Session CC. Speech Communication V: Neural Networks and Hidden Markov Models

Howard Nusbaum, Chairman
Department of Psychology, University of Chicago, 5848 South University Avenue, Chicago, Illinois 60637

Contributed Papers

1:15

CC1. Neural networks for estimating articulatory positions from speech. Bishnu S. Atal and Olivier Rioul (Acoustics Research Department, AT&T Bell Laboratories, Murray Hill, NJ 07974)

This talk describes an application of neural networks for estimating positions of various articulators (such as the tongue, the lips, etc.) in the vocal tract from the speech signal. In general, a neural network consists of a large number of interconnected computational elements. The networks that will be discussed in this paper include an input layer of nodes connected directly or through an intermediate layer of hidden nodes to an output layer. Iterative gradient search procedures are often used for determining the unknown parameters of the neural network, but these procedures are very slow for training neural networks with a large number of hidden nodes. For estimating articulator positions, it was found that the weights in the first layer could be set to fixed random values during the training procedure without degrading the performance of the network. The random fixed weights in the first layer permit the use of a fast noniterative procedure for determining the unknown parameters of the second layer. Tests on a vocal tract model with ten articulatory variables show that the articulator positions can be determined accurately, using a network with 500 hidden nodes in the intermediate layer, from ten LPC parameters derived from the speech output of the model.

1:27

CC2. Acoustic feature development during unsupervised learning by a neural net. Bradley S. Seebach and Nathan Intrator (Center for Neural Science, Box 1843, Brown University, Providence, RI 02912 and Division of Applied Mathematics, Brown University, Providence, RI 02912)

A biologically plausible neural network model that employs unsupervised learning was applied to various sets of CV syllables. This network has been shown to develop recognition of input signals on the basis of distinctive signal features rather than overall signal shape [N. Intrator and B. Seebach, Int. Neural Network Soc. Abstr. 1, Suppl. 1, 299 (1988)]. Syllables pronounced in isolation by male and female speakers were digitized and sampled in short (8–32 ms) overlapping time windows, then filtered into overlapping critical bandwidths [E. Zwicker, J. Acoust. Soc. Am. 33, 248 (1961)] to produce three-dimensional energy surfaces in time and frequency. A portion of these syllabic tokens was used as a training set for the net. Those remaining were used to test generalization of network solutions both within a single speaker's utterances and across speakers. For example, when trained on a single speaker's tokens, and tested for classification of place of articulation in stop consonants, the network might correctly identify approximately 80% of similar tokens from a different speaker.

1:39

CC3. A connectionist model for classifying speech into silence, glottal source, burst friction, or mixed categories. Steven J. Sadoff (Central Institute for the Deaf, 818 South Euclid, St. Louis, MO 63110)

An algorithm for classifying speech into four classes (silence, only glottal source, only burst friction, or mixed) is being developed. This scheme primarily differs from the standard silence/aperiodic/periodic classification in that the defining characteristics of glottal source and burst friction sounds do not depend on periodicity distinctions, but on the locus of the energy concentrations. One male speaker reciting the Rainbow passage has been recorded and analyzed. Utilizing a strictly layered backpropagation network, the automated learning procedure is trained using the first 30 s of the passage; the final 75 s are used for testing. Quantitative results, along with several examples illustrating this model, will be presented. Additionally, the performance and classification strategy of the network will be compared to that of human observers. [Work supported by AFOSR.]

1:51

CC4. Context-modulated discrimination of similar vowels using second-order connectionist networks. Raymond Watrous (Department of Computer Science, University of Toronto, Toronto, Ontario M5S 1A4, Canada)

Discrimination of medial adjacent vowels in the context of voiced and unvoiced stop consonants using connectionist networks is investigated. Separate discrimination networks were generated for one speaker from samples of the vowel centers of [e,æ] for the six contexts [b,d,g,p,t,k]. A single context-independent network was similarly generated. The context-specific error rate was 1%, whereas the context-independent error was 9%. A method for merging isomorphic networks into a single network is described. The method uses singular value decomposition to find a minimal basis for the set of context-specific weight vectors. Context-dependent linear combinations of the basis vectors may then be computed using second-order network units. Compact networks can thus be obtained in which the vowel discrimination surfaces are modulated by the phonetic context. In a preliminary experiment, as the number of basis vectors was reduced from 6 to 3, the error rate increased from 1% to 3%. Experiments with nonlinear optimization of context-modulated, second-order networks for this task are underway and will be reported.

2:03

CC5. Automated language acquisition. A. L. Gorin and S. E. Levinson (Speech Research Department, AT&T Bell Laboratories, Murray Hill, NJ 07974)

A new approach to developing large-vocabulary speech understanding systems, in which the system automatically acquires a language model for its task based on semantic information, is introduced. This is in contrast to previous work, in which language was preprogrammed. An important consequence of this approach is that it leads to *habitable* language models. To accomplish this task, use is made of a medium-grain neural network, together with a novel adaptive training procedure for estimating the set of connection weights. The resulting connection weights have an information-theoretic interpretation, and do not require gradient search techniques for their estimation. A conversational-mode system that serves as a test bed for the network is described. The application scenario is inward-call management, where a customer telephones a large organization and encounters an operator whose function is to forward the call to