A HIGH-LEVEL APPROACH TO CONFIDENCE ESTIMATION IN SPEECH RECOGNITION

Stephen Cox and Srinandan Dasmahapatra

School of Information Systems, University of East Anglia, Norwich NR4 7TJ, U.K.

{sjc,sd}@sys.uea.ac.uk

ABSTRACT

Errors in the output of a speech recogniser can be said to be due to the interaction of inadequate phonetic and language modelling components. We investigate an approach to estimating confidence scores for the words output by a recogniser in which the language modelling and acoustic modelling are decoupled by the use of a phone recogniser working in parallel with the word recogniser. An advantage of such an approach is that it avoids techniques which rely on the use of side-information derived from the decoder: such information may depend heavily on the type and configuration of the decoder used. We have investigated two ways of using the additional information provided by the phone-loop recogniser. One is based on correlating the phone strings from the two recognisers; the other is based on using the phone-loop recogniser output to construct hypotheses for the utterance and correlating these hypotheses with the word recogniser output.

1. INTRODUCTION

There has recently been considerable research activity in the field of confidence estimation, for instance [4–6, 10]. There are several motivations for attaching a measure of confidence to the words output by the recogniser: it can be used to improve the efficiency of a speech dialogue/understanding system by requesting confirmation or re-input only when necessary, for detection of out-of vocabulary (OOV) words, to aid unsupervised speaker adaptation etc.

Many approaches to deriving confidence measures (CM's) for words have been based on using side-information derived from the recogniser, such as likelihoods [10], different decodings [6], number of competitors at the end of a word [5] etc. This information is often used as a feature vector for a classifier that attaches a probability to each output word that it is correct e.g. [5]. However, in some cases, it may not be possible to obtain such information (for instance if one is using a proprietary recogniser). Also, it has been our experience that the performance of some of the features derived from side-information varies from recogniser to recogniser depending on details of the design of the decoder. Our motivation in developing the techniques described here is to make the estimation of confidence measures less recogniser-specific by using extra information in which the acoustic and language models are decoupled. If side-information *is* available from the recogniser, the information generated using these approaches can be combined with it to generate more powerful CM's.

One possibility for a more general approach to confidence estimation for the output of a word recogniser is to use an unconstrained phone recogniser which works in parallel with the word recogniser. The question is then how to make the best use of this extra information. One approach has been to correlate the two phone strings, one obtained from the word recogniser and one from the phone recogniser [1, 2]. We have tested some techniques for correlating the strings and present results in the first part of this paper. However, we have also used a more sophisticated technique in which we form a set of word sequence hypotheses for the utterance using the output of the phone recogniser and compare these with the top hypothesis from the word recogniser. A confidence measure for each word output by the word recogniser can then be derived.

2. TECHNIQUES

2.1. Overview

The two approaches we have investigated to make use of a phone recogniser output working in conjunction with a word recogniser are as follows:

1. The output word string is decomposed into a phone string using the dictionary and is then correlated with the string decoded by the phone recogniser. A high correlation between the phones of a decoded word and the corresponding phones output by the phone recogniser should indicate that the word is correct (section 2.2).
2. The string decoded by the phone recogniser is analysed using a moving windows of fixed length. At each window position, a list of putative words is made, and these lists are analysed to give an estimate of the confidence in the decoded word at that position. A similar approach has been used for selection of words from a phonetic string [9]. (section 2.3).

We benchmarked these methods against a technique which is known to give good performance, the "Nbest-score" [6].

2.2. Phone correlation techniques

Two methods for aligning the output from the word- and phone recognisers were employed.

1. **Frame-level, FL**. The phone segmentation provided by the recogniser was used to tag each frame decoded by the word recogniser with the identity of the phone decoded at this point. Similarly, the frames decoded by the phone recogniser were tagged. Hence the tags of a sequence of frames corresponding to a decoded word could be compared on a frame by frame basis.

2. **Phone-level, PL**. Using dynamic programming, the complete sequence of phones decoded by the word recogniser was aligned to the complete sequence of phones decoded by the phone recogniser. The sequence of phones corresponding to a particular word decoded by the word recogniser was then compared with the corresponding sequence in the output of the phone recogniser. Note that "insertions", by either recogniser, are possible in this case.

Once the two phone-strings were aligned, two methods of comparing them were used. Suppose that the tag-sequence (FL) or phone-sequence (PL) corresponding to the i'th word decoded by the word recogniser w_i is $q_1, q_2, \ldots, q_{N(i)}$ and the sequence decoded by the phone recogniser is $p_1, p_2, \ldots, p_{N(i)}$, where p_j and q_j can take on the values $P_1, P_2, .., \ldots, P_M$, where M is the number of phonemes.

1. **Distance-measure, DM**. The confusion-matrix generated from recognising the training-set using the phone recogniser provides the conditional probability $Pr(Y = P_m | X = P_n)$ of decoding a phoneme as P_m when the actual phoneme is P_n. We used the distance-matrix here to provide a measure of the inverse "distance" between the two aligned phonemes p_j and q_j (without any reference to whether either phoneme was correct or incorrect) on the grounds that confusions between two phonemes which are "close" are more likely than between two which are far apart. Hence we compute $DM(i) = \sum_{j=1}^{N(i)} Pr(q_j = P_m | p_j = P_n)$ as a measure of the inverse distance between the two decodings for word w_i and use this as a confidence measure.
2. **Likelihood-ratio, LR**. It was suggested that there may be common co-occurrences of phones in the two decodings which signal regions in the output which have been correctly or incorrectly decoded. Using the training-set to note the co-occurrences of $q_j = P_m, p_j = P_n$ in both correct (C) and incorrect (I) words, we estimate the likelihood ratio L_j for this pair of phones:

$$\begin{aligned} L_j &= \frac{\Pr(C | q_j = P_m, p_j = P_n)}{\Pr(I | q_j = P_m, p_j = P_n)} \\ &= \frac{\Pr(C)}{\Pr(I)} \frac{\Pr(q_j = P_m, p_j = P_n | C)}{\Pr(q_j = P_m, p_j = P_n | I)}. \end{aligned} \quad (1)$$

Our confidence measure for word w_i is then $LR(i) = \sum_{j=1}^{N(i)} \log L_j$.

Using the training-data, these measures were computed for each word and histograms of the values for 'C' and for 'I' utterances were estimated. These were then used in Bayesian classification of the test-set words as either 'C' or 'I'.

2.3. Use of lists of lexical items

Correlating phone-strings is useful for finding regions in the utterance where the language model has "overruled" the acoustic matching to cause an error, but does not address the problem of a phone string which is correct but which has been incorrectly segmented (e.g. "fee mail" instead of "female"). In the method described in this section, we "invert" the action of the recogniser and reconstruct utterances that would be compatible with the output phoneme streams of the word-recogniser and phone-loop. We first invert our lexicon of 20,000 words as a hash table. The keys to this table are phoneme strings of a certain length and the values are words whose phonemic spellings contain the key. For example, using a window of length 3, the key *ah n y* has the values "*unused 6; unusual 7; unusually 8; netanyahu 9; netanyahu's 10*", where the numbers indicate the number of phonemes in the spelling of the word. We then slide a window over the phoneme string and list the words that match the string at each window, retaining only those that are consistent with the spelling. So, for example, if we encounter the string ...ah n y uw zh ua l iy... we first list the words that match *ah n y*, and for each word, see if it occurs in the following windows. If a word of length n appears in ($n-$windowlength $+1$) windows, then clearly it is a perfect match and a candidate word. This scheme allows for possible segmentations of the phoneme stream coming from the word recogniser, not necessarily a partitioning of the stream.

We use properties of these sets of "cohorts"—putative words that match a segment of the phoneme stream—to generate features for confidence annotation. Instead of using word lattices which include both language model and acoustic likelihood information, we prefer to use a confusion matrix. The motivation here is that confusion matrices encode information on (mis-)classification, unlike a maximum likelihood approach. Furthermore, a large number of words that match a window may be indicative of a region of high confusion. We attempt to incorporate these two pieces of information in the following way.

First we gather statistics on which of the phone- or word-recogniser phoneme is likely to be correct given a DP-aligned pair (generated by the PL method described in section 2.2). This is used to assign a probability to a window (assuming independence). In other words, if win_i is a window of length 3 containing the phoneme sequences $/q_i\ q_{i+1}\ q_{i+2}/$ from the word-recogniser, and $/p_i\ p_{i+1}\ p_{i+2}/$ from the phone-recogniser,

$$\Pr(\text{win}_i) = \prod_{k=i}^{i+2} \Pr(q_k = C | q_k, p_k) \quad (2)$$

gives the probability of a window being correctly recognised. The sum of the the probablities of the windows over which a word survives is a signature of word correctness. Next, we measure the "volume" of word probability space occupied by the cohorts assigned to each window and the contribution of each word relative to this volume. Let

$$\Omega_i = \{\text{words that match win}_i\} \quad (3)$$

and

$$F_i(\beta) = -\frac{1}{\beta} \ln \left(\sum_{w \in \Omega_i} p_w^\beta \right) \quad (4)$$

. Then for each word $w \in \Omega_i$ we compute

$$\delta_i(w) = \ln p_w - F_i, \quad (5)$$

which is positive by convexity of the logarithm. So far, we have only worked with $F_i = F_i(\beta = 1)$ and we intend to investigate optimal estimation of β. The size of the number F_i is a measure of how likely the window is and $\delta_i(w)$ is a measure of the likelihood of the word w within the window.

We also intend to incorporate the simultaneous integration of the window of phonemes from the phone-loop recogniser,

Phonemes in window	Active words	# windows	Window prob	Window entropy
ah n y	unusually unusual a	6 5 < 1	0.038	3.88
n y uw	unusually unusual knew you	6 5 1 < 1	0.012	5.62
y uw zh	unusually unusual usual you	6 5 3 < 1	0.013	6.92
uw zh ua	unusually unusual usual	6 5 3	0.028	9.18
zh ua l	unusually unusual usual	6 5 3	0.029	9.18
ua l iy	unusually e lee li ee	6 < 1 < 1 < 1 < 1	0.015	8.47

Table 2: Sample of sliding window features.

which would give a different set of possible words for each window from the straight re-segmentation routine outlined above. The implementation of this method and a detailed analysis of our sliding window approach is still in progress. A sample of the kind of results that this algorithm provides is included in Table 2. Column one shows the phonemes in the current window, column two the "active" words (words that have a string of phonemes that matches or part-matches the window phonemes and are also consistent with previous and subsequent windows). In column three we show the number of windows that each active word has appeared and will appear in (< 1 indicates a partial match within this window). The active word identities shown in column two are sorted by this number combined with the unigram language model probability (not shown). Columns three and four show the window probability (equation 2) and the window entropy F_i (equation 4).

We are currently investigating combining the window probability with the window entropy to form a confidence measure.

3. DATA AND MODELS

We conducted our experiments with a subset of the WSJCAM0 database [8]. All the speech data used was parameterised to a 39-d vector consisting of 12 MFCC's + velocity + acceleration coefficients and a log energy value. The complete training-set, consisting of a total of about 90 hours of speech from 92 speakers, was used to train a set of three state, left-right HMM triphone models which had a Gaussian mixture model of 8 components associated with each state. Tree-clustering was used to reduce the total number of physical HMM states to about 3000. The language model used was a 20000 word bigram model, with back-off, trained using the CMU-Cam-Toolkit(v2) kit. The model did not include all the words in the WSJCAM0 test-set and there were about 2500 OOV words. We used 1826 sentences from the development set of the database and divided these into 913 for training our confidence estimators and 913 for testing them. Each word in the recognition output from each sentence was tagged as 'C' (correct) or 'I' (incorrect) before being used in the CM experiments. The HMM training/decoding software used for all experiments was HTK v2.2. [7].

4. RESULTS

Various methods for rating confidence measures have been proposed [3], some of them more or less difficult to interpret. A straightforward measure is the classification error-rate (CER) [11], which is the percentage of words misclassified as either 'C' or 'I'. To relate this back to the baseline performance when no confidence measure is used, we use the percentage improvement in CER over guessing provided by the CM.

The baseline performance of our word recogniser is 74% accuracy and 64.2% correct. However, for CM's, we are required to classify each word in the output as 'C' or 'I', and in these terms, the "error-rate" (#I/(#C+#I)) is 31%, which corresponds to the CER which would be obtained by guessing every word as 'C'. Table 1 shows the performance of the systems.

Technique	% improvement in CER
FL + DM	2.7
FL + LR	2.3
PL + DM	3.9
PL + LR	5.7
Nbest	22.1

Table 1: Performance of techniques discussed in section 2

It appears that using a DP phone alignment (PL) of the two phone strings is superior to using a frame-level alignment and this is most effective when used in conjunction with the likelihood ratio (LR) to compare aligned phones. However, none of the techniques using a parallel phone recogniser came close to the performance of the Nbest technique. A result not shown in the table is that when the word recogniser used a different set of phoneme models (which had a slightly lower baseline performance, 34.8% CER) from the models used in the phone recogniser and the PL+LR technique was used, an improvement of 36.5% in the CER was obtained. However, at time of writing, we cannot be sure whether this improvement is due to the independence of the output from the two recognisers, or simply to the fact that the recognition performance of the models in the word recogniser was lower than the models in the phone recogniser and so incorrect words were easy for the phone recogniser to "spot".

5. DISCUSSION

We have discussed the possibility of using a phone recogniser in parallel with a word recogniser to decouple the acoustic and language models and hence provide independent information for use as a confidence measure. Although this seems attractive, our results show that the technique is not as effective as the Nbest technique, which provides different hypotheses with the acoustic and language models integrated. We are investigating the possibility of using a phone recogniser which uses

a different set of phone models from those in the word recogniser and hence provides more independent information. Another factor that may affect the performance of this technique is the relative accuracies of the acoustic and language modelling components in the recogniser and how they are balanced. A more sophisticated approach to combining the outputs from the recognisers in which we attempt to produce word hypotheses from the phone string and correlate these with the output word string looks promising and we are continuing to develop this. This technique is clearly similar to Nbest, but has flexibility in balancing acoustic and language model probabilities and also makes use of information about the probability of confusion of phonemes, which Nbest does not. In the longer term, we are interested in the possibility of using higher levels of information for confidence estimation e.g. syntactic and semantic information.

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] L. Chase. *Error-responsive feedback mechanisms for speech recognisers*. PhD thesis, Carnegie Mellon University, September 1997.

[4] L. Chase. Word and acoustic confidence annotation for large vocabulary speech recognition. In *Proc. 5th European Conference on Speech Communication and Technology*, pages 815–818, September 1997.

[5] 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.

[6] 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.

[7] J. Jansen, J. Odell, D. Ollason, and P. Woodland. *The HTK book.* Entropic Research Laboratories Inc., 1996.

[8] T. Robinson et al. WSJCAM0: A British English speech corpus for large vocabulary continuous speech recognition. In *Proc. IEEE Conf. on Acoustics, Speech and Signal-processing*, pages 81–84, 1995.

[9] R.R. Sarukkai and D.H. Ballard. Phonetic set indexing for fast lexical access. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(1):78–82, January 1998.

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

[11] M. Weintraub et al. Neural-network based measures of confidence for word recognition. In *Proc. IEEE Conf. on Acoustics, Speech and Signal-processing*, April 1997.