for the listener, in initial as opposed to final syllables. The acoustic measures showed greater H/L differentiation in initial than in final syllables; the overall percentage correct was higher for initial than for final syllables; and the correlations between responses and acoustic factors were stronger in initial than in final syllables. This suggests that pitch accent information may be available to listeners in just the position where it would be of most use to them in on-line spoken-word recognition, and that listeners are in a position to exploit the available cues.

One further effect which has not yet been discussed is that scores were lower at the beginning of the experiment (66.3% correct responses for the first quarter of the tape), and higher at the end (79.5% for the last quarter). Thus listeners seemed to have been learning the task. It could be that part of this involved learning about the characteristics of the particular speakers' voices. Certainly the inconsistency among speakers which we observed suggests that listeners cannot rely on clear information being immediately available from all speakers.

These results represent only a beginning in our understanding of how listeners might use pitch accent information in Japanese spoken-word recognition. We have only examined bisyllabic words and can as yet say nothing about the perception of pitch accent in longer words. Our stimuli contained no devoiced syllables; but such syllables complicate both the production and perception of pitch accent (Maekawa, 1990). And our task did not involve on-line spoken-word recognition, therefore we cannot yet say whether listeners do exploit pitch accent in recognising words in normal speech situations. Nevertheless, the greater availability of cues to the H/L distinction in initial than in final syllables in our materials, and the clear exploitation of these cues by our subjects, strongly suggest that prosodic information might play a stronger role in lexical access in Japanese than it does in English.

ACKNOWLEDGEMENTS

This research was supported by a grant from the Human Frontier Scientific Program, by grant number 06610475 from the Japanese Ministry of Education to the second author, and by a grant to the second author from the Max Planck Society. We thank Mark Scholten and Daan Broeder for technical assistance, and James McQueen for comments on the text. Authors' addresses: Anne Cutler, Max-Planck-Institute for Psycholinguistics, PO Box 310, 6500 AH Nijmegen, The Netherlands and Takashi Otake, Faculty of Foreign Languages, Dokkyo University, 1-1 Gakuen-machi, Soka-Shi, Saitama 340, Japan. Email addresses: anne.cutler@mpi.nl and otake@dokkyo.ac.jp.

REFERENCES

Cutler, A. (1986). *Forbear* is a homophone: lexical prosody does not constrain lexical access. *Language and Speech*, 29, 201-220.

Haraguchi, S. (1988). Pitch accent and intonation in Japanese. In H. van der Hulst & N. Smith (Eds.) *Autosegmental Studies on Pitch Accent*. Dordrecht: Foris; 123-150.

Jongenburger, W. & van Heuven, V.J. (1995). The role of lexical stress in the recognition of spoken words: prelexical or postlexical?, *Proceedings of the 13th International Congress of Phonetic Sciences*, 368-371.

Maekawa, K. (1990). Production and perception of the accent in the consecutively devoiced syllables in Tokyo Japanese. *Proceedings of the First International Conference on Spoken Language Processing*, Kobe; Vol. 1, 517-520.

Nishinuma, Y. (1994). How do the French perceive tonal accent in Japanese? Experimental evidence. *Proceedings of the Third International Conference on Spoken Language Processing*, Yokohama; 1739-1742.

Nishinuma, Y., Arai, M. & Ayusawa, T. (1996). Perception of tonal accent by Americans learning Japanese. *Proceedings of the Fourth International Conference on Spoken Language Processing*, Philadelphia; Vol. 1, 646-649.

Otake, T., Hatano, G., Cutler, A. & Mehler, J. (1993). Mora or syllable? Speech segmentation in Japanese. *Journal of Memory and Language, 32*, 358-378.

Vance, T.J. (1987). *An Introduction to Japanese Phonology*. Albany: State University of New York Press.

Walsh Dickey, L. (1996). Limiting-domains in lexical access: Processing of lexical prosody. In M. Dickey & S. Tunstall (eds.) University of Massachusetts Occasional Papers in Linguistics 19: *Linguistics in the Laboratory*; 133-155.