T2R2 東京科学大学 リサーチリポジトリ
Science Tokyo Research Repository

論文 / 著書情報
Article / Book Information

Title	Evaluation Methods for Speech Recognition Systems and Technology
Author	SADAOKI FURUI
Journal/Book name	Workshop on International Cooperation and Standardization of Speech Databases and of Speech I/O Assessment Methods, Vol. , No. , pp.
Issue date	1991,

Powered by T2R2 (Science Tokyo Research Repository)

Abstract for the Workshop on International Cooperation and Standardization
of Speech Databases and of Speech I/O Assessment Methods

Evaluation Methods for Speech Recognition Systems and Technology

Sadaoki Furui
NTT Human Interface Laboratories
Musashino-shi, Tokyo, 180 Japan

This paper outlines major issues in evaluating speech recognition systems and technology. Standards for comparison and evaluation are very important, since technological progress can only be measured and encouraged by quantitative and meaningful measures under the same conditions. The issues can be divided into evaluation of the difficulty of recognition tasks, evaluation of performances of recognition algorithms/systems, and the speech database for evaluation. Evaluating spontaneous-speech recognition is much more difficult than evaluating read-speech recognition.

Compared with the evaluation methods for speech synthesis, those for speech recognition are much more complicated, and they are far behind in standardization, having a lot of difficult problems to be discussed.

1. Evaluation of Tasks

How to evaluate the difficulty of each task is one of the important issues, since in order to compare different systems, the difficulty of the task must be taken into account and recognition performances must be normalized by the difficulty. Several measures have been proposed for evaluating the tasks: static or dynamic branching factor, phoneme/syllable/word perplexity, and task entropy. They all measure the constraint imposed by the grammar, or the level of uncertainty given the grammar. The perplexity can be regarded as an information theory-based average branching factor. The test-set perplexity has also been proposed for the cases where the perplexity measured from the language model does not reflect the uncertainty encountered during recognition. The test-set perplexity is simply the geometric mean of probabilities at each decision point for the test set sentences.

Although the word perplexity is now widely used, it has some problems: (1) since the perplexity does not take the sentence length into consideration, different tasks such as isolated digit and connected digit recognition have the same perplexity of 10, (2) for agglutinative languages such as Japanese, the definition of words is unclear, and the perplexity changes depending on the definition of words, and (3) the difficulty of acoustical level recognition, that is, the inherent confusability of the vocabulary, is not reflected in the perplexity.

2. Evaluation of Algorithms

2.1 Evaluation measures

Several measures have been investigated: percent correct (only the top choice or several choices), error rate, word accuracy, rejection rate, insertion rate, deletion rate, and substitution rate. Although computing the error rate for isolated-word recognition is straightforward, it is not easy to evaluate three types of errors (substitution, deletion, and insertion) in continuous speech recognition. Percent correct and word accuracy differ by the number of insertions. Black-box evaluation versus glass-box evaluation is another issue,

which must be chosen depending on the necessity of the evaluation of system components.

2.2 Units of evaluation

Various units have been used for calculating recognition performances: phonemes, syllables, words, and sentences. Relationships between recognition (error) rates for these units has also been investigated.

2.3 Significance tests

An important viewpoint which is often ignored is the significant test for the results of recognition experiments. Improvements in performance by a "new" technique are often reported without careful consideration of the statistical significance of the difference between the new technique and the old ones.

2.4 Evaluation of spoken language systems

Evaluating spoken language systems (spontaneous speech recognition systems) is difficult because (1) we have to consider semantics or meaning of the input speech to decide whether the output of the system is correct or not, and (2) semantics are far more domain- and context-dependent than phonetics or phonology. Even if the same task and the same database are used, it is still necessary to agree on the basic meaning of terms.

2.5 Evaluation from the viewpoint of human interface

In order to produce commercial systems or equipment, it is very important to evaluate the performance from the viewpoints of usefulness and comfortableness for the users. The performance must also be evaluated by comparing with a non-spoken method of input. The measures of performance include not only the correctness of the output but also the response time, efficiency (e.g. the time to complete a task successfully such as document preparation), mental and physical weariness, ease of correcting errors, and adaptability as well as learning capability of the system.

3. Speech Database for Evaluation

Selection of the speech database is crucial, because it has a large effect on the technology pursued. The task must not be too difficult nor too easy for the progress of technology. The following points must be considered in determining the database: the task including the selection of a domain, speakers, sizes of training and testing sets, recording conditions and quality (bandwidth, S/N, etc.), spontaneous versus read speech, and carefulness and preciseness of the utterances. Variations in speakers and the amount of utterances for each speaker must be determined depending on whether the system is speaker-dependent, speaker-independent, or speaker-adaptive. If the utterances are collected in a conversational mode with a machine (human-machine dialogues), it is also important to decide what kind of responses or answers should be produced by the machine.