The Effect of Pitch, Intensity and Pause Duration in Punctuation Detection

Tal Levy,[1] Vered Silber-Varod, Ami Moyal

ACLP – Afeka Center for Language Processing
Afeka Tel Aviv Academic College of Engineering
218 Bney Efrayim Rd.
Tel Aviv 60107
Israel

Abstract—The purpose of this research is to automatically detect punctuation in speech using only prosodic cues. We aim to integrate prosodic elements such as pauses, changes in f_0 and amplitude range, into an Automatic Speech Recognition engine in order to generate punctuation for read speech, without taking the context of the sentences into consideration. We trained acoustic models of the prosodic features of two Punctuation Marks (PMs): full-stop and comma, which we assume have distinct prosodic characteristics. A Neural Network was used to estimate the weights assigned to each prosodic feature that corresponds to a particular PM, later to be used by a PM classifier. Results show that 87% of full-stops were detected, with only 14% false alarms. Nevertheless, since most commas are realized with no pitch breaks, only 54% of the commas were detected, with 35% false alarms. Our results support the hypothesis that acoustic-prosodic cues provide useful evidence about phrases.

I. Introduction

The format of the standard Automatic Speech Recognition (ASR) engine’s output is currently known as Standard Normalized Orthographical Representation (SNOR) and consists only of single-case letters without punctuation marks. Accurate as the best ASR is in converting speech into words, because its output does not include punctuation marks, it is difficult to process. Similar to a book without punctuation, standard ASR output risks ambiguity, a fact that poses problems in further automatic natural language processing, such as machine translation (MT), summarization, information extraction, and more. In addition, most dictation systems rely on spelled-out punctuation; otherwise the punctuation is not present. Although this method allows for a variety of punctuation marks to be dictated, its disadvantage is in the disfluent manner of natural speech, not to mention the cessation of thought processes. Other applications (such as Nuance’s Dragon, for example) enable automatic punctuation but limited to commas and full-stops.

Various approaches to *punctuation generation* with relation to *prosody models* have been proposed in the literature since the beginning of the 21st century, from [1], [2], and [3] until [4].Yet, studies on *comma prediction* [5], [6], *syntactic chunking with acoustic cues* [7], *automatic speech segmentation* [8], or *prosody-syntax interface* [9] contribute to this field as well.

Prosody is defined as the rhythm and intonational aspect of the language and its characteristic signals are partly used in today's ASRs. Pauses, changes in pitch range and amplitude, global pitch declination, melody and boundary tone distribution as well as speaking rate variation are all prosodic elements that can be used by ASR engines to generate punctuation, and in addition to being trained as “noises”– silence, breaths and other non-speech sounds can be modeled into different punctuation marks, or at least to the most common ones – full stop and comma.

The state of the art experiments show that best results are achieved when using a combination of context (words, syntax, or language-models) and acoustical information as input, rather than each model on its own [inter alia, [1], [6], [7]]. As to the system architecture, most research, e.g., [4], and [8], recovered punctuation at the post-processing stage, after the completion of speech decoding, either by using the generated best scoring hypothesis or the word lattice as input. [10], on the other hand, present a system that produces punctuation and a speech recognition hypothesis simultaneously. However, it is agreed that “prosodic information can be used to improve the punctuation detection” [4: 481].

The overall system architecture used in this research involves prosodic feature extraction, training in Neural Networks (NN) for the PMs and a NN-based classifier to determine whether the input segment is speech or punctuation, and to classify the detected punctuation into two possible punctuation marks – full-stops or commas. Neither ASR output nor textual/content/language analysis was involved in the framework of this research.

II. Data

The American-English speech used in this research for training PMs consisted of two hours (three chapters) reading of George Orwell’s *1984* audiobook by Frank Muller – a professional narrator [12]. This corpus provided us with speech files as well as their “transcription” (the original text) and hence no collecting and transcribing process was needed.

[1] This research is part of Tal Levy’s final project, carried out during his B.Sc. studies in electrical engineering at Afeka – Tel Aviv Academic College of Engineering.

978-1-4244-8682-3/12/$26.00 ©2012 IEEE

70% of the recordings were used for training purposes and the remainder –30% – for the testing phase.

Table 1 shows that the most dominant punctuation marks were full-stops and commas, representing about 90% of the total punctuation in the training material. Due to the lack of other punctuation marks, it was decided to focus on full-stop and comma detection, and leave the question mark and other punctuation marks for a future study. In another study, [4], a different approach was taken: a variety of punctuation marks, including the question mark, were converted to full-stops and the dash, "-", was included as a comma.

Table 1: punctuation and pause distributions in two hours of read audiobook

Punctuation mark	**Occurrences**	**Followed by a pause**
Full stop (.)	1331	96.5%
Comma (,)	1189	38.4%
Question mark (?)	103	89.3%
Other (: ; ! -)	229	76.8%

III. METHOD

A. Segmentation and annotation

Training data consisted of manually annotated and segmented speech files. The annotation and segmentation process was carried out using Praat software [13] according to the punctuation marks in the written version. Although our method required hand-labeling of prosody for training purposes, it was not a process that required agreement between labelers, since all punctuation marks were annotated according to the reference book (and not according to any perceptual or acoustical parameters, and with no subjective intervention).

Most punctuation marks were realized as non-speech events (see Table 1); that is, the reader either inhaled or took a short break (silent pause) while "executing" the punctuation marks. In our research, we called such a non-speech event a "Punctuation Interval" (PI). According to the data gathered, the minimum duration of a PI is 100ms. Yet, in a substantial number of cases, the punctuation marks (mostly commas) were realized in fluent, constant speech. These cases were annotated as "Punctuation Points" (PPs). For the purpose of analysis, since the important aspects of pitch and intensity measurements are not their absolute values but the reset, or change, in the values over time, these PPs were then transformed into PIs, with artificial boundaries inserted at 50ms before and 50ms after the PP (thus creating a 100ms PI).

After manually annotating the speech file for reference, automatic extraction was needed for the training and testing phases. For that purpose, intervals of at least 100ms that did not contain pitch were considered possible punctuation candidates, possible PIs. In the second stage, for punctuation within fluent speech, the audio file was analyzed using a sliding window every 100ms (thereby making each window a possible PI), from which the feature vector was extracted.

The third column in Table 1 shows the percentage of punctuation marks that were followed by a pause of 100ms or longer. This data was taken into account during the two stages of automatic detection of PIs – we expected to find almost all the full-stops but only about 38% of the commas (the percentage of commas followed by a pause in Table 1). In order to find most of the commas, we needed to look for PPs.

B. Prosodic features

The prosodic features that were extracted for each of the punctuation marks are listed below, and are in accordance with the common knowledge that "Underlying the prosodic feature extraction process is the linguistic evidence that pitch contour, boundary tones, energy slopes, and pauses are crucial to delimit SUs (Speech Units) across languages." [4: 482]:

1) PI duration (Min, Max, Mean, SD);

2) f_0 values (Min, Max, Mean, SD, in Hz) at N points *before* the PI and *after* the PI;

3) Energy values (Min, Max, Mean, SD, in dB) at the same N points *before* the PI and *after* the PI as the pitch points;

4) f_0 gap (slope) between the last point before PI and the first point after the PI;

5) energy gap (slope).

The prosodic feature extraction was performed using Praat software acoustic analysis of [13].

C. NN

A two layer feed-forward network with sigmoid hidden and output neurons was used; and the network was trained with scaled gradient conjugate back propagation. In other words, the five prosodic features mentioned above were modeled as input parameters and each received a different weight. The weights were summed up and a transfer function was created. In our case, the transfer function produces a number between zero and one, indicating the probability of having the relevant punctuation mark.

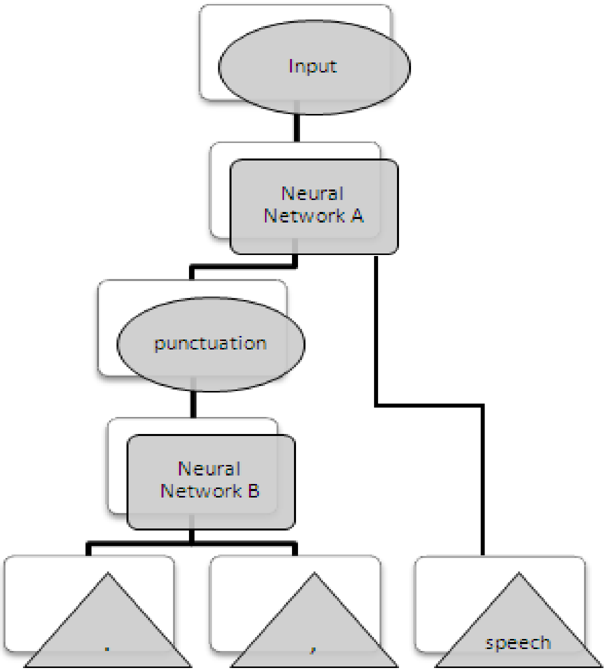

Fig.1 System architecture

D. *System architecture*

Fig. 1 represents the system architecture. Input segments consisting of intervals of over 100ms without pitch are automatically selected. The second stage is NN-A which separates the input segments into two groups – speech and punctuation. The next stage is NN-B which classifies the punctuation segments as commas or full-stops.

IV. RESULTS

A. *Initial Measurements*

The most distinctive feature differentiating between commas and full-stops was the PI duration parameter, where, as shown in Fig. 2, the mean duration for a full-stop (0.84 sec) is 1.5 times longer than for a comma (0.56 sec). As to the question mark interval, the mean value is 1 second but the standard deviation is 0.44 sec, compared to the 0.36 sec standard deviation of the full-stop interval and 0.31 sec of the comma. The "other" punctuation marks are much more diverse.

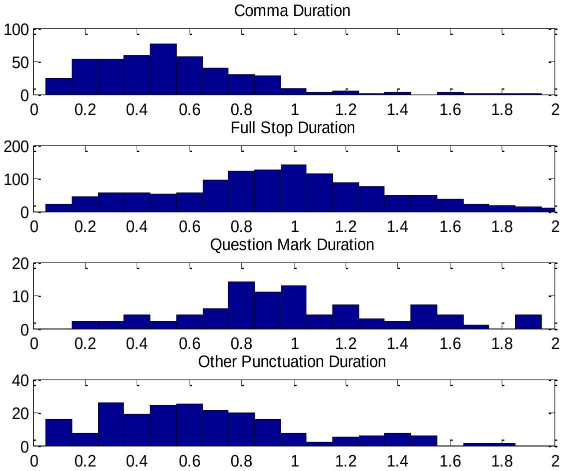

Fig. 2. Punctuation interval duration

The results concerning the pitch and intensity features (the last N values of speech *before* the punctuation interval (PI), the first N values *after* the PI, and the *gap* between the two) were not as clear as for duration, and for some parameters, there is even overlap between the different punctuation marks. Yet, statistics regarding the f_0 gaps (the 4th of the prosodic features in section III-B) and the intensity gaps (the 5th feature in section III-B) show slight differences between full-stops and commas, as shown in Table 2.

Table 2: Values of f_0 gap and intensity gap for full-stops and commas

	f_0 (Hz) gap		**Intensity (dB) gap**	
	Full stop	**Comma**	**Full stop**	**Comma**
Min	0.1	0.1	0.1	0.1
Max	123.3	115.7	23.2	56.5
Mean	26	16.8	6.7	11.4
STDEV	20.4	14.9	4.1	10.5

B. *Recall and precision*

The first NN task was to decide whether the segment at hand was in fact punctuation. The decision was made only for 96.5% of the full-stops and 38.4% of the commas followed by pauses (see Table 1). Table 3 shows a confusion matrix where the rows represent the detection results and the columns represent the correct annotation. In 1674 cases, the NN output was punctuation, and the correct annotation ("Actual punctuation") was indeed punctuation; in 46 cases, the NN output was speech but the actual annotation was punctuation; in 775 cases, the NN output was speech and the actual annotation was speech; and in 73 cases, the NN output was punctuation yet the actual annotation was speech.

Table 3: Speech vs. punctuation (pause duration ≥ 100ms)

	Actual speech	**Actual punctuation ("." , ",")**
Classified as speech	775	46
Classified as punctuation ("." , ",")	73	1674

The parameters that interested us most were the precision and recall parameters for punctuation. Precision (rows) represents the fraction of retrieved elements that are relevant; while recall (columns) relates to the fraction of relevant instances that are retrieved.

$$\text{Precision} = \frac{1674}{1674+73} = 0.958$$

$$\text{Recall} = \frac{1674}{1674+46} = 0.973$$

$$F_1\text{-measure} = 2 \times \frac{(0.958)(0.973)}{0.958+0.973} = 0.97$$

We can see that the NN performed almost perfectly, with 95.8% precision and 97.3% recall.

The next step was taking the segments that were detected as punctuation and passing them through a second NN that divided them between full-stops and commas.

After the first NN to separate punctuation from speech and the second NN to separate commas from full-stops (see Fig. 1), speech was filtered almost perfectly (filtering 96% of all speech segments), and full-stop detection performed nearly as well (90% precision, 86% recall). Comma PI detection, on the other hand, had a high error rate (60% precision and recall), but still performed fairly well.

In parallel to the first two NNs, a separate NN was carried out for punctuation points (PP), i.e., for punctuation marks not followed by a pause. This NN involved only comma points since, as shown in Table 1, full-stops are almost always followed by a pause and therefore did not provide any substantive material for such an NN. Thus, the third NN decision was between speech and comma.

The results of the third NN showed that 61% of the commas were detected with 69% precision, compared to 60% precision and recall for comma PIs.

To sum up, when combining the detection of punctuation with and without pauses, we detected 87% of the full-stops and 54% of the commas with low percentages of false alarms

– 65% precision for commas and 86% precision for full-stops, *using prosodic features only*.

V. Summary and Conclusion

This paper documented an attempt to detect “punctuation events” in read speech based on three prosodic cues only: f_0, intensity, and pauses, with no other lexical information or language model. Moreover, our methodology did not require perceptual annotation or degree of agreement between labelers, since we had the original written version of the book.

The prosodic features were used as inputs in an NN model, which first predicted the appropriate speech-event (speech vs. punctuation), and then predicted the punctuation type (full-stop vs. comma).

As to the effect of the different prosodic features – we showed that when punctuation was uttered without a pause, this reduced the detection rate dramatically. Pitch gaps and intensity gaps provided a distinctive feature for differentiating between speech and punctuation and between full-stops and commas.

Interestingly, full-stops, probably because of their relatively long pauses and distinct pitch and intensity gaps, were much easier to detect. On the other hand, commas were better detected when not followed by a pause, showing that the algorithm better separates commas from speech than from full-stops.

We conclude that even minimal prosodic features provide rich information for the detection of punctuation and that it can be embedded into ASR for the purpose of punctuating and disambiguation, to make the output clearer and more readable.

Still, in keeping with state-of-the-art research that deals with punctuation detection, we plan the following: to add phoneme duration as an extra prosodic feature on the assumption that sentence boundaries are defined by stretching the phonemes at the end of speech units (sentences, for example) and rushing the phonemes at the beginning of the following speech unit. For that we will need an ASR as a pre-processing module; we also plan to widen the training database and add more punctuation marks to our arsenal; and finally, to integrate a language model, including punctuation probabilities, into the system in order to improve its accuracy.

References

[1] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tür, “Prosody-based automatic segmentation of speech into sentences and topics,” *Speech Communication* 32(12) Special Issue on Accessing Information in Spoken Audio, pp. 127–154, 2000.

[2] J. Huang, and G. Zweig, “Maximum entropy model for punctuation annotation from speech,” *Proc. ICSLP*, 2002.

[3] M. Ostendorf, I. Shaffran and R. Bates, "Prosody models for conversational speech recognition," *Proc. of the 2nd Plenary Meeting and Symposium on Prosody and Speech Processing*, pp. 147-154, 2003.

[4] F. Batista, H. Moniz, I.Trancoso, N. J. Mamede, “Bilingual Experiments on Automatic Recovery of Capitalization and Punctuation of Automatic Speech Transcripts,” *IEEE Transactions on Audio, Speech, and Language Processing*, IEEE Signal Processing Society, vol. 20(2), pp. 474–485, 2012.

[5] D. Hillard, Z. Huang, H. Ji, R. Grishman, D. Hakkani-Tur, M. Harper, M. Ostendorf, and W. Wang, “Impact Of Automatic Comma Prediction On Pos/Name Tagging Of Speech” *Spoken Language Technology Workshop, 2006. IEEE*, pp. 58–61, 10-13 Dec. 2006.

[6] B. Favre, D. Hakkani-Tur, and E. Shriberg, “Syntactically-Informed Models for Comma Prediction,” *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2009)*, pp. 4697-4700, Taipei, Taiwan, 2009.

[7] J. K. Pate and S. Goldwater, “Unsupervised syntactic chunking with acoustic cues: Computational models for prosodic bootstrapping. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, 2011.

[8] H. Moniz, F. Batista, H. Meinedo, A. Abad, I. Trancoso, A. I. Mata, N. Mamede, “Prosodically-Based automatic segmentation and punctuation,” *The Fifth International Conference on Speech Prosody, Speech Prosody 2010*, Chicago, USA, 2010.

[9] V. Silber-Varod, “The SpeeCHain Perspective: Prosodic-Syntactic Interface in Spontaneous Spoken Hebrew,” PhD dissertation, Tel Aviv University, 2012.

[10] J-H Kim, and P. C. Woodland, “A combined punctuation generation and speech recognition system and its performance enhancement using prosody,” *Speech Communication*, Vol. 41, Issue 4, pp. 563-577, November 2003.

[11] M. H. Beale, M. T. Hagan, H. B. Demuth, *Neural Network Toolbox*, http://www.mathworks.com/help/pdf_doc/nnet/nnet_ug.pdf, Retrieved at 15/05/2012.

[12] Internet Archive, http://archive.org/details/George-Orwell-1984-Audio-book, retrieved at 13/11/2011.

[13] P. Boersma, and D. Weenink, *Praat: doing phonetics by computer* [Computer program]. Version 5.3.02, retrieved at 7 November 2011 from http://www.praat.org/.