New techniques for automatic speaker verification using telephone speech

S. Furui

Citation: The Journal of the Acoustical Society of America **66**, S35 (1979); doi: 10.1121/1.2017731
View online: https://doi.org/10.1121/1.2017731
View Table of Contents: https://asa.scitation.org/toc/jas/66/S1
Published by the Acoustical Society of America

JASA
THE JOURNAL OF THE ACOUSTICAL SOCIETY OF AMERICA
Special Issue:
Additive Manufacturing and Acoustics
Read Now!

algorithms all assume the test input is an isolated word whose endpoints are known (at least approximately). The major difference in the methods are the global path constraints (i.e., the region of possible paths), the local continuity constraints on the path, and the distance weighting and normalization used to give the overall minimum distance. The purpose of this investigation is to study the effects of such variations on the performance of different algorithms for a realistic speech data base. The performance index is based on speed of operation, memory requirements, and recognition accuracy of the algorithm. Preliminary results indicate, in most cases, only small differences in performance among the various methods.

9:05

R3. On the use of clustering for speaker-dependent isolated word recognition. L. R. Rabiner and J. G. Wilpon (Acoustics Research Department, Bell Laboratories, Murray Hill, NJ 07974)

Speaker-trained, isolated word recognizers have achieved notable success in a wide variety of applications. The training for such systems generally involves a single (or sometimes two) replication(s) of each word of the vocabulary by the designated talker. Word reference templates are then formed directly from these replications. In recent work on speaker-independent word recognition, it has been shown that statistical clustering procedures provided an effective way for determining the structure in multiple replications of a word by different talkers. Such techniques were then used to provide a set of reference templates based on the clustering results. In this talk we discuss the application of clustering techniques to speaker-trained word recognizers. It is shown that significant improvements in recognition accuracy are obtained when using templates obtained from a clustering analysis of multiple replications of a word by the designated talker. It is also shown that recognition accuracy did not change with time (over a 6-month period) for any of the subjects tested, thereby indicating that the reference templates were reasonably stable.

9:20

R4. New techniques for automatic speaker verification using telephone speech. S. Furui[a] (Acoustics Research Department, Bell Laboratories, Murray Hill, NJ 07974)

This paper describes new techniques for automatic speaker verification using telephone speech. The operation of the system is based on a set of functions of time obtained from acoustic analysis of a fixed, sentence-long utterance. These time functions are expanded by orthogonal polynomial representations and compared with stored reference functions. After dynamic time warping, a decision is made to accept or reject an identity claim. Three sets of experimental utterances were used for the evaluation of the system. The first and second sets each comprises 50 utterances by 10 customers each and a single utterance by 40 imposters recorded over a conventional telephone connection. The third set comprises 26 utterances by 21 customers each and a single utterance by 55 imposters recorded over a high quality microphone. The first and third sets were uttered by male speakers, whereas the second set was uttered by female speakers. Reference functions and decision thresholds were updated for each customer. The evaluation indicated mean error rates of 0.19%, 0.36%, and 0.77% for each utterance set, respectively.

[a] Permanent address: Electrical Communication Laboratories, Nippon Telegraph and Telephone Public Corporation, Musashino, Tokyo, Japan.

9:35

R5. A conversational-mode airline information and reservation system using speech input and output. S. E. Levinson and K. L. Shipley (Acoustics Research Department, Bell Laboratories, Murray Hill, NJ 07974)

We describe a conversational-mode speech understanding system which enables its user to make airline reservations and obtain timetable information through a spoken dialog. The system is structured as a three level hierarchy consisting of an acoustic word recognizer, a syntax analyzer, and a semantic processor. The semantic level controls an audio response system making two-way speech communication possible. The system is highly robust and operates on line in a few times real time on a laboratory minicomputer. The speech communication channel is a standard telephone set connected to the computer by an ordinary dialed-up line.

9:50

R6. Automatic acoustic–phonetic segmentation using a hidden Markov model. T. J. Edwards and K. P. Li (TRW Defense and Space Systems Group, One Space Park, Redondo Beach, CA 90278)

In a previous meeting [J. Acoust. Soc. Am. Suppl. 1 **64,** S179(A) (1978)], we presented an evaluation of TRW's automatic segmentation program. More recently, we have sought to improve segmentation performance using a hidden Markov model (HMM) to predict the underlying phonetic segments. Utilizing a hand-transcribed data base, the HMM was initially directly calculated on the segmentation output and transcription and the new segmentation recognition result evaluated. Using an iterative procedure [J. Baker, in *Speech Recognition*, edited by R. Reddy (Academic, New York, 1975), pp. 521–542] the HMM was then trained using only the segmentation output and compared with the previously obtained segmentation result. Using different HMM initial conditions, the model was again trained in a test for convergence. A comparison of each of these HMM for segmentation will be provided and evaluation of the models will be discussed.

10:05

R7. Application of post-correction techniques to connected digit recognition. L. R. Rabiner and C. E. Schmidt (Acoustics Research Department, Bell Laboratories, Murray Hill, NJ 07974)

A scheme is proposed for connected digit recognition in which a set of isolated word templates is used as reference patterns and an unconstrained dynamic time warping algorithm is used to literally "spot" the digits in the string. The recognizer keeps track of a set of candidate digit strings for each test string. The string with the smallest accumulated distance is used as a preliminary string estimate. To help improve the recognition accuracy, we have considered two "post-correction" techniques applied to the entire set of hypothesized digit strings. One technique creates a reference string by concatenating reference contours of the digits of the string and compares this to the test string using a constrained dynamic warp algorithm. The other technique performs a similar comparison using voiced–unvoiced-silence contours instead of the measured features. Small but consistent improvements in recognition accuracy have been obtained using these techniques for both speaker-trained and speaker-independent systems over dialed-up telephone lines.

10:20

R8. A statistical approach to metrics for word and syllable recognition. Melvyn J. Hunt (Bell–Northern Research, 3 Place du Commerce, Verdun, Quebec, Canada, H3E 1H6)

Time-warping pattern-comparison algorithms are widely used in speech recognition. Two words or syllables being compared are described by a series of time frames each containing values of a set of acoustic parameters. After time alignment, the squared distance between the patterns is summed over the parameters within a frame and then across frames. The sum obtained is assumed to be proportional to the log probability of the two patterns having the