# Plasticity in Systems for Automatic Speech Recognition: A Review

*Roger K. Moore[$] and Stuart P. Cunningham[#]*

[$]Dept. Computer Science
[#]Dept. Human Communications Sciences
University of Sheffield, United Kingdom
r.k.moore@dcs.shef.ac.uk, s.cunningham@sheffield.ac.uk

## Abstract

Although the topic 'plasticity in speech perception' is primarily concerned with the malleability of *human* speech perceptual behaviour, it may be illuminating to consider in parallel the degree to which current state-of-the-art '*automatic* speech recognition' (ASR) systems also change their behaviour over time. This paper provides a review of the computational mechanisms underlying contemporary ASR systems with a particular focus on their adaptive and learning behaviour. It is shown how such systems can change dynamically in order to accommodate new speakers, handle unexpected user behaviour and track and compensate for constantly varying acoustic environments.

## 1. Introduction

Recent years have seen a substantial growth in the deployment of practical systems for 'automatic speech recognition' (ASR). These ongoing commercial successes are a direct result of a significant increase in the capabilities of ASR devices over the past thirty years driven by both improvements in the underlying ASR algorithms and the relentless increase in available computer power. In particular, the introduction of 'dynamic programming' (DP) based search techniques in the 1970s, and statistical modelling using sub-word based 'hidden Markov models' (HMMs) in the 1980s, enabled the construction of 'large vocabulary continuous speech recognition' (LVCSR) systems capable of transcribing speech with a level of recognition accuracy that can be useful in a range of practical applications [1][2][3].

Of course many commercial speech applications involve more than just automatic speech recognition. For example, telephone-based 'interactive voice response' (IVR) systems make use of other components such as 'dialogue management' (DM) and perhaps 'text-to-speech synthesis' (TTS). Applications may also incorporate modules for semantic-level 'understanding' and/or 'generation'. All of these system components may exhibit plastic behaviour as they adapt to any given situation. However, this paper concentrates solely on plasticity in ASR.

## 2. Automatic Speech Recognition

The main components of a contemporary ASR system are illustrated in Figure 1. This structure may at first seem a little surprising to someone working on 'human speech recognition' (HSR), in that the architecture does not *explicitly* reflect a hierarchical organisation of processes based on a transformation from acoustic signal through sub-lexical to lexical representations. Indeed it is the failure of such explicit structures to recognise real speech in the 1970s that lead to the development of this much more successful arrangement [4]. Interestingly, this success has started to spawn interest from sections of the HSR community [5], and there are beginning to be attempts at a unification of the principles underlying both ASR and HSR [6][7][8].

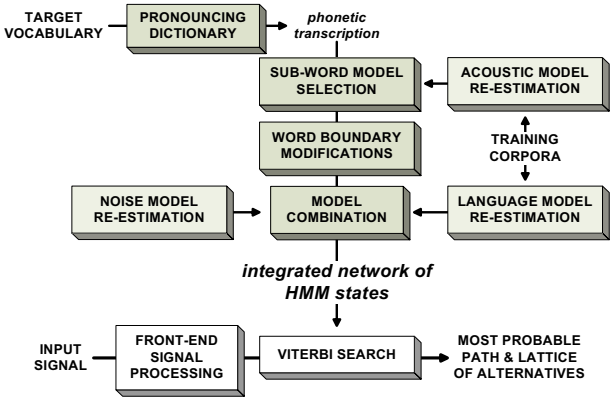


*Figure 1:* Block diagram of a contemporary automatic speech recognition system.

The basic principle underlying the architecture shown in Figure 1 is that all of the *knowledge* about how speech is produced is expressed *probabilistically* in an 'integrated network' of statistical models representing *all* possible utterances (within the constraints of the system's lexicon and syntax). The process of 'recognition' is then defined (mathematically) as a *search* through the network in order to find **the most probable explanation of the incoming signal**. The search process is conducted using an efficient algorithm known as 'dynamic programming' (DP or 'Viterbi' search), and it can be guaranteed to find the *best* explanation. The use of probability and statistics has been found to be a powerful method for *generalising* from seen to unseen data, and has been termed 'ignorance modelling' [9] (in contrast to the knowledge-based approaches of the 1970s).

The architecture shown in Figure 1 is essentially divided into two sections; the upper part represents the steps used to compile knowledge about speech into the integrated network of HMM states, and the lower (simpler) part represents the recognition process itself. The knowledge is derived both from a-priori structure and substantial corpora of training data (which could be hundreds or even thousands of hours of

recorded speech material [10] and millions of words of text [11]).

### 2.1. Components of an ASR system

In order to configure an ASR system, it is first necessary to obtain the largest possible corpora of speech and text data specific to the target application. From the text corpus it is possible to define the vocabulary and thence to train the 'language model'. The speech corpus is used to train the 'acoustic model' and the 'noise model'. A pronouncing dictionary is also required.

#### *2.1.1. Language modelling*

The purpose of the language model is to be able to assign a probability to any word sequence, and it may either be expressed as a probabilistic finite-state syntax (e.g. a regular or context-free grammar) or, for more flexible applications, as a set of statistical 'n-grams'. In the latter case, a typical value for 'n' may be three, in which case it would be referred to as a 'trigram' language model. However, since a finite corpus is unlikely to contain all n-grams, it is usual to estimate the probability of the missing items by 'backing-off' to lower-order n-grams (and ultimately to 'unigram' statistics). The language model can also be thought of as a mechanism for predicting the next word in a sequence.

#### *2.1.2. Acoustic modelling*

The purpose of the acoustic model is to be able to assign a probability to any sound sequence, and it is typically constructed from a set of context-dependent sub-word HMMs, where each HMM represents a particular phone in a particular preceding and following context. If the context is limited to one phone either side - a common arrangement - then the HMMs are referred to as 'triphones'. An example of a triphone might be an acoustic model for /t/ in the context of preceding /s/ and following /r/. Such HMMs would consist of several 'states' (often three) in order to capture the time evolution of the phone.

Word-level HMMs are effectively constructed for each lexical entry by selecting the appropriate sub-word HMMs from the available set. Of course, due to the finite amount of training data, it may not be possible to find a sub-word model with a matching context. In this case it is usual to 'back-off' to a shorter context version, and ultimately to 'context-independent phones' (monophones). This entire process allows models to be constructed for application words that were never spoken in the training data.

The acoustic model is either trained on speech from a known user, in which case it is termed 'speaker-dependent', or it is trained on speech from a wide variety of speakers in which case it is termed 'speaker-independent'.

#### *2.1.3. Noise modelling*

Real-world environments contain many different sound sources, not just the target speaker. Also the incoming speech may have been altered by reverberation due to room acoustics or distorted by being passed over a communications channel (such as a mobile telephone). A practical ASR system thus has to model these effects, for example by deriving an inventory of 'noise models' covering likely non-speech sounds (including lip smacks and other user-generated noises). These models can then be used as extra entries in the lexicon, or combined with the acoustic models using techniques such as 'HMM decomposition' [12] or 'parallel model combination' (PMC) [13]. The latter approaches allow for speech and noise to occur *at the same time*.

#### *2.1.4. Pronunciation modelling*

The purpose of the pronunciation model is to minimise recognition errors that occur due to pronunciation variation. Information about individual and accent variations can be encoded in a lexicon containing both the orthographic and phonetic transcriptions for all the words in a recogniser's vocabulary.

Perhaps the most important task in pronunciation modelling, however, is to determine the potential variants for all the vocabulary items so it is possible to distinguish between various pronunciations in the training data. Obtaining the information about variant pronunciations can be realised using either knowledge based approaches (i.e. pronunciation dictionaries) or data driven approaches (using either manual or automatic transcriptions to generate lists of variants). The use of data driven approaches means that information on variation can be incorporated into the system based on actual pronunciations occurring in the training data.

### 2.2. Training an ASR system

The process of 'training' an ASR system is essentially one of estimating the values of all of the parameters in the language model, acoustic model, noise model and pronunciation model [1]. In a contemporary system this process usually involves substantial quantities of speech and text data coupled with powerful machine learning algorithms. The high volume of data required is primarily due to the statistical nature of the models, i.e. it is necessary to have a sufficient amount of training data in order to get stable estimates of the relevant 'probability density functions' (pdfs). For example, the acoustic model in a typical state-of-the-art LVCSR system may contain 300,000 Gaussian distributions whose mean and a variance have to be derived from the training data.

Interestingly, the statistical nature of the ASR modelling/training process means that, in principle, *all* parameters in a system are capable of being 're-estimated' (i.e. adapted), and this play an important role in the opportunities for plasticity of ASR systems.

### 2.3. Recognition

The process of 'recognition' involves some form of front-end signal processing followed by the Viterbi/DP search.

#### *2.3.1. Font-end signal processing*

Incoming speech is usually transformed into a *cepstral* representation derived on a Mel-frequency scale. The reason for doing this is to derive a independent representation that fits with the assumptions in the hidden Markov models. The resulting data, which may be framed every 10-30 msecs is referred to as 'mel-frequency cepstral coefficients' (MFCCs).

#### *2.3.2. Viterbi search*

The Viterbi search involves finding the route through the integrated network of HMM states that corresponds most

closely to the sequence of incoming speech vectors (where 'closeness' is measured using probability). The resulting most-probable path reveals the sequence of words (and phones) that are most likely to have been spoken, and these are output by the recogniser (together with an n-best list or lattice of alternatives).

## 3. Challenges Facing Contemporary ASR

Outside the controlled environment of the laboratory, a practical ASR system will encounter many situations that can impact on recognition accuracy. Difficulties can arise not only from the prevailing acoustic environment, but also from the characteristics of the users and their potentially unexpected behaviour. In recent years many techniques have been engineered to equip ASR systems with a greater degree of robustness to these challenging situations, often through the introduction of flexible and adaptive (*plastic*) behaviour.

One very practical challenge is that many everyday environments contain additive and convolution noise sources which alter the speech signal transmitted from a speaker to the ASR system. Additive noises such as competing speakers or machinery may mask portions of the input speech signal, while convolution noise induced by the frequency response of the channel or reverberation may alter the spectro-temporal distribution of energy in the signal. Despite the use of noise models, it unlikely that a corpus of training data will capture the wide range of naturally occurring variation in typical acoustic environments.

It is also the case that for speaker independent systems the training data can only represent a sample of potential users of an ASR system. As a consequence, speakers who are not well represented may cause difficulties due to the characteristics of their speech, their differently sized vocal tracts and differences in accent, dialect and speaking style.

Not only can a particular speaker pose a challenge to an ASR system due to the individual characteristics of their speech but the nature of their interaction may also be problematic. For example, a user might well deliver an utterance that contains a word not included in the system's lexicon. Such 'out of vocabulary' (OOV) words are likely to be misrecognised as a word from the within the system's vocabulary unless they can be detected and handled appropriately.

## 4. Plasticity in ASR

The very structure embodied in Figure 1 – the compilation of all prior knowledge into an integrated network – whilst being powerful and effective does encourage a somewhat static perspective of ASR. Any changes require either (i) re-compilation of the network of models, (ii) adaptation of the model parameters to accommodate new situations or (iii) modification of the input representation to better fit the existing model parameters.

### 4.1. Algorithms for learning and adaptation

As has been seen in previous sections, ASR (like any pattern recognition system) is initially *trained* in order to assign a given input to one of a number of target classes [14]. The training data is used to *learn* the parameters of a statistical pattern classifier, which is subsequently used to classify a sequence of input speech vectors into a sequence of words.

Such training, based on the use of labelled training data, is known as *supervised* learning, and several techniques have proved to be popular and found wide application in ASR [15], e.g. 'Maximum Likelihood' (ML), 'Expectation-Maximization' (EM), 'Maximum a-posteriori Probability' (MAP) [16] and 'Maximum Mutual Information' (MMI). Whatever the algorithm, the essential aim is to maximise the probability of the training data being generated by the models.

A second training method, known as *unsupervised* learning, uses data to estimate the underlying distributions without knowledge of the actual identities. The outcome of unsupervised learning is more difficult to control than supervised learning, but the lack of labelled data more closely reflects many practical situations.

Both supervised and unsupervised techniques have been deployed in ASR systems in order to learn new relationships or to adapt to changing circumstances. However, unsupervised adaptation leads to the most obvious plasticity.

### 4.2. Acoustic model adaptation

A number of mathematical algorithms have been investigated for adapting the parameters of the acoustic model in order to accommodate new speakers. Techniques such as 'partial re-estimation', 'maximum likelihood linear regression' (MLLR) [17], 'vocal tract length normalisation' (VTLN) [18] and 'Eigen-voices' [19] have all had some degree of success for speaker-adaptive training.

The basic issue is how to change the multitude of parameters in the system using only a small number of new observations, and all of these techniques tackle this by effectively deriving a low-dimensional *sub-space* that is able to link all the different parameters together and move them towards the new data in a coordinated manner. MLLR has proved to be particularly successful and has been used to adapt to new speakers as well as new acoustic environments.

### 4.3. Noise/environment compensation

Modern ASR systems adapt to additive and convolution noise using techniques such as 'spectral subtraction' (SS) [20] and 'cepstral mean normalisation' (CMN) [21][22]. For example, using a continuously varying estimate of background, SS can be used to remove the energy estimated to be from the background in each input speech frame. Similarly, CMN subtracts a continuously updated estimate of the cepstral mean from each frame to minimise the effects of the communication channel.

Other methods, such as 'relative spectral process' (RASTA) can also be used to counteract the effects of convolutional noise by emphasising energy which has the temporal characteristics of speech [23]. Recent variations of RASTA have employed the use of several different temporal filters which are chosen to best suit the environment [24].

### 4.4. Language model adaptation

Adaptation in language modelling usually involves attempting to integrate a large *background* text corpus that generally relates to the task in hand with a significantly smaller *adaptation* corpus that is much more relevant [25]. Combining the information is posed as a statistical problem, and solutions include methods for 'model interpolation' (i.e.

merging the two sources of information in a balanced way) and more powerful approaches for 'constraint specification' (based on the matching of extracted *features*). Other approaches use *topic* information, *semantic* knowledge or *syntactic* structure to extract relevant data from the adaptation corpus.

Most of these methods are highly mathematical in nature, but of particular interest are techniques such as 'dynamic cache' in which dynamic shifts in word usage are captured by means of a short-term memory that overrides the probabilities being generated by the long-term background model, and 'trigger models' in which dependencies that have a longer-range than the base n-gram are detected and exploited.

### 4.5. Pronunciation model adaptation

Adaptation for pronunciation variation can be incorporated into the lexicon, the acoustic model and even the language model. For a lexicon containing canonical phonetic transcriptions of the vocabulary items, it is possible to re-examine the range of possible pronunciation variations that occur in a training set by obtaining phone-level transcriptions of the data. These automatically derived transcriptions can then be used to generate lists of pronunciation variants that can be incorporated into an updated lexicon [26].

## 5. Summary and Conclusion

This paper has provided a review of the computational mechanisms underlying contemporary ASR systems with a particular focus on their adaptive and learning behaviour. It has been shown how ASR systems can change dynamically in order to accommodate new speakers, handle unexpected user behaviour and track and compensate for a constantly varying acoustic environment.

## 6. Acknowledgement

This work was funded by UK Department of Health 'New and Emerging Applications of Technology' programme. The views expressed in this publication are those of the authors and not necessarily those of the Department of Health.

## 7. References

[1] Rabiner, L. & Juang, B-H., *Fundamentals of Speech Recognition*, Prentice Hall, 1993.

[2] Holmes, J. & Holmes, W., *Speech Recognition and Synthesis*, Taylor & Francis, 2002.

[3] Furui, S. *Digital Speech Processing, Synthesis and Recognition*, Marcel Dekker, Inc., 2001.

[4] Klatt, D. H., "Review of the ARPA speech understanding project", *J. Acoustical Soc. of America*, Vol. 62, pp. 1345-1366, 1977.

[5] Norris, D., "Shortlist: a connectionist model of continuous speech recognition", *Cognition*, Vol. 52, pp. 189-234, 1994.

[6] Moore, R. K. & Cutler, A., "Constraints on theories of human vs. machine recognition of speech", *Proc. Speech Recognition as Pattern Classification workshop*, Nijmegen, pp. 145-150, 2001.

[7] Scharenborg, O., ten Bosch, L., Boves, L. & Norris, D., "Bridging automatic speech recognition and psycholinguistics: extending Shortlist to an end-to-end model of human speech recognition", *J. Acoustical Soc. of America*, Vol. 114(6), pp. 3023-3035, 2003.

[8] Scharenborg, O., ten Bosch, L. & Boves, L., "Recognising 'real-life' speech with SpeM: a speech-based computational model of human speech recognition", *Proc. Eurospeech*, pp. 2097-2100, 2003.

[9] Makhoul, J. & Schwartz, R., "Ignorance modelling", *Invariance and Variability in Speech Processes*, Perkell, J. and Klatt, D. H. (eds.), Erlbaum, 1984.

[10] Everman, G., Chan, H. Y., Gales, M. J. F., Jia, B., Mrva, D., Woodland, P. & Yu K., "Training LVCSR systems on thousands of hours of data", *Proc. IEEE ICASSP*, 2005.

[11] Jelinek, F., *Statistical Methods for Speech Recognition*, MIT Press, 1998.

[12] Varga, A. P. & Moore, R. K., "Hidden Markov model decomposition of speech and noise", *Proc. IEEE ICASSP*, pp. 845-848, 1990.

[13] Gales, M. J. F. & Young, S. J., "Robust speech recognition in additive and convolutional noise using parallel model combination", *Computer Speech and Language*, Vol. 9, pp. 289-307, 1995.

[14] Bishop, C., *Neural Networks for Pattern Recognition*, Oxford University Press, 1995.

[15] Junqua, J-C. & Wellekens, C. (Eds.), "Special Issue on Adaptation Methods for Speech Recognition", *Speech Communication*, Vol. 42(1), 2004.

[16] Bahl, L., Jelinek, F., Raviv, J. & Raviv, F., "Optimal decoding of linear codes for minimising symbol error rate", *IEEE Trans. Information Theory*, Vol. 20, pp. 284-287, 1974.

[17] Leggetter, C. J. & Woodland, P., "Speaker adaptation of continuous density HMMs using linear regression", *Proc. ICSLP*, pp. 451-454, 1994.

[18] Lee, L. & Rose, R. C., "Speaker normalisation using efficient frequency warping procedures", *Proc. IEEE ICASSP*, 1996

[19] Kuhn, R., Junqua, J-C., Nguyen, P. & Niedzielski, N., "Rapid speaker adaptation in Eigenvoice space", *IEEE Trans. Speech & Audio Processing*, Vol. 8(6), pp. 695-707, 2000.

[20] Boll, S., "Suppression of acoustic noise in speech using spectral subtraction", *IEEE Trans. Audio, Speech and Signal Processing*, Vol. 27, pp. 113-120, 1979.

[21] Atal, B., "Effectiveness of linear prediction characteristics of the speech wave for automatic speech recognition", *J. Acoustical Soc. of America*, Vol. 55, pp. 1304-1312, 1974.

[22] Rosenberg, A., Lee, C.-H., & Soong, F., "Cepstral channel normalisation techniques for HMM-based speaker verification", *Proc. ICSLP*, pp. 1835-1838, 1994.

[23] Hermansky, H. & Morgan, N., "RASTA processing of Speech", *IEEE Trans. Speech and Acoustics*, Vol. 2, pp. 587-589, 1994.

[24] Shire, M., "Discriminant training of front-end and acoustic modeling stages to heterogeneous acoustic environments for multi-stream automatic speech recognition", *PhD Thesis, U. California, Berkeley*, 2000.

[25] Bellegarda, J. R., "Statistical language model adaptation: review and perspectives", Speech Communication, Vol. 42(1), pp. 93-108, 2004.

[26] Wester, M. & Fosler-Lussier, E., "A comparison of data-derived and knowledge-based modelling of pronunciation variation", *Proc. ICSLP*, pp. 270-273, 2000.