A Combined Punctuation Generation and Speech Recognition System and its Performance Enhancement Using Prosody

Ji-Hwan Kim and Philip C. Woodland

Cambridge University Engineering Department,
Trumpington street, Cambridge, CB2 1PZ, United Kingdom

Abstract

A punctuation generation system which combines prosodic information with acoustic and language model information is presented. Experiments have been conducted for both the reference text transcriptions and speech recogniser outputs. For the reference transcription, prosodic information of acoustic data is shown to be more useful than language model information. Several straightforward modifications of a conventional speech recogniser allow the system to produce punctuation and speech recognition hypotheses simultaneously. The multiple hypotheses produced by the automatic speech recogniser are then re-scored using prosodic information. When the prosodic information is incorporated, the F-measure (defined as harmonic mean of recall and precision) can be improved. This speech recognition system including punctuation gives a small reduction in word error rate on the 1-best speech recognition output including punctuation. An alternative approach for generating punctuation from the un-punctuated 1-best speech recognition output is also proposed. The results from these two alternative schemes are compared.

Key words: Punctuation generation, Speech recognition, Prosody, Classification And Regression Tree (CART), N-best rescoring

Email address: {jhk23,pcw}@eng.cam.ac.uk (Ji-Hwan Kim and Philip C. Woodland).

1 Introduction

Even with no speech recognition errors, automatically transcribed speech is much harder to read than human generated transcriptions due to the lack of punctuation, capitalisation and number formatting. The format of standard recogniser output is known as Standard Normalised Orthographical Representation (SNOR) (NIST, 1998) and consists of only single-case letters without punctuation marks or numbers. The readability of speech recognition output would be greatly enhanced by generating proper punctuation. When speech dictation is performed, the dictation system can rely on the speakers to say "full stop" or "comma" whenever they are necessary in the dictated text i.e. "verbalised punctuation". However, when speakers are not aware that their speech is automatically transcribed, e.g. broadcast news and conversational speech over the telephone, verbalised punctuation is not present. When the input text comes from speech, the task of punctuation generation becomes more difficult because of corruptions of the input text caused by speech recognition errors.

The objective of this paper is to devise an automatic method of punctuation generation from speech input. This paper consists of eight sections. Section 2 introduces previous work in this area. Section 3 presents a combined system using prosody for punctuation generation and speech recognition. Section 4 describes the experimental setup using broadcast news data and discusses the evaluation measures used for the systems. Section 5 presents the results of punctuation generation for the reference text transcriptions. This section also presents results for applying the same punctuation generation approach to the output of a 1-best list speech recogniser, where the recogniser output does not include any punctuation marks. Section 6 gives the results of punctuation generation combined with speech recognition. Section 7 examines the assumption made when combining punctuation generation with speech recognition, and also discusses the variation in punctuation between annotators. Finally, Section 8 concludes this paper.

2 Previous work

An automatic punctuation system, called Cyberpunc, which is based on only lexical information, was developed in (Beeferman et al., 1998). That system produced only commas, under the assumption that sentence boundaries are pre-determined. A post-processing step added commas to each punctuation-free sentence by applying an extended Language Model (LM) which accounts for punctuation.

A method of speech recognition with punctuation generation based on acoustic and lexical information was proposed and examined using read speech from three speakers in (Chen, 1999). When punctuation generation is performed simultaneously with speech recognition, it is important to assign acoustic "pronunciations" to each punctuation mark. Punctuation marks were treated as words in the speech recogniser, and acoustic baseforms of silence, breath, and other non-speech sounds were assigned to punctuation marks in the pronunciation dictionary.

Since many full stops and question marks are located at the end of a sentence, it is very important in punctuation generation to recognise sentence boundaries correctly. A sentence boundary recogniser using lexical information and pause duration was developed in (Gotoh & Renals, 2000). A sentence boundary recognition test was then developed to find the sequence of sentence boundary classes (either a "last-word" class or a "not-last-word" class) from words in speech recognition output by combining probabilities from a language model and from a pause duration model.

It is generally difficult to disambiguate the meaning of punctuation marks located at the ends of sentences. For example, a full stop can be used for an abbreviation, a decimal point, an end of sentence marker, or an abbreviation at the end of a sentence. A trainable system for the classification of punctuation mark types was presented in (Palmer & Hearst, 1997). In that system, parts-of-speech for words surrounding punctuation marks were estimated, and then the punctuation marks were classified into the different types.

It is known that there is a strong correspondence between discourse structure and prosodic information (Shriberg et al., 1998). A comparison between syntactic and prosodic phrasing was presented in (Fach, 1999). In his study, syntactic structures were generated by Abney's chunk parser (Abney, 1995) and prosodic structures were given by ToBI (Silverman et al., 1992) label files. This work showed that at least 65% of syntactic boundaries are coded in the prosodic boundaries for read speech.

A method using intonation to reduce Word Error Rate (WER) in speech recognition for spontaneous dialogue was described in (Taylor et al., 1998). In their research, a separate intonation model for each Dialogue Act (DA or classification whether an utterance is a statement, question, agreement and etc.) was applied to give a set of likelihoods for an utterance being one or another type of DA. Then a separate language model for each DA was applied to find the most likely DA sequence and the new speech recognition result.

In order to use prosodic information in discourse structure analysis including automatic punctuation, great attention has to be paid to the computational method for obtaining prosodic feature values, how to build a prosodic feature model, and how to combine a prosodic feature model with models for other information sources.

A combination methodology with a language model and a prosodic feature model was discussed in (Shriberg et al., 1998). In their work, the combination methodology was applied to DA classification. A Classification And Regression Tree (CART) (Breiman et al., 1983) was used to construct a prosodic feature model. In order to make the computation tractable, an assumption was introduced that the prosodic features were independent of the word once conditioned on the DA (a similar assumption was introduced in (Taylor et al., 1998)).

3 Automatic punctuation generation

In this section, a method for automatic punctuation generation is described for both the reference transcriptions and the transcriptions generated by automatic speech recognition. When automatic punctuation generation is performed with the reference texts, the sequence of words is already given. Therefore, the experiments are aimed at generating punctuation marks between words.

The broadcast news reference transcriptions contain information marking each speaker turn, mainly for purposes of language modelling. Speaker turns are marked by <s> and </s> symbols. As the speaker turn marks aid finding the location of punctuation marks, it is unrealistic to include this information at the input for punctuation generation. For this reason, the speaker turn marks were removed from the training and test data.

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

3.1 Automatic punctuation generation for reference transcriptions

Let Y be the punctuation mark sequence, W be the word sequence and R be the corresponding discourse structure and prosodic feature sequence, including pause and fundamental frequency. The automatic punctuation system aims to find the maximum *a posteriori* Y, Y_{MAP}, given W and R.

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

Now

$$
\begin{aligned}
P(Y|W,R) &= \frac{P(Y,W,R)}{P(W,R)} & (2)\\
&= \frac{P(Y,W,R)}{P(W,R)}\frac{P(Y,W)}{P(Y,W)}\frac{P(W)}{P(W)} & (3)\\
&= \frac{P(Y,W,R)}{P(Y,W)}\frac{P(Y,W)}{P(W)}\frac{P(W)}{P(W,R)} & (4)\\
&= \frac{P(R|Y,W)P(Y|W)}{P(R|W)} & (5)
\end{aligned}
$$

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

$$P(Y|W,R) \propto P(R|Y,W)P(Y|W) \tag{6}$$

Assuming that R depends only on Y,[1] and $P(R)$ is uniformly distributed,[2]

$$P(R|Y,W) = P(R|Y) = \frac{P(Y|R)P(R)}{P(Y)} \propto \frac{P(Y|R)}{P(Y)} \tag{7}$$

Let y_i be the ith punctuation mark and r_i be the ith prosodic feature. Applying the 1st order Markov assumption yields

$$p(y_i|r_1,...,r_n) \simeq p(y_i|r_i) \tag{8}$$

Also let y_i be conditionally independent i.e.

$$p(y_1,...,y_n|R) \simeq \prod_{i=1}^{n} p(y_i|R) \tag{9}$$

Then $P(Y|R)$ becomes

$$P(Y|R) = \prod_{i=1}^{n} p(y_i|r_i) \tag{10}$$

The probabilities in Equation 10 can be obtained, for instance, from the terminal nodes of classification trees (the classification tree will be described in Section 4.1). $P(Y|W)$ in Equation 6 can be obtained from a statistical language model. $P(Y)$ can be obtained from training data counts.

Among the many kinds of punctuation marks, this study is restricted to the examination of full stops, commas, and question marks, because there are sufficient occurrences of these punctuation marks in the training data to be able to generate models and in the test data to measure the results accurately.

For a string of n words $w_1, ..., w_n$, which does not have punctuation marks, the end of each word is a possible candidate for punctuation. Considering the three types of punctuation marks and No-Punctuation (NP), there are 4^n possible hypotheses for the punctuation of the input $w_1, ..., w_n$. The search for the best hypothesis can be achieved with the Viterbi search algorithm. Using this algorithm, the required time for the search for the best hypothesis is reduced to be linear in the length of input n. Figure 1 shows a sample Viterbi search process for the generation of punctuation for an example reference transcription. The bold line in Figure 1 depicts the best hypothesis. For this hypothesis, commas are generated at the end of the words "pensioners" and "savers". Details of the Viterbi search algorithm are given in (Rabiner & Juang, 1993).

[Figure 1]

Figure 2 illustrates the overall procedure of punctuation generation for reference transcriptions. The raw speech signal is time-aligned with the corresponding reference transcription. During this alignment process, the estimated start time and the end time of each word are found. Prosodic features are generated at the end of each word, and the probabilities $P(Y|R)$ are obtained from the prosodic feature model. $P(Y|W)$, the probability of the sequence of words and the possible punctuation marks, are calculated from a statistical language model. The best hypothesis with punctuation marks is generated using the Viterbi search algorithm.

[Figure 2]

3.2 Automatic punctuation generation combined with speech recognition

In the previous section, a punctuation generation method for the reference transcriptions was presented. It uses prosodic information along with acoustic and language model information. In this section, a speech recogniser which produces punctuation marks and speech recognition hypotheses simultaneously is described. This speech recogniser produces multiple hypotheses and then these hypotheses are re-scored by a prosodic model. In this speech recogniser, punctuation marks are treated as words in the language model and dictionary.

The correlation between punctuation and pauses for read speech was investigated in (Chen, 1999). These experiments showed that pauses closely correspond to punctuation marks. The correlation between pause lengths and sentence boundary marks was studied for broadcast news data in (Gotoh & Renals, 2000). In that 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. In this paper, the pronunciation of punctuation marks is registered as silence in the pronunciation dictionary. The effectiveness of this assumption will be examined in Section 7.1.

A prosodic feature model to predict punctuation can be built in the form of a classification tree. Probabilities from the prosodic feature model can then be incorporated by re-scoring multiple hypotheses each of which includes putative punctuation marks. The probability combination process can proceed as shown in Section 3.1.

Figure 3 illustrates the overall procedure in the generation of punctuation, when combined with speech recognition. Using language models and acoustic models, N-best hypotheses of speech recognition are produced from the raw speech signal. These N-best hypotheses contain punctuation marks. As these hypotheses contain the start time and the end time of every word contained in them, prosodic features are generated at the end of each word. Then the probability of prosodic features are measured from the prosodic feature model. The N-best hypotheses are re-scored using this probability of prosodic features, and the best hypothesis which includes punctuation marks is generated.

[Figure 3]

4 Experimental setup and evaluation measures

Two different sets of data, the Broadcast News (BN) text corpus and the second 100-hour of Hub-4 BN training data set, [3] are available as training data for the experiments conducted in this paper. The BN text corpus (named BNText92_97 in this paper) comprises a 184 million word BN text from the period of 1992-1997 inclusive. [4] The 100-hour BN acoustic training data set released for the 1998 Hub-4 evaluation (named BNAcoustic98) consists of acoustic data and its transcription.

The test data from the NIST 1998 Hub-4 broadcast news benchmark tests are used as test data for the evaluation of the proposed system. This test data is named TestBNAcoustic98. TestBNAcoustic98 comprises 3 hours of acoustic data and the transcription. Table 1 summarises the BN training and test data.

[Table 1]

4-gram LMs were produced by interpolating LMs trained on BNText92_97 and BNAcoustic98, using a perplexity minimisation method. As this LM is used by the speech recogniser, the transcriptions of BNText92_97 and BNAcoustic98 are converted into single-case retaining punctuation marks to produce LM probabilities for punctuation marks. However, only BNAcoustic98 is used for the implementation of a prosodic feature model, because acoustic data are not available for BNText92_97.

Evaluation of a system involves scoring the automatically annotated hypothesis text against a hand annotated reference text. Scoring of a text input is relatively simple because it compares punctuation marks in the reference text to those in the hypothesis text, and counts the number of matched punctuation marks.

However, when the input comes from speech, because of recogniser deletion, insertion and substitution errors, a straightforward comparison is no longer possible (Grishman & Sundheim, 1995). Instead, the reference and hypothesis texts must first be automatically aligned. This is a complex process and involves attempting to determine which part of recogniser output corresponds to which part of the transcript.

Once the alignment is completed, correct/incorrect decisions for all the punctuation marks can be made. We define the symbols as C for the number of correct punctuation marks, S for the number of substitution errors, D for the number of deletion errors, I for the number of insertion errors, N for the number of punctuation marks in reference, and M for the number of punctuation marks in hypothesis. From the above definitions, it is clear that $N = C+S+D$ and $M = C + S + I$.

Two important metrics for assessing the performance of an information extraction system are *recall* and *precision*. These terms are borrowed from the information retrieval community. Recall (R) refers to how much of the information that should have been extracted was actually correctly extracted. Precision (P) refers to the reliability of the information extracted. These quantities are defined as:

$$P = \frac{\text{number of correct punctuation marks}}{\text{number of punctuation marks in hypothesis}} = \frac{C}{M} \tag{11}$$

and

$$R = \frac{\text{number of correct punctuation marks}}{\text{number of punctuation marks in reference}} = \frac{C}{N} \tag{12}$$

Although theoretically independent, in practice recall and precision tend to operate in trade-off relationships. When you try to increase recall, you often lose precision. When you optimise precision, you do so at the cost of recall.

The F-measure (Makhoul et al., 1999) is the uniformly weighted harmonic mean of precision and recall:

$$F = \frac{RP}{(R+P)/2} = \frac{2C}{N+M} \tag{13}$$

Another evaluation metric called Slot Error Rate (SER) was defined in (Makhoul et al., 1999) as follows:

$$\mathrm{SER} = \frac{\text{number of punctuation generation errors}}{\text{number of punctuation marks in reference}} = \frac{S+D+I}{C+S+D} \tag{14}$$

The difference between SER and $(1-F)$ is the weight given to D and I. The value of $(1-F)$ is calculated as:

$$(1-F) = \frac{N+M-2C}{N+M} = \frac{S+(D+I)/2}{(N+M)/2} = \frac{S+(D+I)/2}{C+S+(D+I)/2} \tag{15}$$

4.1 Classification tree setup

Many easily computable prosodic features were investigated for Dialog Act (DA) classification in (Shriberg et al., 1998), for automatic topic segmentation in (Stolcke et al., 1999), and for information extraction in (Hakkani-Tur et al., 1999).

The prosodic features that were found to be most useful for these areas were applied in this paper. By considering the automatic punctuation generation task and the contribution of each prosodic feature for DA classification, a set of 10 prosodic features were investigated for punctuation generation. Table 2 lists these 10 features.

[Table 2]

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 seconds. The left window is the window before the word end, and the right window is just after the word end. “Good” F0 values are those greater than the minimum F0 (50Hz) and less than the maximum F0 (400Hz).

A prosodic feature model is constructed using Classification And Regression Tree (CART) (Breiman et al., 1983). Prosodic features for the classification tree generation are measured from BNAcoustic98. The cross validation method is used in a CART generation.

The overall contribution of different features can be measured by 'feature usage', which is the proportion of the number of times a feature is queried by the test data and can be measured by 'feature appearance', which is the number of times a feature is used as a classifying feature in non-terminal nodes. The degree of overall contribution of each feature is shown in Table 2.

The 'feature usage' of Pau_Len and Eng_Ratio is about 78%. This measure accounts for the position of the feature in the tree. The higher the feature is used in the tree, the greater the feature usage is.

5 Results: Post-processing approach for punctuation generation

In order to generate punctuation marks for the reference transcription, three different systems were developed: a language model only system (LMOnly), a prosodic model only system (CARTOnly), and the combination of these two systems (LM+CART). LMOnly was trained on 185M words of transcriptions (BNText92_97 and BNAcoustic98). As these transcriptions contain punctuation marks, the language models trained on these transcriptions can predict the locations and types of punctuation marks based on word sequences which do not contain punctuation marks. 4-gram LMs were used in LMOnly. CARTOnly was generated on the 10 prosodic features described in Table 2 from a 100 hour broadcast news (BNAcoustic98).

Using the scale factor (α) which is the weighting given to the prosodic feature model, the relative importance of the prosodic feature model and the language model can be controlled. The scale factor is included into the combination of these two models as follows:

$$\alpha \times \log P(R|Y) + \log P(Y|W) \quad (16)$$

In this section, the performance of these three systems are compared for punctuation generation for the reference transcriptions. The language model only system (LMOnly) gives an F-measure of 0.5717 and an SER of 72.25%. When LMOnly generates punctuation for the reference transcription, its precision (0.5966) is a little higher than its recall (0.5488). Surprisingly, the prosodic feature model alone (CARTOnly) outperforms LMOnly by 0.0521 in F-measure and by 0.54% in SER. For CARTOnly, the recall (0.7414) is much higher than the precision (0.5383). These results show that CARTOnly produces a relatively high number of punctuation marks, while the accuracy of the generated punctuation is relatively poor.

As recall is much higher than precision for CARTOnly and precision is slightly higher than recall for LMOnly, the two information sources, one from lexical information and the other from prosodic feature information, are expected to be complementary. By combining these two models, the results are greatly improved. The combined system (LM+CART) produces an F-measure of 0.7830 with an SER of 32.30%, a precision of 0.7638 and a recall of 0.8031. These results are obtained when the scale factor (α) of 2.0 is applied. The F-measure attains a maximum at a scale factor of 2.0. The SER attains a minimum at a scale factor of 1.8. The results of automatic punctuation generation for the reference transcript are summarised in Table 3.

[Table 3]

The performance of LM+CART varies as the scale factor changes. Figure 4 describes how F-measure, precision, recall and SER change with the scale factor. The greater the scale factor for the prosodic feature model, the greater the recall because recall is much higher than precision for CARTOnly. Precision has a maximum value at a scale factor of 1.8. The F-measure attains a maximum of 0.7830 at a scale factor of 2.0. The SER attains a minimum of 32.12% at a scale factor of 1.8.

If the concept of scale factor is not introduced for this experiment, the probabilities from the language model and those from the prosodic feature model are combined 1:1. When a scale factor of 1.0 is applied, the F-measure is 0.7668 and the SER is 34.16%. By the introduction of a scale factor, the F-measure is improved by 0.0162 (2.11% relative) and the SER by 2.04% (5.97% relative).

[Figure 4]

The automatic punctuation generation method can be applied to the 1-best output of a speech recogniser. The 1-best output is first time aligned. Based on the time alignment information, prosodic features were generated. As in the approach applied for the punctuation generation of reference transcripts in Section 3.1, the best sequence of punctuation marks for this 1-best output is found using the prosodic feature model and an LM trained on texts which contain punctuation marks.

The HTK system for Broadcast News (BN) transcription (Woodland, 2002) running under 10 times real time (10xRT) (Odell et al., 1999) was used for the task of speech recognition. The first step of the system is a segmentation stage which converts the continuous input stream into segments with the aim of each segment containing data from a single speaker and a single audio type. Each segment is labelled as being either a wide-band or narrow-bandwidth signal.

The actual recogniser runs in two passes which both use cross-word triphone decision-tree state clustered HMMs with Gaussian mixture output distributions and a N-gram language model. The first pass uses gender-independent (but bandwidth-specific) HMMs with a 60k trigram language model to get an initial transcription for each segment. This transcription is used to determine the gender label for the speaker in each segment by alignment with gender-dependent HMMs. Sets of segments with the same gender/bandwidth labels are clustered for unsupervised Maximum Likelihood Linear Regression (MLLR) (Leggetter & Woodland, 1995) adaptation. The MLLR transforms for each set of clustered segments are computed using the initial transcriptions of the segments and the gender-dependent HMMs used for the second pass. The adapted HMMs along with a 4-gram language model is used in the second stage of decoding and produces the final output. [5]

Implementation details of the HTK BN transcription system (with few constraints on computing power) were given in (Woodland et al., 1998; Woodland et al., 1999), and those of the HTK 10xRT BN transcription system were described in (Odell et al., 1999). In order to speed up the full system, the 10xRT system uses simpler acoustic models and a simplified decoding strategy.

Using the HTK 10xRT system, speech recognition is performed first for TestBNAcoustic98. The WER of the speech recogniser is measured as 16.7%. The difference between the reported performance in (Pallett et al., 1999) and the performance measured in this paper is 0.6%. The system used in this paper differs from the HTK 10xRT system used in the 1998 Hub-4 BN benchmark test in four aspects: the absence of a category-based language model (Niesler et al., 1998), the amount of language model training data, the difference in vocabulary size, and the absence of a procedure to obtain more precise word start and end time information. The HTK 10xRT BN transcription system reported 16.1% overall WER for the NIST 1998 Hub-4 BN benchmark test (Pallett et al., 1999).

The system which generates punctuation marks from the 1-best output of the 10xRT system is named as LM+CART_ASR1Best. The trends of F-measure and SER of LM+CART_ASR1Best are similar to the automatic punctuation generation system for the reference transcription (LM+CART). The SER of LM+CART_ASR1Best reaches a minimum at $\alpha = 1.93$ and its F-measure a maximum at $\alpha = 2.10$. The results of LM+CART_ASR1Best are measured at $\alpha = 2.10$. Table 4 shows the results of LM+CART_ASR1Best.

[Table 4]

6 Results: Integration of punctuation generation within a speech recognition system

In the previous section, punctuation generation results were shown for the reference transcriptions. In this section, experimental results for punctuation generated as part of the speech recognition output are discussed.

Table 5 shows speech recognition results of the HTK 10xRT BN system under 3 different conditions for TestBNAcoustic98. In the first condition, punctuation is not included in the training or test data. The WER of the speech recogniser under this condition is measured at 16.71%. In the second condition, punctuation marks are included in the reference transcriptions and the recogniser output. The WER of the speech recogniser under this condition is increased to 22.73%. This degradation is caused by two factors: additional errors from other words due to the introduction of punctuation marks into the vocabulary, and errors in mis-recognising the punctuation marks themselves.

In order to check which factor contributes more to the degradation, punctuation marks are generated and these marks are then removed from the references and the hypotheses in the third condition. In this condition, the WER of the speech recogniser is measured at 17.04%. Comparing the difference in WER under the first and third conditions and the difference in WER under the second and third conditions, the degradation in WER after including punctuation marks mainly comes from errors in mis-recognising the punctuation marks themselves.

[Table 5]

The second condition is used as the baseline condition for our automatic punctuation generation system with speech recognition. The punctuation generation system with this condition is named as ASR+LMPunc. Using ASR+LMPunc, ASR+LMPunc_H100 generates 100 hypotheses and re-scores these hypotheses on a segment basis using the prosodic feature model. After re-scoring, the best hypotheses for each segment are combined in ASR+LMPunc_H100. The performance of ASR+LMPunc_H100 varies as the scale factor for the prosodic model changes. Figure 5 describes how both the WER and the WER after punctuation is removed from reference and hypothesis (WER′) change with the scale factor. WER is minimised with a scale factor of 0.71, and WER′ is minimised with a scale factor of 0.79.

[Figure 5]

Although the amount of improvement in terms of WER is small, it is very important that these results show that there is a possibility of performance improvement in speech recognition using prosodic feature information. [6] The prosodic feature model used in this paper is focused only on the classification of punctuation marks. Therefore, the punctuation types of the words which do not have punctuation marks at the end of these words are categorised as a single group: No-Punctuation (NP). In spite of this simple categorisation for the punctuation type of words which do not have punctuation marks, the WER after punctuation is removed also decreases.

[Figure 6]

[Figure 7]

Figure 6 shows the variation of F-measure and SER according to scale factor. Figure 7 shows that of precision and recall. The bigger the scale factor for the prosodic feature model, the bigger the recall and the smaller the precision. The value of the F-measure attains its maximum of 0.4400 when the scale factor is 1.93. SER attains its minimum of 83.13% at the scale factor of 0.79.

If re-scoring with prosodic feature model is not performed, the F-measure of the system is 0.3687, and the SER of the system is 85.02%. By the introduction of re-scoring with the prosodic feature model, the F-measure is improved by 0.0713 (19.34% relative) and the SER by 1.89% (2.22% relative).

Table 6 summarises these results. As the punctuation generation is combined with speech recognition, it is worth checking the result of punctuation generation when the best speech recognition performance is achieved. The precision, recall and F-measure are measured as 0.6072, 0.3319, and 0.4292 respectively at the scale factor of 0.79 when WER$'$ attains its minimum. At this scale factor, SER attains its minimum value of 83.13% too. These results show that the result of punctuation generation can be improved by re-scoring multiple hypotheses using a prosodic feature model while also improving speech recognition WER with respect to the results obtained from the baseline speech recognition system which generates punctuation marks with 1-best speech recognition output.

[Table 6]

The results of ASR+LMPunc_H100 are compared with those of LM+CART_ASR1Best. As LM+CART_ASR1Best uses the 1-best output of the speech recogniser without punctuation marks, WER$'$ of LM+CART_ASR1Best is not affected by degradation due to the inclusion of punctuation marks into the vocabulary. LM+CART_ASR1Best shows better performance in terms of F-measure and WER$'$, but poorer in terms of WER and SER. If precision is more important than recall, ASR+LMPunc_H100 is the better system, but if recall is more important than precision, LM+CART_ASR1Best is shown to be better.

The values of precision vary around 0.60 while the values of recall vary around

0.30. Comparing these results to the results of punctuation generation for the reference transcription shown in Section 5, the precision is satisfactory, but the recall is too low. This suggests that an insufficient number of punctuation marks are generated in the hypotheses. As stated previously, in this paper, the pronunciation of punctuation mark is assumed to be silence. This is only a rough approximation. This assumption will be analysed in Section 7.1.

The style of punctuation varies between writers. For example, there is a difference between British and American punctuation of lists: A, B, C, and D versus A, B, C and D (the sentence in Figure 1 is an example of this difference). The inherent uncertainty in punctuation will be discussed in Section 7.2.

7 Discussion

The pronunciation of punctuation marks was assumed to be that of silence. In Section 7.1, the effectiveness of this assumption is examined. In addition, Section 7.2 discusses the variation between different annotators for annotating punctuation marks.

7.1 The effectiveness of the assumption for punctuation mark pronunciation

The reference word sequence of TestBNAcoustic98 was time aligned with its acoustic data. This word sequence does not contain any punctuation mark. Then, the duration of the models 'sp' and 'sil' [7] were measured at the end of each word. Table 7 shows the ratio of the presence of silence for each punctuation mark type. About 90% of full stops and question marks are related to silence, but pauses do not exist at about 40% of commas. In addition, pauses are measured at the end of about 15% of words where no punctuation is located.

[Table 7]

7.2 The variation of punctuation between annotators

The use of punctuation is documented in manuals and in hand-books such as in (University of Chicago, 1993; Shaw, 1993). However, the style of punctuation varies between writers and between type of text (Chen, 1999). In addition, punctuation marks are used to change the meaning of sentences. In this section, the punctuation variation between annotators is measured.

The first 1000 words of TestBNAcoustic98 were used for this experiment. As capitalisation information gives cues to the location of sentence boundaries, these 1000 words were de-capitalised. Three native speakers of British English were asked to add punctuation marks between the test words wherever punctuation is necessary. Only commas, full stops and question marks were permitted as punctuation marks. These three annotators worked on the de-capitalised written text only. They did not listen to the spoken text. Although this experiment was performed with a small amount of text and a small number of annotators, it gives some idea as to the variation in punctuation between different annotators for the domain of broadcast news. Table 8 summarises the experimental conditions. Table 9 gives an example of the variation of punctuation between annotators.

[Table 8]

[Table 9]

In the reference transcription of TestBNAcoustic98, there are 43 commas and 54 full stops between the first 1000 words of TestBNAcoustic98. Table 10 shows the differences between the punctuation in the provided reference transcription and each annotator's transcription. These differences are measured in terms of precision, recall, F-measure and SER. On average, the F-measure for the three annotators was measured as 0.7199.

[Table 10]

Table 11 shows the variations in punctuation between annotators. These variations were measured in terms of precision, recall, F-measure and SER, regarding an annotator's text as the reference and another annotator's text as the hypothesis. On average, the F-measure between annotators was measured as 0.7113.

[Table 11]

The amount of this variation between annotators is quite substantial. Even though the acoustic data for the text is provided when the reference text was transcribed, the single set of punctuation marks provided in the reference text are certainly not a perfect measure. This variation may partly account for reported punctuation generation errors.

8 Conclusions

A punctuation generation system which incorporates prosodic information along with acoustic and language model information has been described. Experiments were conducted first for the reference transcriptions. In these experiments, prosodic information was shown to be more useful than language model information. When these information sources are combined, an F-measure of up to 0.7830 was obtained for adding punctuation to a reference transcription.

This method of punctuation generation can also be applied to the 1-best output of a speech recogniser. The 1-best output is first time aligned. Based on the time alignment information, prosodic features are generated. As in the approach applied in the punctuation generation for reference transcriptions, the best sequence of punctuation marks for this 1-best output is found using the prosodic feature model and an LM trained on texts which contain punctuation marks.

As an alternative, a modified conventional speech recogniser was used to produce punctuation marks and speech recognition hypotheses simultaneously. Multiple hypotheses from the recogniser were re-scored by the prosodic model. Rescoring with the prosodic model increased the F-measure by 19% relative to the 1-best output from the modified speech recogniser. At the same time, a small reduction in word error rate was obtained over the 1-best output from the modified speech recogniser.

This modified speech recogniser is based on the assumption that the pronunciation of punctuation marks is silence. Its results were compared with those from the 1-best output in which punctuation marks were generated by post-processing the 1-best output of standard speech recogniser. If precision is more important than recall, then the modified speech recogniser gives the better results, but if recall is more important than precision, then the method using the 1-best output of standard speech recogniser is shown to be better.

The variation in punctuation annotation between annotators was investigated. Although this experiment was performed with only a small amount of text and three annotators, it gives some indication of the substantial variation between different annotators for punctuation.

9 Acknowledgements

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

Footnote

[1] Similar to the assumption in (Shriberg et al., 1998; Taylor et al., 1998), where it is assumed that prosodic features are independent of the words once conditioned on the dialogue act, we introduced an assumption that R depends only on Y. The assumption in (Shriberg et al., 1998) is a simplification to make the computation tractable. Clearly, the independence assumption in our paper is violated for the energy. However, for practical reasons, we introduce this independence assumption.

[2] A similar assumption was introduced in (Shriberg et al., 1998) as the probability of a prosodic feature sequence is the same for all dialogue act types. In addition, the probability of an acoustic observation sequence is assumed to be independent of the word sequence in speech recognition. It is true that P(R) is not uniformly distributed, but this assumption is introduced to make the computation tractable.

[3] The actual amount of transcribed acoustic data is 71 hours.

[4] The 1992-1996 data was provided by the LDC (http://www.ldc.upenn.edu) and the 1997 data was provided by the Primary Source Media.

[5] Note that the same form of language model is used whether or not the training data contains punctuation marks.

[6] This difference in WER is not statistically significant.

[7] These are the models for silence in the HTK BN system.

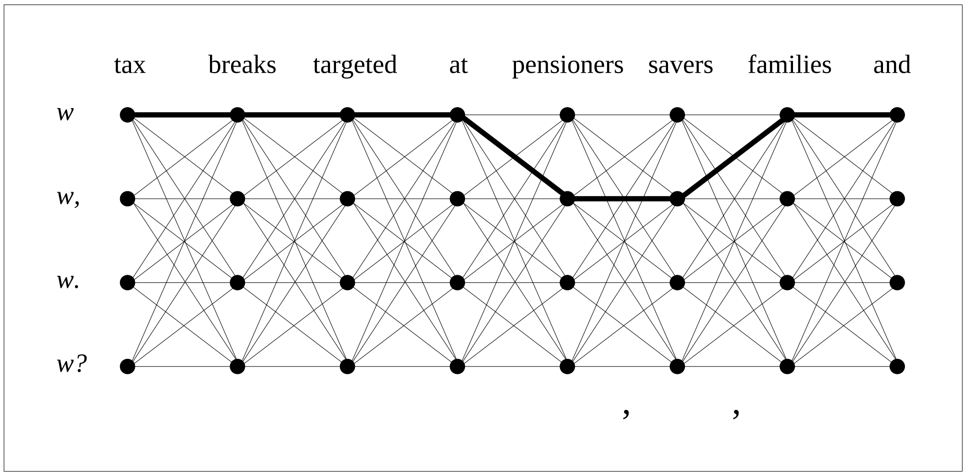

Fig. 1. Viterbi search process for the generation of punctuation for an example reference transcription. The bold line depicts the best hypothesis. Punctuation marks at the bottom are generated according to this best hypothesis.

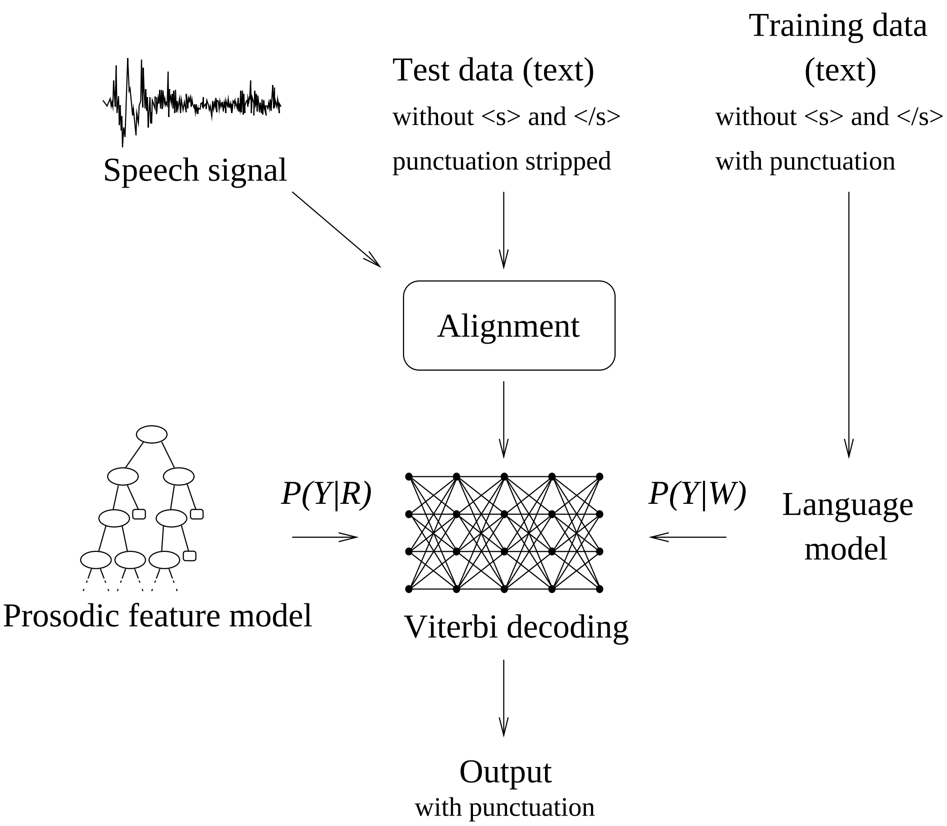

Fig. 2. Overall procedure for punctuation generation from reference transcriptions

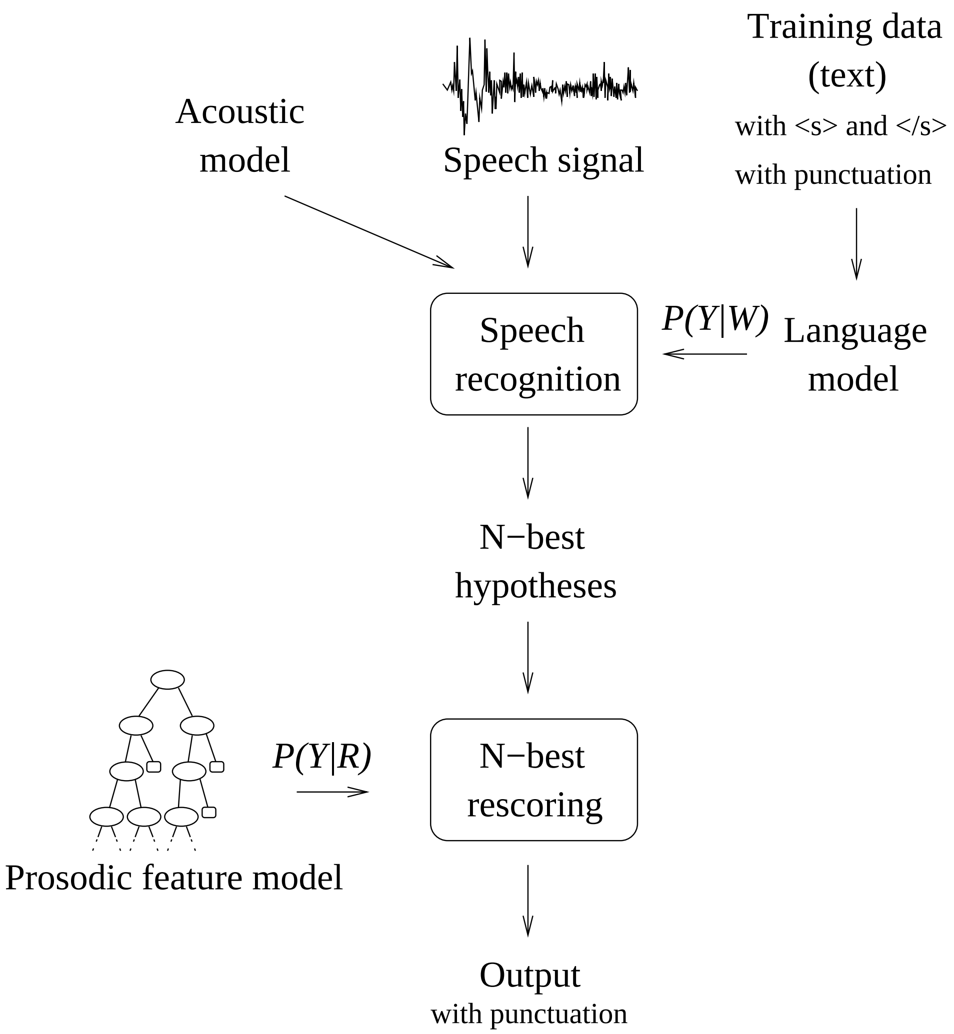

Fig. 3. Overall procedure for punctuation generation combined with speech recognition

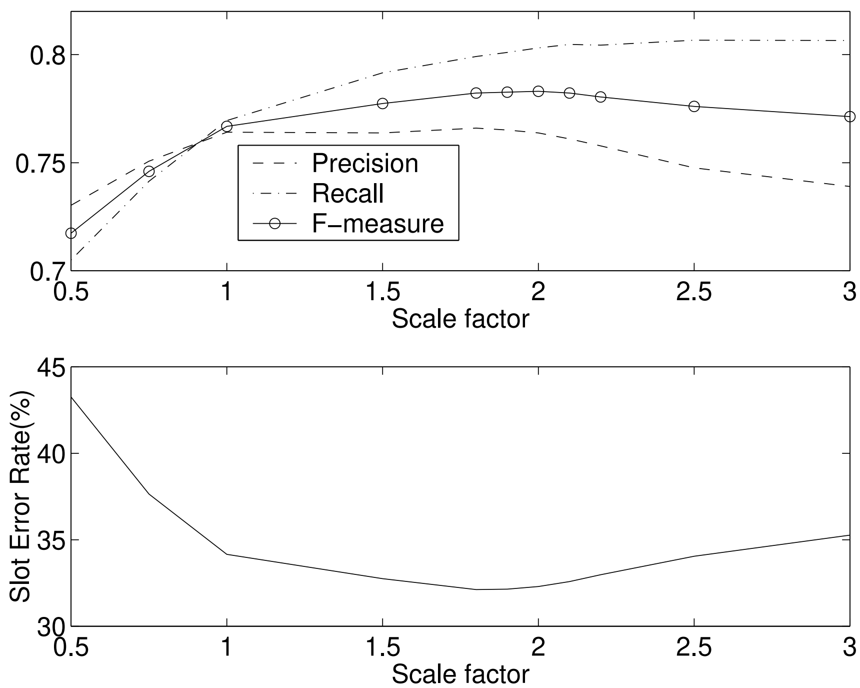

Fig. 4. Automatic punctuation generation results of LM+CART with different scale factors

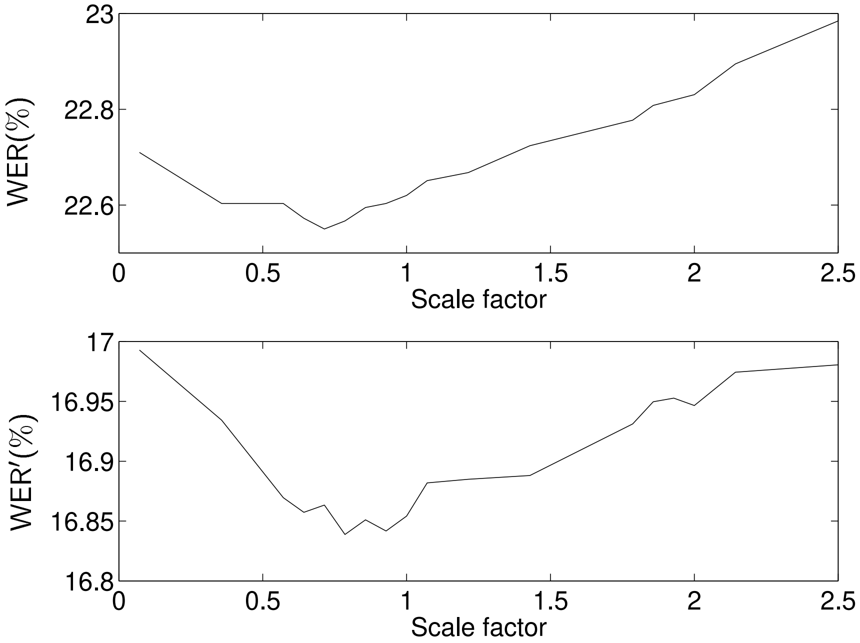

Fig. 5. WER and WER′ of ASR+LMPunc_H100 with different scale factors (ASR+LMPunc_H100: Final hypothesis from re-scored 100 hypotheses; WER: Word Error Rate; WER′: WER after punctuation is removed from a reference and a hypothesis)

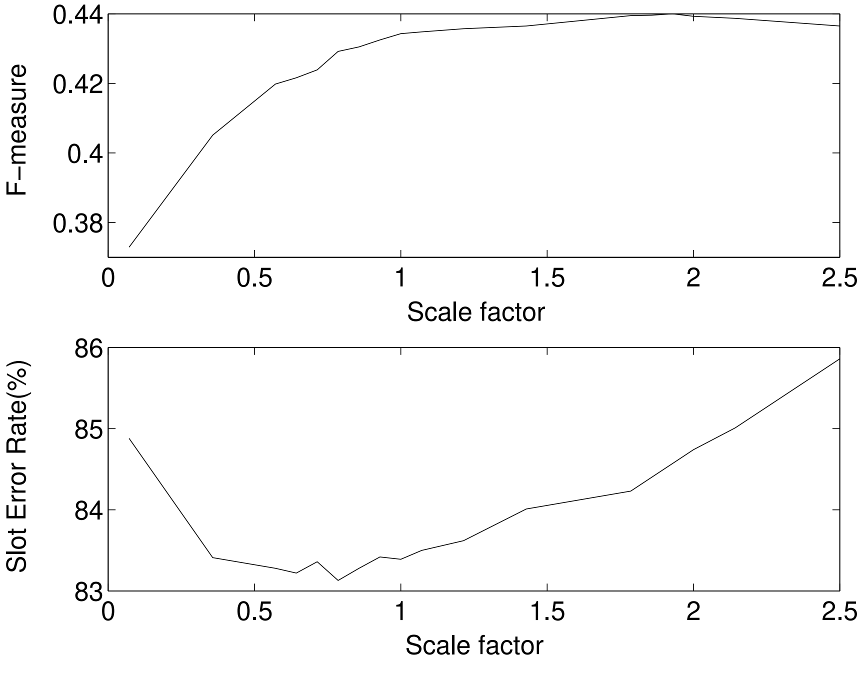

Fig. 6. F-measure and SER of ASR+LMPunc_H100 with different scale factors (ASR+LMPunc_H100: Final hypothesis from re-scored 100 hypotheses; SER: Slot Error Rate)

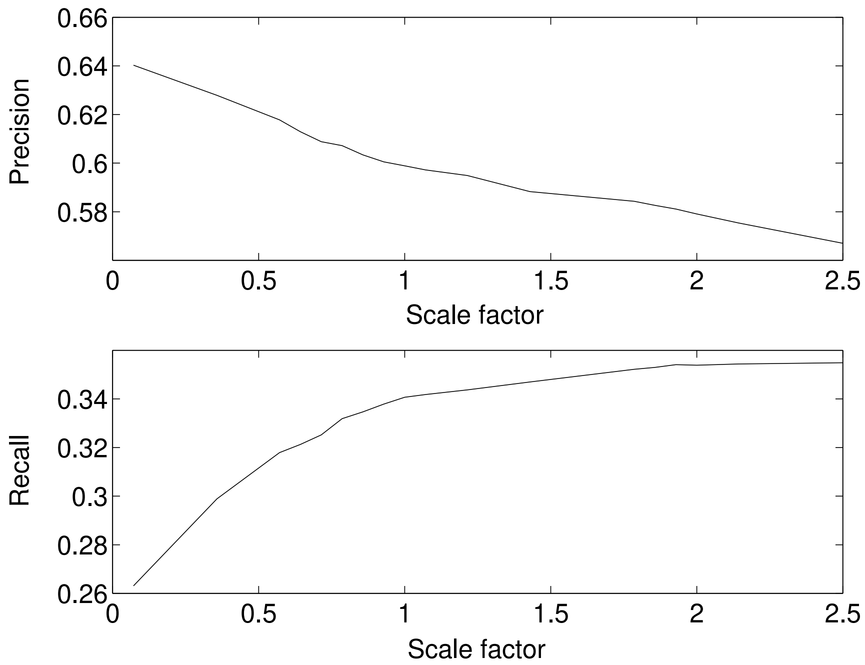

Fig. 7. Precision and recall of ASR+LMPunc_H100 with different scale factors (ASR+LMPunc_H100: Final hypothesis from re-scored 100 hypotheses)

Name	Description	#Words	Purpose	Acoustic data
BNtext92_97	1992_97 BN texts	184M	Training data	Not available
BNAcoustic98	100 hrs of Hub-4 data (1998)	774K	Training data	Available
TestBNAcoustic98	1998 benchmark test data	32K	Test data	Available

Table 1
Description of broadcast news training and test data

Name	Description	Feature appearance	Feature usage
Pau_Len	Pause length at the end of a word	672	0.5799
Dur_fr_Pau	Duration from the previous pause	539	0.0230
Avg_F0_L	Mean of good F0s in left window	342	0.0246
Avg_F0_R	Mean of good F0s in right window	230	0.0363
Avg_F0_Ratio	Avg_F0_R/Avg_F0_L	261	0.0461
Cnt_F0_L	No. of good F0s in left window	204	0.0429
Cnt_F0_R	No. of good F0s in right window	230	0.0176
Eng_L	RMS energy in left window	203	0.0038
Eng_R	RMS energy in right window	160	0.0252
Eng_Ratio	Eng_R/Eng_L	239	0.2006

Table 2
Description of each prosodic feature and its contribution for the CART trained by BNAcoustic98 and tested by TestBNAcoustic98 (Feature usage: proportion of the number of times a feature is queried. Feature appearance: the number of times a feature is used as a classifying feature. Window length = 0.2 sec, 50Hz $\leq$ good F0 $\leq$ 400Hz)

System	Precision	Recall	F-measure	SER(%)
LMOnly	0.5966	0.5488	0.5717	72.25
CARTOnly	0.5383	0.7417	0.6238	71.71
LM+CART (α=2.0)	0.7638	0.8031	0.7830	32.30

Table 3
Automatic punctuation generation results for reference transcripts (LMOnly: language model only; CARTOnly: prosodic feature model only; LM+CART: combination of LMOnly and CARTOnly; α = scale factor to the prosodic feature model; SER: Slot Error Rate)

System	WER	WER$'$	Precision	Recall	F-measure	SER
LM+CART_ASR1Best (α=2.10)	23.08	16.71	0.5329	0.4304	0.4762	88.32

Table 4
Automatic punctuation generation results of LM+CART_ASR1Best (WER: Word Error Rate (%); WER$'$: WER after punctuation is removed from a reference and a hypothesis; SER: Slot Error Rate (%))

Remarks	WER
Punctuation excluded	16.71
Punctuation included	22.73
Punctuation marks are generated and then removed from reference and hypothesis	17.04

Table 5
Speech recognition results of the HTK 10xRT BN system under 3 different conditions (WER = Word Error Rate (%))

System	WER	WER′	Precision	Recall	F-measure	SER
ASR+LMPunc	22.73	17.04	0.6425	0.2585	0.3687	85.02
ASR+LMPunc_H100 (α=0.79)	22.57	16.84	0.6072	0.3319	0.4292	83.13
ASR+LMPunc_H100 (α=1.93)	22.82	16.95	0.5811	0.3541	0.4400	84.57

Table 6
Results of automatic punctuation generation with speech recognition (ASR+LMPunc: Baseline system. No re-scoring; ASR+LMPunc_H100: Final hypothesis from re-scored 100 hypotheses; WER: Word Error Rate (%); WER′: WER after removing punctuation from a reference and a hypothesis; SER: Slot Error Rate (%))

Punctuation mark	Number of silences(%)
NP	4352/28218 (15.42)
,	948/1565 (60.58)
.	1590/1794 (88.63)
?	45/49 (91.84)

Table 7
Number of times silence is present for each punctuation mark type (NP: No-Punctuation)

Condition	Description
Text source	First 1000 words in TestBNAcoustic98
Writing style	Single case. No punctuation mark
Annotator	Three native British English speakers

Table 8
Experimental conditions for investigating the variation in punctuation mark annotation between different annotators. The annotators work on the de-capitalised written text only. They do not listen to the spoken text.

Annotator 1: china, another market with big potential, is also having second thoughts about culinary competition. with american fast food joints, china's domestic food industry recently concluded that in order to become a great world power, a nation needs to conquer the globe with its own fast food.
Annotator 2: china, another market with big potential, is also having second thoughts about culinary competition with american fast food joints. china's domestic food industry recently concluded that, in order to become a great world power, a nation needs to conquer the globe with its own fast food.

Table 9
Example of the variations in punctuation marks between annotators. The same sequence of de-capitalised words was given to each annotator.

Source of hypothesis text	Precision	Recall	F-measure	SER(%)
Annotator 1	0.7558	0.6701	0.7104	49.48
Annotator 2	0.7158	0.7010	0.7083	47.42
Annotator 3	0.7448	0.7371	0.7409	44.85

Table 10
The difference of putting punctuation marks between the provided reference transcription and each annotator's transcription. The provided transcription is regarded as the reference and each annotator's transcription as the hypothesis. (SER: Slot Error Rate)

Source of text		Results of variations			
Reference	Hypothesis	Precision	Recall	F-measure	SER(%)
Annotator 1	Annotator 2	0.6421	0.7093	0.6740	60.47
Annotator 1	Annotator 3	0.7188	0.8023	0.7582	44.19
Annotator 2	Annotator 3	0.6979	0.7053	0.7016	49.47

Table 11
Variations in punctuation between annotators. Results of variations are measured regarding an annotator's text as the reference and another annotator's text as the hypothesis. (SER: Slot Error Rate)

References

[NIST, 1998] NIST, 1998. NIST Hub-4 Information Extraction (Named Entity) Broadcast News Benchmark Test Evaluation. Available at ftp://jaguar.ncsl.nist.gov/csr98/h4iene_98_official_scores_990107/index.htm.

[University of Chicago, 1993] University of Chicago, 1993. The Chicago Manual of Style, 14th Edition. The University of Chicago Press.

[Abney, 1995] Abney, S., 1995. Chunks and Dependencies: Bringing Processing Evidence to Bear on Syntax. Computational Linguistics and the Foundations of Linguistic Theory. pp. 145–164.

[Beeferman et al., 1998] Beeferman, D., Berger, A., and Lafferty, J., 1998. Cyberpunc: A Lightweight Punctuation Annotation System for Speech. Proc. ICASSP. Seattle, pp. 689–692.

[Breiman et al., 1983] Breiman, L., Friedman, J., Olshen, R., and Stone, C., 1983. Classification and Regression Trees. Wadsworth and Brooks.

[Chen, 1999] Chen, C., 1999. Speech Recognition with Automatic Punctuation. Proc. Eurospeech. Budapest, pp. 447–450.

[Fach, 1999] Fach, M., 1999. A Comparison Between Syntactic and Prosodic Phrasing. Proc. Eurospeech. Budapest, pp. 527–530.

[Gotoh & Renals, 2000] Gotoh, Y. and Renals, S., 2000. Sentence Boundary Detection in Broadcast Speech Transcripts. Proc. ISCA ITRW ASR. Paris, pp. 228–235.

[Grishman & Sundheim, 1995] Grishman, R. and Sundheim, B., 1995. Design of the MUC-6 Evaluation. Proceedings of the 6th Message Understanding Conference. pp. 1–11.

[Hakkani-Tur et al., 1999] Hakkani-Tur, D., Tur, G., Stolcke, A., and Shriberg, E., 1999. Combining Words and Prosody for Information Extraction from Speech. Proc. Eurospeech. Budapest, pp. 1991–1994.

[Leggetter & Woodland, 1995] Leggetter, C. and Woodland, P., 1995. Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models. Computer Speech and Language. 9, 171–185.

[Makhoul et al., 1999] Makhoul, J., Kubala, F., Schwartz, R., and Weischedel, R., 1999. Performance Measures for Information Extraction. Proceedings of the DARPA Broadcast News Workshop. Dulles, Virginia, pp. 249–252.

[Niesler et al., 1998] Niesler, T., Whittaker, E., and Woodland, P., 1998. Comparison of Part-Of-Speech and Automatically Derived Category-Based Language Models for Speech Recognition. Proc. ICASSP. Seattle, pp. 177–180.

[Odell et al., 1999] Odell, J., Woodland, P., and Hain, T., 1999. The CUHTK-Entropic 10xRT Broadcast News Transcription System. Proceedings of the DARPA Broadcast News Workshop. Dulles, Virginia, pp. 271–275.

[Pallett et al., 1999] Pallett, D., Fiscus, J., Garofolo, J., Martin, A., and Przybocki, M., 1999. 1998 Broadcast News Benchmark Test Results: English and Non-English Word Error Rate Performance Measures. Proceedings of the DARPA Broadcast News Workshop. Dulles, Virginia, pp. 5–12.

[Palmer & Hearst, 1997] Palmer, D., and Hearst, M., 1997. Adaptive Multilingual Sentence Boundary Disambiguation. Computational Linguistics. 23(2), 241–269.

[Rabiner & Juang, 1993] Rabiner, L. and Juang, B., 1993. Fundamentals of Speech Recognition. Prentice Hall.

[Shaw, 1993] Shaw, H., 1993. Punctuate it Right! Harper-Collins.

[Shriberg et al., 1998] Shriberg, E., Bates, R., Stolcke, A., Taylor, P., Jurafsky, D., Ries, K., Coccaro, N., Martin, R., Meteer, M., and Ess-Dykema, C., 1998. Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech? Language and Speech. 41(3-4), 439–487.

[Silverman et al., 1992] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., and Hirschberg, J., 1992. ToBI: A Standard for Labelling English Prosody. Proc. ICSLP. Banff, Canada, pp. 867–870.

[Stolcke et al., 1999] Stolcke, A., Shriberg, E., Hakkani-Tur, D., Tur, G., Rivlin, Z., and Sonmez, K., 1999. Combining Words and Speech Prosody for Automatic Topic Segmentation. Proceedings of the DARPA Broadcast News Workshop. Dulles, Virginia, pp. 61–64.

[Taylor et al., 1998] Taylor, P., King, S., Isard, S., and Wright, H., 1998. Intonation and Dialog Context as Constraints for Speech Recognition. Language and Speech. 41(3-4), 489–508.

[Woodland et al., 1998] Woodland, P., Hain, T., Johnson, S., Niesler, T., Whittaker, E., and Young, S., 1998. The 1997 HTK Broadcast News Transcription System. Proceedings of the Broadcast News Transcription and Understanding Workshop. Lansdowne, Virginia.

[Woodland et al., 1999] Woodland, P., Hain, T., Moore, G., Niesler, T., Povey, D., Tuerk, A., and Whittaker, E., 1999. The 1998 HTK Broadcast News Transcription System: Development and Results. Proceedings of the DARPA Broadcast News Workshop. Dulles, Virginia, pp. 265–270.

[Woodland, 2002] Woodland, P., 2002. The development of the HTK Broadcast News transcription system: An overview. Speech Communication. 37(1-2), 47–67.