CONFIDENCE MEASURES FOR EVALUATING PRONUNCIATION MODELS

Gethin Williams and Steve Renals

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

1. ABSTRACT

In this paper, we investigate the use of confidence measures for the evaluation of pronunciation models and the employment of these evaluations in an automatic baseform learning process. The confidence measures and pronunciation models are obtained from the ABBOT hybrid Hidden Markov Model/Artificial Neural Network (HMM/ANN) Large Vocabulary Continuous Speech Recognition (LVCSR) system [8]. Experiments were carried out for a number of baseform learning schemes using the ARPA North American Business News (NAB) and the Broadcast News (BN) corpora from which it was found that a confidence measure based scheme provided the largest reduction in Word Error Rate (WER).

2. INTRODUCTION

A confidence measure may be defined as a function which quantifies how well a model matches some acoustic data, where the values of the function must be comparable across utterances. More specifically, an *acoustic* confidence measure is one which is derived exclusively from an acoustic model. As an acoustic confidence measure can be used to measure the quality of the match between a word model and the acoustic realisations of that word, independently of any language model constraints, such a measure is well suited to the evaluation of a pronunciation model.

A common approach to evaluating pronunciation models is to align the subword class sequence output by the recogniser (using full word level decoding constraints) against an alternative subword sequence, obtained without any pronunciation model constraints. In this case, a poor pronunciation model is signalled by a portion of the alignment where the class labels do not match. As it stands this approach suffers from suffers from two problems, however. Firstly, the alignment only signals potential pronunciation variants and does not provide a measure of the quality of model match and, secondly, obtaining an accurate alternative decoding sequence is difficult. One method for obtaining such a decoding sequence is to transcribe the acoustic data with subword class labels by hand [4]. This method is prohibitively labour intensive for large corpora, such as the BN corpus. Another method is to run the recogniser over the data using only phone level decoding constraints, e.g. [7]. Decoding sequences obtained using this method contain many errors, however (typically around 30% error rate for phone classification). Also such decoding sequences should not be attributed as much credence in regions of ambiguous acoustics as in those containing clear examples of distinct acoustic classes.

This work was supported by an EPSRC studentship, the Royal Commission for the Exhibition of 1851, ESPRIT Long Term Research Project 23495 (THISL) and was carried out at the International Computer Science Institute.

If an acoustic confidence measure based approach is adopted, the quality of the model match between each subword class model and the portion of acoustics against which it is aligned can be measured directly. In this case, a poor pronunciation model is signalled by constituent subword models with low confidence estimates.

Once a pronunciation model which *consistently* provides a poor match to the acoustic realisations of some word has been identified, that model should be replaced with one which provides a better match to the acoustics (on average). Such a process requires the proposal of *alternative* pronunciation models. A number of methods for automatically generating an alternative pronunciation model for a word exist [3, 4, 7, 11]. An acoustic confidence measure may be employed to determine whether an alternative model is an improvement upon an existing model.

3. CONFIDENCE MEASURES

An acoustic class model created using a 'traditional' HMM estimates the the likelihood of the acoustic observation sequence $\mathbf{X}$ given the class q_k, $p(\mathbf{X}|q_k)$. Such likelihoods are relative to the probability of the acoustic observations, $p(\mathbf{X})$, and so are not comparable across utterances. Hence likelihoods cannot be used in isolation as confidence measures. One approach to deriving a confidence measure from a likelihood based system is to form a *ratio* between the likelihood of the acoustics given the class model and an estimate of $p(\mathbf{X})$ given by a '*garbage*' or '*filler*' model. The use of such likelihood ratios has been reported in the keyword spotting and utterance verification literature [9, 10]. A problem with the use of such likelihood ratios is that it is very difficult to explicitly estimate $p(X)$ for a wide range of acoustic conditions.

A second approach to deriving a confidence measure is to compare the likelihood of a particular acoustic observation sequence against the frequency distribution of likelihood values for that class, calculated over some data set. A low confidence is given to the match of an acoustic class model if its associated likelihood falls sufficiently far from the mean of the distribution [6]. An objection to this approach is that it is somewhat ad hoc as it does not explicitly accommodate different acoustic conditions.

In contrast to likelihood based recognisers, hybrid HMM/ANN systems are well suited to producing acoustic confidence measures. This is because the acoustic model (ANN) can directly estimate acoustic subword class posterior probabilities which are comparable across utterances [2]. The ABBOT [8] acoustic model is trained to estimate phone class posterior probabilities, based on a local portion of acoustics of the utterance, $p(q_k|\mathbf{x}^n)$.

In this paper we make use of an acoustic confidence measure based on local estimates of posterior phone probabilities, CM_{npost}. In previous studies, we have compared the performance of CM_{npost} to that of a number of other confidence measures for the task of decoding hypothesis verification [12, 13]. We have found CM_{npost} to perform better than the other confidence measures for the task of phone hypothesis verification and to be the least expensive to compute. A description of CM_{npost} and four other confidence measures is given below for a phone q_k with an hypothesised start time n_s and end time n_e. The n prefix in $CM_{npost}(q_k)$, $CM_{nsl}(q_k)$ and $CM_{nolg}(q_k)$ indicates that they are durationally normalised. This counteracts the underestimate of the acoustic probabilities caused by the observation independence assumption. Figure 1 provides a comparison of the performance of the confidence measures for the verification of phone hypotheses obtained for an episode of the BN corpus[1] using phone level decoding constraints.

Posterior Probability $CM_{npost}(q_k)$ is computed by rescoring the Viterbi state sequence obtained for a phone q_k with the local posterior probabilities output by the acoustic model.

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

Scaled Likelihood The 'scaled likelihood' of a phone hypothesis q_k is obtained by dividing the local posterior probability estimate of q_k by its acoustic data prior. The ABBOT system uses scaled likelihoods in the search for the optimal state sequence.

$$\begin{aligned} CM_{nsl}(q_k) &= \frac{1}{n_e - n_s} \sum_{n=n_s}^{n_e} \log\left(\frac{p(\mathbf{x}^n|q_k)}{p(\mathbf{x}^n)}\right) \\ &= \frac{1}{n_e - n_s} \sum_{n=n_s}^{n_e} \log\left(\frac{P(q_k|\mathbf{x}^n)}{P(q_k)}\right) . \end{aligned} \tag{2}$$

Online Garbage The term 'online garbage' [1] refers to the normalisation of the probability of the best decoding hypothesis by the average probability of the n-best decoding hypotheses. This average may be considered to be a form of garbage model probability. $CM_{olg-n}(q_k)$ is $CM_{sl}(q_k)$ normalised by the average of the n-best scaled likelihoods.

$$\begin{aligned} CM_{nolg}(q_k) &= \frac{1}{n_e - n_s} CM_{sl}(q_k) - \\ & \sum_{n=n_s}^{n_e} \log \frac{1}{\mathrm{n-best}} \sum_{l=\mathrm{best}}^{\mathrm{nth-best}} \frac{p(q_l^n|\mathbf{x}^n)}{p(q_l)} . \end{aligned} \tag{3}$$

Per Frame Entropy $CM_{ent}(n_s, n_e)$ is the per frame entropy of the distribution of the K local phone class posterior probabilities, averaged over the interval n_s to n_e.

$$CM_{ent}(n_s, n_e) = -\frac{1}{n_e - n_s} \sum_{n=n_s}^{n_e} \sum_{k=1}^{K} p(q_k^n|\mathbf{x}^n) \log p(q_k^n|\mathbf{x}^n) . \tag{4}$$

Lattice Density $CM_{lat}(n_s, n_e)$ is a measure of the density of competing decodings in an n-best lattice of decoding hypotheses and is computed by averaging the number of unique competing decoding hypothesis (NCH) which pass through a frame over the interval n_s to n_e [5]. $CM_{lat}(n_s, n_e)$ is not an acoustic confidence measure as the n-best lattices of decoding hypotheses from which it is derived are created using both acoustic and language model information.

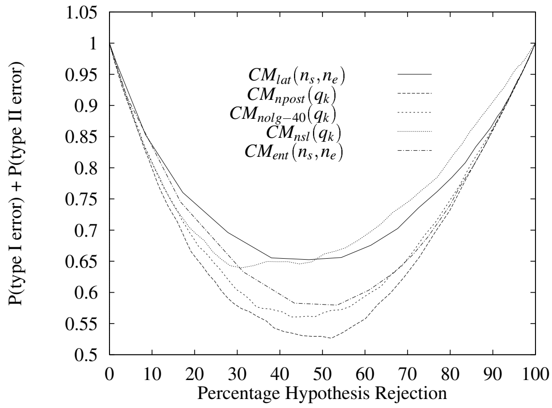

Figure 1: Hypothesis verification performance of five phone level confidence measures for an episode of the BN corpus decoded using phone level constraints.

$$CM_{lat}(n_s, n_e) = \frac{1}{n_e - n_s} \sum_{n=n_s}^{n_e} NCH_n . \tag{5}$$

CM_{npost} can be extended to the word level, $CM_{npost}(w_j)$, by summing the values of $CM_{npost}(q_k)$ over the K constituent phones of a word model and dividing by this number:

$$CM_{npost}(w_j) = \frac{1}{K} \sum_{k=1}^{K} CM_{npost}(q_k) . \tag{6}$$

An example of poor model match signalled by low confidence is illustrated in figure 2. The solid lines in the figure trace how a subset of the outputs of the acoustic model evolve over the duration of an instance of the word 'FUNDS'. The overlayed dashed lines indicate the timings of a forced Viterbi alignment of the pronunciation model /f ah n dcl d z/ to the same portion of acoustics. $CM_{npost}(q_k)$ values for each constituent phone of the pronunciation model are shown next to the timings. From the figure it can be seen that the outputs of the acoustic model do not provide evidence for the occurrence of the phones /dcl/ and /d/ between the 177th and 180th frames of the utterance and that the values of $CM_{npost}(q_k)$ are correspondingly low for the alignment of these two phone class models.

Figure 3 illustrates an improved model match. In this case the quality of model match for the pronunciation model /f ah n z/ to the same acoustic realisation of the word 'FUNDS' is shown. It can be seen from the figure that the improved model fit is accompanied by a commensurate increase in confidence. The model /f ah n z/ was found to match with higher confidence to the majority of examples of the word 'FUNDS' in the 1994 development test set of the ARPA Hub-1 NAB corpus and an episode[2] of the BN corpus.

4. BASEFORM LEARNING

The ABBOT pronunciation lexicon contains mappings between a given word and a sequence of subword acoustic classes. Such

[1] NPR Nightline: 23/05/96

[2] ABC Nightline: 23/05/96

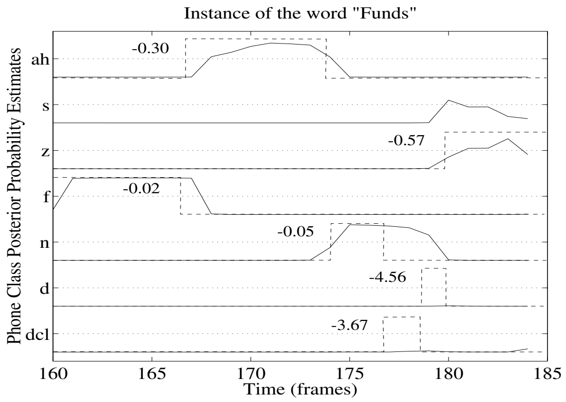

Figure 2: A subset of acoustic model outputs for an instance of the word 'FUNDS', overlayed with timings from a forced Viterbi alignment of the pronunciation model /f ah n dcl d z/ and values of $CM_{npost}(q_k)$ for the aligned model ($CM_{npost}(w_j)$ for the alignment $= -1.53$).

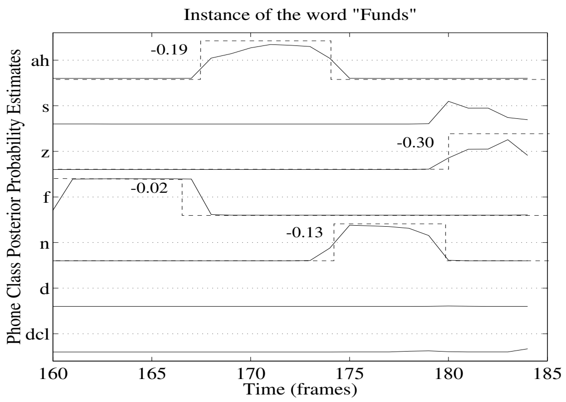

Figure 3: A subset of acoustic model outputs for the same acoustic realisation of the word 'FUNDS', overlayed with timings from a forced Viterbi alignment of the pronunciation model /f ah n z/ and values of $CM_{npost}(q_k)$ for the aligned model ($CM_{npost}(w_j)$ for the alignment $= -0.16$).

mappings are termed *baseforms*. Pronunciation variations are accommodated through the listing of more than one baseform for a given word. In order to investigate the potential of the CM_{npost} confidence measure for evaluating pronunciation models, we incorporated it into a 3 step baseform learning process:

step 1 evaluate entries in the existing pronunciation lexicon based on confidence.

step 2 similarly evaluate alternative models for each word considered in step 1.

step 3 generate a new set of lexical entries based upon a comparison of the evaluations made in steps 1 and 2.

4.1. Evaluation of Existing Baseforms

A good pronunciation model should provide a *consistently* high confidence match to instances of the word it models. Accordingly, our approach to evaluating existing pronunciation models was the following:

1. A forced Viterbi alignment of the reference transcription was made to the training set acoustics. This eliminated any potential complications introduced by decoding errors. Portions of acoustics corresponding to instances of a given word could be located from this forced alignment.
2. A list of words eligible for baseform evaluation was compiled. To be eligible for evaluation, a word must be seen in the training set on a sufficient number of occasions so as to facilitate a reliable evaluation. We arbitrarily set the minimum number of occurrences for eligibility to 10.
3. A given baseform for a word was aligned against *all* instances of the word in the training set and values of $CM_{npost}(w_j)$ were calculated for each alignment.
4. An overall confidence estimate for a given baseform was found by averaging the values of $CM_{npost}(w_j)$ for each of its alignments.

4.2. Evaluation of Alternative Baseforms

For this investigation, we elected to use a phone level constraint decoding as the source for alternative pronunciation models. This approach facilitated a fully automated baseform learning processes using a minimum of linguistic intuitions.

A phone level constraint decoding of the training set was time aligned against the forced Viterbi alignment obtained during the evaluation of the existing baseforms. From this time alignment, sequences of hypotheses generated by the phone level decoding could be mapped to words in the training set. This mapping was carried out using a dynamic programming alignment which makes use of a phonetic feature based distance metric [4]. A list of unique mappings, which can be used as *alternative baseforms*, was then compiled for each eligible word and ordered by the frequency of occurrence of the alternative baseform.

In order to combat the errorful nature of a phone level constraint decoding, only phone sequences which occurred sufficiently frequently were considered as valid alternative baseforms. Indeed, for reasons of computational expense, the number of alternative baseforms evaluated was limited to the 5 most frequent. The set of alternative baseforms for a given word were evaluated in the same fashion as the existing baseforms. This confidence measure based evaluation also serves to eliminate any spurious alternative models.

4.3. Generation of New Lexical Entries

We investigated a number of decision schemes for accepting or rejecting of a baseform.

augment The control scheme was to augment the pronunciation lexicon with the set of most frequently occurring alternative baseforms.

CM-replace1 A list of the existing and alternative baseforms was ordered according to their associated values of CM_{npost} and the baseforms for a given word in the existing pronunciation lexicon were *replaced* with the n-best baseforms drawn from the evaluation list, where n was set equal to the number of baseforms possessed by the word in the existing lexicon. The rationale behind this scheme was to limit the potential confusability introduced into the pronunciation lexicon by the association of a large number of competing baseforms with each word.

CM-replace2 A similarly ordered list of existing and alternative baseforms to that compiled in CM-replace1 was drawn up with the exception that short, frequently occurring function words were omitted from the list. The replacement was carried out in an identical fashion. The rationale behind this scheme is that such function words are subject to increased coarticulation and vowel reduction and so it may be supposed that it is harder to learn baseforms for such words.

CM-augment The existing pronunciation lexicon was *augmented* with all alternative baseforms obtaining a value of CM_{npost} exceeding the lowest value obtained for the set of existing baseforms for the word in question.

The priors for the baseforms added to the pronunciation lexicon were obtained by scaling their associated confidence values or frequencies of occurrence to the range [0,1].

5. RESULTS

An objective evaluation of a modified pronunciation lexicon was made by comparing the WER obtained using the original (baseline) pronunciation lexicon to that obtained using the modified version for some data set. The training set used was composed of 15 episodes of NPR's "Marketplace", constituting an approx. 7 hour subset of the 50 hour BN corpus. Using this training set, 868 words were eligible for baseform evaluation, the most frequently occurring of which was, 'THE', with 3310 occurrences. Words such as 'VALVE', 'TECH' and 'BEGINNING' occurred on 10 occasions. For baseform learning scheme CM-replace2, the 289 most frequent words were omitted from the eligibility list ($1/3$rd).

Pronunciation Lexicon	*WER* (%) Episode from Training Set	Test Set
baseline	18.5	23.5
augment	18.2	23.6
CM-replace1	19.0	24.8
CM-replace2	18.4	24.0
CM-augment	17.6	23.1

Table 1: WER values for the baseform learning schemes calculated for the test set and an episode from the training set.

6. DISCUSSION AND FUTURE WORK

From the table 1, it can be seen that CM-augment scheme performs the best. This scheme provided an approx. 1% absolute (4.9% relative) reduction in WER for an episode drawn from the training set[3], but only an approx 0.5% absolute (1.7% relative) reduction on the test set. It may be speculated that this lack of generalisation is the product of two factors. Firstly, the training and test sets may have been mismatched. The training set was made up of episodes of NPR's "Marketplace", whereas the test set was made up of an episode of ABC's "Nightline"[4] and an episode of NPR's "All Things Considered"[5]. In future experiments the two data sets will contain a better balance of episodes from the 11 shows of the BN corpus. Secondly, the training set is relatively small and work is currently in progress to scale up the evaluations to a training set of 50 hours of BN data. Approx. 5K words will be eligible for evaluation using this larger training set.

An increase in the number of words eligible for evaluation raises an interesting issue. Whilst an augmentation scheme was found to be superior to replacement schemes for a relatively small set of eligible words this may not be the case for a larger set, due to the potential for introducing a detrimental amount of confusability into the pronunciation lexicon. If this is found to be so, replacement based schemes may well provide better performance. It will therefore be interesting to track the relative performance of the different schemes as the they are applied to larger and larger sets of eligible words.

Given that there is potential for introducing confusability into the lexicon, 5K words is still only a small fraction of the 65K that currently reside in the ABBOT BN lexicon. Another issue which may arise is how to propagate the learnt pronunciation models throughout the lexicon. One approach may be to incorporate a baseform learnt for a word into that for a compound word which includes it.

The BN corpus, contains a diverse range of acoustic conditions, such as degraded speech signals, speech mixed with music and non-speech sounds, as well as clean speech. For such corpora, it is important, for a reliable evaluation, to only assess pronunciation models using unambiguous acoustics. $CM_{ent}(n_s, n_e)$ is designed to provide a *general* measure of acoustic model match. As such, it may be used to spot occurrences of ambiguous acoustics or acoustics containing non-speech sounds.

A more complete description of the match of a baseform to a set of realisations of a given word would is the distribution of confidence measure values over that set. Figure 4 is a histogram of the values of $CM_{npost}(w_j)$ for the baseform /hh ah n dcl d r/ over a set of realisations of the word 'HUNDRED'. Typically such a distribution will be bimodal. The fraction of good model matches gives rise to a peak at high confidence and the poor model matches lead to a peak at low confidence, where the relative magnitude of the peaks is dependent upon whether the model provides, on average, a good or poor match to the acoustics. Two schemes are available for accepting/rejecting baseforms using such distributions. Firstly, mixture distributions could be used to model both the performance of a particular baseform and also the average performance of all baseforms for a word. A comparison between the components modelling the high confidence peak could then be used in the acceptance/rejection decision. A simpler approach could be to use

[3]NPR Marketplace: 23/05/96

[4]ABC Nightline: 24/06/96

[5]NPR All Things Considered: 20/05/96(d)

two thresholds to describe the minimum confidence value over a number of occasions required for model acceptance.

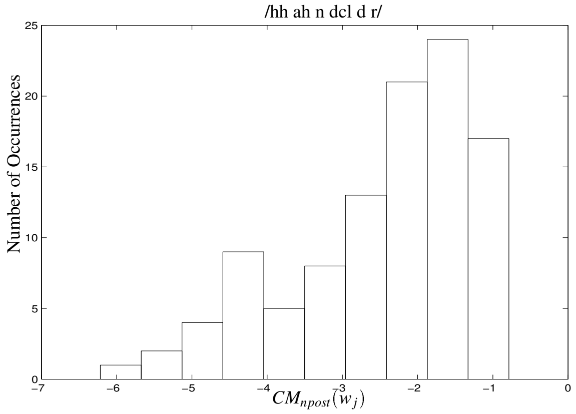

Figure 4: A histogram of $CM_{npost}(w_j)$ values, calculated using the baseform /hh ah n dcl d r/ and the realisations of the word 'HUNDRED' in the training set.

A number of sources for alternative pronunciation models besides phone level constraint decodings exist. For example, a rule based transformational mapping may be applied to existing baseforms to modify phone sequences [11]. It would be desirable to investigate different sources of alternative models and their effect on the baseform learning process.

7. ACKNOWLEDGEMENTS

We would like to thank Eric Fosler-Lussier for enlightening conversations during this work.

8. REFERENCES

[1] J-M. Boite, H. Bourlard, B. D'hoore and M. Haesen. A New Approach Towards Keyword Spotting. In *Proceedings of EuroSpeech*, pages 1273-1276, 1993.

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

[3] F.R. Chen. Identification of Contextual Factors for Pronunciation Networks. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 753-756, 1990.

[4] E. Fosler, M. Weintraub, S. Wegmann, Y-H. Kao, S. Khudanpur, C. Galles and M. Saraclar. Automatic Learning of Word Pronunciation from Data. In *Proceedings of JHU/CLSP Workshop*, Pronunciation Group, 1996.

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

[6] K.L. Markey and W. Ward. Lexical Tuning Based on Triphone Confidence Estimation. In *Proceedings of EuroSpeech*, pages 2479-2482, 1997.

[7] M. Ravishankar and M. Eskenazi. Automatic Generation of Context-Dependent Pronunciations. In *Proceedings of EuroSpeech*, pages 2471-2474, 1997.

[8] 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.

[9] 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*, pages 303-329, Kluwer, 1996.

[10] 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 the International Conference on Acoustics, Speech and Signal Processing*, pages 518-521, 1996.

[11] G. Tajchman, E. Fosler and D.Jurafsky. Building Multiple Pronunciation Models for Novel Words using Exploratory Computational Phonology. In *Proceedings of EuroSpeech*, pages 2247-2250, 1995.

[12] G. Williams and S. Renals. Confidence Measures for Hybrid HMM/ANN Speech Recognition. In *Proceedings of EuroSpeech*, pages 1955-1958, 1997.

[13] G. Williams. A Study the Use and Evaluation of Confidence Measures in Automatic Speech Recognition. Technical Report, *CS-98-02*, Dept. Comp. Sci., Sheffield University.