Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 51 (2009) 76–82
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Development of familiarity-controlled word lists 2003 (FW03) to assess spoken-word intelligibility in Japanese ☆

Shigeaki Amano [a,*], Shuichi Sakamoto [b], Tadahisa Kondo [a], Yôiti Suzuki [b]

[a] NTT Communication Science Laboratories, NTT Corporation, 2-4 Hikari-dai, Seika-cho, Souraku-gun, Kyoto 6190237, Japan
[b] Research Institute of Electrical Communication/Graduate School of Information Sciences, Tohoku University, 2-1-1 Katahira, Aoba-ku, Sendai, Miyagi 980-8577, Japan

Received 27 November 2007; received in revised form 13 July 2008; accepted 13 July 2008

Abstract

A new set of "Familiarity-controlled word lists 2003" (FW03) has been developed for a spoken-word intelligibility test in Japanese. FW03 consists of 20 lists of 50 words in four word-familiarity ranks (i.e., 4000 words in total). The entropy of (a) initial moras and (b) sequences consisting of a vowel and a following consonant was maximized in the word lists within each word-familiarity rank. FW03 is now published with speech files of the 4000 words spoken by two male and two female Japanese. The word intelligibility of FW03 was measured with the speech files at various signal-to-noise ratios. In addition to the signal-to-noise ratio effects, strong word-familiarity effects were observed in terms of word intelligibility, indicating that word familiarity is well controlled in FW03. FW03 enables us to measure word intelligibility in several word-familiarity ranks that correspond to the degree of lexical information.

Keywords: Word intelligibility; Word familiarity; Word list

1. Introduction

Speech intelligibility is assessed using several linguistic units such as phonemes, syllables, words or sentences. Various test materials have been proposed for measuring speech intelligibility with these units. For example, there are monosyllabic word materials such as the phonetically balanced word list (PB) (Egan, 1948), the modified rhyme test (MRT) (House et al., 1965), and the diagnostic rhyme test (DRT) (Voiers, 1983). There are also sentence materials such as the test sentences for speech perception in noise (SPIN) (Bilger et al., 1984; Kalikow et al., 1977), the connected speech test (CST) (Cox et al., 1987), the topic-related everyday sentences developed by The City University of New York (Boothroyd et al., 1988), the hearing in noise test (HINT) (Nilsson et al., 1994), and the semantically unpredictable sentence test (SUS) (Benoit et al., 1996). These test materials relate to the English language.

Fewer speech materials have been proposed for Japanese than for English. However, the Japan Audiological Society proposed syllable lists named 57-S (Japan Audiological Society, 1983) and 67-S (Japan Audiological Society, 1987). 57-S contains five lists of 50 monosyllables and six lists consisting of six digits (2, 3, 4, 5, 6, and 7). 67-S contains eight lists of 20 monosyllables and the same digit lists as 57-S. 67-S is most frequently used to assess hearing abilities and evaluate the fitting of hearing aids.

However, although almost all Japanese monosyllables are meaningful, they are sometimes recognized as nonsense monosyllables especially when they are heard in isolation without a context. This means that an assessment using Japanese monosyllables is not good for controlling lexical information that must be used for speech communication. It is well known that lexical information of a word (e.g., Amano, 1993) as well as contextual information in a

☆ Parts of this research were presented at the 9th Western Pacific Acoustic Conference, Seoul, Korea, June 26–28, 2006.

* Corresponding author. Tel.: +81 774 93 5310; fax: +81 774 93 5345.
E-mail address: amano@cslab.kecl.ntt.co.jp (S. Amano).

0167-6393/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2008.07.002

sentence (e.g., Connine, 1987) usually improve speech intelligibility and play a major role in speech communication. Therefore, an assessment using words or sentences would be more appropriate than one using monosyllables when measuring the ability of central levels of processing in speech communication. This paper focuses an assessment with words.

In consideration of the appropriateness of an assessment with words, Yonemoto et al. (1989) proposed word lists (TY-89) in Japanese. TY-89 contains 50 two-syllable words and 50 three-syllable words. These words were selected from a list of child vocabulary that everyone should know. TY-89 is used as frequently as 67-S for measuring the ability to hear speech.

However, as Yonemoto (1995) pointed out, TY-89 contains certain words that are hard to recognize. That is, some words always have a lower intelligibility score than others. There are two reasons for this problem. One reason is that word familiarity is not well controlled and it is low for some words in TY-89. Word familiarity is a subjective rating value indicating how familiar a person is with each word. It is typically obtained by averaging participants' familiarity ratings for a word using a 7-point scale (1: most unfamiliar, 7: most familiar). Controlling word familiarity is very important in test materials, because word recognition depends heavily on word familiarity. That is, the higher the word familiarity is, the more correctly and quickly a spoken word is recognized (Amano and Kondo, 1999; Amano et al., 1999). The other reason is that the phonetic balance is not controlled in TY-89. Since each phoneme differs in terms of its recognizability and a phoneme is a unit constituting a word, the intelligibility score of a word is affected by the phonemes included in the word. Therefore, to assess the speech intelligibility of words with various phonemes, the phonetic balance should be controlled in test materials.

The two reasons given above probably result in an inequality of word items in TY-89. Some words inherently have low intelligibility. To overcome these problems, a new set of word lists (familiarity-controlled word lists 2003, hereafter FW03) was proposed for spoken-word intelligibility tests. Word familiarity was taken into consideration when developing FW03, because it has a strong effect on spoken-word recognition. That is, there is a strong tendency for the recognition accuracy and speed to become higher as the word familiarity increases (Amano and Kondo, 1999). Although word frequency has the same tendency, it has a weaker effect than word familiarity with regard to word recognition (Amano and Kondo, 2000). Therefore, word frequency was not used when developing FW03.

Word familiarity has a very high correlation ($r = .958$, $N = 10{,}515$) when measured over a number of years and in different places (Amano et al., 2007). This means that word familiarity is fairly stable, although it is a subjective measure. This characteristic is another reason why word familiarity was used for developing FW03.

Phonetic balance was also considered when developing FW03, because the intelligibility score of a word must be related to the variety of phonemes included in the word. Previous word lists were also developed considering the phonetic balance. For example, Egan (1948) controlled the distribution of phonemes when developing the PB word list. A mathematical method was used to achieve the phonetic balance in this study.

By controlling these two factors, a reasonable word intelligibility score can be obtained with FW03. This paper describes the procedure for developing FW03 and evaluating word intelligibility with FW03.

2. Development of FW03

2.1. Word candidates

Word candidates (13,607) were selected from a word-familiarity database (Amano and Kondo, 1999) with about 88,000 word entries, which were derived from all the word entries in a medium sized Japanese dictionary. The conditions for selection were as follows:

1. Word length is usually counted in moras in Japanese (for the definition of mora, see Otake and Cutler, 1996; Vance, 1987). The word length was set at four moras when selecting words because four-mora words are the most frequently occurring type in Japanese. In the word-familiarity database (Amano and Kondo, 1999), four-mora words account for 42.02% of all words ($N = 88{,}569$).
2. The accent type was Low–High–High–High (i.e., the first mora has a low pitch and the following moras have a high pitch) because the Low–High–High–High accent type is most common in four-mora words. In the word-familiarity database (Amano and Kondo, 1999), the Low–High–High–High accent type accounts for 69.97% of four-mora words ($N = 37{,}219$). Words with more than one accent type were excluded to avoid accent type ambiguity, which might affect the word intelligibility score.
3. Homophones (i.e., a set of words that have the same sequence of phonemes) were regarded as a single word because homophones have the same word familiarity and they are not distinguished in the word-familiarity database.
4. Words with a negative image, antisocial words, and disease-related words were excluded because these kinds of words might be affected by social suppression or other kinds of inhibitions, which would result in unexpectedly low word intelligibility scores.

2.2. Word selection

The word candidates were divided into four sets according to word-familiarity rank: low familiarity (1.0–2.5),

lower-middle familiarity (2.5–4.0), upper-middle familiarity (4.0–5.5), and high familiarity (5.5–7.0). These sets contain 2501, 4108, 4885, and 2113 words, respectively. The four word-familiarity ranks were defined by arbitrarily dividing the word-familiarity range (1.0–7.0) into four ranges. However, as seen in the following sections, they function fairly well in classifying word intelligibility.

From each of the four sets, 20 lists of 50 words (i.e., 1000 words) were selected by considering the phonetic balance. This phonetic balance was achieved by taking account of "entropy". Entropy is a measure that indicates the uncertainty associated with random events. If the events happen independently, this measure is equivalent to an average amount of information. If the distribution of the occurrence of events is biased, entropy decreases. Inversely, if the distribution of the occurrence of events is not biased, entropy increases. That is, high entropy means that each event happens almost randomly. Therefore, by maximizing the entropy associated with a phoneme, an unbiased phoneme occurrence can be achieved.

Two kinds of entropy, H_1 and H_2, were used for the phonetic balance. H_1 was calculated as in

$$H_1 = -\sum_m p(m)\log_2 p(m), \tag{1}$$

where $p(m)$ is the occurrence probability of a word-initial mora m. This entropy was introduced because the word-initial part is important in spoken-word recognition (e.g., Marslen-Wilson and Welsh, 1978). It is thought that the word-initial part evokes word candidates that share the initial part, then lexical processing narrows the word candidates down to a single word. In addition, the phoneme pair that distinguishes two similar Japanese words most frequently exists at a word-initial position (Makino and Kido, 1979). Moreover, the word-initial mora has a lower intelligibility score than the other moras when a word is presented at a low sound pressure level (e.g., Sakamoto et al., 2004). This means that the word-initial mora is harder to identify than the other moras. These facts rationalize the introduction of the entropy of a word-initial part.

H_2 was calculated from the transitional probability of two successive phonemes within a word as in

$$H_2 = -\sum_v \sum_c p(v)p(c|v)\log_2 p(c|v), \tag{2}$$

where $p(v)$ is the probability of vowel v, and $p(c|v)$ is the conditional occurrence probability of the consonant c preceded by vowel v. This entropy was introduced to emulate transitions between moras within a Japanese word, which consists of concatenations of consonant–vowel sequences in most cases.

Total entropy is defined as H_{total} as in

$$H_{\text{total}} = H_1 + H_2. \tag{3}$$

For each word-familiarity rank, the lists were obtained by maximizing the H_{total} for each list by employing the "Add and Delete" method (Shikano, 1984). The procedure is as follows:

Step 1. Individually, add a word to each list so that the each list has a maximum gain of H_{total} until a word set reaches 1000 words (i.e., 20 lists of 50 words).
Step 2. Search for a pair of words that gives a maximum gain of H_{total} for a list if one of the words is deleted from the list and the other word is added to the list.
Step 3. Exchange the words found in Step 2.
Step 4. Repeat Steps 2 and 3 until the gain of H_{total} reaches zero.

Since each word-familiarity rank had 20 lists of 50 words, there were 80 lists of 50 words (i.e., 4000 words) in total. The mean word familiarity was 5.81 (SE = 0.039) for the high familiarity, 4.84 (SE = 0.053) for the higher-middle familiarity, 3.15 (SE = 0.066) for the lower-middle familiarity, and 2.16 (SE = 0.024) for the low familiarity. The developed word lists can be found at http://www.ais.riec.tohoku.ac.jp/lab/wordlist/index-e.html.

2.3. Recordings

Four professional narrators (male: "mya" and "mis"; female: "fto" and "fhi") pronounced all the words in FW03. The words were digitally recorded (16-bit quantization rate and 48-kHz sampling frequency) in a soundproof room, and stored in a computer as speech files in the WAV format. The amplitude of the sound waveform of the speech files was adjusted so that each word in FW03 had the same L_{Aeq} level. The recorded word lists can be obtained from the Speech Resources Consortium (http://research.nii.ac.jp/src/eng/index.html).

3. Word intelligibility of FW03

An experiment was conducted to measure the word intelligibility of FW03 at various signal-to-noise ratios. If the development of FW03 is successful, the effect of word familiarity on the word intelligibility of FW03 would be clearly observable as well as the effect of the signal-to-noise ratio.

3.1. Method

3.1.1. Participants

Thirty-two Japanese adults (16 males and 16 females) participated in the experiment. Their average age was 27.1 (SD = 0.8, min = 20, max = 37). Their minimum audible threshold was checked by using an automatic audiometer (Rion, AA79). All participants had normal hearing. To check the participants' language ability, we employed the score of the "Reading ability test for kanji words (100-Rakan)" (Kondo and Amano, 1998). This test measures the ability of a participant to read 100 kanji words, and it is designed to provide a score in the 0–100 range.

The averaged 100-Rakan score of the participants was 88.5 (SD = 1.39, max = 99, min = 67). This means that all the participants had a good language ability. The participants were paid for their participation.

3.1.2. Stimulus

Words in FW03 pronounced by all four narrators were used as original stimuli. The original stimuli were digitally added to random noise with the speech spectral shape (ITU-T Recommendation G.227), which was fixed at 60 dBA. Seven signal-to-noise ratios were set by changing the L_{Aeq} level of the original stimuli. They were −12, −9, −6, −3, 0, 3, and 6 dB for low familiarity, −15, −12, −9, −6, −3, 0, and 3 dB for lower-middle familiarity, and −18, −15, −12, −9, −6, −3, and 0 dB for upper-middle familiarity and high familiarity. These signal-to-noise ratios were decided according to the results of a preliminary experiment with 10 Japanese participants (5 males and 5 females) aged between 20 and 30 with normal hearing ability.

The amplitude of the noise was increased linearly at the beginning and decreased at the end to prevent any audible click. The duration of these amplitude transitions was 50 ms. The noise started at 350 ms before the onset of an FW03 word. The noise continued for 250 ms after the end of the word. Twenty lists with 50 words in four word-familiarity ranks with seven signal-to-noise ratios spoken by four narrators resulted in a total of 112,000 stimuli.

3.1.3. Procedure

Four participants undertook the experiment at the same time in a soundproof room with a background noise of less than 30 dBA. A notebook computer (IBM, R50e) was assigned to each participant for stimulus presentation and response collection. Stimuli were presented from the notebook computer through a D/A converter (Creative Technology, SoundBlaster Audigy2NX) and headphones (Sennheiser, HDA200) to the left ear. The stimulus order was randomized for each participant.

Sixteen participants (8 males and 8 females) were assigned to the "mya" and "fto" stimulus sets. Half of the participants listened to the "mya" stimulus set first and then the "fto" stimulus set. The other half listened to the sets in the reverse order. The other participants (8 males and 8 females) were assigned to the "mis" and "fhi" stimulus sets. Half of the participants listened to the "mis" stimulus set first and then the "fhi" stimulus set. The other half listened to the sets in the reverse order.

The participants typed what they heard in katakana characters (Japanese phonetic symbols). The next stimulus was presented about 1 s after they had confirmed their current answer. The participants performed a 15-min block of experiments that consisted of about 200 trials. After a 5-min break, the next block started. Each participant was assigned 280 blocks. It took about 18 days for a participant to complete all the trials.

3.2. Results and discussion

3.2.1. Word intelligibility

A correct answer was defined as an answer where every katakana character of a participant's response matched that of the presented word. Fig. 1 shows average intelligibility at each word-familiarity rank as a function of signal-to-noise ratio for each narrator. The intelligibility was obtained as a percentage of the correct answers of each participant for 1000 words in each familiarity rank at each signal-to-noise ratio. The effect of the signal-to-noise ratio

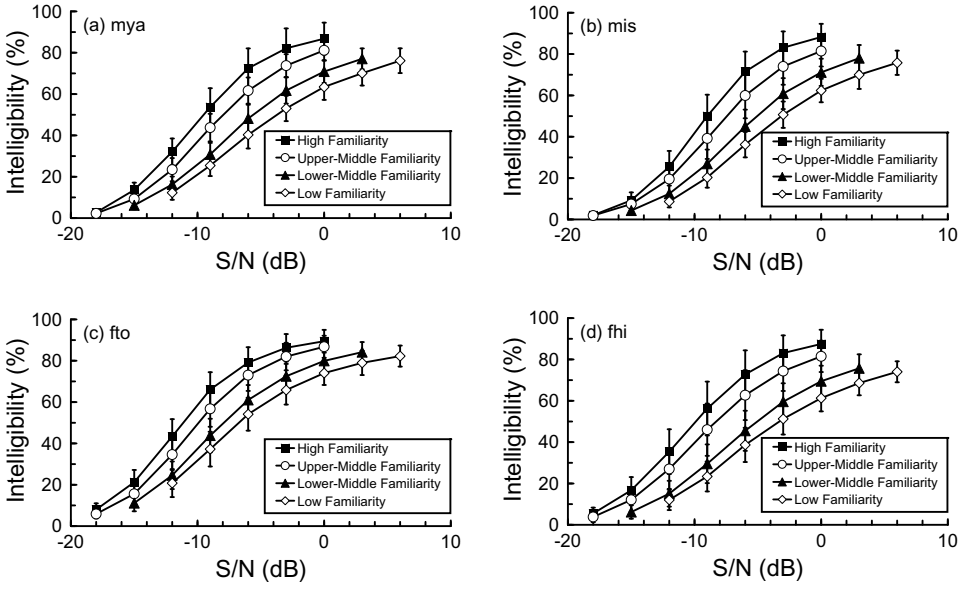

Fig. 1. Word intelligibility of FW03 at each word-familiarity rank as a function of signal-to-noise ratio. The bar represents the standard deviation.

is very clear. That is, word intelligibility decreases as the signal-to-noise ratio decreases.

The strong effects of word familiarity on word intelligibility can be seen in Fig. 1. For example, the word intelligibility is about 70–80% for high-familiarity words but only about 35–40% for low-familiarity words at a signal-to-noise ratio of −6 dB. This means that the difference in word familiarity caused about 30–45% difference in word intelligibility at the same signal-to-noise ratio in the maximum case.

To check the factors that affect word intelligibility, two types of analysis of variance were conducted with three factors, namely narrator, word familiarity, and the signal-to-noise ratio between −12 and 0 dB. One type analyzed the word intelligibility which was obtained as a percentage of the correct answers for 1000 words for each participant (subject analysis), and the other analyzed the word intelligibility which was obtained as a percentage of the correct answers of 16 participants (item analysis). Namely, the participant was a random variable in the subject analysis and there were 16 intelligibility values in each condition, whereas the word was a random variable in the item analysis and there were 1000 intelligibility values in each condition. The narrator was a factor between participants, and the others were factors within the participants in the subject analysis. The word familiarity was a factor between words and the others were factors within words in the item analysis. It was revealed that the interaction among the three factors was significant in both types of analysis [$F_1(36, 720) = 14.82$, $p < .0001$; $F_2(36, 47{,}952) = 12.96$, $p < .0001$; where F_1 is for the subject analysis and F_2 is for the item analysis].

These results mean that the factors of narrator, word familiarity, and signal-to-noise ratio have complicated relationships. Of these three factors, narrator is a confounding factor with participants, because, as described in the procedure section, the participants were equally divided into two groups and each group was assigned to different narrators. This makes it impossible to distinguish between narrator and participant factors in the current experimental design. For this reason, the narrators were dealt with individually in the following analyses of variance.

Two types of analysis of variance with two factors, namely word familiarity and signal-to-noise ratio were conducted for each narrator. They revealed that the interaction between word familiarity and signal-to-noise ratio was significant for all narrators [for mya, $F_1(12, 180) = 40.38$, $p < .0001$; $F_2(12, 15{,}984) = 23.68$, $p < .0001$; for mis, $F_1(12, 180) = 67.13$, $p < .0001$; $F_2(12, 15{,}984) = 40.79$, $p < .0001$; for fto, $F_1(12, 180) = 50.25$, $p < .0001$; $F_2(12, 15{,}984) = 27.13$, $p < .0001$; for fhi, $F_1(12, 180) = 29.94$, $p < .0001$; $F_2(12, 15{,}984) = 23.70$, $p < .0001$; where F_1 is for the subject analysis and F_2 is for the item analysis].

The simple main effect of word familiarity was significant at all levels of signal-to-noise ratio for all narrators ($p < .001$). A Tukey's honestly significant differences (HSD) test revealed that the difference between every word-familiarity pair at each signal-to-noise ratio was significant ($p < .05$) except for the following word-familiarity pairs: low familiarity and lower-middle familiarity at −12 dB for mis, fto, and fhi; and low familiarity and lower-middle familiarity at −9 dB for mya; upper-middle familiarity and high familiarity at 0 dB for mya and fto; and at −3 dB for mya and at −12 dB for fto in subject analysis; and upper-middle familiarity and high familiarity at 0 dB for fto in item analysis.

The simple main effect of the signal-to-noise ratio for the word intelligibility of each participant was also significant at all ranks of word familiarity for all narrators ($p < .001$). A Tukey's HSD test revealed that the difference between every signal-to-noise ratio pair at each word familiarity was significant ($p < .05$) except for the following signal-to-noise ratio pairs: −3 and 0 dB at high familiarity for mya, fto, and fhi; and at upper-middle familiarity for fto only in subject analysis.

These results of an analysis of variance indicate that although word familiarity and signal-to-noise ratio interact, word intelligibility was affected by both of them in all narrators. A paired comparison revealed that almost all pairs of word familiarity have significant differences. This means that word familiarity must be controlled for word intelligibility assessments.

Even if the signal-to-noise ratio was high, the intelligibility of high familiar words appears to peak at about 90% and it does not reach 100% in Fig. 1. This was probably because certain phonemes such as voiceless fricatives and plosives were not correctly identified when noise was added to stimulus words. Using the same FW03 word lists, Sakamoto et al. (2004) conducted an intelligibility-measuring experiment at various sound pressure levels without adding noise. They obtained almost 100% intelligibility for all word-familiarity ranks when the sound pressure level was high. Therefore, word intelligibility will reach approximately 100% if the signal-to-noise ratio is higher than the current conditions.

Learning effects, which facilitate word intelligibility, might exist in this experiment because the participants heard the same word several times at different signal-to-noise ratios. However, the effects would be expected for all words because the presentation times were the same for all words. Therefore, even if learning effects are present, they are not thought to lead to errors in the analysis of variance results, which indicate the significance of word familiarity and signal-to-noise ratio for word intelligibility.

3.2.2. Speech recognition threshold (SRT)

A logistic curve was fitted for word intelligibility as a dependent variable and signal-to-noise ratio as an independent variable. The fitting was performed for the intelligibility of 16 participants for each word with each narrator in each word familiarity. Almost all of the fitting was successful (95.3%), however, some were failed because the fitting program did not converge, or because the estimated speech

recognition threshold (SRT) was too low (less than −20 dB) or too high (more than 10 dB). The number of successful fittings is shown in Table 1.

The SRT was obtained as the signal-to-noise ratio that gives 50% intelligibility on each fitted logistic curve for each word. SRT indicates how difficult it is to recognize a spoken word in a noisy condition. Table 1 shows the mean and standard deviation of the obtained SRT. For the same reason as that mentioned in the previous section, the narrators were treated individually in the following analyses of variance.

An analysis of variance for the SRT of words with one factor, namely word familiarity, revealed that the effect of word familiarity was significant for each narrator [for mya, $F(3, 3842) = 209.4$, $p < .0001$; for mis, $F(3, 3847) = 301.4$, $p < .0001$; for fto, $F(3, 3767) = 187.5$, $p < .0001$; and for fhi, $F(3, 3780) = 335.1$, $p < .0001$]. A Tukey's HSD test revealed that the difference between every word familiarity pair was significant ($p < .05$). These facts mean that the SRT is affected by word familiarity and it becomes lower as the word familiarity becomes higher. This reveals that words are more intelligible when the word familiarity is higher.

The slope at the SRT was obtained for each fitted logistic curve. Table 1 shows the mean and standard deviation of the slope at the SRT.

An analysis of variance of the slope at the SRT with one factor, namely word familiarity, revealed that the effect of word familiarity was significant for each narrator [for mya, $F(3, 3842) = 92.4$, $p < .0001$; for mis, $F(3, 3847) = 98.5$, $p < .0001$; for fto, $F(3, 3767) = 126.4$, $p < .0001$; and for fhi, $F(3, 3780) = 54.4$, $p < .0001$]. This indicates that there is a strong tendency for the slope at the SRT to become steeper as a function of word familiarity. A Tukey's HSD test revealed that the difference between every word-familiarity pair was significant ($p < .05$) except for pairs of high and upper-high word familiarity for all narrators. These results suggest that the slopes at the SRT are different for different familiarity ranks but they tend to be similar when word familiarity is high.

Table 1
SRT and slope at the SRT for each word-familiarity rank

Narrator	Familiarity	N	SRT (dB)		Slope (%/dB)	
			Mean	SD	Mean	SD
mya	High	985	−8.12	4.14	10.88	6.32
	Upper-middle	990	−7.11	4.40	10.54	6.61
	Lower-middle	963	−4.65	5.37	8.57	5.49
	Low	908	−2.92	6.02	7.00	4.37
mis	High	996	−8.31	3.78	12.11	6.86
	Upper-middle	987	−6.77	4.23	11.50	7.24
	Lower-middle	964	−4.17	4.97	9.31	5.16
	Low	904	−2.49	5.44	7.82	4.82
fto	High	958	−10.29	3.92	9.95	4.21
	Upper-middle	970	−8.93	4.02	9.89	5.16
	Lower-middle	956	−6.83	4.96	8.02	4.88
	Low	887	−5.64	5.55	6.36	4.15
fhi	High	971	−9.18	4.04	10.05	4.90
	Upper-middle	982	−7.39	4.63	9.52	4.57
	Lower-middle	947	−4.31	5.15	8.09	4.86
	Low	884	−2.66	5.76	7.33	6.42

The results of the analysis of variance indicate that word familiarity affects the SRT and the slope at the SRT. It means that word familiarity should be treated properly when SRT and the slope at the SRT are measured in an intelligibility assessment with words.

To check the difference between word lists, the intelligibility, SRT, and the slope of SRT of the word list for each participant were obtained for each narrator at each combination of word familiarity and signal-to-noise ratio.

An analysis of variance of intelligibility with one factor, namely word list, revealed that the effect of the word list is significant in all cases in both subject and item analysis ($p < .001$). An analysis of variance of the SRT with one factor, namely word list, revealed that the effect of the word list is significant in all cases in both subject and item analysis ($p < .001$). An analysis of variance of the slope at the SRT with one factor of word list revealed that the effect of word list is significant in all cases in both subject and item analysis ($p < .001$).

These results indicate that the intelligibility, SRT, and the slope at the SRT differ among word lists, even though the word familiarity was controlled with four familiarity ranks and the phonetic balance was considered. The difference might result from the fact that several phonemes, which strongly affect the intelligibility, might be contained only in certain specific lists. Or, the difference might be because words in each word list have slightly different word familiarities. Within each familiarity rank, word familiarity was assumed to have the same strength of effect on word intelligibility. However, a slight difference in word familiarity might actually have a different strength of effect on word intelligibility.

Word familiarity is strongly related to lexical information. That is, the higher the word familiarity is, the richer the lexical information becomes. In other words, word familiarity is related to a mental lexicon. For example, there is a strong tendency for people to give "known" responses to high-familiarity words but not to low-familiarity words, and the response ratio is a function of word familiarity (cf. Amano and Kondo, 1998). Therefore, FW03 enables us to control the degree of lexical information in a word intelligibility test. However, an intelligibility assessment does not always require all four word-familiarity ranks. It depends on the situation in which word intelligibility is measured.

For example, the high-familiarity word lists in FW03 contain rich lexical information. Tests with these lists provide us with word intelligibility including strong facilitation provided by lexical information. These lists would be suitable for an intelligibility assessment related to, for example, hearing-aid fitting or speech hearing ability in everyday conversations, because familiar words are probably fre-

quent in casual speech. On the other hand, the low-familiarity word lists of FW03 contain poor lexical information. Therefore, tests with these lists result in word intelligibility with weak facilitation provided by lexical information. These lists would be suitable for an intelligibility assessment in a lecture room or a public hall, because unknown or unfamiliar words are probably frequent in a lecture or a formal speech. In addition, it is possible to estimate the degree of the effects of lexical information by comparing the intelligibilities of high- and low-familiarity word lists. This adjustability in terms of lexical information is an advantage of FW03.

Although FW03 has these useful characteristics, certain problems were revealed. For example, intelligibility, SRT and the slope at the SRT are different among the word lists in each word-familiarity rank. The difference in word familiarity within the ranks may not be negligible especially when applying FW03 to a clinical situation in which only one list is used for an assessment. The discrepancies among word lists in each word-familiarity rank should be overcome in a future study. However, it might be difficult to overcome the discrepancies by controlling word familiarity, because there is a trade-off between controlling word familiarity and maximizing entropy to achieve a phonetic balance. That is, controlling word familiarity within a small range decreases word variation, whereas maximizing entropy requires the word variation. One possible way to achieve uniform word intelligibility among word lists is to adjust the power of each word according to the SRT. Another possibility is to adjust the loudness of each word.

Another problem is that 50 words in each list of FW03 might be too many for clinical situations where only 10 or 20 items are usually used for a hearing test. A long time is required to complete the answers for the 50 words, which imposes a greater burden on a patient than usual. A smaller number of words would be appropriate in such cases. Reducing the number of words in a list might be necessary in a future study.

4. Conclusion

A set of new word lists, FW03, was developed for a word intelligibility test in Japanese. FW03 consists of 20 lists with 50 words in four word-familiarity ranks. FW03 is a useful set of word lists for the assessment of hearing ability concerning lexical information processing in spoken-word recognition, because the word familiarity and phonetic balance are properly controlled.

References

Amano, S., 1993. Effects of lexicon and coarticulation on phoneme perception. J. Acoust. Soc. Jpn. (E) 14, 91–97.

Amano S., Kondo, T., 1998. Estimation of mental lexicon size with word familiarity database. In: Proc. Internat. Conf. on Spoken Language Processing, Vol. 5, pp. 2119–2122.

Amano, S., Kondo, T., 1999. Lexical Properties of Japanese, Vol. 1. Sanseido, Tokyo (in Japanese).

Amano, S., Kondo, T., 2000. Lexical Properties of Japanese, Vol. 7. Sanseido, Tokyo (in Japanese).

Amano, S., Kondo, T., Kato, K., 1999. Familiarity effect on spoken word recognition in Japanese. Proc. 14th Internat. Congr. of Phonetic Science Vol. 2, 873–876.

Amano, S., Kasahara, K., Kondo, T., 2007. Reliability of familiarity rating of ordinary Japanese words for different years and places. Behav. Res. Methods 39, 1008–1011.

Benoit, C., Grice, M., Hazan, V., 1996. The SUS test: a method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences. Speech Comm. 18, 381–392.

Bilger, R.C., Neutzel, J.M., Rabinowitz, W.M., Rzeczkowski, C., 1984. Standardization of a test of speech perception in noise. J. Speech Hear. Res. 27, 32–48.

Boothroyd, A., Hnath-Chisolm, T., Hanin, L., Kishon-Rabin, L., 1988. Voice fundamental frequency as an auditory supplement to the speechreading of sentences. Ear Hear. 9, 306–312.

Connine, C.M., 1987. Constraints on interactive processes in auditory word recognition: The role of sentence context. J. Memory Lang. 26, 527–538.

Cox, R.M., Alexander, G.C., Gilmore, C., 1987. Development of the connected speech test (CST). Ear Hear. 8, 119s–126s.

Egan, J.P., 1948. Articulation testing methods. Laryngoscope 58, 955–991.

House, A.S., Williams, C.E., Hecker, M.H.L., Kryter, K.D., 1965. Articulation-testing methods: consonantal differentiation with a closed-response set. J. Acoust. Soc. Amer. 37, 158–166.

International Telecommunication Union, Telecommunication Standardization Sector (ITU-T). Recommendation G.227.

Japan Audiological Society, 1983. The 57-S syllable list.

Japan Audiological Society, 1987. The 67-S syllable list.

Kalikow, D.N., Stevens, K.N., Elliott, L.L., 1977. Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. J. Acoust. Soc. Amer. 61, 1337–1351.

Kondo, T., Amano, S., 1998. The reading ability test for kanji words: an estimation method for language ability using word familiarity. In: Proc. 62nd Ann. Meet. of the Japanese Psychological Association, Vol. 711 (in Japanese).

Makino, S., Kido, K., 1979. Properties of phoneme pairs for distinguishing a word pair with a short distance. IEICE Trans. Inform. Systems J62-D, 507–514 (Japanese edition).

Marslen-Wilson, W.D., Welsh, A., 1978. Processing interactions and lexical access during word recognition in continuous speech. Cognit. Psychol. 10, 29–63.

Nilsson, M., Soli, S.D., Sullivan, J.A., 1994. Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise. J. Acoust. Soc. Amer. 95, 1085–1099.

Otake, T., Cutler, A., 1996. Phonological Structure and Language Processing: Crosslinguistic Studies. Mounton de Gruyter, New York.

Sakamoto, S., Amano, S., Suzuki, Y., Kondo, T., Ozawa, K., Sone, T., 2004. The effect of familiarity on mora identification in word intelligibility tests. J. Acoust. Soc. Jpn. 60, 351–357 (in Japanese).

Shikano, K., 1984. Phonetically balanced word list based on information entropy. Proc. Spring Meet. of the Acoustic Society of Japan, 211–212 (in Japanese).

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

Voiers, W.D., 1983. Evaluating processed speech using the diagnostic rhyme test. Speech Technol. 1, 30–39.

Yonemoto, K., 1995. Characteristics of CD (TY-89) for evaluation of hearing-aid fitting. JOHNS 11, 1395–1401 (in Japanese).

Yonemoto, K., Tateishi, T., Kiba, K., Kurauchi, N., 1989. Syllable intelligibility and sound pressure level in TY-89 and 57-S word list. Audiol. Jpn. 32, 429–430 (in Japanese).