

WHY HAVE HMMS BEEN SO SUCCESSFUL FOR AUTOMATIC SPEECH RECOGNITION AND HOW MIGHT THEY BE IMPROVED?

Wendy HOLMES and Mark HUCKVALE

Abstract

Most of the current successful systems for automatic speech recognition are based on hidden Markov models (HMMs). HMMs are basically general-purpose statistical pattern matchers, but have so far proved more successful than approaches which have been based on specific knowledge about speech. This paper discusses the likely reasons for this success. It is argued that, in addition to providing a tractable mathematical framework with straightforward algorithms for training and recognition, HMMs have a general structure which is broadly appropriate for speech: both short-term spectral variability and temporal variability can be modelled. This general structure can be tailored to known characteristics of the sounds being modelled, but the actual parameters are optimized based on training data. Another very important characteristic of HMMs is that they provide a complete model which can also be employed at higher levels (such as syntax), with only a single decision being made based on finding the best match at all levels. All these aspects combine to make HMMs a powerful approach to speech modelling. It is therefore argued that, in order to progress recognition capabilities further, the best approach is to retain all these advantages of the general framework but to overcome the limitations of the current HMM formalism by making it segment-based rather than frame-based.

1. Introduction

Automatic speech recognition (ASR) is a difficult and complex problem. The understanding of unconstrained fluent speech by machine remains a distant objective, but in recent years there have been significant and steady advances in systems operating in more constrained domains. Almost perfect accuracy can be achieved with speaker-independent isolated-digit recognition, with only 2-3% digit string errors when the digit sequence is spoken in a naturally connected manner by non-specific talkers (Juang and Rabiner, 1991). Furthermore, some systems have achieved about 96% word accuracy for speaker-independent connected-speech recognition with a 1000-word vocabulary and some grammatical constraints (Marcus, 1992).

Currently, most of the successful ASR systems are based on hidden Markov models (HMMs). HMM techniques are basically general-purpose pattern recognition algorithms, yet they have been much more successful than the previous systems which focused on the explicit use of speech knowledge (Klatt, 1977). Basic HMMs do not take advantage of many of the well-known characteristics of speech, and in addition they make certain assumptions about the speech signal which are clearly incorrect. In view of these potentially quite serious disadvantages, this paper considers the issue of why these statistical models have actually been very successful for characterizing speech. Section 2 summarizes the main characteristics of spoken language that make speech processing such a challenging problem. The next section briefly discusses the different types of approaches to ASR, together with their associated advantages and limitations. Section 4 then provides further background by briefly considering how approaches to ASR relate to theories of human speech perception. This is followed by a more detailed discussion of the reasons why HMMs are successful as models of speech. The final section considers how the HMM formalism could be extended and improved so that the underlying model is more appropriate for speech, while retaining the advantages of the general approach.

2. Challenges for automatic speech recognition

The task of ASR is to extract the underlying linguistic message from a complex acoustic pattern containing many sources of variability. Types of variability which must be accommodated include:

- Phonetic context effects, both from coarticulation with immediately surrounding sounds and from larger context such as position in a whole sentence.
- Differences between speakers due to sex, age, accent and so on.
- Occasion-to-occasion differences within any one speaker. These variations can arise from speed or loudness of speaking, effects of stress, or from having a cold or sore throat.
- Environmental variations, such as the surrounding noise level. There may be noise which masks some sounds, or even whole words.

A speech recognition system needs to treat the many possible variants of any one phoneme as the same, while being sensitive to the differences between different phonemes spoken by one speaker in the same phonetic environment.

Understanding fluent spoken sentences is not simply a case of extracting the sequence of phonemes from the acoustic signal: the signal contains information at several different but interacting levels, including syntax, semantics and pragmatics. By using information at these different levels, it is possible to take advantage of the redundancy which exists in speech and so resolve any ambiguities at any one level and comprehend a message even when parts of it are missing. Ideally, a system for ASR should use the information at all these levels in order to assist with the difficult problem of coping with all the sources of acoustic variability.

3. Approaches to speech recognition

Generally speaking, approaches to ASR can be divided into two types which can be classed as "knowledge-based" and "self-organizing" (Mariani, 1991). These terms reflect the contrast between systems which are based on explicit formulation of knowledge about the characteristics of different speech sounds, and systems where a much more general framework is used and the parameters are learned from training data.

3.1. Knowledge-based approaches

Efforts to develop speech recognition systems based on the explicit use of speech knowledge began in the early 1970s within the framework of the Advanced Research Projects Agency (ARPA) speech understanding project (Klatt, 1977). Several systems were developed, most of which adopted a two-stage approach whereby the acoustic signal was first segmented and labelled into phoneme-like units, and the resulting string was then subjected to further analysis. The emphasis was on applying "Artificial Intelligence" techniques to use "higher level" knowledge (lexicon, syntax, semantics and pragmatics) to obtain an acceptable recognition rate even if the initial phoneme recognition rate was poor. The "knowledge" used in these systems was generally related to the constraints of the task the system was designed to perform, rather than principled general linguistic knowledge. The resulting systems were computationally expensive, limited to the particular task for which they were designed, and produced quite poor recognition performance. A fundamental

problem with this type of approach is that it is inevitably limited by the accuracy of the acoustic-phonetic decoding: it is still not possible to reliably extract phonetic information from the speech signal.

More recently, several systems have been developed based on "expert systems" modelling the human ability to interpret spectrograms or other visual representations of the speech signal (see O'Brien, 1993, for a review). Such systems separate the knowledge that is to be used in a reasoning process from the reasoning mechanism which operates on that knowledge. The knowledge is usually manually entered, and is based on the existence of particular features (such as "a silence followed by a burst followed by noise" for an aspirated voiceless stop). A vast amount of knowledge would be needed for speaker-independent continuous speech recognition of large vocabularies (Mariani, 1991). The problem with a system based on a large set of rules is that it is difficult to imagine all of the ways in which the rules are interdependent, so inevitably some rules compete with each other to explain the same phenomenon while others are in direct contradiction (Levinson, 1985a).

Some knowledge-based systems have been developed which reduce this problem by only specifying the set of features to be used: the values of the features and the weights given to different features in the recognition process are learned from training data. An example of this type of approach is Allerhand's (1987) knowledge-based statistical classifier. The success of this approach is however still very dependent on appropriate selection of the feature set and on the correct extraction of these features from the acoustic data.

3.2. Self-organizing approaches

The alternative type of approach provides a general structure and allows the system to learn the parameters from a set of training data. The simplest approach of this type is "template matching", whereby a template is generated for each word in the vocabulary to be recognized, based on one or more examples of that word. Recognition then proceeds by comparing an unknown input with each template using a suitable spectral distance measure. The template with the smallest distance is output as the recognized word. To accommodate the inevitable differences in time scale which exist between different examples of the same word, the technique of dynamic programming is used to find the optimal alignment of the input pattern with the template (Bridle, Brown and Chamberlain, 1983). In order to cope with variability in spectrum as well as time scale, each template can be derived from several examples of the relevant word, based on either computing an average over all examples or selecting the most representative example according to some distance criterion. For speaker-independent recognizers, which need to recognize a wide range of speakers' voices, templates can be derived from examples of the words spoken by a variety of different speakers. For each word, a clustering algorithm can be applied to the examples, in order to generate a set of templates which represent the main different pronunciations of that word.

Template matching was the predominant approach to ASR from the late 1970s up until the mid 1980s, and was the basis for most commercial systems during this period (for example, recognizers produced by Votan). In this period almost all commercial systems were speaker-dependent. Good recognition performance has been achieved with constrained tasks where the vocabulary is quite small and words are spoken in isolated form. However, performance deteriorates for tasks involving larger vocabularies or connected speech. This is due to inherent limitations in the nature of the approach: there is some allowance for variations in pronunciation and speaking rate by using multiple-reference templates and applying dynamic programming in recognition. In real speech, however, there will be a range of plausible variants of each speech sound, some of which are more likely than others. The extent and type of possible variation, both in time scale and in spectrum, will be different for different sounds. The ability to model this statistical variability is the main advantage

of the HMM approach (see, for example, Bourlard, Kamp, Ney and Wellekens, 1985, for a general introduction to HMMs). HMMs provide a formal statistical framework that is broadly appropriate for modelling speech patterns: the time-varying nature of spoken utterances is accommodated through an underlying Markov process. Statistical processes associated with the model states define output probability distributions and so encompass the variability which occurs both between and within speakers when producing linguistically equivalent speech sounds. This broadly appropriate framework is combined with computationally useful and rigorous mathematical methods for automatically optimizing the parameters of a set of HMMs relative to training data, and for classifying an unknown speech pattern given a set of HMMs. Using HMM techniques, it is now possible to obtain good recognition performance for connected-word medium vocabulary whole-word matching systems. By using sub-word units and a statistical language model, it is also possible to achieve a useful performance on very large vocabulary isolated-word recognition.

There are also approaches to ASR, with origins in work on artificial intelligence and parallel computation, which use Artificial Neural Networks such as multi-layer perceptrons (MLPs). An MLP consists of a network of interconnecting units, with an input layer, an output layer and one or more hidden layers. The output units represent the set of speech units to be recognized, and the recognition process relies on the weights of the connections between the units. The connection weights are trained in a procedure whereby input patterns are associated with output labels. MLPs are therefore learning machines in the same way that HMMs are, but have the advantage that the learning process maximizes discrimination ability, rather than just accurately modelling each class separately. However, MLPs have a serious disadvantage in that, unlike HMMs, they are unable to deal easily with the time-sequential nature of speech. Good recognition results have been reported for isolated-word recognition (e.g. Peeling and Moore, 1988) by storing the entire input sequence to be recognized in a buffer at the input to the MLP. The problem is that this approach does not generalize to connected speech or to any task which requires finding the best explanation of an input pattern in terms of a sequence of output classes. To circumvent this problem, there has been an interest in approaches which are a hybrid of HMMs and neural networks. These are discussed further in Section 6, as they really represent one way of improving HMM systems. However, as an underlying approach, it is generally agreed that the HMM framework has shown clear performance superiority over alternative recognition structures (Juang and Rabiner, 1991). HMMs are currently the most popular approach to ASR, both for research systems and for new commercial systems such as the DragonDictate (Baker, 1991). The existence of simple-to-implement, elegant and efficient algorithms for both training and recognition has been suggested as the reason for the popularity and success of the approach (Gales and Young, 1993). General appropriateness of the models for characterizing speech is however obviously also important for them to be successful. It is therefore useful to briefly consider models of human speech perception, before discussing in more detail possible justifications for tackling speech recognition by machine from an HMM framework.

4. Relationship between ASR approaches and models of speech perception

Humans are able to easily understand speech from a wide variety of speakers under different conditions of stress and in different acoustic environments, with a recognition performance considerably better than that offered by current technology. An accurate, complete model of human speech perception in a form amenable to computational simulation would obviously be relevant for approaches to ASR. A variety of models for various aspects of speech perception have been proposed from different theoretical viewpoints (see Klatt, 1989, for a review). Many of these models, proposed in the context of speech perception, have strong parallels with a corresponding approach to ASR. A few examples are considered here.

There have been a number of attempts to specify "phonetic features" and ways that they might be extracted from the acoustic signal (e.g. Pisoni and Luce, 1987). These approaches suffer the same limitations as knowledge-based approaches to ASR, in that they rely on the ability to reliably identify meaningful features: there has been only limited success in specifying a useful set of features that can be reliably extracted (Klatt, 1986).

An alternative model of perception, LAFS (Lexical Access From Spectra), was suggested by Klatt (1979, 1986), influenced by some aspects of the computational strategies created in the ARPA speech-understanding project (Klatt, 1977). This model originated from the belief that a pattern-matching scheme can actually be incorporated into a plausible perceptual model. The model proposes that the expected spectral patterns for words and for cross-word-boundary recordings are stored in a very large decoding network of expected sequences of spectra. It is therefore assumed that any phonetic transition can be characterized to an arbitrary degree of accuracy by a sequence of spectra, or by several alternative spectral sequences. A fully expanded decoding network enumerates all possible spectral sequences for all possible word combinations of English. A uniform spectral distance metric is used to compute phonetic differences between the input spectrum and each spectral template in the network. This allows the network to determine matching scores for all possible words in parallel, and perception consists of finding the best match between the input representation and paths through the network. In this model, the first decision made is a lexical one.

Klatt points out the advantage of this delayed-decision model over many other approaches to speech perception, which use a two-stage model involving an intermediate representation based on some form of phonetic features. A separate phonetic-feature stage discards information that might be useful in making lexical decisions, and may introduce errors. As with pattern-matching approaches to ASR, the main problem with the LAFS model is in defining a spectral distance metric that is powerful enough to make fine phonetic discriminations in the face of the kinds of spectral variability seen in real data.

Models based on neural networks are inherently attractive from a perception viewpoint, due to the analogy with the architecture of the human brain. One model which grew out of work in the area of parallel distributed processing (PDP) is the TRACE model (McClelland and Elman, 1986). The PDP philosophy of distributing knowledge through a network of interacting units was seen as a natural way to capture the integration of multiple sources of information in speech perception. The TRACE model involves a very large number of units organized into three levels: the feature, phoneme and word levels. The network is highly interconnected with bi-directional facilitatory connections between levels that share common properties (e.g. between a word and the phonemes composing that word), and inhibitory connections within levels (e.g. if /b/ is activated it will tend to suppress activity for related phonemes, such as /p/). The entire network of units is called the trace, because the pattern of activation left by a spoken input is a trace of the analysis of the input at each of the three processing levels. The positive feedback paths between the phonemic level and the lexical level can be used to explain our perceptual bias towards hearing phonetic patterns that make up real words and the fact that we can sometimes make early decisions about lexical candidates even though the entire word has not yet been heard. A simulation was implemented, and McClelland and Elman have argued that it showed characteristics which are appropriate for what is known about human speech perception. Although there are attractive aspects of the TRACE model, there are also some quite serious problems: as with any feature-based approach, there is dependence on being able to extract appropriate features from the acoustic signal. However, the main problem with the model is that, like neural network approaches to ASR, it does not provide any real modelling of temporal variability in speech.

In summary, there appear to be no engineering demonstrations of substantially complete models of human speech perception, presumably because there are still some very fundamental gaps in our

understanding of perception. In view of this fact, the performance of current state-of-the-art speech recognition systems is really quite impressive.

5. Advantages of HMMs as an approach to speech recognition

This section will discuss the reasons for the success of the HMM approach to ASR in more detail, by considering the issue from three perspectives:

- **Architecture:** Basic characteristics of the mathematical framework that are useful for speech recognition.
- **Completeness:** Advantages of the underlying approach over specific knowledge-based approaches.
- **Flexibility:** Ways in which speech knowledge can be incorporated into HMMs, in the form of constraints on the basic flexible structure.

5.1. Architecture

The HMM methodology provides a tractable mathematical structure that can be examined and studied analytically; it includes the provision of straightforward algorithms for training and recognition. In the model, the short-time spectral characteristics (associated with the individual states) and the temporal relationship among the processes (the Markov chain) are treated as separate aspects of a single dynamic process (i.e. speech) using one consistent framework. The combination of these two aspects is such that the calculation of the probability of a given set of observations being produced by a particular model can be decomposed simply into a summation of the joint probability of the observations and the state sequence, either taken over all state sequences (total likelihood) or just the most likely state sequence (maximum likelihood). Therefore, when performing recognition, the segmentation of an utterance arises automatically as a simultaneous part of the recognition process. Similarly, a set of models can be trained given a set of labelled, but not necessarily pre-segmented, speech data.

The use of probabilities to express the output distributions of the models also has the desirable property of allowing the models to easily generalize to unseen data. This can be accomplished either by smoothing estimated discrete distributions or by parameterizing the observations with a model such as a multi-variate Gaussian. The choice of observation distributions also extends to the case of the HMM itself: sequences of sub-word models can form the "states" of word models, which can in turn act as the "states" in syntax models, which can even be part of a higher level which uses semantics to provide dialogue control. Thus, the HMM framework allows for an integrated recognition system whereby the best match is found at all levels simultaneously, taking into account constraints imposed at each level. As Levinson (1985b) pointed out, the use of constraints in this way has the great advantage of increasing redundancy and so improving reliability of classification. The integration of all the levels also has the property of delayed decision making which, as pointed out by Klatt (1989) in relation to human speech perception, is very desirable as it eliminates problems which might arise if the input to a level contains errors which have been introduced at a separate lower level. By making only one decision, all the information is available to all levels of modelling.

5.2. Completeness

The comparatively poor recognition performance shown by knowledge-based ASR systems is generally agreed to be due to the difficulties involved in specifying all the required knowledge for

performing speech recognition in an appropriate way, combined with the fact that this knowledge is still limited and may even sometimes be erroneous. These observations led Makhoul and Schwartz (1986) to point out the advantages of "ignorance modelling", whereby the parameters of a model are automatically trained using a large amount of training data but very little speech knowledge, apart from a general structure which integrates the modelling of both spectral and temporal variation. The training process determines the important model characteristics based on optimizing the models for the given data. Zue (1985) has argued for the incorporation of speech knowledge to improve standard recognition algorithms, while also pointing out that, where such knowledge is limited, "sophisticated ignorance models can make optimal use of whatever knowledge we do have".

Levinson (1985a) expresses similar ideas from a rather different perspective. He describes the statistical approach as being a "macrotheory" of speech, in contrast to the knowledge-based approach which can be regarded as a "microtheory". Thus, the statistical approach imposes no particular structure but provides sufficient degrees of freedom to acquire the details by optimisation to match training data, whereas the knowledge-based approach attempts to build a general model of speech by listing every important aspect in detail. In contrast, Levinson regards the general model framework provided by HMMs as being a powerful theory of speech. However, he does point out that specialized knowledge about speech can and should be used to constrain the models by limiting the number of degrees of freedom.

5.3. Flexibility

All successful HMM-based recognition systems use knowledge about speech to some extent. For example, it is usual to choose a method of acoustic feature analysis which enhances phonetically important spectral characteristics while being relatively insensitive to spectral differences which do not reflect differences in meaning, such as pitch and overall loudness. It is also common to use a perceptually-motivated frequency scale, such as the Mel scale, which uses a finer frequency resolution for low frequencies than high frequencies. In addition to the static spectral feature vectors, most systems also use some form of dynamic feature vectors, which are typically calculated either by determining the rate of change of the static vectors over a region of about five frames, or by simple differencing between the values for two frames. All the above characteristics of the acoustic feature analysis involve the use of speech knowledge in order that the models are based on the most useful and relevant information from the speech signal.

Speech knowledge is also used when specifying the structure of the models themselves, both in the inventory of units and the model topology: there are many systems which use sub-word units based on some form of context-dependent phone models to incorporate coarticulation effects (e.g. Lee, 1988), and some systems include explicit modelling of known important allophonic effects (e.g. Holmes, Wood and Pearce, 1993). Such systems are applying phonological knowledge to make optimum use of the data, by using the same model for a sound occurring in different words if the immediate context is considered equivalent. Structuring models in this way can actually be advantageous compared with whole-word models, as only the phonetically different parts of words will be represented by different models.

The fact that the HMM is an inherently time-sequential generative model means that the changing speech signal can easily be related to a progression through the states. It is usual to only allow a subset of the possible transitions between the states, by using fairly simple left-to-right models which include self-loop transitions and possibly some skip transitions to allow for time scale variability. Furthermore, it is possible to tailor the structure of the model for each individual sound: the model topology and the number of states in each model can be allocated according to the typical duration, spectral complexity and variability of the sound being modelled (Holmes, 1991). All the above ways of using HMMs are constraining the basic flexible model to reflect known characteristics of speech

signals.

6. Limitations of the HMM formalism and possible improvements

In the previous section it has been argued that, in addition to their practical advantages in providing an easy-to-implement, sound mathematical framework for building ASR systems, HMMs are actually useful as a basis for speech modelling. They provide a flexible framework, which can be constrained by knowledge about the nature of the speech they are being used to model, but with the actual model parameters learned from training data. This type of approach avoids the inevitable problems which arise when trying to actually specify knowledge about speech in a way which is useful for a speech recognizer. In terms of general principle, self-organizing approaches to ASR are comparable with the human language acquisition process: we may be predisposed towards recognizing speech-like sounds, but the actual learning is dependent on hearing spoken language. This learning process results in "knowledge" about speech which is implicit, as it is with trained statistical models.

Although there are many arguments in favour of the HMM approach and HMMs have provided the basis for the most successful ASR systems, the performance is still not as good as we would like on complex tasks, such as those involving a large vocabulary or continuous spontaneous speech. This leads to the question of how performance can be improved. Some improvements will undoubtedly arise with further advancements in the areas of language and dialogue modelling. There are, however, also aspects of acoustic-phonetic modelling which can be improved. With standard HMMs, there are undeniably some aspects of the models themselves which are actually at variance with what we do know about speech. In particular, the following three assumptions which are made by the HMM formalism are clearly inappropriate for modelling speech patterns:

- **Piece-wise stationarity**

The HMM framework assumes that a speech pattern is produced by a piece-wise stationary process, with instantaneous transitions between stationary states. This is in direct contradiction with the fact that speech patterns are derived from signals produced by a continuously moving physical system - the vocal tract.

- **The independence assumption**

It is assumed that the probability that a given acoustic vector corresponds to a given state of the HMM depends only on the vector and the state, and is independent of the sequence of acoustic vectors preceding and succeeding the current vector and state. Thus the model takes no account of the dynamic constraints of the physical system which has generated a particular sequence of acoustic data, except inasmuch as these can be incorporated in the feature vector associated with a state.

- **State duration distribution**

A consequence of the above assumption is that the probability of a model staying in the same state for several frames is determined only by the self-loop transition probability. Thus the state duration in an HMM conforms to a geometric pdf which assigns maximum probability to state duration 1 and successively smaller probabilities to longer durations.

The impact of these inappropriate assumptions can be reduced by, for example, using a fairly generous allocation of states which allows a sequence of piece-wise stationary segments to better approximate the dynamics and also makes a duration of one frame per state more appropriate. The improved recognition performance which has been demonstrated from using variable frame-rate

(VFR) analysis (Peeling and Ponting, 1991) can also be regarded as a way of reducing the impact of the erroneous HMM assumptions. VFR analysis is used to replace sequences of similar vectors by a representative single vector. Vectors in the relatively stationary regions of speech patterns are highly correlated, which is contrary to the independence assumption. Discarding vectors in these regions results in observation sequences which are more consistent with the formalism.

Rather than trying to modify the data to fit the model, it should be better to make the model more appropriate for speech signals. This really requires some form of model which incorporates the concept of modelling segments of speech, rather than individual frames. Artificial Neural Networks such as the MLP provide a potential solution to this problem: unlike HMMs, they do not need to treat features as independent, as they can incorporate multiple constraints and find optimal combinations of constraints for classification. However, as explained in Section 3, they have difficulty in modelling the time-sequential nature of speech and have therefore not been very successful at recognizing connected utterances. To avoid this problem, most of the recent work using neural networks has focused on using hybrid systems which exploit the advantages of connectionist models while preserving the HMM formalism to integrate over time and to segment continuous speech (Boulevard, Morgan and Renals, 1992). One approach has been to use MLPs to compute HMM emission probabilities (Morgan and Boulevard, 1990) with better discriminant properties and without any hypotheses about the statistical distribution of the data. An alternative approach (Austin, Zavaliagkos, Makhoul and Schwartz, 1992) is to use a neural network as a post-processing stage to an N-best HMM system.

Although hybrid HMM/neural net approaches do allow for incorporation of segmental aspects of speech, a more elegant solution is to extend the HMM itself to be segment-based, using a model which actually characterizes the dynamic behaviour of the speech signal over a segment of speech representing some basic speech unit. A model of this type would effectively impose further speech-like constraints on the allowed parameters and should actually require less parameters than standard HMMs or combined HMM/MLP systems.

7. Conclusions

This paper has focused on the advantages of HMMs as a general approach for speech recognition, in providing a robust mathematical framework which is able to characterize many of the properties of speech signals in an integrated manner. By retaining all the good aspects of this approach but improving the general framework to make it more appropriate for the continuous segmental nature of speech, further performance improvements should be possible. A segment-based framework has been developed by Russell (1993), and work is currently in progress to investigate and further develop this new model.

8. References

- Allerhand, M. (1987) *Knowledge-Based Speech Pattern Recognition*. Kogan Page, London.
- Austin, S., Zavaliagkos, G., Makhoul, J. and Schwartz, R. (1992) Continuous Speech Recognition Using Segmental Neural Nets. *Proc. IEEE ICASSP*, San Francisco, 625-628.
- Baker, J.M. (1991) Large vocabulary speaker-adaptive continuous speech recognition research overview at Dragon Systems. *Proc. Eurospeech-91*, Genova, 29-32.

Bourlard, H., Kamp, Y., Ney, H. and Wellekens, C.J. (1985) Speaker-dependent connected speech recognition via dynamic programming and statistical methods. In M.R. Schroeder (Ed.) *Speech and Speaker Recognition*, Karger, Basel.

Bourlard, H., Morgan, N. and Renals, S. (1992) Neural nets and hidden Markov models: Review and generalizations, *Speech Communication*, 11, 237-246.

Bridle, J.S., Brown, M.D. and Chamberlain, R.M. (1983) Continuous connected word recognition using whole word templates. *The Radio and Electronic Engineer*, 53, 167-175.

Gales, M.J.F. and Young, S.J. (1993) Segmental hidden Markov models. *Proc. Eurospeech-93*, Berlin, 1579-1582.

Holmes, J.N. (1991) Use of phonetic knowledge when designing and training stochastic models for speech recognition. *Proc. Eurospeech-91*, Genova, 1257-1260.

Holmes, W.J., Wood, L.C. and Pearce, D.J.B. (1993) Allophone modeling for vocabulary-independent HMM recognition. *Proc. IEEE ICASSP*, Minneapolis, 487-490.

Juang, B.H. and Rabiner, L.R. (1991) Hidden Markov Models for Speech Recognition. *Technometrics*, 33, 251-272.

Klatt, D.H. (1977) Review of the ARPA Speech Understanding Project. *J. Acoust. Soc. Am.*, 62, 1345-1366.

Klatt, D.H. (1979) Speech perception: A model of acoustic-phonetic analysis and lexical access. *J. Phonetics*, 7, 279-312.

Klatt, D.H. (1989) Review of selected models of speech perception. In W. Marslen-Wilson (Ed.) *Lexical Representation and Process*, MIT Press, Cambridge, MA.

Lee, K.F. (1988) Large vocabulary speaker independent continuous speech recognition, *Ph.D. Thesis*, Carnegie-Mellon University.

Levinson, S.E. (1985a) Structural Methods in Automatic Speech Recognition. *Proc. IEEE*, 73, 1625-1650.

Levinson, S.E. (1985b) A unified theory of composite pattern analysis for automatic speech recognition. In F. Fallside and W.A. Woods (Eds.) *Computer Speech Processing*, Prentice-Hall International, London.

Makhoul, J. and Schwartz, R. (1984) Ignorance Modeling. In J.S. Perkell and D.H. Klatt (Eds.) *Invariance and Variability in Speech Processes*, Lawrence Erlbaum Associates, Hillsdale, N.J.

Marcus, M., Ed. (1992) *Proc. of the Fifth DARPA Speech and Natural Language Workshop*, San Mateo.

Mariani, J. (1991) Knowledge-Based Approaches Versus Mathematical Model Based Algorithms: the Case of Speech Recognition. *Proc. 30th Conference on Decision and Control*, Brighton, 841-846.

McClelland, J.L. and Elman, J.L. (1986) The TRACE Model of Speech Perception. *Cognitive Psychology*, 18, 1-86.

Morgan, N. and Bourlard, H. (1990) Continuous speech recognition using multilayer perceptrons with hidden Markov models, *Proc. IEEE ICASSP*, Albuquerque, 413-416.

O'Brien, S.M. (1993) Knowledge-based systems in speech recognition: a survey. *Int. J. Man-Machine Studies*, 38, 71-95.

Peeling, S.M. and Moore, R.K. (1988) Isolated digit recognition experiments using the multi-layer perceptron, *Speech Communication*, 7, 403-409.

Peeling, S.M. and Ponting, K.M. (1991) Variable frame rate analysis in the ARM continuous speech recognition system, *Speech Communication*, 10, 155-162.

Pisoni, D.B. and Luce, P.A. (1987) Acoustic-phonetic representations in word recognition, *Cognition*, 25, 21-52.

Russell, M.J. (1993) A segmental HMM for speech pattern modelling, *Proc. IEEE ICASSP*, Minneapolis, 499-502.

Zue, V.W. (1985) The Use of Speech Knowledge in Automatic Speech Recognition. *Proc. IEEE*, 73, 1602-1615.