META-MODELS FOR CONFIDENCE ESTIMATION IN SPEECH RECOGNITION

Srinandan Dasmahapatra and Stephen Cox

School of Information Systems, University of East Anglia, Norwich NR4 7TJ, U.K.
{sd, sjc}@sys.uea.ac.uk

ABSTRACT

We describe an approach to confidence estimation that attempts to decouple the contributions of the acoustic and language model components to speech recognition output. The output of the acoustic models when decoding phonemes is itself modelled using HMM's to produce a set of models which we term *meta-models*. When benchmarked against a "standard" method for assigning confidence (the N-best score), the meta-models gave a relative improvement of 6.2%. Furthermore, it appears that the N-best and meta-models techniques are complementary, because they tend to fail on different words.

1. INTRODUCTION

Systems which employ speech recognition to facilitate a dialogue with a user are increasingly being deployed. As the tasks which these systems attempt to perform automatically grow in complexity, it becomes imperative that the system responds intelligently to avoid a protracted dialogue. Confidence measures associate a probability of correct decoding with each output item, which aids the system to infer information reliably from the spoken input and to request confirmation or repetition of an item only when there is insufficient confidence in its identity. They can also be used to facilitate adaptation of the speaker's voice to the system.

Many approaches to deriving confidence measures (CM's) for words have been based on using "side-information" derived from the recogniser, such as likelihoods [6], different decodings [5], number of competitors at the end of a word [3] etc. In our own experience, we have found that measures that perform well on a given decoder do not always work when used with another decoder, even though the decoders may be very similar in design. With the appearance of speech API's that are effectively black boxes, we think that an approach that relies less on the details of the recogniser might be useful.

The initial objective in this and in previous work [4] is to isolate the language and acoustic modelling components of the recogniser in order to assess separately the evidence for decoding a particular segment of the speech as a sequence of words. This approach points the way towards a system-independent method for computing a confidence score. Our approach is to use a phone recogniser in parallel with the word recogniser to derive some independent information (a similar approach was used in [1, 2]). In a previous paper [4], we investigated the effectiveness of correlating the phone strings available from the word and phone recognisers, and also of using word hypotheses formed from the phone recogniser output. Here, we have extended this work by modelling the errors made by the phone recogniser within an HMM framework.

A secondary objective is to provide a measure which is in some sense complementary to the currently most robust and consistent technique for obtaining confidence measures, the N-best technique. Most systems now provide "N-best" hypotheses, which enable easy computation of a confidence measure for a word in any hypothesis, based on its frequency of occurrence in corresponding positions in the other hypotheses [5]. The N-best confidence measure relies on the principle that the recognition process optimally incorporates and implements contextual effects over the length of the utterance, combining the language and acoustic modelling components.

The outline of the paper is as follows. We describe the method in section 2 and then in section 3, outline the details of the training and testing procedures, and the data sets used. Section 4 is devoted to results, and we end with a brief discussion and summary.

2. METHOD

For speech recognition, we attempt to find the word sequence $\mathbf{w} := w_1 w_2 \dots w_N$ for which the probability $P(\mathbf{w}|A)$ is largest among all word-sequences from the vocabulary V, conditioned on the front-end speech signal A. If we write $P(\mathbf{w}|A)$ as

$$P(\mathbf{w}|A) = \sum_{\mathbf{p}} P(\mathbf{w}|\mathbf{p}) P(\mathbf{p}|A), \tag{1}$$

where $\mathbf{p}$ is an arbitrary sequence drawn from a discrete alphabet of phonemes $\mathcal{P}$, we note that the two terms in the summand can be estimated from a phoneme recognition problem. The second term $P(\mathbf{p}|A)$ is used in a phoneme classification task, while the first term can be estimated using a language model and Bayes theorem (see equation 5).

Each phoneme $p \in \mathcal{P}$ is given a list of dictionary "pronunciations" which are the labels of the distinct hidden Markov models $P(A|x-p+y)$ used by the complete word recognition system, where $x-p+y$ is a triphone (with appropriate context monophones x and y). The transition probability between every pair of monophones is set equal. This arrangement ignores word-internal phonotactic constraints as well as lexical (uni- or bi-gram) probabilities which are combined in the word decoding task. The phone sequence $\mathbf{p} := p_1 p_2 \dots p_N$ is chosen for which $P(\mathbf{p}|A)$ is larger than for any other such sequence:

$$\mathbf{p}^* = \text{argmax}_{\mathbf{p}} \quad P(\mathbf{p}|A) \tag{2}$$

The right-hand side of equation (1) involves a sum over all sequences of phonemes drawn from the alphabet $\mathcal{P}$. As a way of isolating the word-dependent probabilities (inter-word, given by

the language model, as well as word-internal phonotactics), we approximate eq(1) by

$$P(\mathbf{w}|A) \approx P(\mathbf{w}|\mathbf{p}^*)P(\mathbf{p}^*|A). \quad (3)$$

A comparison of this output with the reference transcription provides an assessment of the global (averaged over all contexts) performance of the acoustic component of the recogniser as a phoneme classifier. (This may be viewed as a "prior" that can be incorporated with the acoustic probabilities for each word in the recognition lattice for the full word recogniser.) The phoneme confusion matrix obtained thus encodes the probability of the actually uttered (reference) phoneme q, given that the decoded phoneme was p^*, $P(\text{reference-}q|\text{decoded-}p^*)$.

If we perform the phonemic expansion of the word-string $\mathbf{w}$ as $\pi(\mathbf{w}) = \pi_1\pi_2\ldots\pi_M$ (*e.g.* for `w she had your ...`, $\pi(\mathbf{w})$ is `sh iy hh ae d y ao ...`), we can now approximate eq (1) by replacing the right-hand-side of eq(3) by

$$P(\mathbf{w}|A) \approx P\big(\pi(\mathbf{w})|\mathbf{p}^*\big)P(\mathbf{p}^*|A). \quad (4)$$

We can evaluate the probabilities to go back and forth between the phone stream for the reference transcription $\pi_1(\mathbf{w})\pi_2(\mathbf{w})\ldots\pi_M(\mathbf{w})$ and the decoded one $p_1^*p_2^*\ldots p_N^*$, as the product of probabilities of making substitutions, insertions and deletions, $P(\pi_k|p_l^*)$, $P(-|p_j^*)$ and $P(\pi_i|-)$. Instead, we regard this problem within a generative framework, and rewrite $P\big(\pi(\mathbf{w})|\mathbf{p}^*\big)$ using Bayes' rule

$$P\big(\pi(\mathbf{w})|\mathbf{p}^*\big) = \frac{P(\mathbf{p}^*|\mathbf{w})P(\mathbf{w})}{P(\mathbf{p}^*)}. \quad (5)$$

(Note, that the above is strictly true only if there is one pronunciation per word, *i.e.*, this requires that we find the most appropriate pronunciation while training.) Our objective being confidence estimation and not word recognition, we constrain the lexical probabilities to coincide with those used in the word recognition task, and this includes the power α in $P(\mathbf{w})^\alpha$. In embedded Baum-Welch re-estimation, a chain of states is formed by linking HMMs in a row. As a result, the transition probabilities between HMMs (those corresponding to the edges linking the in and out states) are averaged over all occurrences of the phonemes in the corpus. The scale factor α is included to override this effect, and is tuned to optimise recognition accuracy. We can now extract the "confusion probabilities" from

$$P\big(\mathbf{p}^*|\pi(\mathbf{w})\big) \quad (6)$$

within a hidden Markov model framework. Since $\mathbf{p}^*$ is obtained from the phonemic classification performed by the acoustic models, the probabilities in equation(6) model the performance of the acoustic models used for recognition. Hence, we term them *meta-models.*

Each phoneme has more than one state associated with it in the underlying Markov chain. Since the output of a phone recogniser always contains many more discrete outputs than the number of phonemes in the reference word string, there are enough reference tokens with which to align the output stream. Insertions are thus easily modelled, while deletions are accommodated if the number of output phonemes exceeds the number for reference phonemes in the framework of embedded Baum-Welch re-estimation. The discrete output probability distributions of the hidden Markov model encode the substitution probabilities, $P(\text{decoded-phone } p_j^*|\text{reference-state of } \pi_k)$. Figure (1) is a schematic representation of the method.

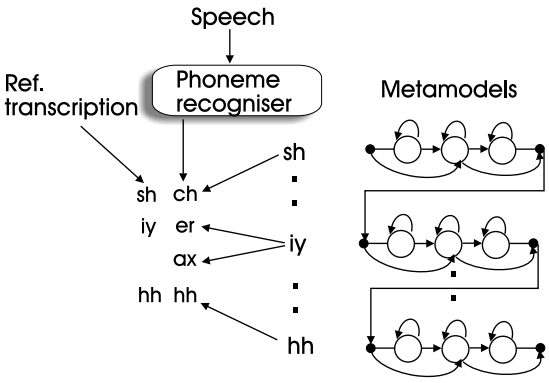

Figure 1: Schematic diagram for the estimation of meta-models from the discrete phonemic output of recogniser. $\mathbf{p}^*$ starts off as `ch er ax hh` while $\pi(\mathbf{w})$ begins with `sh iy hh`.

Once these meta-models have been estimated, we can perform a Viterbi decoding to obtain another N-best list of hypotheses for the best word-strings that match the phoneme output of the unconstrained phone-recogniser. Since our objective in this paper is not so much recognition as confidence estimation, we shall ultimately do all our calculations with the value of α set by the requirement of obtaining the best recognition in the full continuous speech recogniser. We tag the words that are decoded by the speech recogniser as correct or incorrect depending on whether they *appear more than once* in the top 100 decodings of the meta-model decoder.

To summarize the method, we take as input the decoded phoneme stream from a phone recogniser (eq(2)) and train discrete hidden Markov models using embedded Baum-Welch re-estimation. For test data, which is (again) a phoneme stream from a phone decoder, we find N-best word strings that maximize the left-hand side of eq(3). We then find the number of occurrences of the decoded words from the word recogniser to set a confidence tag.

3. DATA

3.1. Speech recogniser

Our baseline recogniser has been trained on speech data from the WSJCAM0 data-set using mainly "standard" techniques implemented in the Entropic HTK package. The specifications of the recogniser are as follows:

1. Trained on the speaker-independent training set `si_tr` of WSJCAM0 (92 talkers, $\sim$ 90 utterances per speaker)
2. Number of words in vocabulary $\sim$ 20000
3. Bigram language model (trained on 60M words from the North American business news corpus), perplexity $\sim$ 160
4. 3500 states created by tree-clustering word-internal triphones; 8 Gaussian mixture components per state
5. 3-state left-to-right models
6. Test set used: the speaker-independent development set `si_dt` in WSJCAM0, $\sim$ 1800 utterances
7. Current performance: 74.0% correct, 68.2% accurate.

3.2. Meta-model confidence measure

1. Trained on recognition performance of $\sim$ 1400 utterances of 15 speakers from the `si_dt` test set chosen above
2. 2- and 3-state hidden Markov models with skips, to capture substitutions and insertions and deletions; discrete output probability functions
3. Tested on $\sim$ 400 decoded utterances of 5 speakers also from `si_dt` set above.

3.3. *N*-best confidence measure

1. Partition the $\sim$ 1800 decoded utterances from `si_dt` in the same way as for the meta-model confidence measure
2. Set threshold for acceptance/rejection by maximising tagging (of correct/incorrect decoding) accuracy on the training set

4. RESULTS

We present the results for the confidence score using meta-models in tandem with those obtained using *N*-best. This will give an indication of the merits of this method, and also indicate ways in which the method can be improved. A measure based on guessing every word as correct would give a tagging error equal to the baseline recognition error of the recogniser, which is 31.8%. The confidence tagging error for the *N*-best measure is 23.9% (24.8% improvement in tagging error) and for the meta-model measure is 21.9% (31.1% improvement in tagging error). On the subset of words for which both measures gave a tag (some of the utterance files had to be discarded because the pruning thresholds for recognition were set too tightly) we list the performance in the Table 1 below.

	meta-model C	meta-model I
N-best C	4413	1120
N-best I	1225	463

Table 1: Comparison of confidence measures.

Table 1 lists the number of words tagged correctly (C=tagged correct for correct decoding, tagged incorrect for incorrect decoding) or incorrectly (I=tagged correct for incorrect decoding, tagged incorrect for correct decoding) for each of the two confidence measures. For example, there are 1225 words which are mis-tagged by *N*-best, but correctly tagged by the meta-model confidence measure. It is promising that only 463 of the 7221 words listed above were mistagged by both measures, indicating an upper bound of 6.4% tagging error over the baseline guessing measure (31.8% error) possible by some combination of the two features. A further breakdown of these figures in order to compare the performance of each tag-pair is given below.

N-best tag	meta-model tag	prob C%	prob I%
C	C	91.8	9.2
C	I	50.3	49.7
I	C	59.2	40.8
I	I	14.5	85.5

Table 2: Percentage of correct and incorrect words (columns 3 and 4) compared with prediction of two confidence measures.

Table 2 shows that when both confidence measures tag a word as correct, there is a 91.8% chance that the word is correct. Conversely, when both tag incorrect, there is a 85.5% chance that the word is incorrect.

We also plotted receiver operating curves for the *N*-best and for the meta-model confidence methods. This is shown in Figure (2) below. Note that the meta-models approach does not eliminate false alarms entirely because not all the words in the word recogniser output string appear in the word strings hypothesized by the meta-models plus language model.

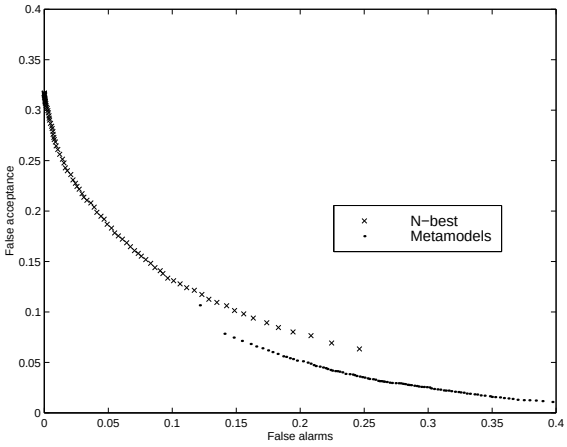

Figure 2: False acceptances *vs.* false alarms at different thresholds. The dots are for the meta-model operating points whereas the pluses (+) are for *N*-best.

5. DISCUSSION

We have described a method of obtaining a confidence score on words output by a recogniser by modelling the output from a parallel phoneme recogniser with a "higher-level" HMM to place a probability on the correctness of each decoded phoneme being correct. The resulting confidence measure is a little better than that obtained using the *N*-best technique we have been using as a benchmark. An obvious way in which this work can be extended is to combine the two methods. However, an analysis of the figures in Table 2 shows that using the tags from the two classifiers and marking words as 'C' or 'I' according to columns 3 and 4 of the table gives an improvement of only 0.5% over the performance obtained using meta-models alone. However, the fact that there is considerable independence between the two classifiers in the tagging (as shown in Table 1) suggests that the outputs could be used together in a scheme that relies on further information about each classifier decision. In both the meta-models and *N*-best score, we only counted the frequency of occurrence of the words, not the probabilities. The differences in log-likelihoods between the hypotheses might give not just a more accurate confidence measure, but also, these scores might provide more useful clues for combining the features of the two approaches outlined.

The approach described here has focussed on examining the performance of the acoustic models to provide confidence measures. In line with our philosophy of decoupling the acoustic and language modelling components of the recogniser, we are

currently examining the semantic coherence of the words decoded as a means of obtaining confidence, an approach that is complementary to the sublexical focus of this work.

ACKNOWLEDGMENT

This work was funded by a grant from the UK Engineering and Physical Sciences Research Council.

6. REFERENCES

[1] A. Asadi, R. Schwartz, and J. Makhoul. Automatic detection of new words in a large vocabulary speech recognition system. In *Proc. IEEE Conf. on Acoustics, Speech and Signal-processing*, pages 125–128, 1990.

[2] M.C. Benitez et al. Word verification using confidence measures in speech recognition. In *Proc. 5th International Conference on Speech Communication and Technology*, pages 1082–1085, November 1998.

[3] S.J. Cox and R.C. Rose. Confidence measures for the SWITCHBOARD database. In *Proc. IEEE Conf. on Acoustics, Speech and Signal-processing*, pages 511–515, 1996.

[4] S.J. Cox and Dasmahapatra S. A high-level approach to confidence estimation in speech recognition. In *Proc. 6th European Conf. on Speech Communication and Technology*, pages 41–44, September 1999.

[5] L. Gillick, Y. Ito, and J Young. A probabilistic approach to confidence estimation and evaluation. In *Proc. IEEE Conf. on Acoustics, Speech and Signal-processing*, April 1997.

[6] T. Schaaf and T Kemp. Confidence measures for spontaneous speech recognition. In *Proc. IEEE Conf. on Acoustics, Speech and Signal-processing*, April 1997.