To appear in: *IEEE Expert*, August, 1996

Neural network speech processing for toys and consumer electronics

Michael C. Mozer
Sensory, Inc. and
University of Colorado, Boulder

The ongoing challenge in speech research is the recognition of continuous, unconstrained speech. In comparison, isolated word recognition with small vocabularies is easy. Many research efforts and commercial ventures are aimed at the high-end problem. Sensory has successfully focused on the low end, producing a family of inexpensive speech-recognition integrated circuits for markets such as interactive toys, consumer electronics, electronic learning aids, telecommunication devices, security systems, and household appliances. These markets are extremely cost conscious and demand solutions in the under five dollar price range. Previously, low-cost solutions were based on coarse properties of the speech signal such as the amplitude envelope or zero crossing rates, and typically yielded poor recognition performance. Adequate performance can be achieved using digital signal processor based approaches, but are beyond the reach of the low-cost market. Sensory uses a neural network based approach that obtains robust performance at a low cost.

Sensory's general-purpose controller, the RSC-164, can perform four speech recognition functions. *Speaker-independent recognition* works regardless of the age, accent, or gender of the user and without user-specific training, and can handle predefined vocabularies of up-to-fifteen alternatives. *Speaker-dependent recognition* requires user training, but can be used for vocabularies of up-to-sixty alternatives that are not specified in advance. Arbitrarily large vocabularies can be realized for both recognizers by switching the set of alternatives based on the context in which a response is requested. The recognizers also have the capability of rejecting an utterance as not belonging to the current recognition set, for example, when the user responds 'maybe' to a yes/no query. *Speaker verification* recognizes a stored password spoken by a particular speaker for security applications. *Word spotting* picks out a key word from continuous speech.

In addition to its speech recognition functions, the RSC-164 can perform polyphonic music synthesis, speech synthesis, sound effects, voice record and playback, and has the capability of driving and communicating with the application product. It thus serves as a single-chip solution for speech recognition, synthesis, and application product control.

The RSC-164 is a four-MIPS eight-bit microcontroller with an instruction set similar to that of the Intel MCS51. It has on-board analog-to-digital and digital-to-analog conversion, digital filtering, automatic gain control, 384 bytes of RAM, 64 kilobytes of ROM, sixteen programmable general-purpose input/output lines, an external memory bus, and can drive a speaker directly via a pulse-width modulator.

The neural network architectures and algorithms that underlie speech signal analysis in the RSC-164 are fairly traditional, but the RSC-164 imposes some interesting challenges. First, RAM, which consumes a large area on a chip, is minimal; the 384 available bytes includes storage for the current speech pattern to be recognized. Consequently, the implementation must minimize the use of intermediate state variables, and on-line learning is quite restricted because back propagation

Figure 1: Overview of speaker-independent recognition algorithm

training, which would require dynamic storage of multiple patterns, is infeasible. Second, although the microcontroller is relatively fast, it does not have hardware floating point arithmetic, and software emulation is quite slow. Thus, the neural net must be implemented primarily with fixed point computation. Third, while the digital filtering provides key information about the speech signal, it is not as rich a representation as the full power spectrum. To make the most of the data available, expert knowledge about salient features of the speech signal is utilized in the construction of the neural net input representation.

Speaker-independent recognition

Figure 1 shows the basic processing stages that occur when an acoustic signal is presented for speaker-independent recognition. A segmentation stage determines where the utterance begins and ends in the continuous input stream, a difficult problem due to background noise in natural environments. Once segmented, the utterance is passed through a set of digital filters. Time normalization is then performed and acoustic features are extracted from the filter outputs. The result of this preprocessing is a fixed-size vector of about a hundred elements which serves as input to a neural network recognizer. Based on the context in which the utterance is made, a network is selected which has the appropriate set of response alternatives. The network models the probability distribution of each word class conditional on the input pattern. Included among the response alternatives is a "none of the above" category, into which any word or sound not a member of the recognition set is classified.

As is typical in neural net applications, an appropriate input representation is essential. Because the input representation is at least as important to the recognizer's success as the type of classifier used, much effort has gone into the selection of optimal filter parameters and feature extraction algorithms. Also crucial to the recognizer's success is having an appropriate corpus of training examples. Sensory employs a staff of linguists who collect speech samples, roughly four hundred per vocabulary word. An effort is made to match the distribution of speakers in the training corpus with the distribution of eventual users of the product, by age, gender, and regional accent. Conditions of data collection are also matched to conditions of eventual usage, e.g., noisy environments, room echoes, and the physical state of the speaker. Neural net training is done off line, using the grab bag of tricks that have become standard in the field, including cross validation and Bayesian methods for model selection, adding weight constraints between neighboring input frames, and ensemble methods. Altogether, over two thousand networks are trained in the process of selecting the one that will end up in the application product.

Speaker-independent recognition with a vocabulary of a dozen alternatives achieves recognition accuracies greater than 95% in actual usage.

Speaker-dependent recognition

In speaker-dependent recognition, the user trains the system by speaking several examples of each vocabulary item. These examples serve as *templates* which are then serially compared to an *un-*

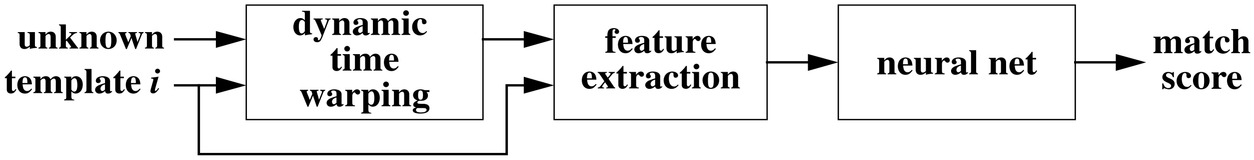

Figure 2: Overview of speaker-dependent recognition algorithm

known input pattern. Figure 2 shows the basic processing stages involved in this comparison. Dynamic time warping is used to align the unknown pattern to the template. The standard approach in speech recognition is to use the alignment score as the measure of how well the unknown pattern matches the template, and then to choose the template with the best match score. However, significant performance improvements have been achieved by comparing the warped unknown pattern to the template using a neural network, as shown in Figure 2. The neural network is trained in advance to produce good match scores when the template and unknown pattern represent the same utterance, or bad scores otherwise. Speaker-dependent recognition achieves accuracies of 98% with thirty word vocabularies.

Speaker verification

The speaker verification function of the RSC-164 operates very much like speaker-dependent recognition, although the training methodology and performance criterion are different. Speaker verification decisions can result in two types of errors. *False accepts* occur when the match score is sufficiently high that the unknown pattern is believed to match the stored password—i.e., it is the same utterance spoken by the same speaker. *False rejects* occur when the match score is too low, even though the speaker and utterance are the same as the stored password. By varying the match threshold, false accepts can be traded off against false rejects. One can select the threshold to minimize either false accepts or false rejects. The equal error rate for speaker verification—i.e., when the threshold is set such that false accept and reject rates are equal—is 1.5%.

Announced products

The RSC-164 was recently awarded EDN Magazine's Innovation of the Year Award in the category of integrated circuits and semiconductors. In addition to the RSC-164, Sensory has developed two special-purpose integrated circuits, one based on speaker verification (Voice Password [TM]) and the other speaker-dependent recognition (Voice Direct [TM]). The first products using Sensory's speech recognition technology were available in late 1995. Current products include electronic learning aids (Fisher-Price, Tiger/Playskool, Tomy, and VTech) and speaker verification for automobile remote security systems (CrimeStoppers, Guardian International, Omega Research). Over a dozen other products are under development. Sensory will sell in the millions of neural network based speech recognition chips in 1996.

Author biography

Michael Mozer received a Ph.D. in Psychology and Cognitive Science at the University of California at San Diego in 1987. Following a postdoctoral fellowship with Geoffrey Hinton, he joined the faculty at the University of Colorado at Boulder and is presently an Associate Professor in the Institute of Cognitive Science and the Department of Computer Science, and also serves as the Director of Neural Network Technologies for Sensory and as the General Chair of the Neural Information Processing Systems conference in 1996. He is the recipient of a 1990 National Science Foundation Presidential Young Investigator Award, and has over 70 publications in neural networks. His research interests include neural network models of human perception, attention, and awareness, and applications of neural network predictive and control techniques for home automation. Reach him at: Department of Computer Science, University of Colorado, Boulder, CO 80309-0430; `mozer@cs.colorado.edu`. `http://www.sensoryc.com`

Author info

Address: Department of Computer Science, University of Colorado, Boulder, CO 80309-0430
Fax: 303 492 2844
Email: `mozer@cs.colorado.edu`