POSTERIOR PROBABILITY DECODING, CONFIDENCE ESTIMATION AND SYSTEM COMBINATION

G. Evermann & P.C. Woodland

Cambridge University Engineering Department,
Trumpington Street, Cambridge, CB2 1PZ, UK.
Email: {ge204,pcw}@eng.cam.ac.uk

ABSTRACT

In this paper the estimation of word posterior probabilities is discussed and their application in the CU-HTK system used in the March 2000 Hub5 Conversational Telephone Speech evaluation is described. The word lattices produced by the Viterbi decoder were used to generate confusion networks, which provide a compact representation of the most likely word hypotheses and their associated word posterior probabilities. These confusion networks were used in a number of post-processing steps. The 1-best sentence hypotheses extracted directly from the networks are shown to be significantly more accurate than the baseline decoding results. The posterior probability estimates were used as the basis for the estimation of word-level confidence scores. A new system combination technique is presented that uses these confidence scores and the confusion networks and performs better than the well-known ROVER technique.

1 INTRODUCTION

Most HMM-based speech recognition systems use the sentence-level maximum a-posteriori (MAP) criterion to decide among the possible word sequence hypotheses. The sentence level MAP probability and associated word sequence can be efficiently found by using Bayes rule and the Viterbi assumption. In this approach it is assumed to be sufficient to consider only the best state-level path instead of summing over all alternative paths corresponding to a word sequence.

Recently there has been renewed interest in alternatives to the sentence MAP criterion. In [10] it was argued that this criterion is only optimal with respect to minimising the *sentence* error rate, whereas the real optimisation aim in ASR development is usually the *word* error rate. A technique for explicitly minimising the word error rate was presented in [6]. This technique uses *word-level* posterior probabilities and relies on post-processing the word lattice generated by a Viterbi decoder.

Estimates of word-level posteriors can also be used as the basis for very accurate confidence tagging of the recognition result [1]. Previously such confidence scores for the 1-best word sequence have been used to combine the output of multiple recognition systems using the ROVER technique [2]. In this paper, a new approach that is a generalisation of the ROVER technique is presented. This new approach, unlike ROVER, takes alternative recognition hypotheses into account.

The results of experiments using word posteriors are presented in the context of the CU-HTK system used in the March 2000 Conversational Telephone Speech evaluation. In the following section a brief overview of this system is given. In section 3 an overview of word posterior probabilities and their estimation from word lattices is given. Experimental results of the application of word posteriors in an improved decoding scheme are presented in section 4. The estimation of confidence scores is discussed in section 5. Finally the new system combination scheme used in the CU-HTK evaluation system is presented in section 6.

2 SYSTEM DESCRIPTION

In this section a brief overview of the CU-HTK system used in the March 2000 Hub5 evaluation is given (see [4] for details). All the experiments reported are based on this system.

The acoustic models used are triphone and quinphone HMMs trained on data from the Switchboard and CallHome corpora. Two different training criteria were used, namely the conventional maximum likelihood estimation (MLE) criterion and the maximum mutual information estimation (MMIE) criterion. The resulting four model sets (two each for triphones and quinphones) are used in different stages of recognition. Vocal tract length normalisation (VTLN) was used in training and testing. During recognition, maximum likelihood linear regression (MLLR) based speaker/channel adaptation was performed and a full-variance transform applied.

A 4-gram language model was trained on the transcripts of the acoustic training data and additional broadcast news transcriptions. The word 4-gram was smoothed with a class-based trigram that used automatically derived classes.

The system operates in a number of different stages with more complex models being applied in later stages. The first two stages are only used to determine the gender of the speaker, select a VTLN warp factor and to generate an initial transcription for use in the MLLR adaptation. In the third stage lattices are generated using the MMIE triphones and the 4-gram language model. These lattices are then individually rescored using the four different acoustic model sets (after application of MLLR and a full-variance transform), resulting in four lattices for each utterance. The lexicon used for rescoring contained pronunciation variants with unigram probabilities. The lattices generated by these four systems (referred to as P4a, P4b, P6a and P5b in [4]) are the basis for all experiments reported in this paper. All experiments were performed on the test sets used in the September 1998 and March 2000 Hub5 evaluations (eval98 and eval00, respectively).

3 WORD POSTERIOR PROBABILITY ESTIMATION

Word-level posterior probabilities are the basis for the techniques discussed in this paper. Estimates of these posteriors are derived from the acoustic and language model (LM) likelihoods of the word sequences hypothesised by a Viterbi decoder.

3.1 Lattice-based Posterior Estimation

The estimation of word-level posterior probabilities is based on the word lattices generated by a conventional Viterbi decoder. The word lattices represent the most likely part of the search space for each utterance and contain scores for a large number of competing word hypotheses. In the lattices used here, each node corresponds to a point in time and each link is labelled with a word (pronunciation) hypothesis and the associated log likelihoods from all the models used (acoustic, pronunciation and language model). As some of these scores may depend on the surrounding context (e.g. cross word acoustic models or n-gram language models) many of the links have to be duplicated, i.e. there are multiple links with the same word label and the same start and end times.

The estimation of word posteriors is performed in two stages. First the link posterior probability $p(l|\boldsymbol{X})$ is estimated for each link l. These probabilities are then combined to form word posteriors for the set of links that are considered to correspond to the same word.

The joint probability of a lattice path $\boldsymbol{q}$ (corresponding to word sequence $\boldsymbol{w}$) and the acoustic observations $\boldsymbol{X}$ is the product of the scores from the three models:

$$p(\boldsymbol{q}, \boldsymbol{X}) = p_{acc}(\boldsymbol{X}|\boldsymbol{q})^{\frac{1}{\gamma}}\ p_{lm}(\boldsymbol{w})\ p_{pr}(\boldsymbol{q}|\boldsymbol{w}) \tag{1}$$

where γ is the factor that is used to *scale down* the acoustic scores (contrary to normal practice in Viterbi decoding of scaling the LM scores) because otherwise the resulting distribution would typically be dominated by the best path. The scaling used here results in a much "flatter" posterior distribution. This form of scaling is also more appropriate from a theoretical point of view since the main effect, that scaling attempts to compensate for, is the underestimation of the acoustic likelihoods due to invalid independence assumptions.

For each link l, the joint probabilities of all paths through the link (set $\boldsymbol{Q}_l$) are summed to yield the link posterior:

$$p(l|\boldsymbol{X}) = \frac{\sum_{\boldsymbol{Q}_l} p(\boldsymbol{q}, \boldsymbol{X})}{p(\boldsymbol{X})} \tag{2}$$

This summation can be performed efficiently using a variant of the forward-backward algorithm on the lattice.

The second step in the estimation of word posteriors is the combination of links that correspond to the same word in the utterance. This decision is non-trivial and corresponds to the problem discussed in [11] in the context of N-best lists.

The approach used in the following experiments and the final evaluation system is based on the clustering procedure introduced in the framework of "consensual lattice post-processing" in [6]. An alternative texhnique based on time-conditioned word posteriors was introduced in [1].

3.2 Confusion Network Generation

The confusion network decoding technique relies on a clustering procedure that transforms a word lattice produced by a conventional Viterbi decoder into a linear graph, called a *confusion network*. All paths through this graph pass through all nodes in the same order. The links are grouped into *confusion sets* and every path contains exactly one link from each such set. The clustering is performed in two stages. In the first stage, links that correspond to the same word and overlap in time are combined (i.e. their posteriors are added and the graph topology is updated). The result of this stage is a graph that contains word posteriors.

In the second stage, links corresponding to different words are clustered into confusion sets. These sets represent competing hypotheses corresponding to the same part of the utterance. The order of clustering is based on the phonetic similarity, the time overlap and the posteriors of the words. The clustering is constrained by the order of links encoded in the original lattice and is performed until the linear graph structure is achieved. A detailed description of the clustering procedure is given in [5] and an example of such a confusion network is shown in Figure 1.

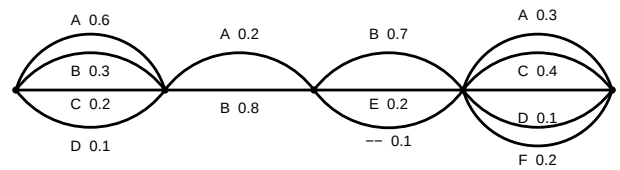

Figure 1: Example Confusion Network

Confusion networks offer a very compact representation of the most likely word hypotheses and will be used in the processing steps described in the following three sections.

4 POSTERIOR PROBABILITY DECODING

The confusion networks generated by the procedure outlined in the previous section are used in a decoding scheme that aims to find an improved 1-best sentence hypothesis (cf. [6]).

In a confusion network, each word hypothesis is labelled with its word posterior probability, i.e. the sum of the link posteriors that were combined in the clustering. The sentence hypothesis found by picking the word with the highest posterior from each confusion set can be shown to have the lowest expected word error rate (according to the posterior distribution represented in the network).

The confusion network decoding technique was evaluated in the CU-HTK Hub5 evaluation system described in section 2. For each of the four systems (triphone/quinphone and MMIE/MLE) confusion networks were generated and the hypothesis with the minimum expected word error rate was found.

triphones		eval98	eval00	
		WER	WER	SER
MMIE	Viterbi	38.5	28.4	65.7
	Confnet	37.1	27.2	65.8
MLE	Viterbi	39.3	28.8	65.4
	Confnet	38.0	27.8	65.4

quinphones		eval98	eval00	
		WER	WER	SER
MMIE	Viterbi	37.2	27.3	66.0
	Confnet	36.0	26.5	66.2
MLE	Viterbi	38.2	27.6	65.0
	Confnet	37.0	26.9	65.4

Table 1: Word Error Rates (WER) and Sentence Error Rates (SER) using Confusion Network and Viterbi Decoding

In Table 1 the word error rates for the confusion network decoding technique (Confnet) are compared against the baseline sentence-level MAP results (Viterbi). Confusion network decoding achieves a

consistent improvement of more than 1% absolute over the baseline and as expected the sentence error rate stays constant or increases as a side effect of the minimisation of the word error rate.

A relevant detail of the lattice transformation procedure as used here is that pronunciation variants of the same word (which result in multiple links in the word lattice) are recombined. As a result the word posteriors estimated in the first stage of the above procedure do not just represent the probability of the most likely pronunciation of a word but the sum over all variants. This summation over variants is very desirable from a theoretical point of view but is difficult to implement directly in a Viterbi decoder. Therefore typically only the most likely variant is taken into account. A modification to the Viterbi decoding procedure that offers a limited version of this summation has been suggested recently in [7], but only variant hypotheses ending in the same time frame are considered and the acoustic models are constrained to ignore cross-word contexts. The confusion network framework provides a more general solution to this problem.

5 CONFIDENCE SCORES

As the speech recogniser is not perfect it is often useful to annotate the words in the 1-best hypothesis with a measure of how certain the recogniser is in its decision. These word level confidence scores have many applications in the post-processing of the recogniser output (e.g. syntactic parsing, information extraction, etc.). For example all words with a confidence score below a threshold could be considered as unreliable and discarded. If such a scheme is used only the relative order of word hypotheses is relevant. In other applications making a hard decision is not appropriate and a confidence score is assumed to be the posterior probability that a word is correct[1]. Therefore it is important that the absolute values are in the correct range. The metric most commonly employed to asses the accuracy of a confidence scoring procedure is the normalized cross entropy (NCE).

The NCE is an information theoretic measure of how much additional information the confidence tags provide over the trivial baseline case of setting all scores to the (optimal) constant value p_c (corresponding to the ratio of correct words in the hypothesis: $p_c = 1.0 - \text{sub} - \text{ins}$, where sub and ins are the substitution and insertion error probabilities respectively). An NCE of zero means no additional information is contained in the confidence scores and positive values mean they provide useful extra information. See [8] for a more detailed discussion of this metric.

The word posterior probabilities that result from the confusion network clustering procedure can be used directly as confidence scores but they tend to overestimate the probabilities of correct recognition. This is due to the fact that the lattices used as the basis for the posterior estimation only represent part of the posterior distribution and a significant amount of the probability mass is "missing". Consistent with this explanation, it was found that this effect is more pronounced in systems with higher error rates and on smaller lattices. If the system has a low overall error rate then the models are able to distinguish relatively well between the correct hypothesis and incorrect alternatives, whereas for systems with high error rates the probability mass is more evenly distributed over a large number of competing hypotheses.

[1] In this context "correct" refers to the result of the standard Levenshtein scoring procedure, i.e. it depends on the exact alignment procedure and the context of the reference and hypothesis word sequences.

To compensate for the over-estimation effect we applied a piece-wise linear mapping to the lattice based posterior estimates to map them to confidence scores. This mapping function is based on a decision tree (see [1]). An alternative to this is the use of a neural network for the mapping as suggested in [9].

	posteriors		+mapping	
	eval98	eval00	eval98	eval00
Triphone MMIE	-0.034	0.191	0.238	0.294
Triphone MLE	-0.034	0.195	0.236	0.287
Quinphone MMIE	-0.132	0.135	0.224	0.284
Quinphone MLE	-0.097	0.180	0.229	0.292

Table 2: NCEs with and without Mapping

Table 2 gives the NCEs before and after the mapping. It can be seen that for the eval00 test set the unmapped posteriors perform much better than for the eval98 set. This can be explained by the fact that the system has has a much lower error rate on the eval00 set and therefore the lattices contain a larger part of the probability mass in the same number of paths.

The normalised cross entropies achieved using the lattice based estimation clearly outperform other techniques. The confidence estimation scheme used in the 1998 CU-HTK system [3] relied on an N-best homogeneity based measure and resulted in an NCE of 0.143 on the eval98 set.

The piece-wise linear mapping used is based on a small decision tree (eight leaf nodes). Table 3 shows the average confidence scores and the optimal constant score p_c. The tree used for this mapping was trained on the eval98 data. It can be seen that the discrepancy between the average confidence score and p_c is bigger on the (much easier) eval00 data, which implies that a better NCE could have been achieved by using more appropriate training data for the mapping.

	Triphones		Quinphones	
	eval98	eval00	eval98	eval00
avg. confidence	0.722	0.763	0.729	0.763
p_c=1.0-sub-ins	0.750	0.809	0.751	0.810

Table 3: Average Confidence Scores after Mapping for MMIE Systems

6 SYSTEM COMBINATION

A technique that has become very popular in recent years is the combination of the recognition output of multiple systems to produce a hypothesis that is more accurate than any of the original systems.

The most widely used technique is based on the ROVER program [2] and uses the 1-best word sequence from the different systems. These word sequences are aligned using a dynamic programming (DP) procedure similar to the one used in scoring recognition results. Based on this alignment a decision is made among the words aligned together. This decision can either be based on a simple voting scheme or take confidence scores into account. If simple voting is used, then very frequently "ties" are encountered where the same number of systems favoured two competing words. In such cases an arbitrary decision has to be made. If reliable confidence scores are available this situation is avoided and a far more accurate decision can be made.

A limitation inherent in ROVER is the restriction to the 1-best word sequences in the alignment as well as the decision procedure.

Thus only words hypotheses that were chosen by one of the systems can be picked as the final result.

Alternative hypotheses can be taken into account by using confusion networks instead of word strings in the DP alignment procedure. The local scoring function used in ROVER's DP algorithm to compare two words (simple test for word equality) is replaced by a version that calculates the probability of a word match given two confusion sets. It was found that the use of alternatives and their associated posteriors improved the quality of the DP alignment significantly.

Given the alignment of the confusion networks a generalisation of ROVER's decision procedure is used. The probability of each candidate word in the composite system is calculated as the sum of the posteriors from the component systems. In this calculation a weight can be associated with each system although in practice this was found to make only a very small difference. The candidate word with the largest weighted sum of component posteriors is picked as the final system output:

$$\hat{w} = \underset{w}{\operatorname{argmax}} \sum_{i=1}^{N} P(S_i) P(w|\boldsymbol{X}, S_i) \tag{3}$$

The confidence score used for the final word is just the average of the (mapped) posteriors of the component systems. Table 4 shows results of different system combination techniques. The four main systems used in the CU-HTK evaluation system were used (MMIE/MLE and triphone/quinphones, see [4] for details). The single best system was based on the quinphone MMIE models. It can be seen that the use of confidence scores consistently gives better performance than the simple voting scheme. The confusion network combination (CNC) technique presented here gave a a further small improvement over the use of ROVER.

		eval98	eval00	
		WER	WER	NCE
single system	Quin MMIE	36.0	26.5	0.284
2-way	Rover conf	35.6	25.7	0.267
(MMIE)	CNC	35.2	25.6	0.278
	Rover vote	35.8	25.9	
4-way	Rover conf	35.4	25.5	0.262
	CNC	35.0	25.4	0.271

Table 4: System Combination Results

The improvement over the single best system achieved by ROVER are rather disappointing especially on the eval98 test set. To investigate the interaction between the confusion network decoding and the system combination, the ROVER experiments were run on the word hypotheses produced by the Viterbi decoder (i.e. without applying the confusion network decoding).

	eval98		eval00	
	cn	no-cn	cn	no-cn
Quin MMIE	36.0	36.9	26.5	27.3
4-way ROVER vote	35.8	36.6	25.9	26.6
4-way ROVER conf	35.4	36.1	25.5	26.2

Table 5: Effect of Confnet Decoding on System Combination

The results in Table 5 show that the improvements due to ROVER are consistently slightly bigger for the case where no confusion network decoding was applied. Nevertheless the gains are almost additive.

7 CONCLUSIONS

We have discussed the estimation of word posterior probabilities and investigated applications in large vocabulary decoding, the estimation of confidence scores and system combination. A generalisation of the ROVER technique was presented that takes alternative hypotheses and their posterior probabilities into account. Experimental results were presented based on the CU-HTK conversational telephone speech evaluation system.

Acknowledgements

This work was in part supported by GCHQ. Gunnar Evermann is supported by grants from EPSRC and the Cambridge European Trust.

References

1. G. Evermann and P.C. Woodland. Large Vocabulary Decoding and Confidence Estimation using Word Posterior Probabilities. In *Proc. ICASSP 2000*, pp. 2366–2369, Istanbul.
2. Jonathan G. Fiscus. A Post-Processing System to Yield Reduced Word Error Rates: Recogniser Output Voting Error Reduction (ROVER). In *Proc. IEEE ASRU Workshop*, pp. 347–352, Santa Barbara, 1997.
3. T. Hain, P.C. Woodland, T.R. Niesler, and E.W.D. Whittaker. The 1998 HTK System for Transcription of Conversational Telephone Speech. *Proc. ICASSP'99*, pp. 57–60, Phoenix.
4. T. Hain, P.C. Woodland, G. Evermann, and D. Povey. The CU-HTK March 2000 Hub5E Transcription System. In *Proc. Speech Transcription Workshop*, 2000.
5. L. Mangu. *Finding Consensus in Speech Recognition*. PhD Thesis, Johns Hopkins University, 2000.
6. L. Mangu, E. Brill, and A. Stolcke. Finding Consensus Among Words: Lattice-Based Word Error Minimization. In *Proc. Eurospeech'99*, pp. 495–498, Budapest.
7. H. Schramm and X. Aubert. Efficient Integration of Multiple Pronunciations in a Large Vocabulary Decoder. In *Proc. ICASSP 2000*, pp. 2169–2172, Istanbul.
8. M. Siu, H. Gish, and F. Richardson. Improved Estimation, Evaluation and Application of Confidence Measures for Speech Recognition. In *Proc. Eurospeech'97*, pp. 831–834, Rhodes.
9. A. Stolcke, H. Bratt, J. Butzberger, H. Franco, V. R. Rao Gadde, M. Plauché, C. Richey, E. Shriberg, K. Sönmez, and J. Zheng F. Weng. The SRI March 2000 Hub-5 Conversational Speech Transcription System. In *Proc. Speech Transcription Workshop*, 2000.
10. A. Stolcke, Y. König, and M. Weintraub. Explicit Word Error Minimization in N-Best List Rescoring. In *Proc. Eurospeech'97*, pp. 163–166, Rhodes.
11. M. Weintraub. LVCSR Log-Likelihood Ratio Scoring for Keyword Spotting. In *Proc. ICASSP'95*, pp. 297–300, Detroit.