A word model based endpoint detector for isolated utterunce recognition

James L. Hieronymus

Citation: The Journal of the Acoustical Society of America **73**, S101 (1983); doi: 10.1121/1.2020205
View online: https://doi.org/10.1121/1.2020205
View Table of Contents: https://asa.scitation.org/toc/jas/73/S1
Published by the Acoustical Society of America

2:45

TT6. Solutions of some problems in duct acoustics, using nonseparable modes. P. G. Vaidya (Mechanical Engineering Department, Washington State University, Pullman, WA 98104)

Almost all closed form solutions of the problems in duct acoustics (and in other branches of acoustics) are obtained by assuming that the eigenfunctions depend on each of the coordinates independently of the other. Such "separable" solutions are impossible if, for example, the boundary conditions are nonuniform. Under these conditions, instead of resorting to *ad-hoc* numerical procedures, closed form solutions could still be obtained by using nonseparable modes. Examples of circumferentially and/or axially nonuniform boundary conditions are worked out to illustrate this procedure.

FRIDAY AFTERNOON, 13 MAY 1983 BALLROOM A, 1:00 TO 5:05 P.M.

Session UU. Speech Communication VIII: Recognition and Intelligibility

Gary M. Kuhn, Chairman
Institute for Defense Analyses, Thanet Road, Princeton, New Jersey 08540

Chairman's Introduction—1:00

Contributed Papers

1:05

UU1. Speaker independent recognition of isolated digits in a telephone environment. J. G. Wilpon and L. R. Rabiner (Bell Laboratories, Acoustics Research Department, Murray Hill, NJ 07974)

A field study was initiated to learn about the effects of various telephone transmission and switching conditions on the algorithms currently used in the Bell Laboratories, LPC-based, isolated word recognizer. Digit recordings were obtained from customers over a variety of transmission facilities. During a 23-day recording period a total of 11 035 isolated digits were recorded. For each recording statistics were recorded about the line condition, the background environment, and the customer's ability to speak his telephone number as a sequence of isolated digits. Also recorded was information about the ability of the automatic work endpoint detector to find each spoken digit and to accurately determine the correct endpoints. The results of several recognition tests are presented—one in which a previously defined set of laboratory created digit reference templates was used, and several others where new sets of reference templates were created from a subset of the recorded digits. It is shown that the performance of the recognizer is poor (average digit accuracy of 77.4%) using the laboratory template set, but improves substantially (average digit accuracy 93.1%) for a template set created from the field recordings. An explanation of the reasons for this improvement in digit recognition accuracy is presented along with their implications to future work in isolated word recognition.

1:20

UU2. The use of clustering result in speaker-independent, isolated-word recognition. Nien-chih Wei (Bell Laboratories, IH6D317, Naperville, IL 60566)

Currently, a test utterance is recognized as word "A" when the distance score of the test utterance to a reference template for word "A" is the lowest among the measured scores between test and all reference utterances. The reference template is generated by: (1) clustering utterances from many speakers and (2) choosing the cluster centers as reference templates. Thus, the distance between the test utterance and the reference template is the distance of the test utterance to the cluster center. This type of distance measure ignores the fact that clusters have different sizes. Using a Gaussian model for the distribution of utterances inside a cluster, this paper proposes a new distance measure based on the size of the cluster. The correlation between the Itakura distance measure and the new distance measure for many utterance ($\sim$1500) is fairly low ($\sim$0.2). This suggests that the new distance measure represents new information. Incorporating this new distance measure can reduce the recognition error rate in half.

1:35

UU3. A new task-oriented conversational mode speech recognition system. S. E. Levinson, L. R. Rabiner, S. G. Terrace, and K. L. Shipley (Acoustics Research Department, Bell Laboratories, Murray Hill, NJ 07974)

A conversational mode-connected speech recognition system which simulates the function of an airline ticket agent is described. The system which allows a complete spoken dialog over dialed-up telephone lines comprises a syntax-directed dynamic programming temporal alignment algorithm [C. S. Meyers and S. E. Levinson, IEEE Trans. Acoust. Speech Signal Process. **ASSP-30**, 561–565 (1982)] for acoustic pattern recognition and a semantic processor [S. E. Levinson and K. L. Shipley, Bell Syst. Tech. J. **59**, 119–137 (1980)] that controls an audio response system making two-way speech communication possible. The system is highly robust and operates on-line in ten times real time in speaker trained mode (or 60 times in speaker-independent mode) on a 32-bit minicomputer/array processor combination. The heterarchical nature of the system allows it some intelligent processing such as the ability to recognize incorrect, incomplete, improbable, and/or conflicting information in the context of the transaction and to pose questions which identify the difficulty thereby allowing communication to proceed in a natural way.

1:50

UU4. A word model based endpoint detector for isolated utterance recognition. James L. Hieronymus (National Bureau of Standards, Institute for Computer Sciences and Technology, Bldg. 225, Rm. A216, Washington, DC 20234)

Endpoint detection is a critical issue for several types of isolated utterance recognizers, because improper endpoints often result in recognition errors. Endpoint errors often stem from nonspeech artifacts, namely lip smacks, tongue and teeth clicks, and breath noise. Endpoint detectors based only on energy thresholds cannot correctly reject these artifacts, but adding a word model allows most of these artifacts to be properly rejected. The rules which implement the word model are (1) the word cannot begin or end with two released plosives, (2) word initial stop gaps are less than 120 ms and word final ones less than 200 ms, (3) a word must contain a vocalic nucleus and be at least 100 ms in length, (4) word final sounds containing only mid-frequency energy are breath noise. The detection algorithm has been implemented on a Heuristics Speech Recognizer and tested using the Texas Instruments isolated word data base. The word model based system substantially reduced the error rate relative to an energy threshold based system.