A Study of the Use and Evaluation of Confidence Measures in Automatic Speech Recognition

Gethin Williams
Technical Report CS-98-02
Department of Computer Science
University of Sheffield
G.Williams@dcs.shef.ac.uk

February 18, 1998

Abstract

Confidence measures have been found to be useful for a number tasks within the field of Automatic Speech Recognition (ASR). For example, the use of confidence measures has been reported in the utterance verification, keyword spotting and Out-of-Vocabulary (OOV) word spotting literature. In this report, it is shown that so called 'hybrid Artificial Neural Network/Hidden Markov Model' (HMM/ANN) systems are well suited to the task of generating confidence measures, due to their ability to provide local phone class posterior probability estimates which may be used to generate confidence measures in a computationally efficient manner. A number of evaluation metrics are also described and the performance of five confidence measures derived from the ABBOT hybrid HMM/ANN system for the tasks of utterance verification and OOV word spotting are evaluated using these metrics.

Besides the tasks described above, confidence measures may also be used for tasks such as filtering the acoustics for a number of conditions prior to decoding, guiding the search over the space of alternative decodings and also as diagnostic tools. The use of confidence measures as diagnostic tools can indicate how components of an ASR system can be improved. For example, confidence measures can be used to identify poor pronunciation models and also to guide the search for improved models, where the generation of accurate pronunciation models remains an important challenge in the creation of practical recognition systems.

Contents

1 Introduction **2**

1.1 The Statistical Modelling Approach to Automatic Speech Recognition . . . 2

1.2 Building a Speech Recognition System . . . 4

1.2.1 Generative vs. Acceptor Models . . . 5

1.3 Overview . . . 8

2 Statistical Hypothesis Testing **10**

2.1 Classical Hypothesis Testing . . . 10

2.2 Evaluation of Hypothesis Tests . . . 11

2.2.1 Evaluation Metrics based on a $\mathbf{2} \times \mathbf{2}$ Confusion Matrix . . . 11

2.2.2 Probability Distribution Based Evaluation Metrics . . . 16

3 Confidence Measures **17**

3.1 Introduction . . . 17

3.2 Potential Uses of Confidence Measures . . . 18

3.3 Literature Review . . . 19

3.3.1 Filtering . . . 20

3.3.2 Search . . . 20

3.3.3 Testing Hypotheses About Decodings . . . 20

3.3.4 Remarks . . . 23

3.4 The ABBOT system . . . 24

3.5 Five Confidence Measures . . . 25

4 Experiments **28**

4.1 Method . . . 28

4.2 Results . . . 29

5 Discussion **43**

5.1 Marking Schemes . . . 43

5.2 Normalised Unconditional Error Rate . . . 43

5.3 Providing Insight . . . 43

5.4 Summary . . . 45

Chapter 1

Introduction

1.1 The Statistical Modelling Approach to Automatic Speech Recognition

Currently, the most successful approach to Automatic Speech Recognition (ASR), is the application of statistical modelling techniques and it is this approach which will be considered throughout this report. The key properties of a statistical model are that it should have a number of adjustable parameters and that the values of these parameters should be inferred directly from a some observed data, in an attempt to model the underlying distribution of the data. An application of statistical modelling is the field of statistical decision theory, which has three fundamental concepts:

- A set of possible states of nature, $\Omega = \{\omega_1, \omega_2, \ldots, \omega_N\}$.
- Some observations of the world, possibly coupled with some prior knowledge.
- A set of possible actions to take, $\mathcal{A} = \{\alpha_1, \alpha_2, \ldots, \alpha_M\}$.

A branch of the wider area of statistical decision theory is the field of statistical pattern recognition. The goal of a pattern recognition process is to *classify* a given entity, that is to declare it as an example of some class, taken from a set of predefined classes. The set of possible states of nature which are relevant to a pattern recognition process are, therefore, the set of predefined classes to which an entity may belong, $\Omega = \{\mathcal{C}_1, \ldots, \mathcal{C}_k, \ldots, \mathcal{C}_K\}$.

The use of any prior knowledge in conjunction with some observations of an entity for its classification is dependent upon which of the two approaches to statistical inference is adopted. The *Frequencist* approach allows only sample data to be employed in an inferential process and so the Frequencist approach to statistical pattern recognition only uses the observations of an entity are to classify it. The *Bayesian* approach to statistical inference, on the other hand, allows prior knowledge to be used in conjunction with any observations in order to make an inference. It should be noted that this prior knowledge may be either *subjective*, i.e. based on some intuition, or *objective*, i.e. based on some previously collected data. In the absence of any observations, the Bayesian approach allows an entity to be classified using prior knowledge alone. The observations of an entity may be represented using a vector valued random variable known as either a *feature vector*, an *observation vector* or a *pattern*, $\mathbf{X} = \{x_1, x_2, \ldots, x_d\}$. A consequence of using a feature vector is that the observations of a number of entities may be compared by plotting the vectors in *feature space*.

The minimal set of actions for a pattern recognition process comprises the classification of a given entity as an example of each of the predefined classes, $\mathcal{A} = \{\text{classify}(\mathcal{C}_1), \ldots, \text{classify}(\mathcal{C}_k), \ldots, \text{classify}(\mathcal{C}_K)\}$. The decision taken is based upon the values of a set of functions of the observed data, possibly in conjunction with some prior knowledge, called *discriminant functions*. These functions have a number of adjustable parameters, Θ, the values of which are determined through a process of statistical inference from some data sample. The number of discriminant functions is equal to the number of classes, $\mathcal{F} = \{f_1, \ldots, f_k, \ldots, f_K\}$, and they may be used to discriminate between the different classes using the following decision rule:

$$\text{classify}(\mathcal{C}_k) \text{ if } f_k(\mathbf{X}; \Theta) > f_j(\mathbf{X}; \Theta) \;\forall\; k \neq j \;. \tag{1.1}$$

The set of discriminant functions define a partition of feature space, where the decision boundaries are described by:

$$f_k(\mathbf{X};\Theta) = f_j(\mathbf{X};\Theta) \ \forall \ j,k \ . \tag{1.2}$$

As there is a one-to-one correspondence between a class and a discriminant function, a discriminant function may be considered as a *class model*, for the purposes of classification. In this case, a 'strongly matching' class model is a discriminant function which produces a relatively large value given the data, whereas a 'weakly matching' model is a function which produces a relatively small value.

It should also be noted that different decisions carry different consequences. The relative weightings of such consequences may be encoded by a *loss function*, $L_{k,j}$. Only 'zero-one' loss functions will be considered in this report, however, where a zero loss is assigned to all correct decisions whereas a loss of unity is assigned to all incorrect decisions.

$$L_{k,j} = \begin{cases} 0 & \rightarrow \mathbf{X} \in \mathcal{C}_k, \text{classify}(\mathcal{C}_k) \\ 1 & \rightarrow \text{otherwise.} \end{cases}$$

In order to recognise an utterance, the task is to classify examples of speech sounds, where an utterance is typically composed of a sequence of words. Due to the infinite number of possible word combinations, it is clearly infeasible to build a model of every possible word sequence for its recognition. An alternative is to construct a model for an utterance by concatenating models of individual words. This approach, however, is only feasible for small vocabulary recognition tasks due to the computational resources required to build individual word models. In order to reduce the resource requirements for large vocabulary tasks, the hierarchical nature of speech may be further exploited, where it is noted that words themselves are composed of a concatenation of subword speech sounds. For large vocabulary tasks it is common practice, therefore, to define a set of word class models in terms of a shared set of subword class models. In this way the parameters of the subword class models are shared by the models higher up the hierarchy and the computational resource requirements are correspondingly reduced. This hierarchical approach to model building is also efficient in the use of training data as the parameters for models of rarely seen words in the training data are trained using data drawn from instances of more frequent words. A number categories of subword models may be defined including context independent phone units, context dependent phone units and also syllable models. It should be noted that this approach assumes that words are created by articulating a sequence of distinct speech sounds and the method in which such speech sounds are articulated is not explicitly addressed in this case.

An utterance results in an acoustic waveform in which variations in frequency and amplitude *encode* the words which were spoken. The task of a speech recogniser is, therefore, to *decode* the acoustic signal back into the words which gave rise to it through a process of classifying portions of waveform. The first stage of the ASR process is to digitise the acoustic signal, typically at a rate of the order of 16KHz. Before classification is attempted, a *feature extraction* or *pre-processing* stage is carried out. There are two criteria for a useful pre-processing stage. The first is to reduce the data rate, again for computational resource reasons, and the second is to highlight the information in the acoustic signal which is useful for classification. It is important, however, not to reduce the data rate beyond that at which important events in the speech signal occur, as this will clearly degrade classification performance. A common form of pre-processing is to perform a short time spectral analysis of the digitised signal using a sliding window of samples called a *frame*, typically producing a feature vector for a time period of the order of every 16ms of the signal.

Examples of a given class of speech sound exhibit variation due to a number of factors. These include inter-speaker variations such as dialect regions and physiological differences between speakers, intra-speaker variations such as the mood of a speaker or whether the speaker has a cold and also extra-speaker variations, such as differences in the channel across which a speech signal is transmitted. This variation causes examples of a class of speech sound to form a *cluster* of points in feature space, where the region occupied by one cluster may overlap with the region occupied by another. If any such regions do overlap, a completely error free classification is not possible. In order to accommodate such variation, probabilistic class models are often adopted, where class membership is modelled as a probability distribution. Bayes' Decision Rule may be used for classification if the posterior probability of a class given the observations, $P(\mathcal{C}_k|\mathbf{X})$, is known:

$$\text{classify}(\mathcal{C}_k) \text{ if } P(\mathcal{C}_k|\mathbf{X}) > P(\mathcal{C}_j|\mathbf{X}) \; \forall \; k \neq j \; . \tag{1.3}$$

As it has been stated above, the probability of misclassification will not be zero if the posterior probability distributions for the classes overlap. A schematic representation of two overlapping class posterior probability distributions is given in figure 1.1.

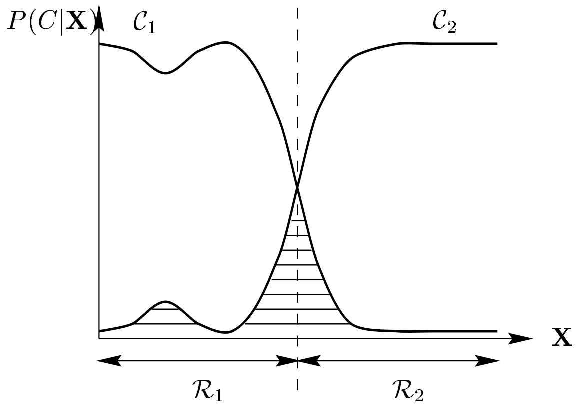


Figure 1.1: A schematic representation of two overlapping class posterior probability distributions.

The probability of error in this case is proportional to the area of the hatched region in figure 1.1 and can be calculated using equation 1.4 [4]:

$$\begin{aligned} P(\text{error}) &= P(\mathbf{X} \in \mathcal{R}_2, \mathcal{C}_1) + P(\mathbf{X} \in \mathcal{R}_1, \mathcal{C}_2) \\ &= P(\mathbf{X} \in \mathcal{R}_2|\mathcal{C}_1)P(\mathcal{C}_1) + P(\mathbf{X} \in \mathcal{R}_1|\mathcal{C}_2)P(\mathcal{C}_2) \\ &= \int_{\mathcal{R}_2} p(\mathbf{X}|\mathcal{C}_1)P(\mathcal{C}_1)\, d\mathbf{X} + \int_{\mathcal{R}_1} p(\mathbf{X}|\mathcal{C}_2)P(\mathcal{C}_2)\, d\mathbf{X} \\ &\propto \int_{\mathcal{R}_2} P(\mathcal{C}_1|\mathbf{X})\, d\mathbf{X} + \int_{\mathcal{R}_1} P(\mathcal{C}_2|\mathbf{X})\, d\mathbf{X} \; , \end{aligned} \tag{1.4}$$

where the area of this hatched region, and hence the probability of misclassification, is minimised by placing the feature space partition at the intersection of the two distributions, which is equivalent to assigning a pattern to the class with the highest posterior probability.

A number of excellent text books, including [13, 4, 26], cover the field of statistical decision theory and statistical pattern recognition in particular in much greater detail than is given above.

1.2 Building a Speech Recognition System

From the foregoing discussion, it can be seen that when attempting to recognise an utterance, it is desirable to assign it to the class with the highest posterior probability given the acoustics:

$$\mathcal{H}^* = \underset{\mathcal{H}}{\operatorname{argmax}}\, P(\mathcal{H}|X_1^N) \; , \tag{1.5}$$

where the acoustics of the utterance are represented as a sequence of acoustic observation vectors, $X_1^N = \{\mathbf{X}^1, \ldots, \mathbf{X}^n, \ldots, \mathbf{X}^N\}$ and the class with the highest posterior probability, $\mathcal{H}^*$, is often termed the most probable *decoding hypothesis*. The use of this term emphasises the fact that a recogniser can only *hypothesise* about the content of an utterance. The term may be applied to the recogniser's output given an entire utterance or to the output corresponding to some portion of an utterance which may constitute a word or subword speech sound.

It is common practice to combine two distinct components to create a speech recogniser, namely an *acoustic model* and a *language model*, where it can be shown that this 'divide-and-conquer' approach to building a recogniser is motivated by Bayes' Theorem:

$$\begin{aligned} P(\mathcal{H}|X_1^N) &= \frac{p(X_1^N,\mathcal{H})}{\sum_{j=1}^{J} p(X_1^N,\mathcal{H}_j)} \\ &= \frac{p(X_1^N|\mathcal{H})P(\mathcal{H})}{\sum_{j=1}^{J} p(X_1^N|\mathcal{H}_j)P(\mathcal{H}_j)} \\ &= \frac{p(X_1^N|\mathcal{H})P(\mathcal{H})}{p(X_1^N)}\ , \end{aligned} \tag{1.6}$$

where the summation used to calculate the denominator is over all possible decoding hypotheses, J.

In the case of so called '*traditional' Hidden Markov Model*[1] (HMM) based recognisers, the language model provides an estimate of the a priori probability of a given word sequence, $P(\mathcal{H})$, whereas, the acoustic model provides an estimate of the likelihood of the acoustics given the particular word sequence model, $p(X_1^N|\mathcal{H})$. These two quantities may then be multiplied to obtain the joint probability of the acoustics and the word sequence model, $p(X_1^N,\mathcal{H})$. As the probability of the acoustics themselves, $p(X_1^N)$, is constant for all possible word sequence decodings, J, this joint probability is proportional to posterior probability of the word sequence model given the acoustics:

$$\mathcal{H}^* = \underset{\mathcal{H}}{\operatorname{argmax}}\, P(\mathcal{H}|X_1^N) \propto \underset{\mathcal{H}}{\operatorname{argmax}}\, p(X_1^N,\mathcal{H})\ . \tag{1.7}$$

As it will be seen later in the context of confidence measures, it is important to note that both $p(X_1^N,\mathcal{H})$ and $p(X_1^N|\mathcal{H})$ are relative to $p(X_1^N)$, and so are not comparable across utterances. Traditional HMM based recognisers can be contrasted, however, with so called '*hybrid Hidden Markov Model/Artificial Neural Network*' (HMM/ANN) based recognisers, which can be shown to be able to produce an estimate of the likelihood of the acoustics given the word sequence model, scaled by the probability of the acoustics themselves, $p(X_1^N|\mathcal{H})/p(X_1^N)$, which can be combined with an estimate of $P(\mathcal{H})$ to produce an estimate for the full posterior probability of the word sequence model given the acoustics:

$$\mathcal{H}^* = \underset{\mathcal{H}}{\operatorname{argmax}}\, P(\mathcal{H}|X_1^N) = \underset{\mathcal{H}}{\operatorname{argmax}}\, \frac{P(X_1^N|\mathcal{H})}{P(X_1^N)} P(\mathcal{H})\ . \tag{1.8}$$

When using either a traditional HMM or a hybrid HMM/ANN based recogniser, the acoustic model is typically composed of a number of probabilistic subword acoustic class models, whereas the language model is typically composed of a number of n-gram probabilities[2] for word sequences drawn from the recogniser's vocabulary, together with a pronunciation lexicon specifying each word in the vocabulary is thought to be articulated in terms of the subword acoustic models, where it should be noted that a word may have more than one pronunciation model. The decoding process, in this case, amounts to finding the most probable word sequence given the acoustics and the models.

1.2.1 Generative vs. Acceptor Models

Generative Models

A model which provides an estimate of the likelihood of the acoustic observations given the model, $p(X_1^N|\mathcal{H})$, may be termed a *generative model*, as it provides an estimate of the probability that the model *generated* the acoustics. An example of such a model is the formulation of the acoustic model in terms of traditional HMMs. In general, an HMM has a number of states, each of which have a set of transition links to other states, possibly including a link to the same state, called a 'self-transition'. Adhering to

[1] where this and the terms 'hybrid Hidden Markov Model/Artificial Neural Network' (HMM/ANN), 'scaled likelihood' and 'conflict of priors' are taken from [8].

[2] where, for example, a a *tri*gram is the probability a word given the two previous, $P(w_3|w_2,w_1)$.

a common formulation of such a model, each transition is assigned a '*transition probability*' and each state is assigned an '*emission probability distribution*'[3]. In the case of *discrete* HMMs, these emission probability distributions take the form of a distribution of likelihoods that a given symbol, drawn from a discrete set of symbols, will be emitted by the state. In the case of a *continuous* HMM the conditional density for some continuous random variable is associated with each state. The discrete case will be considered below, although a similar description could also be developed for the continuous case.

The left hand panel of figure 1.2 is a schematic representation of a first order Markov model[4] in which the three states are labelled as q_i, q_j and q_k respectively and the emission probability distribution for state q_k is labelled as $p(\mathbf{X}|q_k)$. The likelihood that, for example, a first order HMM emitted some symbol sequence can then be found by summing the likelihoods of all possible state sequences, Γ, through the model that could give rise to the observation sequence:

$$\begin{aligned} p(X_1^N|\mathcal{H}) &= \sum_{\Gamma} p(X_1^N|q^1,\ldots,q^N,\mathcal{H}) \\ &\approx \sum_{\Gamma} \prod_{n=1}^{N} p(\mathbf{X}^n|q_k^n) P(q_k^n|q_j^{n-1},\mathcal{H}) \ , \end{aligned} \tag{1.9}$$

where $q^1,\ldots,q^N$ denotes a state sequence of length N and q_k^n denotes the passage of such a state sequence through state q_k at time n.

The *Viterbi* approximation to equation 1.9 is given by the value of the most likely path:

$$p(X_1^N|\mathcal{H}) \approx \max_{\Gamma} \left[\prod_{n=1}^{N} p(\mathbf{X}^n|q_k^n) P(q_k^n|q_j^{n-1},\mathcal{H}) \right] \ . \tag{1.10}$$

It is common practice for ASR to use 'left-to-right' HMMs with self loops, that is HMMs which have no transitions to a previously visited state and which allow transit to the same or a following state. Using such a formulation, a minimal durational constraint for the class of speech sound being modelled is enforced by the minimum number of states that must be visited to traverse a model, whereas the temporal variability exhibited by examples of the class is accommodated by the self loops. The model shown in the left hand panel of figure 1.2 is an example of a left-to-right HMM. By concatenating subword acoustic class HMMs, as prescribed by the pronunciation lexicon, the likelihood of the acoustics given some word sequence may be found. Through the comparison of the likelihoods of competing word sequences, the utterance may be decoded as the most likely word sequence.

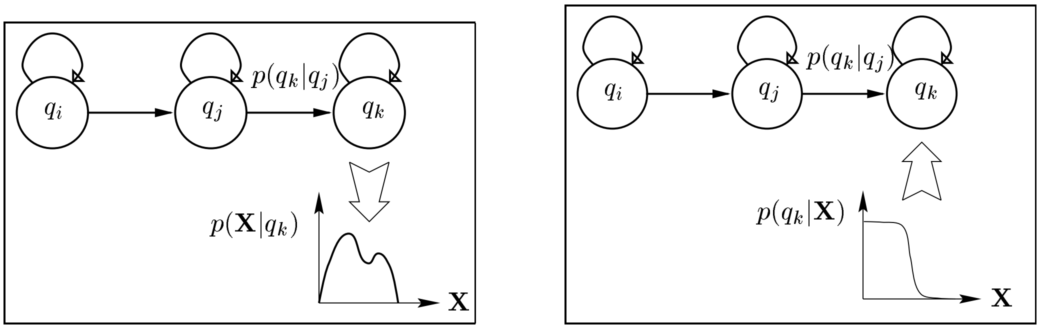


Figure 1.2: *Left*: A schematic representation of a three state, left-to-right, first order HMM with self-transitions. *Right*: A schematic representation of a finite state acceptor, again with a three state, left-to-right topology including self-transitions.

[3]where such a formulation of an HMM, which may be termed a 'Moore machine', and may be contrasted with a so called 'Mealy machine', where the latter makes use of '*emission-upon-transition*' probability distributions [12].

[4]where a first order Markov model is one in which the probability of transiting to a state is conditioned only upon the previous state. A second order Markov model is one in which the transition probability is conditioned upon the last two states from the preceding state sequence, and so on.

Acceptor Models

An *acceptor* model, on the other hand, is one which directly estimates the posterior probability of the model given the observations, $P(\mathcal{H}|X_1^N)$, as it provides an estimate of the probability of the model *accepting* the observations. It can be seen from Bayes' Theorem (1.6) that, in theory, it is possible to estimate $P(\mathcal{H}|X_1^N)$ using the components of a traditional HMM based system with the following equation:

$$P(\mathcal{H}|X_1^N) = \frac{p(X_1^N|\mathcal{H})P(\mathcal{H})}{\sum_{j=1}^{J} P(X_1^N|\mathcal{H}_j)P(\mathcal{H}_j)} \; . \tag{1.11}$$

It can be seen, however, that this is still an application of generative models as $P(\mathcal{H}|X_1^N)$ is estimated via some estimate of the data distribution. Also, the summation used to calculate the denominator is far too computationally expensive to compute in practice for most recognition tasks. An example of an acceptor model is provided by an hybrid HMM/ANN based recogniser. In the case of the hybrid HMM/ANN approach, $P(X_1^N, \mathcal{H})$, is estimated in the following way:

$$P(\mathcal{H}|X_1^N) = \sum_{\Gamma} P(\mathcal{H}, q^1, \ldots, q^N|X_1^N) \; , \tag{1.12}$$

where the RHS of equation 1.12 may be factorised:

$$P(\mathcal{H}, q^1, \ldots, q^N|X_1^N) = P(q^1, \ldots, q^N|X_1^N)P(\mathcal{H}|q^1, \ldots, q^N, X_1^N) \; . \tag{1.13}$$

Using this factorisation, the second term on the RHS of equation 1.13 may be estimated by a language model, whereas the first term may be estimated by an acoustic model. Disregarding the language model term for the moment, the acoustic model term may be further factorised:

$$P(q^1, \ldots, q^N|X_1^N) = \prod_{n=1}^{N} p(q^n|q^1, \ldots, q^{n-1}, X_1^N) \; . \tag{1.14}$$

Making some simplifying assumptions, the RHS of equation 1.14 may be reduced to a continued product of state posterior probabilities conditioned upon some *local* portion of the complete acoustic observation sequence, where a suitably trained ANN can be used to provide estimates of these *local posterior probabilities*. For example, a Multilayer Perceptron may be trained to estimate the posterior probability of a state conditioned on a window of acoustic observation vectors, centred on frame n and with c frames of context on either side, $P(q^n|X_{n-c}^{n+c})$, whereas a Recurrent Neural Network (RNN) may be trained to estimate $p(q^n|X_1^n)$. The right hand panel of figure 1.2 is a schematic representation of a finite state acceptor model which may be used to formulate acoustic class models in much the same way as an HMM. The posterior probability of a residing in state given some observations, $p(q_k|\mathbf{X})$, is shown.

A consequence of estimating local posterior probabilities directly is that the ANN inherently estimates the prior probability for each state, where this estimation is made from the acoustic data used to train the ANN. These priors may thus be called the acoustic data priors. It can also be seen, however, that the language model also specifies a prior probability for each state, where this prior is a consequence of the n-gram word sequence probabilities and the pronunciation model(s) for a word, where should a word have multiple pronunciation models, they will typically be assigned different probabilities. These priors may be termed language model priors. As the language model parameters are not derived from the same data used to train the acoustic model, these two will priors conflict. In order to combine the probabilities given by the acoustic and language models, the local posterior probabilities may be divided by the acoustic data priors, providing the 'scaled likelihoods' referred to above.

The combination of the acoustic and language models also requires the estimation of the state transition probabilities, which may be retrieved from the language model term in equation 1.13. The factorisation of this term, using Bayes' Theorem and assuming an independence from X_1^N, also provides a more formal treatment of the conflict of priors:

$$P(\mathcal{H}|q^1,\dots,q^N) = \frac{P(q^1,\dots,q^N|\mathcal{H})P(\mathcal{H})}{P(q^1,\dots,q^N)} \; . \tag{1.15}$$

By making 1st order Markov assumptions, equation 1.15 may be rewritten as equation 1.16, or as equation 1.17, if zeroth order Markov assumptions are made.

$$P(\mathcal{H}|q^1,\dots,q^N) \approx \left[\prod_{n=1}^{N} \frac{P(q^n|q^{n-1},\mathcal{H})}{P(q^n|q^{n-1})}\right] P(\mathcal{H}) \tag{1.16}$$

$$P(\mathcal{H}|q^1,\dots,q^N) \approx \left[\prod_{n=1}^{N} \frac{P(q^n|\mathcal{H})}{P(q^n)}\right] P(\mathcal{H}_i) \tag{1.17}$$

Considering the RHS of 1.17, $P(\mathcal{H})$ may be estimated, as it is for traditional HMM systems, from a large text only corpus. $P(q^n|\mathcal{H})$ is the prior probability of a state, conditioned on the model for some word sequence and therefore constitutes a prior specified by the language model. This probability can, in principle, be estimated from the pronunciation specifications of the word sequence represented by the model $\mathcal{H}$. The denominator of equation 1.17 can be considered to be the prior probability of a state given by the pronunciation dictionary as a whole, where the acoustic data priors may also be considered to be a sample estimate of this probability. This term is computationally expensive to estimate from the language model, but may be approximated by the priors calculated from the acoustic training data. By using the acoustic data approximation to $P(q^n)$, the conflict of priors can be resolved and the computational expensive of computing the term from the language model can be avoided.

Local posterior probability estimates, provided by an RNN, may be combined with the language model according to equation 1.18, if first order Markov models are used, or according to equation 1.19, in the case of zeroth order Markov models [47], where both equations make use of the Viterbi approximation.

$$P(H|X_1^N) \simeq \max_{\Gamma} \left[\prod_{n=1}^{N} P(q_k^n|\mathbf{X}_1^n) \frac{P(q_k^n|q_j^{n-1},\mathcal{H})}{P(q_k|q_j^{n-1})}\right] P(\mathcal{H}) \tag{1.18}$$

$$\simeq \max_{\Gamma} \left[\prod_{n=1}^{N} P(q_k^n|\mathbf{X}_1^n) \frac{P(q_k^n|\mathcal{H})}{P(q_k)}\right] P(\mathcal{H}) \; . \tag{1.19}$$

Through the use of local posterior probability estimates, hybrid HMM/ANN based recognisers are able to produce estimates for $P(H|X_1^N)$ in a very computational efficient manner. An additional benefit of using acceptor based models is that as they are trained to estimate posterior probabilities directly, their use of parameters is optimised with regard to recognition and so is more efficient than that of generative models for this task. The topic of HMMs has an extensive literature and is covered in more detail than is given above in [8, 42, 12, 43]. Hybrid HMM/ANNs, such as the ABBOT system [50], may be considered to be finite state acceptors and are described in more detail in [47, 34, 33, 35, 7, 8].

1.3 Overview

Chapter 2 provides a description of the process of hypothesis testing, with examples from the field of ASR, together with descriptions of a number of metrics for the evaluation of an hypothesis test. In this chapter it is important to be clear about the use of the word 'hypothesis'. In the context of ASR, there are firstly decoding hypotheses. Secondly, however, it is possible to make subsequent hypotheses about these decodings, such as whether or not a decoding is correct, where it is possible to test these subsequent hypotheses. In order to reduce any potential confusion, the term 'decoding' will be used wherever possible in chapter 2 in preference to the term 'decoding hypothesis'.

Chapter 3 firstly provides a definition of the term '*confidence measure*' and a description of a number of ways in which such confidence measures could potentially be used to improve ASR systems. Secondly, a review of the literature describing how confidence measures have been used in the field of ASR to date. Lastly, this review is contrasted with a brief description of the ABBOT hybrid HMM/ANN Large Vocabulary Continuous Speech Recognition (LVCSR) system and the introduction of five confidence measures derived from this system.

Chapter 4 provides a description of two sets of experiments which were carried out using speech read from North American business news sources, such as the Wall Street Journal, and television and radio news broadcasts, to investigate the performance of the five confidence measures for decoding hypothesis verification and Out-of-Vocabulary (OOV) word spotting tasks.

Chapter 5 contains a discussion of some issues that are relevant to these experiments and their results and to the field of confidence measures in general. Specifically, various schemes for marking decoding hypotheses as, for example, correct or incorrect, the importance of choosing an effective evaluation metric for an hypothesis test and how some confidence measures can also provide insight into how an ASR system can be improved. A summary of the key points of the report is also provided.

Chapter 2

Statistical Hypothesis Testing

2.1 Classical Hypothesis Testing

The process of statistical hypothesis testing is another example of the application of statistical decision theory. As an example of a statistical hypothesis test, consider a procedure designed to determine whether the *hypothesis* that the most probable decoding output by a recogniser is correct, is true or false. In general, there are two states of nature that are relevant to the test of some hypothesis, H, i.e. whether the hypothesis is *true* or *false*. These two states of nature can be represented by the set, $\Omega = \{\omega_1, \omega_2\} = \{\text{true}(H), \text{false}(H)\}$. The classical approach to statistical hypothesis testing requires that a *null hypothesis*, H_0, and an *alternative hypothesis*, H_1, be specified. H_0 represents an *assumption* about the state of nature, where this assumption may be either subjective or objective. In our example, H_0 may encode the assumption that the most probable decoding is correct, where this may be based upon an intuition or upon a previously calculated Word Error Rate (WER) for the recogniser. If the above example of H_0 were adopted then H_1 would correspond to the hypothesis that the decoding is incorrect. The set of possible states of nature in terms of H_0 and H_1 is therefore:

$$\Omega = \{\omega_1, \omega_2\} = \{\text{true}(H_0), \text{true}(H_1)\} = \{\text{true}(H_0), \text{false}(H_0)\} \ . \tag{2.1}$$

For this example, the observations of the world may include the probability of the decoding, the number of unique decoding hypotheses considered by the search algorithm, the probabilities of these competing decodings etc., where these observations may be represented using a feature vector. A statistical hypothesis test uses a function of this vector, $f(\mathbf{X})$. In the language of statistical inference, a function of some data sample is called a *statistic*. In the case of statistical hypothesis testing, such a statistic is called a *test statistic*.

The result of a classical hypothesis test is the decision to either accept or to reject H_0. The possible values which the test statistic may take are partitioned into two regions, namely the *acceptance* region and the *critical* region. In general, the acceptance and critical regions may lie in n-dimensional Euclidean space, $\mathbb{R}^n$, and may or may not be disjoint. Only regions in $\mathbb{R}^1$ space, corresponding to a test statistic which gives a single value, which are not disjoint will be considered in this report, however. Such regions are delimited by a single *critical value* which takes the form of a threshold. A value of the test statistic falling on one side of the threshold will result in the hypothesis being accepted, whilst a value of the test statistic falling on the other will result in the hypothesis being rejected. The set of possible actions is therefore:

$$\mathcal{A} = \{\alpha_1, \alpha_2\} = \{\text{accept}(H_0), \text{accept}(H_1)\} = \{\text{accept}(H_0), \text{reject}(H_0)\} \ . \tag{2.2}$$

Two types of error can be made when performing a classical hypothesis test:

- A *type I error* occurs when H_0 is rejected when it is true, whereas,
- A *type II error* occurs when H_0 is accepted when it is false.

The probability of a type I error is defined as the probability that H_0 is rejected, given that it is true, $P(\text{reject}(H_0)|\text{true}(H_0))$, whereas the probability of a type II error is defined as the probability that H_0 is accepted, given that it is false, $P(\text{accept}(H_0)|\text{false}(H_0))$. The probability of a type I error is also called the *size* or *significance* of the test. The probability that H_0 is rejected, given that it is false, $P(\text{reject}(H_0)|\text{false}(H_0))$, is called the *power* of the test.

The power of a test quantifies the ability of the test statistic to distinguish between true and false instances of the null hypothesis. A common method for describing the performance of a classical hypothesis test is to report the power of the test for a given significance.

An important result for classical hypothesis testing is the *Neyman-Pearson Lemma*, which states [23]:

> Given the likelihood functions, $p(f(\mathbf{X})|H_0)$ and $p(f(\mathbf{X})|H_1)$, the most powerful test for a given size of H_0 vs. H_1 has a critical region of the form:
>
> $$\frac{p(f(\mathbf{X})|H_1)}{p(f(\mathbf{X})|H_0)} \geq T, \tag{2.3}$$
>
> for some non-negative constant T.

In this case the test takes the form of a *likelihood ratio*. The threshold, T, may also be called the *operating point*, where this operating point controls the tradeoff between the power and size of the test.

2.2 Evaluation of Hypothesis Tests

2.2.1 Evaluation Metrics based on a 2×2 Confusion Matrix

Error Rates

The results of a series of applications of an hypothesis test may be recorded in a 2×2 *Confusion Matrix*. The cells of the matrix contain counts of the number of times that the application of the hypothesis test resulted in a particular action when a particular state of nature was known to be true. Such a confusion matrix for the application of a classical hypothesis test is shown in figure 2.1, where, for example, $N(\text{accept}(H_0), \text{true}(H_0))$, is the number of times that the test resulted in the action 'accept(H_0)' and where the state of nature was 'true(H_0)'.

		actions	
		accept(H_0)	reject(H_0)
states of nature	true(H_0)	$N(\text{accept}(H_0), \text{true}(H_0))$	$N(\text{reject}(H_0), \text{true}(H_0))$
	false(H_0)	$N(\text{accept}(H_0), \text{false}(H_0))$	$N(\text{reject}(H_0), \text{false}(H_0))$

Figure 2.1: A confusion matrix recording actions against states of nature.

A number of metrics can be used to summarise the data in the matrix. The simplest metric is the *unconditional error rate* which is given by:

$$P(\text{error}) = \frac{N(\text{reject}(H_0), \text{true}(H_0)) + N(\text{accept}(H_0), \text{false}(H_0))}{N(H)}, \tag{2.4}$$

where $N(H)$ is the total number of hypotheses tested and is equal to the sum of the counts in all four cells of the matrix. It is also possible to calculate error rates conditioned on the respective states of nature, where these correspond to the probability of a type I error and the probability of a type II error:

$$P(\text{type I error}) = \frac{N(\text{reject}(H_0), \text{true}(H_0))}{N(\text{true}(H_0))} \quad (2.5)$$

$$P(\text{type II error}) = \frac{N(\text{accept}(H_0), \text{false}(H_0))}{N(\text{false}(H_0))} \quad (2.6)$$

It is important to note at this point that the unconditional error rate of an hypothesis test will be influenced by the relative frequencies of the two states of nature, which may be thought of as representative of the *difficulty* of the task. To see this, consider the task of recognising only spoken digits where a recogniser for such a task will typically achieve a very low error rate, say 0.3%. In this case a simplistic verification test which merely accepts the hypothesis that the output of the recogniser is correct on all occasions will receive an unconditional error rate of 0.3%. Consider also, however, the task of recognising spontaneous speech drawn from a large vocabulary, where the WER of the recogniser will be much higher, say 30%. Here, obtaining a low unconditional error rate for a verification test is much more difficult. Whilst the unconditional error rate may be useful for determining how well some test performs upon some task of a given difficulty, it is much more desirable to be able to evaluate the performance of some test independently of the difficulty of the task when developing and comparing test statistics. This may be done through the use of conditional error rates.

ROC Curves

A *Receiver Operating Characteristic* (ROC) curve may be defined as a plot of the *detection* rate, as ordinates, against the *false alarm* (FA) rate, as abscissas [14] and in the context of decision theory can be used to plot the probability of performing some correct action conditioned on one state of nature against the probability of performing the same action conditioned, as it must be, on the other state of nature. An ROC curve may thus be created by plotting either:

- $P(\text{accept}(H_0)|\text{true}(H_0))$ on the ordinate against $P(\text{accept}(H_0)|\text{false}(H_0)) \equiv P(\text{type II error})$ on the abscissa, or
- $P(\text{reject}(H_0)|\text{false}(H_0)) \equiv$ power on the ordinate against $P(\text{reject}(H_0)|\text{true}(H_0)) \equiv P(\text{type I error})$ on the abscissa

The choice depends upon whether it is desired to plot the rate of detecting instances of $\text{true}(H_0)$ or $\text{false}(H_0)$ against the respective false alarm rate.

From the above, it can be seen that:

$$P(\text{accept}(H_0)|\text{true}(H_0)) = 1 - P(\text{reject}(H_0)|\text{true}(H_0)) \text{ and} \quad (2.7)$$

$$P(\text{reject}(H_0)|\text{false}(H_0)) = 1 - P(\text{accept}(H_0)|\text{false}(H_0)) \quad (2.8)$$

and so that the information contained in either of the two possible ROC plots is the same. It should be noted at this point that an ROC curve is different to a plot of *recall* against *precision*, where using the above context, recall may be defined as $P(\text{accept}(H_0)|\text{true}(H_0))$ and precision may be defined as $P(\text{true}(H_0)|\text{accept}(H_0))$.

A *perfect* hypothesis test is one with an error rate of zero, i.e. a detection rate of 1.0 for all FA rates, and is characterised by an ROC curve which passes through the top left-hand corner of the plot, whereas a *guessing* hypothesis test is one with a maximum error rate[1], i.e. where the detection rate is equal to the FA rate, and is characterised by an ROC curve which is a straight line, passing through the origin and the top right-hand corner of the plot. A typical ROC curve will therefore fall between these two extremes. The left hand panel of figure 2.2 illustrates the different shapes of these three possible ROC curves.

[1] excluding the case where the test becomes 'pathological' in the sense that it rejects more instances a true hypothesis than it accepts and vice versa.

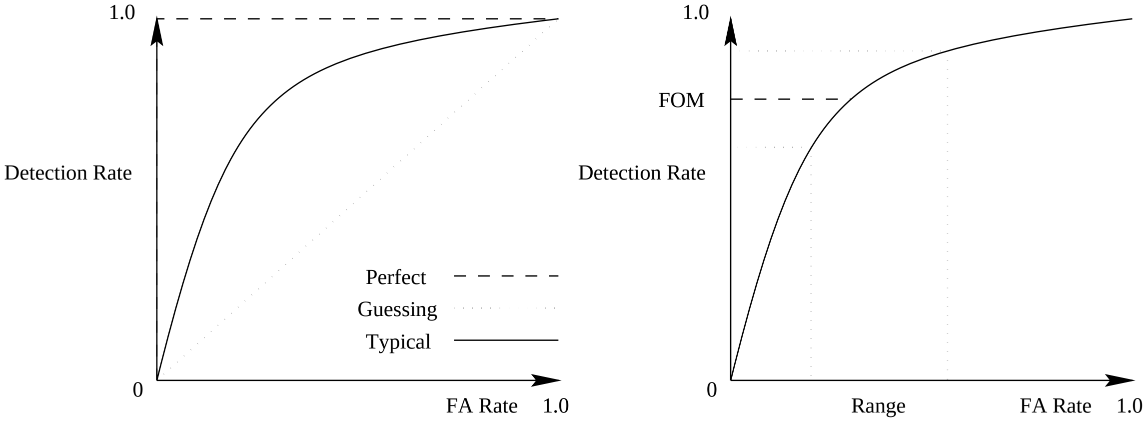


Figure 2.2: *Left*: A schematic representation of ROC plots for a perfect, guessing and typical hypothesis test. *Right*: A schematic illustration of the FOM for an ROC curve and a given range of FA rates.

The information conveyed in an ROC curve itself can be summarised. One statistic which gives a single value to characterise the curve is the *area beneath an ROC curve*[2]. Another statistic which gives a single value to characterise an ROC curve is the *Figure-of-Merit* (FOM), which is defined as the average detection rate calculated over a range of FA rates [36]. The right hand panel of figure 2.2 provides a schematic representation of a FOM for a given range of FA rates. The area beneath an ROC bounded by a lower and an upper FA rate is related to the FOM calculated over this range of FA rates by:

$$area = FOM \times \{max(\text{FA rate}) - min(\text{FA rate})\} . \tag{2.9}$$

In a similar vein to an ROC curve the probability of a type I error can be plotted against the probability of a type II error. The left hand panel of figure 2.4 provides a schematic representation of such curves for a perfect, guessing and typical hypothesis test. From equations 2.7 and 2.8, it can be seen that such a curve presents equivalent information to that presented by an ROC curve. One method for summarising a plot of $P(\text{type I error})$ against $P(\text{type II error})$ using a single value is to record the *equal error rate* (EER), where for some value of the threshold for the test:

$$P(\text{type I error}) = P(\text{type II error}) = \text{EER} . \tag{2.10}$$

DET Curves

The relationship between type I and type II errors may also be plotted using a Detection Error Tradeoff (DET) curve [39]. Here, it is assumed that the likelihood distributions for the values of the test statistic, given the two states of nature, are Gaussian. A schematic illustration of such a case is given in figure 2.3, where the means for the two distributions are labelled μ_1 and μ_2 respectively, and the operating point for the test is marked T. The hatched regions indicate the areas corresponding to $P(\text{type I error})$ and $P(\text{type II error})$.

If the Gaussian deviates are plotted instead of $P(\text{type I error})$ and $P(\text{type II error})$ for a range of operating points, a plot such as that schematically illustrated in the right hand panel of figure 2.4 is obtained, where the plot for a given test statistic will be linear if the two distributions are indeed Gaussian. The plots for a 'good' and guessing hypothesis test are shown, where it can be noted that the perpendicular distance, d, between these two lines has the property:

$$d = \sqrt{\mu_1 - \mu_2} . \tag{2.11}$$

[2] If the detection rate and FA rate cover the range [0,1], this area is equivalent to the value of the Mann-Whitney version of the non-parametric two-sample statistic [65] and has a value equal to 1.0 for a perfect hypothesis test and has a value equal to 0.5 for a guessing test. In the case of a Frequencist test, where a high test statistic value is indicative of true(H_0), the value of this area is equal to the probability that an hypothesis drawn at random from the set true(H_0) has a test statistic value that is larger than that for one drawn at random from the set false(H_0).

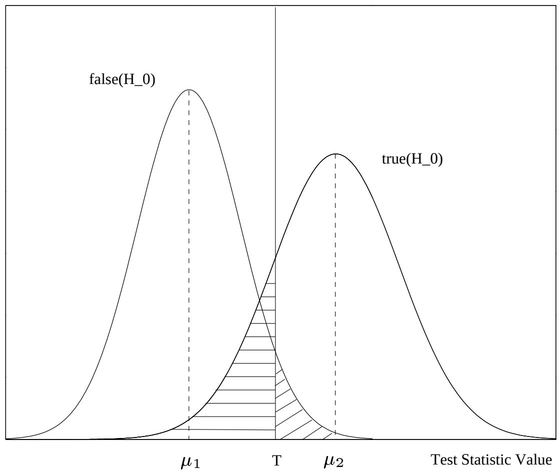


Figure 2.3: A schematic illustration of two likelihood distributions for the value of the test statistic given the two states of nature, false(H_0) and true(H_0). The operating point for the hypothesis test is labelled T and the means of the two distributions are labelled μ_1 and μ_2 respectively. The hatched regions indicate the areas corresponding to P(type I error) and P(type II error).

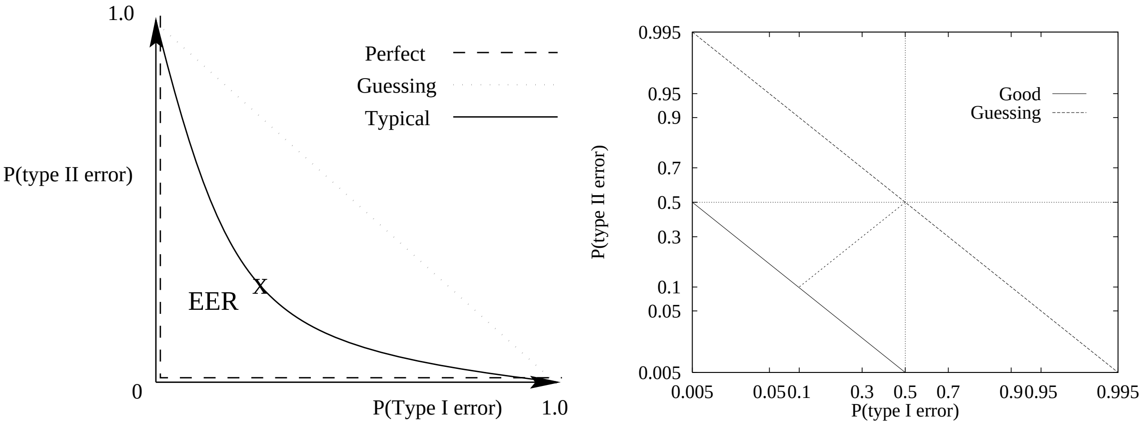


Figure 2.4: Schematic illustrations of plots of P(type I error) against P(type II error). *Left*: Curves are shown for guessing, typical and perfect hypothesis tests on linearly scaled axes. The point of Equal Error Rate (EER) is also indicated for a typical test. *Right*: A schematic illustration of a Detection Error Tradeoff (DET) curve for a good and guessing hypothesis test. In this case P(type I error) is plotted against P(type II error) on a Gaussian deviate scale.

Information Theory Based Metrics

Another evaluation metric which is derived from information theory and can be computed using the values in a 2×2 confusion matrix is the *mutual information* between the state of nature and the chosen action.

In order to compute this metric, an hypothesis test is considered to be analogous to the transmission of a signal across a noisy binary channel. Specifically, the state of nature is considered to be analogous to the source value of the signal and the action resulting from the test is considered to be analogous to the received value of the signal after transmission across the noisy channel.

The state of nature may be encoded using a binary random variable Z:

$$Z = \begin{cases} 1 & \rightarrow \omega_1 \\ 0 & \rightarrow \omega_2 \end{cases} ,$$

and the action chosen by the test may be encoded using another binary random variable A:

$$A = \begin{cases} 1 & \rightarrow \alpha_1 \\ 0 & \rightarrow \alpha_2 \end{cases} .$$

The mutual information between the values of these two random variables is given by [10]:

$$\begin{aligned} I(Z;A) &= H(Z) - H(Z|A) = H(A) - (A|Z) \\ &= -\sum_{z\in\{0,1\}} p(z)\log p(z) - \left\{ -\sum_{z,a\in\{0,1\}} p(z,a)\log p(z|a) \right\} \\ &= -\sum_{z,a\in\{0,1\}} p(z,a)\log p(z) + \sum_{z,a\in\{0,1\}} p(z,a)\log p(z|a) \\ &= \sum_{z,a\in\{0,1\}} p(z,a)\log\frac{p(z|a)}{p(z)} \\ &= \sum_{z\in\{0,1\}}\sum_{a\in\{0,1\}} p(z,a)\log\frac{p(z,a)}{p(z)p(a)} , \end{aligned} \tag{2.12}$$

where $H(Z)$ is the *entropy* of the random variable encoding the state of nature and $H(A)$ is the entropy of the random variable encoding the action. The entropy of a random variable may be defined as the *uncertainty* in the value of that random variable. $H(A|Z)$ and $H(Z|A)$ are *conditional entropies* and may be interpreted as the uncertainty in the action given the state of nature and vice versa, respectively. From equation 2.12, it can be seen that $I(Z;A)$ may thus be interpreted as the *reduction* in uncertainty in the action, given the state if nature, which is equal to the reduction in uncertainty in the state of nature given the action.

The value of $I(Z;A)$ will be equal to the entropy of the state of nature in the case of a perfect hypothesis test and equal to zero in the case of a guessing hypothesis test. The left hand panel of figure 2.5 shows how the value of the function varies over the range of possible values for the entropy of the state of nature and the error rate of the hypothesis test. The entropy of the state of nature may be considered as representative of the difficulty of the task as it is a measure of the uncertainty of the state of nature.

It can be seen from the left hand panel of figure 2.5 that value of the mutual information function, is larger for a test that performs at a certain error rate on a hard task than one which has the same error rate for an easier task.

A related metric, termed *Normalised Mutual Information* or *Efficiency* [11], is defined as:

$$E(Z;A) = \frac{H(A) - H(A|Z)}{H(A)} = \frac{H(Z) - H(Z|A)}{H(A)} , \tag{2.13}$$

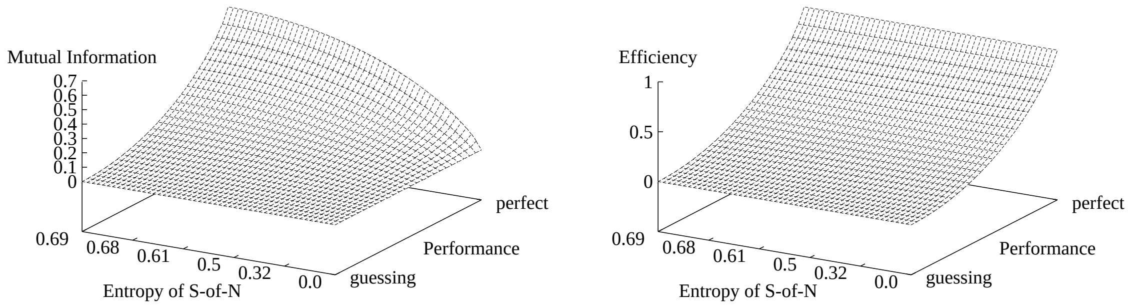


Figure 2.5: *Left*: Values of $I(Z;A)$ over the range of possible values for the entropy of the state of nature (S-of-N) and the error rate of the hypothesis. *Right*: Values of $E(Z;A)$ over the same ranges of the entropy of the state of nature and hypothesis error rate.

and is illustrated in the right hand panel of figure 2.5 over the same ranges of the entropy of the state of nature and hypothesis test error rate that were used in left hand panel for the mutual information function. In this case the normalisation ensures that an efficiency value obtained for a test with some error rate will be equal for all values of the entropy of the action. It should be noted that normalisation by $H(Z)$ would ensure equal efficiency values for all values of the entropy of the state of nature.

It should be noted that any summarisation of the data recorded in a 2×2 confusion matrix results in a loss of information. Thus information is lost when an ROC plot is used to summarise such a confusion matrix and further information is lost when an ROC curve is itself described using the area beneath it, for example. It should also be noted that the calculation of any probabilities from a 2×2 confusion matrix requires the assumption that each trial, that is application of the hypothesis test, is independent. This assumption is not strictly warranted when verifying individual components of a decoding output by a continuous speech recognition system as errors in the decoding tend to highly context dependent.

2.2.2 Probability Distribution Based Evaluation Metrics

Given a set of observation vectors, $X_1^N = \{\mathbf{X}^1, \ldots, \mathbf{X}^n, \ldots, \mathbf{X}^N\}$, the distribution of test statistic values for a particular state of nature can be plotted. If the distributions for both states of nature are plotted on the same axes, an impression of the separability of the two distributions can be gained. Ideally, the test statistic should provide values such that the distributions for the two states of nature are both narrow and well separated. Such a plot therefore provides an impression of how well the value of the test statistic distinguishes the two states of nature.

A number of *inseparability metrics* can be used to quantify how inseparable the two distributions are. As these metrics measure *in*separability, a low value indicates that the distributions are separable, whereas a high value indicates that the distributions are highly correlated. These metrics may be defined for either posterior probability distributions or likelihood distributions. Two such metrics are the *Kolmogorov Variational Distance*, D_{Kol}, and the *Bhattacharyya Distance*, D_{Bhatt}. These may be calculated for a classical hypothesis test using:

$$D_{\mathrm{Kol}} = -\sum_{n=1}^{N} \left| \frac{p(f(\mathbf{X})^n|\mathrm{true}(H_0)) - p(f(\mathbf{X})^n|\mathrm{false}(H_0))}{2} \right| \text{ and} \tag{2.14}$$

$$D_{\mathrm{Bhatt}} = \sum_{n=1}^{N} \sqrt{p(f(\mathbf{X})^n|\mathrm{true}(H_0))\; p(f(\mathbf{X})^n|\mathrm{false}(H_0))} \tag{2.15}$$

Chapter 3

Confidence Measures

3.1 Introduction

In general, a *confidence measure* may be defined as a function which quantifies how well a model matches some acoustic observations, where its values are comparable across utterances[1]. More specifically, an *acoustic* confidence measure may be defined as one derived exclusively from an acoustic model, whereas a *combined* confidence measure may be defined as one which is derived from both the acoustic and language models. The notion of the quality of a model match is illustrated figure 3.1. The left hand panel, which illustrates a poor model match, is a match between an instance of the word, “Bedouin”, and the model for the words, “Better One”, whereas the right hand panel, which illustrates a good model match, is a match of the appropriate word model to the same acoustics. The timings for the illustrated model matches were obtained from a forced alignment of each of the models to the acoustics, where the timings are derived from the most probable (Viterbi) state sequence.

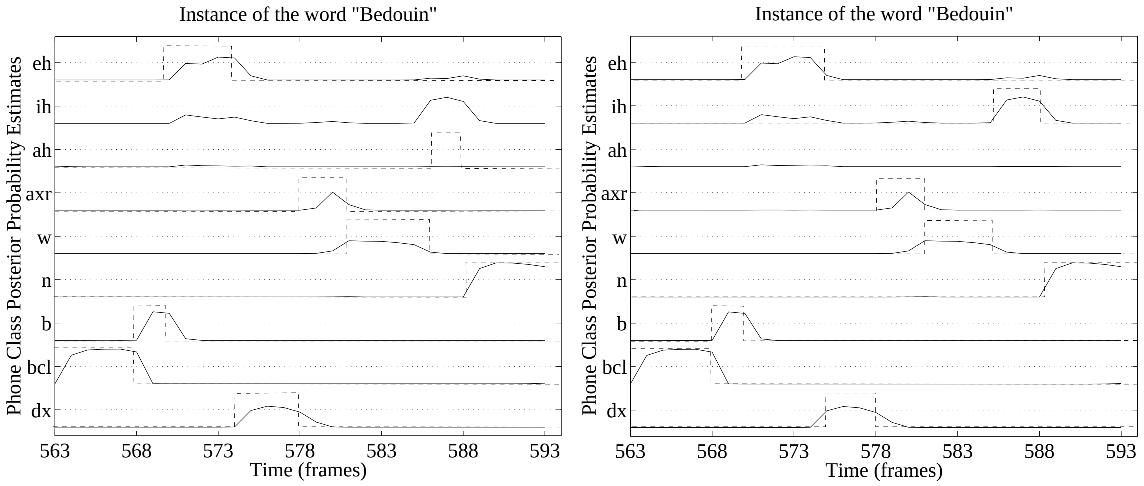


Figure 3.1: The relevant subset of acoustic model outputs generated by an instance of the word, “Bedouin”, overlaid with timings obtained from a forced alignment of the model for the words “Better One”, /bcl b eh dx axr w ah n/ (*Left*) and for “Bedouin” /bcl b eh dx axr w ih n/ (*Right*).

It can be seen from figure 3.1 that the acoustic model provides no evidence for the occurrence of the phone class /ah/ between the 586th and the 588th frame of the utterance, whereas the model for the phone class /ih/ provides a strong model match over this period. A reliable confidence measure should be able to highlight the differences in the quality of the model match for these two models.

[1] where it should be noted that this definition of a confidence measure is not restricted to the statistical modelling approach to ASR.

3.2 Potential Uses of Confidence Measures

Listeners are able to perform a number of sophisticated tasks when attempting to decode an utterance. For example, they are able to detect when:

- an acoustic signal contains speech, but is too noisy for the speech to be decoded,
- an acoustic signal has been transmitted across a low fidelity channel, such as over a telephone line,
- an utterance is produced by a male as opposed to a female speaker, or by a child as opposed to an adult speaker, or even by a particular individual,
- an utterance is spoken with a particular dialect,
- a word is OOV, spoken in a foreign language or in an unfamiliar dialect,
- a word is spoken with a disfluency or when the articulation of a word differs to the listener's model for its pronunciation.

A mechanism may be proposed for each of these capacities which necessitates the use of some form of confidence measure on the behalf of the listener. The possible uses for confidence measures in ASR mirror this list of capacities to some extent and may be described using five broad categories:

Filtering Confidence measures may be used to distinguish between a number of conditions of the input signal before the search for the most probable word sequence is begun. For example, assuming that a speaker-independent acoustic model of sufficient quality is available, portions of a speech signal with a low signal to noise ratio may be spotted if that portion is not sufficiently well matched by any of the acoustic class models. In this case, an accurate decoding is unlikely to be made and so the portion of acoustics may be simply ignored or any decoding attributed a low confidence estimate. Also, given such a model, it may be possible to determine when an acoustic signal has been transmitted across a low fidelity channel, such as the telephone system, where the acoustic model classes should provide an intermediate level of model match in this case.

Given gender specific, age specific, dialect specific and speaker specific models, again of sufficient quality, it may be possible, using the same principles, to distinguish between male and female and young and old speakers, between one dialect and another and between one speaker and another. In this case, it would be possible to invoke more specific models for the decoding process and so potentially increase recogniser performance.

Search It is possible to guide the search over the space of competing decodings with the use of a confidence measure. In this case, the partial decodings may be ranked according to their confidence and the most confident partial decoding search path extended first. The result of this search would be the most confident decoding of the utterance.

Testing Hypotheses About Decodings As it has been stated in section 1.3, it is important to be clear about the use of the word 'hypothesis' in the context of hypothesis testing and ASR. Firstly, there are decoding hypotheses that are output by the recogniser itself and secondly there are subsequent hypotheses about these outputs, such as whether or not they are correct, which may be either accepted or rejected by an hypothesis test which takes place after the decoding process.

Confidence measures are well suited to the role of a test statistic in a test of the hypothesis that an output of the recogniser is correct. This is because a confidence measure quantifies model match and a poor model match is likely to result in an incorrect decoding hypothesis. In addition to a general test of the truth or falsity of a decoding hypothesis, more specific tests can be made using a confidence measure as a test statistic. For example, a common cause of an incorrect decoding hypothesis is the occurrence of an OOV word. In this case, it is quite likely that the acoustic realisation of the OOV word will be well matched by a sequence of subword acoustic class models, but this sequence will not be a good match to any of the pronunciation specifications contained within the system's pronunciation lexicon. Another possible cause of an incorrect decoding hypothesis is the occurrence of a disfluency. In this case, the above scenario as for an OOV word may be true, or more likely, some but not all of the subword acoustic class models in the best sequence will provide a good match to their aligned portion of acoustics. Other potential causes of incorrect decoding hypotheses are noise or a badly modelled dialect, which may be tested for in ways described above.

Information Sources for Subsequent Processes Confidence estimates for decoding hypotheses can act as an additional knowledge source in a number of processes which use the output of a recogniser. One area in which confidence estimates can be particularly useful is in spoken language interfaces to command and control applications. In this case, it is useful to have a confidence estimate for a decoding hypothesis, such as "Delete all my files"!, so that an appropriate recourse can be taken given a decoding hypothesis with a suitably low confidence estimate[2].

Other areas in which confidence estimates are useful are automatic model adaptation and information retrieval from archives of spoken documents. In the case of automatic model adaptation, it is undesirable to adapt a model to some example acoustics unless it is known with reasonable certainty that the correct model is being adapted. A confidence measure for some decoding hypothesis can be used a such measure of certainty and so model adaptation can be limited to those decodings that have a sufficiently high level of confidence. Information retrieval from archives of spoken documents requires a search for targets contained in the decoding of such data. As the decoding of such data is imperfect, it is useful to weight the 'hit' rate statistics with the confidence estimates of those hits.

Diagnostics As a confidence measure quantifies model match, it can provide useful feedback to a system developer. In this case, confidence measures can be used to spot models in a system which do not provide a good match to the system's input and so highlight components which are most likely to contribute to any system error. An example of the use of such feedback is the identification of poor pronunciation models.

It can be seen from the above that confidence measures are potentially useful *throughout* the recognition process, that is before, during and after the search for the most probable decoding hypotheses. From an implementational viewpoint, it is imperative that any confidence measure used to either filter the acoustics input to a recogniser or to guide the search must be relatively inexpensive to compute in comparison to the decoding process itself. If it is not the case that computational savings are made over an unmodified decoding process and the subsequent rejection of any low confidence decoding hypotheses, then the usefulness of a confidence measure in a filtering or search guiding capacity is lost.

3.3 Literature Review

In the vast majority of ASR systems, the acoustic model is formulated in terms of traditional HMMs. In this case, the acoustic model produces estimates of the likelihood of the observed acoustics given the hypothesised decoding, $p(X_1^N|\mathcal{H})$, and the combination of the acoustic and language model provides an estimate of the joint probability of the acoustics and the decoding, $p(X_1^N, \mathcal{H})$. As both these quantities are related to the quality of the match of the model to the acoustics, they are at first sight potential candidates for the role of a confidence measure. Both $p(X_1^N|\mathcal{H})$ and $p(X_1^N, \mathcal{H})$, however, are relative to the probability of the observed acoustics themselves, $P(X_1^N)$ and so are not comparable across utterances, which an essential feature of any confidence measure. Some further processing is therefore required to convert the probabilities output by a 'traditional' HMM based system into useful confidence measures. This is in contrast to 'hybrid HMM/ANN' or 'finite state acceptor' based systems, which are capable of estimating the posterior probability of the hypothesised decoding given the observed acoustics, $P(\mathcal{H}|X_1^N)$, in a direct and computationally efficient manner, where such posterior probabilities are implicitly scaled by $P(X_1^N)$ and so are comparable across utterances.

The confidence measure literature is dominated by descriptions of methods for converting the likelihoods and joint probabilities output by traditional HMM based recognisers into useful confidence measures. Two popular methods are to either normalise these quantities in some way so that they are comparable across utterances or to train some application specific post-classifier which takes as input a number of features, including likelihoods, language model probabilities and information from n-best decoding lists.

The use of confidence measures has been reported for a number of tasks which may be summarised using a subset of the categories outlined above:

[2] A simplistic approach to such a situation would be to prompt the user to repeat the relevant utterance, possibly providing a playback of the relevant acoustics as feedback, in the hope that the second example will yield a more confident decoding than the first. A more sophisticated approach, however, would be to take steps to eliminate the uncertainty in the decoding by prompting the user with relevant questions.

3.3.1 Filtering

Barker et. al. [3] used an acoustic confidence measure derived from the ABBOT hybrid HMM/ANN LVCSR system for segmenting BN data into portions of acoustics which either matched the acoustic model with high confidence, which it was hoped would correspond to clean speech, or with low confidence, which it was hoped would correspond to music and speech mixed with music. The confidence measure is based upon a per-frame entropy measure which quantifies the degree of confusion as to the acoustic class of the observations. It was found that it could be used to successfully filter out portions of acoustics which would otherwise be recognised with a high WER, whilst incurring little extra computational cost, and thus decrease both decoding time and WER.

3.3.2 Search

Neti et. al. [41] trained a post-classifier to estimate the probability that a decoding hypothesis drawn from the search path is correct, given a set of features derived from a number of information sources, such as the likelihood of the hypothesis and n-best lists of hypotheses generated both with and without language model information. The post-classifier was formulated in terms of a binary decision tree and its output was used to weight the language model component of the decoding score. In this way, the search could be influenced by the confidence estimates for components of a path. It was found that the WER could be reduced for data from the ATIS corpus but that no performance increase could be obtained for data from the SwitchBoard corpus using this technique. A possible explanation for this discrepancy may that the quality of the language model match may be reduced for the more challenging SwitchBoard task and so the language model will have less influence upon the differences in cost of the competing search paths and so any weighting by confidence will also have a correspondingly reduced effect.

3.3.3 Testing Hypotheses About Decodings

As it has been described in the previous section, the term 'testing hypotheses about decodings' encompasses a number of related but slightly different tasks. Three such tasks, which dominate the confidence measure literature are:

Utterance Verification Here, the task is to determine whether or not a given decoding hypothesis is correct, where it is assumed that all words which are input to the recogniser are within its vocabulary. Possible causes of an incorrect decoding hypothesis, if this assumption holds, include the occurrence of non-speech sounds, a noisy channel or disfluencies.

Keyword Spotting In this case, the recogniser acts as an interface to an application which only requires the detection of the occurrence of members of a relatively small set of '*keywords*'. The assumption that the recogniser will encounter an unconstrained vocabulary, containing non-keywords together with keywords, silence, non-speech sounds and disfluencies is also made. An example of such an application is a repatory dialling service. Here a user may associate some keyword with a telephone number which the system dials when the associated keyword is spotted. An example of an input to such a system may be, "<cough> please dial, umm, Phil", where 'Phil' is a designated keyword.

OOV Word Spotting Here, the task is to detect the occurrence of any OOV words, where it is assumed that the vast majority of the words encountered by the recogniser are within its vocabulary. Given that such an assumption holds, and where the input to the recogniser is predominantly clean speech, the occurrence of an OOV word is a prime candidate for the cause of an incorrect decoding hypothesis.

Utterance Verification

The task of utterance verification may be formulated in terms of a test of the hypothesis that a given decoding hypothesis is correct. A popular approach to generating a confidence measure which may be used as the required test statistic is to normalise the likelihood of the observed data given the decoding hypothesis by the likelihood of the observations given by some '*alternate*' model. This approach may be motivated from either a Frequencist or a Bayesian viewpoint. Taking the Frequencist view, H_0 may be set to the hypothesis that a given decoding hypothesis is correct, in which case H_1 would correspond to the hypothesis that the given decoding hypothesis is incorrect. Given two models which estimate $p(X_1^N|H_0)$ and $p(X_1^N|H_1)$ respectively, a likelihood ratio may be formed and used as the test statistic. In practice, the likelihood of the acoustic observations given the decoding hypothesis is often used as an approximation to $p(X_1^N|H_0)$ and some very general model of the acoustics that may be observed is used as an approximation to $p(X_1^N|H_1)$. Taking the Bayesian view, Bayes' Theorem, given in equation 1.6, indicates that the likelihood of some acoustic observations given a decoding hypothesis should be normalised by the unconditional probability of the acoustics themselves, $P(X_1^N)$. Likelihoods 'scaled' by the unconditional probability of the acoustics[3] are comparable across utterances and so may be used as confidence measures.

One method for normalising the likelihood of the decoding hypothesis is to create an explicit alternate model. For example, the use of an explicit alternate model is reported in [44, 38, 11, 61, 57] and [60]. Rahim et. al. [44], Lleida & Rose [38], Sukkar et. al. [61] and Setlur et. al. [57] also report the use of composite alternate models and discriminative training techniques. The composite alternate models described are composed of general models of extraneous acoustics together with 'anti-keyword' or 'imposter' models which are trained upon examples of members of the recogniser's vocabulary other than a specific keyword. The discriminative training techniques reported involve optimising the parameters of the vocabulary models and/or alternate model using the ratio of the likelihood of the observations given the decoding hypothesis and the likelihood of the observations given the alternate model as the training criterion. From a Frequencist viewpoint, this criterion can be seen as based upon the Neyman-Pearson Lemma, whereas if the alternate model is considered to be sufficiently general so as to provide an estimate of $P(X_1^N)$, the criterion may be compared to the Maximal Mutual Information (MMI) criterion.

Another method for generating the alternate likelihood is to use the likelihoods of competing decodings. Cox & Rose [11] investigated the performance of a likelihood ratio formed by normalising the likelihood of the Viterbi path for the decoding hypothesis in question by the likelihood of the best Viterbi path for the period spanned by the decoding hypothesis. This confidence measure was not found to perform as well as one derived using an explicit alternate model, however. Rivlin et. al. [49] describe using the posterior probability of a given context independent phone class as a phone level confidence measure, where the probability was computed using Bayes Theorem and the likelihoods and prior probabilities for all phone classes. Word level confidence estimates could then be found using the confidence estimates for the phones constituent to the word. Caminero et. al. [9] describe using the 'on-line garbage' technique for an utterance verification task, where the technique involves averaging the likelihoods of the n-best decoding hypotheses on a per-frame basis and using the result as the alternate model likelihood. The term 'on-line garbage' was coined by Boite et. al. [5], where they introduce the technique for the task of keyword spotting.

Another approach to the task of decoding hypothesis verification is to train a post-classifier to estimate the probability that a given decoding hypothesis is correct given some features of that hypothesis. The use of post-classifiers taking features such as the likelihood of the decoding hypothesis, the language model probability, speaking rate, the length of the hypothesis and information from n-best decoding lists is reported in [15, 63, 24] and [56]. Eide et. al. [15] describe the use of a binary decision tree as a post-classifier which may be used for diagnostic as well as confidence measure purposes. From a diagnostic viewpoint, they report that the presence of short words and an increase in speaking rate are correlated with an increase in WER. Weintraub et. al. [63] describe the use of an ANN as a post-classifier. In contrast to the previous study, the choice of post-classification model in this case meant that the cause of any low confidence estimates was much harder to trace. This said, they did report, like the previous study, that shorter decoding hypotheses are more likely to be incorrect. An algorithm for marking decoding hypotheses as either correct or incorrect which makes use of the timings of hypotheses together with a number of metrics for evaluating the performance of a confidence measure are also described. Gillick et. al. [24] report using a Generalised Linear Model (GLM) as the model choice for the post-classifier as well as a cross-entropy based evaluation metric. Schaaf & Kemp [56] report the investigation of a number of different features and post-classification models which they evaluated using the unconditional error rate and an entropy based evaluation metric. An ANN was found to be the best performing post-classification model. All the above studies found that information derived from n-best decoding lists was the most valuable with regard to providing a confidence estimate for a decoding hypothesis.

Yet another approach is to the task is to calculate probability distributions for the likelihoods of both correct and incorrect decoding hypotheses. The decision to either accept a given decoding hypothesis as correct or to reject it as incorrect is then based upon the relative probabilities drawn from the two

[3]termed '*scaled likelihoods*' in [8].

distributions for the estimated likelihood of hypothesis. Both Rivlin [48] and Fetter et. al. [19] report using this approach to generating confidence estimates.

Keyword Spotting

A common approach to the task of keyword spotting is to employ a set of '*filler*' models, sometimes known as a '*garbage*' models, to compete with the keyword models in a decoding, where a filler model is defined as a model of any extraneous acoustics with regard to a given set of keywords. Examples of such extraneous acoustics may include silence, non-speech sounds and extraneous speech. The generation of keyword hypotheses is often followed by a keyword hypothesis verification stage. It should be noted that in the absence of a language model, competition between two generative models, such as between a keyword and a filler model, in a decoding is equivalent to the formation of a likelihood ratio between the two, where, for example, the filler model is included in the Viterbi decoding should the likelihood ratio between it and all other keyword models provide a value which lies in the acceptance region.

The use of explicit models of extraneous acoustics has been reported in [51, 64, 55, 52, 40, 37, 30, 52, 36, 53] and [54]. One of the earliest descriptions of the use of an explicit filler model is given by Rohlicek et. al. [51] who report the use of whole word keyword models and a single filler model created by combining the Gaussian emission probability distributions obtained from all the keyword models used in the system. Wilpon et. al. [64], on the other hand, report one of the first uses of filler models trained on examples of extraneous acoustics. Lleida et. al. [37] report an investigation into various forms of filler models, including phonetic fillers, syllabic fillers and word-based fillers. The use of a set of syllabic fillers was found to provide the best performance.

An additional verification stage for hypothesised keywords was introduced by Rose & Paul [55] where the likelihood of the acoustic observations given by an alternate model, separate to the filler model, was used to form a ratio with the likelihood of the observations given by the keyword model. The addition of a verification stage was found to reduce the number of false alarms given by the keyword spotter. In a later study [52], Rose introduces the use of a discriminative training technique, based upon the Neyman-Pearson Lemma, to optimise the parameters of the keyword models with respect to the filler model and so further increase the performance of the keyword spotter in question. Mazor & Feng [40] also report the use of an additional keyword verification stage. In this case, the application of a threshold to various quantities, including the likelihood of the keyword hypothesis and the difference between the likelihood of the keyword hypothesis and the next best model likelihood, was investigated. Lippmann et. al. [36] report the use of a discriminative training criterion based upon the FOM of a keyword spotter to optimise the parameters of the keyword models.

The increased discrimination between the two states of nature by a test which is brought about through the use of a training criterion based upon the Neyman-Pearson Lemma can be seen to increase the probability of rejecting H_0 given that it is false, i.e. the power of the test, and so the probability of rejecting H_0 given that it is true, i.e. the probability of a type I error, must correspondingly be reduced. Likewise it can be seen that the probability of accepting H_0 given that it is true will be increased and so the probability of a type II error will correspondingly be reduced. Training according to a criterion based upon the FOM of a test, on the other hand, can be seen to increase the detection rate for a given range of false alarm rates. For one of the two ROC curves that it is possible to plot for a test, this will correspond to increasing the power of the test for a given range of the probability of a type I error. Although it has been shown previously that the information contained within the two ROC plots is equivalent, this is not the same as increasing the probability of accepting H_0 for a given range of the probability of a type II error.

A number of approaches to the keyword spotting task have been reported that do not make use of explicit filler models, where the filler model is derived from the keyword models instead. The on-line garbage approach to keyword spotting is reported in Boite et. al. [5] and Bourlard et. al. [6]. In these studies, on-line garbage was found to perform favourably in comparison to the use of various forms of explicit filler models. Klemm et. al. [32] report the use of fillers based upon word, syllable and phone models. In the case of word models the keyword spotter is equivalent to a large vocabulary recognition system. The performance of syllabic fillers was found to compare favourably to that of the large vocabulary recogniser based approach in this study. A similar approach and results is also reported by El Méliani & O'Shaughnessy [16, 18, 17]. The application of large vocabulary recognition techniques is also described by Weintraub [62].

Jeanrenaud et. al. [30, 29] extend the task of keyword spotting to 'event spotting', where they define an event as, "a collection of phrases or clauses, of one or more words, that relate to a single concept". They give an example of such an event as, "time-of-the-day", which may be instantiated as, "one thirty", or, "one thirty in the morning on Tuesday". They report the investigation to two approaches to the task, one based on keyword detection and sub-grammars of keywords which describe events and the other based upon the use a large vocabulary recogniser. The sub-grammar based approach was found to perform favourably in comparison to the large vocabulary recognition based approach.

OOV Word Spotting

Here, the task may be formulated in terms of a test of the hypothesis that a given word input to the recogniser is within its vocabulary. It is in the nature of the task that all the words in the training set will be incorporated into the recogniser's lexicon and so it is not possible to train a filler model on examples of extraneous acoustics. A proposed solution to this problem is to use a 'looped phone' alternate model described by Asadi et. al. [1]. In this case an HMM is constructed so that the phone models used to construct the word models of the recogniser's vocabulary are arranged in parallel with transitions such that any phone may follow any other. If the likelihood of the observations given the looped phone model is greater than the likelihood given by any of the word models from the recogniser's vocabulary, the inference is that the observations constitute an OOV word. A similar approach using syllable models is reported by Kemp & Jusek [31]. In a later study reported by Asadi et. al. [2], the incorporation of any identified OOV word into the recogniser's vocabulary is addressed. In this case, a phone level decoding of an identified instance of an OOV word was found not be sufficiently reliable to provide a pronunciation model of the previously OOV word and text-to-speech rules were found to perform better, where it assumed that the spelling of the new word can be provided to the system. Haeb-Umbach et. al. [25], on the other hand, report that given several examples of a new word, the 'average' phone level decoding compares favourably with text-to-speech rules as a pronunciation model for that word.

An interesting and quite different approach to the problem is described by Hetherington [28]. In this case it is acknowledged that a recogniser will attempt to match a model from its vocabulary to an instance of an OOV word. It is further proposed that in such a case an n-best word lattice will reveal the attempted match of many different word models to the offending portion of acoustics, all of which with a low likelihood. This is in contrast to the case for a well modelled word within the recogniser's vocabulary, where it is assumed that the lattice will reveal the repeated match of the correct word, all with a high likelihood. Hetherington therefore proposes a measure which he terms the 'active word count' which is defined as, "the number of unique words competing with each other during the recognition process at each point in the utterance", which may be used to identify the occurrence of an OOV word.

The language model constitutes an especially useful information source for the identification of OOV words where, for example, Suhm et. al. [59] report that 27% of OOV words are proper nouns. Asadi et. al. [1] report the use a statistical class grammar where the probability of encountering an OOV word is estimated for each open class, such as 'SHIP-NAME' in the grammar. Suhm et. al. [59] describe a fairly simple approach where a standard n-gram language model is augmented with an OOV word class. Fetter et. al. [20, 21], on the other hand describe a much more complicated approach to generating a language model probability for OOV words, which they term 'iterative substitution'. In this case, it is argued that the majority of the words in the language model training data are incorporated into the recogniser's vocabulary, and so a mismatch arises between the frequency of OOV words in the training data and in the test data. To address this problem an initially small subset of the training data is declared as within vocabulary and all other words are declared as OOV. N-gram statistics are computed for the OOV word class, the within vocabulary subset enlarged and the process iterated, where the statistics derived from successive iterations are merged. It is claimed that this process exerts a smoothing effect upon the calculated n-gram statistics. Gallwitz et. al. [22] report the use of a category based n-gram where the probability of an OOV word is calculated for each category, where categories such as 'CITY' may be created manually or where some clustering algorithm may be used to derive categories automatically.

3.3.4 Remarks

The methods for deriving confidence measures described in the literature reviewed above suffer from a number of problems:

Normalising Likelihoods Creating a sufficiently general alternate model to accurately model the diverse range of acoustics required, such as silence, non-speech sounds, disfluencies and non-keywords, is very difficult and would seem to require an approach somewhat more sophisticated approach than the use of a fairly arbitrary set of HMMs that is adopted in the vast majority of the literature.

Discriminative Training Criteria Although the MMI criterion and one based upon the Neyman-Pearson Lemma are discriminative, the MAP criterion may be preferred as it does not assume equal priors.

Post-Classifiers It is important to choose a useful set of features in order to be able to generate an accurate estimate of the probability of some hypothesis given the feature values. For example, as likelihoods are not comparable across utterances, it is important to provide such a post-classifier with information besides likelihoods. According to Bayes' Theorem, information should be provided regarding the prior probability of the the hypothesis together with the unconditional probability of the observations, in addition to the likelihood of the hypothesis.

Likelihood Distributions As likelihoods are not comparable across utterances, basing a decision upon frequency distributions for the likelihoods of decoding hypotheses seems to be rather ad hoc . It may further be speculated that this approach may only work for very constrained tasks in which $P(X_1^N)$ is broadly similar for all utterances.

Keyword Models as Fillers The use of keyword models as fillers has several attractive properties. Firstly, as it has been stated above, designing and training explicit models of a diverse range of extraneous acoustics is difficult. Secondly, when subword models are employed as fillers, as opposed to applying large vocabulary recognition techniques, the filler model is task independent and also may be used to detect and provide a phone level transcription of OOV words. A problem with deriving the filler model from the keyword models, however, is that without the use of a principled approach such as Bayes' Theorem, no accommodation is made for non-speech sounds and disfluencies, which will inevitably occur in practical keyword spotting applications. The use of large vocabulary speech recognition techniques for keyword spotting is also computationally expensive in comparison to the use of filler models.

Diagnostics Although the study reported by Eide et. al. [15] goes some way to providing some diagnostics as to the cause of incorrect decoding hypotheses, short words and a fast speaking rate were found to be a common feature of many incorrect hypotheses, it does not provide any direct indication as to how a recognition system may be improved. For diagnostic tools to be genuinely useful, they must not only indicate where a recogniser is failing, but also suggest ways in which to remedy any problems.

The next two sections provide a contrast to the literature reviewed above, which is dominated by descriptions of methods for generating confidence measures from 'traditional' HMM based systems. Section 3.4 describes the ABBOT hybrid HMM/ANN LVCSR system [50], which is able to directly estimate the posterior probability of a decoding hypothesis given some acoustics. Section 3.5 introduces five confidence measures which may be derived from the ABBOT system.

3.4 The ABBOT system

The structure of the ABBOT system is illustrated in figure 3.2. The digitised acoustic waveform is first preprocessed using frame based Perceptual Linear Prediction (PLP) analysis, to produce an acoustic feature vector, typically for every 16ms of the input signal. Using an appropriately trained RNN, a vector of local posterior probability estimates for a set of context independent phone classes[4] is then obtained for each frame of the acoustics. As it has been stated previously, a key property of these local phone class posterior probability estimates is that they provide a means to obtain posterior probability estimates for a decoding hypothesis given some acoustics in a computationally efficient manner. As such posterior probabilities are implicitly scaled by the probability of the acoustics themselves, they are comparable across utterances and thus may be used as acoustic confidence measures. A number of confidence measures are described in the next section which make use of this key property.

[4] It should be noted that a set of context dependent phone classes may be derived from these, however.

The output vectors of the network over time form a phone class posterior probability space which is searched using the NOWAY stack based decoder for the optimal HMM state sequence [45, 46]. As part of the decoding process, the posterior probabilities output by the RNN are converted to scaled likelihoods by dividing by the acoustic data priors, where these scaled likelihoods are used in place of the emission probabilities estimated using 'traditional' HMMs. The HMMs used in the ABBOT system are called 'repeated state' HMMs as the scaled likelihoods of the acoustics given each state of the HMM are derived from the estimated posterior probabilities of the same phone class given the appropriate portion of acoustics. The repeated states of these HMMs effectively provide duration models for the context independent phone classes. The conversion to scaled likelihoods circumvents the conflict between the acoustic data priors and those specified by the HMM topologies and the language model [8, 47]. Using, word level decoding constraints, the optimal word or phone sequence can be obtained by supplying NOWAY with a word n-gram language model, a set of phone HMMs and pronunciation specifications for the words in the language model. The optimal phone sequence, given phone level constraints, can be obtained using a phone n-gram language model, phone HMMs and no pronunciation lexicon. NOWAY is also capable of producing n-best decoding lattices.

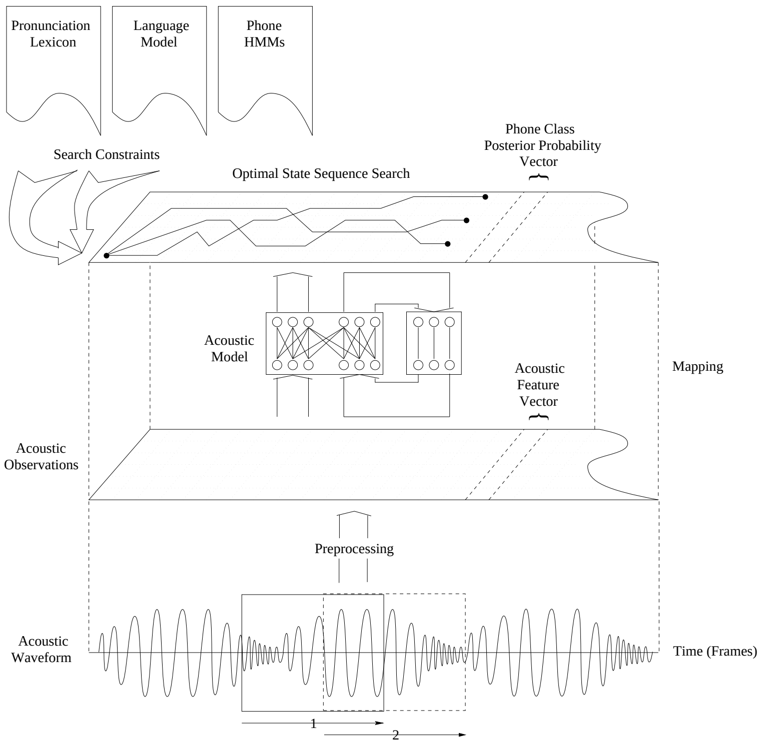


Figure 3.2: A schematic representation of the ABBOT hybrid HMM/ANN LVCSR system.

The ABBOT system is based on 1.19, except that the first order Markov prior, $P(q_k^n|q_j^{n-1}, \mathcal{H})$ instead of the zeroth order prior, $P(q_k^n|\mathcal{H})$. It should also be noted that repeated state HMMs are used in which each state shares the same probability distribution for a context independent phone class and so both state and phone probabilities may use the label q_k.

3.5 Five Confidence Measures

The four acoustic confidence measures and one combined confidence measure investigated are described below. Three of the confidence measures made use of the optimal state sequence and so may be used to test the model match for a particular class model. The remaining confidence measures, namely $CM_{ent}(n_s, n_e)$ and $CM_{lat}(n_s, n_e)$, do not make use of the optimal state sequence. $CM_{ent}(n_s, n_e)$ is an acoustic confidence measure and so may be considered as a *general* measure of acoustic model match and was designed to spot portions of acoustics that are not well matched by *any* of the acoustic class models. $CM_{lat}(n_s, n_e)$ is a combined confidence measure and so may be considered to be a general measure of the combined acoustic and language model match. Equations for the five confidence measures using phone level decoding hypotheses, q_k, are given below, where n_s and n_e are the start and end frames of an hypothesis respectively:

Scaled Likelihood $CM_{sl}(q_k)$ is the log scaled likelihood of the phone hypothesis q_k, as used in the decoding process.

$$\begin{aligned} CM_{sl}(q_k) &= \sum_{n=n_s}^{n_e} \log\left(\frac{p(\mathbf{X}^n|q_k)}{p(\mathbf{X}^n)}\right) \\ &= \sum_{n=n_s}^{n_e} \log\left(\frac{P(q_k|\mathbf{X}^n)}{P(q_k)}\right) \quad . \end{aligned} \tag{3.1}$$

Posterior Probability $CM_{post}(q_k)$ is computed by rescoring the optimal state sequence using the local posterior probability estimates estimated by the RNN, and differs from $CM_{sl}(q_k)$ by the division by the acoustic data priors.

$$\begin{aligned} CM_{post}(q_k) &= \sum_{n=n_s}^{n_e} \log\left(p(q_k|\mathbf{X}^n)\right) \\ &= CM_{sl}(q_k) + (n_e - n_s)\log p(q_k) \quad . \end{aligned} \tag{3.2}$$

Online Garbage The term 'online garbage' [5, 6] is used to refer to the normalisation of the probability of the best decoding hypothesis by the average probability of the n-best decoding hypotheses. This average may be considered to be a form of garbage model probability and so a separate garbage model is not required. $CM_{olg}(q_k)$ is $CM_{sl}(q_k)$ normalised by the average of the n-best scaled likelihoods.

$$CM_{olg}(q_k) = CM_{sl}(q_k) - \sum_{n=n_s}^{n_e} \log \frac{1}{\mathrm{n-best}} \sum_{l=\mathrm{best}}^{\mathrm{nth-best}} \frac{p(q_l^n|\mathbf{X}^n)}{p(q_l)} \quad . \tag{3.3}$$

Per Frame Entropy $CM_{ent}(n_s, n_e)$ is the per frame entropy of the K phone class posterior probabilities estimated by the RNN, averaged over the interval n_s to n_e:

$$CM_{ent}(n_s, n_e) = -\frac{1}{n_e - n_s} \sum_{n=n_s}^{n_e} \sum_{k=1}^{K} p(q_k^n|\mathbf{X}^n) \log p(q_k^n|\mathbf{X}^n) \quad . \tag{3.4}$$

A schematic illustration of $CM_{ent}(n_s, n_e)$ is given in the left hand panel of figure 3.3.

Lattice Density $CM_{lat}(n_s, n_e)$ is a measure of the density of competing decodings in an n-best lattice of decoding hypotheses and is computed by averaging the number of unique decoding hypothesis which pass through a frame over the interval n_s to n_e:

$$CM_{lat}(n_s, n_e) = \frac{1}{n_e - n_s} \sum_{n=n_s}^{n_e} NCH_n \quad , \tag{3.5}$$

where, NCH_n is the number of competing decoding hypotheses which pass through the nth frame of the lattice. $CM_{lat}(n_s, n_e)$ may be computed from an n-best lattice of word hypotheses and at the word level, NCH_n is equivalent to an 'active word count' described in [28, 27]. $CM_{lat}(n_s, n_e)$ is a combined confidence measure as an n-best lattice of decoding hypotheses is created using both language model and acoustic model information. A schematic illustration of $CM_{lat}(n_s, n_e)$ is given in the right hand panel of figure 3.3.

Duration normalised forms of $CM_{sl}(q_k)$, $CM_{post}(q_k)$ and $CM_{olg}(q_k)$, denoted $CM_{nsl}(q_k)$, $CM_{npost}(q_k)$ and $CM_{nolg}(q_k)$ respectively were used throughout the experiments. Duration normalisation is desirable as otherwise a bias toward decoding hypotheses with short durations will exist.

$$CM_{nsl,npost,nolg} = \frac{1}{n_e - n_s} CM_{sl,post,olg} \quad . \tag{3.6}$$

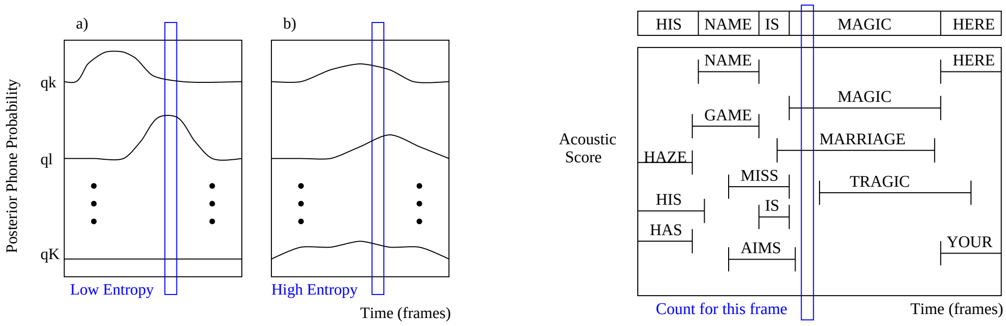


Figure 3.3: *Left*: A schematic illustration of $CM_{ent}(n_s, n_e)$. *Right*: A schematic illustration of $CM_{lat}(n_s, n_e)$.

As the phone level confidence measures $CM_{sl}(q_k)$, $CM_{post}(q_k)$ and $CM_{olg}(q_k)$ make use of the optimal state sequence, equivalent confidence measures for a word hypothesis, w_j, may be obtained by combining the confidence estimates for the phone hypotheses that are constituent to w_j. $CM_{ent}(n_s, n_e)$ and $CM_{lat}(n_s, n_e)$ may be extended to the word level by simply averaging the per frame values for the interval over which w_j is hypothesised to occur.

Chapter 4

Experiments

4.1 Method

Three data sets were used to calculate the results of the following experiments. Two data sets were composed of read North American Business (NAB)[1] data whilst the other was composed of Broadcast television and radio News (BN)[2]. Whilst, the NAB data contains examples of clean, read speech, the BN data set, contains examples of spontaneous and conversational speech, including disfluencies and noise. The performance of the confidence measures were evaluated for utterance verification and OOV word spotting tasks.

Each of the data sets were decoded under two conditions. The first condition used the word level decoding constraints of a word pronunciation lexicon and a trigram word language model, whereas as the second condition used only the phone level decoding constraints of a phone bigram language model and no pronunciation lexicon. Clearly only phone level decoding hypotheses could be obtained when using phone level decoding constraints, whereas both word and phone level hypotheses could be obtained using word level decoding constraints[3]. N-best decoding lattices could be created, for word level hypotheses using word level decoding constraints and for phone level hypotheses using phone level decoding constraints, but not for phone level hypotheses using word level decoding constraints. Confidence estimates were then calculated for each decoding hypothesis using each of the 5 confidence measures described in section 3.5, with the one exception that clearly confidence estimates using $CM_{lat}(n_s, n_e)$ could not be computed for phone level hypotheses decoded using word level constraints.

Time aligned reference word and phone sequences were then obtained for the three data sets by performing a forced alignment of the reference transcriptions to the acoustic data, using the same pronunciation lexicons used for the decodings, augmented to accommodate any words from the reference transcription which would otherwise be OOV. The decoding and reference sequences were then aligned so that each decoding hypothesis could be marked using a marking algorithm. In the case of the decoding hypothesis verification experiments, each decoding hypothesis was marked as either true or false, and in the case of the OOV word spotting experiments, each word decoding was marked as being matched to a reference word which would either be in vocabulary (IV) or OOV with respect to the *non*-augmented pronunciation lexicon. In the case of the decoding hypothesis verification experiments, a number of marking algorithms were investigated and were all found to produce broadly similar results, although an algorithm similar to that proposed in [63], which makes use of timing information, was found to produce slightly better results than the rest. According to this algorithm, in addition to insertions and substitutions, decoding hypotheses could be marked as incorrect due to bad time alignments. In order for a decoded word hypothesis, w_{hyp}, for example, to be marked as correctly time aligned, there must exist an identical word in the time aligned reference transcription, w_{ref}, such that greater than 50% of the duration of w_{hyp} overlaps with w_{ref} and vice versa. This 50% overlap criterion is illustrated in figure 4.1, which is taken from [63]. All the reported results were computed using this algorithm. A similar 50% overlap criterion

[1]The 1994 Hub-1 development test set and the 1995 Hub-3 evaluation test set of the ARPA CSRNAB evaluation data.

[2]ABC Nightline, episode 05/23/96, from the 1996 Hub-4 ARPA evaluation data.

[3]At the word level, a 20K word pronunciation lexicon and language model was used to decode the Hub-1 data, a 60K word pronunciation lexicon and language model was used to decode the Hub-3 data and a 65K word pronunciation lexicon and language model was used to decode the Hub-4 data. At the phone level, the same phone bigram language model (54 phone classes) was used to decode all three data sets.

was used to mark decoding hypotheses in the case of the OOV word spotting experiments. Once the decoding hypotheses were marked, the performance of each of the confidence measures as hypothesis test statistics could be evaluated.

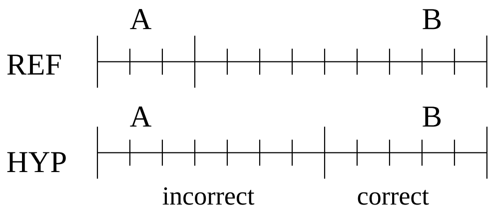


Figure 4.1: A schematic representation of the 50% overlap criterion that was used to assess the time alignment of a decoding hypothesis. Taken from [63].

Given an appropriately marked set of decoding hypotheses, a confidence estimate for each member of the set and some threshold on the values for the particular confidence measure, an hypothesis test could be applied. In the case of the verification experiments, H_0 was set to the hypothesis that a given decoding hypothesis is correct, whereas for the OOV word spotting experiments, H_0 was set to the hypothesis that a given decoding hypothesis was matched against an IV word, with respect to some pronunciation lexicon. The results of a series of applications of an hypothesis test, to each member of the set of decoding hypotheses was recorded in 2×2 confusion matrix such as that shown in figure 2.1. Given such a matrix, the performance of the test could be evaluated using a number of the metrics described in chapter 2.

The unconditional error rate of a test, that is the probability of either rejecting H_0 when it was true, or accepting H_0 when it was false, was computed using equation 2.4. The conditional error rates, that is $P(\text{type I error})$ and $P(\text{type II error})$, on the other hand, were calculated using equations 2.5 and 2.6. Other probabilities conditioned on the state of nature required for the plotting of ROC curves were obtained through the use of equations 2.7 and 2.8. The joint, marginal and conditional probabilities required for computing the mutual information between the states of nature and the actions resulting from the test using equation 2.12 were computed in a similar fashion. A value for the efficiency function could then be obtained through the use of equation 2.13. Thresholds across the entire range of values that a given confidence measure could take were investigated, where the application of each threshold was recorded using a separate 2×2 confusion matrix. A series of values for the above metrics could then be computed, allowing graphs to be plotted showing how each metric value varied across the range of 0 to 100% hypothesis rejection.

Values for the two evaluation metrics D_{Kol} and D_{Bhatt} could be computed without application of a threshold or the use of a 2×2 confusion matrix. In this case, the set of decoding hypotheses and their associated confidence estimates was partitioned according the two possible states of nature. Using this partition, equations 2.14 and 2.15 were used to compute the required metric values. These two sets were also used in conjunction with software supplied by NIST [58] to plot DET curves.

4.2 Results

Figures 4.2 and 4.3 show the results of the word level decoding hypothesis verification experiments for the development test set, whereas figures 4.4, 4.5 and 4.6 and 4.7, 4.8 and 4.9 show the results of the word and phone level verification experiments for the NAB and BN evaluation sets respectively. Figures 4.10 and 4.11 show the results of the OOV word spotting experiments for the NAB development and evaluation test sets respectively. There were not enough OOV words in the BN data set, with respect to the 65K word lexicon, to obtain any meaningful OOV word spotting results for the BN data.

The top panel of figure 4.2 shows the effect of duration normalisation upon the performance of the three confidence measures $CM_{post}(w_j)$, $CM_{olg}(w_j)$ and $CM_{sl}(w_j)$ as test statistics. The plot shows the sum of the probability of a type I error and the probability of a type II error against the percentage of rejected test hypotheses. It can be seen from the plot that duration normalisation produces a marked improvement in performance. Duration normalised versions of these three confidence measures, denoted by an 'n' prefix, were therefore adopted throughout the rest of the experiments.

The two centre panels of figure 4.2 provide a comparison of an ROC curve, which in this case is the plot of power against $P(\text{type I error})$, and the plot of $P(\text{type I error})$ against $P(\text{type II error})$. It can be seen from these two panels that equivalent information is provided by the two plots, and so only ROC plots are given for the remainder of the experiments.

The bottom left and right panels provide a comparison of the relative performance of the 5 word level confidence measures as measured by the mutual information, $I(Z;A)$, and efficiency, $E(Z;A)$, metrics respectively. Again the values of metrics are plotted against the percentage of rejected test hypotheses. $E(Z;A)$ was computed using the entropy of the state of nature as normalisation factor for all experiments. It can be seen from the two plots that, for a particular value of the entropy of the state of nature, both $I(Z;A)$ and $E(Z;A)$ provide identical information, where the only difference is a scaling of the ordinate values. Only the values of $E(Z;A)$ are therefore provided for the rest of the experiments.

Figure 4.3 shows the distributions of the five word level confidence measure values for the two states of nature of interest, i.e. whether the test hypothesis is true, true(H), or false, false(H). Clearly, the greater the separation of the two distributions, for the values of some test statistic, the better the test will be able to distinguish between the two states of nature. It can be seen that there is good correlation between the separation of the distributions and the values of the two inseparability metrics, D_{Kol} and D_{Bhatt}, where a lower value for both metrics indicates greater separability of the two distributions. The values of the two inseparability metrics only are therefore reported for the rest of the experiments.

Figure 4.4 shows the relative performance of the five word level confidence measures for the verification test using the NAB evaluation data and word level decoding constraints. The top left panel of the figure shows the performance of $CM_{nolg}(w_j)$ for the for an increasing number of n-best decoding hypotheses used to calculate the measure. From this plot, it can be seen that a slight improvement in performance is gained by increasing n from 1 to 5, but that a negligible improvement is gained thereafter. It may be hypothesised that this is due to the fact that the scaled likelihoods used to compute $CM_{nolg}(w_j)$ are derived from the local posterior probabilities output by the ANN and that, due to the discriminative nature of its training, the probability distribution over the output vector of the network is typically quite 'sharp'. In this case, only a small number of phone classes will have high posterior probabilities, and thus high scaled likelihoods, at any one time. The inclusion of any additional class likelihoods in the computation will therefore add little to the sum used for the garbage likelihood.

The top right hand panel of the figure shows the range of unconditional error rates over the range of the percentage of testing hypothesis that were rejected by the test. It can be seen from this plot that a very low unconditional error is obtained when no testing hypotheses are rejected, and similarly that a very high unconditional error rate is obtained if all of the testing hypotheses are rejected. This reflects the fact that the recogniser itself has a very low error rate for this data set. In this case it is clear that rejecting H_0 for an increasing percentage of the samples will increase the error rate of the test. The plot also shows a very similar curve for all five confidence measures. This indicates that when the prior probability of one state of nature is much greater than the other the information provided by the test statistic has little effect upon the unconditional error rate statistics for that test.

The upper and lower right hand panels of the figure show the relative performance of the five confidence measures as measured by the sum of the probability of a type I error and the probability of a type II error and by $E(Z;A)$ respectively, where both are plotted against the percentage hypothesis rejection. There is a striking similarity between the two plots, where one is very close to an inverted version of the other, indicating that the information theory based metrics provide the same information as those based on conditional error rates. A higher value of $E(Z;A)$ indicates improved performance.

The upper left hand panel of the figure show the relative performance of the confidence measures as illustrated by an ROC curve, whereas the lower left hand panel provides similar information in the form of a DET curve. The table at the bottom of the figure provides values for D_{Kol} and D_{Bhatt}, the area under the ROC curve and the EER for the various confidence measures.

In general, it can be seen that the 7 metrics, P(type I error) + P(type II error), $E(Z;A)$, D_{Kol}, D_{Bhatt}, the ROC and DET curves and the EER, agree in the information that they provide, and that in this particular case, $CM_{lat}(n_s, n_e)$ performs significantly better than CM_{npost}, CM_{nolg} and CM_{nsl} and that $CM_{ent}(n_s, n_e)$ performs the worst. $CM_{ent}(n_s, n_e)$ was conceived as a confidence measure for *general* acoustic model match, where a high confidence value will be given if *any* of the acoustic class models provide a good match to some sample acoustics and will produce a low confidence estimate if *all* of the acoustic class models provide a poor match to the acoustics. In the case of clean speech, $CM_{ent}(n_s, n_e)$ will therefore always give low values and so will not be useful distinguishing correct and incorrect decoding hypotheses or IV and OOV words. It should be noted at this point that as $CM_{ent}(n_s, n_e)$ is computed on a per-frame basis from only the outputs of the RNN and so is relatively computationally inexpensive to compute. It is thus well suited to the role of a confidence measure for filtering the acoustics that are input to a recogniser [3]. $CM_{lat}(n_s, n_e)$ on the other hand draws upon more information sources than a purely acoustic confidence measure, which makes it more expensive to compute, and so can be expected to perform better when these information sources are reliable. Low values of $CM_{lat}(n_s, n_e)$ will be obtained when a decoding hypothesis provides a good match to *both* the acoustic and language models and vice versa. As the NAB corpus contains only clean speech, the majority of decoding errors may be attributed to OOV words. A decoding hypothesis matched against an OOV word will typically provide a bad acoustic model and language model match which will be reflected in the value of $CM_{lat}(n_s, n_e)$.

Figure 4.5 shows the relative performance of four of the phone level confidence measures for the verification test using the NAB evaluation data and word level decoding constraints. $CM_{lat}(n_s, n_e)$ could not ne calculated for this decoding condition. From this figure, it can be seen that the unconditional error rate follows a similar pattern to that shown previously and that the other evaluation metrics agree on a slight performance increase for the majority of the confidence measures on this test.

Figure 4.6 provides the corresponding performance indicators for the verification test of phone level hypotheses decoded using phone level constraints. This figure shows a increase in the unconditional error rate but also some general improvement in hypothesis test performance over the previous two decoding conditions. It can also be seen from this figure that the performance of $CM_{lat}(n_s, n_e)$ is markedly reduced for this decoding condition, only performing better than $CM_{ent}(n_s, n_e)$. It may be hypothesised that this is due to a decrease in the quality of the information provided by the language model as we move from a word trigram to a phone bigram.

Figures 4.7 and 4.8 and 4.9 show the relative performance of the five confidence measures the same for same decoding hypothesis verification tests using as computed using the BN evaluation data. The top left panel from figure 4.7 shows the performance of $CM_{nolg}(w_j)$ for the for an increasing number of n-best decoding hypotheses for this particular data set. In this case it can be seen that performance improves sufficiently for increasing n to warrant the use of $CM_{nolg-40}(w_j)$.

From these figures, some general performance trends can be identified:

1. It can be seen that the unconditional error rate of the tests using the BN data is higher than for those using the NAB data. This is a reflection of the lower error rate of the recogniser for data from the NAB corpus, which contains clean read speech, than for data from the more challenging BN corpus, which contains, amongst other conditions, spontaneous speech and speech mixed with music.

2. The universal improvement given by duration normalisation to three of the acoustic confidence measures can be attributed to the tendency to unfairly penalise long decoding hypotheses otherwise.

3. From the conditional error rate and information theory based metrics, it can be seen that a lower percentage of test hypotheses must be rejected to obtain the optimal hypothesis test performance using the NAB data than when using the BN data, but that the maximum level of performance is broadly similar for both data sets.

4. It can be seen that $CM_{lat}(n_s, n_e)$ is the best performing confidence measure for word level hypotheses on the NAB data, but the *worst* performing measure on the BN data. It was initially speculated that the usefulness of the measure using BN data was reduced due to the saturation of the n-best lattices , reflecting the increased difficulty of decoding this data. The lattices, however, were found not to be saturated. Two further hypotheses may be proposed as to why $CM_{lat}(n_s, n_e)$ performed better on the NAB than on the BN data. The first hypothesis is that the quality of the language model as a model of the underlying generator of the data is reduced as we move from the highly grammatically constrained read speech of the NAB corpus to the more spontaneous and conversational speech of the BN corpus. As the usefulness of a confidence measure is entirely dependent upon the quality of the model used to derive it, if the model is not a good model of the underlying generator of the data, the value of the confidence measure will not be particularly informative. In the case of a poor language model the value of $CM_{lat}(n_s, n_e)$ will be artificially raised for *all* decoding hypotheses and so the performance of the confidence measure as a test statistic will be reduced, where the reduced information supplied by the $CM_{lat}(n_s, n_e)$ reflects the reduced information supplied by the language model. This hypothesis is also supported by the reduced performance of the measure under phone level decoding constraints, where again the reduction in performance may be supposed to be due to a decrease in the quality of the language model information. The second hypothesis is that the poor performance of $CM_{lat}(n_s, n_e)$ on BN data was caused by the use of an insufficient beam width for the search process that created the lattices.

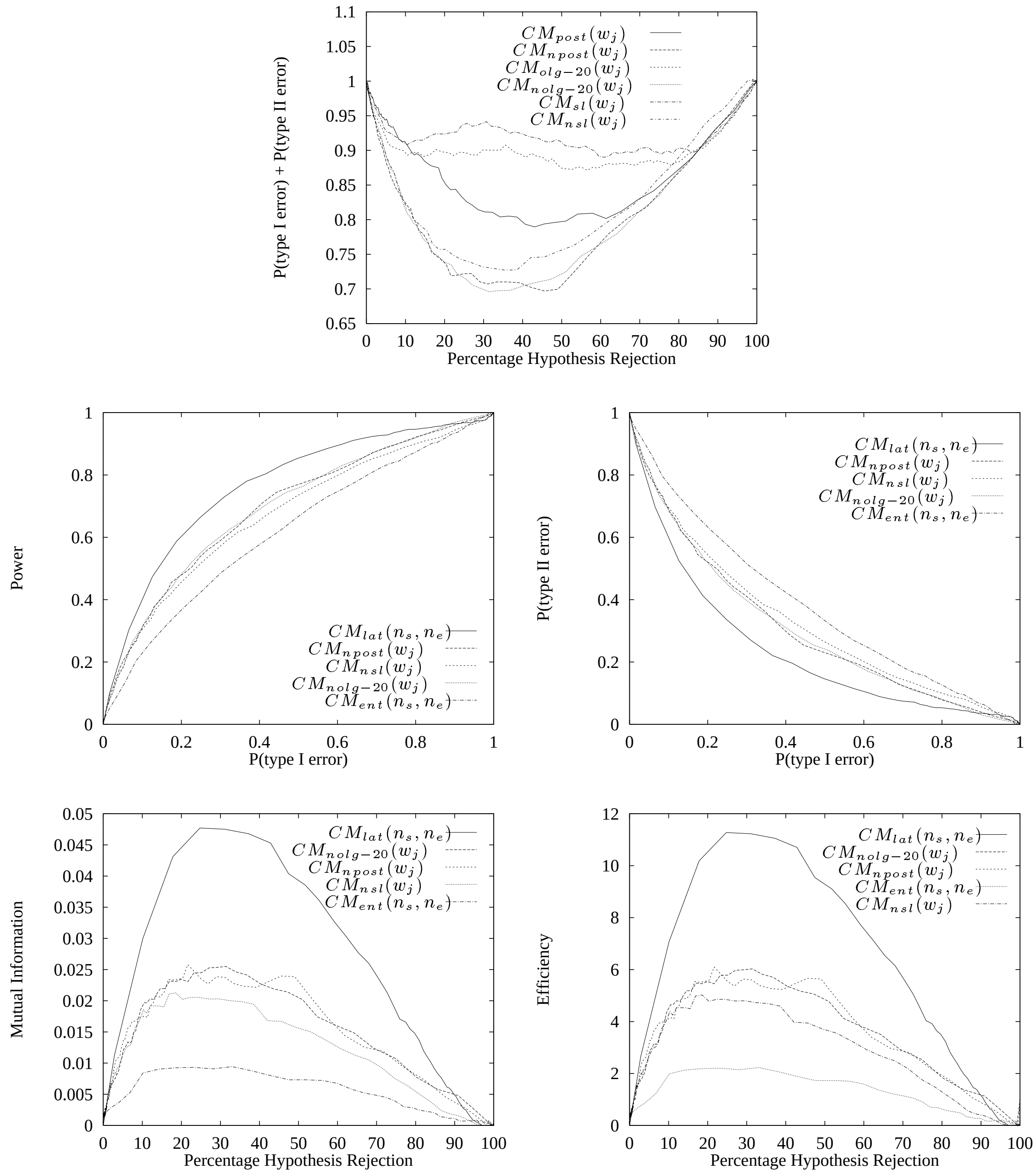
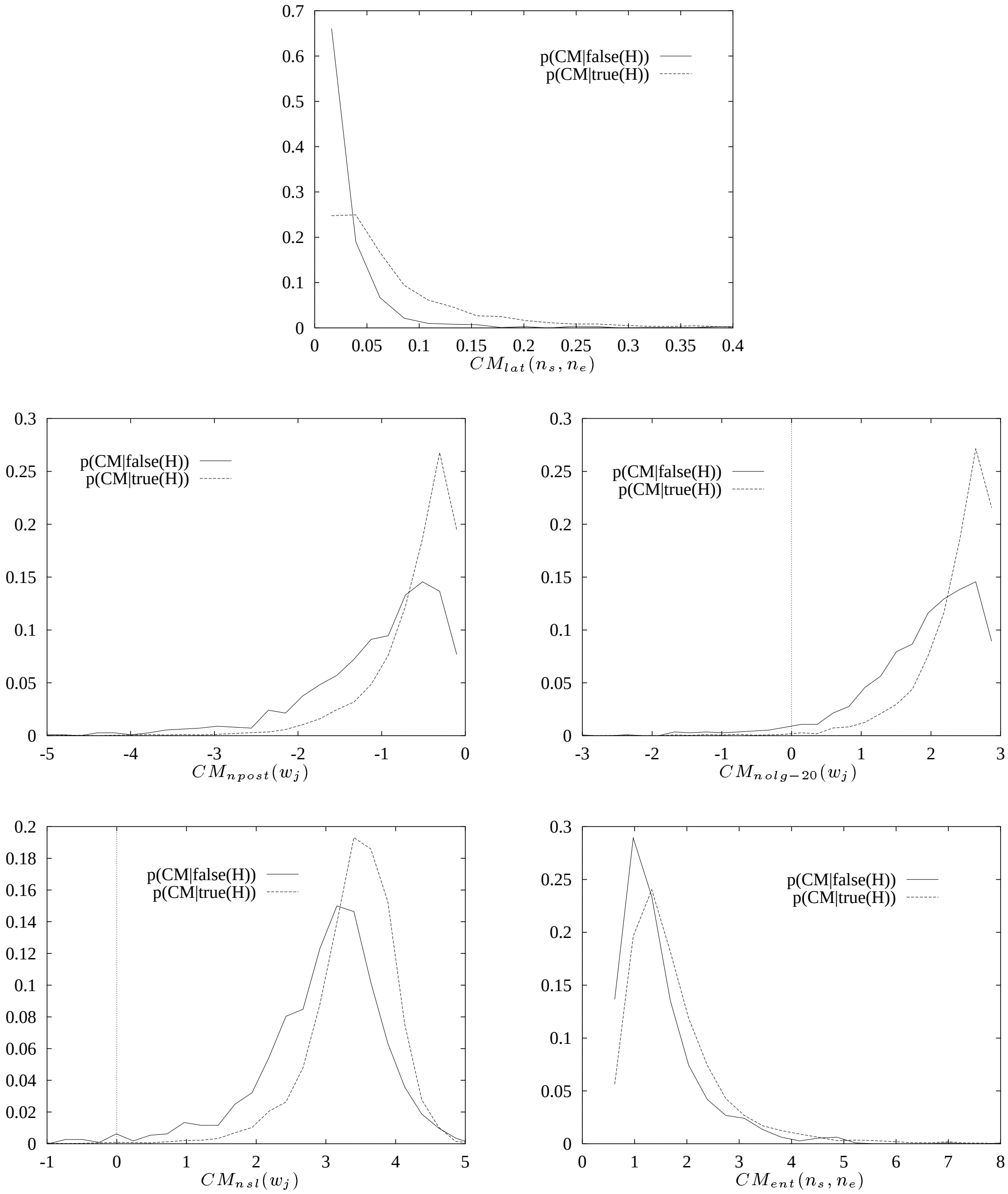
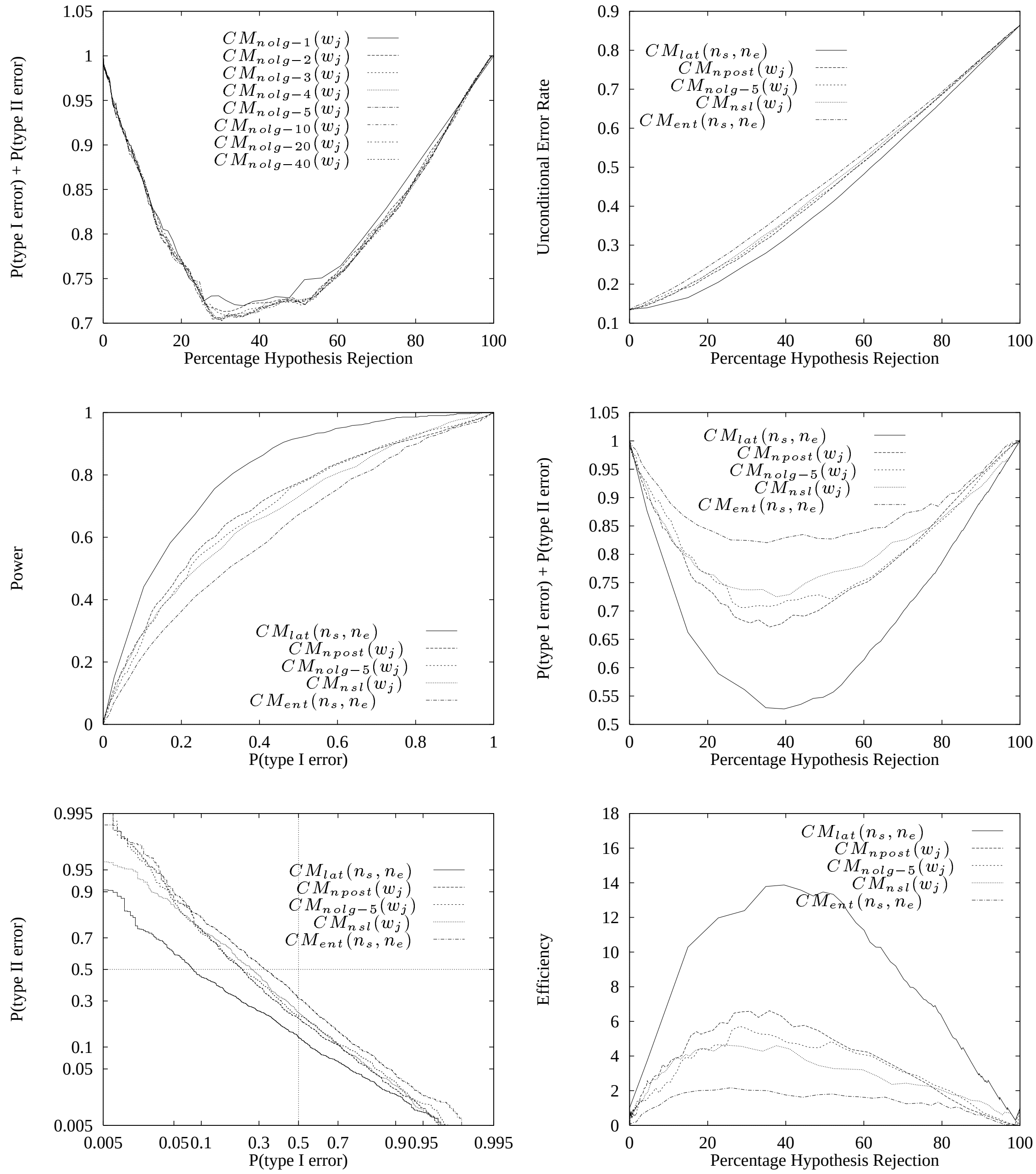


Figure 4.2: Verification test performance for word level hypotheses (6331 correct and 1122 incorrect) from the Hub-1 '94 development test set, decoded using word level decoding constraints. *Top*: The effect of duration normalisation. *Centre Left*: The ROC curve of power against the probability of a type I error. *Centre Right*: The probability of a type I error plotted against the probability of a type II error. *Bottom Left*: Mutual information, $I(Z;A)$, against percentage test hypothesis rejection. *Bottom Right*: Efficiency, $E(Z;A)$, against percentage test hypothesis rejection.



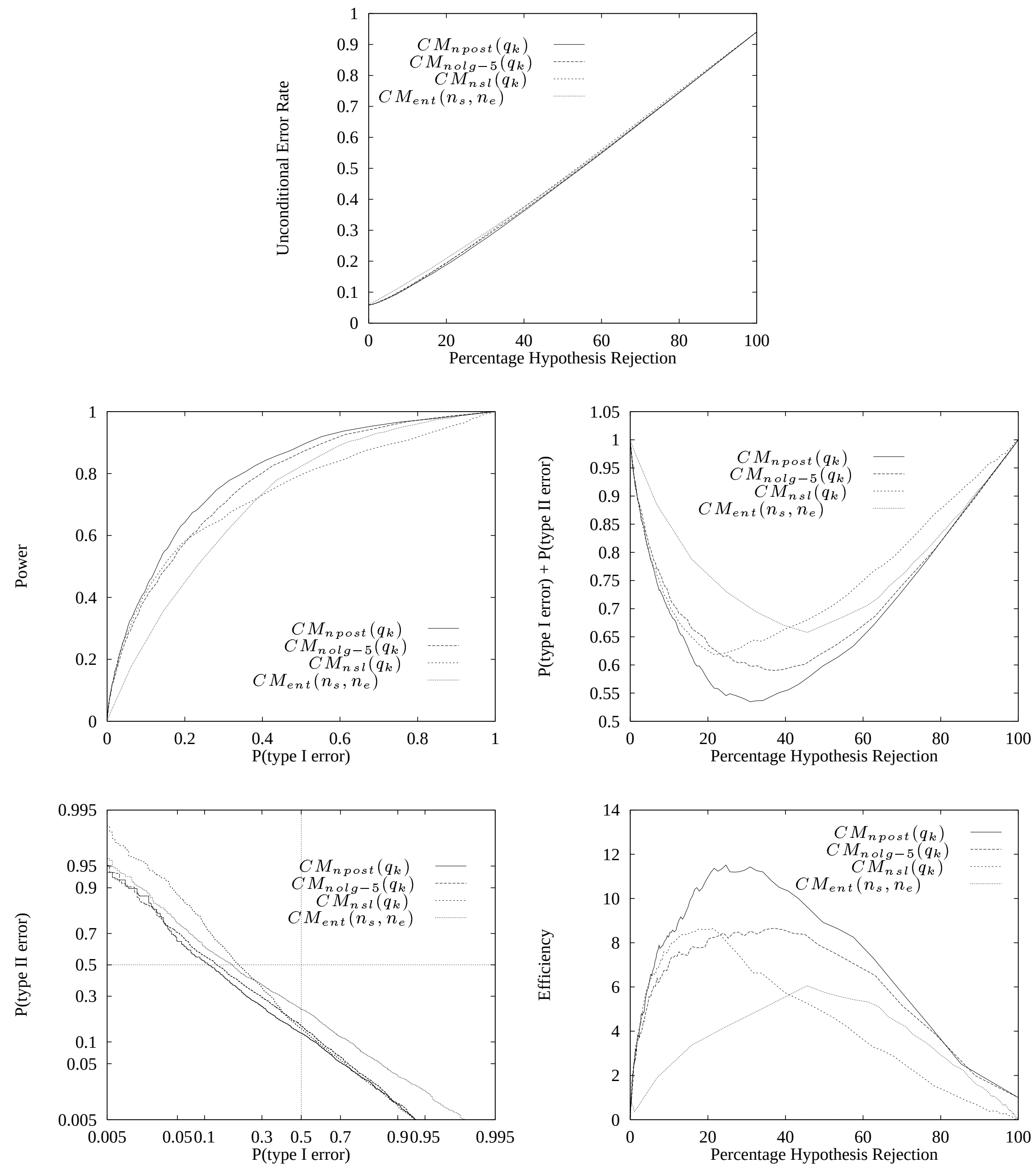
Confidence Measure	D_{Kol}	D_{Bhatt}
$CM_{lat}(n_s, n_e)$	-0.4313	0.8872
$CM_{npost}(w_j)$	-0.3016	0.9318
$CM_{nolg-20}(w_j)$	-0.2897	0.9361
$CM_{nsl}(w_j)$	-0.2691	0.9437
$CM_{ent}(n_s, n_e)$	-0.1808	0.9722

Figure 4.3: Verification test performance for word level hypotheses from the Hub-1 '94 development test set, decoded using word level decoding constraints. *Upper Panels*: The distributions of the five confidence measure values for the two states of nature of interest, i.e. whether the test hypothesis is true, true(H), or false, false(H). *Lower Panel*: The values of the two inseparability metrics, D_{Kol} and D_{Bhatt}, for the distributions associated with each of the five confidence measures.



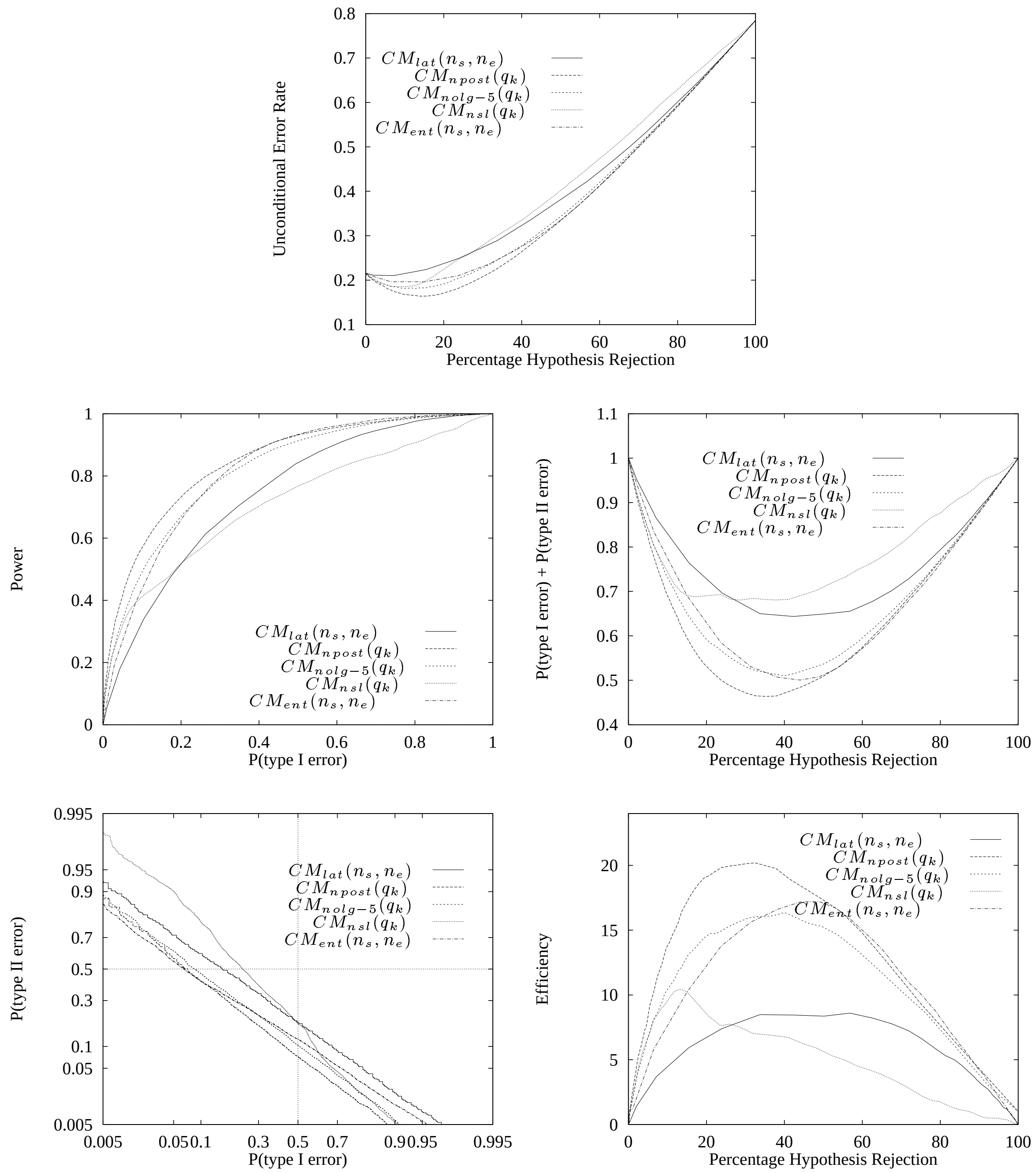
Confidence Measure	D_{Kol}	D_{Bhatt}	area under ROC curve	EER
$CM_{lat}(n_s, n_e)$	-0.4744	0.8330	0.7772	0.2647
$CM_{npost}(w_j)$	-0.3120	0.9295	0.6907	0.3387
$CM_{nolg-5}(w_j)$	-0.2897	0.9368	0.6805	0.3591
$CM_{nsl}(w_j)$	-0.2581	0.9419	0.6694	0.3646
$CM_{ent}(n_s, n_e)$	-0.1832	0.9717	0.6096	0.4174

Figure 4.4: Verification test performance for word level hypotheses (5211 correct and 818 incorrect) from the Hub-3 '95 evaluation test set, decoded using word level decoding constraints. *Top Left*: The performance of $CM_{nolg-1-40}$. *Top Right*: Unconditional error rates. *Upper Left*: An ROC curve (power against the probability of a type I error). *Upper Right*: The sum of the probability of a type I error and the probability of a type II error against percentage test hypothesis rejection. *Lower Left*: A DET curve (probability of a type I error against the probability of a type II error on a Gaussian deviate scale). *Lower Right*: Efficiency, $E(Z; A)$, against percentage test hypothesis rejection. *Bottom*: Values of D_{Kol} and D_{Bhatt}, the area below the ROC curve and the EER.



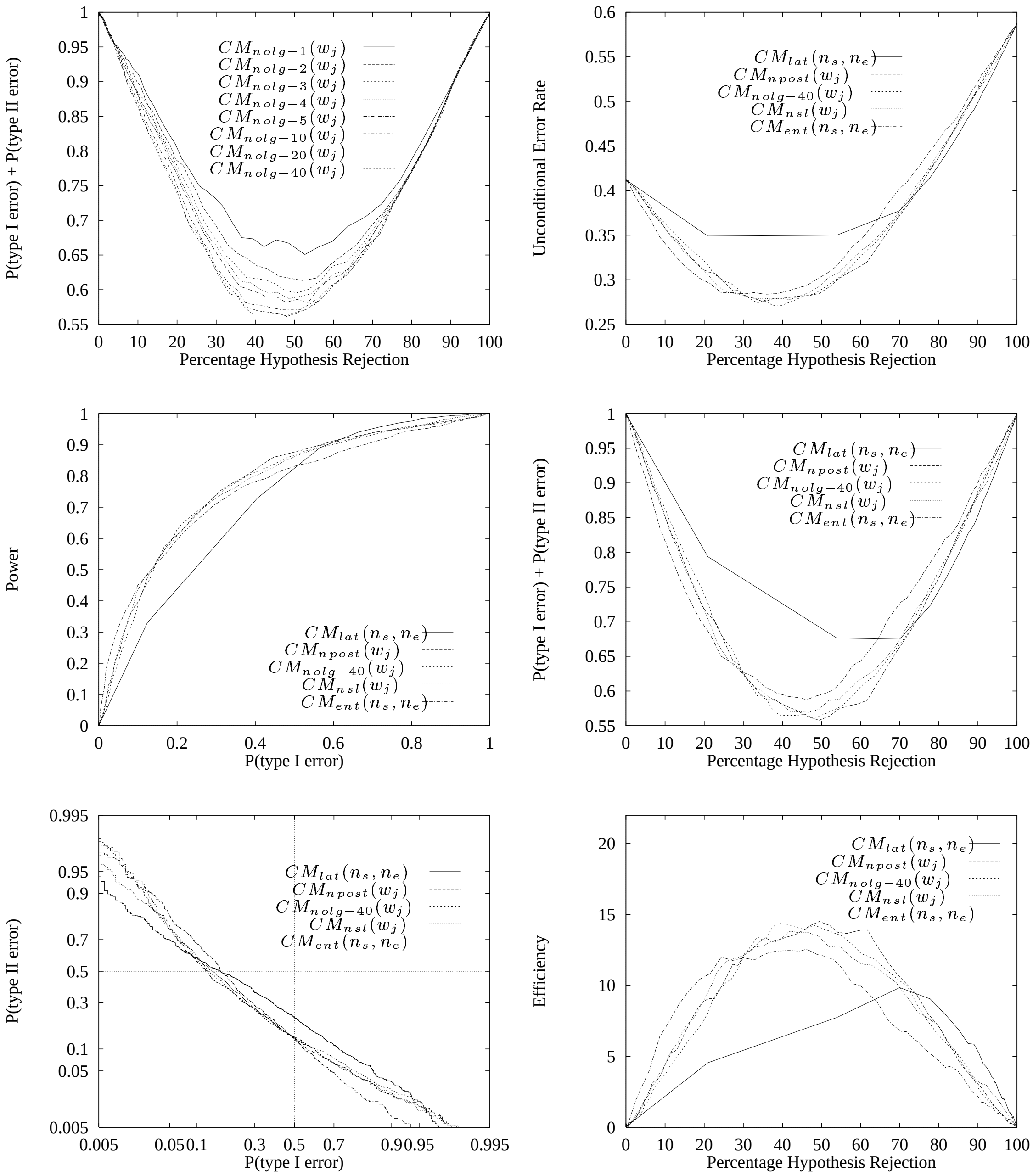
Confidence Measure	D_{Kol}	D_{Bhatt}	area under ROC curve	EER
$CM_{npost}(q_k)$	-0.4635	0.8497	0.7908	0.2697
$CM_{nolg-5}(q_k)$	-0.4025	0.8792	0.7647	0.2993
$CM_{nsl}(q_k)$	-0.3724	0.8985	0.7287	0.3724
$CM_{ent}(n_s, n_e)$	-0.3402	0.9191	0.6860	0.3402

Figure 4.5: Verification test performance for phone level hypotheses (25570 correct and 1614 incorrect) from the Hub-3 '95 evaluation test set, decoded using word level decoding constraints. *Top*: Unconditional error rates. *Upper Left*: An ROC curve (power against the probability of a type I error). *Upper Right*: The sum of the probability of a type I error and the probability of a type II error against percentage test hypothesis rejection. *Lower Left*: A DET curve (probability of a type I error against the probability of a type II error on a Gaussian deviate scale). *Lower Right*: Efficiency, $E(Z; A)$, against percentage test hypothesis rejection. *Bottom*: Values of D_{Kol} and D_{Bhatt}, the area below the ROC curve and the EER.



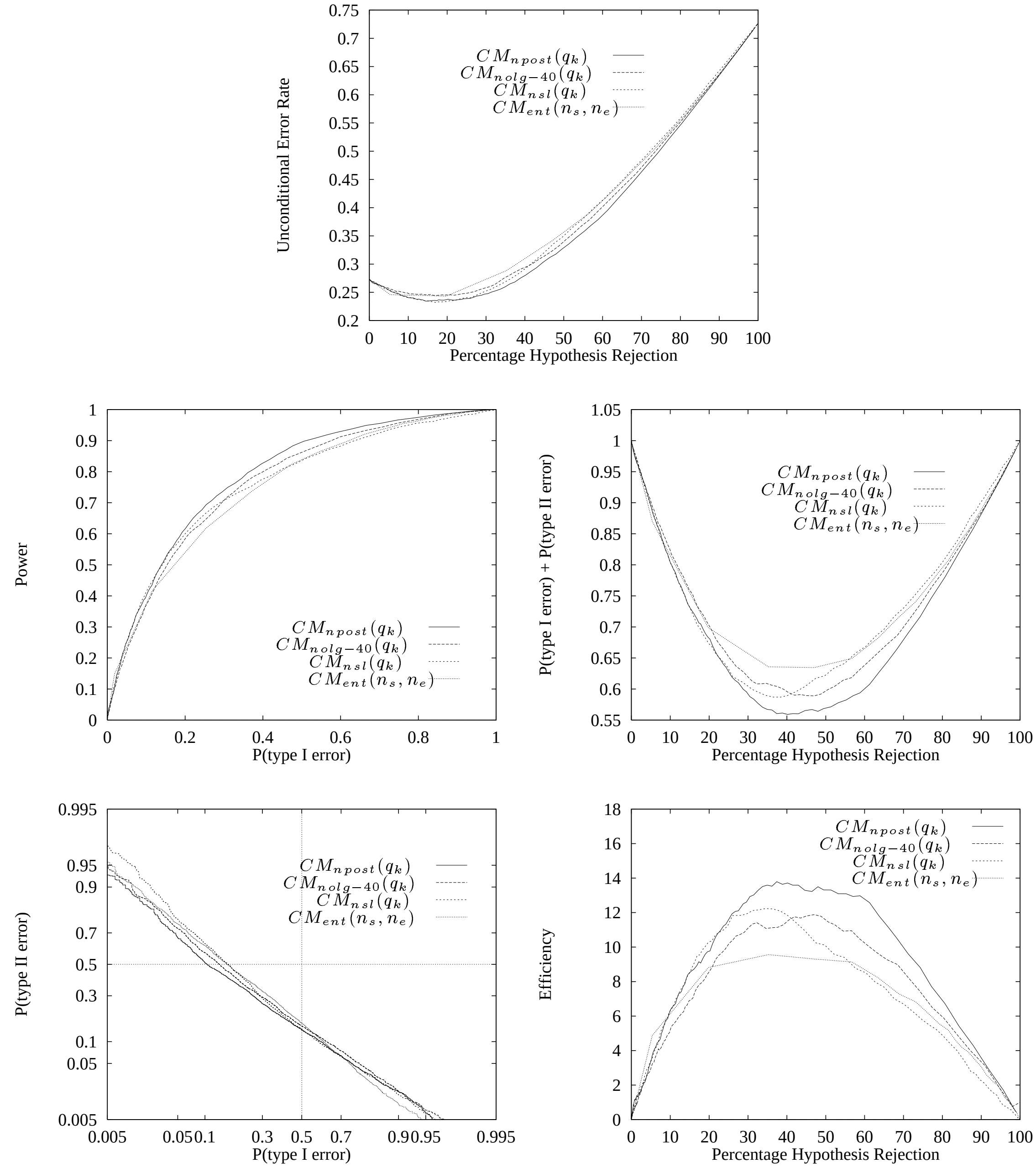
Confidence Measure	D_{Kol}	D_{Bhatt}	area under ROC curve	EER
$CM_{lat}(n_s, n_e)$	-0.3553	0.9061	0.7147	0.3218
$CM_{npost}(q_k)$	-0.5171	0.7929	0.8376	0.2318
$CM_{nolg-5}(q_k)$	-0.4816	0.8252	0.8117	0.2589
$CM_{nsl}(q_k)$	-0.3176	0.9065	0.7056	0.3410
$CM_{ent}(n_s, n_e)$	-0.5008	0.8240	0.7960	0.2533

Figure 4.6: Verification test performance for phone level hypotheses (19519 correct and 5358 incorrect) from the Hub-3 '95 evaluation test set, decoded using phone level decoding constraints. *Top*: Unconditional error rates. *Upper Left*: An ROC curve (power against the probability of a type I error). *Upper Right*: The sum of the probability of a type I error and the probability of a type II error against percentage test hypothesis rejection. *Lower Left*: A DET curve (probability of a type I error against the probability of a type II error on a Gaussian deviate scale). *Lower Right*: Efficiency, $E(Z; A)$, against percentage test hypothesis rejection. *Bottom*: Values of D_{Kol} and D_{Bhatt}, the area below the ROC curve and the EER.



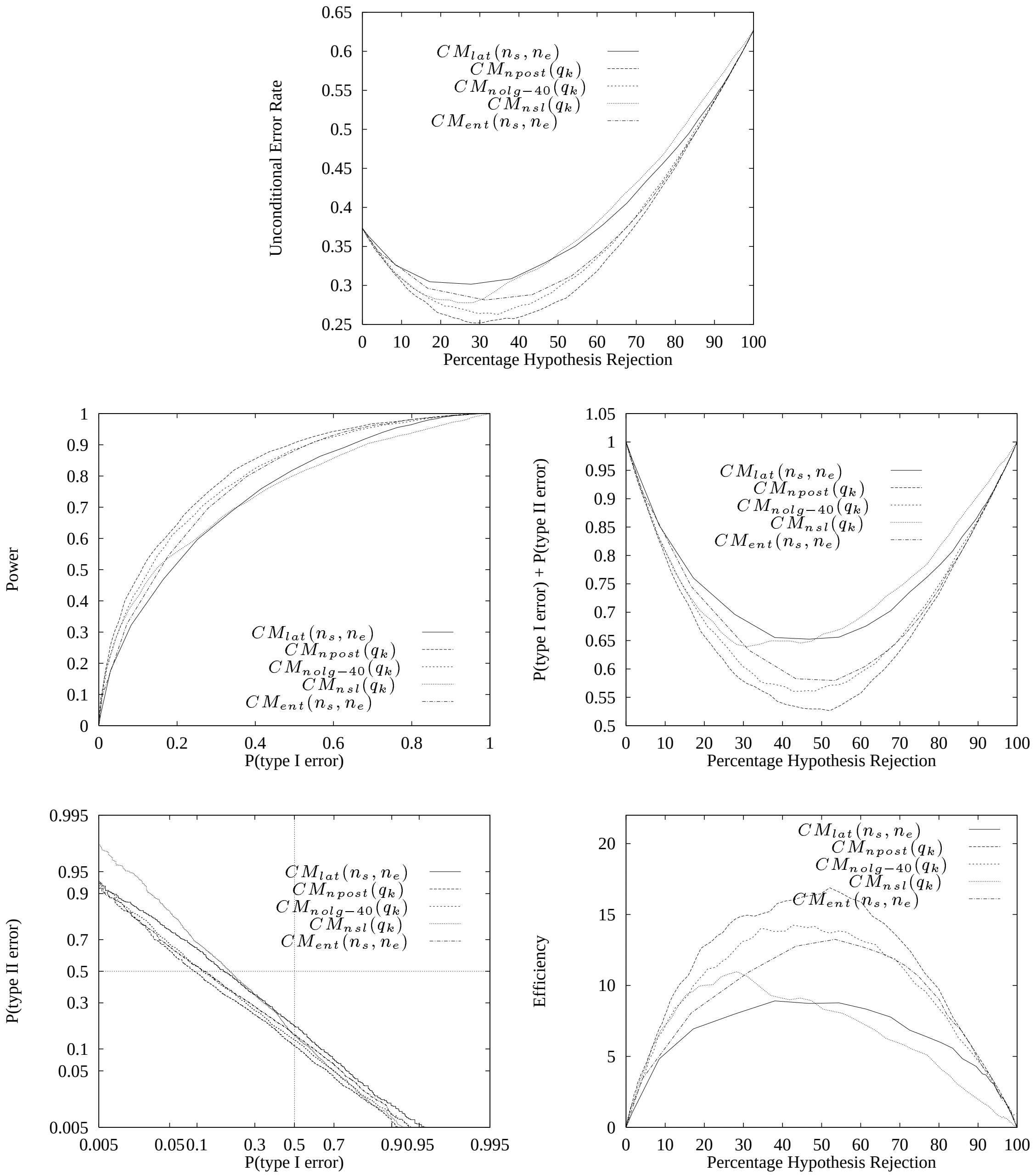
Confidence Measure	D_{Kol}	D_{Bhatt}	area under ROC curve	EER
$CM_{lat}(n_s, n_e)$	-0.1824	0.9441	0.6233	0.3385
$CM_{npost}(w_j)$	-0.4286	0.8678	0.7709	0.2801
$CM_{nolg-40}(w_j)$	-0.4331	0.8724	0.7655	0.2827
$CM_{nsl}(w_j)$	-0.4134	0.8725	0.7667	0.2847
$CM_{ent}(n_s, n_e)$	-0.4123	0.8772	0.7547	0.2942

Figure 4.7: Verification test performance for word level hypotheses (1820 correct and 1277 incorrect) from ABC Nightline, episode 05/23/96, decoded using word level decoding constraints.



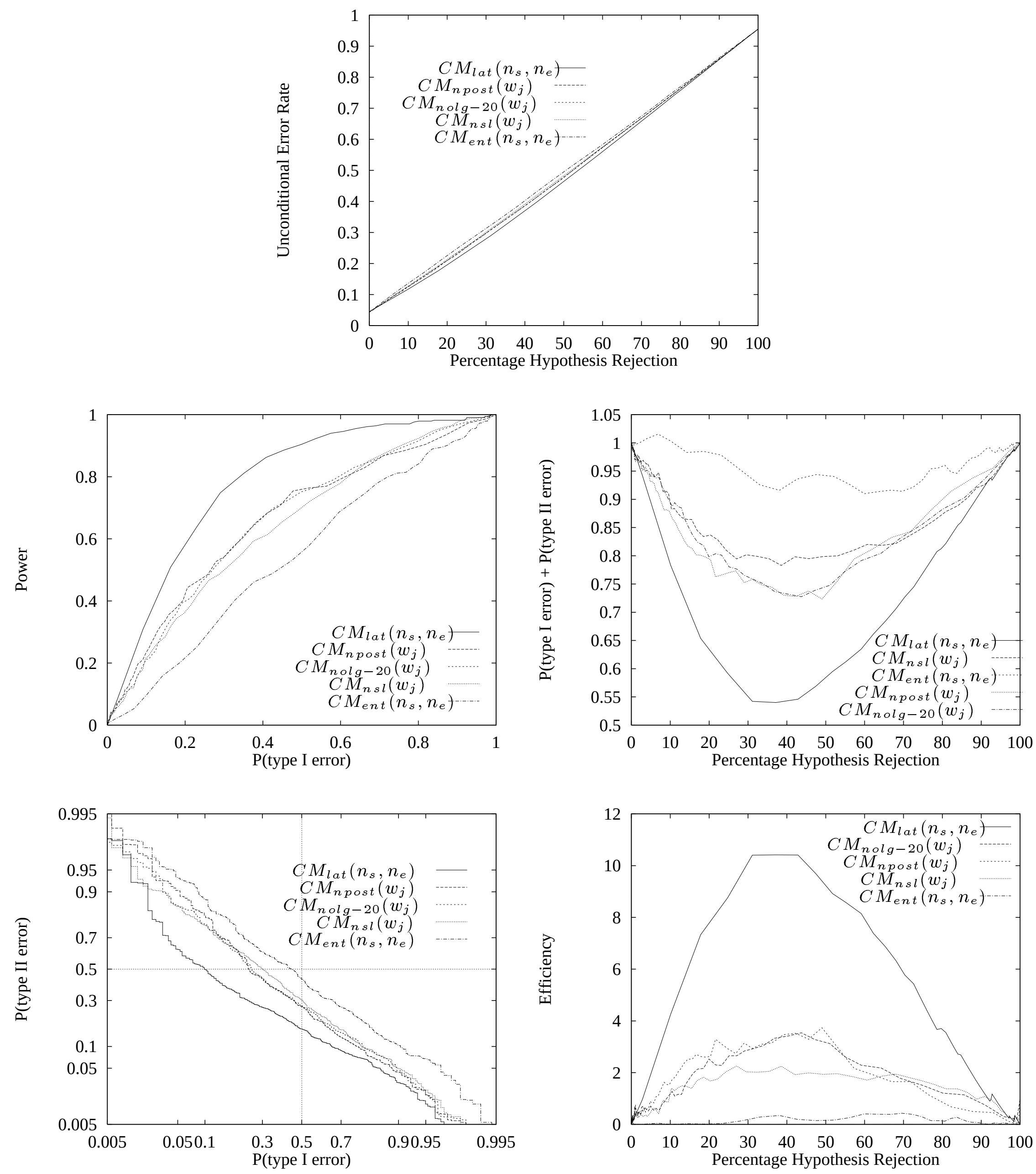
Confidence Measure	D_{Kol}	D_{Bhatt}	area under ROC curve	EER
$CM_{npost}(q_k)$	-0.4354	0.8550	0.7835	0.2796
$CM_{nolg-40}(q_k)$	-0.4108	0.8797	0.7621	0.2969
$CM_{nsl}(q_k)$	-0.4082	0.8835	0.7567	0.2947
$CM_{ent}(n_s, n_e)$	-0.2948	0.9332	0.7081	0.3174

Figure 4.8: Verification test performance for phone level hypotheses (8614 correct and 3237 incorrect) from ABC Nightline, episode 05/23/96, decoded using word level decoding constraints.



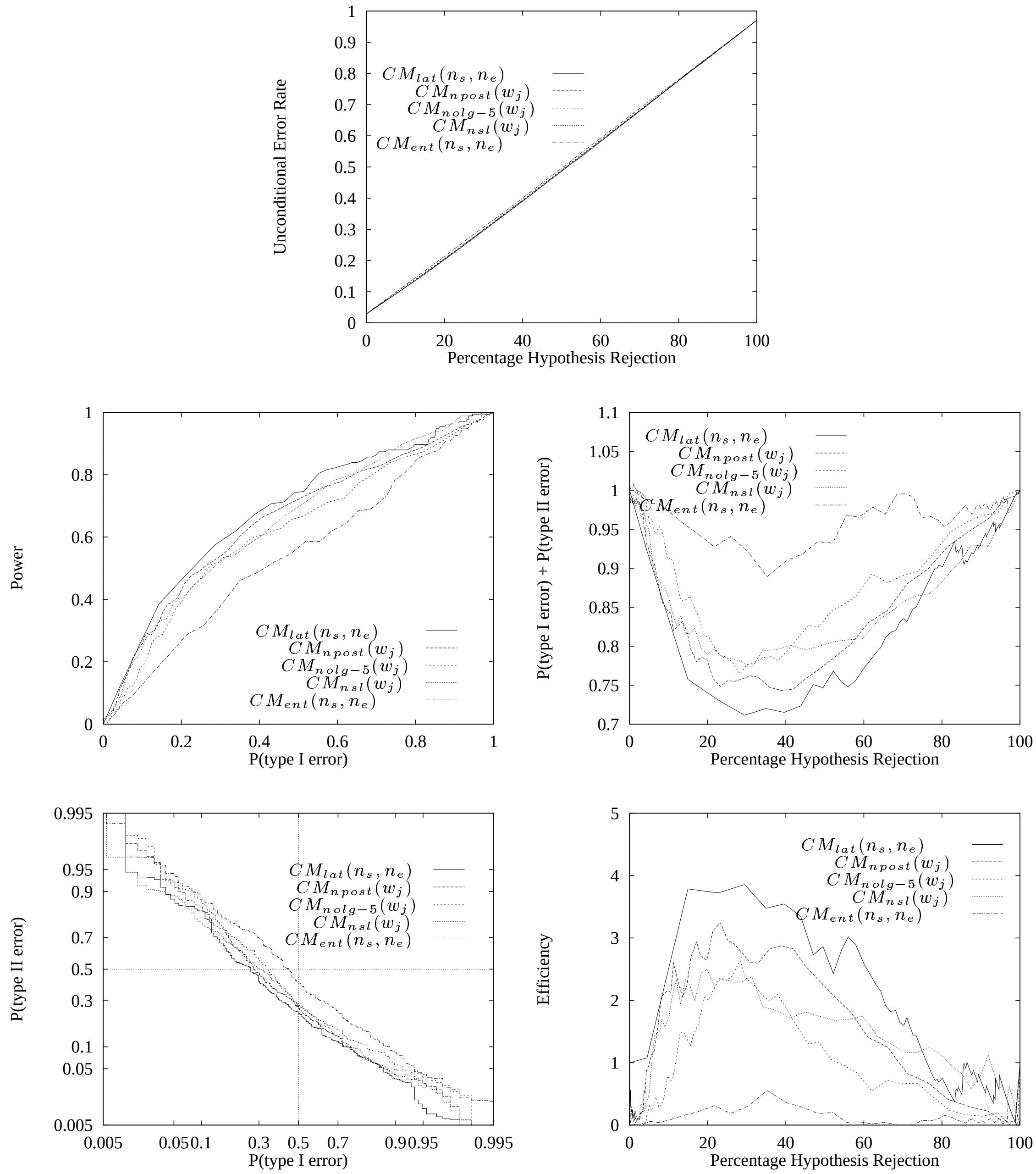
Confidence Measure	D_{Kol}	D_{Bhatt}	area under ROC curve	EER
$CM_{lat}(n_s, n_e)$	-0.3497	0.9174	0.7095	0.3261
$CM_{npost}(q_k)$	-0.4669	0.8256	0.8061	0.2653
$CM_{nolg-40}(q_k)$	-0.4339	0.8498	0.7874	0.2803
$CM_{nsl}(q_k)$	-0.3513	0.8960	0.7339	0.3229
$CM_{ent}(n_s, n_e)$	-0.3792	0.8942	0.7427	0.2914

Figure 4.9: Verification test performance for phone level hypotheses (6197 correct and 3689 incorrect) from ABC Nightline, episode 05/23/96, decoded using phone level decoding constraints.



Confidence Measure	D_{Kol}	D_{Bhatt}	area under ROC curve	EER
$CM_{lat}(n_s, n_e)$	-0.4617	0.8348	0.7535	0.2704
$CM_{npost}(w_j)$	-0.2660	0.9491	0.6528	0.3651
$CM_{nolg-20}(w_j)$	-0.2792	0.9487	0.6507	0.3684
$CM_{nsl}(w_j)$	-0.2150	0.9601	0.6344	0.3991
$CM_{ent}(n_s, n_e)$	-0.0916	0.9871	0.5273	0.4719

Figure 4.10: OOV word spotting performance (7119 IV and 334 OOV) for the Hub-1 '94 development test set.



Confidence Measure	D_{Kol}	D_{Bhatt}	area under ROC curve	EER
$CM_{lat}(n_s, n_e)$	-0.3129	0.9008	0.6549	0.3600
$CM_{npost}(w_j)$	-0.2539	0.9469	0.6426	0.3716
$CM_{nolg-5}(w_j)$	-0.2282	0.9554	0.6102	0.3969
$CM_{nsl}(w_j)$	-0.2188	0.9523	0.6387	0.4005
$CM_{ent}(n_s, n_e)$	-0.1403	0.9780	0.5291	0.4671

Figure 4.11: OOV word spotting performance (5855 IV and 174 OOV) for the Hub-3 '95 evaluation test set.

5. It can be noted that the size of n in the n-best required for optimal performance of CM_{olg} increases also increases as we move from the NAB data to the BN data. It can be speculated that this increase is brought about by a typically 'flatter' posterior probability distribution for a frame of BN data than for one of NAB data, where this flat distribution indicates that some sample acoustics from the BN corpus is typically harder to classify than that from the NAB corpus.

6. It can be seen that in the case of the NAB development test set the performance of the five confidence measures for the task of OOV word spotting is very similar to that for the decoding hypothesis verification task for the same data. The correlation in performance is not so strong in the case of the NAB evaluation test set, but the same general trend in relative confidence measure performance may be observed (this may be due to an insufficient number of OOV words in the evaluation set). This correlation together with the poor performance of $CM_{ent}(n_s, n_e)$ for the verification task using NAB data, supports the hypothesis that the majority of decoding errors on the NAB data are caused by OOV words. Additionally, the reduced number of OOV words and the improved performance of $CM_{ent}(n_s, n_e)$ for the BN data, indicates that the majority of decoding errors in this case are caused by degraded acoustics.

Chapter 5

Discussion

5.1 Marking Schemes

During the course of the experiments, it became clear that the marking of a decoding hypothesis as either correct or incorrect was not a trivial problem. Several marking algorithms were investigated and none produced an ideal result. In the case of the marking scheme described in [63], a deletion in the decoding sequence causes one of the decoding hypotheses adjacent to this deletion be marked as incorrect. It was found during the course of the experiments, however, that the vast majority of deletions were of small function words, such as "of" and "a". In this case, a typically long adjacent content word must be marked as incorrect despite the fact that the hypothesised model provides a good match to the majority of the acoustics against which it is aligned. The result of this is that a word which justifiably has a high confidence estimate, as on the whole the model match is good, is marked as incorrect and thus the distinguishing between correct and incorrect decoding hypotheses on the basis of their confidence becomes more difficult. One solution to this problem is to ignore deletions, the result of which is an increase in the performance of a confidence measure as a test statistic for the decoding hypothesis verification task.

5.2 Normalised Unconditional Error Rate

As it was stated in section 2.2.1, the unconditional error rate, UER, is a performance metric which does not reflect the *difficulty* of the hypothesis testing task, i.e. the relative frequencies of instances of the two relevant states of nature. In an attempt to generate a metric from the unconditional error rate which does reflect the difficulty of the task, a normalised version, NUER, is proposed in [63][1]. Assuming a Frequencist approach and further that $1 - P(\text{true}(H_0)) \geq P(\text{false}(H_0))$, NUER may be defined as:

$$NUER = \frac{[1 - P(\text{true}(H_0))] - P(\text{error})}{[1 - P(\text{true}(H_0))]} \tag{5.1}$$

In figure 5.1, the left hand panel shows the shape of the UER function over the possible ranges of task difficulty (entropy of the state of nature) and hypothesis test performance, whereas the right hand panel shows the shape of the NUER function over the same ranges. This figure shows that the shape of the NUER is uninformative over the majority of the space spanned by the function. The figure also shows that the value of the NUER function increases for improved hypothesis test performance, where it would be more desirable for an *error rate* function to decrease, given an improvement in performance.

5.3 Providing Insight

A reliable confidence measure that is comparable across utterances is useful from the perspective of improving the usability of an existing recogniser. A much more useful confidence measure, however, is one which is additionally capable of providing feedback about how well each of the components of a

[1] It should be noted that this unconditional error rate was called *classification error rate* in this paper.

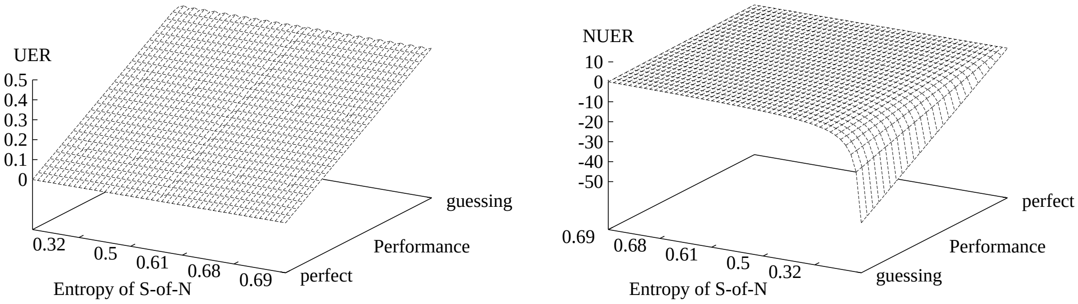


Figure 5.1: *Left*: Unconditional error rate over the range of possible values of the entropy of the state of nature and hypothesis test error rate. *Right*: A normalised unconditional error rate function over the same ranges of the entropy of the source and hypothesis test error rate.

system perform their matching tasks. In this way insight can be gained into to where both the system and also the confidence measure itself can be improved. Given a system, there is a temptation to look for a large set of confidence indicators from a host of different sources and to combine them to give a single confidence measure statistic, where the quality of the confidence measure is based purely upon its performance on, say, the decoding hypothesis verification task. The danger here, however, is that the confidence measure becomes somewhat ad hoc and that the theoretical well-foundedness required for confidence measures to provide insight is lost.

An example of an insight provided by the confidence measures investigated in this study is provided by the search for the cause of the residual error in the hypothesis tests. There are a number of possible causes for this residual error. A likely candidate, however, is the existence of poor pronunciation models in the system's lexicon. At the word level, it is clear that a poor pronunciation lexicon will result in the reduced confidence of a correct decoding hypothesis as the model will match the acoustics as well as it might. Indeed, poor pronunciation models could even lead to the hypothesis of incorrect models, if the erroneous models provide a better match to the acoustics than the correct models. The effect of poor pronunciation models also extends to the phone level. In this case low confidence phones may be marked as correct and vice versa, where this may come about as the reference phone sequence is created through a forced alignment of the reference word transcription to the acoustics, where the word models for the alignment are drawn from the pronunciation lexicon. All of these factors introduce noise into the hypothesis testing experiment and limit the performance that can be obtained using any given confidence measure.

It should be noted at this point that improved pronunciation models will also result in improved targets for acoustic model training, as these targets are derived from a forced alignment of the reference transcription to the acoustics, which requires the use of a pronunciation lexicon. Improved targets will clearly improve the quality of the probability estimates provided by the acoustic model. As these probabilities form the basis of the confidence measures described in this report, any improvement in pronunciation modelling will be propagated through the system and will result in improved confidence measures. This point illustrates that the components of an ASR system are heavily interdependent and that it is a mistake to consider them as otherwise. Improving an existing ASR system should therefore be considered as an iterative process, where an improvement in one component will often facilitate improvements in others.

Some initial investigations show that the acoustic confidence measure $CM_{post}(q_k)$ can be used to reliably evaluate pronunciation models. In this way poor pronunciation models may be highlighted and better models inserted into the lexicon. The left hand panel of figure 5.2 illustrates a poor pronunciation model which contains some constituent phones which provide a bad model match to portions of an acoustic realisation of the word “Funds”, whereas the right hand panel illustrates a good pronunciation model which provides a better fit to the same acoustics.

The use of acoustic confidence measures for pronunciation model evaluation opens up the whole area of pronunciation model adaptation. Through the use of confidence measures pronunciation models may be used to adapt pronunciation lexicons to particular dialects or speakers, where this approach to model adaptation is more theoretically pleasing and may well be more effective than attempting to adapt the

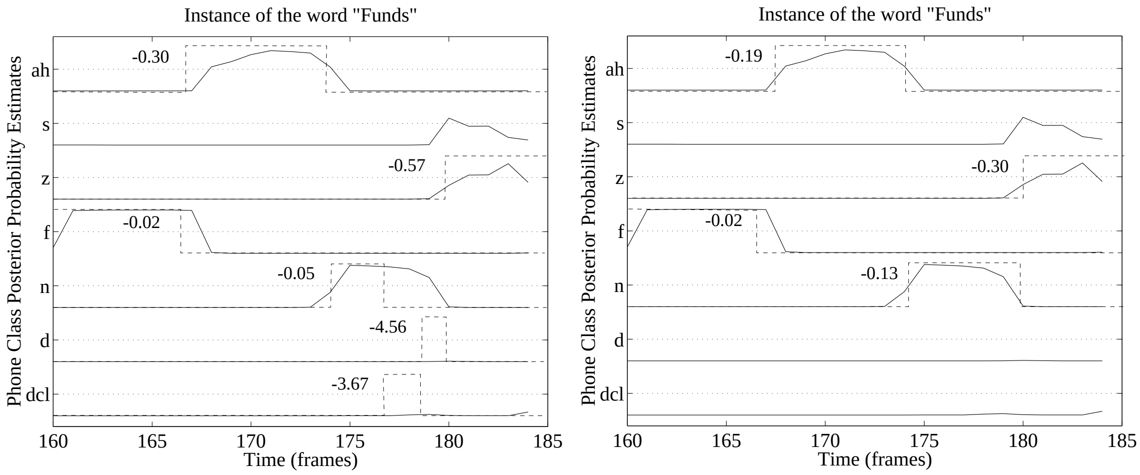


Figure 5.2: The relevant subset of acoustic model outputs generated by an instance of the word, "Funds", overlaid with timings and values of $CM_{npost}(q_k)$ obtained from a forced alignment of the pronunciation model /f ah n dcl d z/ (*Left*) and of the model /f ah n z/ (*Right*).

subword acoustic class models.

The application of acoustic confidence measures to pronunciation model evaluation highlights the distinction between acoustic and combined confidence measures. Whilst combined confidence measures can be expected to perform better than acoustic confidence measures, for tasks such as decoding hypothesis verification and OOV word spotting, as they bring more information to bear, there are clearly defined areas in acoustic confidence measures are superior, for example evaluating pronunciation models and filtering the input acoustics.

5.4 Summary

- So called 'hybrid HMM/ANN' based recognisers are well suited to the task of generating confidence measures as local phone class posterior probability estimates facilitate the computation of the posterior probability of a decoding hypothesis in a computationally efficient manner.
- Five confidence measures derived from the ABBOT hybrid HMM/ANN LVCSR system have been introduced. Experiments have been carried out using clean, read speech data and also data obtained from recordings of broadcast television and radio news shows, to investigate the performance of the confidence measures for utterance verification and OOV word spotting tasks.
- A diverse range of metrics were used to evaluate the performance of the confidence measures, where it was found that all the metrics investigated provided evaluations that were in agreement.
- At the word level, it was found that the combined confidence measure $CM_{lat}(n_s, n_e)$ provided the best performance for clean read speech, for both the verification and OOV word spotting tasks, suggesting that the majority of decoding errors in this case were caused by OOV words. In contrast, the acoustic confidence measure $CM_{npost}(w_j)$ provided the best performance for the broadcast news data, where this discrepancy is suspected to be caused by the reduction in information provided by the language model is reduced for the spontaneous speech found in the broadcast news data.
- Acoustic confidence measures have a number of uses which are not available to combined confidence measures:
  - Acoustic confidence measures can be used evaluate pronunciation models and so can be used to spot poor pronunciation models and guide the search for improved models. The superior performance of $CM_{npost}(q_k)$ at the phone level for both data types and the reduction in residual error for all confidence measures at the phone level provided an insight into this use for acoustic confidence measures.

– Computationally efficient confidence measures can be used to filter the acoustics input to a recogniser prior to the decoding stage. A general measure of acoustic model match, such as $CM_{ent}(n_s, n_e)$, can be used to filter out portions of acoustics that are not well modelled by the recogniser. Such a filter can reduce both WER and decoding time.

Investigations of both these uses for acoustic confidence measures are currently in progress.

Bibliography

[1] A. Asadi, R. Schwartz, and J. Makhoul. Automatic detection of new words in a large vocabulary continuous speech recognition system. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 125–128. IEEE, 1990.

[2] A. Asadi, R. Schwatrz, and J. Makoul. Automatic modeling for adding new words to a large-vocabulary continuous speech recognition system. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 305–308. IEEE, 1991.

[3] J. Barker, G. Williams, and S. Renals. Acoustic confidence measures for segmenting broadcast news. November 1997.

[4] C.M. Bishop. *Neural Networks for Pattern recognition.* Oxford University Press, Walton Street, Oxford, 1995.

[5] J-M. Boite, H. Bourlard, B. D'hoore, and M. Haesen. A new approach towards keyword spotting. In *Proceedings of EUROSPEECH*, pages 1273–1276. ESCA, 1993.

[6] H. Bourlard, B. D'hoore, and J-M. Boite. Optimizing recognition and rjection performance in wordspotting systems. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 373–376. IEEE, 1994.

[7] H.A. Bourlard, Y. Konig, and N. Morgan. REMAP: Recursive estimation and maximisation of posterior probabilities. Technical Report TR-94-064, International Computer Science Institute (ICSI), EECS Department, University of California, Berkeley, Ca, March 1995.

[8] H.A. Bourlard and N. Morgan. *Connectionist Speech Recognition: A Hybrid Approach.* Kluwer Academic Publishers, Norwell, Massachusetts, 1994.

[9] J. Caminero, C. de la Torre, L. Villarrubia, C. Matín, and Hernández. On-line garbage modeling with discriminant analysis for utterance verification. In *Proceedings of the International Conference on Spoken Language Processing*, 1996.

[10] T.M. Cover and J.A. Thomas. *Elements of Information Theory.* Wiley Series in Telecommunications. Wiley-Interscience, New York, 1991.

[11] S. Cox and R. Rose. Confidence measures for the switchboard database. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 511–515. IEEE, 1996.

[12] J.R. Deller, J.G. Proakis, and Hansen J.H.L. *Discrete-Time Processing of Speech Signals.* Macmillan, 1993.

[13] R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis.* John Wiley & Sons, Inc., 1973.

[14] J.P. Egan. *Signal Detection Theory and ROC Analysis.* Academic Press, 1975.

[15] E. Eide, H. Gish, P. Jeanrenaud, and A. Mielke. Understanding and improving speech recognition performance through the use of diagnostic tools. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 221–224. IEEE, 1995.

[16] R. El Méliani and D. O'Shaughnessy. Lexical fillers for task-independent-training based keyword spotting and detection of new words. In *Proceedings of EUROSPEECH*, pages 2129–2132. ESCA, 1995.

[17] R. El Méliani and D. O'Shaughnessy. New efficient fillers for unlimited word recognition and keyword spotting. In *Proceedings of the International Conference on Speech and Language Processing*, 1996.

[18] R. El Méliani and D. O'Shaughnessy. Accurate keyword spotting using strictly lexical fillers. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 907–910. IEEE, 1997.

[19] P. Fetter, F. Dandurana, and P. Regel-Brietzmann. Word graph rescoring using confidence measures. In *Proceedings of the International Conference on Spoken Language Processing*, 1996.

[20] P. Fetter et al. Detection of unknown words in spontaneous speech. In *Proceedings of EUROSPEECH*, pages 1637–1640. ESCA, 1995.

[21] P. Fetter, A. Kaltenmeier, T. Kuhn, and Regel-Brietzmann. Improved modeling of OOV words in spontaneous speech. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 534–538. IEEE, 1996.

[22] F. Gallwitz, E. Nöth, and H. Niermann. A category based approach for recognition of out-of-vocabulary words. In *Proceedings of the International Conference on Spoken Language Processing*, 1996.

[23] P.H. Garthwaite, I.T. Jolliffe, and B. Jones. *Statistical Inference.* Prentice-Hall, 1995.

[24] L. Gillick, Y. Ito, and J. Young. A probabilistic approach to confidence estimation and evaluation. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 879–883. IEEE, 1997.

[25] R. Haeb-Umbach, E. Beyerlein, and E. Thelen. Automatic transcription of unknown words in a speech recognition system. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 840–843. IEEE, 1995.

[26] D.J. Hand. *Construction and Assessment of Classification Rules.* Wiley Series in Probability and Statistics. John Wiley & Sons Ltd., 1997.

[27] I.L. Hetherington. *A Characterisation of the Problem of New, Out-of-Vocabulary Words in Continuous-Speech Recognition and Understanding.* PhD thesis, Massachusetts Institute of Technology, 1994.

[28] L. Hetherington. New words: Effect on recognition performance and incorporation issues. In *Proceedings of EUROSPEECH*, pages 1645–1648. ESCA, 1995.

[29] M. Jeanrenaud, P.and Siu, J. R. Rohlicek, M. Meteer, and H. Gish. Spotting events in continuous speech. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 381–384. IEEE, 1994.

[30] P. Jeanrenaud, K. Ng, J.R. Siu, M. Rohlicek, and H. Gish. Phonetic-based word spotter: Various configurations and application to event spotting. In *Proceedings of EUROSPEECH*, pages 1057–1060. ESCA, 1993.

[31] T. Kemp and A. Jusek. Modelling unknown words in spontaneous speech. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 530–534. IEEE, 1996.

[32] H. Klemm, F. Class, and U. Kilian. Word- and phrase spotting with syllable-based gaebage modelling. In *Proceedings of EUROSPEECH*, pages 2157–2160. ESCA, 1995.

[33] Y. Konig. *REMAP: Recursive Estimation and Maximisation of A Posteriori Probabilities in Transition-based Speech Recognition.* PhD thesis, University of California at Berkeley, 1996.

[34] Y. Konig, H. Bourlard, and N. Morgan. REMAP - experiments with speech recognition. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 3351–3355. IEEE, 1996.

[35] Y. Konig, H.A. Bourlard, and N. Morgan. Recursive estimation and maximisation of a posteriori probabilities - application to transition-based connectionist speech recognition. In *Proceedings of the Conference on Neural Information Processing Systems '95*, 1995.

[36] R.P. Lippmann, E.I. Chang, and C.R. Jankowski. Wordspotter training using figure-of-merit back propagation. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 389–392. IEEE, 1994.

[37] E. Lleida, J.B. Mariño, J. Salavedra, A. Bonafonte, E. Monte, and A. Martínez. Out-of-vocabulary word modelling and rejection for keyword spotting. In *Proceedings of EUROSPEECH*, pages 1265–1268. ESCA, 1993.

[38] E. Lleida and R.C. Rose. Efficient decoding and training procedures for utterance verification in continuous speech recognition. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*. IEEE, 1996.

[39] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Pryzybocki. The DET curve in assessment of detection tak performance. In *Proceedings of EUROSPEECH*, pages 1895–1898. ESCA, 1997.

[40] B. Mazor and M-W Feng. Improved a-posteriori processing for keyword spotting. In *Proceedings of EUROSPEECH*, pages 2231–2234. ESCA, 1993.

[41] C. Neti, S. Roukos, and E. Eide. Word-based confidence measures as a guide for stack search in speech recognition. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 883–887. IEEE, 1997.

[42] L.R. Rabiner and B-H. Juang. *Fundamentals of Speech Recognition*. Prentice-Hall, 1993.

[43] L.R. Rabiner and B.H. Juang. An introduction to hidden markov models. *IEEE ASSP Magazine*, 3(1):4–16, January 1986.

[44] M.G. Rahim, C-H. Lee, and B-H. Juang. Robust utterance verification for connected digits recognition. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 285–288. IEEE, 1995.

[45] S. Renals and M.M. Hochberg. Efficient evaluation of the LVCSR search space using the NOWAY decoder. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*. IEEE, 1996.

[46] S.J. Renals and M.M. Hochberg. Decoder technology for connectionist large vocabulary speech recognition. Technical Report CS-95-17, University of Sheffield, 1995.

[47] C. Ris, J. Hennebert, H. Bourlard, S. Renals, and N. Morgan. Estimation of global posteriors and forward-backward training of hybrid HMM/ANN systems. In *Proceedings of EUROSPEECH*, page Submitted. ESCA, 1997.

[48] Z. Rivlin. A confidence measure for acoustic likelihood scores. In *Proceedings of EUROSPEECH*, pages 523–526. ESCA, September 1995.

[49] Z. Rivlin, M. Cohen, V. Abrash, and T. Chung. A phone-dependent confidence measure for utterance rejection. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 515–518. IEEE, 1996.

[50] A.J. Robinson, M.M Hochberg, and S.J. Renals. The use of recurrent networks in continuous speech recognition. In C-H. Lee, F.K. Soong, and K.K. Paliwal, editors, *Automatic Speech and Speaker Recognition*, Engineering and Computer Science, chapter 10, pages 233–258. Kluwer Academic Publishers, 101 Philip Drive, Assinippi Park, Norwell Ma., 1996.

[51] J.R. Rohlicek, W. Russell, S. Roukos, and H. Gish. Continuous hidden markov modeling for speaker-independent word spotting. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 627–630. IEEE, 1989.

[52] R.C. Rose. Discriminant wordspotting techniques for rejecting non-vocabulary utterances in unconstrained speech. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 105–108. IEEE, 1992.

[53] R.C. Rose. Keyword detection in conversational speech utterances using hidden markov model based continuous speech recognition. *Computer Speech and Language*, 9:309–333, 1995.

[54] R.C. Rose. Word spotting from continuous speech utterances. In C-H. Lee, F.K. Soong, and K.K. Paliwal, editors, *Automatic Speech and Speaker Recognition*, chapter 13, pages 303–329. Kluwer Academic Publishers, 1996.

[55] R.C. Rose and D.B. Paul. A hidden markov model based keyword recognition system. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 129–132. IEEE, 1990.

[56] T. Schaaf and T. Kemp. Confidence measures for spontaneous speech recognition. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 875–878. IEEE, 1997.

[57] A.R. Setlur, R.A. Sukkar, and J. Jacob. Correcting recognition errors via discriminative utterance verification. In *Proceedings of the International Conference on Spoken Language Processing*, 1996.

[58] NIST Speech Software. http://www.nist.gov/speech/software.htm. URL, 1998.

[59] B. Suhm, M. Woszczyna, and A. Waibel. Detection and transcription of new words. In *Proceedings of EUROSPEECH*, pages 2179–2182. ESCA, 1993.

[60] R.A. Sukkar and C-H. Lee. Vocabulary independent discriminative utterance verification for non-keyword rejection in subword based speech recognition. In *IEEE Transactions on Speech and Audio Processing*, volume 4, pages 420–429. IEEE, 1996.

[61] R.A. Sukkar, A.R. Setlur, M.G. Rahim, and C-H. Lee. Utterance verification of keyword strings using word-based minimum verification error (WB-MVE) training. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*. IEEE, 1996.

[62] M. Weintraub. Keyword-spotting using sri's decipher large-vocabulary speech-recognition system. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 463–466. IEEE, 1993.

[63] M. Weintraub, F. Beaufays, Z. Rivlin, Y. Konig, and A. Stolcke. Neural - network based measures of confidence for word recognition. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 887–890. IEEE, 1997.

[64] J.G. Wilpon, L.R. Rabiner, Lee C-H., and E.R. Goldman. Automatic recognition of keywords in unconstrained speech using hidden markov models. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(11):1870–1878, November 1990.

[65] M.H. Zweig and G. Cambell. Reciever-operating characteristic (roc) plots: A fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, 39(4):551–577, 1993.