CONFIDENCE MEASURES FOR HYBRID HMM/ANN SPEECH RECOGNITION

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 introduce four acoustic confidence measures which are derived from the output of a hybrid HMM/ANN large vocabulary continuous speech recognition system. These confidence measures, based on local posterior probability estimates computed by an ANN, are evaluated at both phone and word levels, using the North American Business News corpus.

1. INTRODUCTION

A reliable measure of the confidence of a speech recogniser's output is useful in many circumstances. A word may be hypothesised with low confidence when an out-of-vocabulary (OOV) word is encountered or when the word model is matched against unclear acoustics caused by disfluencies or noise. Both OOV words and unclear acoustics are a major source of recogniser error. A confidence measure based on can be used to reject those hypotheses which are likely to be erroneous (i.e., have a low confidence) in a hypothesis test.

Additionally, a reliable confidence measure may be of practical use in recognition search (confidence estimates may be used to order partial decoding hypotheses) [5] and in further processing of the recognition output. Confidence estimates can also be used in addition to error rate statistics when assessing the quality of the recognition model.

This paper is concerned with the use of confidence measures for hypothesis verification. Hypothesis testing and the use of a confidence measure as a test statistic are described in section 2. Confidence measures are discussed in section 3, where we define the term 'confidence measure' and describe the use of likelihood ratios for the generation of confidence measures. We introduce four acoustic confidence measures based on the estimates of local posterior probabilities produced by a hybrid Hidden Markov Model/Artificial Neural Network (HMM/ANN) large vocabulary continuous speech recognition system [7], and investigate the use of these confidence measures at both the word and phone level. Experiments were carried out using the North American Business News corpus.

2. HYPOTHESIS TESTING

A hypothesis test is a procedure which results in a decision to either accept some null hypothesis, H_0, or to reject it in favour of an alternative hypothesis, H_1. The null hypothesis is rejected if the value of the test statistic with which it is associated falls within some critical region and is accepted otherwise. In the case of a one-tailed test, the acceptance and critical regions are delineated by a single threshold value of the test statistic. Two types of error are possible when performing a hypothesis test. Firstly, the null hypothesis may be rejected when the it is in fact true—a Type I error. Secondly, the null hypothesis may be accepted when it is false—a Type II error.

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

To formulate a hypothesis test for the output of a speech recognition system, the recogniser output may be declared as the null hypothesis (H_0). In this case the alternative hypothesis H_1 is the hypothesis that a decoded phone or word did not occur. A stronger test is made if a decoding hypothesis is declared as the alternative hypothesis H_1. In this case, the null hypothesis (H_0) must be rejected for H_1 to be accepted. In order to carry out such a test, a test statistic is required. A confidence estimate for a decoding hypothesis can be used as the required test statistic. To assess the performance of a test statistic as a predictor of truth or falsity, a set of hypotheses which are known to be either true or false is required.

3. CONFIDENCE MEASURES

A confidence measure may be defined as a statistic which quantifies how well a model matches the data. In the case of speech recognition, a confidence measure may be derived from the output of both the acoustic and language models, or from either model separately. An acoustic confidence measure is one which is derived exclusively from the acoustic model. (Note that in this paper we are not concerned with computing confidence intervals, or error bars, on the output of the acoustic or language models.)

3.1. Likelihood Ratios

The use of likelihood ratios has been proposed as a method for converting the output of a 'traditional' HMM-based recogniser into a statistic suitable for use as a confidence measure [8]. A traditional HMM-based recogniser will find the word sequence model, H, which maximises the joint probability , $P(X,H)$, of the acoustic observation sequence $X = \{x^1, \dots, x^n, \dots, x^N\}$, and the model. This joint probability is related to the posterior probability of the model given the acoustics, $P(H|X)$, and the likelihood of the acoustics given the model, $P(X|H)$, by Bayes Theorem:

$$P(H|X) = \frac{P(X,H)}{P(X)} = \frac{P(X|H)P(H)}{P(X)} \quad . \tag{1}$$

When it is only required to find the model which best matches the acoustics, $P(X,H)$ is assumed to be proportional to $P(H|X)$ since the denominator, $P(X)$, is independent of the model. A consequence of omitting $P(X)$, however, is that the recogniser estimates a probability which is relative to the particular acoustic observations, X. Thus the output of a traditional HMM-based recogniser is not comparable across utterances and is therefore not an appropriate statistic to use as a confidence measure.

This difficulty has been addressed by normalising $P(X|H)$ by the likelihood of the acoustics given a 'filler' or 'garbage' model, $P(X|H_f)$. If $P(X|H)$ is considered to be the likelihood of the acoustics given the null hypothesis, H_0, and $P(X|H_f)$ is considered to be the likelihood of the acoustics given the alternative hypothesis, H_1, a hypothesis test can use the likelihood ratio shown below as a test statistic:

$$\frac{P(X|H_0)}{P(X|H_1)} = \frac{P(X|H)}{P(X|H_f)} \geq A \quad , \tag{2}$$

where A is some suitably chosen threshold, or operating point. If the likelihood of the acoustics given H_0 is sufficiently small relative to the likelihood of the acoustics given H_1, H_0 is rejected in favour of H_1, and vice versa.

In the case of a keyword spotting task, a filler model may be used to model extraneous acoustics, such as non-keywords and noise. In the case of an hypothesis verification task, a garbage model may be used as a more general acoustic model and may be trained using instances of keywords also. To increase the effectiveness of a likelihood ratio statistic as a confidence measure, a discriminative training criterion has been proposed [8]. If the garbage model is assumed to be sufficiently general so as to estimate $P(X)$, this discriminative training criterion can be compared to a Maximal Mutual Information (MMI) criterion.

3.2. Posterior Probability Estimates

Hybrid HMM/ANN recognisers are well suited to generating confidence measures. It has been shown that both multilayer perceptrons and recurrent neural networks are capable of providing good estimates of the posterior probability of a phone given some acoustic data, $P(q_k^n|x^n)$ [1, 7]. As shown in [6], these local posterior probability estimates may be combined to produce a Viterbi estimate of the global posterior probability of a word sequence given the acoustic observations, $P(H|X)$:

$$P(H|X) \simeq \max_{\text{state}-\text{seq}} \left[\prod_n P(q_k^n|q_j^{n-1}, x^n) \frac{P(q_k^n|q_j^{n-1}, H)}{P(q_k|q_j^{n-1})} \right] P(H) \tag{3}$$

$$\simeq \max_{\text{state}-\text{seq}} \left[\prod_n P(q_k^n|x^n) \frac{P(q_k^n|H)}{P(q_k)} \right] P(H) \tag{4}$$

The ABBOT system [7] is based on (4), except that the first order Markov model prior $P(q_k^n|q_j^{n-1}, H)$ is used in place of $P(q_k^n|H)$. $P(H)$ is provided by the language model.

In the search process, the posterior $P(q_k^n|x^n)$ is divided by the acoustic data prior of the phone, $P(q_k)$, giving a "scaled likelihood":

$$\frac{P(q_k^n|x^n)}{P(q_k)} = \frac{P(x^n|q_k^n)}{P(x^n)} \tag{5}$$

We have used a repeated state phone model topology is used with all transition probabilities set to either 0.5 or 0. In this case the transition probabilities are used to provide a ("pseudo-Poisson") duration model. Since the outputs of the network probability estimator are implicitly scaled by $P(X)$ they are comparable across utterances without the need for normalisation by the output of any additional garbage or filler models.

3.3. Acoustic Confidence Measures

We have used a number of confidence measures from the output of a hybrid HMM/ANN system without the need for additional filler or garbage models. Four acoustic confidence measures, derived exclusively from the acoustic model, are defined below. These confidence measures may be applied at both the phone and word levels. For convenience, we define them at the phone level (recalling that we are using repeated state phone models), with each measure providing a confidence estimate for a phone q_k which a hypothesised start time n_s end time n_e. The language model is used to constrain the search for the optimal state sequence but is not used in the computation of the confidence estimates.

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

$$\begin{aligned} CM_{sl}(q_k) &= \sum_{n=n_s}^{n_e} \log \left(\frac{p(x^n|q_k)}{p(x^n)} \right) \\ &= \sum_{n=n_s}^{n_e} \log \left(\frac{p(q_k|x^n)}{p(q_k)} \right) \end{aligned} \tag{6}$$

2. **Posterior** $CM_{post}(q_k)$ is computed by rescoring the optimal state sequence using the local posterior probability estimates, and differs from $CM_{sl}(q_k)$ by the division by the priors. This amounts to the assumption that the training data priors are correct.

$$\begin{aligned} CM_{post}(q_k) &= \sum_{n=n_s}^{n_e} \log \left(p(q_k|x^n) \right) \\ &= CM_{sl}(q_k) + (n_e - n_s) \log p(q_k) \end{aligned} \tag{7}$$

3. **Normalised Posterior** $CM_{npost}(q_k)$ is $CM_{post}(q_k)$ normalised by the duration of the phone in frames. This counteracts the underestimate of the acoustic probabilities caused by the observation independence assumption. Phone duration constraints are applied during the decoding, but do not contribute to the confidence measures.

$$\begin{aligned} CM_{npost}(q_k) &= \frac{1}{n_e - n_s} \sum_{n=n_s}^{n_e} \log \left(p(q_k|x^n) \right) \\ &= \frac{CM_{post}(q_k)}{n_e - n_s} \end{aligned} \tag{8}$$

4. **Entropy** $CM_{ent}(q_k)$ is the entropy of the K posterior phone probability estimates output by the ANN each time frame, averaged over the duration of the phone.

$$CM_{ent}(q_k) = -\frac{1}{n_e - n_s} \sum_{n=n_s}^{n_e} \sum_{k}^{K} p(q_k) \log p(q_k) \tag{9}$$

The scaled likelihood, posterior and normalised posterior probability confidence measures are based on the most probable state sequence (obtained by the Viterbi algorithm). Thus, to extend these measures to the word level, time-aligned phone hypotheses which are constituent to the word and their timings are required. $CM_{ent}(q_k)$ does not make use of the optimal state sequence; it may be extended to the word level by summing over the duration of a word hypothesised to start at time n_s and to end at time n_e.

A fifth confidence measure $CM_{lat}(q_k)$ may be calculated from a phone or word lattice [4], and is a measure of the lattice density at frame n — the number of competing phone (word) hypotheses, NCH_n. As both the language and acoustic models contribute to the lattice, this is not purely an acoustic confidence measure. The performance of $CM_{lat}(q_k)$ was used as a benchmark against which to compare the performance of the other four confidence measures.

$$CM_{lat}(q_k) = \frac{1}{n_e - n_s} \sum_{n=n_s}^{n_e} NCH_n \tag{10}$$

4. EXPERIMENTS

The Hub1 development test set of the 1994 CSRNAB evaluation was decoded using the ABBOT hybrid HMM/ANN system [7], under two conditions. The first decoding condition was performed using a 20K word pronunciation lexicon and a trigram word grammar. Both the word sequence and the corresponding phone sequence were found for this condition, using the Viterbi criterion. The second decoding condition was performed using a bigram phone grammar and no word pronunciation lexicon. Only the optimal phone sequence could therefore be found for this condition. Confidence estimates were then calculated for each decoding hypothesis using the five confidence measures described in section 3.

A time aligned reference word and phone sequence was then obtained by performing a forced alignment of the reference word transcription. This was done using the same 20K word pronunciation lexicon used for the decodings, augmented to accommodate any words from the reference transcription which would otherwise be OOV, and the local posterior phone probabilities output by the ANN at each time frame. The word and phone sequences hypothesised during the decoding were then aligned to the reference word and phone sequences so that each decoding hypothesis could be labeled as either correct or incorrect, by a marking algorithm. A number of marking algorithms were implemented. They were all found to provide similar results, although the algorithm proposed in [10], which makes use of timing information, was found to work slightly better than the rest. This algorithm compares the recogniser output with a time aligned reference transcription: in addition to substitutions, insertions and deletions, recogniser output could be marked incorrect due to bad time alignments (using a 50% overlap criterion). All the reported results were calculated using this algorithm.

Once the truth or falsity of each decoding hypothesis was known, the performance of each confidence measure as a test statistic in a test of the decoding hypotheses could be evaluated. We assessed the performance of the different confidence measures by computing the overall probability of error (Type I + Type II) in a hypothesis test. This error probability is proportional to the Classification Error Rate (CER) described in [10].

In order to make the performance differences clear between the different confidence measures, the number of true and false hypotheses in the test set were equalised for each condition. This was done by counting the number of false hypotheses for a condition and randomly selecting the same number from the set of true hypotheses for that condition. Equalising the number of true and false hypotheses had the effect of artificially raising the recogniser error rate to 0.5 for each condition. The actual error rates for the three conditions were 0.16, 0.08 and 0.28 respectively. In order to plot the graphs shown in figures 1, 2 and 3, a number of thresholds across the range of possible values for each confidence measure were tried. No instances of silence were included in the test set.

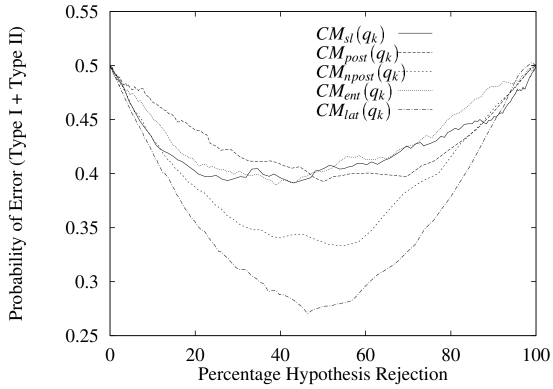

Figure 1: Hypothesis verification performance of the confidence measures at the word level using a 20K word pronunciation dictionary and a trigram word grammar on a test set of 1122 true and 1122 false hypotheses.

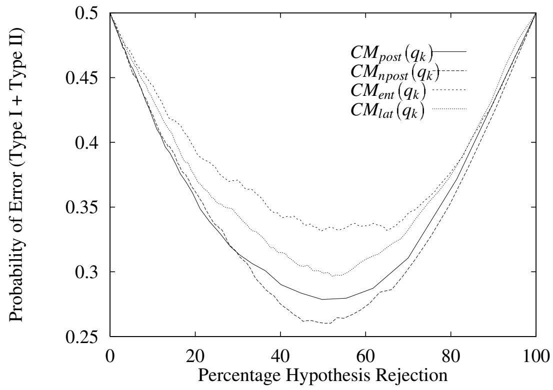

Figure 2: Hypothesis verification performance of the confidence measures at the phone level using a 20K word pronunciation dictionary and a trigram word grammar on a test set of 1965 true and 1965 false hypotheses.

5. DISCUSSION

Figure 1 shows that CM_{lat} gives the best hypothesis verification performance at the word level. This is consistent with results reported in [2, 3], where information extracted from lattices of n-best word decodings was found to be a good indicator of word confidence.

Figures 2 and 3 show that CM_{npost} is the best performing test statistic for both conditions at the phone level. These figures also show that a reduced probability of error is obtained at the phone level, where CM_{npost} gives a probability of error of 0.26 when word level constraints were used and 0.22 when only phone bigrams were used to constrain the search.

It is possible that the performance of CM_{npost} at the word level is limited due to the existence of crude pronunciation models in the pronunciation lexicon. This theory can be illustrated using an example. Figure 4 shows the local posterior phone probabilities output by the ANN over the duration of an instance of the word 'usual'. The pronunciation model for the word used in the experiments is the phone sequence, 'y uw zh uw el'. The output of the acoustic model, shown in figure 4 , suggests, however, that a better pronunciation model might be the phone sequence 'y uw zh el'. Despite the crude pronunciation model, the word is correctly hypothesised by the decoder. The confidence estimate for this correct hypothesis is thus lower than it might be, had a more refined pronunciation model been used. It can therefore be seen that the performance of CM_{sl}, CM_{post} and CM_{npost} at the word level are dependent upon the quality of the pronunciation models.

The effect of a pronunciation model also extends to the phone level. This can be illustrated using the same example. Figure 5 shows the confidence estimates for the constituent phone hypotheses of the same hypothesised word, where this phone sequence is specified by the pronunciation model. It can be seen from figure 5 that there is a very low confidence for the second hypothesised instance of the phone 'uw' in stark contrast to the good confidence estimates for the other phone hypotheses. This poor confidence is due to a poor acoustic match suggested in figure 4. A consequence of using the same pronunciation lexicon to perform the forced alignment, from which the reference phone sequence is obtained, is that the second hypothesised instance of the phone 'uw' is marked as correct, despite its low confidence. Instances such as this will compromise the per-

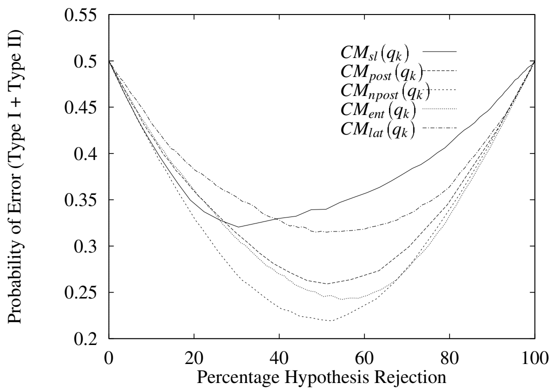

Figure 3: Hypothesis verification performance of the confidence measures at the phone level using a bigram phone grammar (and no word level lexicon or language model) on a test set of 5950 true and 5950 false hypotheses.

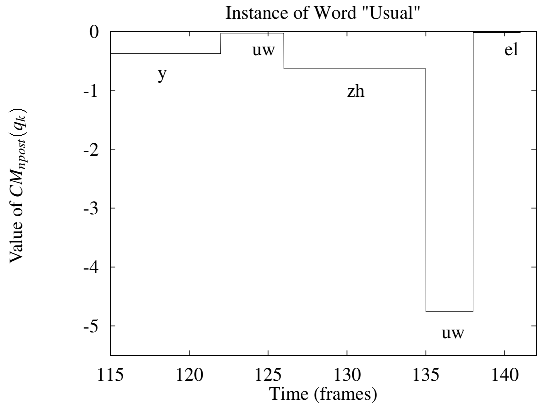

Figure 5: Confidence estimates provided by $CM_{npost}(q_k)$ for the constituent phones of the same instance of the word 'usual'.

formance of CM_{sl}, CM_{post} and CM_{npost} at the phone level and may well be responsible for a large portion of the residual error observed for the hypothesis verification experiments.

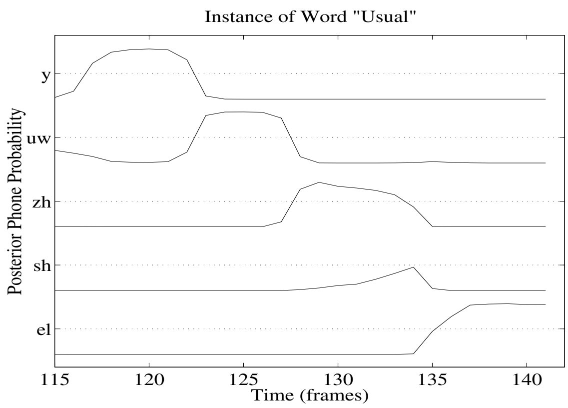

Figure 4: The local posterior phone probabilities output by the ANN over the duration of an instance of the word 'usual'.

6. SUMMARY

There are two advantages of using hybrid HMM/ANNs over likelihood ratios for computing confidence estimates. The first is that no additional models are required to normalise the output of the acoustic model since they are automatically normalised due to the discriminative training criterion. The second is that the acoustic model is trained according to the Maximum A Posteriori (MAP) criterion, which is naturally discriminative and can be preferred over the MMI criterion as it does not assume equal priors.

The best hypothesis verification performance at the word level was a probability of error (Type I + Type II) of 0.27. This was obtained using $CM_{lat}(q_k)$. The best performance at the phone level was obtained using $CM_{npost}(q_k)$. Using this confidence measure, a probability of error of 0.26 was possible for a decoding using word level constraints and 0.22 for a decoding using only phone level constraints. It is possible that the acoustic confidence measures described may be used to indicate poor pronunciation models, where improved pronunciation models lead to a reduction in the residual error for the hypothesis tests.

7. REFERENCES

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

[2] S. Cox and R.C. Rose. Confidence measures for the switchboard database *Proceedings of ICASSP–96*, 511–515. 1996.

[3] L. Gillick, Y. Ito and J. Young. A probabilistic approach to confidence estimation and evaluation *Proceedings of ICASSP–97*, 879–882. 1997.

[4] L. Hetherington. New words: Effect on recognition performance and incorporation issues. *Proceedings of EUROSPEECH–95*, 1645–1648. 1995.

[5] C. Neti, S. Roukos and E. Eide. Word-based confidence measures as a guide for stack search in speech recognition *Proceedings of ICASSP–97*, 883–886. 1997.

[6] C. Ris, J. Hennebert, H. Bourlard, S. Renals, and N. Morgan. Estimation of global posteriors and forward-backward training of hybrid HMM/ANN systems. To appear in *Proceedings of EUROSPEECH–97.*

[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*, 233–258. Kluwer, 1996.

[8] 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*, 303–329. Kluwer, 1996.

[9] 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 ICASSP–96*, 518–522. 1996.

[10] M. Weintraub, F. Beaufays, Z. Rivlin, Y. Konig, A. Stolcke. Neural - network based measures of confidence for word recognition. In *Proceedings of ICASSP–97*, 887-890. 1997.