Feature extraction segmentation and labeling in the Harpy and Hearsay-II systems

H. G. Goldberg, and R. Reddy

Citation: The Journal of the Acoustical Society of America **60**, S11 (1976); doi: 10.1121/1.2003140

View online: https://doi.org/10.1121/1.2003140

View Table of Contents: https://asa.scitation.org/toc/jas/60/S1

Published by the Acoustical Society of America

tences (containing 1580 words) of about 3-sec duration each for a single speaker. The system requires about 12 Mipss (million instructions per second of speech) and uses about 200 000 words of memory on a PDP-10 system. More complete results, including several speakers and additional sentences, will be reported. [Research supported by the Defense Advanced Research Projects Agency.]

9:20

E3. The Hearsay-II speech understanding system. L. D. Erman, F. Hayes—Roth, V.R. Lesser, and R. Reddy (Department of Computer Science, Carnegie—Mellon University, Pittsburgh, PA 15213)

The Hearsay-II System has as its design goal recognition, understanding, and responding to connected speech utterances, particularly in situations where sentences cannot be guaranteed to agree with some predefined, restricted language model, as in the case of the Harpy System. Further, it attempts to view knowledge sources as different and independent which cannot always be integrated into single representation. It is based on the blackboard model [V. R. Lesser, R. D. Fennell, L. D. Erman, and D. R. Reddy, IEEE Trans. Acoust. Speech and Signal Process. ASSP-**23**, 11—23 (1975) with knowledge sources as a set of parallel processes which are activated asynchronously depending on data events. The system performs on the Information Retrieval task with accuracy comparable to that of the Harpy system, but runs about 2 to 20 times slower. More complete performance results will be reported. As we get closer to unrestricted vocabularies and nongrammaticality of spoken languages, it will be necessary to have systems which have the flexibility of Hearsay-II and the performance of Harpy. [Research supported by the Defense Advanced Research Projects Agency.]

9:30

E4. Feature extraction, segmentation, and labeling in the Harpy and Hearsay-II systems. H. G. Goldberg and R. Reddy (Department of Computer Science, Carnegie—Mellon University, Pittsburgh, PA 15213)

Goldberg [J. Acoust. Soc. Am. **59**, S97(A) (1976)] has shown that uniform techniques for segmentation and labeling can provide the initial signal-to-symbol transformation for speech recognition systems with reasonable accuracy and efficiency. Furthermore, the choice of parametric representation was not found to be critical for most commonly accepted representations. However, for efficiency, the computationally simplest techniques should be used to segment the utterance before more accurate (and expensive) spectral representations are used for labeling [R. Reddy, J. Acoust. Soc. Am. **42**, 329—47 (1967)]. To provide an initial symbolic input for both the Harpy and Hearsay-II systems, an hierarchical, feature-extraction based segmenter, using the ZAPDASH parameters, has been developed. After segmentation, labeling is done by a modified LPC minimum distance [F. Itakura, IEEE Trans. ASSP-**23**, 67—72 (1975)]. Labeling proceeds by comparing the midpoint of each segment with stored templates (acquired by an iterative learning process from speaker-specific training corpus) and adjusted with weights according to features obtained from the segmenter. The use of the highly efficient segmentation procedures and parameters provides approximately a factor of 5 speedup over uniform techniques which were previously used with both Harpy and Hearsay-II [Research supported by the Defense Advanced Projects Agency.]

9:40

E5. Connected Digit Recognition using symbolic representation of pronunciation variability. G. Goodman, B. Lowerre, R. Reddy, and D. Scelza (Department of Computer Science, Carnegie—Mellon University, Pittsburgh, PA 15213)

Most connected speech recognition systems such as Harpy and Hearsay-II use some form of symbolic representation alternative pronunciations of the vocabulary whereas most isolated word recognition systems use word templates. In an attempt to compare relative performance of systems that use symbolic representations of words, the Harpy system was run on a connected digit task requiring the recognition of random three-digit sequences. Each of ten speakers (seven male and three female) spoke 30 training sentences and 100 test sentences over a period of two weeks in a computer terminal room environment (approximately 65 dBA). Using speaker-dependent phoneme templates, the word error rate over all the ten speakers was about 2%. Using speaker-independent phoneme templates computed from the training data for all the speakers (male and female), the word error rate was about 8% for a test data set of 1200 random connected three-digit sequences from 20 speakers (including ten new speakers). The recognition time is about 4.5 Mipss (million instructions per second of speech). [Research supported by the Defense Advanced Research Projects Agency.]

9:50

E6. Parametric representation of speech. G. Gill and R. Reddy (Department of Computer Science, Carnegie—Mellon University, Pittsburgh, PA 15213)

As digital processing of speech becomes commonplace, it becomes desirable to have a parametric representation of speech which is simple, fast, accurate, and directly obtainable from the PCM representation of speech. The ZAPDASH representation of speech (Zerocrossings And Peaks of Differenced And Smooth waveforms) is one such. The PCM data is used to generate a different waveform and a down sampled, smoothed waveform (for 10-kHz sampling rate, the smoothing FIR filter coefficients were −1 0 1 2 4 4 4 2 1 0 −1, used every fourth point). Peak-to-peak distances and number of zerocrossings are calculated each 10 msec, resulting in 400 8-bit parameters per second of speech. ZAPDASH can be done in 15—20 computer instructions per sample and can be extracted in less than a 1/3 real time on minicomputers with 2 μsec instruction time. Although this representation is not noticeably different other similar proposals, it seems to be fairly robust and accurate, and is used in the feature extraction, segmentation, and labeling parts of the Harpy and Hearsay-II systems. Fortran and PDP-11 machine language versions are available from the authors. [Research supported by the Defense Advanced Research Projects Agency.]

10:00

E7. The HWIM speech understanding system—Overview and performance. Jared J. Wolf and William A. Woods (Bolt Beranek and Newman Inc., 50 Moulton St., Cambridge, MA 01238)

HWIM (for Hear What I Mean), the speech understanding system developed at BBN as part of the recent five-year ARPA Speech Understanding Research Project, is designed to "understand" naturally spoken utterances relevant to a task domain of travel budget management. Its vocabulary is over 1000 words, and its grammar permits a habitable subset of natural English. HWIM contains sources of knowledge at the levels of acoustic-phonetics, phonology, vocabulary, syntax, semantics, factual knowledge, and discourse. This paper describes the system as a whole and presents its performance results at the end of the ARPA project. [This research was supported by the Advanced Research Projects Agency of the Department of Defense and was monitored by ONR under Contract No. N00014-75-C-0053.]

10:10

E8. Phonetic and lexical processing in the HWIM speech understanding system. Richard M. Schwartz, John W. Klovstad, Victor W. Zue, John I. Makhoul, and Jared J. Wolf (Bolt Beranek and Newman Inc., 50 Moulton St., Cambridge, MA 02138)

The "front end" of HWIM, the BBN speech understanding system, is that part of the system that governs the formation