THE USE OF PROSODY IN A COMBINED SYSTEM FOR PUNCTUATION GENERATION AND SPEECH RECOGNITION

Ji-Hwan Kim and P. C. Woodland

Cambridge University Engineering Department
Trumpington Street, Cambridge, CB2 1PZ, United Kingdom
{jhk23, pcw}@eng.cam.ac.uk

ABSTRACT

In this paper, we discuss a combined system for punctuation generation and speech recognition. This system incorporates prosodic information with acoustic and language model information. Experiments are conducted for both the reference transcriptions and speech recogniser outputs. For the reference transcription case, prosodic information is shown to be more useful than language model information. When these information sources are combined, we can obtain an F-measure of up to 0.7830 for punctuation recognition.

A few straightforward modifications of a conventional speech recogniser allow the system to produce punctuation and speech recognition hypotheses simultaneously. The multiple hypotheses are produced by the automatic speech recogniser and are re-scored by prosodic information. When prosodic information is incorporated, the F-measure can be improved by 19% relative. At the same time, small reductions in word error rate are obtained.

1. Introduction

As the sentence is the basic unit for natural language understanding systems such as those in information retrieval, automatic punctuation from speech is a crucial step in making the transition from speech recognition to speech understanding. Also, automatic punctuation can greatly improve the readability of transcriptions.

An automatic punctuation system, which was based on only lexical information, was developed in [1]. Their system only produced commas under the assumption that full stops and question marks were pre-determined.

When automatic punctuation is simultaneously performed with speech recognition, it is important to assign acoustic pronunciations to each punctuation mark. Acoustic baseforms of silence, breath, and other non-speech sounds were assigned to punctuation marks, and an automatic punctuation experiment with speech recognition performed for 3 speakers in [2].

It is known that there is a strong correspondence between discourse structure and prosodic information [3]. A sentence boundary recogniser using lexical information and pause duration was developed in [4]. This development reported that a pause duration model when used alone performs better than a language model, and the result can be improved by combining these two information sources.

Prosodic features can be modelled using a classification tree. Combination methodologies for probabilities from a prosodic model and a language model were discussed in [3, 5]. We adopt their methodologies in this paper.

In Section 2, we will present a methodology for automatic punctuation. Then, experimental setups will be described in Section 3. In Section 4 and Section 5, experimental results will be presented and discussed. Finally, we conclude this paper in Section 6.

2. Punctuation generation

We will describe automatic punctuation experiments for both the reference transcriptions and with speech recognition. When automatic punctuation is performed with the reference texts, the sequences of words are already given. Therefore, experiments aim at generating punctuations between words. As sentence boundary marks (<s> and </s>) provide a lot of information for locating punctuation near to them, it is unrealistic to include this information at the input for punctuation generation. Therefore, the sentence boundary marks are removed from the training and test data.

When automatic punctuation is performed simultaneously with speech recognition, the approximate sentence boundary marks are generated by recogniser segmentation. Sentence boundary marks are therefore not removed in this case, because the recogniser is part of the automatic punctuation generation system.

2.1. Automatic punctuation generation for reference transcriptions

Let Y be the punctuation mark sequence, W be the word sequence and F be the corresponding prosodic feature sequence. The automatic punctuation system aims to find the maximum *a posteriori* Y, Y_{MAP}, given W and F.

$$Y_{MAP} = \arg_Y \ \max P(Y|W,F) \tag{1}$$

Now

$$P(Y|W,F) \quad = \quad \frac{P(F|Y,W)P(Y|W)}{P(F|W)} \tag{2}$$

Since Y is independent of the evidence $P(F|W)$,

$$P(Y|W,F) \quad \propto \quad P(F|Y,W)P(Y|W) \tag{3}$$

Assuming that F depends only on Y, and $P(F)$ is uniformly distributed,

$$P(F|Y,W) = P(F|Y) = \frac{P(Y|F)P(F)}{P(Y)} \propto \frac{P(Y|F)}{P(Y)} \tag{4}$$

Let y_i be the ith punctuation mark and f_i be the ith prosodic feature. Apply the 1st order Markov assumption i.e. $p(y_i|f_1,...,f_T) = p(y_i|f_i)$ and also let y_i be conditionally independent ie $p(y_1,...,y_T|F) = \prod_{i=1}^{T} p(y_i|F)$,

$$P(Y|F) = \prod_{i=1}^{T} p(y_i|f_i) \tag{5}$$

The probabilities in Equation 5 can be obtained, for instance, from the terminal nodes of classification trees, and $P(Y|W)$ in Equation 3 can be obtained from a statistical language model. $P(Y)$ can be obtained from training data counts.

2.2. Combined automatic punctuation and speech recognition

The correlation between punctuation and pauses was investigated in [2]. These experiments showed that pauses closely correspond to punctuations. The correlation between pause lengths and sentence boundary marks was studied for broadcast news data in [4]. In their study, it was observed that the longer the pause duration, the greater the chance of a sentence boundary existing. Although some instances of punctuation do not occur at pauses, it is convenient to assume that the acoustic pronunciation of punctuation is silence.

A prosodic feature model to predict punctuation can be built by a classification tree. Probabilities from the prosodic feature model can then be incorporated by rescoring of multiple hypotheses each of which includes putative punctuation marks. The probability combination process can proceed as shown in Section 2.1.

3. Experiments

Broadcast News (BN) provides a good test-bed for speech recognition, because it requires systems to handle unanticipated speakers, a large vocabulary, and various domains. In this paper, BN texts comprising 211 million words which were broadcast from 1992 to 1997 and a 100-hour 1998 Hub-4 BN data set (acoustic data and its transcription) are used as training data. We also use 3 hours of test data from the NIST 1998 Hub-4 broadcast news benchmark tests. Table 1 summarises the training and test data. Among the many kinds of punctuation mark, this paper is restricted to the examination of full stops, commas, and questions marks, because there are sufficient occurrences of these punctuation marks in the training and test corpora.

Name	Description	#Words
DB92_97	1992_97 BN texts	211M
DB98	100 hrs of Hub-4 data (1998)	767K
TDB98	1998 benchmark test data	35710

Table 1: Database descriptions

4-gram language models are trained by interpolating language models trained on DB92_97 and DB98 using a perplexity minimisation method. The test data, TDB98, is provided as two separate parts. When we perform automatic punctuation for one part of the test data, we use the other part of the test data as the development set to estimate the language model mixture ratios.

The different systems are evaluated using the agreement between the punctuations in the hypothesis file and those in the reference file. Precision and Recall are used as metrics for assessing the performance. These are defined as:

$$P = \frac{\text{number of correct punctuations}}{\text{number of hypothesised punctuations}} \tag{6}$$

and

$$R = \frac{\text{number of correct punctuations}}{\text{number of punctuations in reference}} \tag{7}$$

A half score is given when a punctuation is located correctly, but recognised as a different type of punctuation. The F-measure is the uniformly weighted harmonic mean of Precision and Recall:

$$F = \frac{PR}{(P+R)/2} \tag{8}$$

The F-measure is also used as a metric for assessing performance. As a scorer of speech recognisers, the NIST HUB-4 scoring pipeline is used.

3.1. Classification tree setup

Many easily computable prosodic features are investigated for Dialog Act (DA) classification in [3]. By considering the automatic punctuation task and the contribution of each prosodic feature for DA classification, a set of 10 prosodic features has been investigated for punctuation generation.

The end of each word is a possible candidate for punctuation, and so all prosodic features are measured at the end of a word. The window length is set at 0.2 secs. The left window is the window to the left of the word end, and the right window to the right. Good F0 values are those greater than the minimum F0 (50Hz) and less than the maximum F0 (400Hz). Table 2 explains these features.

Name	Description
Pau_Len	Pause length at the end of a word
Dur_fr_Pau	Duration from the previous pause
Avg_F0_L	Mean of good F0s in left window
Avg_F0_R	Mean of good F0s in right window
Avg_F0_Ratio	Avg_F0_R/Avg_F0_L
Cnt_AgvF0_L	No. of good F0s in left window
Cnt_AgvF0_R	No. of good F0s in right window
Eng_L	RMS energy in left window
Eng_R	RMS energy in right window
Eng_Ratio	Eng_R/Eng_L

Table 2: Description of the prosodic feature set (Window length = 0.2 sec, 50Hz $\leq$ good F0 $\leq$ 400Hz)

Prosodic features for classification tree generation are measured from DB98 because it is the only database in the training set with acoustic data. Nodes in the classification tree are split according to the entropy reduction criteria.

4. Results: Automatic punctuation for reference transcription

We have developed a language model-only system (S_LM) and a prosodic model-only system (S_CART), and have also formed a combination of these two systems (S_LM+CART). Table 3 summarises these systems.

The results of automatic punctuation for the reference transcripts are shown in Table 4. S_LM gives an F-measure of 0.57. Surprisingly, S_CART outperforms S_LM by 0.05. By combining these two models, the result can be improved to give an F-measure of up to 0.78 when a scale factor of 2.0 is applied. The scale factor (α) is the weighting given to the prosodic feature model i.e. $\alpha \times \log P(F|Y,W) + \log P(Y|W)$. From these results, we conclude that lexical information and prosodic information are very complementary in an automatic punctuation task with reference transcriptions. The performance of S_LM+CART varies as the scale factor changes. Figure 1 describes how F-measure, Precision and Recall change with the scale factor. The F-measure attains a maximum at a scale factor of 2.0.

System	Description
S_LM	Language model-only
S_CART	Prosodic feature model-only (by classification tree)
S_LM+CART	Combination of S_LM and S_CART

Table 3: Description of automatic punctuation systems for reference transcripts

System	P	R	F
S_LM	0.60	0.55	0.57
S_CART	0.54	0.74	0.62
S_LM+CART (α=2.0)	0.76	0.80	0.78

Table 4: Automatic punctuation results for reference transcripts (α = scale factor)

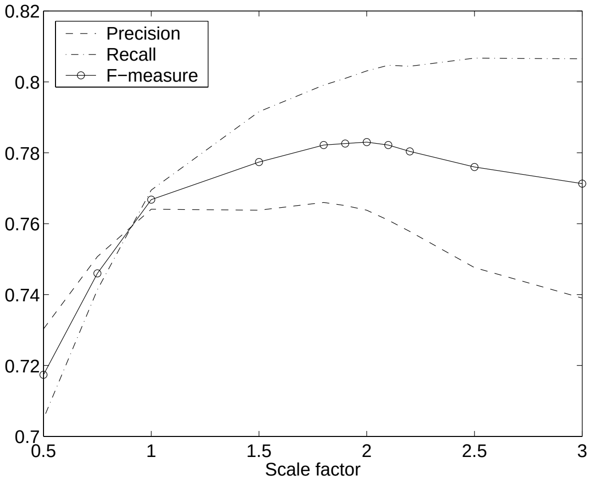

Figure 1: Recognition results of S_LM+CART with different scale factors

5. Results: Combined automatic punctuation for speech recognition

Table 5 shows speech recognition results under 3 different conditions. When punctuation is not included in training and test data, the word error rate (WER) of the speech recogniser (S_woP) is 16.71%. After including punctuation marks, the WER of the speech recogniser (S_Base) is increased to 22.73%. This degradation is caused by two factors: the additional error from other words due to the introduction of punctuation marks into the word list, and the error in mis-recognising the punctuation marks themselves. In S_rmP, punctuation marks are generated by S_Base and these marks are then removed from the reference and the hypothesis. Using the degradation from S_woP to S_rmP, the error from other words due to adding punctuation marks in the word list can be measured at 0.33%; the other factor is therefore measured at 5.69%.

We use S_Base as the baseline automatic punctuation system with speech recognition. Using S_Base, 100 hypothe-

System	WER	Remarks
S_woP	16.71	Punctuation excluded
S_Base	22.73	Punctuation included
S_rmP	17.04	Punctuations removed from reference and S_Base's result

Table 5: Speech recognition results (WER = Word Error Rate (%))

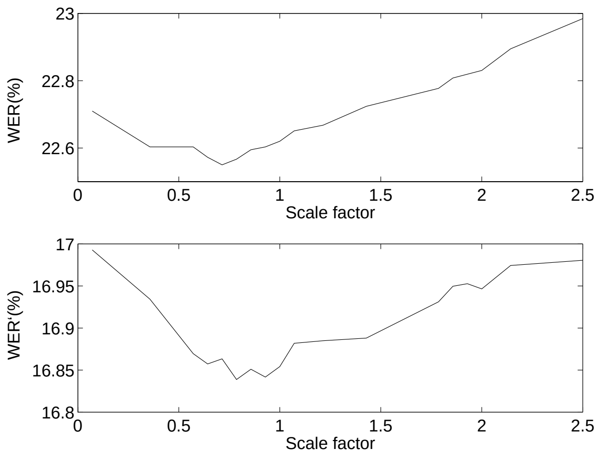

Figure 2: WER (Word Error Rate) and WER' (WER after punctuation is removed from a reference and a hypothesis) of S_H100 with different scale factors

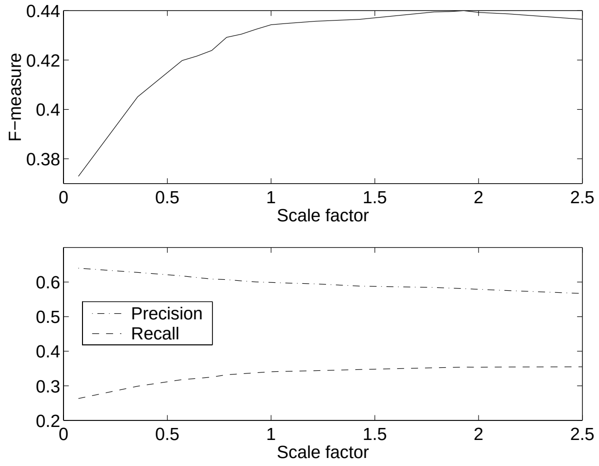

Figure 3: F-measure, Precision and Recall of S_H100 with different scale factors

ses are generated and re-scored on a segment basis using the classification tree prosodic feature model. After re-scoring, the best hypotheses for each segment are combined. Table 6 summarises these systems.

System	Description
S_Base	No re-scoring (baseline. WER = 22.73%)
S_H100	Final hyp. from re-scored 100 hypotheses

Table 6: System descriptions

The performance of S_H100 varies as the scale factor to prosodic model changes. Figure 2 describes how both the WER and the WER after punctuation is removed from reference and hypothesis (WER') change according to scale factor. WER is minimised with a scale factor of 0.71, and WER' is minimised with a scale factor of 0.79.

Figure 3 shows the variation of F-measure, Precision and Recall according to scale factor. The value of F-measure attains its maximum when the scale factor is 1.93. Table 7 summarises these results.

System	WER	WER'	P	R	F
S_Base	22.73	17.04	0.6425	0.2585	0.3687
S_H100 α=0.79	22.57	16.84	0.6072	0.3319	0.4292
S_H100 α=1.93	22.82	16.95	0.5811	0.3541	0.4400

Table 7: Results of automatic punctuation with speech recognition (WER = Word Error Rate (%). WER' = WER after removing punctuations from a reference and a hypothesis, P = Precision, R = Recall, F = F-measure)

6. Conclusions

In this paper, we present an automatic punctuation method which generates punctuations simultaneously with speech recognition output. This system produces multiple hypotheses and uses prosodic features to re-score the hypotheses. Given the reference transcription, using prosodic information alone outperforms using lexical information alone. As these two information sources are shown to be very complementary, further improvements can be achieved by combining these two information sources. When punctuations are generated simultaneously with speech recognition output, the F-measure can be improved up to 0.44 by utilising prosodic information. At the same time, we achieve reductions in word error rate.

7. Acknowledgements

Ji-Hwan Kim is supported by the British Council, LG company and GCHQ.

8. References

1. D. Beeferman, A. Berger, and J. Lafferty. Cyberpunc: A Lightweight Punctuation Annotation System for Speech. In *Proc. ICASSP*, pages 689–692, 1998.
2. C. Chen. Speech Recognition with Automatic Punctuation. In *Proc. Eurospeech*, pages 447–450, 1999.
3. E. Shriberg et al. Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech? *Language and Speech*, 41(3-4):439–487, 1998.
4. Y. Gotoh and S. Renals. Sentence Boundary Detection in Broadcast Speech Transcripts. In *Proc. International Workshop on Automatic Speech Recognition*, pages 228–235, 2000.
5. D. Hakkani-Tur, G. Tur, A. Stolcke, and E. Shriberg. Combining Words and Prosody for Information Extraction from Speech. In *Proc. Eurospeech*, pages 1991–1994, 1999.