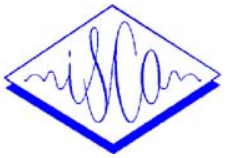

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

SPEECH RECOGNITION WITH AUTOMATIC PUNCTUATION

C. Julian Chen
IBM Thomas J. Watson Research Center, PO Box 218, Yorktown Heights, NY 10598, USA
email: juchen@us.ibm.com

10.21437/Eurospeech.1999-115

ABSTRACT

We present a method of speech recognition with automatic punctuation based on a combination of acoustic and lexical evidence. In the recognizer vocabulary, punctuation marks are treated as word entries. By assigning the acoustic baseforms of silence, breath, and other non-speech sounds to punctuation marks, and using a properly processed N-gram language model, unpronounced punctuation marks of various types (commas, periods, etc.) appear naturally in the recognizer output. This technology can be used in dictation systems to improve usability, in commercial broadcast transcription systems to reduce editing time, and in information retrieval systems to provide phrasing information to facilitate natural language understanding.

1. INTRODUCTION

Punctuation is an indispensable element of modern writing. In current speech recognition systems, in order to have punctuation marks appear in the transcribed text, each one must be pronounced by name, such as PERIOD, COMMA, QUESTION MARK, etc. However, in natural speech, punctuation marks are usually not pronounced. To transcribe natural speech into an orthographically adequate text, a method of automatically inserting punctuation marks in the transcribed text is essential. The practical need for automatic punctuation is evidenced in the following situations:

1. When using dictation systems for spontaneous speech recognition, or free-hand composing, punctuation is often not originally in the thought. It is more natural to dictate the idea first, then add punctuation later. In this case, if the dictation system can punctuate automatically, even with limited accuracy, the result is useful in the subsequent editing stage.

2. Automatic transcription of radio and television broadcast, or public speech. Pronounced punctuation is not allowed in such types of speech. Automatic punctuation, even if it is not very accurate, will greatly improve the readability of the transcription.

3. In natural language systems, such as information retrieval, automatic banking, or travel service systems, the phrasing or punctuating of the transcriptions of the input speech will facilitate natural language understanding.

In this paper, we present a method of speech recognition with automatic punctuation, using the phonetic characteristics of punctuation marks, and properly processed lexical information.

2. AUTOMATIC PUNCTUATION USING THE LANGUAGE MODEL

Several authors have attempted automatic linguistic segmentation or automatic punctuation by post-processing the unpunctuated recognized text [2, 5]. Stolcke and Shriberg [5] report results of experiments on linguistic segmentation of conversational speech using N-gram language models. Good progress has been reported in the prediction of segmentation boundaries from the dictated text. Beeferman, Berger, and Lafferty [2] developed an annotation system to insert intra-sentence punctuation from a unpunctuated text; in their study, the sentence boundaries were predetermined. Only one type of punctuation mark is predicted, the comma. Using a trigram language model and a straightforward application of the Viterbi algorithm, the results are promising.

Although the lexical structure of the unpunctuated text provides certain information about punctuation, in most cases it is not sufficient. For example, the following unpunctuated sentence

```
Woman without her man is nothing
```

can have different meanings according to punctuation:

```
Woman!  Without her, man is nothing.
```

Or:

```
Woman without her man, is nothing.
```

In the above case, the unpunctuated sentence does not provide a clue to punctuation. However, by reading

the punctuated sentences aloud, the linguistic segmentation becomes unambiguous, because prosody (pauses and intonation) provides the clue. This example clearly indicates that acoustic information is essential for the automatic prediction of punctuation. As we will show, although acoustic segmentations do not correspond one-to-one with linguistic segmentations, the combination of acoustic and lexical information permits an adequate automatic prediction of punctuation marks.

3. CORRELATION BETWEEN PUNCTUATION AND PAUSES

The possibility of using pauses to predict punctuation marks has been discussed previously [2, 5]. The difficulty is that pauses of various lengths often do not directly correlate to the positions and types of punctuation marks. To date, the acoustic information has not been used to predict punctuation marks.

The following statistics, however, indicate that the pauses are closely related to punctuation, and the information provided by pauses is vital. An experiment was conducted by asking three speakers to read a 330-word business letter, each containing 31 punctuation marks. The total number of punctuation marks is 93. The speech was then decoded with an IBM speech recognition system. The position of each pause, totally 115, was identified. The positions of the pauses were compared with the positions of punctuation marks. Three cases were identified, as shown in Table 1:

Item	Count	%
Number of pauses	115	123.6
Pauses related to punctuation	87	93.5
Pauses unrelated to punctuation	28	30.1
Punctuation marks without pause	6	6.5

Table 1. Punctuation and Pauses

The last column is the percentage with respect to the total number of punctuation marks, 93. As the data demonstrate, there is not a one-to-one correspondence between pauses and punctuation marks. However, the number of pauses is only about 12% of the number of words. In the pure language-model approach to automatic punctuating, every space between two words is a candidate for a punctuation mark. By restricting the insertion of punctuation marks to pauses, we can narrow down the search space by one order of magnitude. Of course, this restriction will cause us to miss inserting punctuation at some places where it may belong. Nevertheless, the percentage of punctuation marks without a pause, 6.5%, is small.

4. CORRELATION BETWEEN PUNCTUATION AND DISFLUENCIES

In addition to pauses, other acoustic phenomena are also closely related to punctuation marks. The most important one is breath-sound. Lipsmacks, to a much lesser extent, are also related to punctuation. This is evidenced in Table 2, which lists statistics gathered from broadcast speech of 12 announcers:

Item	Count	%
Number of words	178394	
Number of punctuation marks	16644	
Punctuation marks with breath	5686	34.2
Punctuation marks with lip-smack	518	3.1
Standalone breath	4926	
Standalone lip-smack	700	
Punctuation without disfluencies	10440	62.7

Table 2. Punctuation and Disfluencies

The number in the last column refers to the percentage of punctuation marks related to various acoustic phenomena with respect to the total number of punctuation marks. As shown, about one third of punctuation marks are associated with breaths, and about one half of breaths are associated with punctuation marks.

5. THE METHOD

Since both acoustic and lexical information are available during the decoding process, automatic punctuating can be executed during decoding. The method is as follows:

5.1. Recasting the Acoustic Model

The first step is to restrict the search space of punctuation marks to places with acoustic evidence of punctuation. This is done by assigning acoustic baseforms to punctuation marks (including comma, colon, semicolon, period, question mark, and exclamation mark) as a silence, a double silence, or a disfluency (such as breath or lipsmack). In detail, in the phonetic baseform list, we add some of the following lines (X denotes silence):

Word	Baseform
,	X
,	X X
,	X X X
,	X HH X
,	X F X
,	X K X
,	X P X
,	X UM X

Table 3. Baseforms of Comma

Similar lines are added for other punctuation marks, such as colon, semicolon, period, question mark, and exclamation mark.

By using a different list of acoustic baseforms for punctuation marks, the minimum lengths of the pauses to be identified as candidates of punctuation marks can be adjusted. In the current recognition system, each frame is set to be 10 msec, and each phoneme has three HMM states. Thus, the minimal time duration of one phoneme is 30 msec. By excluding the single silence from the list of acoustic baseforms, the minimal time duration for a pause to be a candidate of a punctuation mark is 60 msec. By doing so, some of the punctuation marks will be missing, but the number of false insertions will be lower. This statement is verified by experiments, as shown below.

5.2. Recasting the Language Model

During decoding, the most likely word series $\widehat{W}$ is obtained by maximizing the combined probability,

$$\widehat{W} = argmax_W Prob(Y|W)Prob(W),$$

where $W = w_1 w_2 ... w_n$ is the sequence of words, $Prob(Y|W)$ is the conditional probability from acoustic models, and the a priori probability $Prob(W)$ is the language model score. Therefore, once a pause or a non-speech sound in the baseform list of punctuation marks is detected by the acoustic unit of the decoder, the search process looks for the score of the language model. If the language model score does not favor a punctuation mark, then the silence is interpreted as a silence. In other words, there is no output. If the language model score favors a punctuation mark, a punctuation mark will be returned. In deed, as shown in Figure 1, many silences are decoded as silences, not punctuation marks.

However, the commonly used method of building the language model has a major deficiency which is not suitable for predicting the type of punctuation marks. It is a common practice that in the counting process of building a language model, the text is separated into sentences. Each full stop (or terminal punctuation, including period, question mark, and exclamation mark) is followed by a BOUNDARY-WORD. After the BOUNDARY-WORD, the counting process starts again. Therefore, each terminal punctuation should be followed by a BOUNDARY-WORD, and unrelated to the words in the next sentence. If the speech to be decoded is spoken continuously across the sentence boundary without signifying an end of the sentence, the terminal punctuation marks (periods, question marks, and exclamation marks) would never appear, because it always expects an end of the record, or a BOUNDARY-WORD.

To recast the language model to predict sentence-ending punctuation marks, the text corpus for counting is rebuilt by combining sentences into paragraphs, removing BOUNDARY-WORD between sentences in a paragraph, but retains terminal punctuation. Paragraphs are still terminated by BOUNDARY-WORD. The length of the paragraphs (not necessary the natural paragraphs) is typically 200 to 500 words.

The effect of recasting the language model into paragraphed format is demonstrated by direct experiments, as shown in the following section.

6. EXPERIMENTAL RESULTS

The decoding system is the IBM Research experimental continuous speech recognition system. The acoustic database is trained on 1,800 speakers. Speaker adaptation is implemented. The language model is built for IBM's legal dictation system, from 250 million words. The test text is a business letter with 333 words, including 31 punctuation marks. Three speakers read the letter, without verbalizing the punctuation marks.

Two methods of building the language models are tried:

a, Using paragraphed text for counting. The length of the paragraph is about 200 to 500 words.

b, Using the commonly applied method, where each sentence is a unit for counting.

Two sets of acoustic baseforms are tested:

a, The baseforms of punctuation marks include the single silence, the double and triple silence, and various disfluencies.

b, The baseforms of punctuation marks include the double and triple silence, and various disfluencies, but excluding the single silence.

The results are summarized in Table 4.

%	A	B	C	D
Place OK, type OK	57.0	48.4	37.6	25.8
Place OK, type bad	25.8	22.5	40.9	35.5
Place OK, total	82.8	70.9	78.5	71.3
Place bad	31.2	18.3	26.9	13.9

Table 4. Summary of Results

The notations are as follows:

Place OK, type OK: Punctuation mark at correct place and of correct type.

Place OK, type bad: Punctuation mark at correct place but of wrong type.

Place OK, total: Total number of punctuation marks at correct places.

Place bad: Extra punctuation marks, all commas.

Case A: Paragraphed text, single silence included in the baseform list. This case gives the

Dear Barbara: We finally received in the mail on Friday the enclosed letter from Cohen's counsel setting forth the people who she considered to be the comparisons for her equal pay act claim. She also rejects our settlement proposal. Here is our plan of action. We proceed, as I mentioned to you last week, with moving to strike or dismiss as many allegations of her complaint as possible in order to limit the issues, and answer the remaining allegations. Michael is in the process of drafting this pleading.

Dear Barbara. We finally received in the mail on Friday the enclosed letter * from Colon's counsel setting forth the * people who she can * consider to be the * comparisons for her equal pay * act claim. See also rejects our settlement proposal. Here is that plan of action. We proceeded * as I mentioned to you last week, without moving to strike or dismiss * as many allegations of * a complaint as possible. In order to limit the issues, any answer the remaining allegations, Michael is in the process of drafting this pleading.

Figure 1. Example of Operation. Left: original text. Right: as decoded, with automatically inserted punctuation. The * marks show pauses present in the audio that did not lead to insertion of punctuation marks. Capitalization of words after a period is added for easy reading.

best result for predicting punctuation marks originally in the text. The number of extra punctuation marks (all commas) is high. However, many of the extra commas are acceptable and possibly also useful, if the decoded text is used as a starting point of a subsequent editing process.

Case B: Paragraphed text, single silence excluded from the baseform list. The exclusion of the single silence from the baseform of the punctuation mark sets the minimal silence length requirement for a punctuation mark to 60 milliseconds. This reduces the number of extra commas. However, many punctuation marks that should appear are missing.

Case C: Sentenced text, single silence included in the baseform list. The number of punctuation marks with correct type is markedly reduced.

Case D: Sentenced text, single silence excluded from the baseform list. The error rate becomes markedly higher by using the sentenced text corpus for counting. Many periods now become commas.

Figure 1 shows an example of the original text and decoded text. It uses the arrangement of Case B. As shown, many pauses (marked as *) are not decoded as punctuation marks. The lexical information made the right selection in most cases.

7. DISCUSSION

Historically, punctuation emerged much later than the alphabetic writing system, which was considered exclusively as a record of spoken words [3]. It was not until the sixth century that punctuation was introduced in the writing of religious documents to record the pauses and intonation in speech. The modern system of punctuation was started in the Renaissance period. Its purpose is not solely to transcribe pauses and intonation; it also serves to disambiguate and clarify meaning. The use of punctuation has evolved remarkably in the last century. The contemporary use of punctuation is documented in standard style manuals, such as "The Chicago Manual of Style" [1], or dedicated handbooks [4]. However, the style of punctuation varies from author to author, and from field to field. Punctuation is also used to alter the meanings of the text. Thus, absolute accuracy of punctuation regarding a given unpunctuated text is not a totally fair measure. The numbers presented in this paper are for reference only.

8. CONCLUSIONS

We have described and tested a speech recognition system with automatic punctuation, based on the combination of acoustic and lexical evidence. This technique is useful for dictation products, broadcast transcription, and information systems to facilitate natural language understanding.

9. ACKNOWLEDGMENTS

The author appreciates inspiring discussions with Harry Printz, Michael Picheny, and David Nahamoo.

REFERENCES

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

[2] D. Beeferman, A. Berger, and J. Lafferty. "Cyberpunc: A Lightweight Punctuation Annotation System for Speech". *Proceedings ICASSP 1998*, 2:693–696.

[3] M. P. Parkes. *Pause and Effect: Punctuation in the West.* University of California Press, Berkeley, Los Angeles, 1993.

[4] Harry Shaw. *Punctuate it Right!* Harper-Collins, Dunmore, PA, 1993.

[5] A. Stolcke and E. Shriberg. "Automatic Linguistic Segmentation of Conversational Speech". *Proceedings of EuroSpeech 97*, 2:1005–1008.