New 20-word lists for word intelligibility test in Japanese

SAKAMOTO Shuichi†, *YOSHIKAWA Tadahiro*†, *AMANO Shigeaki*‡,
SUZUKI Yôiti†, *KONDO Tadahisa*‡

†Research Institute of Electrical Communication / Graduate School of Information Sciences, Tohoku University
{saka,tada,yoh}@ais.riec.tohoku.ac.jp
‡NTT Communication Science Laboratories, NTT Corporation
amano@cslab.kecl.ntt.co.jp, tkondo@brl.ntt.co.jp

Abstract

"Familiarity controlled Word lists 2003 (FW03)" is a set of 20 lists of 50 Japanese words each in four ranks of word familiarity (for a total of 4000 words). FW03 is useful for evaluating the ability of a person to hear human speech. However, it has been pointed out that the FW03 lists have too many words to evaluate the ability of a person to hear human speech in a short time. This can cause problems, such as when evaluating the ability of an elderly person to hear human speech during a medical exam. To remedy this problem, we propose new word lists that are a subset of FW03. The new lists were created by considering both phonetic balance and word familiarity. Each of the new word lists comprises only 20 words. This reduces the duration of an examination to around two-fifths of a conventional FW03-based examination. Word intelligibility tests suggest that these new lists provide word intelligibility scores nearly equal to those resulting from FW03.
Index Terms: word intelligibility test, word familiarity, speech reception/recognition threshold

1. Introduction

Word or sentence intelligibility tests are useful when evaluating the ability of an individual to hear human speech. Various word and sentence lists for this purpose have been proposed by researchers, audiologists and others [1, 2]. Unfortunately, most such lists are insufficient for evaluating the ability to hear speech because they do not properly reflect the difficulty of recognizing a given word or sentence.

To cope with these problems, we published lists of Japanese words, "Familiarity controlled Word lists 2003" (FW03)[3]. In FW03, word familiarity was the index used to determine the difficulty of word recognition. By controlling word familiarity, 4000 four-mora-words were selected for FW03 across four ranks of familiarity. Each rank of familiarity consists of 20 lists of 50 words each. Not only did we control for word familiarity, we also controlled for phonetic balance. That is, we controlled for the probability of a word's initial phoneme occurring as well as the transitional probabilities of the two successive phonemes. It is critical to control for phoneme balance to evaluate the efficacy of a hearing aid. Word intelligibility scores obtained using FW03 suggest that it is possible to evaluate not only the perceptual level of a person's ability to hear speech, but also the degree of cognitive recognition [4].

Despite these successes, it has been pointed out that the large numbers of words on each FW03 list make it difficult to evaluate ability of a person to hear speech in a short time. A medical examination of an elderly person is an example of a situation in which this is problematic. Merely removing words from a FW03 list to shorten the test duration might cause problems, as well. For example, using only a portion of the list might result in poor phonetic balance. Moreover, reducing the list might increase the variance of word familiarity in that list. This, in turn, might increase the variance of word intelligibility.

In this study we develop new word lists that are subsets of FW03. Just as was done when creating FW03, we controlled these new lists for both word familiarity and phonetic balance. We also controlled for the variance of word familiarity. We then evaluated the performance of these new word lists according to word intelligibility.

2. Outline of FW03

To control for word familiarity, we used a comprehensive word-familiarity database developed by Amano and Kondo [5]. The database contains values for word familiarity ranging from 1 (low familiarity) to 7 (high familiarity) for all the entries in the Fourth Edition of the Shinmeikai Japanese Dictionary (around 80000 words) [6].

The FW03 was developed as follows: First, LHHH-accent-type words (Types 0 & 4) consisting of four moras were selected. These words are the most common Japanese words. Next, that population of words was divided into four ranks of familiarity: 7.0–5.5 (high familiarity), 5.5–4.0 (middle-high familiarity), 4.0–2.5 (middle-low familiarity), and 2.5–1.0 (low familiarity). We then selected 1000 words for each rank of familiarity to create 20 lists containing 50 words each by maximizing the phoneme entropies to achieve optimum phonetic balance.

Four professional narrators (male: "mya" and "mis"; female: "fto" and "fhi") pronounced all the words in FW03. The spoken words were recorded digitally (using 16-bit and 48-kHz sampling) in a soundproof room, and were saved as speech files in the WAV format. The amplitude of the waveform for each spoken word was adjusted so that all speech files have an equal L_{Aeq} level.

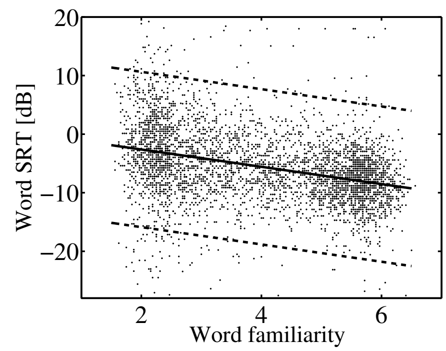
3. Development of new lists

3.1. Word selection

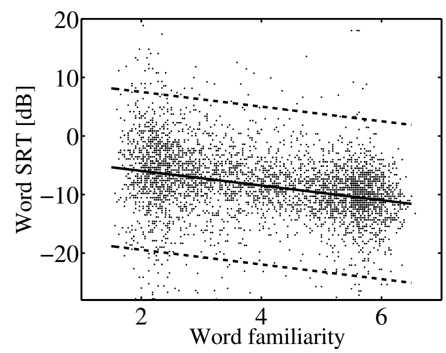
All 4000 of the words in FW03 were used as candidates for the new, abbreviated lists. When developing FW03, we used a β version of the word-familiarity database. As a result, the familiarity rank of several words has changed with the release of the latest word-familiarity database. These words were excluded as candidates for the new lists. (Only two high-familiarity words and three low-familiarity words needed to be excluded for this reason).

To clarify the recognition difficulty of each word contained in FW03, we measured the speech reception/recognition threshold (SRT) of all 4000 words as pronounced by two of the speakers, "mya" and "fto." SRT is the signal-to-noise ratio (SNR) at which a listener can recognize a word 50% of the time. In this experiment, 16 adults (eight males and eight females) with normal hearing acuity participated. Their average age was 29.4 ($SD = 4.8$, $Min = 23$, $Max = 37$) and they were all native Japanese speakers. Speech signals were replayed at a volume of 60 dBA. Speech spectral shape noise was added to the speech signal digitally. The SNRs used were -12, -9, -6, -3, 0, 3, and 6 dB for low familiarity words, -15, -12, -9, -6, -3, 0, and 3 dB for middle-low familiarity words, and -18, -15, -12, -9, -6, -3, and 0 dB for middle-high familiarity and high familiarity words.

Figure 1 shows the SRT as a function of word familiarity. In each panel, the solid line shows a regression line of SRT predicted on familiarity, while the dotted line shows 95% confidence limits. These results demonstrate that SRT values are widely distributed even if words are similarly familiar. These results were used to exclude words with an SRT outside of the 95% confidence limits. This resulted in the exclusion of 222 low-familiarity words, 101 middle-low-familiarity words, 40 middle-high-familiarity words and 60 high-familiarity words.



(a) *speaker: mya*



(b) *speaker: fto*

Figure 1: *Relationship between word SRT and word familiarity. Solid line: regression line of SRT, Dotted line: 95% confidence limits.*

3.2. Controlling of phonetic balance and distribution of word familiarity

In FW03 each list has 50 words. The new lists were comprised of only 20 words to shorten the duration of tests. The traditional list of nonsense monosyllables proposed by the Japan Audiology Society consists of 20 monosyllables. This list is widely used, especially in clinical situations. From this we infer that 20 words are sufficient, even for elderly listeners. We made 20 lists of 20 words each for all four familiarity ranks.

When we developed FW03, we considered both phonetic balance and word familiarity. When developing these new lists, we controlled for these two factors as well as the distribution of word familiarity in each list (something that was not considered in the construction of FW03). Because word familiarity influences word intelligibility (e.g.[4]), the distribution of word familiarity should be equalized. However, this restricts availability of usable words and degrades phonetic balance. Therefore, we searched for the optimum compromise between phonetic balance and well-distributed word familiarity.

To construct an index for the distribution of word familiarity, we defined two parameters. One parameter is the standard deviation (SD) of word familiarity across the 20 lists for each familiarity rank (Parameter A). The other is the SD of word familiarity within each list (Parameter B). First, average word familiarity was calculated for each list. Parameter A was calculated from the 20 average values of word familiarity for each familiarity rank. To determine the value of Parameter B, the SD of each list was calculated from the familiarity of the 20 words in each list. Parameter B was then calculated from the SDs determined for each familiarity rank.

As the index of phonetic balance, we defined two entropies: H_1 and H_2. H_1 was calculated from the probability of occurrence of a word's initial phoneme (Eq. (1)).

$$H_1 = -\sum_i p_i \log_2 p_i \tag{1}$$

Therein, p_i represents the occurrence probability of phoneme i. H_2 was calculated from the transitional probability of two successive phonemes within a word (Eq. (2)).

$$H_2 = -\sum_j \sum_i p_i p(i|j) \log_2 p(i|j) \tag{2}$$

In that equation, p_j is the probability of vowel j and $p(i|j)$ is conditional occurrence probability of the consonant i preceded by vowel j. The total entropy is then defined as H_{total} (Eq. (3)).

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

For each rank of familiarity, the lists were created by maximizing a sum of H_{total} for the 1000 words using the "Add & Delete" method described by Shikano [7]. This method was applied within the limits of a value of Parameter A and Parameter B to control for the variance of word familiarity.

Figure 2 portrays the relationship between H_{total}, Parameter A and Parameter B. This figure indicates that H_{total} becomes lower when Parameter A or Parameter B becomes smaller. Setting Parameter A or Parameter B to a low value means that the population of words is limited. Therefore, H_{total} becomes lower.

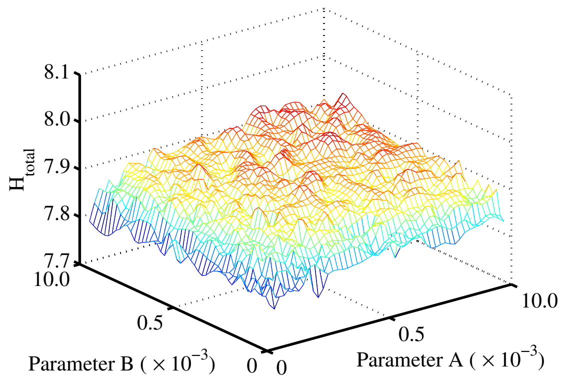


Figure 2: *H_{total} as a function of the SD of word familiarity between lists (Parameter A) and the SD of familiarity within a list (Parameter B) (Familiarity: 7.0-5.5).*

Table 1: *Optimum value of Parameter A and Parameter B.*

Word familiarity rank	Parameter A ($\times 10^{-3}$)	Parameter B ($\times 10^{-3}$)
7.0 – 5.5	2.1	2.2
5.5 – 4.0	2.3	2.1
4.0 – 2.5	2.0	1.9
2.5 – 1.0	1.9	2.3

Figure 3 shows the effects of Parameter A and Parameter B separately. The following equation was fit to the calculated results using a regression function (Eq. (4)).

$$y = a\{1 - \exp(-bx)\} + c \tag{4}$$

where y is H_{total} and x is Parameter A or Parameter B. Using a regression analysis, a, b and c were calculated.

The following equation (Eq. (5)) was applied to obtain the optimum compromise between phonetic balance (H_{total}) and the distribution of word familiarity (Parameter A and Parameter B).

$$\exp(-bx) = 1/e^2 \tag{5}$$

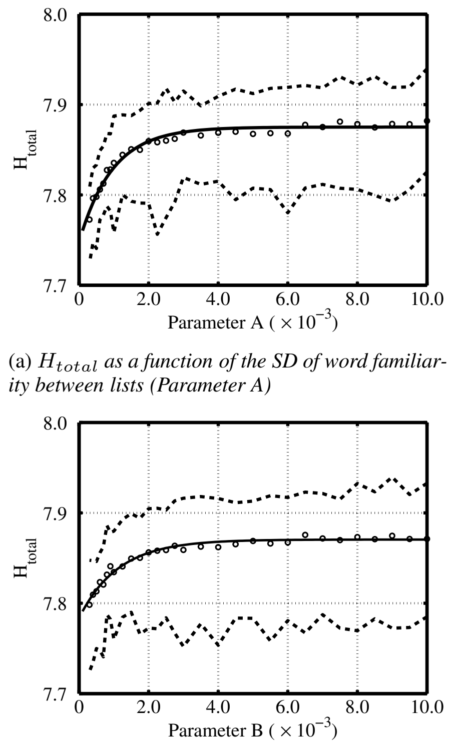
For this equation, b (which is calculated from Eq. (4)) was substituted for Eq. (5) to calculate x. This x was decided as the optimum value of Parameter A or Parametar B.

Table 1 shows the optimum value of Parameter A and Parameter B for each familiarity rank.

4. Evaluation of new lists

The intelligibility of each list was calculated by taking the average intelligibility of each of the 20 words selected from FW03. We then calculated the SD of the intelligibility across all word lists within a familiarity rank. The level of variance so calculated indicates the uniformity of the new word lists. As reference lists, we produced the word lists without controlling for word familiarity distribution and only controlled for phonetic balance.

Figure 4 shows word intelligibility scores for the three different types of word lists: FW03, the new lists and reference lists. This figure indicates that all these different word lists have similar intelligibility.



(a) *H_{total} as a function of the SD of word familiarity between lists (Parameter A)*

(b) *H_{total} as a function of the SD of word familiarity within a list (Parameter B)*

Figure 3: *Relationship between H_{total} and Parameter A or Parameter B (Familiarity: 7.0-5.5). Cycles show calculated data. The solid line shows the regression line. Dotted lines show the max and min of H_{total}.*

Figure 5 shows the SDs of the three types of word lists. Using ANOVA, the difference between the calculated SDs of the three types of word lists was statistically significant for almost all speakers, word familiarity ranks and SNRs ($p < .05$). The results of Tukey's HSD test show that the SD of FW03 was significantly smaller than that of other lists ($p < .05$) because each FW03 word list includes 50 words, whereas the other lists include only 20 words. Also, the difference between the SDs of the new lists and the reference lists was statistically significant for almost all SNRs ($p < .05$). However, there was no consistent trend to the differences that crossed speakers, word familiarity ranks and SNRs.

5. Discussion

Figure 4 demonstrates that the intelligibility scores for the new word lists were similar to the intelligibility scores of FW03 despite the fact that the new lists only contain 20 words. This suggests that the new lists are useful for testing because similar results are achieved in two-fifths the time required by FW03. As a result, these new lists should be useful for measuring word intelligibility for elderly persons who are unable to undergo long tests.

Interestingly, controlling word familiarity variance in the list did not clearly tend to reduce the SD of word intelligibility scores. Also, when the process is controlled for the variance of word familiarity, the available pool of usable words is limited. For that reason, it is difficult to conduct equally balanced phonetic word

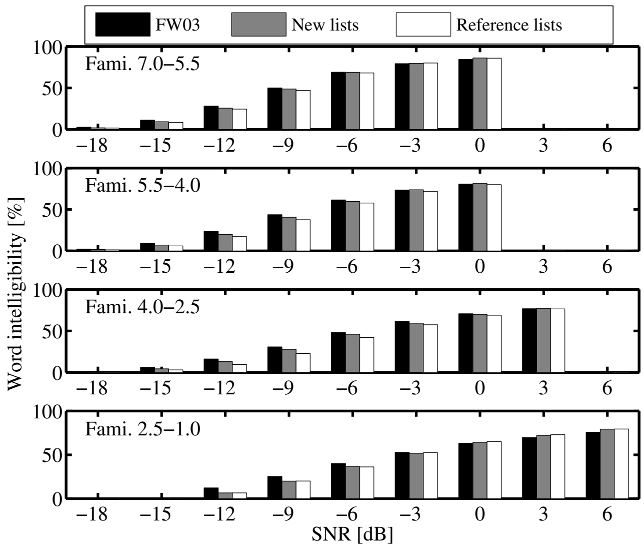


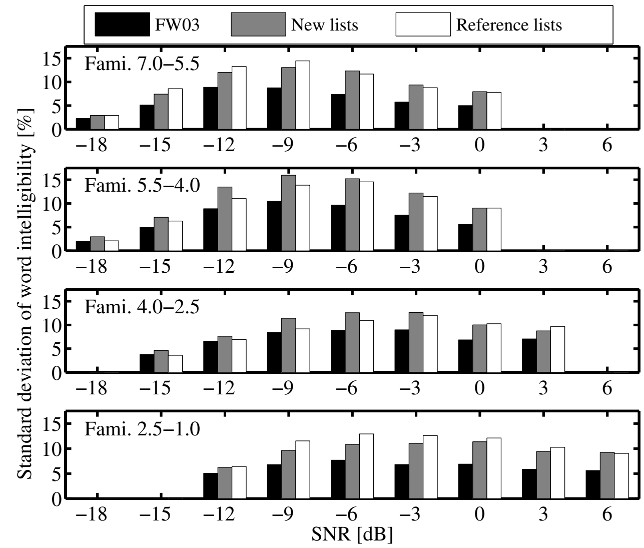
Figure 4: *Word intelligibility scores using three types of word lists (speaker: mya).*

lists. This is one reason that the SD of word intelligibility scores for the new lists increases under certain conditions. This result suggests that it might be impossible to equalize the intelligibility of each word list in the same familiarity rank by only controlling for the variance of word familiarity distribution. To remedy this problem, it might be useful to change the presentation level of each word according to the difficulty of word recognition.

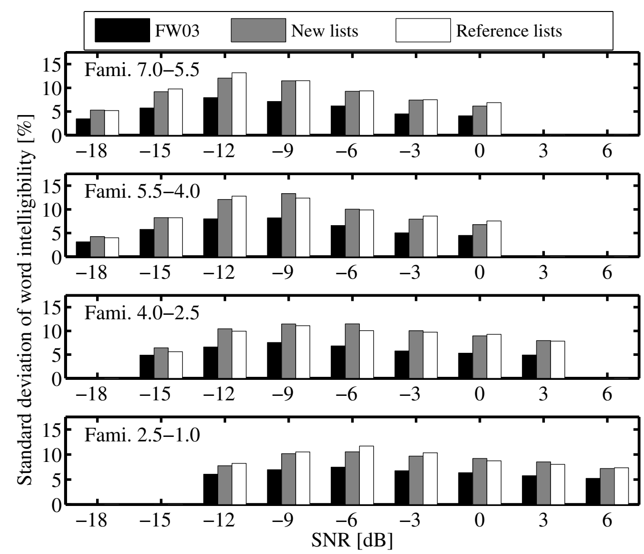
In this study, we measured the SRT of whole words recorded in FW03. Using this data, there are possible methods to compensate for the previous problems: high-SRT words could be set at high presentation level, whereas low-SRT words could be set at low presentation level. A previous study [8] confirmed that controlling for the presentation level of words according to the difficulty of word recognition was useful in decreasing the difference between word lists of word intelligibility.

6. Conclusions

We proposed new word lists for word intelligibility tests. Word intelligibility tests suggest that these new lists are as intelligible as lists from FW03 while also reducing test duration to around two-fifths of that of FW03.



(a) *Results of Speaker:mya*



(b) *Results of Speaker: fto*

Figure 5: *Standard deviation of word intelligibility between word lists in same familiarity rank.*

7. References

[1] S. D. Soli, M. Nilsson, and J. A. Sullivan, "Development of hearing in noise test for the measurement of speech reception thresholds in quiet and in noise," J. Acoust. Soc. Am., Vol.102, 2412–2421, 1994.

[2] T. Iwaki, M. Shiroma, T. Kubo, and S. D. Soli, "HINT-Japanese: Development of the hearing in noise test," Audiology Japan, Vol. 42, 421–422, 1999.

[3] S. Amano, S. Sakamoto, T. Kondo, and Y. Suzuki, "Development of FW03: Japanese word lists for spoken-word intelligibility test," Proc. of the 9th Western Pacific Acoustics Conference (WESPAC IX), 2006.

[4] S. Sakamoto, Y. Suzuki, S. Amano, K. Ozawa, T. Kondo, and T. Sone, "New lists for word intelligibility test based on word familiarity and phonetic balance," J. Acoust. Soc. Jpn., Vol. 54, 842–849, 1998.

[5] S. Amano and T. Kondo, Lexical Properties of Japanese Vol. 1, Sanseido, Tokyo, 1999.

[6] K. Kindaichi, T. Shibata, A. Yamada, and T. Yamada, Shinmeikai Japanese Dictionary (Fourth Ed.), Sanseido, Tokyo, 1989. S. Amano and T. Kondo, Lexical Properties of Japanese Vol. 1, Sanseido, Tokyo, 1999.

[7] K. Shikano, "Phonetically balanced word list based on information entropy," Proc. of Spring Meeting of the Acoust. Soc. Jpn., 211-212, 1984.

[8] S. Sakamoto, N. Iwaoka, Y. Suzuki, S. Amano, and T. Kondo, "A compensation method for word-familiarity difference with SNR control in intelligibility test," Proc. INTERSPEECH2004 (ICSLP), TuB901o., 2004.