Proceedings of the International Speech Communication Association (ISCA) Workshop: Automatic Speech Recognition: Challenges for the New Millennium (ASR-2000), pages –, Paris, September 2000

Sentence Boundary Detection in Broadcast Speech Transcripts

Yoshihiko Gotoh *Steve Renals*

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

ABSTRACT

This paper presents an approach to identifying sentence boundaries in broadcast speech transcripts. We describe finite state models that extract sentence boundary information statistically from text and audio sources. An n-gram language model is constructed from a collection of British English news broadcasts and scripts. An alternative model is estimated from pause duration information in speech recogniser outputs aligned with their programme script counterparts. Experimental results show that the pause duration model alone outperforms the language modelling approach and that, by combining these two models, it can be improved further and precision and recall scores of over 70% were attained for the task.

1. INTRODUCTION

Spoken audio data is a rich information source. Extensive research efforts during past decades have resulted in automatic speech transcription systems that can perform certain tasks (*e.g.*, large vocabulary dictation from a cooperative speaker) with a high degree of accuracy. However, it is the case that much more information may be extracted from the audio, for example sentence boundary information[1].

Conventional automatic speech recognisers, in most cases, rely on a language model component that contains a sentence marker, <s>. Such systems may be able to identify sentence boundaries to some extent, but their detection performance is generally not always satisfactory[2]. This reflects the emphasis in large vocabulary speech recognition research to obtain the correct sequence of words, with little concern for the overall structure.

In the area of natural language processing, sentence boundary disambiguation of textual data has been well investigated by fully utilising case, punctuation, and other structural information (*e.g.*, [6]), but it is a relatively new problem for spoken data. Stevenson and Gaizauskas [12] have discussed the difficulty in detecting such boundaries in the absence of case information. Shriberg *et al.* [11] investigated an approach to utilising prosodic features from speech, such as the pause duration at boundaries and the phone duration preceding a boundary. They argued that prosody was less susceptible to speech recognition errors and achieved good results using the Broadcast News and the Switchboard corpora. Hirschberg and Nakatani [4] also tested various acoustic indicators (*e.g.*, $f0$, voicing, energy component) for topic and phrase boundary identification.

The objective of this paper is to identify sentence boundaries in broadcast speech using statistical finite state models derived from news transcripts (textual data) and speech recogniser outputs (audio data). It is a step towards the production of structured speech transcriptions, which may include punctuation or content annotation. Recently trainable statistical models have been developed for extracting named entities from television and radio news broadcasts [2, 3, 7]. In this paper, such models are further extended to incorporate information from both textual and audio sources for the sentence boundary detection task.

The experiments reported in this paper have used the THISL collection of BBC news broadcasts and scripts (§ 2). An n-gram type language model is constructed from news transcriptions annotated with sentence break markers (§ 3). An alternative model may be estimated from pause information (*e.g.*, sentence and silence markers having durational information) in speech recogniser outputs aligned with their programme script counterparts (§ 4). By combining these two models from different sources (§ 5), experimental results show that over 70% of precision and recall score has been attained for sentence boundary detection in broadcast speech transcripts (§ 6).

2. EXPERIMENTAL DATA

The THISL data collection consists of a large amount of scripts, audio data, and some human generated reference transcriptions of a variety TV and radio news and current affairs programmes broadcast by the BBC since 1997 [8]. Table 1 presents some statistics of the data used in this study.

Programme scripts are not precise transcriptions of the broadcast audio, but are prepared texts preceding broadcast. There is an inevitable mismatch between the scripts and the actual broadcast, but there remains a strong similarity, and we use these pre-broadcast scripts as training material for constructing statistical models in the experiments. There ex-

[1] In spoken language, the natural unit seems to be the phrase, rather than the sentence, and phenomena such as disfluencies, corrections and repetitions are common. Throughout this paper, we refer ambiguously to such natural unit for segmenting speech as a 'sentence', and do not necessarily indicate a sentence in textual data.

[2] Using the same task as in the experiment, we have calculated that sentence boundary detection performance of our speech recogniser is slightly lower than 60% precision and recall.

	#shows	#words	#' ; '	#words per ' ; '
all programme scripts:				
BBC 1	1160	4 280 071	268 283	16.0
Radio 4	1050	4 748 831	234 456	20.3
total	2210	9 028 902	502 739	18.0
programme scripts with audio data counterpart:				
BBC 1	337	1 397 656	87 628	15.9
Radio 4	265	1 276 323	63 632	20.1
total	602	2 673 979	151 260	17.7

Table 1: Programme scripts in the THISL data collection consist of approximately nine million words from over two thousand news programmes. They cover *BBC 1* television news and *Radio 4* news broadcast between 1997 and 1999. The column, #' ; ', indicates the total number of sentence breaks in the scripts. An average sentence contains 16 words for television news and 20 words for radio news. Audio data was available for about 30% of the programme scripts.

	#shows	#words	#' ; '	#words per ' ; '
BBC 1	18	87 926	5064	17.4
Radio 4	14	70 054	3326	21.1
total	32	157 980	8390	18.8

Table 2: Reference transcriptions from the THISL data collection consist of slightly less than 160 000 words manually transcribed for 32 half-an-hour news programmes, broadcast between 1997 and 1999. They do not overlap with the news shows in table 1. The column, #' ; ', indicates the total number of sentence breaks in the transcriptions. An average sentence contained 17 words for television news and 21 words for radio news. Corresponding audio data counterpart exists for all reference transcriptions.

ist about 300 hours of audio data counterparts, which may also be transcribed using automatic speech recognisers.

Table 2 gives statistics of the human generated reference transcriptions in the THISL data collection. These transcriptions, totalling 160 thousand words (32 programmes), were obtained from careful hand transcription and included repeated and incorrect speech, or imperfect speech (*e.g.*, 'um', 'er'). These reference transcripts were used as the evaluation data set in our experiments.

A sample reference transcription and corresponding speech recogniser output are shown in appendix A.

3. TEXTUAL CLUES

Some textual clues may be used for identifying sentence boundaries in transcripts (either textual data or speech recogniser output). The following example is extracted from the BBC news reference transcription (table 6 in appendix):

> ... million pounds a year in the UK; And worldwide it's an industry worth several billion; But it's thought ...

Ignoring case information (as it does not exist in typical speech recogniser output), some words — such as 'and' and 'but' — often indicate the beginning of a new sentence[3].

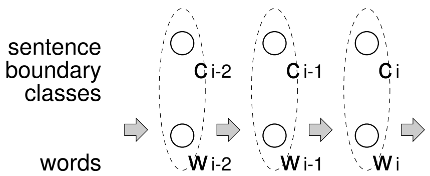

Figure 1: Topology for the sentence boundary language model. The arrows represent the evolution of the states, consisting of word and sentence boundary components, rather than explicit probabilistic dependences.

Generally, a sentence break is attached to the end of the last word of a sentence (*e.g.*, a semicolon, ' ; ', in the THISL data collection). We formulate the problem as the identification of the last word of each sentence, given a sequence of words. Each word in any text belongs either to a "last-word" class, or to a "not-last" class (denoted by <b> and ' • ', respectively). Given this notation, a corresponding class sequence for the above example is

... • • • • • • <b> • • • • • • • <b> • • • ...

In this section, we describe a finite state machine that models the joint probability of the current word and sentence boundary class conditioned on the previous words and classes (figure 1). Recently, similar architectures have been successfully applied to the named entity identification task from speech transcripts [2, 3, 7]. The formulation in this paper explicitly models constraints at the class level, compensating for the fundamental sparseness of n-gram tokens in the training material. It is mostly analogous to the formulation presented in [3], however the smoothing scheme needs to be modified to some extent. A bigram level model is presented in this paper, although it is straightforward to extend to higher level n-gram modelling. In the experiments reported here, we tested up to 5-gram models.

Sentence Boundary Language Modelling

Let $\mathcal{V}$ denote a vocabulary and $\mathcal{C}$ be a set of sentence boundary classes. We consider that $\mathcal{V}$ is similar to a vocabulary for conventional speech recognition systems (*i.e.*, typically containing tens of thousands of words, and no case information or other characteristics). $\mathcal{C}$ contains two classes[4], <b> and ' • ' as described earlier. Then we consider the joint probability of a sequence of words, $w_1^m = \{w_1, \ldots, w_m\}$, and corresponding class tokens, $c_1^m = \{c_1, \ldots, c_m\}$:

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

The superscript, $p^{[L]}$, indicates a language modelling probability in contrast to the prosody models presented in section 4. Once a sentence boundary language model (or,

[3]Note that the "sentence structure" of broadcast speech is rather different to that of written language.

[4]There is little reason that the formulation should be limited to the sentence boundary detection problem. It may be applied to a more general punctuation identification problem by arbitrarily selecting a set of punctuation classes for $\mathcal{C}$, and without any other alteration.

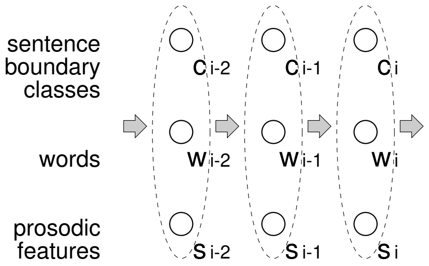

Figure 2: Extended finite state sentence boundary model. The model shown in figure 1 is augmented with prosodic features. Again the arrows do not represent the explicit probabilistic dependences.

simply a 'language model' when there is no confusion) is constructed, sentence boundaries can be identified by searching the Viterbi path such that

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

for a novel sequence of words, w_1^m.

Formulation (1) treats words and class tokens independently. Using bigram level constraints, (1) is reduced to

$$p^{[L]}(w_1^m, c_1^m) = \prod_{i=1\dots m} p^{[L]}(w_i, c_i | w_{i-1}, c_{i-1}) \ . \tag{3}$$

The right side of (3) may be decomposed as

$$\begin{aligned} &p^{[L]}(w_i, c_i | w_{i-1}, c_{i-1}) = \\ &\quad p^{[L]}(w_i | c_i, w_{i-1}, c_{i-1}) \cdot p^{[L]}(c_i | w_{i-1}, c_{i-1}) \ . \end{aligned} \tag{4}$$

In (4), the conditional current word and current class probabilities, $p^{[L]}(w_i|c_i, w_{i-1}, c_{i-1})$ and $p^{[L]}(c_i|w_{i-1}, c_{i-1})$, are in the same form as a conventional n-gram and may be estimated from annotated text data.

We used programme scripts, whose sentence breaks are annotated with semicolons, ' **;** ', as training material. The amount of text data obtainable from programme scripts is orders of magnitude smaller than that typically used to estimate n-gram models for large vocabulary speech recognition. Smoothing the maximum likelihood probability estimates is therefore essential to avoid zero probabilities for events that are not observed during the training. We have applied standard techniques in which more specific models are smoothed with progressively less specific models (see [3] for further details).

4. PAUSE DURATION

Pause information in speech recogniser outputs is less susceptible to speech recognition errors (indicated in [11], also observed in appendix A). In this section, we first consider a statistical model of prosody, then describe the specific case of pause duration features. The finite state machine of section 3 may be extended using a unigram probability model of prosodic features (figure 2).

token	occurrence (%)	' **;** ' (%)
<s>	1.79	1.06
<sil>	7.94	2.44
word	90.27	0.78
all	100.00	4.28

Table 3: The speech recogniser used in the experiment produced over three million tokens from about 300 hours of audio data. This table shows occurrences of pauses (<s> and <sil>) and word tokens, along with the occurrences of corresponding sentence breaks when aligned with programme scripts (column with ' **;** ').

Modelling Prosody

We assume that the current word, class and prosodic tokens are statistically independent from the previous tokens (although they are dependent on each other). This rather drastic assumption enables the construction of a unigram model of the joint probability for the sequences of prosodic features ($s_1^m = \{s_1, \dots, s_m\}$), words ($w_1^m$) and sentence boundary classes (c_1^m) :

$$p^{[P]}(s_1^m, w_1^m, c_1^m) = \prod_{i=1\dots m} p^{[P]}(s_i, w_i, c_i) \ . \tag{5}$$

We note that s_i is defined as any prosodic feature corresponding to a word w_i, and between w_i and w_{i+1}. The superscript, $p^{[P]}$, indicates a prosodic probability model.

Prosodic features may be extracted from audio data using signal processing algorithms or the acoustic model of a speech recogniser. Alignment with the corresponding transcription enables the identification of sentence boundary locations in the recognised word sequence. Because prosodic features are often strongly affected by the current word or the existence of a sentence boundary, a unigram level model may not be totally unwarranted.

There exist a few approaches to decomposing the joint probability, $p^{[P]}(s_i, w_i, c_i)$, in the right side of (5). For example, it may be decomposed as

$$p^{[P]}(s_i, w_i, c_i) = p^{[P]}(s_i|w_i, c_i) \cdot p^{[P]}(w_i|c_i) \cdot p^{[P]}(c_i) \ . \tag{6}$$

Each term in the right side of (6) may be estimated from speech recogniser outputs aligned with corresponding transcriptions. If necessary, discounting and smoothing schemes may also be applied.

Pause Duration Model

We have constructed a pause duration model within the prosody modelling framework. As noted earlier, outputs from our speech recogniser include sentence boundary and silence markers. Each duration is indicated in seconds, which is quantised to units of 0.1s. We make the approximation that the current word does affect the existence or duration of pauses[5]. Then, the right side of (5) is reduced to

$$p^{[P]}(s_i, w_i, c_i) \sim p^{[P]}(s_i, c_i) \ .$$

[5]This approximation is not true for many occasions; for example, some word followed by a pause, such as imperfect speech, 'um' or 'er', does not very often imply existence of a sentence boundary.

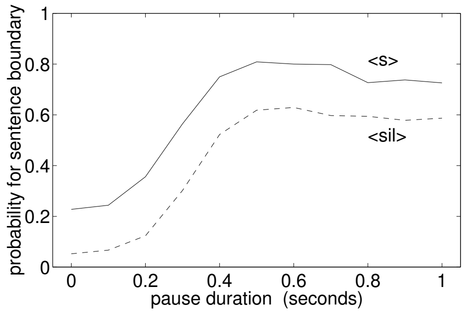

Figure 3: Plot of the probability that a pause in the speech recogniser output corresponds to a sentence break in the programme script. Pauses are marked by <s> and <sil>. The probability was estimated by aligning the speech recogniser output with the programme script, then counting the (mis)matches between sentence breaks (' **;** ') and pause markers.

Our speech recogniser produced over three million tokens from about 300 hours of audio data (*i.e.*, those pre-broadcast scripts with corresponding audio in table 1), of which nearly 10% were pauses (table 3). After alignment with recogniser outputs, it was found that less than a quarter of the sentence boundaries in the scripts were identified by sentence markers, <s>, and that many more sentence boundaries were hidden behind silence markers, <sil>. For less than 20% of sentence boundaries, there did not exist a matched pause marker. This seems to indicate that pauses (*i.e.*, both <s> and <sil>) are good predictors of sentence boundaries.

Finally, figure 3 shows $p^{[P]}(c_i|s_i)$, the probability that a pause marker in the speech recogniser output corresponds to a sentence break in the programme script. It is observed in the figure that the longer the pause duration, the more chance there exists a corresponding sentence break in the programme script.

5. MODEL COMBINATION

In this section, we discuss schemes for combining a language model, $p^{[L]}$, and a prosody model, $p^{[P]}$, for sentence boundary identification. This is not straightforward since the two models are estimated from different training material: $p^{[L]}$ is derived from text scripts and $p^{[P]}$ is estimated from aligned speech recogniser output.

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) :

$$p(s_1^m, w_1^m, c_1^m) = \prod_{i=1\ldots m} p(s_i, w_i, c_i | s_1^{i-1}, w_1^{i-1}, c_1^{i-1}) \ . \quad (7)$$

Using this combined model, sentence boundaries can be identified by

$$<\hat{c}_1^m> = \underset{c_1^m}{\operatorname{argmax}} \ p(s_i^m, w_1^m, c_1^m) \quad (8)$$

given sequences of word and prosody tokens (w_1^m and s_1^m).

We assume that previous prosodic features do not affect the current word, class and prosody. Further, we assume that the current prosody is independent from previous word and class tokens. With these independence assumptions, and using bigram level constraints, the joint probability (7) is reduced to

$$p(s_1^m, w_1^m, c_1^m) = \prod_{i=1\ldots m} p(s_i, w_i, c_i | w_{i-1}, c_{i-1}) \ , \quad (9)$$

although it is straightforward to extend to higher order n-grams.

In this paper, we restrict our attention to a pause duration model. The right side of (9) may be decomposed into the language model component, $p^{[L]}$, and the pause duration model component, $p^{[P]}$, in two different ways. Decomposition **A** is mathematically motivated. Decomposition **B** is heuristic and works particularly well for pause duration.

Decomposition A

The first approach assumes that a prosodic feature is independent of previous words and classes, given the current word and class, applying the following conditional probability relation

$$p(s_i, w_i, c_i | w_{i-1}, c_{i-1}) \sim \left\{ p^{[P]}(s_i | w_i, c_i) \right\} \times \left\{ p^{[L]}(w_i, c_i | w_{i-1}, c_{i-1}) \right\} \ . \quad (10)$$

When combining models, $p^{[L]}$ and $p^{[P]}$ from different sources, the probability estimate may be distorted. In order to compensate for this, the pause duration model may be factored using some real number a :

$$p(s_i, w_i, c_i | w_{i-1}, c_{i-1}) \sim \left\{ p^{[P]}(s_i | w_i, c_i) \right\}^a \times \left\{ p^{[L]}(w_i, c_i | w_{i-1}, c_{i-1}) \right\} \ . \quad (11)$$

To some extent, this formulation is similar to the language model match factor widely used in large vocabulary speech recognition. In (11), the pause duration model may be obtained using:

$$p^{[P]}(s_i | w_i, c_i) = \frac{p^{[P]}(s_i, w_i, c_i)}{p^{[P]}(w_i, c_i)} \ . \quad (12)$$

Decomposition B

An alternative approach is heuristically derived, decomposing the right side of (9) as

$$p(s_i, w_i, c_i | w_{i-1}, c_{i-1}) \sim \left\{ p^{[P]}(w_i, c_i | s_i) \right\}^a \times \left\{ p^{[L]}(w_i, c_i | w_{i-1}, c_{i-1}) \right\} \ . \quad (13)$$

In (13), the pause duration model is estimated by

$$p^{[P]}(w_i, c_i | s_i) = \frac{p^{[P]}(s_i, w_i, c_i)}{p^{[P]}(s_i)} \ . \quad (14)$$

It may be observed in figure 3 that, for both sentence and silence markers, the probability that corresponding sentence break exists is not very high when pause duration is short. Such unlikely cases are effectively filtered out by a large enough a because $p^{[P]}(w_i, c_i | s_i) < 1$ in decomposition (13). It also compensates the distortion problem described earlier. When the pause duration is longer (say, 0.4 seconds and over), $p^{[P]}$ is flatter and $p^{[L]}$ becomes a dominant factor for making a decision.

	#shows	*WER* (%)
BBC 1	18	32.0
Radio 4	14	19.2
total	32	26.3

Table 4: Summary of word error rates for the evaluation data (table 2).

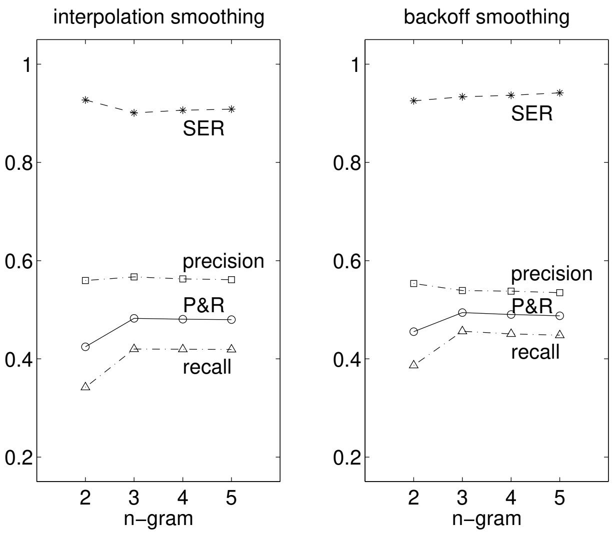

Figure 4: Sentence boundary detection performance for reference transcriptions (after conversion to single case text, and removal of punctuation). N-gram level models ($n = 2, 3, 4, 5$) are compared for both interpolation (left) and backoff (right) smoothing schemes.

6. EXPERIMENTS

Sentence boundary detection performance was evaluated using an unseen set of evaluation data (table 2). 16 hours of audio data for 32 half-an-hour news shows was processed using the ABBOT large vocabulary speech recognition system [10] and the CHRONOS decoder [9]. Table 4 summarises the *WER* (word error rate) for the evaluation data.

As noted in section 4, we have simplified the problem so that the current word does not affect the existence or duration of pauses. Thus, instead of pause duration models (12) and (14), approximations, $p^{[P]}(s_i|w_i, c_i) \sim p^{[P]}(s_i|c_i)$ and $p^{[P]}(w_i, c_i|s_i) \sim p^{[P]}(c_i|s_i)$, have been used in the experiments. In this section, after a brief summary of the performance measures we have employed, experimental results using language models, pause duration models, and combined models are described.

Performance Measures

In the experiments, hypothesised sentence boundaries in a speech recogniser output were compared with those annotated in a reference transcription. In order to measure the performance, we adopt four types of measures from the recent named entity evaluation organised by NIST [1].

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

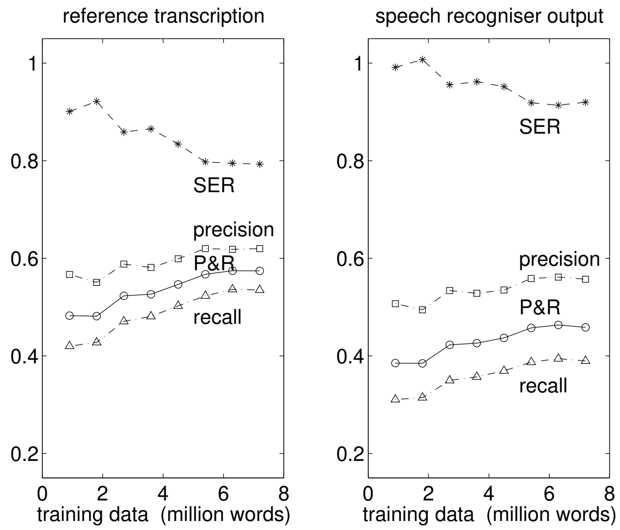

Figure 5: Sentence boundary detection performance improves as the amount of training data increases. For reference transcriptions (left — single case, no punctuation) and speech recogniser outputs (right), trigram level models were derived from the amount of training data varying between 0.9 to 7.2 million words of BBC programme scripts, then smoothed with interpolation.

statistic:

$$P\&R = \frac{2RP}{R + P} \ .$$

Although recall and precision are useful and informative measures, Makhoul *et al.* [5] 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)[6].

Sentence Boundary Language Models

In the first experiment, language models were estimated using over 0.9 million words of BBC programme scripts. Figure 4 shows the sentence boundary detection performance for reference transcriptions. They were initially converted to single case text and all punctuation was removed. The accuracy of sentence boundary identification performance improved when a trigram level model was used instead of a bigram. However, there was hardly any improvement using n-grams higher than 3. Further, the two smoothing schemes (interpolation and backoff) achieved a very similar level of performance.

A second experiment investigated the relationship between detection accuracy and the size of the training set. Figure 5 shows how sentence boundary identification varies as the training data varies between 0.9 to 7.2 million words of programme scripts. Not surprisingly, sentence boundaries

[6] *SER* is analogous to *WER*. It is obtained by $SER = (I + M + S)/(C + I + M)$ 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 as $R = C/(C+I+M)$ 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.

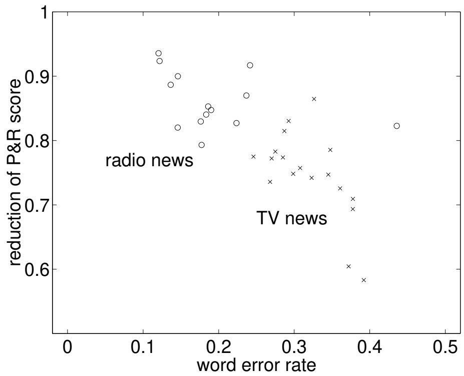

Figure 6: Precision and recall of sentence boundary detection with respect to *WER*. The graph shows how *P&R* changes when moving from reference transcriptions (single case, no punctuation) to speech recognition output with varying *WER*.

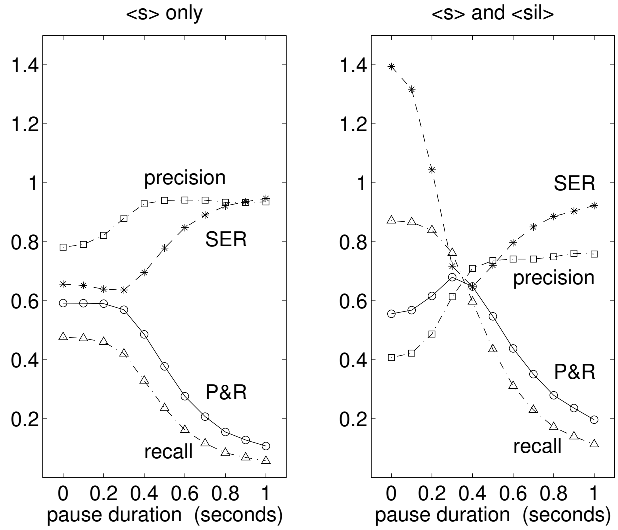

Figure 7: Sentence and silence markers (<s> and <sil>) and their duration in speech recogniser outputs are good indicators for detecting sentence boundaries. These graphs show the performance when a certain pause duration threshold is selected for <s> only case (left) and for using both <s> and <sil> (right). For example, a pause duration threshold of 0.2 seconds indicates that markers having pause duration less than 0.2 seconds are rejected.

from recogniser outputs were identified with less accuracy than those from their reference transcription counterparts.

We note that the n-gram models were estimated from scripts, but were applied to speech recogniser output. Figure 6 shows how detection accuracy (with respect to the reference transcription) for the speech recognition output varies with *WER*. As expected, *P&R* declines approximately linearly with *WER*.

As a side note, it is interesting that *BBC 1* television news and *Radio 4* news shows were clustered in different areas of the figure. This is probably because the former contained some broadcasts with lower audio quality (*e.g.*, outside broadcasts), while radio programmes feature a larger proportion of material broadcast from the studio.

Pause Duration Model

The pause duration model was estimated using speech recogniser outputs from 300 hours of audio data, aligned with their programme script counterparts. Figure 7 shows that pause markers (<s> and <sil>) and their duration in speech recogniser outputs are good predictors of sentence boundaries. When using <s> only, the recall rate was relatively low although boundaries were detected with high precision. By using both sentence and silence markers and with a good choice of pause duration threshold (around 0.4 seconds in this task), it achieved a higher *P&R* (and lower *SER*) than <s> alone. Although a simple method, this model achieved substantially better accuracy than the language modelling approach described earlier.

Pause duration models cannot be applied to the reference transcriptions because they do not contain pause markers.

Combined Models

Finally, we describe experiments using both the sentence boundary language model and the pause duration model. Figure 8 shows that each decomposition performed differently as the weight for the pause duration model varied. Table 5 summarises the performance for speech recogniser output.

Perhaps decomposition **A** is the more straightforward method for combining two models, but the result is less easy to evaluate. Compared with the pause model alone, *P&R* was improved from 65% to 68%; however the *SER* was worse than the pause duration model alone.

	R	*P*	*P&R*	*SER*
language model	.39	.56	.46	.92
pause duration model	.58	.74	.65	.63
combined models:				
decomposition **A**	.71	.66	.68	.65
B	.62	.80	.70	.54

Table 5: Sentence boundary detection performance for speech recogniser output. The language model and the pause duration model were estimated as in figure 8. For combined models, numbers were extracted for $a = 1.5$ (decomposition **A**) and $a = 20$ (**B**).

The construction of decomposition **B** was heuristic, but it resulted in the better model combination approach. By choosing a sufficiently large a (say, 10 or greater in this case), it consistently outperformed the pause duration model. In particular when $a = 20$, the *P&R* was over 70% and the *SER* fell to 54%.

7. SUMMARY

In this paper, we have described approaches to identifying sentence boundaries in transcriptions of broadcast speech produced by a large vocabulary speech recognition system. Two statistical models were developed: one was an n-gram type language model derived from programme scripts (textual data), and the other was a pause duration model estimated from recogniser outputs (audio data) aligned with their programme script counterparts. Experimental results indicated that the pause duration model alone outperformed the language modelling approach, and that it could be improved

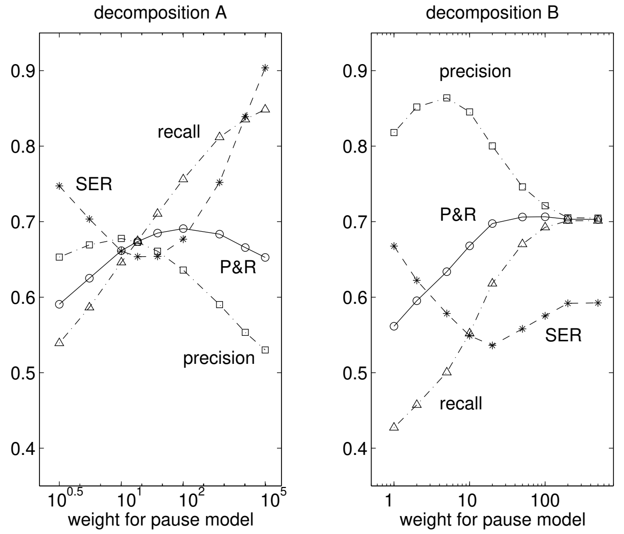

Figure 8: Sentence boundary detection accuracy for the combined models, using both decompositions. The language model was a trigram estimated from 5.4 million words of text scripts and smoothed using interpolation. The pause duration model was based on 300 hours speech recogniser output aligned with their scripts.

further by appropriately combining those two models.

Several approximations have been made in the models presented here. In particular, the pause duration model is not conditioned on the current word; this condition may be removed in future work. The only prosodic feature that we have used is pause duration; it will be interesting to investigate the use of other prosodic features, such as phone duration [11], in this framework.

ACKNOWLEDGEMENTS

We have benefited greatly from cooperation and discussions with Dave Abberley, Mark Stevenson, and Robert Gaizauskas. This work was funded by UK EPSRC grant GR/M36717, *Structured Transcription of Broadcast Speech*. Development of ABBOT speech recognition system and collection of the BBC data was funded by ESPRIT Long Term Research Project THISL (EP23495).

REFERENCES

[1] Defense Advanced Research Projects Agency. Proceedings of DARPA broadcast news workshop. Herndon, VA, February 1999.

[2] Daniel M. Bikel, Scott Miller, Richard Schwartz, and Ralph Weischedel. Nymble: a high-performance learning name-finder. In *Proceedings of the 5th ANLP*, pages 194–201, Washington, DC, April 1997.

[3] Yoshihiko Gotoh and Steve Renals. Information extraction from broadcast news. *Philosophical Transactions of the Royal Society, series A*, 358:1295–1310, April 2000.

[4] Julia Hirschberg and Christine H. Nakatani. Acoustic indicators of topic segmentation. In *Proceedings of ICSLP-98*, volume 4, pages 1255–1258, Sydney, November 1998.

[5] John Makhoul, Francis Kubala, Richard Schwartz, and Ralph Weischedel. Performance measures for information extraction. In *Proceedings of DARPA Broadcast News Workshop*, pages 249–252, Herndon, VA, February 1999.

[6] Andrei Mikheev. Feature lattices for maximum entropy modelling. In *Proceedings of COLING/ACL-98*, pages 848–854, Montreal, August 1998.

[7] David D. Palmer, Mari Ostendorf, and John D. Burger. Robust information extraction from automatically generated speech transcriptions. *Speech Communication*, 32(1/2), September 2000.

[8] S. Renals, D. Abberley, D. Kirby, and T. Robinson. Indexing and retrieval of broadcast news. *Speech Communication*, 32(1/2), September 2000.

[9] Tony Robinson and James Christie. Time-first search for large vocabulary speech recognition. In *Proceedings of ICASSP-98*, volume 2, pages 829–832, Seattle, May 1998.

[10] Tony Robinson, Mike Hochberg, and Steve Renals. The use of recurrent networks in continuous speech recognition. In C. H. Lee, K. K. Paliwal, and F. K. Soong, editors, *Automatic Speech and Speaker Recognition — Advanced Topics*, chapter 10, pages 233–258. Kluwer Academic Publishers, 1996.

[11] Elizabeth Shriberg, Andreas Stolcke, Dilek Hakkani Tür, and Gükhan Tür. Prosody modeling for automatic sentence and topic segmentation from speech. *Speech Communication*, 32(1/2), September 2000.

[12] Mark Stevenson and Robert Gaizauskas. Experiments on sentence boundary detection. In *Proceedings of the 6th ANLP*, Seattle, April 2000.

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

A. SAMPLE BBC DATA

Table 6 shows the reference transcription and the corresponding speech recogniser output excerpted from a 30 minute television news broadcast. For the reference transcription, mixed case text was used and sentence breaks and speaker change information were inserted by transcribers. The speech recogniser output (described in section 6) produced sentence markers (obtained by treating sentence boundaries as a word with an acoustic realisation as a pause) and other periods of silence identified by the acoustic model.

This particular sample was a report from a local supermarket scene, and thus the quality of audio was much worse than the other part broadcasted from inside the television studio. As a consequence, the speech recogniser output contained more than the average number of errors; its overall *WER* was nearly 60%, contrasting with the average *WER* of below 27% for the whole programme. In particular, the seventh sentence of the first speaker (reporter), "He played stereotypical German music . . . and French wines;" was overlapped with loud music, which the speech recogniser could not handle properly. Further, all sentences by the second speaker (professor) was affected by the relatively high background sound in the local supermarket.

Despite this less than average performance by the speech recogniser, there is a good correlation between pause markers in the recogniser output and sentence breaks in the reference transcription. This supports the observation by Shriberg *et al.* [11] that prosody is not strongly affected by speech recognition errors.

	reference transcription:	**speech recogniser output:**
	<speaker: Nicola Carslaw>	
•	This in store radio station provides music for Asda;	davies install radio stations <sil:0.128> provide music has said <sil:0.400>
•	It's a handy way of advertising and they clearly believe it lifts shoppers' spirits;	it had where advertising <sil:0.272> and deeply believe eclipse up as britain's <sil:0.720>
•	Royalties payments alone for instore music are about fifty five million pounds a year in the UK;	most payments alone be in store music about fifty five million pounds a year in the uk <sil:0.400>
•	And worldwide it's an industry worth several billion;	or worldwide for an industry worth seven billion <sil:0.768>
•	But it's thought no one had independently tested whether it affects how shoppers spend their money;	but it's thought no i think independently tested whether it affects how shoppers spend their money <sil:0.384>
•	Until this Leicester University psychologist carried out research here at his local supermarket;	into this leicester university psychologist carried out research here at his local supermarket <sil:0.192>
•	He played stereotypical German music and on alternate days over a fortnight traditional sounding French music next to the display of German and French wines;	them home newsroom of the new labour stereotypical german using adult ten days over a fortnight <sil:0.224> tradition of sounding french to the communication lines <sil:0.944>
	<speaker: Dr Adrian North>	
•	Well we find that when you played French music then French wine outsold German wine by about five bottles to one;	or five week life issues and french wine outsold german wine by five officers were on <sil:0.272>
•	Whereas when we played German music then German wine outsold French wine by about two bottles to one;	was repaid german music <s:0.224> the german whitehouse of french wine by back to top this was <sil:0.384>
•	The first implication is simply it shows for the first time that music in supermarkets does actually work;	the first indication simply shows the first time that we can see what is this actually work <sil:1.136>

Table 6: These excerpts are the reference transcription and the corresponding speech recogniser output from the *BBC Nine O'Clock News* (a half-an-hour programme broadcast on 12th of November 1997 on *BBC 1* television). Sentence by sentence alignment (shown by bullets ' • ') was done manually by authors. For the reference transcription, mixed case text was used and semicolons (' **;** '— indicating sentence breaks) were inserted by transcribers (left). It also contains speaker change information marked with <speaker: $name$> where '$name$' indicates a speaker name. On the other hand, this speech recogniser produced a single case sequence of words and occasional pauses, but no punctuation (right). Pauses include sentence boundary and silence markers and are shown as <s:$d.ddd$> and <sil:$d.ddd$> where '$d.ddd$' indicates pause duration in seconds. This particular excerpt (*WER* of 59.7%) contains a greater than average number of errors. (For comparison, *WER* for the whole half-an-hour news show is 26.8%.)