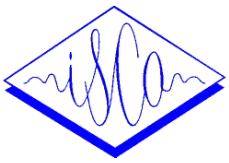

3rd European Conference on
Speech Communication and Technology
EUROSPEECH'93
Berlin, Germany, September 19-23, 1993

MODELS OF SPEECH RECOGNITION

Personal Perspectives on Particular Approaches

John S Bridle

Dragon Systems UK Ltd, Millbank, Stoke Road, Bishops Cleeve, Cheltenham, GL52 4RW, England.

ABSTRACT

The aim of this keynote tutorial is to explore the nature of Automatic Speech Recognition and its relationships to human speech communication. We then examine various approaches evident in the papers presented at this meeting, based on consistency with linguistic theory, with statistical properties of speech data, or with theories of brain function.

10.21437/Eurospeech.1993-9

1. HUMAN COMMUNICATION AS A MODEL

Communication by speech is a wonderful and very important process. We hope that our machines would be easier to use and more effective helpers if we could communicate with them using speech as easily as we can communicate with other people. We are therefore driven to explore the possibilities of Automatic Speech Recognition: one of the crucial technologies needed for future Interactive Spoken Language Systems.

There is some debate about whether Automatic Speech Recognition can usefully be dealt with separately from the complete system of which it is a part, and this may well be true of the type of system featured in Star Trek. However, the job of interpretation, or *understanding*, and the decisions about how to respond, will not be my concern. (In some applications, including virtually all that have been deployed to date, there is very little interpretation to do: examples are data entry, command-and-control of machinery, and dictation of text.) I shall simply assume that the job of the ASR sub-system is to deliver one or more explanations of each utterance, based on the acoustics, and information about

• Acoustic-phonetic properties of the language (the sounds)

• The Lexicon (the possible words, and their phonetic spellings)

• The Grammar (the way the words are likely to be used together)

All three may have generic component for any particular language, and a speaker-specific component that is adapted to suit the speaker.

Before we go on, it is worth pausing to question our assumptions.

• Is human speech communication so wonderful? I often find it painfully slow and error prone, particularly over the telephone.

• Can we not imagine something much more effective? Asimov's description of spaceship controls in *Foundation's Edge* is an inspiration. The possible consequences of direct connections between brains and computers are vividly presented in Marvin Minsky's novel *The Turing Option*.

• If we are to use speech, does it have to be much like human conversation? I suggest that a much more terse style may be more appropriate- more like that used in text-based adventure games. We may be inspired by the challenge of creating a system intended to respond appropriately to something like "I'm planning a trip to Berlin in September – can you tell me about travel possibilities", but the general public may rightly assume that they should use a style more like "From London to Berlin in September please."

• Are we limited to merely human performance? We already have connected digit recognizers that can respond correctly to long rapidly spoken digit strings that would tax the short-term memory of any normal person. There are also large-

vocabulary dictation systems that can deal with a range of specialized medical vocabulary better than all but a specialist medical secretary. Of course, the system behind the recognizer – the system we are talking to – is likely to have some super-human powers, such as the ability to access information very fast from a vast, dynamically-changing database.

Surely the answer is that human speech communication serves as a reference point, and a starting point, for any new user of a speech-operated computer system. It is up to the designers of the system to indicate the ways that the system is unable to meet expectations, and also to encourage the user to exploit the ways that performance is un-human or super-human.

2. STOCHASTIC MODELS

Virtually all working automatic speech recognition systems can be understood in terms of a very important approach based on stochastic models. A Hidden Markov Model (HMM) is a type of Stochastic Model (SM). The model is a way of encoding the speech and language information that the recognizer uses. The model itself can be thought of a probability distribution, or as a machine for generating synthetic example speech patterns. (In fact most current SMs make rather poor speech synthesizers, for reasons that may become apparent.)

An HMM of a word would typically consist of a set of hidden *states*, probabilities of *transitions* from one state to another (thus defining a Markov chain), and parameters of *output distributions*, one for each state. The generating machine is in one of these states at a time, and can make state transitions at regular intervals, typically 100 times per second. For each interval, the machine generates a symbol or a vector of numbers, corresponding to an acoustic *observation*. Each observation is chosen from an *output distribution* which depends on the current state, or in some systems on the details of the state transition just made.

We set up the model initially from prior knowledge, based on general acoustic-phonetic knowledge or just what has worked before. Examples are the number of states, the pattern of possible state transitions, and the form and initial parameters of output distributions.

The crucial step is to adjust the parameters based on example speech. The idea is that the distribution of patterns produced by the model should be as close as possible to the distribution of real patterns that correspond to that model.

The SM models more than knowledge – it also models ignorance! We may care to think of some aspects of the structure as knowledge (whether this was built in by an expert or learnt from data) All variability of the acoustic pattern that is not explained by such knowledge is treated as random variation. Even though in many cases we know that there is rule-governed variation not captured by the structure of the model, this is treated as noise. Apparently this can be better than ignoring the source of variability or of modelling it badly. An example might be the differences in the speech of different people, which are often dealt with as if the variation between voices could occur at the same rate that the model changes states!

The recognition process uses the Stochastic Model as a static data structure while interpreting the incoming acoustic pattern in terms of the model. In a sense, the recognition, or *search*, process is un-interesting, because it should not alter the answers from the recognizer; only the time, memory and processor power used to find them. However, it is often the need for efficient search methods which limits our choice of model structure. *Learning, training* or model parameter *estimation*, usually relies on a process of the same general form as the recognition process, and this too can limit our choice of model.

The most important principle used in recognition and training of most current speech recognition systems is that of *dynamic programming*: we rely on a property that we usually build into our models (the markov property) that the only information about the past behaviour of the system needed for predicting its future behaviour is the state. The idea of stochastic modelling does not depend on the markov property, but in practice it is almost always used in some form.

It is important to keep in mind that a stochastic model does not recognize speech: it is a model of the process that *generated* the acoustic pattern, and the repository of information used by the (separate) recognition process.

3. NEURAL NETWORK MODELS

At the same time as the increasing success of stochastic model based speech recognition systems, there has been a great deal of excitement in recent years about the potential, for ASR and many other pattern processing tasks, of a variety of ideas inspired in various ways by analogies with the structure and supposed principles of functioning of nervous systems. Such ideas go by names such as *connectionist*, *PDP*, *sub-symbolic*, and most popular (although often misleadingly) *artificial neural networks* (ANNs).

The type of ANN most frequently used for ASR, often known as a *multi-layer perceptron* (MLP), is a continuous function which maps input vectors, such as acoustic spectrum shapes, to output vectors that often serve as indicators of probabilities of alternative class labels such as phonetic classes.

The function is composed of a cascade of much simpler functions; typically each constituent, or *model neuron*, is a non-linear scalar function of a weighted combination of inputs to the network and/or outputs of other neurons. The whole function is parameterized by continuous variables, typically the *weights*.

MLPs have been used for speech recognition in many ways, often used together with dynamic programming and stochastic models. There are also many other varieties of artificial neural network, and almost all of them have been applied to speech recognition by somebody – not always conclusively.

Neural networks used directly as recognizers are often referred to as models, but we should remember that they are models of the recognition process, rather than the generation process.

Neural network ideas have challenged the stochastic modellers to consider a wider variety of configurations, and to understand important issues such as *discriminant training* and *generalization*.

Most researchers would now deny that stochastic modelling and neural networks are distinct approaches to speech recognition, and we are in a period of cross-fertilization and better understanding.