CONFIDENCE MEASURES DERIVED FROM AN ACCEPTOR HMM

Gethin Williams and Steve Renals

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

ABSTRACT

In this paper we define a number of confidence measures derived from an *acceptor* HMM and evaluate their performance for the task of utterance verification using the North American Business News (NAB) and Broadcast News (BN) corpora. Results are presented for decodings made at both the word and phone level which show the relative profitability of rejection provided by the diverse set of confidence measures. The results indicate that language model dependent confidence measures have reduced performance on BN data relative to that for the more grammatically constrained NAB data. An explanation linking the observations that rejection is more profitable for noisy acoustics, for a reduced vocabulary and at the phone level is also given.

1. INTRODUCTION

We define a confidence measure as a function which quantifies how well a model matches some spoken utterance, where the values of the function must be comparable across utterances. More specifically, an *acoustic* confidence measure is one which is derived exclusively from the acoustic model, whereas a *grammatical* confidence measure is derived solely from the language model (LM). A *combined* confidence measure is derived from both the acoustic and language models. The above definition is less restrictive than an often used alternative which formulates a confidence measure as the posterior probability of word correctness given a set of 'confidence indicators' [8]. The former definition has the advantage of allowing confidence measures to be applied at the state, phone and word levels; the pursuit of the latter is often characterised by the conglomeration of multiple potential causes of low confidence, typically through a postclassifier, obscuring their individual contributions.

Following [9], three acoustic confidence measures are presented in section 2. These measures are based on local phone posterior probability estimates produced by an HMM/ANN system [7, 3]. We refer to such systems as *acceptor* HMMs to contrast with the generative modelling approach adopted in most HMM systems. We have demonstrated in [9] that acceptor HMMs are well suited to producing computationally efficient acoustic confidence measures. One grammatical and two combined confidence measures are also presented in section 2 for comparison.

The results of the application of these confidence measures to the task of utterance verification at the word and phone level using the North American Business News (NAB) and Broadcast News (BN) corpora[1] are given in section 3. This section also contains a brief outline of the range of metrics available for evaluating confidence measures together with a discussion of the insights into the causes of low confidence provided by the results.

This work was supported by an EPSRC studentship and by ESPRIT Long Term Research Project 23495 (THISL).

[1] Linguistic Data Consortium: http://www.ldc.upenn.edu/

2. CONFIDENCE MEASURES

2.1. Acoustic Measures

Three purely acoustic confidence measures are defined below for an hypothesised phone q_k with a duration $D = n_e - n_s + 1$, where n_s and n_e are the start and end frames respectively. The acoustic observation vector at frame n is denoted $\mathbf{x}^n$.

Posterior Probability $\mathrm{nPP}(q_k)$ is computed by rescoring the Viterbi state sequence using the local posterior probability estimates produced by the acceptor HMM acoustic model and duration normalising:

$$\mathrm{nPP}(q_k) = \frac{1}{D}\sum_{n=n_s}^{n_e} \log\left(p(q_k|\mathbf{x}^n)\right) . \tag{1}$$

Scaled Likelihood The 'scaled likelihood' of q_k is obtained by dividing the local posterior probability estimate by the class prior, obtained from the acoustic data:

$$\frac{p(\mathbf{x}^n|q_k)}{p(\mathbf{x}^n)} = \frac{P(q_k|\mathbf{x}^n)}{P(q_k)} . \tag{2}$$

$\mathrm{nSL}(q_k)$ is the duration normalised log scaled likelihood of q_k:

$$\begin{aligned}\mathrm{nSL}(q_k) &= \frac{1}{D}\sum_{n=n_s}^{n_e} \log\left(\frac{p(q_k|\mathbf{x}^n)}{p(q_k)}\right) \\ &= \mathrm{nPP}(q_k) - \log\left(p(q_k)\right) .\end{aligned} \tag{3}$$

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:

$$\mathrm{S}(n_s,n_e) = -\frac{1}{D}\sum_{n=n_s}^{n_e}\sum_{k=1}^{K} p(q_k^n|\mathbf{x}^n)\log\left(p(q_k^n|\mathbf{x}^n)\right) . \tag{4}$$

2.2. Grammatical and Combined Measures

Equations for the one grammatical and two combined confidence measures are:

N-gram Probability $\mathrm{nNG}(q_k)$ is computed by rescoring the optimal phone sequence using the probability of q_k conditioned upon its n-gram history h, as estimated by the LM, and durationally normalising:

$$\mathrm{nNG}(q_k) = \frac{1}{D}\log\left(p(q_k|h)\right) . \tag{5}$$

N-gram based Posterior Probability $\text{nPP}_{ng}(q_k)$ results from replacing the acoustic class prior inclusive in $\text{nPP}(q_k)$ with the n-gram probability of that class:

$$\text{nPP}_{ng}(q_k) = \frac{1}{D}\sum_{n=n_s}^{n_e} \log\left(\frac{p(\mathbf{x}^n|q_k)}{p(\mathbf{x}^n)}.p(q_k|h)\right) \quad (6)$$

$$= \text{nSL}(q_k) + \log\left((q_k|h)\right) . \quad (7)$$

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

$$\text{LD}(n_s, n_e) = \frac{1}{D}\sum_{n=n_s}^{n_e} \text{NCH}_n \ , \quad (8)$$

where NCH_n is the number of competing decoding hypotheses which pass through the nth frame of the lattice. If $\text{LD}(n_s, n_e)$ is computed from an n-best lattice of word hypotheses, NCH_n is equivalent to the 'active word count' described in [5].

$\text{nPP}(q_k)$ and $\text{nSL}(q_k)$ may be extended to a word hypothesis w_j by summing their values over the L phone hypotheses constituent to w_j and normalising by L [2]. $\text{S}(n_s, n_e)$ and $\text{LD}(n_s, n_e)$ may be derived at the word level by simply matching the period over which they are calculated to the duration of the word hypothesis. $\text{nPP}_{ng}(w_j)$ and $\text{nNG}(w_j)$ make use of word level LM statistics.

3. EXPERIMENTS

3.1. Utterance Verification

The task of utterance verification maybe cast as a *statistical hypothesis test*, where the decision to accept or to reject the null hypothesis H_0, regarding the correctness of the recogniser output, is based upon a threshold on the confidence estimate. A number of metrics are available to evaluate the performance of the confidence measure in this case. The simplest of these is the *unconditional error rate* of the hypothesis test, where an error will occur if H_0 is rejected when it is true (a type I error) or accepted when it is false (a type II error). The individual probabilities of type I and type II errors provide error statistics *conditioned* upon a particular state of nature (the truth or falsity of H_0). The unconditional error rate evaluates the performance of an hypothesis test relative to a particular task, whereas the two conditional error rates can be used to evaluate the performance of the test independently of the prior probabilities of the two states of nature. Whilst conditional error rates (task independent) are useful for confidence measure development, only unconditional error rates are reported here due to space constraints and the desire to provide task dependent results.

In addition to unconditional and conditional error rates, a range of evaluation metrics including mutual information [4], ROC (Receiver Operating Characteristic) [11] and DET (Detection Error Tradeoff) [6] curves and distributional separability have been investigated in [10]. Two findings of this investigation were that the diverse set of metrics broadly agree in their evaluations and that duration normalisation was beneficial for all confidence measures.

Utterance verification experiments were performed using the Hub-3 1995 evaluation test set of the NAB corpus and seven episodes from the 1996 training set of the BN corpus. The ABBOT large vocabulary continuous speech recognition (LVCSR) system [7] was used to decode each data set under two conditions. The first used the *word level decoding constraints* of a pronunciation lexicon and a word n-gram LM; 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 may be recorded for the first condition, whereas only phone level output may be recorded for the second.

3.2. Word vs. Phone Level

A broad trend in two dimensions can be seen in figures 1 and 2. Firstly, rejection is more profitable for BN than for NAB data and secondly rejection is more profitable for phone constraint decoding hypotheses than at the word level. To explain this, consider the range of values which a confidence measure may take: non-speech sounds will cause gross model mismatches and so lead to large reductions in confidence in comparison to that for correctly recognised clean speech. Conversely, the occurrence of OOV words, for example, will cause more subtle model mismatches as a pronunciation model from the lexicon of a LVCSR system may be incompatible with an OOV word by perhaps only a single phone. Such disparities will give rise to correspondingly small reductions in confidence.

This pattern of confidence reduction is complicated, however, by the presence of crude pronunciation models in the lexicon. Such models subtly reduce the confidence with which words are correctly recognised. This 'noise' on the confidence measure values masks the range of confidence associated with, OOV words. This masking makes it difficult to set a threshold for profitable rejection for the clean speech of the NAB corpus, whereas the presence of non-speech sounds in the BN data facilitates profitable rejection (figure 1).

Mismatches will be more distinct for phone constraint decodings, facilitating more profitable rejection, as there is no correlate of a crude pronunciation model at this level. It should be noted, however, that the effect of crude pronunciation models extends to the phone level to some degree, as phone hypotheses are marked against Viterbi alignments derived using an imperfect pronunciation lexicon. Noise will be manifest at the phone level, therefore, as incorrect phone hypotheses will be marked as correct and vice versa.

Additional experiments carried out to investigate this theory further are described in the sections below. It should be noted that the point of minima on the unconditional error rate curves, independent of the profitability of rejection, is indicative of the degree of difficulty of the recognition task.

In addition to these broad trends, it can be seen from figures 1 and 2 that both $\text{LD}(n_s, n_e)$ and $\text{nNG}(w_j)$ perform badly at the phone level and on BN data at the word level, whereas they perform at least as well as the other measures on NAB data at the word level. At the word level the LM is far more informative for the highly grammatically constrained NAB data than it is for the relatively less constrained BN data, whilst at the phone level only a bigram language model was used.

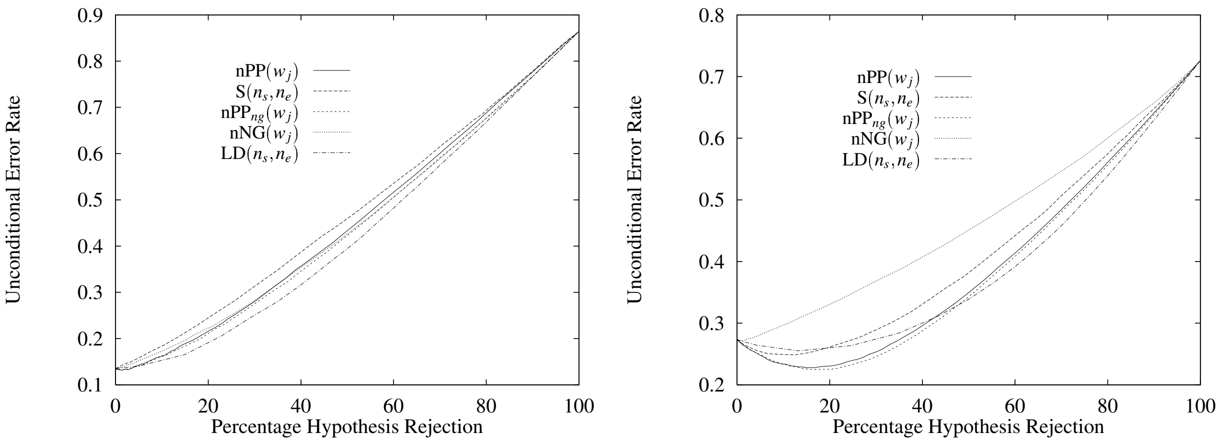

Figure 1: Word level unconditional error rate. NAB (*Left*) BN (*Right*).

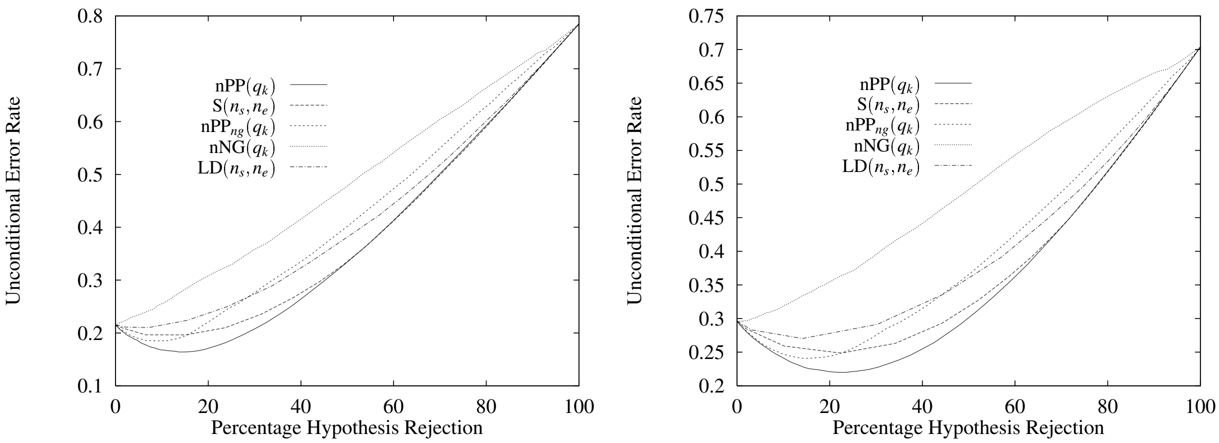

Figure 2: Unconditional error rate for phone level constraint decodings. NAB (*Left*) BN (*Right*).

3.3. Differing Acoustic Conditions

The gross model mismatches caused by non-speech sounds, facilitating profitable rejection, is clearly illustrated in figure 3. The BN corpus is partitioned into 7 acoustic conditions. The plot on the left side of the figure is for the F0 condition, which is composed of clean, planned speech similar to that found in the NAB corpus, decoded using word level constraints. The similarity of NAB and F0 data is borne out by the likeness between the plots for the two data types. The plot on the right side of the figure is for the FX condition, which can contain very noisy speech and non-speech sounds. The arrow heads below the abscissa of the two plots in figure 3 indicate the position of the overall best threshold (i.e. calculated over all 7 acoustic conditions). These two reference markers indicate that this threshold is beyond the best value for F0 data and below that for FX data, as one might expect. The second plot also highlights the completely uninformative nature of the language model and the good performance of $S(n_s, n_e)$ for the FX condition. $S(n_s, n_e)$ is designed to give high confidence for clean speech and low confidence in the presence of non-speech sounds, irrespective of the actual decoding hypothesis. For this reason the measure performs badly on NAB data, but is useful for portions of the BN corpus. We have successfully employed $S(n_s, n_e)$ for filtering out portions of unrecognisable acoustics from the input stream to the recogniser [1].

3.4. Rich vs. Sparse Lexicon

An increasing degree of mismatch for incorrect decoding hypotheses, will occur as the 'richness' of the lexicon is progressively reduced. A marked improvement in the profitability of rejection is seen between the plot on the left side of figure 1, which is for the 60k word baseline lexicon (OOV rate: 0.58%), and figure 4, which is for a 5k word decoding vocabulary (OOV rate: 8.56%). This result clearly indicates that utterance verification performance is dependent upon the vocabulary size.

4. CONCLUSIONS

The confidence measures that we have presented have a simple and explicit link to the models, allowing them to extract more subtle information regarding the cause of low confidence. Effects that we have found include:

- Crude pronunciation models limit confidence measure performance by masking relatively subtle reductions in confi-

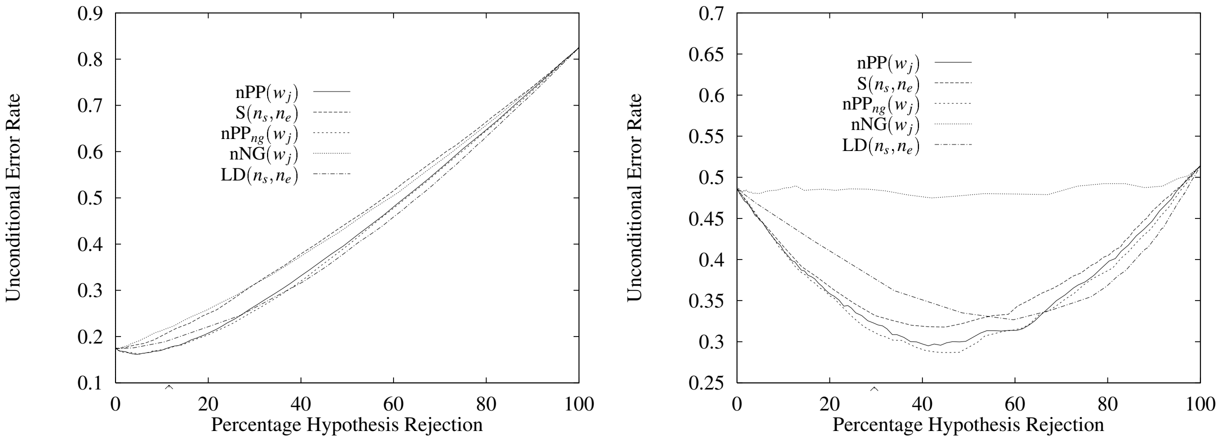

Figure 3: Word level unconditional error rate on BN data for the F0 (*Left*) and FX acoustic conditions (*Right*).

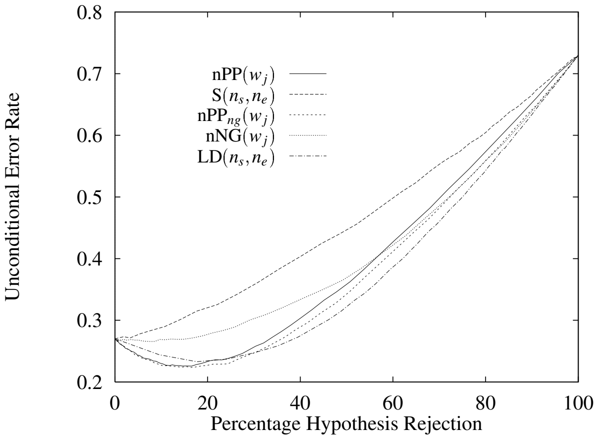

Figure 4: Word level unconditional error rate on NAB data for a 5k word decoding vocabulary.

dence caused by, for example, OOV words.

- Non-speech sounds cause gross model mismatches, beyond the range masked by crude pronunciation models, and so may be profitably rejected.
- The degree of model mismatch associated with incorrect decoding hypotheses for clean, read speech is increased as the vocabulary size is decreased. If the vocabulary size is sufficiently reduced, model mismatch may eventually be pushed past the range masked by crude pronunciation models, facilitating profitable rejection.
- The pattern of confidence reduction is subject to less 'noise' at the phone level allowing for more profitable rejection.
- Reduced quality of LM fit limits the performance of LM based confidence measures for the move from highly grammatically constrained read speech to broadcast news data.
- A set of complimentary confidence measures can be designed which respond to various causes of low confidence. For example, $S(n_s, n_e)$ is designed to signal low confidence for noisy acoustics and high confidence for clean.

5. REFERENCES

[1] J. Barker, G. Williams and S. Renals. "Acoustic confidence measures for segmenting broadcast news". In *these proceedings*.

[2] G. Bernardis and H. Bourlard. "Improving posterior based confidence measures in hybrid HMM/ANN speech recognition systems". In *these proceedings*.

[3] H. Bourlard and N. Morgan. *Connectionist Speech Recognition: A Hybrid Approach*. Kluwer, 1994.

[4] S. Cox and R. Rose. "Confidence measures for the switchboard database". In *Proceedings of ICASSP*, pages 511-515, 1996.

[5] L. Hetherington. "New words: Effect on recognition performance and incorporation issues". In *Proceedings of EuroSpeech*, pages 1645-1648, 1995.

[6] A. Martin, G. Doddington, T. Kamm, M. Ordowski and M. Pryzybocki. "The DET curve in assessment of detection Task performance". In *Proceedings of EuroSpeech*, pages 1895-1898, 1997.

[7] 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*, pages 233-258. Kluwer, 1996.

[8] M. Weintraub, F. Beaufays, Z. Rivlin, Y. Konig and A. Stolcke. "Neural - network based measures of confidence for word recognition". In *Proceedings of ICASSP*, pages 887-890, 1997.

[9] G. Williams and S. Renals. "Confidence measures for hybrid HMM/ANN speech recognition". In *Proceedings of EuroSpeech*, pages 1955-1958, 1997.

[10] G. Williams "A study of the use and evaluation of confidence measures in automatic speech recognition". *Technical report CS-98-02*, Department of Computer Science, University of Sheffield, 1998. http://www.dcs.shef.ac.uk/people/G.Williams.

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