ISCA Archive
http://www.isca-speech.org/archive

Speech Prosody 2012
Sixth International Conference
Shanghai, China
May 22-25, 2012

A Quantitative Study on Information Contribution of Prosody Phrase Boundaries in Chinese Speech

Jinsong Zhang[1,2] , *Wei Li*[2], *Yanlu Xie*[2], *Wen Cao*[1]

[1] Center for Studies of Chinese as a Second Language,
[2] College of Information Science,
Beijing Language and Culture University, Beijing, China
E-mail: jinsong.zhang@blcu.edu.cn, {blculiwei,xieyanlu}@gmail.com, tsao@blcu.edu.cn

10.21437/SpeechProsody.2012-140

Abstract

In speech, acoustic cues are used to manifest a number of linguistic events including segmental phonemes and supra-segmental ones such as tones, prosodic phrasing structure, intonation, etc. It has been an interesting topic to quantitatively compare the importance of different linguistic events. However, previous studies have been mainly confined to segmental or segment-like units. No studies could be found to show quantitatively the importance of supra-segmental events of prosody boundaries. From the view of information transmission, both segmental and supra-segmental events make indispensable contributions to natural human speech communication. This paper presents a novel way to quantitatively estimate the information contribution of prosody boundaries, taking the same way to estimate those of segmental phonetic contrasts and syllable tones. Experiments were done using a Chinese newspaper text corpus and a conversation speech corpus. Preliminary results show that prosody boundaries carry much more information than phonetic segments. Hence, they are much more important than segments for human speech communication.

Index Terms: functional load, prosody boundary, mutual information, information contribution

1. Introduction

In speech, acoustic cues are used to manifest a number of linguistic events [1] including segmental phonemes and supra-segmental ones such as tones, prosodic phrasing structure, intonation, etc. Phonemes are differentiated by various kinds of phonetic contrasts such as voicing vs. voiceless, aspiration vs. un-aspiration, etc, which are realized through regular movements of vocal tract and changes in articulation mode. From the view of information transmission, all these events are coding methods of speakers' message, and have to be decoded by the listeners. Therefore, they play important contributions to natural human speech communication.

Questions have been raised in speech studies [2,3] about information contributions of different phonetic events: how to measure them reliably and what they are like? The information contribution of a speech event was measured using Functional Load (FL) in early studies [2,3]. The measurement of FLs provides a quantified way to order any phonetic contrasts in a language. The order is meaningful and applicable to many domains of research and applications, such as language evolution, speech recognition, language acquisition, phonetics, phonology, etc [2-4]. It was predicted that perceptually similar pairs of phonemes with low FL would merge as the language evolved [2]. Phonetic events with high FLs should arouse more attention when incorporated in human-machine interface systems. Phonemes with low FL might be merged to reduce the size of sub-word unit set for developing an automatic speech recognition system [7]. In the area of 2nd language education, FL provides guidance for the importance of phonetic contrasts to be learned.

There have been a few methods proposed to compute the FLs of various kinds of phonetic contrasts [2-5]. The most used ones were frequency counts [3] and entropy based measurements [2,4,5]. These methods are computable by using a large scale text corpus [2-5]. Quantitative results have been provided for phoneme pairs, phonetic contrasts, lexical tones, etc. [2-5]. The results helped to improve our understanding of phonetic parts of human languages.

However, almost all the previous studies were confined to the FLs of segmental contrasts or segmental-like ones such as tones. No studies have worked on the issue of evaluating the FLs of supra-segmental events like prosody boundaries. Prosodic phrasing is one of the most important prosodic features, which gives rise to a segmentation of the speech chain into groups of syllables and words, or in the other word, chunks. The segmentation of chunks was found to be important for speech planning and language perception [1]. Therefore, it is reasonable and interesting to ask: How about the information contribution of prosody boundaries? How about it when compared to other phonetic events?

One major reason for the lack of studies on FLs of phrase boundaries might be attributed to the fact that there have been no appropriate methods for evaluating their information contributions. Most proposed ones depend on frequency counts of speech events [2,3,4]. Since phrase boundaries are not on the same level as segmental events, it is inappropriate to evaluate their FLs using frequency counts or any derived entropies. Consequently, it is impossible to compare their information contribution to the segmental events one.

In our previous study [5], we proposed a novel method to estimate FLs of phonetic events, which is based on the mutual information (MI) of text transcriptions and their phoneme representations. The phoneme representations vary according to the availability of the studied phonetic events. Usually, the unavailability of a phonetic event might lead to more uncertainty in the decoding space from a phoneme representation to its text. The FL of a phonetic contrast, for example, [p] vs. [p^h], is computed as the relative reduction of uncertainty by the availability of the contrast of [p] and [p^h], when compared to none. The method takes into account the contextual effects including lexicon and word concatenations, therefore it is able to model more accurately human language process than the other ones [2-4].

The decoding uncertainty from a phoneme representation to its text usually decreases with the availability of more phonetic information. Either segmental phonemes and tones or supra-segmental prosodic events can help to reduce the uncertainty.

Therefore, the different levels of segmental and supra-segmental phonetic events can be viewed to have the same role in human speech communication: uncertainty reduction. From this view, the FL of prosody boundaries can be estimated in the same way of segmental events, i.e., they can be computed using the MI based FL estimator [5].

In [5], the decoding uncertainty is represented by a word hypothesis graph (WHG), with all the word paths in the WHG sharing the same phoneme representation. Whenever more phonetic information (events) is incorporated into the phoneme representation, the size of WHG tends to decrease, indicating a reduction of uncertainty. The information of a phrase boundary position can prohibit word formation among syllables across the boundary, so that it helps to reduce the size of a WHG. The amount of uncertainty reduction is computed by the probability change of word paths in the two WHGs with or without a phonetic event.

The following is organized as follows: Section II introduces theory backgrounds including source channel model of Text-Phoneme-Text transmission, the MI based on functional Load estimator, and the Word Hypothesis Graph to represent decoding uncertainty. Section 3 describes the experiment setup and experimental results. Section 4 concludes the study and suggests future directions.

2. Fl based on MI

2.1. Text-Phoneme-Text Transmission Model

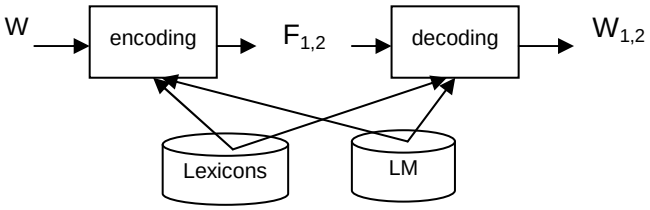

Figure1: Text-Phoneme-Text transmission model

As in [5], speech communication is modeled using a Text-Phoneme-Text model, illustrated in Figure 1, where W stands for a language and appears as a text corpus. The availability or unavailability of specific phonetic events can be modeled by different lexicons and different phonetic representations $F_{1,2}$ of W. If the studied phonetic events are segmental contrasts, two lexicons having or not having the specific contrasts are used to encode W into two phoneme representations. If the studied phonetic events are supra-segmental events like phrase boundaries, F_1 may stand for the phoneme representation without phrase boundaries, while F_2 has the same phoneme representations as F_1 plus boundary information. The encoding from W to F depicts the transmission of a message from a speaker to a listener via phonetic coding.

The conversion from F to W represents the interpretation of a phoneme sequence into a word sequence based on high level knowledge, including lexicon and language model, by a listener. In the model, it was realized as a word lattice scoring process as done in most automatic speech recognition systems. Different phoneme transcriptions $F_{1,2}$ might be interpreted into different word sequences $W_{1,2}$.

2.2. Mutual Information (MI)

The mutual information of is W and F is defined as $I(W,F)$:

$$I(W,F) = H(W) - H(W \mid F) \qquad (1)$$

$H(W)$ is the entropy of text corpus W, which depicts for sequence of words $\{w_1, w_2, \cdots, w_n\}$. It is usually calculated as the word average entropy by:

$$H(W) = \lim_{n\to\infty} -\frac{1}{n}\log p(w_1, w_2, \cdots, w_n) \qquad (2)$$

Where

$$p(W) = p(w_1, w_2, \cdots, w_n) = \prod_{i=1}^{n} p(w_i \mid w_1, \cdots, w_{i-1}) \qquad (3)$$

$I(W,F)$ measures the relation between the sentence W and its phoneme transcription F. Given two kinds of phoneme transcriptions F_i, i=1,2, when other conditions including lexicon and language model are the same, the bigger the $I(W,F_i)$ is, the stronger the relation between W and F_i is, and the less ambiguity is issuing from F_i to deduce W.

After derivation and ignorance of polyphone words as done in [3], the $I(W,F)$ can be calculated as follows:

$$I(W,F) = -\log \sum_{all\ W_j\ \text{with}\ F} p(W_j) \qquad (4)$$

W_j stands for all word sequences that have the same phoneme transcription F.

2.3 Functional Load (FL) based on MI

In [5], Functional Load was defined for a phonetic contrast θ: x vs. y, based on a relative loss of mutual information before and after the deletion of the θ.

$$FL(\theta) = \frac{I(W, F_{\text{with }\theta}) - I(W, F_{\text{without }\theta})}{I(W, F_{\text{with }\theta})} \qquad (5)$$

2.4. Word Hypothesis Graph (WHG)

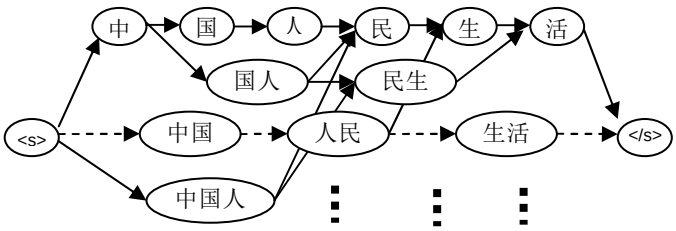

Figure 2: An example of a partial WHG for the phonetic transcription "*zhong guo ren min sheng huo*".

The decoding uncertainty from a phonetic representation to its text is represented by a word hypothesis graph (WHG), with all the word paths in the WHG sharing the same phonetic representation. Figure 2 illustrates a part of a WHG of a phoneme transcription of

"*zhong guo ren min sheng huo*". The nodes <s> and </s> represent the start and end of all candidate sentences. The other nodes stand for candidate words sharing the same phoneme transcription. For examples, a candidate word for the syllable "*zhong*" is "中"(middle), a bi-syllable candidate word for the syllables "*zhong guo*" is "中国"(China). Due to the limited size, Figure 2 only shows a part of the whole WHG for the example phoneme transcription. The MI of this WHG is computed through a summarization of probabilities of all routes from <s> to </s> nodes, as defined by the equation (4).

If the studied phonetic events are prosody boundaries, they can be added into the phonetic representation. For example, if the word boundary information is assumed known for the example phoneme transcription in Figure 2, it becomes a string like "zhong guo | ren min | sheng huo", where "|" depicts a word boundary. With the availability of word boundaries, the WHG will decrease into only one path as shown by dashed line in Figure 2. This indicates clearly the reduction of uncertainty by prosody boundaries. The quantitative estimation of the information contribution can be performed by the equation (5).

3. Experiments and results

3.1 Experimental set-up

The training corpus consists of 500k sentences randomly extracted from the People Daily newspaper. After word segmentation, we used the CMU LM toolkit [8] to train a bi-gram word based language model (LM).

Table 1. *Statistics of the LM training corpus.*

Number of	Number
Sentences	500,000
Words	2,655,469
Characters	4,673,383
Avg. Chars per sentence	9.35
Words in lexicon	46,558

The testing data for prosody phrases consisted of text transcriptions of a natural speech corpus, in which prosody boundaries have been labeled. The information of testing corpus is shown in Table 2. After POS tagging, the testing corpus was converted to standard Pinyin transcriptions with boundary information.

Table 2. *The information of testing corpus.*

Statistics of testing corpus (per sentence)	
Number of characters	25.2
Number of word boundaries	16.5
Number of prosody word boundaries	5.7
Number of prosody phrase boundaries	2.8

3.2 Results

Experiments have been carried out to estimate FLs of a number of phonetic contrasts, among them including segmental contrasts, lexical tones and prosody boundaries. Figure 3 shows FLs of some pairs of consonants, and these results were based on newspaper data and adopted from [5].

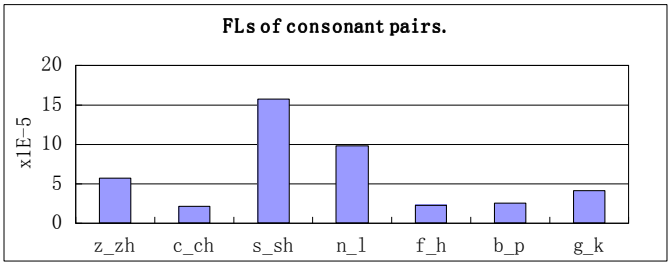

Figure 3: FLs of consonant pairs.

Figure 4 shows FLs of some pairs of lexical tones, together with that of consonant pair "s sh" which is the highest one in Figure 3. These results were adopted from [5] also.

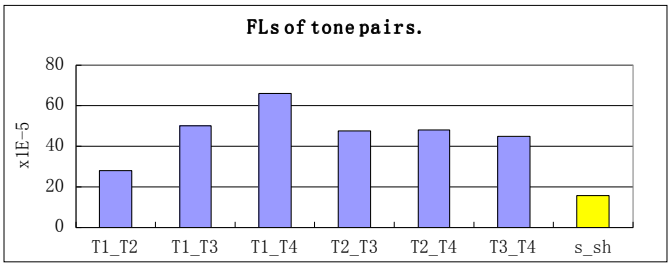

Figure 4: FLs of lexical tone pairs. Here T[1-4] stand for lexical tones [1-4].

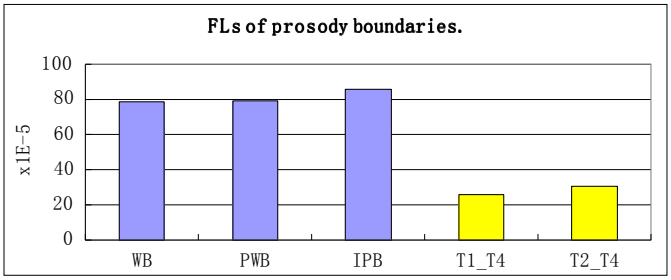

Figure 5: FLs of prosody phrases. Here WB/PWB/IPB stand for word boundary/prosody word boundary/intonational phrase boundary respectively.

Figure 5 shows FLs of prosody boundaries, together with lexical tone pairs of Tone 1 and 4, Tone 2 and 4. These results were computed on the testing data of text transcriptions of a natural speech corpus.

3.3 Discussions

Based on the results of Figure 3, 4, 5, we can observe:

- The MI based FL estimator is able to offer a uniform way to measure the information contributions by phonetic events at different levels.
- Different FLs in the three figures show that different phonetic events do make different contributions based on the source channel information transmission model.
- A look at Figure 3 shows that the consonant pair "s vs. sh" and "n vs. l" own the highest FL values when compared with other

checked pairs.

- Values in Figure 4 show that tone pairs “T1 vs. T4”, “T1 vs. T3”, “T2 vs. T4” had the highest FLs. As “s vs. sh” had the highest FL among the consonant pairs studied, we can say that tone contrasts usually have significantly high FLs than segmental contrasts (FLs of vowel contrasts are usually smaller than those of consonants [5]).
- FLs of two tone pairs “T1 vs. T4 ” and “T2 vs. T4” in Figure 5 were computed from the spoken corpus data, and were the highest FLs among all tone pairs in the corpus data. They are different from those in Figure 4, but the difference is not too much. Because the two pairs are also among the highest ones in Figure 4.
- Figure 5 shows that FLs of all boundaries, including word boundaries, prosodic word boundaries and intonational phrase boundaries, are significantly higher than those of the two representative tone pairs. Based on these, we suggest that boundaries carry relatively more information than the comparison tone and phonetic segment pairs for human speech communication.
- Figure 5 also shows a non-significant phenomenon: the FLs of the three kinds of boundaries are to a little extent correlated with the levels of boundaries in the prosody hierarchy: WB < PWB < IWB. However, we need further study to check this issue.

4. CONCLUSIONS

In this paper, we proposed to use mutual information between text and its phoneme transcription to measure functional loads of prosody boundaries. Experimental results showed that the estimator is able to offer a uniform way to measure the information contributions by segmental contrasts, tones and supra-segmental prosody boundaries. Preliminary results showed that prosody boundaries have the highest contributions when compared to segmental contrasts and tones.

The results will shed some light on the study of incorporating prosody information processing into man-machine interface systems: prosody phrasing structure carries relatively more information than segments, while they are usually ignored in ASR studies. In the future work, we will improve our models and do further studies checking the relationship between FLs and the levels of prosody boundaries.

5. Acknowledgements

The authors would like to thank Endong Xun of the college of information science of BLCU, Aijun Li and Ziyu Xiong of China Academy of Social Science for their help and for kindly providing the text and speech data.

We also would like to acknowledge the financial support by the China MOE Project 07JJD740060 of Key Research Institute of Humanities and Social Sciences, and the Youth Independent Research Program Projects 10JBT01 of Beijing Language and Culture University (Special Funds of Basic Research Costs for the National University).

6. References

[1] Thierry Dutoit, Automatic prosody generation, in An Intorduction to Text-to-Speech Synthesis, Kluwer academic publishers, 1997.

[2] W. S-Y. Wang, “The measurement of functional load”, Phonetica, 16, 1967, pp. 36-54.

[3] C. F. Hockett, “A manual of phonology”, International Journal of American Linguistics, Vol. 21-4, Indiana University Publications, 1955.

[4] D. Surendran, P. Niyogi, “Measuring the usefulness (Functional Load) of phonological contrasts”, Technical Report TR-2003-12, Dept. of Comp. Science, Univ. of Chicago.

[5] Jinsong Zhang, Wei Li, Yuxia Hou, Wen Cao, Ziyu Xiong, “ A Study On Functional Loads of Phonetic Contrasts Under Context Based On Mutual Information of Chinese Text And Phonemes”, The 7th International Symposium on Chinese Spoken Language Processing (ISCSLP), Tainan, Nov. 2010.

[6] Wang Pei, Yang Yufang, “Prosodic Structure and Syntactic Structure”, The proceedings of the third international Conference on Cognitive Science, PP491-496 2001.

[7] J.-S. Zhang, X.-H. Hu, S. Nakamura, “Using mutual information criterion to design an efficient phoneme set for Chinese speech Recognition”, IEICE Trans on Inf. & Syst. Vol. E91-D, No 3, March, 2008, pp.508-513.

[8] The CMU-Cambridge Statistical Language Training Modeling Toolkit http://mi.eng.cam.ac.uk/~prc14/toolkit.