CONFIDENCE MEASURES FROM LOCAL POSTERIOR PROBABILITY ESTIMATES

Gethin Williams and Steve Renals

Department of Computer Science
University of Sheffield
Sheffield S1 4DP UK
{g.williams,s.renals}@dcs.shef.ac.uk

Abstract

In this paper we introduce a set of related confidence measures for large vocabulary continuous speech recognition (LVCSR) based on local phone posterior probability estimates output by an acceptor HMM acoustic model. In addition to their computational efficiency, these confidence measures are attractive as they may be applied at the state-, phone-, word- or utterance-levels, potentially enabling discrimination between different causes of low confidence recognizer output, such as unclear acoustics or mismatched pronunciation models. We have evaluated these confidence measures for utterance verification using a number of different metrics. Experiments reveal several trends in 'profitability of rejection', as measured by the unconditional error rate of a hypothesis test. These trends suggest that crude pronunciation models can mask the relatively subtle reductions in confidence caused by out-of-vocabulary (OOV) words and disfluencies, but not the gross model mismatches elicited by non-speech sounds. The observation that a purely acoustic confidence measure can provide improved performance over a measure based upon both acoustic and language model information for data drawn from the Broadcast News corpus, but not for data drawn from the North American Business News corpus suggests that the quality of model fit offered by a trigram language model is reduced for Broadcast News data. We also argue that acoustic confidence measures may be used to inform the search for improved pronunciation models.

1 Introduction

A reliable measure of confidence in the output of a speech recognizer is useful under many circumstances. For example, a low degree of confidence should be assigned to the outputs of a recognizer presented with an out-of-vocabulary (OOV) word or some unclear acoustics, caused by noise or a high level of background music. Both OOV words and unclear acoustics are a major source of recognizer error and may be detected by employing a confidence measure as a test statistic in a statistical hypothesis test.

Our approach to generating confidence measures is based upon local estimates of phone posterior probabilities produced by a hidden Markov model/artificial neural network (HMM/ANN) system. We refer to such models as *acceptor HMMs*, to contrast with the generative modelling approach adopted in most HMM systems (section 3). If phone posterior probability estimates constitute a suitable basis for confidence measures (see e.g. Williams and Renals (1997)) then it is apparent that acceptor HMMs which produce local estimates of these values directly are well suited to producing computationally efficient measures of confidence. The computational efficiency of the measures arises as their computation requires little more than the forward pass of some acoustic observations through a suitably trained phone classifier. In this paper, a set of related confidence measures are introduced. These confidence measures are purely acoustic: they are based on the acceptor HMM acoustic model and do not require the incorporation of language modelling constraints. For comparison we also apply a 'combined' confidence measure derived from both the acoustic and language models.

The confidence measures have been applied to the output of the ABBOT large vocabulary continuous speech recognition (LVCSR) system (Robinson et al., 1996) for the task of utterance verification at the word- and phone-levels. Several probabilistic metrics were used for evaluation. In addition to their computational efficiency, an attractive property of these acoustic confidence measures is their simple and explicit links to the underlying acoustic model, allowing them to be used to extract more subtle information related to the acoustic confidences. For example, it is of interest to develop confidence measures that are able to discriminate between low confidence due to a mismatched pronunciation model (potentially arising from the occurrence of an OOV word) and low confidence owing to unclear acoustics.

The remainder of the paper is structured as follows. Section 2 is concerned with statistical hypothesis testing and methods for evaluating such tests based on conditional error probabilities, information theory and distributional separability. Section 3 introduces the statistical models of speech used in this paper and distinguishes between generative and acceptor modelling. Section 4 defines a set of confidence measures derived from the acceptor HMM acoustic model which may be used as a test statistic in a hypothesis test. Section 5 describes a set of utterance verification experiments using two large vocabulary continuous speech recognition databases: North American Business News and Broadcast News.

		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 1: A confusion matrix recording the actions resulting from a classical hypothesis test against the corresponding states of nature.

2 Statistical Hypothesis Testing

A series of applications of a classical hypothesis test may be summarized in a 2×2 confusion matrix such as that given in figure 1. From the figure it is apparent that there are two types of error:

- *Type I errors* occur when the null hypothesis (H_0) is true but rejected;
- *Type II errors* occur when H_0 is false but accepted.

Empirical probability estimates $\hat{P}$ calculated from the 2×2 confusion matrix may be used to evaluate a test statistic (confidence measure). The simplest of these metrics is the *unconditional error rate* (UER) $P(\text{error})$ which may be estimated by:

$$\hat{P}(\text{error}) = \frac{N(\text{reject}(H_0), \text{true}(H_0)) + N(\text{accept}(H_0), \text{false}(H_0))}{N(H)} , \tag{1}$$

where $N(H)$ is the total number of hypotheses tested.

The probability of error conditioned on a particular state of nature may be similarly estimated:

$$\hat{P}(\text{type I error}) \; = \; \hat{P}(\text{reject}(H_0)|\text{true}(H_0)) \; = \; \frac{N(\text{reject}(H_0), \text{true}(H_0))}{N(\text{true}(H_0))} . \tag{2}$$

$$\hat{P}(\text{type II error}) \; = \; \hat{P}(\text{accept}(H_0)|\text{false}(H_0)) \; = \; \frac{N(\text{accept}(H_0), \text{false}(H_0))}{N(\text{false}(H_0))} . \tag{3}$$

Conditional and unconditional error rate statistics are complimentary. Consider an utterance verification task for which H_0 (representing a prior assumption regarding the domain under consideration) is set to the hypothesis that a given recognizer output is correct. If the task is applied to an isolated digit recognizer, then the percentage of recognition errors is likely to be very small. In this case, a test which simply accepts H_0 for each recognizer output will yield a low UER. However if the task of interest is more difficult (e.g. spontaneous speech recognition) resulting in a much higher word error rate from the recognizer, then this simplistic strategy will yield a much increased UER. In contrast, the conditional error rates can be used to evaluate the performance of a test independent of the prior probabilities of the two states of nature (true(H_0) and false(H_0)).

One method for plotting conditional probability statistics for a hypothesis test is to use an ROC (Receiver Operating Characteristic) curve (Egan, 1975). Such a curve is created by plotting the 'hit' rates (ordinates) against the 'false alarm' rates (abscissas) over the range of possible operating points on the test statistic. For example, an ROC curve may be obtained by plotting $P(\text{accept}(H_0)|\text{true}(H_0))$ against $P(\text{type II error})$ or (equivalently) by plotting $P(\text{reject}(H_0)|\text{false}(H_0))$ against $P(\text{type I error})$.

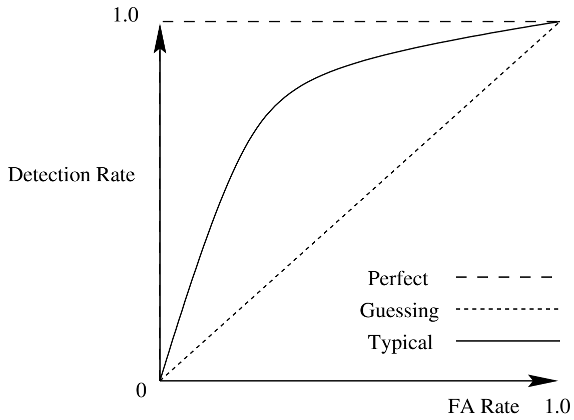


Figure 2: A schematic illustration of ROC plots for perfect, guessing and typical hypothesis tests.

$P(\text{reject}(H_0)|\text{false}(H_0))$ is called the *power* of a hypothesis test. ROC curves for perfect, "guessing" and typical hypothesis tests are schematically illustrated in figure 2.

Figure 3 illustrates the distributions of the values of the test statistic (confidence measure) $\text{nPP}(w_j)$ (see section 4) conditioned upon $\text{true}(H_0)$ and $\text{false}(H_0)$ respectively. If these distributions are assumed to be Gaussian then the probabilities $P(\text{type I error})$ and $P(\text{type II error})$ may be plotted over the range of operating points of a test as a detection error tradeoff (DET) curve (Martin et al., 1997). In this case, the axes are warped according to the deviations of the tails, corresponding to the probabilities, from the mean of the Gaussian. This logarithmic warping of the axes has the effect of accentuating any differences between well performing test statistics, clustered in the lower left quadrant of the plot of $P(\text{type I error})$ against $P(\text{type II error})$.

An alternative approach to evaluating a hypothesis test regards the state of nature (hypothesis true or false) and the action resulting from the test (accept or reject) as binary random variables, denoted Z and A respectively, and computes the mutual information $I(Z;A)$ between them:

$$I(Z;A) = H(Z) - H(Z|A) = H(A) - H(A|Z)\ , \tag{4}$$

where $H(\cdot)$ and $H(\cdot|\cdot)$ denote the entropy of a random variable and the conditional entropy of a random variable given the value of another, respectively. $H(Z)$ measures the uncertainty in the state of nature, reflecting the difficulty of the hypothesis testing task. As before, empirical probability estimates obtained from a 2×2 confusion matrix may be used to compute the value of the metric.

By normalizing $I(Z;A)$ by $H(A)$, equal values can be obtained for a particular level of hypothesis testing performance, irrespective of the task difficulty. This normalized mutual information metric $E(Z;A)$ is known as the *efficiency* (Cox and Rose, 1996) of a test:

$$E(Z;A) = \frac{I(Z;A)}{H(A)} = \frac{H(A) - H(A|Z)}{H(A)} = \frac{H(Z) - H(Z|A)}{H(A)}\ . \tag{5}$$

The above evaluation metrics result in a set of curves covering a range of operating points for a hypothesis test. To obtain a scalar valued evaluation metric either a particular operating point can be chosen, or it is possible to integrate over a range of operating points. An example of the first approach is the equal error rate (EER) condition which specifies the point when $P(\text{type I error})$ and $P(\text{type II error})$ are equal. The area under the ROC curve is an example of the second approach and has a well-defined statistical interpretation. If the detection and false alarm rates both cover the range $[0,1]$, then the area has a value equal to 1.0 for a perfect hypothesis test and has a value equal to 0.5 for a guessing test.[1] In the case where

[1] This area is equivalent to the value of the Mann-Whitney version of the non-parametric two-sample statistic (Zweig and Cambell., 1993).

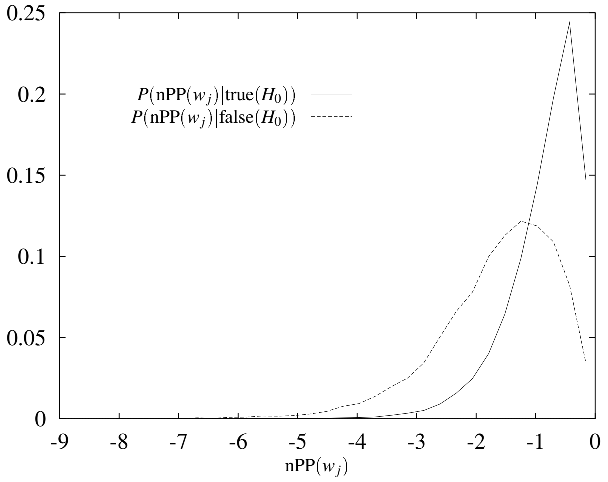


Figure 3: Distributions of values of the test statistic $nPP(w_j)$ (see section 4) conditioned on the two states of nature (true(H_0) and false(H_0)) computed for a word-level decoding of some Broadcast News data.

a high test statistic value is indicative of true(H_0), the value of the area is equal to the probability that a 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).

A third method for obtaining a scalar valued evaluation metric is an estimation of the separability of the test statistic (confidence measure) distributions conditioned on the two states of nature (figure 3). An ideal test statistic will yield distributions which are completely separable, where such distributions would facilitate perfect hypothesis testing. The separability of the two distributions may be estimated using a number of metrics, two of which are the *Kolmogorov Variational Distance*, d_{Kol}, and the *Bhattacharyya Distance*, d_{Bhatt} (Hand, 1997). Another is the 'symmetric Kullback-Leibler distance', d_{KL2}. Since the Kullback-Leibler distance between two distributions A and B is asymmetric, the symmetric version sums the divergence of the distributions measured in both directions.[2] We define these for the nPP(w_j) test statistic (introduced in section 4):

$$
\begin{aligned}
d_{\text{Kol}} &= -\sum_{m=1}^{M} \left| \frac{p(\text{nPP}(w_j^m)|\text{true}(H_0)) - p(\text{nPP}(w_j^m)|\text{false}(H_0))}{2} \right| , && (6) \\
d_{\text{Bhatt}} &= \sum_{m=1}^{M} \sqrt{p(\text{nPP}(w_j^m)|\text{true}(H_0)) \; p(\text{nPP}(w_j^m)|\text{false}(H_0))} , && (7) \\
d_{\text{KL2}} &= -\sum_{m=1}^{M} p(\text{nPP}(w_j^m)|\text{true}(H_0)) \log \left\{ \frac{p(\text{nPP}(w_j^m)|\text{false}(H_0))}{p(\text{nPP}(w_j^m)|\text{true}(H_0))} \right\} \\
&\quad -\sum_{m=1}^{M} p(\text{nPP}(w_j^m)|\text{false}(H_0)) \log \left\{ \frac{p(\text{nPP}(w_j^m)|\text{true}(H_0))}{p(\text{nPP}(w_j^m)|\text{false}(H_0))} \right\} , && (8)
\end{aligned}
$$

where nPP(w_j^m) denotes the value of the confidence measure for the mth word decoding under consideration.

[2] As data points with values of zero are problematic for this metric, any terms of (8) containing zero probability values were ignored during the computation of the metric.

3 Generative HMMs and Acceptor HMMs

A complete probability model provides a probability distribution for all variables in the system. In the case of speech recognition, such a model will provide a joint distribution over the word sequence model M and the acoustics X, with parameters Θ, $P(M,X;\Theta)$. This is usually decomposed into an acoustic model term $P(X|M;\Theta)$, and a prior language model term $P(M;\Theta)$, which is independent of the acoustics:

$$P(M,X;\Theta) = P(X|M;\Theta)P(M;\Theta)\ . \tag{9}$$

This is a *generative model* since the acoustic model specifies a probability distribution over acoustic vectors X generated by M. The parameters of this model are usually optimized according to a maximum likelihood criterion. If an HMM is used for the acoustic model, then it is convenient to perform the parameter optimization using the EM algorithm.

At recognition time we are concerned with finding the most probable model M^* to account for the observed data:

$$M^* = \underset{M}{\operatorname{argmax}} P(M|X;\Theta) = \underset{M}{\operatorname{argmax}} \frac{P(X|M;\Theta)P(M;\Theta)}{P(X;\Theta)}\ . \tag{10}$$

Since $P(X;\Theta)$ is independent of M, the denominator may be ignored for the purposes of finding M^*. The remaining numerator is identical to the right hand side of (9) and so may be estimated using a generative model. Discarding the estimation of $P(X;\Theta)$ also facilitates substantial computational savings. If the "correct" model is in the space of models under investigation, then a generative model will result in an optimally performing system (providing certain other conditions are met) (Bahl et al., 1986). Since the correct model is not known for speech recognition, and M^* depends on the posterior $P(M|X;\Theta)$, recognition accuracy could be maximized by optimizing a criterion directly related to the posterior. An example of this is the maximum mutual information (MMI) criterion (Bahl et al., 1986) for generative HMMs.

Acceptor HMMs are based on the observation that an estimate of $P(X)$ is not required to directly optimize $P(M|X;\Theta)$, using the following factorization:

$$\begin{aligned} P(M|X;\Theta) &= \sum_{Q_M} P(M,q^1,\dots,q^N|X;\Theta) && (11)\\ &= \sum_{Q_M} P(q^1,\dots,q^N|X;\Theta)P(M|q^1,\dots,q^N,X;\Theta)\ , && (12) \end{aligned}$$

where Q_M is the set of all possible state sequences $\{q^1 \dots q^N\}$ through model M, and $X = \{x^1 \dots x^N\}$ is the acoustic data. This does not require a generative model, but rather the posterior probability of each possible state sequence given the acoustic data. We refer to these as acceptor HMMs since they may be regarded as (stochastic) finite state acceptors; the usual generative HMM may be regarded as a (stochastic) finite state generator.

The first term on the right hand side of (12) may be expressed as a product of conditional probabilities and further simplified assuming a first order Markov process:

$$\begin{aligned} P(q^1,\dots,q^N|X;\Theta) &= \prod_{n=1}^{N} P(q^n|q^1,\dots,q^{n-1},X;\Theta) && (13)\\ &\simeq \prod_{n=1}^{N} P(q^n|q^{n-1},X;\Theta)\ . && (14) \end{aligned}$$

This acoustic model probability can be estimated by an artificial neural network such as a multilayer perceptron (Bourlard and Morgan, 1994) or a recurrent network (Robinson, 1994), making an assumption about the dependence on the acoustic input. In the case of the recurrent network used in this work, we assume no dependence on the previous state or future acoustics (Robinson et al., 1996):

$$P(q^1,\dots,q^N|X;\Theta) = \prod_{n=1}^{N} P(q^n|X_1^n;\Theta) \ . \tag{15}$$

For both network architectures, the acoustic model has traditionally been trained as a (context dependent or independent) phone classifier. The estimated probability distribution for the kth phone class q_k is then tied across all states of the corresponding HMM. As the observation distributions are identical, multiple state phone HMMs serve only to provide durational constraints in this case.

Although (15) is a zeroth order Markov process, the overall system is still first order Markov. This is reflected in the prior (or language model) term which appears as the second term on the right hand side of (12); assuming that the probability of M is conditionally independent from X given the state sequence $\{q^1,\dots,q^N\}$:

$$\begin{aligned} P(M|q^1,\dots,q^N;\Theta) &= \frac{P(q^1,\dots,q^N|M;\Theta)P(M;\Theta)}{P(q^1,\dots,q^N;\Theta)} && (16) \\ &\simeq \prod_{n=1}^{N}\left[\frac{P(q^n|q^{n-1},M;\Theta)}{\sum_{M'} P(q^n|q^{n-1},M';\Theta)P(M';\Theta)}\right] P(M;\Theta) \ . && (17) \end{aligned}$$

$P(q^n|q^{n-1},M;\Theta)$ may be regarded as a prior specified by the model (pronunciations and language model); the denominator of (17) is this prior summed over all models and amounts to $P(q^n|q^{n-1};\Theta)$. This summation is difficult to perform but a sample estimate of its value $P(q^n|q^{n-1})$ may be calculated using the relative frequencies of the phone labels in the acoustic training data. Combining (12), (15) and (17) we have:

$$P(M|X;\Theta) \simeq \sum_{Q_M}\prod_{n=1}^{N}\left[P(q^n|X_1^n;\Theta)\frac{P(q^n|q^{n-1},M;\Theta)}{P(q^n|q^{n-1};\Theta)}\right] P(M;\Theta) \ . \tag{18}$$

The ABBOT recurrent network-based system makes use of a further zeroth-order Markov assumption in the denominator. The Viterbi approximation to the full model probability is also made during decoding:

$$P(M|X;\Theta) \simeq \max_{Q_M}\prod_{n=1}^{N}\left[P(q^n|X_1^n;\Theta)\frac{P(q^n|q^{n-1},M;\Theta)}{P(q^n;\Theta)}\right] P(M;\Theta) \ . \tag{19}$$

The direct optimization of such a model (Bengio et al., 1992) is a computationally expensive process. However, using similar factorizations and assumptions as above, Bourlard et al. (1996) and Hennebert et al. (1997) demonstrated that a generalized EM algorithm exists for the optimization of the parameters of acceptor HMMs. The E-step consists of estimating the posterior state/time probabilities given the acoustic data; the M-step involves the parameter optimization of the local posterior probability estimators (typically artificial neural networks). This is a generalized EM algorithm since the M-step is not a direct maximization, but an iterative optimization. In the case of ABBOT, the Viterbi criterion is used for training as well as recognition.

4 Confidence Measures

A *confidence measure* is a function which quantifies how well a model matches some speech data, where the value of the function must be comparable across utterances. An *acoustic* confidence measure is derived exclusively from an acoustic model, whereas a *combined* confidence measure is derived from both the acoustic and language models. As poor model fit is indicative of unclear acoustics or the occurrence of an OOV word, a confidence measure is an ideal candidate for a test statistic in some hypothesis test regarding the output of a speech recognizer. A more restrictive definition of a confidence measure (Weintraub et al., 1997; Gillick et al., 1997) is the posterior probability of word correctness given a set of "confidence indicators" for the recognizer output, such as acoustic and language model probabilities, the duration of the word hypothesis and information from a word graph. The latter definition has the disadvantage of conglomerating multiple potential causes of low confidence, typically through a post-classifier, and thus obscuring their individual contributions.

Defining confidence measures in terms of model fit allows utterance verification based on posterior probabilities of speech sound units, but also opens the possibility for characterizing confidence at different levels (state, phone, word and utterance). For example, low confidence at the state-level implies a mismatched acoustic model (most likely due to unclear acoustics), whereas a low confidence at word-level, but a high confidence at state-level implies a mismatched pronunciation model.

We have investigated four acoustic confidence measures based on the local posterior probabilities estimated by an acceptor HMM system, arising from (19). Each of the equations below refer to phone-level outputs q_k of the recogniser with start and end frames n_s and n_e respectively. The duration of q_k in frames is thus $D = n_e - n_s + 1$.

Posterior Probability $\mathrm{PP}(q_k)$ is computed by rescoring the Viterbi state sequence using the local posterior probability estimates estimated by the acceptor HMM acoustic model,

$$\mathrm{PP}(q_k) = \sum_{n=n_s}^{n_e} \log\{p(q_k|X_1^n)\} \ . \tag{20}$$

Scaled Likelihood The acceptor HMM system of section 3 is regarded as a "pseudo-generative" model, in which the likelihoods of a generative model are replaced by likelihood ratios or *scaled likelihoods* (Renals et al., 1994), which following (19) may be obtained by dividing the local posterior probability by the class prior estimated from the relative frequencies of the phone labels in the acoustic training data:

$$\frac{p(X_1^n|q_k)}{p(X_1^n)} = \frac{P(q_k|X_1^n)}{P(q_k)} \ . \tag{21}$$

This may be used to define $\mathrm{SL}(q_k)$, the log scaled likelihood of a phone hypothesis q_k:

$$\begin{aligned} \mathrm{SL}(q_k) &= \sum_{n=n_s}^{n_e} \log\left\{\frac{P(q_k|X_1^n)}{P(q_k)}\right\} \\ &= PP(q_k) - D\log\{P(q_k)\} \ . \end{aligned} \tag{22}$$

Online Garbage The term "online garbage" (Boite et al., 1993; Bourlard et al., 1994) is used to refer to the normalization of the probability of the best decoding hypothesis by the average probability of the m-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. $\mathrm{OLG}_m(q_k)$ is $\mathrm{SL}(q_k)$ normalized by the average of the m-best scaled likelihoods.

$$\mathrm{OLG}_m(q_k) = \mathrm{SL}(q_k) - \sum_{n=n_s}^{n_e} \log\left\{\frac{1}{\text{m best}} \sum_{l=\text{best}}^{\text{mth best}} \frac{p(q_l^n|X_1^n)}{p(q_l)}\right\} \ . \tag{23}$$

Per-Frame Entropy $\mathrm{S}(n_s, n_e)$ is the per-frame entropy of the K phone class posterior probabilities estimated by the acceptor HMM acoustic model, averaged over the interval n_s to n_e:

$$\mathrm{S}(n_s, n_e) = -\frac{1}{D}\sum_{n=n_s}^{n_e}\sum_{k=1}^{K} p(q_k^n|X_1^n)\log\{p(q_k^n|X_1^n)\} \ . \tag{24}$$

Duration normalized versions of $\mathrm{SL}(q_k)$, $\mathrm{PP}(q_k)$ and $\mathrm{OLG}_m(q_k)$, may be obtained by dividing by D. $\mathrm{S}(n_s, n_e)$ is already normalized for duration. A consequence of the observation independence assumption is that the probability of a decoding hypothesis is always underestimated. Duration normalization counteracts the bias toward shorter decoding hypotheses created by this underestimate. The duration normalized version of, for example, $\mathrm{PP}(q_k)$ is denoted $\mathrm{nPP}(q_k)$:

$$\mathrm{nPP}(q_k) = \frac{1}{D}\mathrm{PP}(q_k) \; . \tag{25}$$

$\mathrm{SL}(q_k)$, $\mathrm{PP}(q_k)$ and $\mathrm{OLG}(q_k)$ may be extended to the word-level by averaging their values over the phones that are constituent to the word hypotheses (Bernardis and Bourlard, 1998). $\mathrm{S}(n_s, n_e)$ may be derived at the word-level by simply matching the period over which it is calculated to the duration of the word hypothesis.

We have also investigated a combined confidence measure that incorporates both language model and acoustic model information.

Lattice Density $\mathrm{LD}(n_s, n_e)$ is a measure of the density of competitors in an m-best lattice of decoding hypotheses and is computed by averaging the number of unique decoding hypotheses which pass through a frame over the interval n_s to n_e:

$$\mathrm{LD}(n_s, n_e) = \frac{1}{D}\sum_{n=n_s}^{n_e} \mathrm{NCH}_n \quad , \tag{26}$$

where, NCH_n is the number of competing decoding hypotheses which pass through the nth frame of the lattice.

If $\mathrm{LD}(n_s, n_e)$ is calculated from an m-best lattice of word hypotheses, NCH_n is equivalent to an 'active word count' described by Hetherington (1995). $\mathrm{LD}(n_s, n_e)$ constitutes a combined confidence measure since an m-best lattice of decoding hypotheses from which it is calculated is created using both language model and acoustic model information. As with the entropy measure, $\mathrm{LD}(n_s, n_e)$ is already normalized for duration.

5 Experiments

Utterance verification experiments were performed using the North American Business News (NAB) and Broadcast News (BN) corpora.[3] The NAB corpus consists of a set of business news sentences dictated in a quiet office environment. In these experiments the Hub-3 1995 evaluation test set (Hub-3E-95) comprising 310 utterances was used. The BN corpus is composed of a set of recordings of broadcast radio and television news shows. This data encompasses a variety of speaking styles and acoustic conditions, and includes acoustic phenomena such narrow band transmission channels as well as background noise and music. The Hub-4 1997 evaluation test set (Hub-4E-97) constituting approximately 3 hours of data was used.

For the experiment using NAB data, the ABBOT speech recognizer used two recurrent networks (trained forwards and backwards in time, to exploit the temporal asymmetry of the recurrent network) with a perceptual linear prediction (PLP) front end. These networks were trained on the WSJ0 corpus of around 7 200 utterances. For the BN corpus the outputs of two similar recurrent networks, trained on approximately 39 hours of the Hub-4 1996 training set, were merged with the output of a 4 000 hidden unit multilayer perceptron, trained on the same data using modulation-filtered spectrogram features (Cook et al., 1999). A backed-off trigram language model was used, trained on the 200 million word NAB text corpus in the NAB case and on the 132 million word BN text corpus for the BN system. Vocabularies of 60 022 and 65 532 words, obtained from the most common unigrams in the text corpora, were used respectively.

[3] Both corpora are available from the Linguistic Data Consortium: http://www.ldc.upenn.edu/

Each data set was decoded under two conditions. The first condition used the *word-level decoding constraints* of a pronunciation lexicon and a word n-gram language model; the second used neither of these and so was governed only by the *phone-level decoding constraints* of a bigram defined over the phoneset (estimated from the acoustic training data). Recognition output at the word- and phone-levels was recorded for the first condition; only phone-level output could be recorded for the second.

Given the outputs of the speech recognizer, marked as either correct or incorrect and tagged with their respective confidence estimates, a hypothesis test was formed using the confidence estimates as values of the test statistic. To determine whether the recognizer output is correct or incorrect, the output was aligned with the transcript. In addition to considering errors due to substitutions and insertions, poor time alignment was also considered to be an error (Weintraub et al., 1997). Specifically, for a segment of the recognition output to be considered well time aligned, an identical reference segment was required with greater than 50% of its duration overlapping with that of the recognition segment and vice versa. An example of good and poor time alignment is schematically illustrated in figure 4.

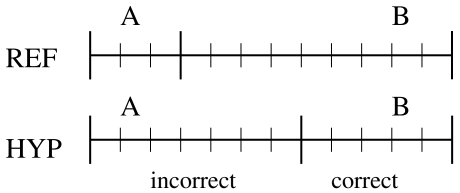


Figure 4: A schematic illustration of the 50% overlap criterion used to assess the time alignment of the recognition output. After Weintraub et al. (1997).

The results of applying a hypothesis test to the recognition output was recorded in a 2×2 confusion matrix, such as that illustrated in figure 1. H_0 was defined to be the hypothesis that a given segment of the recognition output is correct. From such a matrix, the unconditional error rate of the test was calculated using (1). The probabilities of type I and type II errors were computed using (2) and (3). The mutual information $I(Z;A)$ and the efficiency $E(Z;A)$ were calculated using (4) and (5) respectively and the values for d_{Kol}, d_{Bhatt} and d_{KL2} were calculated for the various confidence measures following (6), (7) and (8). In all cases empirical probability estimates derived from a 2×2 confusion matrix were used to calculate the values of the evaluation metrics.

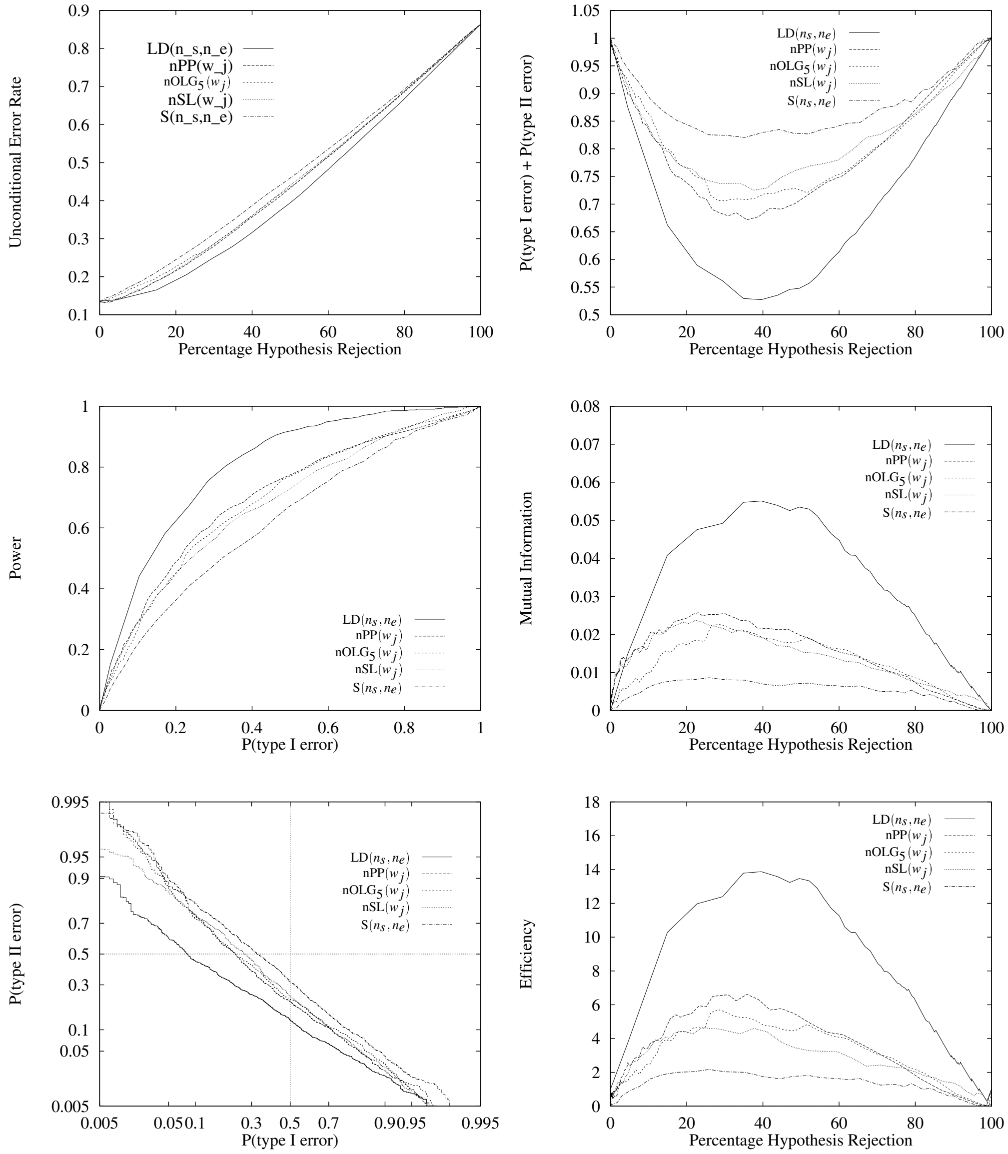
The results of the experiments are presented in figures 5, 6 and 7 and in table 1.

From figure 5, it can be seen that, in broad terms, the diverse range of evaluation metrics investigated all agree in their rankings of the five confidence measures. The task independent metrics, $E(Z;A)$, the ROC curve and the sum of $P(\text{type I error})$ plus $P(\text{type II error})$, provide greater distinctions between the performance the various confidence measures that the UER task dependent metric. The UER curve shows (a) that none of the confidence measures facilitate ‘profitable rejection’ on NAB data, *i.e.* the rejection of any fraction of the decoding hypotheses based upon their associated confidence estimates only leads to an increase in the UER of the test; and (b) that the recognition error rate on this particular data set is relatively low (15.4%), evidenced by the low y-axis intercept. The curves in the DET plot do not fall far enough into the lower left quadrant to benefit from the increased separation offered by the axis warping. The plots of $E(Z;A)$ and $I(Z;A)$ are very similar, bar a scaling of the y-axis.

The performance of the $\mathrm{nOLG}_m(w_j)$ confidence measure, as measured by the sum of $P(\text{type I error})$ plus $P(\text{type II error})$ on NAB data, for various values of m is given in figure 6. It can be seen from the left panel of the figure that small performance improvements are seen as m is increased from one to five. The graph in the right panel shows that little further gains are obtained as m is increased from 5 to 40. Although equivalent graphs are not shown, it was found that $m = 40$ was best for BN data, supporting the intuition that the phone class posterior probability distributions are not as ‘sharp’ for BN data.

A comparison of unconditional error rates over the two data types and decoding conditions is given in figure 7.

- At the word-level, the small amount of profitable rejection for BN data can be contrasted against none for NAB data. The combined confidence measure $\mathrm{LD}(n_s, n_e)$ provides the best performance



Confidence Measure	d_{Kol}	d_{Bhatt}	d_{KL2}	area under ROC curve	EER
$\text{LD}(n_s, n_e)$	-0.4744	0.8330	1.9185	0.7772	0.2647
$\text{nPP}(w_j)$	-0.3114	0.9269	1.8757	0.6937	0.3507
$\text{nOLG}_5(w_j)$	-0.2897	0.9369	1.8356	0.6805	0.3591
$\text{nSL}(w_j)$	-0.2822	0.9292	1.8243	0.6851	0.3556
$\text{S}(n_s, n_e)$	-0.1832	0.9717	1.6643	0.6096	0.4174

Figure 5: Assessments provided by various evaluation metrics for utterance verification at the word-level on Hub-3E-95. *Top Left*: Unconditional error rates. *Top Right*: The sum of $P(\text{type I error})$ plus $P(\text{type II error})$ against percentage hypothesis rejection. *Upper Left*: ROC curves. *Upper Right*: Mutual information, $I(Z;A)$, against percentage hypothesis rejection. *Lower Left*: DET curves, *i.e.* $P(\text{type I error})$ vs. $P(\text{type II error})$ on a Gaussian deviate scale. *Lower Right*: Efficiency, $E(Z;A)$, against percentage hypothesis rejection. *Bottom*: Values of the scalar valued evaluation metrics, d_{Kol}, d_{Bhatt}, d_{KL2}, the area below the ROC curve and the EER.

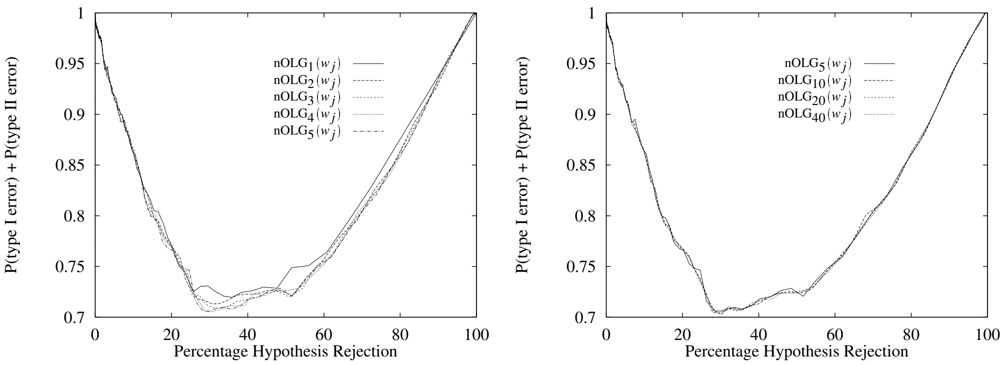


Figure 6: The utterance verification performance of $\mathrm{nOLG}_m(w_j)$ on Hub-3E-95 for various values of m, as measured by the sum of $P(\text{type I error})$ plus $P(\text{type II error})$ plotted against percentage hypothesis rejection.

on NAB data but is outdone by the purely acoustic measure $\mathrm{nPP}(w_j)$ (offering an 11.6% relative decrease in UER at a rejection rate of 10.2%) on BN data. The worst performing measure at this level is $\mathrm{S}(n_s, n_e)$.

- For the phone-level decodings made using word-level constraints, a similar pattern of a small amount of profitable rejection for BN data and none for NAB data can be seen. For both these graphs, the curves for all of the confidence measures are tightly clustered except for $\mathrm{S}(n_s, n_e)$, which again performs the worst, offering no profitable rejection for either data type.
- Profitable rejection is witnessed for both data types for phone-level decodings made with phone-level constraints. The $\mathrm{nPP}(q_k)$ measure provides the best performance on both data types at this level (offering a 23.9% relative reduction in UER at a rejection rate of 17.8% on NAB data and a 25.0% reduction at a rate of 17.8% on BN data). $\mathrm{S}(n_s, n_e)$ is promoted to the third best performing confidence measure whilst $\mathrm{LD}(n_s, n_e)$ is relegated to the worst performing measure.

Similar performance trends can be seen from the values for the area under the ROC curve given in table 1.

	Hub-3E-95	**Hub-4E-97**
$\mathrm{LD}(n_s, n_e)$	0.7772	0.7695
$\mathrm{nPP}(w_j)$	0.6937	0.7499
$\mathrm{nOLG}_{5/40}(w_j)$	0.6805	0.7362
$\mathrm{nSL}(w_j)$	0.6694	0.7443
$\mathrm{S}(n_s, n_e)$	0.6096	0.6643
$\mathrm{nPP}(q_k)$	0.7908	0.8394
$\mathrm{nOLG}_{5/40}(q_k)$	0.7647	0.8380
$\mathrm{nSL}(q_k)$	0.7287	0.8168
$\mathrm{S}(n_s, n_e)$	0.6860	0.6853
$\mathrm{LD}(n_s, n_e)$	0.5243	0.6919
$\mathrm{nPP}(q_k)$	0.8376	0.8465
$\mathrm{nOLG}_{5/40}(q_k)$	0.8117	0.8285
$\mathrm{nSL}(q_k)$	0.7056	0.7404
$\mathrm{S}(n_s, n_e)$	0.7960	0.7645

Table 1: The area under the ROC curve for utterance verification on Hub-3E-95 and Hub-4E-97. Values for word- (*Top*) and phone-level (*Mid*) hypotheses derived from a word-level constrained decoding, and for hypotheses derived from a phone-level constrained decoding (*Bottom*).

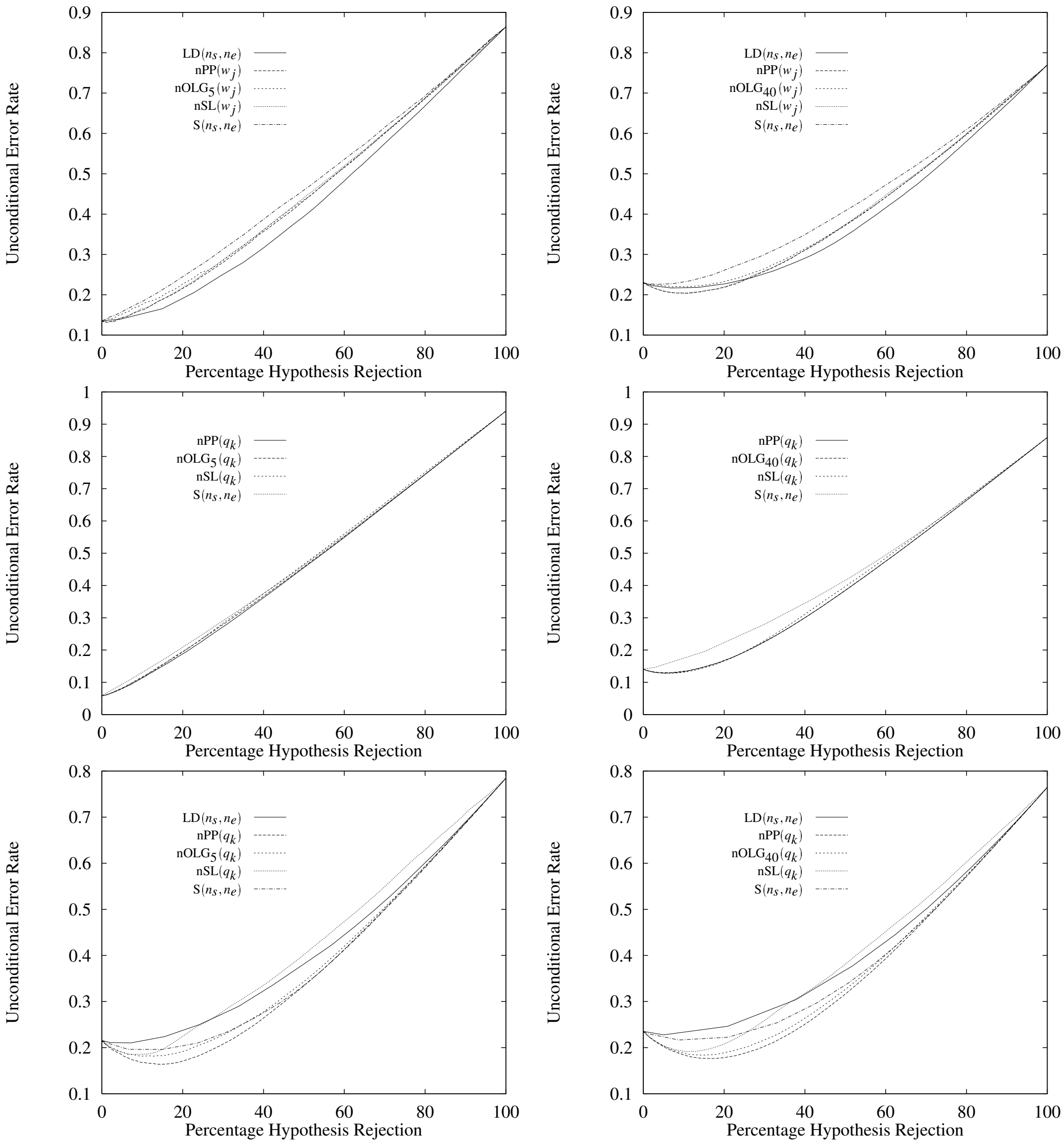


Figure 7: Unconditional error rates for utterance verification on Hub-3E-95 (*Left*) and Hub-4E-97 (*Right*). Plots for word- (*Top*) and phone-level (*Mid*) hypotheses derived from a word-level constrained decoding, and for hypotheses derived from a phone-level constrained decoding (*Bottom*).

6 Discussion

The NAB corpus is composed of clean, read speech, whereas the BN corpus contains speech 'found' under a variety of acoustic conditions, such as narrow band channels and in the presence of background music or noise. When decoding NAB data, it may thus be expected that a large proportion of the errors are due to OOV words and disfluencies. The range of potential sources of error for BN, on the other hand, is much wider.

An explanation as to why profitable rejection is possible at the word-level on BN but not on NAB data is that crude pronunciation models mask the relatively subtle reductions caused by OOV words and disfluencies but not the gross model mismatches elicited by non-speech sounds. The effects of crude pronunciation models extend to both the word- and the phone-levels: At the word-level, words with crude pronunciation models which are nevertheless correctly decoded will suffer from reduced confidence. At any level, decodings must be marked against a reference. At the phone-level, this reference is typically a forced Viterbi alignment of the reference word transcription which relies upon a set of relevant pronunciation models. Crude pronunciation models will result in phones constituent to the word constraint decoding being erroneously marked as correct, despite their low confidence, and components of the phone constraint decoding being marked as incorrect, not withstanding their high confidence.

An example of a crude pronunciation model is illustrated in lower left panel of figure 8. The plot shows the relevant outputs of the acoustic model (solid lines) evolving over the duration of an instance of the word *funds* (drawn from the Hub-3E-95 dataset and correctly decoded despite its crude pronunciation model) overlaid with timings (dashed lines) for the Viterbi alignment of the baseform [f ah n dcl d z] to the acoustics. Also given in the plot are values $\mathrm{nPP}(q_k)$ for the alignment of each of the constituent phones. The output of the acoustic model clearly suggests the absence of the phones [dcl] and [d] between the 177th and 180th frames of the utterance and the alignment of these phones over this interval are assigned correspondingly low confidence estimates (-3.67 and -4.56 respectively). The alignment of the baseform [f ah n dcl d z] receives an overall value of $\mathrm{nPP}(w_j) = -1.53$ which is improved to -0.16 for the alignment of the baseform [f ah n z], shown in the lower right panel of the figure. Support for the notion that the reduction in confidence due to crude pronunciation models is similar to that for OOV words is provided by the plots in the upper panels of figure 8. The upper left panel of the figure provides the plot for the alignment of the model for the word pair *better one* [bcl b eh dx axr w ah n] to an instance of the word *bedouin* (again drawn from the Hub-3E-95 and decoded as *better one* when *bedouin* is OOV). The upper right panel plots the alignment of the correct model. It can be seen that the two models differ by only a single phone and that the confidence for [ih] ($\mathrm{nPP}(q_k) = -0.41$) between the 586th and the 588th frames of the utterance is much higher than that for [ah] ($\mathrm{nPP}(q_k) = -5.67$) over a similar period. Overall, the confidence values for *better one* [bcl b eh dx axr w ah n] ($\mathrm{nPP}(w_j) = -1.35$) and *bedouin* [bcl b eh dx axr w ih n] ($\mathrm{nPP}(w_j) = -0.68$) are comparable to those for the two pronunciation models for *funds*. An interesting remedy to the problem can potentially be crafted from the sensitivity of confidence measures to crude pronunciation models itself, by turning the phenomenon 'on its head' and using acoustic confidence measures to evaluate potential pronunciation models in alignment against example acoustics. These evaluations may then be used to inform the search for improved models.

As the confidence measure $\mathrm{S}(n_s, n_e)$ is based upon the per-frame entropy of the K phone class posterior probability estimates, it will only indicate low confidence if none of these class models provides a good fit to a given frame of data. $\mathrm{S}(n_s, n_e)$ may thus be used to identify regions which are not clean speech. The poor performance of $\mathrm{S}(n_s, n_e)$ for utterance verification on the word constraint decodings is therefore not surprising as the value of the measure is independent of the actual decoding hypotheses. The improved performance of $\mathrm{S}(n_s, n_e)$ for the phone constraint decodings reflects the absence of correlates for OOV words and crude pronunciation models for this condition. Decoding errors are much more likely to be caused by unclear acoustics in this case.

The reduced performance at the word-level of the $\mathrm{LD}(n_s, n_e)$ measure on BN data may be attributed to a reduction in the quality of language model fit on this data type. Whereas NAB data is composed of read newspaper text and so has a constrained grammar, the portions of more spontaneous speech which are included in the BN corpus will have a relatively unconstrained grammar which is harder to capture with a simple trigram language model. The poor performance to the $\mathrm{LD}(n_s, n_e)$ measure for the phone constraint decodings may be explained by the reduced quality of language model used for this condition.

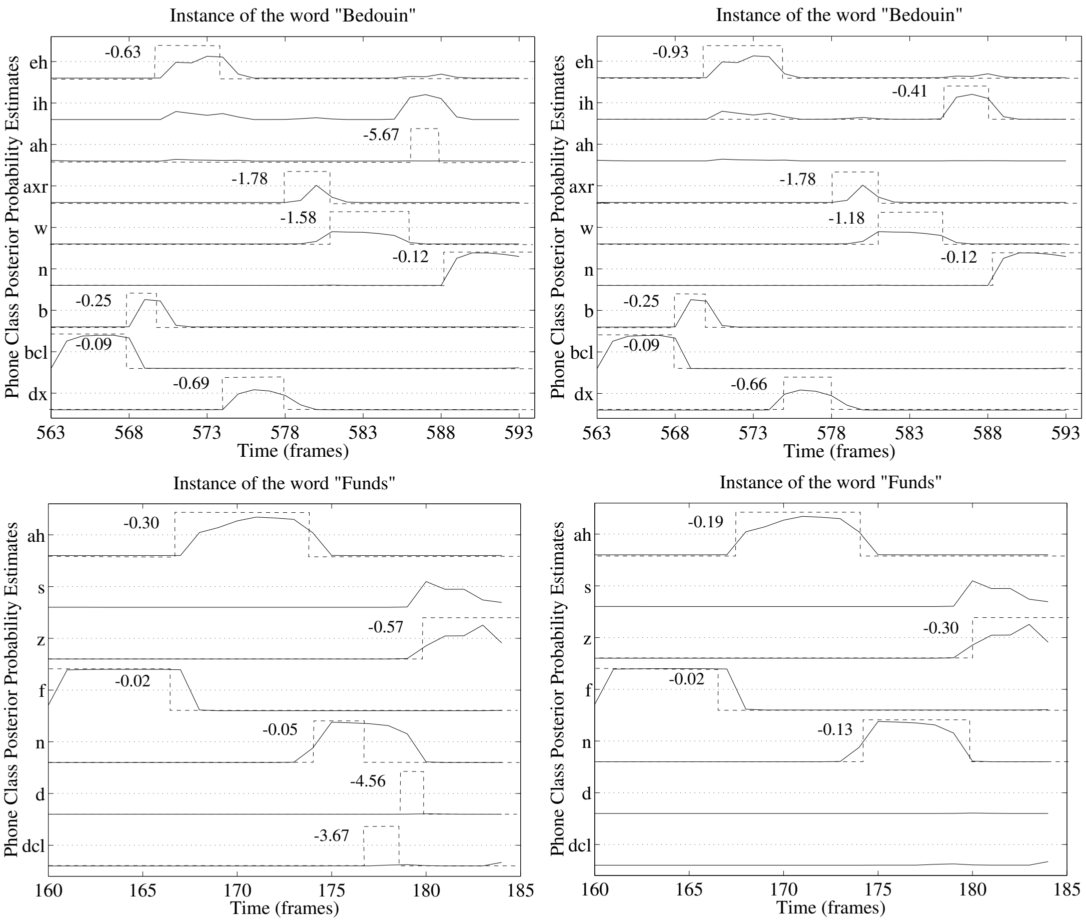


Figure 8: The relevant subset of acoustic model outputs (solid lines) generated by an instance of the word *Bedouin* (*Top*) and *Funds* (*Bottom*), overlaid with timings (dashed lines) and values of nPP(q_k) obtained from a forced Viterbi alignment of the model for *Better One* [bcl b eh dx axr w ah n] (*Top Left*), *Bedouin* [bcl b eh dx axr w ih n] (*Top Right*), *funds* [f ah n dcl d z] (*Bottom Left*) and *funds* [f ah n z] (*Bottom Right*) to the relevant acoustics.

7 Summary

In this paper we have shown that acceptor HMMs, which directly estimate posterior probabilities, are well suited to producing computationally efficient measures of confidence. We have evaluated a set of related acoustic confidence measures and a combined confidence measure for utterance verification using a number of evaluation metrics. Our experiments have revealed several trends in 'profitability of rejection', as measured by the unconditional error rate of a hypothesis test. These trends suggest that crude pronunciation models can mask the relatively subtle reductions in confidence caused by OOV and disfluencies, but not the gross model mismatches elicited by non-speech sounds. The observation that a purely acoustic confidence can provide improved performance over a measure based upon both acoustic and language model information for data drawn from the Broadcast News corpus, but not for data drawn from the North American Business News corpus suggests that the quality of model fit offered by a trigram language model is reduced for Broadcast News data. We have also argued that our definition of a confidence measure provides a focus for investigating whether low confidence is caused by, for example, an OOV word or some unclear acoustics. Lastly, we have suggested that purely acoustic confidence measures may be useful in the search for improved pronunciation models.

Acknowledgements

This work was supported by ESPRIT Long Term Research Project THISL (LTR23495) and an EPSRC studentship awarded to GW.

References

Bahl, L. R., Brown, P. F., de Souza, P. V., and Mercer, R. L. (1986). Maximum mutual information estimation of hidden Markov model parameters for speech recognition. In *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, pages 49–52, Tokyo.

Bengio, Y., de Mori, R., Flammia, G., and Kompe, R. (1992). Globally optimization of a neural network—hidden Markov model hybrid. *IEEE Trans. Neural Networks*, 3:252–259.

Bernardis, G. and Bourlard, H. (1998). Improving posterior confidence measures in hybrid HMM/ANN speech recognition systems. In *Proceedings of the International Conference on Spoken Language Processing*, pages 775–778.

Boite, J.-M., Bourlard, H., D'hoore, B., and Haesen, M. (1993). A new approach towards keyword spotting. In *Proc. Europ. Conf. Speech Communication and Technology*, pages 1273–1276, Berlin.

Bourlard, H., D'hoore, B., and Boite, J.-M. (1994). Optimizing recognition and rejection performance in wordspotting systems. In *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, volume 1, pages 373–376, Adelaide.

Bourlard, H., Konig, Y., and Morgan, N. (1996). A training algorithm for statistical sequence recognition with applications to transition-based speech recognition. *IEEE Signal Processing Lett.*, 3:203–205.

Bourlard, H. and Morgan, N. (1994). *Connectionist Speech Recognition—A Hybrid Approach*. Kluwer Academic Publishers.

Cook, G., Christie, J., Ellis, D., Fosler-Lussier, E., Gotoh, Y., Kingsbury, B., Morgan, N., Renals, S., Robinson, T., and Williams, G. (1999). The SPRACH system for the transcription of broadcast news. In *Proceedings of the DARPA Broadcast News Workshop*.

Cox, S. and Rose, R. C. (1996). Confidence measures for the Switchboard database. In *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, pages 511–515, Atlanta.

Egan, J. P. (1975). *Signal Detection Theory and ROC Analysis*. Academic Press.

Gillick, L., Ito, Y., and Young, J. (1997). A probabilistic approach to confidence estimation and evaluation. In *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, pages 879–882, Munich.

Hand, D. J. (1997). *Construction and Assessment of Classification Rules*. John Wiley & Sons Ltd.

Hennebert, J., Ris, C., Bourlard, H., Renals, S., and Morgan, N. (1997). Estimation of global posteriors and forward-backward training of hybrid HMM/ANN systems. In *Proc. Europ. Conf. Speech Communication and Technology*, pages 1951–1954, Rhodes, Greece.

Hetherington, L. (1995). New words: Effect on recognition performance and incorporation issues. In *Proc. Europ. Conf. Speech Communication and Technology*, pages 1645–1648, Madrid.

Martin, A., Doddington, G., Kamm, T., Ordowski, M., and Pryzybocki, M. (1997). The DET curve in assessment of detection task performance. In *Proc. Europ. Conf. Speech Communication and Technology*, pages 1895–1898, Rhodes, Greece.

Renals, S., Morgan, N., Bourlard, H., Cohen, M., and Franco, H. (1994). Connectionist probability estimators in HMM speech recognition. *IEEE Trans. Speech and Audio Processing*, 2:161–175.

Robinson, A. J. (1994). The application of recurrent nets to phone probability estimation. *IEEE Trans. Neural Networks*, 5:298–305.

Robinson, T., Hochberg, M., and Renals, S. (1996). The use of recurrent networks in continuous speech recognition. In Lee, C. H., Paliwal, K. K., and Soong, F. K., editors, *Automatic Speech and Speaker Recognition – Advanced Topics*, chapter 10, pages 233–258. Kluwer Academic Publishers.

Weintraub, M., Beaufays, F., Rivlin, Z., Konig, Y., and Stolcke, A. (1997). Neural-network based measures of confidence for word recognition. In *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, pages 887–890, Munich.

Williams, G. and Renals, S. (1997). Confidence measures for hybrid HMM/ANN speech recognition. In *Proc. Europ. Conf. Speech Communication and Technology*, pages 1955–1958, Rhodes, Greece.

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