Word recognition using confusion matrices for computing similarities among phonemes

S. Makino, J. Miwa, and K. Kido

Citation: The Journal of the Acoustical Society of America **64**, S181 (1978); doi: 10.1121/1.2004063

View online: https://doi.org/10.1121/1.2004063

View Table of Contents: http://asa.scitation.org/toc/jas/64/S1

Published by the Acoustical Society of America

[Research supported by NSF Grant MSC-76-81034 and the USAECOM Contract MDA 904-77-C-0157.]

[a] Present address: Signal Technology, Inc., Santa Barbara, CA 93101.

2:42

NNN15. A real-time recognition system for spoken words. Hiroya Fujisaki, Keikichi Hirose, Mikio Mizutani (Faculty of Engineering, University of Tokyo, Bunkyo-ku, Tokyo, 113 Japan), and Yasuo Sato (Fujitsu Limited, Kawasaki, Japan)

A method is presented for reducing the memory capacity and the amount of computation necessary in automatic recognition of spoken words based on the registration of word templates by individual speakers. The reduction is achieved by representing the spectral information of speech in terms of a small number of feature parameters and by sampling them at a variable rate determined by their rate of variation. A special-purpose hardware unit was designed and constructed for the extraction of the speech parameters, using microprocessors and microprogramming techniques for high-speed processing. On the other hand, a general-purpose minicomputer was utilized for word registration and recognition in order to determine detailed specifications for microprocessor implementation. The performance of the total system for real-time word recognition was tested by using three kinds of vacabulary items, including digits, arithmetic symbols, and railway station names, giving an average rate of correct recognition of 98.1%. [Work supported by Ministry of Education Grant-in-Aid for Scientific Research No. 285070.]

2:45

NNN16. A phonetically based isolated word recognition system. M. E. Blomberg and K. O. E. Elenius (Department of Speech Communication, Royal Institute of Technology, 100 44 Stockholm 70, Sweden)

The principal object in using a phonetic approach is the reduction of the influence on recognition rate caused by the intra- and interspeaker speech variations. The system is implemented on a 16-K minicomputer and uses a filter bank delivering spectral sections, 0–5 kHz, every 10 ms. Estimates of the first three formants are calculated and energies in different spectral bands are used to segment the speech signal into broad classes. The following measures are calculated depending on the segmental class and the speech parameter: mean values, steady-state values, durations, transition rates, and some distances between formants. In a learning phase the statistics of the measures of the used vocabulary are automatically calculated by a program given the quasiphonetic spelling of the input words. The statistics are based on phoneme pairs, i.e., diphones. In the recognition phase the program uses the statistics and the quasiphonetic spelling to recognize the input words. Six male speakers were used for calculating the statistics of a 41-word vocabulary. Their mean recognition rate was 98%, using a new recording. The rate decreased to 96.3% using four male talkers, unknown to the system. [Work supported by STU, Sweden.]

2:48

NNN17. A spoken word recognition system for unspecified male speakers. K. Kido, J. Miwa and S. Makino (Research Institute of Electronic Communication, Tohoku University, Sendai, 980 Japan)

A spoken word recognition system for unspecified male speakers is outlined. The system operates in the following four stages. (I) Seven acoustic parameters are extracted every 10 ms from the outputs of the filter bank. The parameters are the frequencies of three spectral local peaks, the speech power and three parameters expressing the gross pattern of the spectrum. (II) Segmentation and the phoneme recognition are carried out. (III) Errors in the segmentation and phoneme recognition are corrected by means of phoneme connecting rules. (IV) The recognized sequence is determined to be the item of the dictionary having a maximum similarity to the recognized phonemic sequence. The similarity between the item of the dictionary and the recognized phonemic sequence is computed using confusion matrices made from the recognition of 17 040 phonemes. Every item of the dictionary is written in phonemic symbols derived from the word in Japanese "kana" letters by simple rules, so that the contents of the dictionary can be easily changed. The scores for word recognition were found to be 84.0% for 166 words uttered by 25 male speakers and 95.0% for 51 words selected from these 166.

2:51

NNN18. Word recognition using confusion matrices for computing similarities among phonemes. S. Makino, J. Miwa, and K. Kido (Research Institute of Electronic Communication, Tohoku University, Sendai, 980 Japan)

A method of word recognition using confusion matrices has been developed. In this method, the likelihood of additions, omissions and erroneous phoneme recognitions given by the confusion matrices are used for the computation of the similarity between the recognized phonemic sequence to every item of the word dictionary. Three confusion matrices are made for the initial and final phoneme and the medial phoneme(s) of words. The dynamic programming was adopted to increase the efficiency of the computation. The item of the dictionary having the maximum similarity to the input sequence is chosen as the output of the recognition. The score of the word recognition was 84.0% for 166 words uttered by 25 male speakers, including 15 speakers whose utterances were used for making the confusion matrices, where the averaged score of the phoneme recognition was 64.7%. In experiments using a small number of words, the scores were 95.0% and 97.2% for 51 and 20 words, respectively.

2:54

NNN19. Speaker-independent recognition of isolated words using clustering techniques. L. R. Rabiner, S. E. Levinson, and A. E. Rosenberg (Acoustics Research Department, Bell Laboratories, Murray Hill, NJ 07974)

Levinson *et al.* have recently described a powerful set of techniques for clustering multiple replications of words spoken by different talkers into a set of composite reference templates. These techniques have been incorporated into a speaker-independent, isolated word recognition system. The vocabulary which we have tested consists of 39 words including the alphabet, the digits 0–9, and the three cueing words STOP, ERROR, and REPEAT. This vocabulary is of great utility for a wide range of applications of automatic word recognition. The features used for recognition are an eight-pole LPC set measured every 15 ms over a standard telephone line. The distance measure is the log likelihood ratio as originally proposed by Itakura. Several variations of a dynamic time warping algorithm have been incorporated into and tested in this system. Using the clustering analysis we have obtained from 2 to 12 clusters per word. The decision rule is a generalized K-nearest neighbor (KNN) rule. Recognition accuracies comparable to those of speaker-dependent isolated word recognition systems have been obtained.

2:57

NNN20. Automatic recognition of spoken spelled names using speaker-independent templates. A. E. Rosenberg, L. R. Rabiner, and J. G. Wilpon (Bell Laboratories, Murray Hill, NJ 07974)

An automatic word recognizer is used to accept strings of spelled letters spoken in isolation. The output of the recognizer is a set of