Statistical Methods for Speech Recognition

Frederick Jelinek
(The Johns Hopkins University)

Cambridge, MA: The MIT Press
(Language, speech, and communication
series), 1997, xvii+283 pp; hardbound,
ISBN 0-262-10066-5, $35.00

Reviewed by
Eric Neufeld
University of Saskatchewan

Current practitioners in the area of speech recognition who are familiar with the approach of Jelinek and others will find this a compact, concise, and useful overview of the state of the art in statistical approaches to speech recognition. Readers already familiar with Rabiner and Juang (1993) will find it an excellent companion volume.

Computational linguists will also find this book to be engaging reading. For one thing, Jelinek's lucid and well-organized survey might well have been written with the mathematical backgrounds of computational linguists and computer scientists in mind. Apart from a smattering of probability, the reader needs only some basic discrete mathematics.

This is partly because Jelinek does not discuss signal processing, the crucial first step in speech recognition, where continuous vectors representing waveforms are matched with discrete phonetic symbols. Some may think this omission surprising, but from Jelinek's quick sketch of the vector quantization algorithm, it seems reasonable to assume that although signal processing is critical ("bad [signal] processing means loss of information: there is less of it to extract"), developments in that discipline and advances in statistical aspects of speech recognition can proceed relatively independently of each other. (Readers wishing to explore signal processing in the context of speech recognition might consult Rabiner and Juang [1993].)

Thus, Jelinek initially concentrates on the problem of building a speech recognizer that can be trained to construct the most probable hypothesis (a string of English text originating in a speaker's mind) that explains a perceived string of phonetic symbols reaching the hearer's ear. These strings of phonetic symbols are outputs of an acoustic processor: finite sequences of discrete symbols from a finite phonetic alphabet such as might be used in a dictionary.

In his introduction, Jelinek writes "I am fascinated by the idea that while system structure and parametrization should come from intuitive understanding of the process, the parameter values are best extracted from the data." The speech recognizers that he describes use the assumption that natural language at the phonetic and semantic levels is an output-generating Markov process and use the outputs (sounds) to infer the text that produced it. This, of course, is the celebrated method of hidden Markov models (HMMs), and while it seems to oversimplify the structure of language, this method gives the recognizer the ability to exploit context to resolve ambiguities. (Some speech recognizers use neural net models; Jelinek does not discuss these.)

Although by now most people in CL know about HMMs and understand in broad terms how they are used, many would be surprised to see how remarkable a first cut at a solution they provide. Even when the English text in the speaker's mind is pronounced idiosyncratically and convolved with ambient noise before it reaches the hearer, contextual relationships can still provide the (computer) hearer with sufficient clues to decode the message. (As a side note, there may also be cognitive or psychological validity to this approach: Saffran, Aslin, and Newport [1996] suggest that segmentation of words from fluent speech can be accomplished by 8-month-old infants solely on the basis of the statistical relationships between neighboring speech sounds.)

The first half of the book presents the basic algorithms needed to build and parametrize an HMM-based speech recognizer. Jelinek devotes a chapter to the Viterbi algorithm and another to hypothesis search on trees. He also gives the Baum-Welch (forward-backward) algorithm that is guaranteed to improve any initial estimate of parameters (though it may converge to a local maximum). He also describes the method of optimal linear smoothing that ensures that sequences not seen during training are not assigned zero probabilities when encountered after training. A larger training set does not solve this problem because no training set can contain every possible sequence that may eventually occur in practice.

The last half of the book refines the ideas of the first half. It requires more mathematical knowledge from the reader and Jelinek provides a chapter of required background. He discusses the EM algorithm from which the Baum-Welch algorithm can be derived and which allows the generalization of HMMs that output vectors of normally distributed real numbers. This lets a speech recognizer exploit waveform information that vector quantization loses. He also discusses the problem of coarticulation, the influence of phones on one another. For example, the waveform for *i* in *king* more closely resembles the *o* in *moves* than the *i* in *bishop*. This is due to the influence of the nasal *ng* in *king* and the *m* in *moves*. The conflict therefore can be resolved by considering the phonetic context, i.e., the most probable *sequence* of phones.

The lifetime of experience that Jelinek brings this presentation provides a great foundation upon which AI and CL practitioners can build. Recently, Zweig and Russell (1998) report successes using dynamic Bayesian nets, a natural generalization of HMMs, to handle the coarticulation problem in isolated-word speech recognition.

References

Rabiner, Lawrence R. and Biing-Hwang Juang. 1993. *Fundamentals of Speech Recognition*, Prentice-Hall, Englewood Cliffs, NJ.

Saffran, Jenny R., Richard N. Aslin, and Elissa L. Newport. 1996. Statistical learning by 8-month-old infants. *Science*, 274(5294), 13 December 1996, pages 1,926–1,928.

Zweig, Geoffrey and Stuart Russell. Speech recognition with dynamic Bayesian networks. *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-98)*, Madison, Wisconsin: AAAI Press, pages 173–180.

Eric Neufeld is interested in probabilistic approaches to artificial intelligence, including natural language processing. With Greg Adams, he has written several papers on HMM-based approaches to part-of-speech tagging of text. Neufeld's address is Department of Computer Science, University of Saskatchewan, Saskatoon, SK, Canada S7H 3A8; e-mail: eric@cs.usask.ca