Trainable grammars for speech recognition

J. K. Baker

Citation: The Journal of the Acoustical Society of America **65**, S132 (1979); doi: 10.1121/1.2017061
View online: https://doi.org/10.1121/1.2017061
View Table of Contents: https://asa.scitation.org/toc/jas/65/S1
Published by the Acoustical Society of America

HEAR acoustic processor, wholly developed and tested for continuous speech recognition at the IBM Thomas J. Watson Research Laboratory in Yorktown Heights, NY, applies Baum's algorithm for prototype selection. Peformance on two tasks with extensive training on a single speaker, in a quiet, high quality recording environment are reported for HEAR, used in conjunction with the standard training and decoding programs developed by the IBM Research Continuous Speech Recognition Group. Sentence recognition on a set of 125 test sentences (1010 words) of the artificial language 250 word "New Raleigh Language" is 100%. The word recognition rate on a set of ten test sentences (282 words) of the natural language, 1000 word "Laser-1000 Task" is 80% and reflects the performance of all the system components; e.g., search strategy, language model, pruning procedures, acoustic processor, etc. For only 3.2% of the words is the correct word hypothesized, fully matched, and rejected in favor of an incorrect word. These errors are comprised exclusively of substitutions and deletions of short function words (e.g., "of," "the," etc.), 2.2%, and deleted commas (realized acoustically only by optional interword pauses), 1.0%.

3:40

YY11. Trainable grammars for speech recognition. J. K. Baker (Dialog Systems, 32 Locust Street, Belmont, MA 02178)

Algorithms which are based on modeling speech as a finite-state, hidden Markov process have been very successful in recent years. This paper presents a generalization of these algorithms to certain denumerable-state, hidden Markov processes. This algorithm permits automatic training of the stochastic analog of an arbitrary context free grammar. In particular, in contrast to many grammatical inference methods, the new algorithm allows the grammar to have an arbitrary degree of ambiguity. Since natural language is often syntactically ambiguous, it is necessary for the grammatical inference algorithm to allow for this ambiguity. Furthermore, allowing ambiguity in the grammar allows errors in the recognition process to be explicitly modeled in the grammar rather than added as an extra component.

3:50

YY12. Talker-independent speech recognition in commercial environments. S. Moshier (Dialog Systems, 32 Locust Street, Belmont, MA 02178)

A machine which performs long distance voice telephone dialing and charge account verification for 4000 civil service employees of the State of Illinois was installed in April 1978. This and similar installations now perform routine daily services for thousands of users. An overview of the development of this system is presented, including discussion of human factors studies, computer engineering, voice data base development, software operating systems, and the speech recognition method *per se*.

4:00

YY13. The human factor in isolated word data entry. R. W. Phelps and R. A. Wiesen (Dialog Systems, 32 Locust Street, Belmont, MA 02178)

At present, most of the applications at Dialog Systems involve numeric input of one kind or another. Applications accomodate telephone speech from a large population and range from telephone toll management and switching to telephone bill paying which involves rather lengthy data entry. All entry is by isolated word. This paper will concentrate on users of these systems. Training and user format design along with human and system performance will be discussed in the context of a system that has been in operation for about nine months with some 4000 users. Since the user is affected by constraints of voice recognition protocol and parameters of the voice recognition system, these topics will also be addressed. Particular emphasis will be given to two styles of digit input. In the first, each digit is verified on input. In the second, short strings of digits are input and then verified by the user. Data from about 60 subjects will be presented and trade-offs between the two systems will be discussed. Dependent upon recognition rate, a mixture of the two styles gives an optimum solution.

4:10

YY14. Small sample statistics for likelihood-ratio tests. L. Bahler, S. Moshier, and T. Rey (Dialog Systems, Inc., 32 Locust Street, Belmont, MA 02178)

Classification by likelihood ratio of a one-dimensional random variate drawn randomly and with equal probability from one of two independent Gaussian populations is known to be optimal if misclassification costs are equal and statistics (true means and variances) are available; the expected classification error is then a known function of the statistics. In practice, likelihood functions are computed with sample statistics estimated from training sets of n samples. The effect of sample statistics in lieu of true statistics on the expected classification error is calculated to order n^{-1}. The error is found to increase by a term proportional to n^{-1} over a wide range of statistics, unless the random variate is from the training set; in the latter case, the error decreases by that term. Monte Carlo experiments were performed with results supporting the theory.

4:20

YY15. A language and multitasking operating system to support an eight channel speech input terminal. S. Glazer (Dialog Systems, Belmont, MA 02178)

A speech input terminal must be able to carry on a natural conversation with a user. This presents several special problems for the designer of the terminal's operating system. Multiple channels must be handled independently and must not interact with other channels. Timing must be sufficiently precise so that there are no unnecessary pauses before listening or speaking. Changes must be easy to implement in order to experimentally optimize the human factors aspects of the conversation. The system must lend itself to modularization and easy documentation. These issues are discussed and the language and operating system used at Dialog are presented.

4:30

YY16. A versatile vector processor for multichannel speech recognition. R. R. Osborn (Dialog Systems, 32 Locust Street, Belmont, MA 02178)

A 32-bit 122-ns cycle time computer has been developed for use as both a speech research tool and as the basis of a speech recognition product. It is structured as a single instruction stream device with three data buses independently controllable during each cycle. Function modules are connected between the buses. The timing and architecture are straightforward, allowing the addition of new functions as independent modules. Pipelines can be configured under software control and the instruction set is well structured and small, eliminating some of the complications of microcoded implementations. A standardized configuration is described which allows eight simultaneous channels of isolated word recognition in real time. The standardized configuration is currently being produced as part of a speech recognition product line.