CONFIDENCE SCORING BASED ON BACKWARD LANGUAGE MODELS

Jacques Duchateau, Kris Demuynck and Patrick Wambacq

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

ABSTRACT

In this paper we introduce the backward N-gram language model (LM) scores as a confidence measure in large vocabulary continuous speech recognition.

Contrary to a forward N-gram LM, in which the probability of a word is dependent on the preceding words, a word in a backward N-gram LM is predicted based on the following words only. So the backward LM is a model for sentences read from the end to the beginning.

We show on the benchmark 20k word Wall Street Journal recognition task that the backward LM scores contain information for the confidence measure that is complementary to the information in forward LM scores. The normalised cross entropy metric for confidence measures increases significantly from 18.5% to 23.3% when backward LM scores are added to a confidence measure which includes the forward LM scores.

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.

Some of these information sources are used in the recognition engine itself to produce the recognition result, as for instance the acoustic models and the language model (LM).

But in recent confidence measures, other information sources are exploited as well. Examples are the phoneme duration [1, 2], the speaking rate [3], the prosody pattern of the sentence [1], sentence parsing [4] or the dialogue manager [4].

It is often difficult or impossible to use these information sources directly in the recognition engine itself. In a sense, for some information sources one has even the choice where to use them. Applying them directly into the recognition engine will make the engine typically slower but better. But as shown in [2] it is also possible to rescore recognition results based on a confidence measure in order to improve the recognition rate of the recogniser.

In this paper, we introduce an extra information source for confidence scoring: the backward LM, in which the probability of each word is estimated depending on the following words. We investigated the usefulness and complementarity of this information source by adding it to a confidence measure which includes the forward LM as an information source.

If the results are promising enough, we can still decide to follow the more laborious way and apply the backward LM directly in the recognition engine to investigate if this can also reduce the error rate of the recogniser.

The paper is structured as follows: in section 2, the different confidence measures we use are described, especially the measure based on the backward LM scores. 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 proposed 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 reference single measures we used, and the way in which we construct a combined confidence measure based on single measures.

Then we introduce the new single measure based on the backward LM which will be added to the reference measures in a combined confidence measure in order to investigate if the backward LM contains complementary information.

Research funded by IWT in the STWW programme, project ATraNoS (Automatic Transcription and Normalisation of Speech), home page: http://atranos.esat.kuleuven.ac.be

2.1. Reference Measures

In this paper, the reference confidence measures are based on only three basic information sources:

- the acoustic model likelihood, normalised by the estimated unconditional acoustic probability (see [5], page 44),
- the beam width (this is the number of tokens in the search beam) after pruning, and
- the – forward – LM probability for the word in the recognised sentence.

The language model immediately results in a score at the word level: the logarithm of the LM probability. The acoustic model 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.

As for the combination of different single measures, the *logit* model is used [6]. 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 probability by a sigmoid function.

2.2. Measure based on the Backward LM

Based on the backward N-gram LM, a single confidence measure at the word level can be calculated in the same way as for the forward N-gram LM. The backward N-gram LM can be estimated on the same training texts as the forward N-gram LM, but with the words in each sentence in reverse order (from the last to the first word).

We compared a forward trigram LM and a backward trigram LM on the WSJ large vocabulary recognition task with a 20k word open vocabulary. Both Katz-smoothed [7] trigrams were estimated on the standard 38.9 million words of WSJ texts (years 1987 until 1989).

The perplexities of both models on the November 92 evaluation test set are approximately the same, the perplexities for the backward LM being slightly better: 137.9 versus 137.5 including the 1.9% out-of-vocabulary (OOV) words, or 143.2 versus 141.5 without considering the OOV words. In both language models about 61% of the trigrams is seen in the training texts, for 33% of the cases only a bigram can be found and the remaining 6% fall back on the unigram.

In figure 1 the correlation plot is shown between the forward LM probability for a word given the two previous words, and the backward LM probability for that word given the next two words. It can be seen that the difference between both probabilities can be very large, even several orders of magnitude.

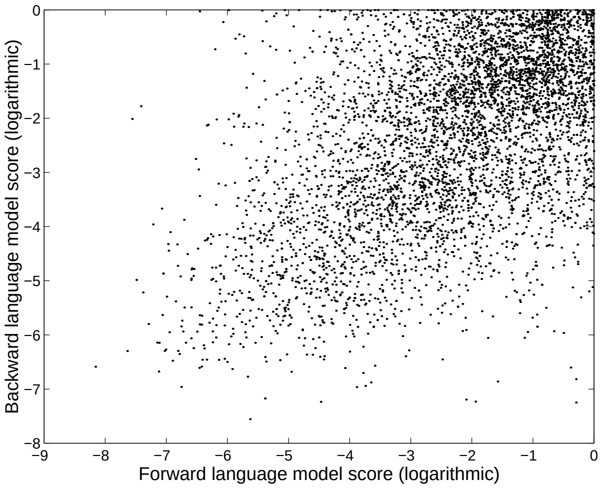

Fig. 1. Correlation between forward and backward LM

For example in a sentence that ends in *“... the company, he said.”*, the word *“he”* has a rather low forward trigram probability but a high backward trigram probability. Especially when one of both trigrams is unseen, the difference between both LM probabilities can be high.

3. RECOGNITION SYSTEM

We evaluated the backward trigram LM 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 modeling can be found in [8, 9], the search module is described in [10, 5].

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 [5].

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 [8, 5] 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.

4. EXPERIMENTS

In this section, the different single confidence measures and several combined confidence measures are evaluated. As described in section 2, the combination of confidence measures is based on the *logit* model. The weights for the linear combination in this model are estimated on the WSJ November 92 development test set, the results are given on the November 92 evaluation test set.

The different measures are evaluated in the following two ways.

- With receiver operating characteristic (ROC) curves: on these curves, the performance of the confidence measure as binary tagger of correct and wrong words can be seen at the different operating points.
- With 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.

Both metrics are reviewed in detail in [6].

The ROC curves for the different confidence measures are given in figure 2.

The single confidence measures based on the acoustic model, on the beam width and on both language models clearly have a different behaviour. The acoustic model score is the worst, the beam width results in the best score for a high false rejection rate and the language models are best for a low false rejection rate. The difference between both language models is small.

Also two combined confidence measures are given, one with the acoustic model, the beam width and the forward LM only, and a second, better measure where the backward LM is also included.

It can be seen that the backward LM adds information to the other measures including the forward LM. For instance for a false rejection rate of 10% of the correctly recognised words, the number of detected recognition errors increases from 45.9% to 55.1%.

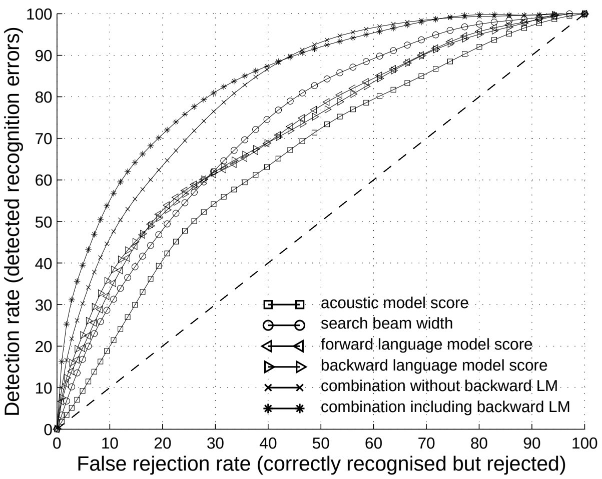

Fig. 2. ROC curves for several confidence measures

The corresponding normalised cross entropy metric for the same single and combined confidence measures is given in table 1. Again an improvement – from 18.5% to 23.3% – is found by adding the backward LM as information source.

Confidence measure	NIST metric
Acoustic model score	2.4%
Search beam width	8.5%
Forward language model score	8.4%
Backward language model score	8.8%
Combination without backward LM	18.5%
Combination including backward LM	23.3%

Table 1. Normalised cross entropy for the different confidence measures

One reason for the improvements found with the addition of a backward LM can be the enlarged word context that is used (2 preceding *and* 2 following words). Therefore we checked if the same effect can be found using a (forward) 5-gram LM.

To do so, we estimated a – word-based – 5-gram on 115.6 million words of WSJ texts (years 1987 until 1994), resulting in a LM with perplexity 113.4 and a normalised cross entropy of 9.8% for the single measure. Combining the acoustic model score, the search beam width and this 5-gram LM score, a normalised cross entropy of 20.0% is found. This is clearly worse than the 23.3% we found with the backward LM.

Moreover adding the backward LM used before (the trigram LM based on only 38.9 million words) to this confidence score increases the normalised cross entropy significantly from 20.0% to 24.2%.

Note that the results with the 5-gram LM are better than the ones with the trigram in table 1 simply because the LM itself is better, as can be seen from the perplexity. Using this 5-gram directly in the single pass recogniser also decreases the WER from 8.0% (for the trigram provided by Lincoln Laboratory) to 7.3%. This improvement is in fact mainly due to the use of more WSJ training texts as a trigram based on the same texts also improves the WER to 7.3%.

5. CONCLUSIONS

In this paper we introduced the backward N-gram LM as an additional information source for confidence measures. Experiments on the WSJ recognition task showed that the backward LM contains information that is complementary to the information in the forward LM. Adding this information to the confidence measure (which is also based on the forward LM) results in an increase of the normalised cross entropy from 18.5% to 23.3%.

Future work on this topic is the implementation of the backward N-gram LM directly into the recognition engine. Then it can be investigated if the use of the backward LM can also reduce the WER of the recogniser.

6. REFERENCES

[1] Denis Jouvet, Katarina Bartkova, and Guy Mercier, "Hypothesis dependent threshold setting for improved out-of-vocabulary data rejection," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, Phoenix, U.S.A., Mar. 1999, vol. II, pp. 709–712.

[2] Dimitra Vergyri, "Use of word level side information to improve speech recognition," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, Istanbul, Turkey, June 2000, vol. III, pp. 1823–1826.

[3] Gustavo Hernández-Ábrego and José B. Mariño, "Contextual confidence measures for continuous speech recognition," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, Istanbul, Turkey, June 2000, vol. III, pp. 1803–1806.

[4] Paul Carpenter, Chun Jin, Daniel Wilson, Rong Zhang, Dan Bohus, and Alex Rudnicky, "Is this conversation on track?," in *Proc. EUROSPEECH*, Aalborg, Denmark, Sept. 2001, vol. III, pp. 2121–2124.

[5] Kris Demuynck, *Extracting, Modelling and Combining Information in Speech Recognition*, Ph.D. thesis, K.U.Leuven, ESAT, February 2001, Available from http://www.esat.kuleuven.ac.be/˜spch.

[6] Manhung Siu and Herbert Gish, "Evaluation of word confidence for speech recognition systems," *Computer Speech and Language*, vol. 13, no. 4, pp. 299–318, Oct. 1999.

[7] Slava M. Katz, "Estimation of probabilities from sparse data for the language model component of a speech recognizer," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 35, pp. 400–401, Mar. 1987.

[8] Jacques Duchateau, Kris Demuynck, and Dirk Van Compernolle, "Fast and accurate acoustic modelling with semi-continuous HMMs," *Speech Communication*, vol. 24, no. 1, pp. 5–17, Apr. 1998.

[9] Jacques Duchateau, *HMM Based Acoustic Modelling in Large Vocabulary Speech Recognition*, Ph.D. thesis, K.U.Leuven, ESAT, November 1998, Available from http://www.esat.kuleuven.ac.be/˜spch.

[10] Kris Demuynck, Jacques Duchateau, Dirk Van Compernolle, and Patrick Wambacq, "An efficient search space representation for large vocabulary continuous speech recognition," *Speech Communication*, vol. 30, no. 1, pp. 37–53, Jan. 2000.