Guidelines for Japanese Speech Synthesizer Evaluation

Shuichi ITAHASHI

Institute of Information Sciences and Electronics, University of Tsukuba
1-1-1 Tennodai, Tsukuba, 305-8573, Japan
itahashi@milab.is.tsukuba.ac.jp

Abstract

Speech synthesis technology is one of the most important elements required for better human interfaces for communication and information systems. This paper describes the "Guidelines for Speech Synthesis System Performance Evaluation Methods" created by the Speech Input/Output Systems Expert Committee of the Japan Electronic Industry Development Association (JEIDA). JEIDA has been investigating speech synthesizer evaluation methods since 1993 and previously reported the provisional version of the guidelines. The guidelines comprise six chapters: General rules, Text analysis evaluation, Syllable articulation test, Word intelligibility test, Sentence intelligibility test, and Overall quality evaluation.

1. Introduction

Evaluating synthetic speech quality is an important part of promoting the research and development of speech synthesis. To date many evaluation methods have been investigated. However, as various speech synthesizers with improved speech quality have recently come on the market, more objective and quantitative criteria are required, not only by developers but also by users. With this background, the Speech Input/Output Systems Expert Committee of JEIDA (Japan Electronic Industry Development Association) began reviewing the performance evaluation of speech synthesis-by-rule in fiscal 1992, and published a provisional version of "Guidelines for Speech Synthesizer Evaluation Methods" in fiscal 1994 (Nishi & Itahashi, 1998).

Recently, a revised version entitled "Guidelines for Speech Synthesis System Performance Evaluation Methods" (JEIDA-G, 2000) was issued, with many examples of evaluation lists added, so that the guidelines can be directly applied to the performance evaluation of speech synthesizers. We also added a new chapter on "Evaluation of text analysis". Along with these guidelines, we prepared a "Commentary on the Guidelines for Speech Synthesis System Performance Evaluation Methods" (hereinafter "Commentary") to explain to users the background against which the guidelines were created (JEIDA-C, 2000).

Speech synthesis systems give an intermediate output of analysis results of the text input as well as the final synthesized speech based on that analysis. Here we present a summary of our guidelines for evaluation of each of these two kinds of output.

2. Text Analysis Evaluation

When using a text-to-speech synthesis system, errors occurring in the text analysis phase can make it difficult for a user to understand the meaning of the speech message, even when the intelligibility of the phoneme-to-speech conversion is quite high. Therefore, evaluation of the text analysis part of a text-to-speech synthesis system is as important as the intelligibility test.

A pause usually indicates a clause boundary which reflects the meaning of the clause, so that insertion of an unnatural pause may cause misunderstanding of a sentence. However, since the placement of pauses varies from one person to another and from one speech tempo to another, it is very difficult to judge pause errors. Therefore, the evaluation of pause errors was not included in the guidelines.

2.1. Text analysis metrics

Since Japanese text is written in Chinese characters ("Kanji") and "Kana" syllabaries, it is quite important to assign a suitable reading to each Chinese character that have multiple readings, as is usually the case.

Accent in Japanese indicates the grouping of word sequences into accentual phrases, and different meanings of homonyms are sometimes also distinguished by a difference in accent. Accent errors are therefore likely to result in misunderstandings.

Rates for correct reading and correct accentuation assignments are defined using the "bunsetsu phrase", which consists of one or more content words followed by zero or more function words.

X = number of bunsetsu phrases containing one or more reading errors

Y = number of bunsetsu phrases with no reading errors, but with one or more accent errors

Z = number of bunsetsu phrases with no reading or accent errors

$X+Y+Z$ = total number of bunsetsu phrases

Correct reading rate = $1-X/(X+Y+Z)$

Correct accent rate = $1 - Y/(Y+Z)$

2.2. Materials for text analysis evaluation

Example sentences are given in the "Commentary" (Appendix E, sentence set 2). Generally, the items listed below are difficult to correctly analyze, so texts which contain them are well suited for use in evaluation.

(1) "katakana" words with their variations in orthography: Although a text analysis system can easily infer the pronunciation of katakana words, because katakana denotes a phonemic description, the correct accent is idiosyncratic and cannot be computed.

(2) kana used for inflectional endings with their variations in orthography

(3) Sound changes at morpheme boundaries in a compound word

(4) Long compound words: Although component words are listed in the dictionary, morphological analysis errors sometimes occur in long compound words, because Japanese text does not indicate word boundaries. Misanalyses often cause pronunciation errors, particularly errors in accent

(5) Unlisted words: If the text analysis system cannot find a word in the dictionary, it is very difficult to infer the correct pronunciation. When an unlisted word is detected in the text, the system must infer its pronunciation from the default reading for each kanji or kana character. However, this is not an especially effective approach, especially for proper nouns. Moreover, the unlisted word can begin a chain of errors in segmenting the words that follow it.

(6) Others: Differences in pronunciation between common nouns and proper nouns expressed with the same Chinese characters; sentences with ambiguous syntactic structure; reading of kana text; and reading of special symbols, numbers (digit by digit, or the digit string as a whole quantity), and acronyms.

2.3. Text analysis evaluation procedures

The evaluators (see section 3.1) should be persons familiar with the particular notation system of the text analysis results of the synthesizer. They should:

a) Determine the boundaries of bunsetsu phrases and count the total number of bunsetsu phrases before the test.

b) Obtain the text analysis output from the synthesizer.

c) Calculate the correct reading/accent rates as described in Section 2.1.

The evaluation repot should include:

a) the whole text used for the evaluation

b) the result of phrase boundary detection

c) the number of evaluators, and their age groups and evaluation experience.

3. Intelligibility Evaluations

Following the text analysis evaluation, final outputs of the speech synthesis itself are evaluated for intelligibility, focusing on speech units of different lengths: syllables, words, and sentences. These intelligibility tests have a number of conditions and procedures in common, which are described in this section.

3.1. Evaluators

Evaluators are people who listen to the synthesized speech and evaluate the articulation and intelligibility. The word "subject" is often used in this context, but "subject" implies someone who is tested psychologically or psycho-acoustically. Since the purpose here is to judge the quality of the synthesized speech itself, we prefer the term "evaluator".

Evaluators should be selected according to the following conditions:

(1) Basic Requirement: An adult having normal hearing ability and stable judgment.

(2) Sex and Age: Not specified. However, to be most representative, evaluators should be both men and women and of wide range of ages.

(3) Experience in Listening to Synthesized Speech: Not specified. It depends on the application purposes. However, because it greatly affects the test results, the previous listening experience of each evaluator should be stated in the test report.

(4) Number of Evaluators: Five or more. The larger the number, the more objective the results will be.

3.2. General testing procedures

Tests for intelligibility should be carried out as follows:

(a) Prepare synthesized speech as described in the section "Materials for evaluation".

(b) Present the synthesized speech to the evaluators, and have the evaluators record their responses according to the "Test procedures" section.

(c) Analyze the results according to the section on metrics.

(d) Finally, report the results and the conditions of the tests (Kasuya, 1992; Higuchi et al., 1989).

(1) Preliminary Training: Less experienced evaluators should receive preliminary training, so they will be familiar with the testing procedures. To prevent the evaluators from becoming too familiar with synthesized speech, natural human speech is used in the preliminary training.

(2) Number of Test Trials: The number of trials is not specified, but the more trials, the better the reliability.

(3) Testing Equipment: The type of testing equipment, i.e., loudspeakers, headphones, or telephones, and the acoustical environment is not specified because the choice depends on the application being tested.

(4) Method of Presenting Test Speech: The interval between adjacent test speech samples is five seconds. Each sample should be presented only once. To avoid evaluator fatigue, the number of test speech samples in one session should be not more than two hundred, so the total time for one session will be less than twenty minutes. When repeating tests in the same set, the presentation order should be changed.

(5) Instructions for Reporting: The evaluator must be instructed as to how to write down what is heard. For example, when the synthesized speech is not intelligible, the evaluator must know whether he or she can skip that item or whether it must be written down.

(6) Learning Effects: The following items must be checked because of significant habituation effects. Show the results of the first test trial, or show the results of a test trial after the learning effects have become saturated; show the number of test trials (the total number of words presented) at a stage before saturation.

3.3. Test conditions to be reported

The following testing conditions should be indicated in the test report:

(1) Evaluators: The number of evaluators and, for each one, their sex, age, previous experience with synthesized speech, and hearing ability.

(2) Sound Output Apparatus: The type and model of the loudspeakers or headphones (single-ear or both) should be listed.

(3) Acoustic Environment(s): The environment should consist of a soundproof room, or meeting room, private office, or the like. The sound level of the test speech sample and the ambient noise level should be recorded. Also, the type of background noise should be recorded, such as stable white noise (as of air conditioner), human voices, or machine noises.

(4) Speech Synthesizer Specifications: If available, the voice (whether male or female) and sampling frequency of the synthesizer should be reported. The speech synthesis method, e.g. formant rule-based synthesis, LPC parameter concatenation, or waveform concatenation, should be reported. Finally, the basic synthesis unit, such as diphone or demisyllable, should be reported.

4. Syllable Articulation Test

Syllable articulation is the most basic criterion for evaluating speech information transmission. A test of syllable articulation evaluates the intelligibility of each phoneme diagnostically.

4.1. Syllable articulation metrics

The following items should be included in the test report:

(1) Average syllable articulation: For each evaluator, the average score of all test speech samples in a set, i.e. the average for each manner of articulation, is calculated.

(2) Confusion matrix: The confusion matrix is calculated for each consonant.

(3) Syllable position: In the two-syllable test of Levels 1 and 3, each syllable's articulation should be scored separately.

(4) Other: Any special characteristics of the synthesizer being evaluated should be described.

Weighting the syllable articulation score according to the frequency of syllable occurrence is also an effective way to evaluate the actual performance of a synthesizer.

4.2. Materials for syllable articulation evaluation

Several sets of speech units have been prepared as test stimuli, monosyllables and CVCV and VCV two-syllable combinations (nonsense words). The number of basic Japanese monosyllables is about one hundred, or 211 if foreign syllables and historical syllables are included. Because the total number is relatively small, the monosyllable articulation test is convenient for evaluation. This test is effective in evaluating the articulation of the initial syllable of a word, but it is not suitable for medial or final syllables. For this purpose, a two-syllable combination test works better.

However, the total number of two-syllable combinations is about ten thousand, which is too many for a practical test, so we have divided the speech unit sets into three levels according to ease of testing; i.e. level 1 can be tested easily, while levels 2 and 3 require a diagnostic test. Recently, many text-to-speech synthesizers have adopted speech unit concatenation methods, and various kinds of unit selection algorithms that consider the phoneme environment have been developed. However, it is nearly impossible to evaluate all phoneme environments. Therefore, one test level should be chosen, according to the purpose or the size of the experiment.

No accentuation is specified for two-syllable combinations like VCV.

Tables in the Guideline
Table 1. List of 100 syllables
Table 2. List of 211 syllables
Table 3. List of 102 CVCV-type nonsense words
Table 4. List of 101 VCV-type nonsense words (1)
Table 5. List of 101 VCV-type nonsense words (2)
Table 6. List of 101 VCV-type nonsense words (3)

Level 1 (for simple evaluation): The sets of 100 and 211 syllables, and a subset of the CVCV two-syllable combinations.

Level 2 (for diagnostic evaluation): The sets of 100 and 211 syllables, and the VCV two-syllable combination set (Yoshikawa et al., 1985; Higuchi et al., 1989). Not all possible VCV combinations are included, as this would result in too large a test. Therefore, we limited this set to combinations of one of the vowels /a/, /i/ or /u/ plus one of the one hundred monosyllables, for a total of three hundred and three.

Level 3 (for diagnostic evaluation) The purpose of this level is to evaluate synthesizability for various co-articulations. The test set includes all 6 sets of syllables and nonsense words, and is based on the frequency distribution of phonemes in the Japanese language, e.g. the Phonetically Balanced Word List for Japanese (Torii, 1956).

5. Word Intelligibility Test

5.1. Word intelligibility metrics

(1) Mean Intelligibility: Mean intelligibility score across all words for each evaluator and the mean of all evaluators should be given.

(2) Others: The special characteristics of the synthesizer to be evaluated should be described.

5.2. Materials for word intelligibility evaluation

5.2.1. Isolated words

When a word list is created, phonemes and other attributes should be balanced to make the statistical characteristics of the list similar to those of the original population. Furthermore, the number of words should be as small as possible while still satisfying the above requirements; the testing duration per evaluator should be no more than 20 minutes.

The word sets should be selected based on Watanabe's test-word selection algorithm (Watanabe et al., 1988), which uses 42,385 words from the dictionary "Shin Meikai Kokugo Jiten [New Concise Japanese Dictionary], second edition". This method classifies words with

respect to five attributes, and the range of values for each attribute is divided into several regions:

(a) Word length: It is well known that the maximum word length a person can process in short-term memory is 8 moras (Higuchi et al., 1985), so the whole word set is classified into three categories by length: 2-3 moras, 4 moras, and 5-8 moras.
(b) Familiarity: The words are divided into five categories, ranging from 1 (very familiar) to 5 (very unfamiliar). They are categorized into three regions by familiarity: 1-2, 2-3 and 3-5.
(c) Similarity: Similarity is defined as the number of familiar words whose word distance (defined as the number of different phonemes) is '1'. There are two regions of similarity, one being "1" or more and the other being zero.
(d) Accent type: Words whose length is four moras or less are divided into two classes, one being the 0 or N type and the other being the 1 to N-1 type. Words of five moras or more are not classified as to accent, because few words have the same accent type in these words.
(e) Phoneme Categories: The first mora of each word is grouped into seven categories. The second mora is divided into nine categories by adding syllabic nasals and doubled consonants, for a total of 63 regions.

The result of this classification is a five-dimensional matrix with 1,701 cells in total:

Length	2-3 moras	4 moras	5-8 moras
Familiarity	3 regions	3 regions	3 regions
Similarity	x 2	x 2	x1
Accent	x 2	x 2	x1
Phonemes	x 63	x 63	x 63
Total regions	756	756	189

When the whole Japanese word population is classified in this manner, it is found that only 1,523 of these cells are actually occupied. Words are selected from the population so as to preserve as far as possible the distribution of the whole across these 1,523 cells.

According to Watanabe (Watanabe et al., 1988), about 1,200 words are necessary to satisfy this condition below, and the word set size per evaluator per session should be limited to 200 words or less. The test can thus be divided into six test sessions.

For the primary selection, one word or more from each cell, to a total of 900-1200 words.
For secondary selection, to improve similarity to the original population, 300 words are suitable, for a grand total of 1200-1500 words.

5.2.2. Semantically unpredictable sentences

In a word intelligibility test using sentences, sentences which make sense seem close to the actual application situation, but word perception can be easily affected by the meaning of the whole sentence if it makes sense. The listener can be influenced to hear words incorrectly in order to fit the sentence in context. This is not a sentence intelligibility test, so semantically anomalous sentences should be used, where the syntactic structure is correct but the overall meaning is not natural. Then the results are based on actual perception of the target word in a reasonable phonological and syntactic context, but not a semantic one.

The following considerations are recommended when creating the sentences for evaluators.

1) Taking human short-term memory into consideration, sentence length should be four phrases or less.
2) Pitch patterns should be considered.
3) Pause positions should be stable.
4) Phoneme occurrences should be balanced.
5) Words should be selected from a large population.

A computer can be used for effective composition of semantically unpredictable sentences, but these items need to be checked manually.

6) Senseless or unfamiliar words should not be included.
7) Sentences should be syntactically correct.

A semantically unpredictable sentence for evaluation should consist of four phrases with the dependency structure shown below.

1st phrase → 2nd phrase   3rd phrase → 4th phrase
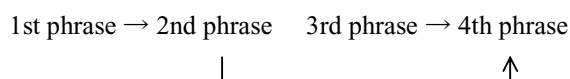
Phrases should be composed according to the following patterns:
1st phrase: a noun + particle *no*, or an adjective ending with *i* or *na*
2nd phrase: a noun + particle *ga*
3rd phrase: an adverb ending with *to*, or an adjective ending with *ku*, or an adjective ending with *ni*, or *to*
4th phrase: a verb or an adjective (final-form)

Semantically unpredictable sentences should include words chosen according to the methods described above. For an easy test, the example sentences listed in the "Commentary" (Appendix C) can be used.

5.3. Word intelligibility test procedures

Test with isolated words and tests with semantically unpredictable sentences both have listening error items for "substitution", "deletion", and "insertion". In addition to the general procedures given in Section 3.2, the following points should be noted in the test with semantically unpredictable sentences:

a) The evaluators should be informed in advance that they will be dealing with nonsense sentences, to ensure that they will not try to interpret them as natural sentences.
b) The evaluators should be informed that the sentences may occasionally make sense, depending on the interpretation one applies to them.

6. Sentence Intelligibility Test

Guidelines for texts, questions and response methods for the intelligibility evaluation of synthesized speech are described below.

The method of testing sentence intelligibility is to have evaluators listen to synthesized sentences in synthesized

speech and then to write the answer to a question on its content.

6.1. Sentence intelligibility metrics

We define an index for sentence intelligibility:
Sentence Understandability = (Number of correct answers) / (Number of all items to be answered).

6.2. Materials for sentence intelligibility test

With respect to content, an objective text such as a news story is desirable and it is easy to make questions from it by "5W1H"method. Guide messages from major areas of the speech synthesis application are also desirable from a practical application viewpoint. Essay-type texts must be simple enough for everyone to understand easily. Content requiring special knowledge, or those like a language examination are not good. Short texts, typically no more than one minute long are desirable. Any style of text, either polite or normal style, is acceptable.

1) Questions must be easy enough so that anyone could provide the correct answers if he/she reads the text instead of hearing it.
2) Several answering styles, such as yes-no and multiple-choice answers, are possible. About four choices are desirable for the multiple-choice style.

6.3. Sentence intelligibility test procedures

(1) Number of Test Trials: The number of trials is not specified, but the more trials the better the reliability.
(2) Testing Equipment: The type of testing equipment; i.e., loudspeakers, headphones or telephones, and the acoustical environment is not specified because the choice depends on the application being tested.

To avoid testing the human memory, it is better for the evaluator to listen to the synthesized speech after reading the questions.

The evaluation of naturalness of speech must be done comprehensively because the concept of naturalness is considered as a combination of several quality factors of speech. Moreover, the level of the quality of naturalness, i.e. the degree of deterioration, needs to be evaluated diagnostically. For this purpose, the synthesized speech should be compared with human speech on various items (Watanabe, 1991). This method differs from overall evaluation, i.e., "a method for evaluating the subjective impressions of synthesized speech" or "a method for evaluating its acceptability for the purpose of utilizing speech synthesis." Currently, no chapter on naturalness evaluation is included in the guidelines, as it is difficult to distinguish this from overall evaluation.

7. Overall Quality Evaluation

The speech synthesizer is a system component of the human-machine interface. Therefore, in addition to the diagnostic evaluation (evaluation of intelligibility and naturalness) of the technical achievement of the speech synthesizer in itself, an overall quality evaluation is also necessary, one which considers user purposes, operating environments, utilization method, and human factors (Kasuya et al., 1991). Actually, however, no method for evaluating overall quality has yet been established. We can only recommend evaluating the subjective impression or the suitability to the user's purposes under the closest situation possible to the actual operating environment (Watanabe, 1989; Higuchi et al., 1989). Since objective evaluation methods are now being studied, the rating scale method is used, which has already been established as a subjective evaluation method.

7.1. Overall quality evaluation metrics

Select one to three pairs of words to describe each of the following items (see Appendix D in the "Commentary"):

(1) Intelligibility: Intelligibility and misreading the synthesized speech message.
(2) Speech sound quality: The sound quality and the voice quality.
(3) Temporal factors: The sense of rhythm inherent in the Japanese language, the speaking rate, and the continuity of speech sound.
(4) Intonation: Naturalness and fluency of intonation and accent.
(5) Overall goodness: Similarity to the human voice, and the preference and the quality of the synthesized speech.
(6) Suitability: Suitability to the user's purposes in terms of voice quality, speaking rate, and intonation. Evaluate the impression of fatigue or comfortableness. Evaluate whether there are any obstacles to the user's purposes. Evaluation terminology can be added according to the user's purposes.
(7) Others: Appropriate items added as necessary to meet the user's special purposes.

7.2. Materials for overall quality evaluation

Use the text actually used or one written in the same style, or use the evaluation text recommended by JEIDA in the "Commentary" (Appendix E). The length of the spoken text should be approximately 30 seconds.

7.3. Overall quality evaluation procedures

Evaluate the speech after listening carefully to the test sample several times. If it is hard to evaluate all the items at one time because of the large number of items, they may be broken into groups of two or three. If there are many objects, prepare a set of test materials arranged at random and listen to it repeatedly (in this order).

7.3.1. Evaluators

Select evaluators who will actually (or possibly) use the system. The number of evaluators should be 20 or more (the more, the better), since a small number of evaluators can produce biased data. If less than 20 evaluators are available, each one should evaluate the sample two or three times to improve statistical reliability.

7.3.2. Conditions to be reported

Report the environment or conditions under which the system is actually used, or similar environments or

conditions. Indicate the user's purposes, the listening equipment (loudspeakers or headphones), the sound pressure level, the distance between the listener and the loudspeaker(s), the ambient noise level, the reverberation time of the testing room and any other relevant factors. Show the contents of the text used for the evaluation. Indicate the number of evaluators employed as well as their age group, sex, mother tongue, birth place (dialect), hearing disabilities, and any other relevant factors.

Appendices in the "Commentary"
Appendix A. Example list of Japanese Phonetically Balanced Words
Appendix B. Example list of words for Intelligibility Test
Appendix C. Example list of semantically unpredictable sentences
Appendix D. Words describing speech quality
Appendix E. List of sentences for Overall Quality Test

8. Conclusion

As stated in the Introduction, the Guidelines for Speech Synthesizer Evaluation have been created to evaluate primarily speech synthesizers. However, not all of the items in the Guidelines are necessary for each individual test. Evaluations can be made by selecting only some of the items, depending on the purpose. The present Guidelines are not complete; some issues remain unsolved. We hope that the study of the evaluation methods themselves will be continued and that these Guidelines will be improved according to the outcome of such study. It will be gratifying to us if these Guidelines, as well as the development of further applications and techniques, prove to be helpful to both developers and users.

9. References

JEIDA-C, 2000, Commentary on the guideline for speech synthesis system performance evaluation methods (in Japanese).

JEIDA-G, 2000, Guidelines for speech synthesis system performance evaluation methods (in Japanese).

Hieda, I., Kuchinomachi, Y., and Nagamura, N, 1985, Speech quality evaluation using short-term memory paradigm, *Preprints of Tech. Group on Acoustics, IEICE*, Paper S85-33 (in Japanese).

Higuchi, N., Yamamoto, S., and Shimizu, T., 1989, Evaluation of intelligibility and naturalness of the synthetic speech generated with a Japanese speech synthesizer by rule, *Trans. Inst. Electron. Inform. & Commun. Engnrs. Jpn. (IEICE)*, **J72-D-II**, 8, 1133-1140 (in Japanese).

Kasuya, H., Kasuya, S., and Morita, K., 1991, Roles of the speaking rate of synthetic speech produced by rule as an aid for proofreading, *J. Acous. Soc. Jpn.*, **47,** 2, 96-98 (in Japanese).

Kasuya, H., 1992, Assessment of speech synthesizer technology, *J. Acous. Soc. Jpn.*, **48**, 1, 46-51 (in Japanese).

Nishi, H. and Itahashi, S., 1998, JEIDA guidelines for speech synthesizer evaluation, *Proc. First International Workshop on East-Asian Language Resources and Evaluation (EALREW)*, 57-64.

Torii, N., 1956, Phonetically balanced word list for Japanese, *Technical Report of NTT Electrocommunications Laboratories*, No. 484 (in Japanese).

Watanabe, T., Nagabuchi, H., and Kitawaki, N., 1988, A word selection method for intelligibility assessment of synthesized speech by rule," *IEICE Trans.*, **J71-A**, 3, 616-623 (in Japanese).

Watanabe, T., 1989, Intelligibility assessment method for synthesized speech using word intelligibility score, *IEICE Trans.*, **J72-A**, 10, 1503-1509 (in Japanese).

Watanabe, T., 1991, Global assessment method for synthesized speech", *IEICE Trans.*, **J74-A**, 4, 599-609 (in Japanese).

Yoshikawa, Yazu, Morimoto, and Yamada, 1985, Evaluation of speech synthesizer by rule, *Preprints of Fall Meeting of Acous. Soc. Jpn.*, Paper 2-3-15 (in Japanese).