Temporal energy contour in isolated word recognition

E. L. Bocchieri

Citation: The Journal of the Acoustical Society of America **76**, S46 (1984); doi: 10.1121/1.2021875
View online: https://doi.org/10.1121/1.2021875
View Table of Contents: https://asa.scitation.org/toc/jas/76/S1
Published by the Acoustical Society of America

Session U. Speech Communication IV: Automatic Speech Recognition

Gregory H. Wakefield, Chairman
Department of Psychology, University of Minnesota, Minneapolis, Minnesota 55455

Chairman's Introduction—12:45

Contributed Papers

12:50

U1. Continuous density hidden Markov models for speaker-independent recognition of isolated digits. B. H. Juang, S. E. Levinson, L. R. Rabiner, and M. M. Sondhi (Acoustics Research Department, AT&T Bell Laboratories, Murray Hill, NJ 07974)

In a recent report [L. R. Rabiner, S. E. Levinson, and M. M. Sondhi, J. Acoust. Soc. Am. Suppl. 1 **74**, S17 (1983)], it was shown that the quantization error inherent in discrete-symbol hidden Markov models introduces a small but consistent degradation into speaker-independent isolated word recognition. Our present study demonstrates that this degradation can be eliminated by a number of techniques based on the theory of hidden Markov models with continuous density functions [L. R. Liporace, IEEE Trans. Inform. Theory **IT-28**, 729–34 (1982)]. By using these methods we have achieved 98% speaker independent recognition accuracy on isolated English digits. This rate is statistically indistinguishable from the best results obtained with the LPC/DTW system [L. R. Rabiner, S. E. Levinson, A. E. Rosenberg, and J. G. Wilpon, IEEE Trans. Acoust. Speech Signal Process. **ASSP-27**, 134–141 (1979)].

1:05

U2. Continuous phonetic speech recognition. John Makhoul, Richard Schwartz, Yen-Lu Chow, Owen Kimball, Salim Roucos, and Michael Krasner (Speech Signal Processing Department, Bolt Beranek and Newman Inc., 10 Moulton Street, Cambridge, MA 02238)

We report on research to develop an automatic phonetic recognition system for continuous speech. The system is based on a hidden Markov model (HMM) representation of phonemes in context. That is, the model parameters depend on the left and right phonetic contexts for each phoneme. The HMM structure is the same for all phonemes, but the model parameter values are different for different phonemes and different contexts. Automatic training procedures are used to adjust the model parameters, using a given set of training speech data. A major focus of our work is to maximize the robustness of the phonetic models given a limited set of training data. An initial system is now operational. Results of phonetic recognition accuracy in continuous speech will be presented. [Work supported by ARPA and monitored by ONR.]

1:20

U3. Network-based isolated digit recognition using vector quantization. Marcia A. Bush, Gary E. Kopec, and Marie E. Hamilton (Fairchild Laboratory for Artificial Intelligence Research, MS 30-888, 4001 Miranda Avenue, Palo Alto, CA 94304)

This talk will describe a network-based system for speaker-independent, isolated-digit (*one–nine, oh,* and *zero*) recognition and will discuss the results of an extensive series of system tuning and evaluation experiments. The digits are modeled by pronunciation networks whose arcs represent classes of acoustic-phonetic segments. Each arc is associated with a *matcher* for rating an input speech interval as an example of the corresponding segment class. The matchers are based on vector quantization of LPC spectra. Recognition involves finding minimum quantization distortion paths through the networks by dynamic programming. The system has been tested using nearly 6000 tokens of speech by 250 talkers, including a subset of a large database developed by Texas Instruments [G. Leonard, Proc. 1984 IEEE ICASSP]. The best recognizer configurations achieved accuracies of 97–99%. Performance over 21 geographically defined talker groups included in the TI database will be discussed.

1:35

U4. Speaker-independent recognition of vocalic segments. Alexander I. Rudnicky (Department of Computer Science, Carnegie-Mellon University, Pittsburgh, PA 15217)

Speaker variations produce substantial differences in vocalic spectra: Vowel templates generated from one speaker's voice will not accurately match another voice. It is possible, however, to impose transformations on the spectrum that factor out speaker differences. The current work presents two sets of experiments that examine such transformations. The first set of experiments used steady-state vowels (/i e a o u/); for ten male and ten female talkers, the results indicate that a log-ratio transformation that incorporates pitch and formants into a three-dimensional *L space* $[\log(F_1/F_0),\log(F_2/F_1),\log(F_3/F_2)]$ allows 93% classification accuracy. By comparison, spectral matching gives 44% accuracy. Extended vocalic segments (e.g., as in "away," /əwe^Y/) have dynamically varying formant patterns. In *L* space, these patterns appear as tracks. The second set of experiments investigates a recognition technique that uses the encoded track shape as part of the representation. Speaker-independent performance on isolated words was better than that obtained through DTW matching using a spectral (mel scale) representation. All experiments were run using automatic pitch and formant trackers developed at C-MU.

1:50

U5. Temporal energy contour in isolated word recognition. E.L. Bocchieri (C.E.C., M/S 439, Texas Instruments, Dallas, TX 75265)

An analysis of recognition errors in a speaker-independent digit recognition experiment has demonstrated that the temporal energy contour of the test tokens is an important feature for recognition and perception. This finding supports the conclusions of other recent studies [Rabiner *et al.*, Proc. ICASSP **17.1.1** (1984)], [Rabiner *et al.*, J. Acoust. Soc. Am. Suppl. 1 **75**, S93 (1984)]. A conventional LPC analysis with a likelihood ratio distance measure and dynamic time warping was used. This recognizer ignores the temporal energy contour. To help us understand the nature of the recognition errors, we played the test tokens through an LPC synthesizer after time alignment with the correct and with the successful (but incorrect) reference templates. Examples on tape demonstrate that the percept of either the correct or of the successful reference word can be evoked by replacing the input energy contour with the reference energy contour.