Competition in the Perception of Spoken Japanese Words

Takashi Otake[1,2], *James M. McQueen*[1,3,4], *Anne Cutler*[1,4,5]

[1]Max Planck Institute for Psycholinguistics, 6500 AH Nijmegen, The Netherlands
[2] E-Listening Laboratory, Tokorozawa 359 0021, Japan
[3] Behavioural Science Institute, Radboud University Nijmegen, The Netherlands
[4] Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, The Netherlands
[5] MARCS Auditory Laboratories, University of Western Sydney, NSW 1797, Australia

otake@e-listeninglab.com, j.mcqueen@pwo.ru.nl, anne.cutler@mpi.nl/a.cutler@uws.edu.au

Abstract

Japanese listeners detected Japanese words embedded at the end of nonsense sequences (e.g., *kaba* 'hippopotamus' in *gyachikaba*). When the final portion of the preceding context together with the initial portion of the word (e.g., here, the sequence *chika*) was compatible with many lexical competitors, recognition of the embedded word was more difficult than when such a sequence was compatible with few competitors. This clear effect of competition, established here for preceding context in Japanese, joins similar demonstrations, in other languages and for following contexts, to underline that the functional architecture of the human spoken-word recognition system is a universal one.

Index Terms: spoken-word recognition, Japanese, vowels, devoicing, competition.

1. Introduction

The recognition of spoken words across languages is based on the same universal architecture, but this can be significantly modulated to cope with language-specific structure. The result is that there are some measurable differences in the way that listeners from varying language backgrounds process speech input. Cross-language comparisons have thus been very helpful in the quest to understand how words are recognised.

Psycholinguists now have a detailed picture of the universal architecture of word recognition (see [12] for a review). Speech input is processed in a continuous manner, whereby multiple candidate interpretations of the input are concurrently considered and evaluated. The more candidate words are available for a particular portion of the speech input, the harder it can be to recognise a word [7,16,21]. This is referred to as lexical competition [7,13,16]. The process of competition assists listeners to arrive at the correct parse for a speech sequence of which parts are potentially ambiguous.

Although the competition process plays this role in the recognition of words in continuous speech, listeners also have many further techniques at their disposal to facilitate the process of segmenting speech. The latter techniques draw on the properties of the vocabulary and of the phonology, and as phonological structure and vocabulary makeup differ widely across languages, these techniques are in large part language-specific. For example, listeners are very sensitive to sequence constraints, and recognise words more rapidly when these are aligned to a necessary boundary arising from such constraints [11, 22]. Thus because the sequence [mr] cannot be syllable-internal in English, *rock* would be recognised more rapidly after a syllable ending [m] than after one ending [g]. This would not be the case in a language in which [mr] can occur within a syllable (e.g., Czech); i.e., it is language-specific.

Similarly, listeners use rhythmic patterns in segmenting speech, and these patterns too are highly language-specific. In English and Dutch, the predominant rhythm is based on stress, and listeners use stress in segmenting speech [2,21], while in French, rhythm and segmentation are syllable-based [15], and in Japanese, rhythm and segmentation are mora-based [3,18].

There are also segmentation procedures in use nearly universally across languages, such as, for instance a constraint making a word hard to recognise if it leaves a vowelless residue of the speech signal, i.e., a portion that could not itself be a word [17]. This constraint, known as the Possible Word Constraint (PWC) is useful for inhibiting spurious recognition of embedded words, such as *ring* in *bring*; although such embedded competitors may be briefly available in the process of recognition, they can be quickly discarded (in this case because *ring* leaves *b*, which has no vowel and so could not be a word). Even in languages with quite complex constraints on what may be a stand-alone word, this simple rule holds [1].

The way in which listeners use these different procedures in conjunction to segment speech has been revealed by cross-language comparisons. Rhythmically indicated boundaries, for example, serve as anchor points against which the existence of a PWC-relevant residue can be assessed [14,17]. Apparent exceptions to the PWC can arise as a result, e.g., where an entire rhythmic unit such as a mora is itself vowelless [14].

Japanese offers a further crucial comparison of relevance to the operation of these procedures, in that some vowels are phonologically present but not in fact pronounced, presenting listeners with a vowelless sequence. This devoicing affects the high vowels [i] and [u] between some voiceless consonants or before a pause. When the effect of this on word recognition was examined [6] it proved to be unhelpful to listeners – the PWC operation was largely unaffected, so that, for instance, *asa* 'morning' was recognised more easily in *asau* than in *asaf*, even though [f] in the latter could arise from a devoiced *fu*.

This experiment, like those in many segmentation studies [1,2,13,14,16,22] employed the word-spotting task [10], in which listeners hear nonsense utterances that may contain a real word. In *obzel*, *foogrock*, *crithnish*, *bookvig*, for example, the second and fourth items contain words (*rock, book*). When a word is spotted, the listener signals this by pressing a response key, and then repeats the word aloud. With this task, the effect on word recognition of changes in the immediately adjacent context of a word can be examined. The context can precede the target word (*foogrock*) or follow it (*bookvig*). As listeners process incoming speech in a continuous manner, a preceding context will activate many lexical candidates which will compete with the target word for recognition, while a following context should offer less competition because the (complete) target word itself will receive the strongest support.

In [6], recognition was harder for targets with devoiced contexts either following (e.g., *asa* in *asau* vs. *asaf*) or preceding (e.g., *sake* 'salmon' in *nyagusake* vs. *nyaksake*), but following contexts actually caused the greater difficulty. One possible reason for this is that prior context activated some competitors that in turn made the devoiced portion potentially acceptable (e.g., in *nyaksake*, *ksa-* activated words beginning *kusa-*, e.g., *kusari* 'chain'). In fact, investigations of the effect of competition using word-spotting or closely related tasks have usually manipulated following contexts. The more words that a following context activates to compete for the final part of a target word, the harder that target is to recognise, both in English [16] and Dutch [21], while the more words the context activates which do not compete for part of the word, the easier it is to segment the target from the context [9]. The present experiment was therefore designed to assess the effect of competition induced by context preceding a target word; at the same time, it constituted the first such study of competition in Japanese spoken-word recognition, and also provided further evidence on the processing of devoiced sequences in speech.

2. Method

2.1. Stimuli

The targets were 72 bimoraic (CVCV) words in four sets of 18. Half of these (experimental targets) began with a voiceless consonant (e.g. *kaba*, 'hippopotamus'); the other half (control targets) began with a voiced consonant (e.g. *nabe*, 'pan'). Each target was paired with two preceding bimoraic (CCVCV) nonsense contexts; the first consonant of the second mora of the context was either voiceless (e.g., *gyachi*) or voiced (e.g., *gyagu*), and the last vowel was either /i/ or /u/. When preceded by voiceless consonants, these vowels could be devoiced; this would be legal before voiceless-initial targets (e.g. in *gyachikaba*) but illegal before control, voiced-initial targets (e.g. in *gyachinabe*). The vowels cannot be devoiced after voiced consonants for either type of target (e.g., in *gyagukaba* and *gyagunabe*). As this example shows, experimental and control targets were yoked, such that the same pairs of nonsense sequences were used across pairs of targets.

In half of the devoiceable sequences, the second-to-third mora portion was consistent with many competitor words (e.g., for *chika* in *gyachikaba*, there are 13 Japanese words beginning *chika*). In the other devoiceable sequences this portion was consistent with few competitors (e.g., for *tsuha*, in *gyatsuhada*, with the target *hada*, 'skin', there are 0 Japanese words beginning *tsuha*). The number of competitors per sequence was ascertained by counting the number of words beginning that way in [20]. In all other sequences (with control targets, or with experimental targets and voiced initial consonant of the second mora) the number of competitors matching the second-to-third mora portion (e.g. *guka* in *gyagukaba*) was always low. These conditions, with measures of the competitor environments, are summarized in Table 1.

Besides the 144 target-bearing sequences (two for each of the 72 targets), there were 144 four-mora filler sequences. In the fillers, the final two (CVCV) morae were not real Japanese words. The first two morae all had CCVCV structures. In analogy to the target-bearing sequences, half of the onset consonants of the second morae of the fillers were voiceless, and half were voiced; this was also true for the onset consonants of the third morae. Finally, there were 20 practice items with matched structures (four ending in real words).

Table 1. *Design, examples, and competitor measures.*

Vowel in 2nd mora	Competitor condition	Example (*target* in *sequence*)	# of competitors from 2nd mora	
			Mean	Range
Experimental stimuli (voiceless-initial targets)				
legally devoiced	many	*kaba* in *gyachikaba*	13.94	10-18
voiced	few	*kaba* in *gyagukaba*	0.72	0-2
legally devoiced	few	*hada* in *gyatsuhada*	1.06	0-3
voiced	few	*hada* in *gyazuhada*	1.06	0-2
Control stimuli (voiced-initial targets)				
illegally devoiced	few	*nabe* in *gyachinabe*	1.56	0-2
voiced	few	*nabe* in *gyagunabe*	0.39	0-2
illegally devoiced	few	*moji* in *gyatsumoji*	1.11	0-3
voiced	few	*moji* in *gyazumoji*	0.50	0-3

Multiple tokens of each stimulus sequence were recorded by the first author (a phonetically trained native Tokyo speaker) in a sound-damped booth onto Digital Audio Tape (DAT, sampling rate 48 kHz). All legally devoiceable second vowels in the sequences were devoiced. The materials were transferred to computer (down-sampling to 16 kHz, 16 bits) for measurement/manipulation using a digital speech editor.

Target-bearing sequences were prepared by cross-splicing, at the second-to-third mora boundary, within the yoked pairs of experimental and control items. Tokens of target words were taken from recordings in voiced contexts (e.g., *kaba* from *gyagukaba, nabe* from *gyagunabe*), tokens of the contexts with devoicing were taken from recordings where devoicing was legal and had indeed occurred (e.g., *gyachi* from *gyachikaba*), and tokens of the contexts with voiced vowels were taken from recordings with voiced-initial (control) targets (e.g., *gyagu* from *gyagunabe*). Note that as a result all presented stimuli were cross-spliced, and none included portions where the speaker had attempted illegal devoicing.

Splices were made at zero-crossings at the onset of the initial consonants of the targets, as located using both visual criteria (from spectrograms and waveforms) and auditory criteria. The resulting stimuli were checked and only those without audible discontinuities were used in the experiment. Those which had noticable disruptions were either remade using different tokens and/or different splice points, or were discarded (the initial stimulus set was in fact based on 80 target words, eight of which were discarded for this reason).

Half of the fillers were cross-spliced in a similar manner. Those that had legal devoicing environments were made by splicing different tokens of the same sequence. Those with illegal devoicing were made by splicing across tokens. The other filler items and the practice items were unedited.

Two counter-balanced lists were made. All 72 targets appeared once on each list, in either one of their two contexts, such that each list had an equal number of trials in each condition. All 144 fillers appeared on each list. Target and filler items were mixed pseudo-randomly, such that at least one filler occurred between any pair of target items. These lists (and a practice list) were recorded to DAT. A timing pulse was aligned with the onset of each target item.

2.2. Participants and Procedure

Thirty two undergraduate students at Dokkyo University, Tokyo, Japan, participated for course credit. They were told that they would hear a list of nonsense sequences, some of which would contain real Japanese words at their offset. They were asked to try to spot those real words as quickly and as accurately as possible. They were instructed to press a response button with their preferred hand as fast as possible if they heard a real word, and then to say aloud what that word was. They were not told prior to the trials what the target words were. Listeners heard first the practice list and then one of the two experimental lists (16 participants per list).

Listeners were tested in a quiet room in separate sound-attenuating carrels either in pairs or individually. Stimuli were presented over headphones. A computer running NESU experiment control software logged Reaction Times (RTs, measured from the timing pulse on each trial to the listener's button press). Prior to analysis the total stimulus duration on the trial was subtracted from each RT, to give RTs from target offset. The listeners' spoken responses were recorded.

3. Results

The spoken responses were analyzed first. All trials where listeners misidentified a target word (or failed to give a spoken response) were treated as errors. Two target words were missed by all listeners. A further five words were missed, across both context conditions, by more than two thirds of the listeners. The data from these seven items (one experimental item and six control items) were excluded from the analysis. The results for the remaining 65 items are shown in Table 2.

Analyses of variance (ANOVAs) were carried out on the RT and error data separately for the experimental and control conditions with listeners (F1) and words (F2) as the repeated measure. In the analyses of the experimental conditions there were two factors: context (ending in a devoiced or a voiced vowel) and number of competitors (in the devoiced context; many or few). As predicted, listeners spotted words in voiced contexts faster than in devoiced contexts, by 28 ms on average, but this difference was not statistically significant ($F1(1,30) = 3.29$, $p = 0.08$; $F2<1$). As also predicted, listeners consistently spotted words faster when there were few words overlapping with the first syllable of the targets than when there were many such competitors in the devoicing environment. This difference (average 80 ms) was significant ($F1(1,30) = 21.47$, $p < 0.001$; $F2(1,33) = 10.42$, $p < 0.005$). The interaction of these two factors was not significant. Nevertheless, pairwise comparisons across targets within each context condition showed a significant competition effect: In devoiced contexts, targets with many competitors (e.g., *kaba* in *gyachikaba*) were spotted more slowly (by 99 ms, on average) than targets with few competitors (e.g., *hada* in *gyatsuhada*; $F1(1,30) = 21.72$, $p < 0.001$; $F2(1,33) = 6.92$, $p < 0.05$). The difference for the same targets in voiced contexts with competitor environment controlled (e.g., *gyagukaba* vs. *gyazuhada,* a mean difference of 61 ms) was significant only by participants ($F1(1,30) = 6.60$, $p < 0.05$; $F2(1,33) = 2.31$, $p = 0.14$). These differences are plotted in Figure 1.

In the corresponding analyses of the errors in the experimental condition there was no effect of context (words were spotted, on average, 1% more accurately in voiced than in devoiced contexts; both F1 and $F2<1$). But listeners spotted words more accurately (by 6%, on average) when there were few competitors than when there were many competitors ($F1(1,30) = 7.03$, $p < 0.05$; $F2(1,33) = 2.47$, $p = 0.13$).

Table 2. *Mean word-spotting RTs (ms, from target offset) and percentage error rates.*

Vowel in 2nd mora	Competitor condition	Example (*target* in *sequence*)	RT	Errors
Experimental stimuli (voiceless-initial targets)				
legally devoiced	many	*kaba* in *gyachikaba*	631	25%
voiced	few	*kaba* in *gyagukaba*	584	20%
legally devoiced	few	*hada* in *gyatsuhada*	532	14%
voiced	few	*hada* in *gyazuhada*	523	18%
Control stimuli (voiced-initial targets)				
illegally devoiced	few	*nabe* in *gyachinabe*	632	43%
voiced	few	*nabe* in *gyagunabe*	609	28%
illegally devoiced	few	*moji* in *gyatsumoji*	610	30%
voiced	few	*moji* in *gyazumoji*	568	31%

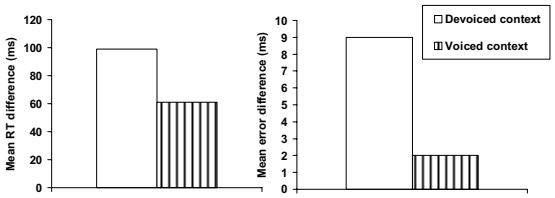

Figure 1: *Competition-induced differences in mean RT (left) and mean errors (right) for experimental targets in devoiced (light bars) vs. voiced (dark bars) contexts.*

The interaction of these two factors was significant by participants ($F1(1,30) = 5.09$, $p < 0.05$; $F2<1$). Follow-up pairwise comparisons like those on the RTs (see Figure 1) showed that, in the devoiced contexts, target words with many competitors were spotted less accurately (by 9%, on average) than targets with few competitors ($F1(1,30) = 10.33$, $p < 0.005$; $F2(1,33) = 3.33$, $p = 0.08$). The difference in accuracy on the same targets in voiced contexts with matched numbers of competitors (a mean difference of 2%) was not significant (both F1 and $F2<1$).

No effects were statistically significant in the RT analyses of the control conditions (note that here the competition factor was a dummy variable). In the error analyses of the control stimuli there was an effect of context by participants only ($F1(1,30) = 13.52$, $p < 0.001$; $F2(1,33) = 2.47$, $p = 0.13$): unsurprisingly, listeners found it harder (by 10%, on average) to spot targets in illegally devoiced than in voiced contexts.

4. Discussion

This study has revealed a clear effect of competition: when many other words are competing for the same portion of the incoming speech signal, listeners are slower and less accurate in recognising words. This is the first demonstration of such an effect of preceding context competing for initial portions of words, and it joins the effects already shown in English and Dutch for following context competing for final portions of words [9,16,21]. Moreover, this is the first demonstration of competition effects in the recognition of Japanese speech.

We did not observe any advantage for contexts involving devoiced vowels, so that the present study confirms the result of [6] that devoicing, like vowel deletion in general, is not helpful for listeners. As has long been known, consonants are easier to process separated by vowels than in sequence [8].

Evidence from Japanese has been important in furthering the psycholinguistic understanding of how listeners recognise spoken words. Across languages, rhythmic structure is used to assist in segmenting continuous speech into its component words, and the moraic rhythm of Japanese provided a crucial case study for the establishment of this knowledge [3,18]; the mora is, after all, the smallest unit on which linguistic rhythm is based. Because of this latter characteristic, Japanese also provided vital evidence for establishing how segmentation procedures (rhythmic and otherwise) work together [14], and how rhythmic categories interface with the continuous nature of lexical candidate activation [5]. Further, the phonology of Japanese offers several unique or at least rare features which further filled out the language-universal account of human speech recognition. One is the vowel devoicing case which has been discussed above, which made possible a deeper picture of the relationship between surface and underlying forms in the recognition of spoken words [6]; another is the pitch accent system of Japanese, which filled out our understanding of the role of lexical prosody in spoken-word recognition [4,19] to a greater extent than would have been possible with information only from languages with different prosodic systems.

By comparing evidence across many languages in this way, we can eventually understand the universal functional architecture of human speech recognition, and how it is modulated by the structural properties of different languages. We now have evidence from many languages confirming that what matters for listeners during speech recognition is the population of words potentially compatible with the incoming signal. Speech is processed continuously as it arrives, and each additional unit of phonetic information allows further adjustment of the set of candidate words under evaluation. Rhythmic evidence [2,3,14,15,18,21] and boundary sequence evidence [9,11,22] can all contribute to signalling the probability of a word boundary at each point in the incoming stream, but such cues do not result in any disruption of the continuous evaluation of the input. That is, listeners in stress languages do not process speech in stress-based chunks [2], and nor do listeners in moraic languages process speech in moraic chunks [5]; in all languages, the universal architecture, continuous probabilistic evaluation of the speech signal, holds in exactly the same way.

Accordingly, the process of competition is fundamental to spoken-word recognition. Even though listeners have at their disposal powerful methods for getting rid of competitors that are only spuriously present in the speech input [1,14,17], the interplay among concurrently activated lexical candidates, and the constant modulation of this competitor set in the light of the continuously arriving speech information, constitute the heart of the operation of listening to speech. Our new results add to growing literature demonstrating how subtle changes in the competitor population have immediate consequences for the facility with which listeners can identify the lexical content of speech signals. In Japanese, just as in English and Dutch and presumably across all languages, the more other words are competing for a stretch of speech, the harder it is for listeners to identify a word that is really there.

5. Acknowledgements

This research was supported by an NWO-SPINOZA award to the third author. We thank Maarten Jansonius and Jessica Pass for extensive speech editing work.

6. References

[1] Cutler, A., Demuth, K., and McQueen, J. M., "Universality versus language-specificity in listening to running speech", Psychol. Sci., 13: 258-262, 2002.

[2] Cutler, A. and Norris, D. G., "The role of strong syllables in segmentation for lexical access", J. Exp. Psychol.: Hum. Perc. and Perf., 14: 113-121, 1988.

[3] Cutler, A. and Otake, T., "Mora or phoneme? Further evidence for language-specific listening", J. Mem. Lang., 33: 824-844, 1994.

[4] Cutler, A. and Otake, T., "Pitch accent in spoken-word recognition in Japanese", J. Acoust. Soc. Am., 105: 1877-1888, 1999.

[5] Cutler, A. and Otake, T., "Rhythmic categories in spoken-word recognition", J. Mem. Lang., 46: 296-322, 2002.

[6] Cutler, A., Otake, T., and McQueen, J. M., "Vowel devoicing and the perception of spoken Japanese words", J. Acoust. Soc. Am., 125: 1693-1703, 2009.

[7] Goldinger, S. D., Luce, P. A., & Pisoni, D. B. "Priming lexical neighbors of spoken words: Effects of competition and inhibition." J. Mem. Lang., 28: 501-518, 1989.

[8] Liberman, A. M., Delattre, P. C., Cooper, F. S., and Gerstman, L. J. "The role of consonant-vowel transitions in the perception of the stop and nasal consonants", Psychol. Monog., 68: 1-13, 1954.

[9] Lugt, A. van der., "The use of sequential probabilities in the segmentation of speech", Percep. Psychophys., 63: 811-823, 2001.

[10] McQueen, J., "Word spotting", Lang. Cognitive Proc., 11: 695-699, 1996.

[11] McQueen, J. M., "Segmentation of continuous speech using phonotactics", J. Mem. Lang., 39: 21-46, 1998.

[12] McQueen, J. M., "Eight questions about spoken-word recognition", in M. G. Gaskell (Ed), The Oxford Handbook of Psycholinguistics, 37-53, Oxford Univ. Press, 2007.

[13] McQueen, J. M., Norris, D., and Cutler, A., "Competition in spoken word recognition: Spotting words in other words", J. Exp. Psychol.: Learn. Mem. Cogn, 20: 621-638, 1994.

[14] McQueen, J. M., Otake, T., and Cutler, A., "Rhythmic cues and possible-word constraints in Japanese speech segmentation", J. Mem. Lang., 45: 103-132, 2001.

[15] Mehler, J., Dommergues, J. Y., Frauenfelder, U. H., and Segui, J., "The syllable's role in speech segmentation", J. Verb. Learn. Verb. Behav., 20: 298-305, 1981.

[16] Norris, D., McQueen, J. M. and Cutler, A., "Competition and segmentation in spoken word recognition", J. Exp. Psychol.: Learn. Mem. Cogn, 21: 1209-1228, 1995.

[17] Norris, D., McQueen, J. M., Cutler, A., and Butterfield, S., "The possible-word constraint in the segmentation of continuous speech", Cognitive Psychol., 34: 191-243, 1997.

[18] Otake, T., Hatano, G., Cutler, A., and Mehler, J., "Mora or syllable? Speech segmentation in Japanese", J. Mem. Lang., 32: 258-278, 1993.

[19] Sekiguchi, T. and Nakajima, Y., "The use of lexical prosody for lexical access of the Japanese language", J. Psychol. Research, 28: 439-454, 1999.

[20] Sugito, M. Osaka-Tokyo Akusento Onsei Jiten (Osaka-Tokyo Accent Pronunciation Dictionary), Maruzen,1995.

[21] Vroomen, J. and Gelder, B. de, "Metrical segmentation and lexical inhibition in spoken word recognition", J. Exp. Psychol.: Hum. Percep. and Perf., 21: 98-108, 1995.

[22] Weber, A. and Cutler, A., "First-language phonotactics in second-language listening", J. Acoust. Soc. Am., 119: 597-607, 2006.