Recognition of Punctuation in Voiced And Unvoiced Speech For Ib-CET[*]

Jiangang Liu
School of Foreign Languages, Southeast University, Nanjing, Jiangsu Province 210096, China
Institute of Modern Educational Technology
Southeast University, Nanjing, Jiangsu Province 210096, China
(Contact E-mail: jhonliunj@163.com)

Chen Li
School of Foreign Languages, Southeast University, Nanjing, Jiangsu Province 210096, China
Institute of Modern Educational Technology
Southeast University, Nanjing, Jiangsu Province 210096, China

Li Zhao
Institute of Modern Educational Technology
Southeast University, Nanjing, Jiangsu Province 210096, China
School of Information Science and Engineering
Southeast University, Nanjing, Jiangsu Province 210096, China

Abstract

The paper discusses the feasibility of punctuation recognition in oral delivery through voiced and unvoiced speech detection. The paper is firstly focused on introduction of iB-CET, in an attempt to clear away possible illiteracy of linguistic perception and speech recognition technology. Secondly, the paper providess some touches on how linguists interpret punctuation, thus inducing the third part in interpreting engineer's concept on punctuation relying on Bark wavelet transform parameters and subspace analysis in speech punctuation recognition. Finally, through experiments, the paper convinces readers that punctuation in speech can be accomplished, hence leaves much thought for the notion of punctuation recognition in speech for oral delivery.

Keywords: punctuation; voiced and unvoiced; iB-CET

I. Introduction

The present Internet-base College English Test (iB-CET for short) is getting popular, though some scholars are still arguing against its credibility and feasibility. Designed and planned in May, 2007(Jin Yan, & Wu Jiang, 2009), the iB-CET got started among 53 colleges and universities in 20th of December, 2008 so that China made a milestone in computer-based test rather than paper-based test for high education in history(Liu Jian-gang, & Dong Jing, 2011). The main feature of iB-CET aims at oral delivery in the exam. This marks the very breakthrough in oral test to paper-based CET, owing to the contribution of speech recognition technology based on auto-computer scoring system. The oral test takes a proportion of 10% in the whole test marks in iB-CET, making a revolution in the paper-based test system, hence the iB-CET system comes into life, a wonder in Chinese education for English teaching. Anyhow, for whatever reasons and senses, the design and plan worked and became accepted by the authorities and faculties in educational fields(Jin Yan, & Wu Jiang, 2009).

However, the most challengeable argumentation results in the credibility of speech recognition of test contents in oral delivery, because nobody can find a satisfactory explanation on the technology and software design of speech recognition model in oral delivery on the official Website of CET(www.ccets.org). All doubts come from the

[*] This paper is written for the programs of "Research on the Production of Patent from Teaching Mode "(2242018S10036) supported by the Fundamental Research Funds for the Central Universities, and "A Study on English Education Planning in the Transition from Secondary to Tertiary Education" (17YYD001) supported by Jiangsu Social Science Foundation, China.

ignorance of linguistic perception and speech recognition on oral delivery detection model of the target language, leaving much to be explained by linguists and computer engineering scientists.

This paper attempts to clear away possible illiteracy of linguistic perception and speech recognition technology through an illustration by oral test scoring method with recognition on punctuation in voiced and unvoiced sound in oral delivery, trying to convince people involved that iB-CET could work in some ways.

II. Linguistic Perception on punctuation

Linguists feel strongly the movement in higher education to replace or supplement traditional pedagogical methods (eg paper-based or face-to-face learning) with online learning has seen considerable acceleration in the last few years, especially in relation to distance learning(Lisa Emerson, & Bruce MacKay, 2011). In traditional pedagogical methods, students can easily respond to teachers according to the contents in written form, deliberately noticing all kinds of punctuation while uttering in perfect pauses, making no syntactic or semantic mistakes. For example, in paper-based oral test, students are usually required to read the text as to be tested about their fluency and accuracy in understanding of the content. Hereby, we have to notice that students can make perfect contextual understanding of the reading material with assistance of punctuation such as comma, full stop or exclamation marks. Since the material is entirely text based, the students can easily take care of the structure of the content, making a perfect answer to examiner, which leaves us with a conception that punctuation is very important in syntactic and semantic understanding of the material in paper-based test. So, a question emerges into people's mind what is the role of punctuation in learning and test?

Throughout its history, punctuation has been employed for varying purposes(Robert J. Scholes, 1990). What was originally used to signal breath breaks in reading aloud has in the course of time come to signal for the most part division into structural units in both silent reading and reading aloud, thus serving to disambiguate meaning and provide clues for coherence of the written. Punctuation in the written is not an adequate substitute for prosody in spontaneous spoken discourse; punctuation has, therefore, only a remote relationship to prosody. However, in the case of reading aloud, the relationship is much closer. And even in the case of written texts themselves, punctuation can be used to simulate spontaneous spoken discourse, "What the hell!" for example.

We know that the written was originally without punctuation; such a script is formally designated as ***scriptio continua*** (continuous speech). The last chapter of James Joyce's Ulysses, Molly Bloom's soliloquy, is written almost entirely without paragraphs, capital lettering (except for proper names), and punctuation. The text thereby becomes extremely difficult to read, whether silently or aloud(Daniel C. O'Connell, & Sabine Kowal, 2008). This can be well met with the ancient Chinese book "Origin of Chinese Characters".

Hereby we come to a question: in what way can we define punctuation? We can define it in elocutionary and syntactic functions. In elocutionary function, punctuation serves as a set of instructions for reading a text aloud, more specifically as one aspect of written speech(written form and a kind of phonetic transcription(sound form) for prosody (stress, pause, and intonation). In syntactic function, punctuation serves to convey meaning. It does so by identifying lexical elements and clausal, phrasal, and sentential structure(Robert J. Scholes, 1990). In these two functions, we can then logically come into the definition of punctuation: punctuation is interpreted as notation for breathing. Merriam-Webster's Collegiate Dictionary(11thed., 2003, p. 1009) defines punctuation as "the act or practice of inserting standardized marks or signs in written matter to clarify the meaning and separate structural units", and "to break into or interrupt at intervals". The Modern Chinese Dictionary(1sted., 1978, p.69) defines punctuation as the written form marking sentence reading and intonation. The Usage of Punctuation (National Standard T) established by the Chinese Standard Bureau claims a precise description that punctuation serves as a written from of records, punctuation is used for the purpose of making pauses, indicating intonation and sentencing or paragraphing.

So linguists will easily come into a conclusion of punctuation that in our times, every sentence is made of words and a specific punctuation mark. A sentence does not make any sense without a punctuation mark. This can be perfectly demonstrated in early Greek texts of the fourth century BC, where punctuation was employed for paragraphs that marked the beginning of a new topic. Latin documents of the sixth century, for example, employed no critical signs, marked neither words nor sentences. The Usage of Punctuation (National Standard T) established by the Chinese Standard Bureau defines a sentence as "a linguistic unit making sense of expression through intonation with pauses in front and end". Form the description, we can get the message that punctuation is highly related with sentence and intonation. Though there is none special punctuation mark for intonation, still we can feel it by the function in punctuation marks of full-stop, question and exclamation.

In conclusion, when we define punctuation, we have to take into consideration of diction, sentence, paragraph, logic, syntactics, semantics, pragmatics and intonation as well, making sense of easy comprehension.

This is very practical concept of linguistics, which is fundamental for applied study and scientific research. As for the computer processing, engineers can also follow the concept of linguists that punctuation marks be detected by voiced and unvoiced speech based on syntactic analysis, semantic analysis, pragmatic analysis and emotional analysis. How do engineers follow the concept of linguists?

III. Engineering Perception on punctuation

Two scholars Reeves and Nass, from Standford University, found that people treat modern communication media as if they are human beings, so established principles of interpersonal communication also predict human responses to computers and television(Zhao Li, 2009). Before phonograph was invented, people could only keep their memory on paper, called written words. Things changed when people invented recorder, and people can keep their memory in sound, called signal speech which can be divided into the voiced and the unvoiced.

In auto-translation research, computer engineers employed corpora in written words to build up natural language processing systems so that machines can intelligently find the punctuation through sentence and paragraph boundary from perspectives of syntactics, semantics, pragmatics. Mainly for the sake of the text-to-speech system, in early stage, people found the development of a sentence boundary disambiguator(Romportl J., Tihelka D., & MatousekJ., 2003). Later, people designed an appropriate classifier where the essential task is to set up its internal parameters when they began to make sure of sentence boundary detection. For instance, B. Say, & V. Akman illustrated an example in their paper(1996) that punctuation can be served as a boundary detection device as follows:

He was happy. He found his book.

(*He was happy to find his book. / He was happy because he found his book.*)

Meanwhile, during 1976-1986, come computer scientists undertook punctuation detection on written form based natural language processing on the probabilistic analysis of a large corpus. They treated all sentence delimiters plus non-letter and non-digit characters as specially-marked, individual words which might have features and referred to by constraints. In this way, punctuation marks are used to detect clause boundaries or lists of similar categories. Also, in recognizing subjects, punctuation marks such as dashes to the left of a finite verb dramatically decrease the probability of the preceding word to be a subject. In the corpus studies, of all the finite verbs preceded by a punctuation mark, less than 5% have been found to have the preceding word as the subject. Thus we enjoy the fruit of translation soft-wear all around the world.

In recent decades, people are rushing into the research on punctuation based on sheep recognition in oral delivery. This is the follow-up of punctuation based on written form. Engineers use different ways to accomplish their goal, such as statistic language model based on Bintree and voiced or unvoiced speech detection.

1. Statistic language model based on Bintree(Qian Yili1, Xun Endong, & Song Rou, 2006)

Bintree is based on Statistic language Model to realize its task. In mainly analyzes a sentence into two in matching with punctuation. The practice takes it for granted that every word i is among all the words $w\ i$ ($1 \leq i \leq n$) on assumption of $W = w_1\ w_2 \ldots\ldots\ wn$， when the sentence is already divided into separate words. Bintree can show that between each word $wi\text{-}_1\ wi$, there is a potential linguistic pause, thus for every sentence there are n words consisting $n\text{-}_1$ possible pauses. These pauses from left to right might be pos， pos $\in$ ($1n - 1$). If a pause mark or notice is inserted into the potential pause point, say comma "," in form a new sentence in $W' = w_1\ w_2 \ldots wi\text{-}_1\ \Delta\ wi \ldots wn$， based on language training model, we can meet the function to figure out the probability on pauses in the sentences $P(W') = P(w_1),\ P(w_2\ |w_1) \ldots\ P(wi\ |wi\text{-}_1\Delta) \ldots P(wn|wn\text{-}_2\ wn\text{-}_1 1),$ which means finding out the best possibilities $arg\ max\ pos\ P(W')$.

2. Voiced and unvoiced speech detection

There are very few articles talking about punctuation detection on sound form (speech). Speech pauses ale considered as punctuation marks of spoken language. Lots of linguists take speech pauses as voiced punctuation marks, indicating intervals in speech. Computer engineers have been making full use of sound corpus, especially of voiced or unvoiced detection. The following table indicating the characters of punctuation marks(Table 1 characters of unvoiced speech[1])

[1] Supported by The National Natural Science Foundation of China: programs of 60572159, 60573184, 60473139. the experiments are undertaken with more than 3,000 sentences from sound corpus (slow speed for 2.59 syllables /s, the average length for each syllable between voiced and unvoiced takes 114．21ms.

Table 1 characters of unvoiced speech

characters	Punctuation marks						
	Punctuation marks in end				Punctuation marks in middle		
	.	?	!	、	,	;	:
Length coverage of unvoiced (ms)	1315- 2294	513- 2800	433- 3311	106- 1481	165- 2277	566- 2764	404- 2803
Average length of unvoiced (ms)	1794	1088	1132	628	738	1066	882

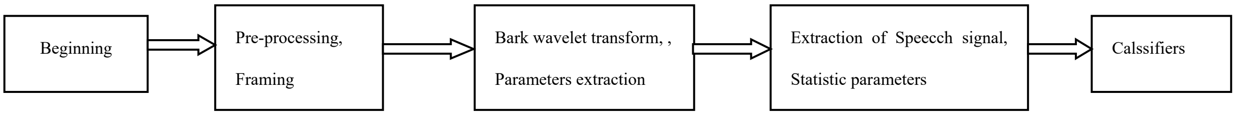
Therefore, we can draw a conclusion that we, at present, can arrive at a very practical method to realize sound-based punctuation according to table 1 by matching punctuation characters. For the very reason, the paper puts more touches on the solution of voiced and unvoiced speech.

For the detection on voiced and unvoiced speech, there are several methods to get accomplished, such as the one based on Bark wavelet transform parameters, the one based on subspace analysis, and ones in daptive thresholding approach or using fuzzy rules, only to mention a few. As an illustration, let us take a glance at the two followings.

2.1. based on Bark wavelet transform parameters

The detection on voiced and unvoiced based on Bark wavelet Transform parameters is to make use of the ability of frequency segmentation and energy focusing on Bark wavelet to extract statistic parameters of speech signals in sub-bands. The processing flow chart can be shown in conclusion as follows(chart 1).

Chart 1 The unvoiced speech processing flow



2.2. based on subspace analysis

The detection on voiced and unvoiced based on sub-analysis is to work out redundant information in speech signals. The algorithm is based on statistical analysis of the above mentioned wavelet-based frequency distribution of the average energy, zero-crossing rate, and average energy of short-time segments of the speech signal. The algorithm first classifies the input speech into voiced, unvoiced and uncertain parts by comparing features with predetermined thresholds. Then, the uncertain parts are treated in three conditions and dynamic thresholds are computed by extracted features of the input signal. Finally, the dynamic thresholds are used to classify the uncertain parts. The performance of the algorithm has been evaluated using a large speech database. The algorithm is shown to perform well in the cases of both clean and noise-degraded speech.

IV. Experiment and result for detection

Only facts can count. In order to get best performance for the detection on voiced and unvoiced, we tried our experiment in the location of language lab, gate of Southeast University and open space.

1. Experimental subject

Our research aims at the solution to the detection on voiced and unvoiced for programs of “Model and Application for Internet-based English Oral Test System”, “Teaching Model Reform under the Platform of Internet-based English Oral Test System” and “Model and Application Network for CET”. As the first step to realize auto-scoring performance, we did the experiment in the hope to solve punctuation recognition problem.

2. Examinee

In the first experiment, we asked 10 post-graduates(6 male, 4 female) in English class in the Foreign Languages Learning Center to read 210 sentences from their textbook. So in two periods, English teachers could get 2100 sentences. According to the computer-engineers’ design, 1000 sentences were for training purpose while 1100 sentences were for recognition data mining. The average speed of reading were designed at 8.2words/s, a bit slower than the speed of VOA in Special English, in a sample of frequency at 12KHZ, with a filter of window in 2133ms length and 10ms offset. We asked the students to do the other experiments with the same requirements, both in the gate and in the open space.

3. Condition

The experiments subjoined natural noise so as the meet the real situation in language lab in the iB- Cet. Only in this way can we get the best result we were expecting before. The noise was made by the computer, designed for the sound/noise ratio at 20db, 10db, 5db and 0db.

4. Procedure and result

Our experiments applied subspace analysis to detection on voiced and unvoiced for punctuation recognition in programs of "Model and Application for Internet-based English Oral Test System", "Teaching Model Reform under the Platform of Internet-based English Oral Test System" and "Model and Application Network for CET".

At the beginning, we tried to find out the shortest path for algorithm in characteristic space by using Nearest Feature Line(NFL in short), which is illustrated in the following chart(chart 2 and 3).b

Chart 2 characteristic points on NFL

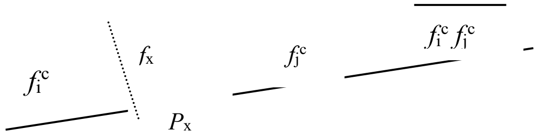
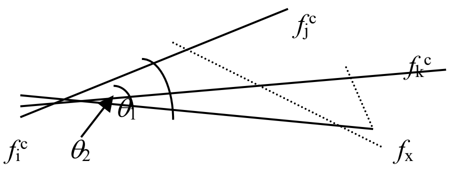


Chart 3 the shortest path in algorithm



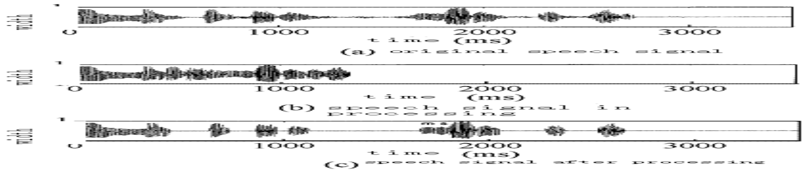
Secondly, relying on subspace analysis, we input recorded sound hinted into the subspace characteristic of sound, figuring out voiced and unvoiced according to the volume of recorded sound hinted(Zhao Li, Zheng Yumin, & et al., 2001).

Thirdly, we tried to figure out the related matrix based on training data, thus finding out the features and volume of the related matrix. To make it simple, the threshold value(θ) is the decisive value, and when the final value is bigger than θ, it means voiced, and the other way round when θ is smaller： $\frac{X^t \mathrm{B} X}{X^t X} > \theta$. Hereby, X stands for Feature vector in each frame, with *i* times in subspace. B stands for matrix of recorded sound hinted. The experiments results are demonstrated in the following(table2 and Chart 4).

Table 2 experiments results illustration

SNR[2]	20db	10db	5db	0db
Beginning	100	100	98.7	97.7
End	100	100	98.3	96.7
Average	100	100	98.5	97.2

Chart 4 voiced and unvoiced detection



During the experiments, the detection went on in each frame, and the detection result comes from the figure with tolerance in 10 frames without mistakes. From table 2, we can arrive at a satisfactory result as expected with a average recognition of 97.2%, and in chart4, the voiced and unvoiced detection is well and clearly demonstrated.

V. Punctuation recognition

After we accomplished the task in detection on voiced and unvoiced, we can build a set of rules to disambiguate a punctuation mark in composing software to visualize punctuation marks. The rules are defined as follows:

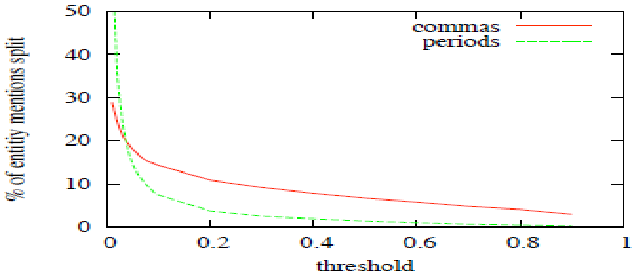
1. result := T (true, punctuation mark is the unvoiced before the voiced)
2. if the preceding token is a number sound, result := F (false)
3. if the following token is with the voiced and followed by the unvoiced, result := T
4. if the preceding token is voiced in rising tone and followed by unvoiced, result := T
5. if the preceding token is voiced in falling tone and followed by unvoiced, result := T
6. return the actual value of result

As for special punctuation, such as comma, period, exclamation or question mark, we apply the data described in Table 1 in matching the length in "ms". Besides, we can apply emotional speech recognition to accomplish exclamatory and questionable marks in punctuation(Wang Zhiping, Zhao Li, & Zou Cairong, 2006). We

[2] Sound and noise ratio

can also use a method to split a noun-phrase(NP in short) merging for missed punctuation so as to choose a lower threshold in order to reduce NP merging, resulting in more splits that shorter speech sentences. Thus, comma and period for example, in shorter sentences, are more likely to break entities, especially if they involve long phrases(Chart 5)(Benoit Favre, Ralph Grishman, & et al., 2008).

Chart 5 comma and period inserted in threshold



VI. Conclusion

The concept presented in this paper focuses on detection of voiced and unvoiced so as to accomplish punctuation recognition. The paper also made some touches on speech NP splitting and emotional speech recognition. The voiced and unvoiced detection has great to do with punctuation recognition in speech. It is very important and vital for the iB-CET in oral test, carrying realistic and practical value in Chinese English teaching.

The importance of doing this research lies in the concept, which was based on paper rather than on speech itself. Since the notion of a sentence is very different in speech compared to written text, it is very crucial in education when technology has reached the peak to match human's intelligence.

Acknowledgment

We appreciate our great thanks to those postgraduates fighting for their academic papers in figuring out precise data in the programs of "Research on the Production of Patent from Teaching Mode "(2242018S10036) supported by the Fundamental Research Funds for the Central Universities, and "A Study on English Education Planning in the Transition from Secondary to Tertiary Education" (17YYD001) supported by Jiangsu Social Science Foundation, China.

References

Jin Yan, & Wu Jiang. (2009). A Study on Language Identification in Call English Oral Test. *Foreign language World:133(4)*, 61–68. (In Chinese).

Liu Jian-gang, & Dong Jing. (2011). The design principles for iB-CET. *Chinese Journal of Electron Devices:34(4)*, 482–484. (In Chinese).

Lisa Emerson, & Bruce MacKay. (2011). A Comparison between Paper-based and Online Learning in Higher Education. *British Journal of Educational Technology:42(5)*,727–735.

Robert J. Scholes. (1990). Prosodic and Syntactic Functions of Punctuation: A Contribution to the Study of Orality and Literacy. *Interchange: 21(3)*,13-20.

Daniel C. O'Connell, & Sabine Kowal. (2008). Cognition and Language: A Series in Psycholinguistics. Germany: Springer, 1-9.

Zhao Li. (2009). Speech Signal Processing(Second Edition). *Beijing: Chian Machine Press*: 261(In Chinese).

Romportl J., Tihelka, D., & Matousek, J. (2003). Sentence Boundary Detection in Czech TTS System Using Neural Networks. *Signal Processing and Its Applications, Proceedings, Seventh International Symposium:2*, 247 – 250.

B. Say, & V. Akman. (1996). Current Approaches to Punctuation in Computational Linguistics. *Computers and the Humanities:30(6)*,457-469.

Qian Yili1, Xun Endong, & Song Rou. (2006). Application of Bintree Based on SLM in Speech Pauses' Prediction. *Computer Engineering*:*32(19)*,23-28.

Zhao Li, Zheng Yumin, & et. al. (2001). A Quiet Speech Identification Based on Subspace Analysis. *CCSP-2001*:148-151(In Chinese).

Wang Zhiping, Zhao Li, &Zou Cairong. (2006). Emotional speech recognition based on modified parameter and distance of statistical model of pitch. *ACTA ACUSTICA: 31(1)*, 28-34(In Chinese).

Benoit Favre, & Ralph Grishman, & et. al. (2008). Punctuating speech for Information Extraction. Acoustics, Speech and Signal Processing. *2008 ICASSP, IEEE International Conference:* 5013-5016.