UNCONSTRAINED VERSUS CONSTRAINED ACOUSTIC NORMALISATION IN CONFIDENCE SCORING

Jacques Duchateau and Patrick Wambacq

Katholieke Universiteit Leuven, ESAT - PSI
Kasteelpark Arenberg 10, 3001 Heverlee (Leuven), Belgium
e-mail: Jacques.Duchateau@esat.kuleuven.ac.be

ABSTRACT

In HMM-based recognition systems for large vocabulary, the observation likelihoods provided by the acoustic models are useful in confidence measures if they are properly normalised. This paper compares two normalisation methods for the acoustic model likelihoods: unconstrained normalisation, based on the unconditional observation likelihood, and constrained normalisation, based on the observation likelihoods in a phoneme recognition system in which the phoneme strings are constrained by an N-gram phoneme sequence model. We found on the benchmark 20k word Wall Street Journal recognition task that both normalisations perform equally well at first sight. However, their behaviour depends on the length of the word: constrained normalisation outperforms unconstrained normalisation for long words and the opposite holds for short words. With a confidence measure that exploits this fact, the normalised cross entropy metric for confidence measures can be increased from a reference 21.9% (with unconstrained normalisation) to 23.5%.

1. INTRODUCTION

The aim of a confidence measure in automatic recognition systems is to predict if the recognised words are correct or not: the confidence measure estimates the a posteriori probability that a word in the recognition result is correct.

These estimates are based on different sources of information about the recognition task which all give some indication on recognised words being correct or wrong.

The most commonly used information sources are the models on which the recognition engine itself is based: the acoustic models and the language model.

In state-of-the-art, elaborate confidence measures however, many other information sources are exploited as well. Examples are the phoneme duration [1, 2], properties of the search during the recognition and of the resulting word graph [3, 4], the distance between phoneme strings obtained from word recognition and from phoneme recognition [5], the speaking rate [6], the prosody pattern of the sentence [1], sentence parsing [7] and the dialogue manager [7].

In this paper one of the basic information sources is investigated: the observation likelihoods from the acoustic models. Because the raw likelihoods are usually not useful as confidence measures, they are normalised in current systems [8, 5, 9, 10, 11].

This work compares two ways of acoustic normalisation: unconstrained normalisation, based on – an approximation of – the unconditional likelihood of the observation (as used in most of the above papers) and constrained normalisation, based on the likelihoods in a phoneme recognition system. The usefulness and possible complementarity of both normalisations was investigated on the benchmark 20k word Wall Street Journal recognition task.

The paper is structured as follows: in section 2, the different confidence measures we use are described, especially the measure based on the observation likelihood and its normalisation. Next our large vocabulary continuous speech recogniser is reviewed in section 3. In section 4 the experiments on the Wall Street Journal (WSJ) recognition task are given and discussed, and finally section 5 draws some conclusions from the presented research.

2. CONFIDENCE MEASURES

Confidence measures are typically based on several information sources. Each source results in what we call a single confidence measure and these single measures are then put together into one combined confidence measure.

In this section we first describe the single measures we used, and the way in which we construct a combined confidence measure based on single measures. Then we look in detail to the single measure that is based on the observation likelihood, more precisely the two investigated ways of normalising the likelihoods are explained.

In this paper, the confidence measures we investigated are based on only three information sources:
- the acoustic model likelihood, normalised as explained below,
- the beam width (this is the number of tokens in the search beam) after pruning, and
- language model probabilities for the word in the recognised sentence.

The (normalised) acoustic model likelihood and the beam width are evaluated for each frame. These values at the frame level are combined into one score at the word level in two steps: first a phone level score is calculated as the average of the logarithm of the values for all frames aligned to that phone, and then the word level score is found as the average phone level score.

The language models immediately result in scores at the word level: the logarithm of the language model probability. Two language models are used: the common forward N-gram language model in which the probability of a word is dependent on the preceding words, and the backward language model in which a word is predicted based on the following words only. We introduced this backward language model in [12], and showed that it provides information for the confidence measure that is complementary to the information in the forward language model.

To construct a combined confidence measure based on the above single measures, the logit model is used [13]. In the logit model, a linear combination of the single measures (of which the weights have to be estimated on a development test set) is turned into a score by a sigmoid function.

The normalisation of the observation likelihoods is done by dividing the likelihood for a frame by an other likelihood. In unconstrained normalisation, this other likelihood is the unconditional likelihood for the frame, which can be estimated very efficiently for the acoustic models with tied gaussians that we use (as explained on page 44 of [14]).

The other likelihood in constrained normalisation is the observation likelihood in a phoneme recogniser which uses the same acoustic models as the word recogniser and an N-gram phoneme sequence model. This normalisation is called constrained because the observation likelihood for a frame depends on the state sequence and this state sequence is restricted by the (left-to-right) phoneme models and by the phoneme sequence model.

In figure 1 the correlation plot is shown between the word scores with unconstrained normalisation and with constrained normalisation (using a 2-gram phoneme sequence model). It can be seen that although there is a strong correlation, for example for two words with the same score with unconstrained normalisation, the difference in word score with constrained normalisation can be quite large. Note that the clear line for zero word scores when constrained normalisation is used corresponds to words for which both word

![Fig. 1. Correlation between word scores with unconstrained and constrained normalisation](image)

and phoneme recogniser produce the same phoneme string and alignment.

## 3. RECOGNITION SYSTEM

We compared the acoustic normalisations on the well-known speaker independent WSJ recognition task with a 20k word open vocabulary. Results are given on the November 92 evaluation test set with non verbalised punctuation.

For the experiments described in this paper, the speaker independent large vocabulary continuous speech recognition system developed at the ESAT-PSI speech group at the K.U.Leuven is used. An overview of the acoustic modelling can be found in [15, 16], the search module is described in [17, 14].

The signal processing calculates 12 Mel scaled cepstral coefficients and the log energy, all of them mean normalised and augmented with first and second order time derivatives. The resulting 39 features are decorrelated using the algorithm described in [14].

The acoustic modelling, estimated on the SI-284 (WSJ1) training data with 69 hours of speech, is gender independent and based on a phone set with 45 phones, without specific function word modelling. No cross word phonetic rules are used to adapt phonetic transcriptions depending on the neighbouring words.

A global phonetic decision tree defines the 6559 tied states in the cross word context dependent and position dependent models. Each tied state is modelled as a mixture of on average 116.6 gaussians which are tied over the different states, the total number of tied gaussians being 62554.

The standard trigram language modelling provided by Lincoln Laboratory for the 20k word open (1.9% OOV rate)
vocabulary is used in the recognition system.

With the above recognition system for the WSJ task, a word error rate (WER) of 8.0% was found on the November 92 evaluation test set. Due to the efficient evaluation of gaussians with the FROG system [15, 14] and the efficient single pass time synchronous beam search algorithm, this 8.0% WER was found with real time recognition on a single 1.7 GHz Pentium 4 processor running Linux.

To evaluate the confidence measure, forward and backward (Katz-smoothed [18]) trigram language models were estimated on the standard 38.9 million words of WSJ texts (thus the forward language model is approximately the same as the Lincoln standard used in the word recogniser).

The phoneme recogniser on which one likelihood normalisation is based, uses the same recognition engine and acoustic models as the word recogniser. We evaluated two phoneme recognisers with different (again Katz-smoothed) N-gram phoneme sequence models: a 2-gram and a 5-gram. They were both estimated on a 8.5M phoneme database that was developed through a forced alignment on the SI-284 database, allowing for multiple pronunciations per word as given by the pronunciation dictionary.

4. EXPERIMENTS

In this section, the different ways of acoustic normalisation in a confidence measure are compared. As described in section 2, the combination of single confidence measures is based on the logit model. The weights for the linear combination in this model are estimated in all cases on the WSJ November 92 development test set, the results are given on the November 92 evaluation test set.

The different confidence measures are evaluated based on the normalised cross entropy – or the normalised mutual information – between the correctness of the recognised words and the confidence scores for them (normalising by the maximum cross entropy). This metric was chosen as a NIST standard to assess confidence measures in an application independent way. It is reviewed in detail for instance in [13].

The normalised cross entropy metric for a confidence measure based on beam width and both language models is 18.3%. When the observation likelihood is added as information source, this number improves to 21.9% both for unconstrained normalisation and for constrained normalisation using a 2-gram in the phoneme recogniser. Using a 5-gram in the phoneme recogniser gives a worse result: 20.7%. Forcing phoneme strings that are typical for the language does not seem to be a good idea, it for instance results in more zero word scores (as the phoneme string from the phoneme recogniser is more often the same as the phoneme string for the word in the pronunciation dictionary).

However the behaviour of the confidence measure depends on the length of the word, as can be seen in table 1. The NIST metric is given for the reference confidence measure (based on beam width and language models), and for the confidence measures when a normalised observation likelihood is added to the information sources.

The results in the first column are obtained when the confidence measure is both made and evaluated specifically for short words (3 phonemes or less, this way about half of the words are short). The second column gives the corresponding results for long words (more than 3 phonemes). In the last column the NIST metric is given for the total test set when the confidence for short and long words is estimated with their respective word length specific confidence measures.

In general it is clear from the table that the confidence of long words can be predicted more easily, and that again the normalisation based on a phoneme recogniser with a 5-gram is always outperformed by the other normalisations.

More importantly the table also shows that constrained normalisation outperforms unconstrained normalisation for long words and that the opposite holds for short words. A possible explanation is the following. For a long word, it should be checked whether there is a phoneme string that is typical for the language and that fits better to the acoustic data, because this indicates that an out-of-vocabulary word occurred. This is exactly what the constrained normalisation does. On the contrary for short words it should for instance be checked if the addition of the word to the recognition result (possibly pushed by the language model) does not give rise to a difficult alignment (due to the use of left-to-right acoustic models), and the unconstrained normalisation may be better suited for this task.

This different behaviour of the normalisation methods for short and long words can be exploited in a combined confidence measure: for short words the prediction of the measure with unconstrained acoustic normalisation is used, and the confidence of long words is predicted with the measure with constrained acoustic normalisation. With this combined confidence measure, the overall normalised cross entropy metric can be increased from 21.9% (for the measure that does not incorporate the use of the word length) to 23.5%, this is a 7% relative improvement.

In figure 2, the receiver operating characteristic (ROC)
curve for this combined confidence measure is given, along with the curves for both short and long words separately. With ROC curves the performance of the confidence measure as binary tagger of correct and wrong words can be seen at the different operating points (see [13] for more details).

![ROC curves for several confidence measures](image)

Fig. 2. ROC curves for several confidence measures

It can for instance be seen that for a false rejection rate of 10% of the correctly recognised words, on average 60% of the recognition errors can be detected, but for short and long words, this number is 53% and 68% respectively.

5. CONCLUSIONS

In this paper we compared different normalisation methods for observation likelihoods when they are to be used in confidence measures. Experiments on the WSJ recognition task showed that constrained normalisation outperforms unconstrained normalisation for long words and the opposite holds for short words. With a combined measure that exploits this fact, the normalised cross entropy could be improved from 21.9% to 23.5%.

6. REFERENCES


