# Punctuation Annotation using Statistical Prosody Models

# Punctuation Annotation using Statistical Prosody Models

*Heidi Christensen* *Yoshihiko Gotoh* *Steve Renals*

University of Sheffield, Department of Computer Science
Regent Court, 211 Portobello St., Sheffield S1 4DP, UK
e-mail: {h.christensen, y.gotoh, s.renals}@dcs.shef.ac.uk

## Abstract

This paper is about the development of statistical models of prosodic features to generate linguistic meta-data for spoken language. In particular, we are concerned with automatically punctuating the output of a broadcast news speech recogniser. We present a statistical finite state model that combines prosodic, linguistic and punctuation class features. Experimental results are presented using the Hub–4 Broadcast News corpus, and in the light of our results we discuss the issue of a suitable method of evaluating the present task.

## 1. Introduction

Basic automatic speech recognition (ASR) systems transform audio to raw streams of words. To make these word transcripts more usable in applications such as information retrieval and speech understanding systems, some structuring in the form of meta-data needs to be inserted into the word stream. This has initiated an interest in automatic structuring techniques such as topic segmentation, sentence boundary identification, named entity classification and automatic punctuation generation. To fully exploit the nature of the available data, many researchers have sought to combine the linguistic information provided by the ASR generated transcripts (albeit error prone), with complementary prosodic information extracted from the audio data.

In this paper we are particularly concerned with investigating the usefulness of different prosodic features for the identification of three types of punctuation marks: the full stop, the comma, and the question mark. Data are taken from the Hub–4 Broadcast News (BN) corpus, for which manually punctuated and processed transcripts are available.

The remainder of this section summarises the motives for making use of prosodic and linguistic information in our system and reviews recent progress concerning the automatic structuring of broadcast and spontaneous speech. Section 2 outlines the models we have employed, and our experimental setup is described in section 3. The results of these experiments are presented in sections 4 and 5. Finally we discuss evaluation measures for this task in section 6.

### 1.1. Prosodic and linguistic clues to structuring speech

Several studies have indicated that speakers use prosody (i.e. pitch, speech unit durations, and pausing) to impose structure on both spontaneous and read speech [1]. There is a high correlation between acoustic cues to discourse structure and the positions of punctuation marks in transcripts. Human speakers (and automatic speech synthesis systems) use prosody extensively to add meaning to word sequences, such as distinguishing questions from statements.

Since it provides information complementary to the word sequence, prosody is a potentially valuable source of additional information. In particular, prosodic information is extracted directly from the audio and is therefore independent of the specific performance level of the automatic speech recogniser.

Some textual clues may be used for identifying punctuation marks in transcripts (either textual data or speech recogniser output). The following example is extracted from the BN data collection:

> *That may be the case, but that is not the truth.*
> *Okay, thanks for your points. Politics as usual?*

In this and many other examples words like 'but' often follow a comma. Also note that in the last sentence, a purely linguistic model would not be able to detect this sentence as a statement or a question. However, prosodic clues would enable a listener to easily disambiguate the two cases.

### 1.2. Related work

Previous work concerning the structuring of speech recogniser transcripts has focused predominantly on the identification of topic changes, sentence boundaries, named entities and dialogue acts.

A common way of exploiting prosodic information (e.g. duration, pause, $F0$, and speaking rate) for these tasks has employed CART-style decision trees. Shriberg, Stolcke and coworkers investigated the capability of decision trees (both alone and in conjunction with word-level models) for the classification of dialogue acts in Switchboard [2] and the identification of topic and sentence boundaries in broadcast news [3]. The latter study indicated that although topic and sentence segmentation benefitted somewhat from the incorporation of prosodic information, named entity identification did not. It was concluded that the use of prosodic cues is task and corpus dependent; in particular, it was found that pause and pitch features were useful for segmenting news speech, whereas pause, duration and word-based cues dominated for natural conversation.

The Verbmobil project made extensive use of prosody to derive more information about discourse structure [4]. In that speech-to-speech translation system, prosody (mainly related to $F0$ and duration) was used to guide the rescoring of an $n$-best list of word hypotheses produced by the speech recogniser. A rather long feature vector (276 dimensions) was computed for each word, which was used in the identification of clause boundaries by a multi-layer perceptron (MLP). In [5] prosodic features and language models trained on the Verbmobil corpus were successfully used to search word-lattices for the most likely segmentation and the classification of dialogue acts.

Hirschberg and Nakatani [6] also investigated topic- and sub-topic segmentation on Broadcast News data. Prosodic features were used in a system for predicting whether a given frame

of speech belonged to an intonational phrase or to a break between phrases. This information was then used to identify intonational phrase boundaries, that begin and end 'topics'.

Automatic punctuation has only been investigated to a limited degree. Beeferman et al [7] presented an approach in which $n$-best lists were rescored using a comma-aware trigram language model. Chen [8] reported some small-scale experiments in which punctuation marks were treated as dictionary words, each with a predefined pronunciation. Kim and Woodland [9] combined prosodic and lexical information in a system designed for the identification of full stops, question marks and commas in Broadcast News data. A prosodic decision tree was tested alone and in combination with a language model, with some improvements reported through the use of the combined model.

Some of our previous work concerned the detection of sentence boundaries in broadcast speech transcripts [10]. An $n$-gram language model was combined with an alternative model estimated from pause duration information obtained from speech recogniser outputs. On a collection of British English broadcast news, experimental results showed that the pause duration model alone outperformed the language model, and that a combined model resulted in further improvement.

# 2. Punctuation Models

## 2.1. Prosodic information

We have used three classes of prosodic feature in this work.

- **Pause durations.** The speech recogniser models sentence boundaries and silence using an acoustic model corresponding to a pause. These two markers are also predicted by the trigram language model. The recogniser output also includes pause duration information, which are used as features on their own and in conjunction with the corresponding marker identity. This prosodic feature was the only prosodic feature investigated in our investigations of sentence boundary identifications [10].
- **Phoneme durations.** Full phoneme level alignments are provided by the speech recogniser, and this durational information has been extracted into two types of features: 1) the average duration of the phonemes in the preceding word; and 2) the average duration of the vowels in the preceding word. Both are normalised with typical phoneme durations for each individual show.
- **Pitch.** Pitch information is extracted from the acoustic data using the Edinburgh Speech Tools [11]. Features describing slope, range, and mean of the $F0$ regression line over the preceding word.

## 2.2. Linguistic information

The modelling of the linguistic information is based on word and punctuation mark sequences similar to that of language models in conventional ASR systems.

Generally, punctuation marks can be seen as being attached to the preceding word in the sentence. We formulate the problem of identifying punctuation marks as that of identifying the last word before a punctuation mark, given a sequence of words and prosodic features. Each word in a text will belong to either one of the punctuation classes or a "not-preceding-punctuation" class (denoted by $<$ c$_i$$>$ and ' •', respectively). Given this notation, a corresponding class sequence for the example given in the introduction is:

• • • • <,> • • • • • <.>
<,> • • • <.> • • <?>

Let $\mathcal{V}$ denote a vocabulary and $\mathcal{C}$ be a set of punctuation mark classes. We consider that $\mathcal{V}$ is similar to vocabulary for a conventional speech recognition system, typically containing tens of thousands of words, and no case information or other characteristics. Here, $\mathcal{C}$ contains four classes, <.>, <,>, <?> and ' •' as described above. We consider the joint probability of a sequence of words, $w_i^m = w_1, \ldots, w_m$, and corresponding punctuation class tokens, $c_i^m = c_1, \ldots, c_m$ :

$$p^{[L]}(w_1^m, c_1^m) = \prod_{i=1\ldots m} p^{[L]}(w_i, c_i | w_1^{i-1}, c_1^{i-1}). \qquad (1)$$

The superscript, $p^{[L]}$, indicates a linguistic modelling probability in contrast to the prosody models presented in section 2.1. Once a linguistic model is constructed, punctuation markings can be identified by searching the Viterbi path such that:

$$<\hat{c}_1^m> = \underset{c_1^m}{\operatorname{argmax}}\ p^{[L]}(w_1^m, c_1^m) \qquad (2)$$

for an unseen sequence of words, $w_1^m$.

We have investigated two different approaches that incorporate linguistic and prosodic information into an automatic punctuation system.

## 2.3. Finite state model approach

The first approach is similar to the one used in [10] for identifying sentence boundaries. A compound statistical finite state model is developed, where each state represents a joint occurrence of a word, a punctuation mark class and a set of prosodic features. We consider the joint probability of a sequence of prosodic features, $s_1^m$, along with corresponding sequences of word and class tokens ($w_1^m$ and $c_1^m$):

$$p(s_1^m, w_1^m, c_1^m) = \prod_{i=1\ldots m} p(s_i, w_i, c_i | w_{i-1}, c_{i-1}) \ , \qquad (3)$$

where we assume that the previous prosodic features do not influence the current prosodic features, or the current word and punctuation class tokens (figure 1).

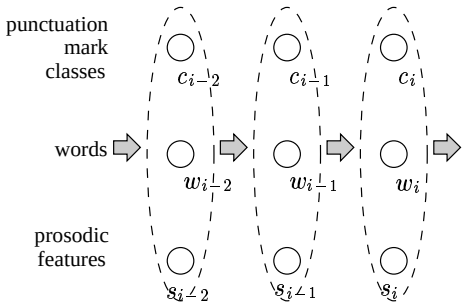


Figure 1: Topology for the punctuation mark model. The arrow represent the evolution of the states, consisting of punctuation mark classes, word, and prosodic components, rather than explicit probabilistic dependences.

The right side of (3) may be decomposed into the linguistic model component, $p^{[L]}$, and the prosodic model component, $p^{[P]}$, in two different ways. Decomposition **A** is mathematically

motivated, whereas decomposition **B** is heuristic and has proven to work well for sentence boundary identification [10]:

**Decomposition A:**

$$p(s_i, w_i, c_i|w_{i-1}, c_{i-1}) \sim \left\{p^{[P]}(s_i|w_i, c_i)\right\}^{\alpha} \times \left\{p^{[L]}(w_i, c_i|w_{i-1}, c_{i-1})\right\} , \quad (4)$$

where $\alpha$ is a weight factor introduced to compensate for distortions in the probability estimates occuring because $p^{[L]}$ and $p^{[P]}$ arise from different sources.

**Decomposition B:**

$$p(s_i, w_i, c_i|w_{i-1}, c_{i-1}) \sim \left\{p^{[P]}(w_i, c_i|s_i)\right\}^{\alpha} \times \left\{p^{[L]}(w_i, c_i|w_{i-1}, c_{i-1})\right\} . \quad (5)$$

These decompositions are discussed in [10].

### 2.4. MLP based approach

A second approach was motivated by the possible limitations of the finite state model approach to incorporate numerous prosodic features. We use MLPs to provide estimates for posterior probabilities, $p(c_i|w_i, s_i)$, and have investigated two main architectures.

In the **joint_MLP** system each MLP takes a feature vector containing a number of features (prosodic and linguistic) and attempt to classify one of the four punctuation mark classes <.>, <,>, <?> and ' •' (figure 2 *(a)*). This configuration allows us to investigate the usefulness of each feature, providing us with a confusion matrix corresponding to the particular subset of input features.

The **binary_MLP** system delegates the classification job even further, letting each MLP deal only with the decision between $c_i$ and $\neg c_i$ (figure 2 *(b)*). A set of MLPs are trained, each specialised in the identification of a particular punctuation mark, and allowing for a detailed analysis of strengths and weaknesses of the particular features.

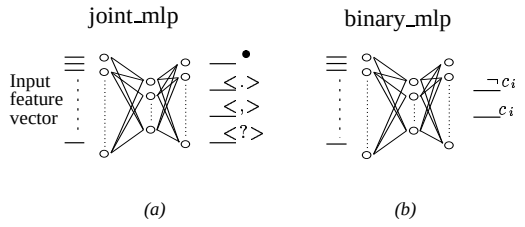


Figure 2: Schematic overview of the joint_MLP and binary_MLP network configurations.

## 3. Experimental setup

In our experiments, we have used audio and punctuated transcripts from a subset of the Hub–4 acoustic data[1]. On the 1998 evaluation set, the recogniser produced a Word Error Rate of about 21% [12]. Table 1 summarises the statistics for the training and testing sets.

[1]Obtainable from LDC `http://www.ldc.upenn.edu/`.

| Data part | # shows | #words | #. | #, | #? |
|---|---|---|---|---|---|
| Train | 101 | 675342 | 37526 | 26465 | 2169 |
| Test | 6 | 44714 | 2249 | 1498 | 143 |

Table 1: Statistics for the Hub–4 broadcast news subset.

We have employed two evaluation metrics based on recall/precision and the slot error rate.

Recall (*R*) and precision (*P*) are calculated in the usual way. A weighted harmonic mean (*P*&*R*), sometimes called the F-measure [13], may be calculated as a single summary statistic:

$$P\&R = \frac{2RP}{R+P}. \quad (6)$$

Although recall and precision are useful and informative measures, Makhoul *et al.* [14] have criticised the use of *P*&*R*, since it implicitly deweights missing and spurious identification errors compared with incorrect identification errors. They proposed an alternative measure, referred to as the slot error rate (*SER*), that equally weights three types of identification error (incorrect, missing, and spurious). *SER* is analogous to word error rate. It is obtained by

$$SER = \frac{I+M+S}{C+I+M} \quad (7)$$

where $C$, $I$, $M$, and $S$ denote the numbers of correct, incorrect, missing, and spurious identifications. Using this notation, recall and precision scores may be calculated at $R = C/(C+I+M)$ and $P = C/(C+I+S)$, respectively. In general, the lower the *SER* score, the better; for the others the higher the score, the better.

## 4. Results using finite state approach

Table 2 presents results from performing automatic punctuation on the test set. The table is organised so that the first column lists the scores when considering all punctuations in the full task. The subsequent columns two to four give the details for how well each of the individual punctuation marks were recognised. Results from two different features are listed. The top half of the table concerns the combination of the linguistic model and the model for the pause duration features. The second half displays results from using features describing the average phone duration. The very first entry in the table present the results from the testing using only linguistic information.

Concentrating first on the overall scores concerning the task of recognising all the three types of punctuation marks (column one) it is clear that both types of prosodic model are able to further increase the performance when combined with the linguistic model. Specifically for the *SER* is decreased from 1.04 for the linguistic model alone to 0.89 when in combination with the pause duration features, and to 0.90 when using the phone duration features. A similar pattern is described by the development in *P*&*R* scores that increase from 0.25 (linguistic) to a maximum of 0.42 and 0.40 for the pause and phone duration models respectively.

Moving on to looking in more detail at how the different punctuation marks contribute to the overall results (columns two-four) shows a large variation in how much each punctuation mark benefits from the use of prosodic information. Initially, identifying question marks in this particular test set was very unsuccessful. This might to some extent be due to lack of sufficient training examples. Comparing results for full stops

| | All punctuation marks | | | | Full stop | | | | Comma | | | | Question mark | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *P* | *R* | *P&R* | *SER* | *P* | *R* | *P&R* | *SER* | *P* | *R* | *P&R* | *SER* | *P* | *R* | *P&R* | *SER* |
| Ling. | 0.46 | 0.17 | 0.25 | 1.04 | 0.78 | 0.21 | 0.34 | 0.79 | 0.52 | 0.12 | 0.19 | 0.88 | 0 | 0 | - | - |
| + Pause Dur. | | | | | | | | | | | | | | | | |
| A, $\alpha = 1$ | 0.49 | 0.33 | 0.39 | 0.89 | 0.84 | 0.44 | 0.57 | 0.57 | 0.48 | 0.17 | 0.25 | 0.83 | 0 | 0 | - | - |
| A, $\alpha = 10$ | 0.39 | 0.45 | 0.42 | 1.05 | 0.78 | 0.63 | 0.70 | 0.41 | 0.39 | 0.19 | 0.25 | 0.81 | 0 | 0 | - | - |
| B, $\alpha = 0.01$ | 0.38 | 0.18 | 0.24 | 1.02 | 0.78 | 0.21 | 0.34 | 0.79 | 0.52 | 0.12 | 0.19 | 0.88 | 0 | 0 | - | - |
| B, $\alpha = 0.1$ | 0.42 | 0.17 | 0.24 | 1.00 | 0.79 | 0.21 | 0.33 | 0.79 | 0.54 | 0.11 | 0.19 | 0.89 | 0 | 0 | - | - |
| + Phone dur. | | | | | | | | | | | | | | | | |
| A, $\alpha = 0.1$ | 0.49 | 0.33 | 0.39 | 0.90 | 0.50 | 0.60 | 0.50 | 0.78 | 0.28 | 0.16 | 0.20 | 1.10 | 0 | 0 | - | - |
| A, $\alpha = 1$ | 0.50 | 0.33 | 0.40 | 0.89 | 0.60 | 0.44 | 0.51 | 0.77 | 0.17 | 0.29 | 0.21 | 1.07 | 0 | 0 | - | - |
| B, $\alpha = 0.01$ | 0.65 | 0.15 | 0.24 | 0.90 | 0.95 | 0.22 | 0.36 | 0.78 | 0.41 | 0.04 | 0.08 | 0.96 | 0 | 0 | - | - |
| B, $\alpha = 0.1$ | 0.67 | 0.14 | 0.23 | 0.90 | 0.94 | 0.21 | 0.34 | 0.79 | 0.40 | 0.04 | 0.07 | 0.96 | 0 | 0 | - | - |

Table 2: Results from automatic punctuating the test data using the **finite state model approach**. **A** and **B** refer to the two ways of combining the models, and $\alpha$ is the weight term used in the expressions. The first column of the table shows the results from a scoring of the hypothesised punctuated text, where all punctuation are taken into consideration. In columns two to four the results are split into individual results for the different punctuation marks. Figure 3 illustrates the results using the pause duration features.

to those for commas shows that the linguistic model does a little better in detecting the full stops than the commas. However, including prosodic information has a very different effect on the two types of punctuation marks. For the full stop features *SER* is almost halved from 0.79 (ling.) to 0.41 (ling+pause) and *P&R* doubled from 0.34 to 0.70. The phone duration features also increase performance, albeit slightly less drastically. Comparing this to the similar figures for the comma, only a very modest improvement is observed; from a *SER* of 0.88 to 0.81 and 0.83 for pause and phone duration features respectively and similarly for the *P&R* values. In particular for the pause duration features, it is not surprising that this feature does to a large extent only provide discriminant, complementary information for full stops (and presumably also question marks), whereas the position of commas would be far less accompanied by pauses in the acoustic stream. It is interesting to see that similar conclusions can be drawn for the phone duration features.

Table 2 presents a few results for different values of $\alpha$ for each type of decomposition. Figure 3 gives a more comprehensive graphical presentation of the test results involving the pause duration features. *P*, *R*, *P&R*, and *SER* values are plotted against the $\alpha$ weights and it is clear that the outcome of both decomposition **A** and **B** is highly sensitive to the weight given to each of the two model sources. Similar patterns were observed when using other prosodic features.

## 5. Performance of extracted features

One of the objectives of the current work has been to qualify to what extent a particular prosodic feature can predict an individual punctuation mark. Training MLPs in various configuration helped us pursue this, and MLPs have been employed in two different experimental setups.

Table 3 presents results for all of the pause, duration and pitch based features when tested using the **joint_MLP**, where each MLP is trained to distinguish between all four punctuation mark classes. Contrasting the different features shows a large difference in ability to discriminate. The pause duration features perform much better than any of the other features, many of which completely fail to recognise any punctuation marks. Comparing to the results reported above for the finite state approach, the pause duration based MLP performs reasonably with *SER* and *P&R* of 1.06 and 0.32 respectively compared to 1.04 and 0.25 for the linguistic model, but somewhat worse than the linguistic and pause duration models combined with *SER* = 0.89 and *P&R* = 0.42.

| | *P* | *R* | *P&R* | *SER* |
|---|---|---|---|---|
| pause dur. | 0.40 | 0.27 | 0.32 | 1.06 |
| vowels dur. | 0.00 | 0.00 | 0.00 | 1.00 |
| phone dur. | 0.00 | 0.00 | 0.00 | 1.00 |
| pitch - slope | 0.04 | 0.15 | 0.07 | 4.01 |
| pitch - range | 0.00 | 0.00 | 0.00 | 1.00 |
| pitch - mean | 0.00 | 0.00 | 0.00 | 1.00 |

Table 3: Single features results from classifying using the **joint_MLP** approach (figure 2*(a)*).

The **binary_MLP** systems can provide some more insight into the relationship between the individual prosodic features and the different punctuation marks. Binary decision type MLPs were trained for all the prosodic features, and in Table 4 representative results for two of the features are presented, pause duration and average phone duration[2]. MLPs are susceptible to uneven distribution of training examples in the training data material, which the current data suffer from (see table 1). To compensate for this, these results are therefore evaluated as scaled likelihoods (obtained by dividing the MLP estimated posterior probabilities through with the corresponding priors), and the results should therefore reflect a more differentiated picture.

The results confirm the conclusions derived from the results from both the finite state approach and the joint_MLP approach. When looking at the different types of prosodic information, the pause duration features are by far the strongest candidate for automatic punctuation. Comparing how well each punctuation mark is recognised, (the question marks aside for the previously mentioned reasons), the well performing pause duration feature does significantly better in detecting full stops than commas. Ignoring the prior information an *SER* of 1.22 and *P*, *R*, and *P&R* of 0.40, 0.41, and 0.40 respectively are obtained (not shown in table). These compare to the values previously obtained with

[2]A collection of the full results are available via ftp at `ftp://ftp.dcs.shef.ac.uk/share/spandh/SToBS/docs`.

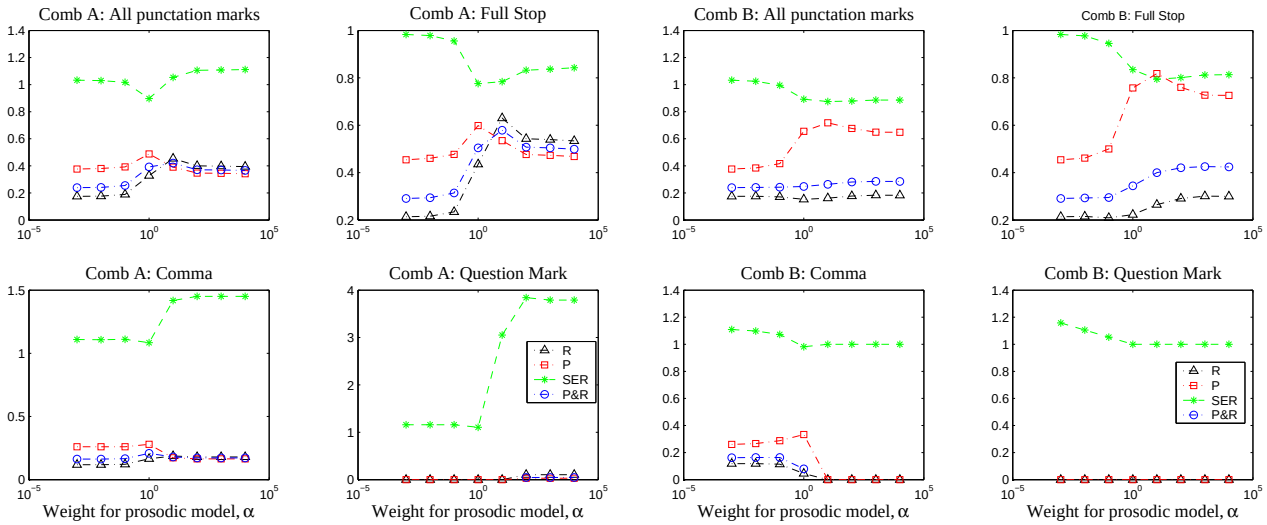


Figure 3: Results from automatic punctuating the test data using the **finite state model approach**, using linguistic and pause duration features. See table 2 for further details.

the other approaches, and confirms that using MLPs is indeed a valid means of gaining the knowledge about prosodic features and punctuation marks we are seeking.

| Pause duration features | | | | |
|---|---|---|---|---|
| | *P* | *R* | *P&R* | *SER* |
| Full stop | 0.26 | 1.00 | 0.41 | 2.85 |
| Comma | 0.11 | 0.73 | 0.19 | 6.04 |
| Question mark | 0.02 | 0.92 | 0.03 | 53.09 |
| Collective punc | 0.51 | 0.40 | 0.45 | 0.99 |

| Average phone duration features | | | | |
|---|---|---|---|---|
| | *P* | *R* | *P&R* | *SER* |
| Full stop | 0.08 | 0.48 | 0.13 | 6.45 |
| Comma | 0.06 | 0.20 | 0.09 | 4.14 |
| Question mark | 0.00 | 0.00 | 0.00 | 1.05 |
| Collective punc | 0.13 | 0.22 | 0.17 | 2.22 |

Table 4: Results from testing the pause duration features and the average phoneme duration features using the **binary_mlp** approach. The 'Collective punc' entry refers to an MLP trained on a data set, where the different punctuation marks are labelled as being part of a collective punctuation class, i.e. all punctuation marks are considered identical.

## 6. Discussion of evaluation methods

The way to evaluate automatic punctuation tasks is not obvious. In this paper, we have chosen to score all results using both the *SER* and the *P&R* measures. The *P&R* measure is a harmonic weight between the precision and recall scores. The *SER* score is computed as the total number of slot errors (missing, spurious and incorrect) divided by the total number of of slots in the reference, which is fixed for a given task. It can be compared to the word error rate score commonly used in speech recognition tasks. As mentioned earlier the different in the two scores lies in the fact that the *SER* score is a linear function of S, M and I, whereas the *P&R* score deweights spurious and missing errors, and Makhoul *et al.* argues that the error represented by *P&R* (i.e. $E = 1 - P\&R$) under-represents the total error.

For a task like automatic punctuation, the picture is even more complex. The task has clear similarities to NE recognition tasks, where both the extent and the identity of the NE is taken into consideration when scoring. For the current task, correct placement of the punctuation mark is crucial, but one can argue, that the scoring of the identity of the punctuation mark could be more gentle, than that of the *SER* measure used here. Substituting a comma for a full stop will still aid the structuring of the text to a great extent, and be better than having no punctuation marks at all. On the other hand, spurious question marks are likely to cause confusion.

Kim and Woodland [9] chose to count correctly placed but wrongly identified punctuation marks as half an error, resulting in an increased *P&R* (and a decreased *SER*[3]). Re-scoring our results along these lines changes the scores for the linguistic + pause duration **A**, $\alpha = 1$: The *P&R* increases from 0.392 to 0.422 and the *SER* decreases from 0.890 to 0.837. These results are similar to those in [9], although no direct comparison can be made since different test sets were used.

In general, the choice of evaluation metrics is very task dependent, and as for all classifier performance evaluations, it comes down to a trade-off between requirement to the True Positive rate and the False Positive rate, i.e. which area in the ROC space is acceptable.

## 7. Conclusions

In this paper, we have described approaches to automatic punctuation in transcriptions of broadcast news data produced by a large vocabulary speech recognition system. Statistical models were developed making fully use of the available data. One was based on linguistic information (based on sequences of words

[3]We propose to use a deweighted count for incorrect punctuations in the *SER* formulation as follows: $SER = (0.5I + M + S)/(C + I + M)$, i.e. keeping the denominator constant across the task as is fundamental to the *SER* score.

and punctuation marks as in a conventional language model) using the textual data from available transcripts of the news shows. The other group of statistical models were extracted from the acoustic data and were based on various extracted prosodic features. Features such as pause duration, phone duration, and pitch related values.

Two different approaches was used to obtain experimental results for automatic punctuating with full stops, commas and question marks on a subset of the Hub–4 collection of broadcast news data. Combining the linguistic model with a prosodic model significantly reduced the slot error rate and increased the related *P&R* measure. Particularly when using the pause duration model. It was shown how this increase in performance could be almost purely ascribed to the increased recognition rate for full stops, whereas the other punctuation marks were very little affected by the additional prosodic information.

The second approach made use of multi layered perceptrons to model the prosodic features. Various configurations made us conclude that there is a large difference in the usability of the different features for the current task, as well as a large variation in how much discriminant information the prosodic information carry.

Overall we have indicated that durational features may be used in a combined linguistic/prosodic model for punctuation annotation. The results of this exploratory work have resulted in a *P&R* of 0.42, although it is apparent that the recall/precision tradeoff is rather dependent on the model and interpolation scheme. It is notable that these results were obtained using a small amount of training data (less than a million words), compared with the corpora usually employed for language model estimation.

## 8. Acknowledgements

This word was funded by UK EPSRC grant GR/M36717, *Structured Transcription of Broadcast Speech.*

## 9. References

[1] R. Kompe, *Prosody in Speech Understanding Systems*, Springer-Verlag, 1996.

[2] E. Shriberg and A. Stolcke, "Prosody-based automatic segmentation of speech into sentences and topics," *Speech Communication*, vol. 32, no. 1-2, Sept. 2000.

[3] D. Hakkani-Tür, Gökhan Tür, A. Stolcke, and E. Shriberg, "Combining words and prosody for information extraction from speech," in *Proc. Eurospeech '99*, Budapest, Hungary, Sept. 1999.

[4] E. Nöth, A. Batliner, A. Kießling, R. Kompe, and H. Niemann, "Suprasegmental modelling," in *Computational Models of Speech Pattern Processing*, K. Ponting, Ed., pp. 182–199. Springer-Verlag, Berlin, 1999.

[5] V. Warnke, R. Kompe, H. Niemann, and E. Nöth, "Integrated dialog act segmentation and classification using prosodic features and language models," in *Proc. Eurospeech '97*, Rhodes, Greece, Sept. 1997.

[6] J. Hirschberg and C. Nakatani, "Acoustic indicators of topic segmentation," in *Proc. ICSLP '98*, Sydney, Australia, Dec. 1998.

[7] D. Beeferman, A. Berger, and J. Lafferty, "Cyperpunc: A lightweight punctuation annotation system for speech," in *Proc. ICASSP '98*, Seattle, WA, USA, May 1998.

[8] C. J. Chen, "Speech recognition with automatic punctuation," in *Proc. Eurospeech '99*, Budapest, Hungary, Sept. 1999.

[9] J-H Kim and P. C. Woodland, "The use of prosody in a combined system for punctuation generation and speech recognition," in *Proc. Eurospeech '01*, Aalborg, Denmark, Sept. 2001.

[10] Y. Gotoh and S. Renals, "Sentence boundary detection in broadcast speech transcripts," in *Proc. ASR-2000*, Paris, France, Sept. 2000.

[11] P. Taylor, R. Caley, W. A. Black, and S. King, "The Edinburgh speech tools library version 1.2.0," available from ftp://ftp.cstr.ed.ac.uk.

[12] A. Robinson, G. Cook, D. Ellis, E. Fosler-Lussier, S. Renals, and G. Williams, "Connectionist speech recognition of broadcast news," *Speech Communication*, To appear, Available from http://www.dcs.shef.ac.uk/∼sjr/pubs/2001/sprach01-preprint.html.

[13] C. J. van Rijsbergen, *Information Retrieval*, Butterworths, London, 2nd edition, 1979.

[14] J. Makhoul, F. Kubala, R. Schwartz, and R. Weischedel, "Performance measures for information extraction," in *Proc. of DARPA Broadcast News Workshop*, Herndon, VA, Feb. 1999.