A PROSODIC LABELING SYSTEM FOR MANDARIN SPEECH DATABASE

Chiu-yu Tseng and Fu-chiang Chou
Institute of Linguistics (Preparatory Office), Academia Sinica
Taipei, Taiwan R. O. C.
e-mail: ctyling@sinica.edu.tw

ABSTRACT

A working database needs tools to transcribe and label at both phonetic and prosodic levels. While the proposed phonetic transcription system is a simplified from of the International Phonetic Alphabet (IPA) following the SAMPA guidelines; the prosodic labeling system is an elaborated form of the ToBI (Tone and Break Indices) framework adopted for Mandarin. In particular, the proposed prosodic labeling system possesses the capability to label two kinds of prosody related phenomena: one correlates directly to specific acoustic parameters that can be measured automatically, namely, speech rate, speech volume, pitch level and pitch range; the other correlates to perceived temporal grouping and relative prominence, namely, breaks and emphases. Some preliminary findings with respect to the labeling system will be reported; these findings will also be utilized for subsequent design of software that aimed automatic transcription and labeling of Mandarin speech. Though the design is language dependent, we believe the findings may very well have cross-linguistic implications.

1. INTRODUCTION

The transcription of speech samples collected in a speech database usually aims at labeling and segmentation in layers. Our system aims to transcribe speech data in two major layers, the phonetic/phonemic layer and the prosodic layer. The citation-phonemic or phonetic transcription contains the output phoneme string derived from the orthographic form (by lexical access, by letter-to-sound rules, or both) [1] In order to label the segments, one may develop a platform that possesses the facility to display the full range of IPA symbols, or one may design an alphabetic and/or numeric representation of the IPA symbols instead. Since we set the prerequisite of our system for keyboard input by human transcribers that would correspond directly to ASCII specifications, it seemed feasible to avoid possible notational complexity both ways. We therefore chose to adopt the SAMPA system designed originally for major European languages, and adjusted it to a language-specific set of alphabetic phoneme symbols for Mandarin. The result is a less sophisticated system for segmental symbols; something we define as a broad phonetic system [2]. At this level, our system is capable of segmental transcription without tonal specifications or running speech related phenomena.

Leaving tonal representations and running speech for future work, we chose to focus our attention to transcribe and label our speech database at the prosodic level. Since there are less clear acoustic correlates to prosodic phenomena, this level is less straightforward than phonemic annotation. A basic distinction may be drawn between a prosodic labeling system that annotates the boundaries of units (analogous to the method used in phonemic annotation) and a system that annotates the occurrence of isolated prosodic events, such as F0 peaks. The former theoretical orientation, i.e., the use of boundaries, resulted in approaches that used intonation categories proposed by [3] to process suprasegmental information, such as intonation phrase, phonological phrase, phonological word, foot, and syllable. Alternatively, it could mark the more traditional units of “minor tone-unit” and “major tone-unit”, as in the MARSEC database [4]. The latter theoretical orientation, i.e., the occurrence of isolated prosodic events, resulted in the marking of these occurrences of high and low tones of various kinds. The recently formulated ToBI transcription system [5] appears to be the most well-known system of this kind, and was reported to work for non-tonal languages such as English and Japanese, where the prosodic units are annotated at the “break index” level rather than the “tone” level. Therefore, we chose to adopt a ToBI-like framework to design the prosodic transcription of our system, but needed to modify it to suit our target languages. Needless to say, the modification is somewhat elaborate due to the intrinsic differences between intonation languages and tonal languages.

2. PROSODIC TRANSCRIPTION

The prosodic transcription of our system is represented on a separate level following ToBI-like notations. The tone and break-index tiers represent the core prosodic part of the ToBI system. The difference in the break-index tier between ToBI and the prosodic level of our system is rather little. In the ToBI system, the break-index tier marks the prosodic grouping within an utterance by labeling the end of each word for its subjective strength in association with the following word on a scale from 0 (for the strongest perceived conjoining) to 4 (for the most disjointed boundaries). Our system followed the same rationale but offered a slightly more elaborate scale of break indices from 0 to 5. As a result, the following six boundaries were proposed instead, i.e., reduced syllabic boundary (0), normal syllabic boundary (1), minor-phrase boundary (2), major-phrase boundary (3), breath group boundary (4), and prosodic group boundary (5). The speech segments between the break indices then form a set of five units, namely, prosodic units, minor prosodic phrase, major prosodic phrase, breath group and prosodic group [6].

The most noted difference between our system and ToBI lies in the tonal and prosodic tiers. ToBI was originally designed for English, an intonation language. It consists of labels for distinctive pitch events, transcribed as a sequence of high (H) and low (L) tones marked with diacritics to indicate their intonation functions. Whereas when dealing with tonal languages, the interaction between lexical tone and intonation is not only more complex but also not well understood, yet. Since both of which involve deliberate manipulation of fundamental frequency patterns, it is difficult to differentiate one from another. We proposed to label the speech data in more detail at the prosodic domain for the time being while leaving the tonal aspects for future studies. Our reason was again the fact that the text for our speech data was readily available for reference.

The physical aspects that corresponds directly to volume, rate, pitch level and pitch range of each prosodic unit can be measured and labeled by developing software readily. However, both the perceived changes of breaks and emphasis that may or may not correspond directly to the physical signals have to be manually added on by transcribers. A transcriber has to note with acceptable consistency of these changes in the utterances.

When the contents of the prosodic transcription is decided, a more standardized method of representation would be the next feasible step. We believe the Java Speech Markup Languages (JSML)[7] could be a good choice for this purpose. There are two elements in JSML, namely, empty elements and container elements. An empty element has only one tag and is suitable for the representation of break indices. A container element has a balanced start tag and end tag and is suitable for the representation of the other factors. These tags are inserted into the phonemic representation of the syllables sequence to form the prosodic transcription. For example:

dz\in1 tien1 <BREAK level=2/> <EMP level=2> tien1 ts\i1 </EMP> h@n3 hau3 <BREAK level=4/>

The example shows that “dz\in1 tien1” is a minor prosodic phrase and “tien1 ts\i1” is emphasized at a moderate level. When the break index is “1”(normal syllabic boundary), it will not be marked to reduce the number of tags used. In other words, normal syllabic boundary will be unmarked. This unmarked convention is held constant for all other prosodic parameters whenever the perceived level is the normal one. The meaning of the levels for each marker are listed in Table 1.

3. EXPERIMENTS

The proposed system is evaluated on the basis of a Mandarin speech corpus that is designed to be phonetically and prosodically rich. There are about 600 short paragraphs in the corpus. To test the segmental labeling system, the major task was to verify the capability of transcription of speech variations at the segmental level. To test the prosodic labeling system, the major task was to define a standard for the transcriber and at the same time maintain the consistency between the transcribers. Another important issue that should be included in the investigation was the convenience factor for both humans and the computer. The corpus was labeled by two transcribers. At first, the two transcribers labeled a small set of identical speech data in order to discuss the standard used for transcription. After several such sessions, a set of one hundred sentences was labeled by each transcriber for comparison. The comparison was focused on the consistency of break indices. The transcription tool is a package called "Waves+" from the Entropic and is shown in Figure 1.

Level Tags	0	1	2	3	4	5
BREAK	reduced syllabic boundary	normal syllabic boundary	minor-phrase boundary	major-phrase boundary	breath group boundary	prosodic group boundary
EMP	reduced	normal	moderate	strong		
RATE	very slow	slow	normal	quick	very quick	
VOLUME	very low	low	normal	high	very high	
PITCH	very low	low	normal	high	very high	
RANGE	very small	small	normal	large	very large	

Table 1. The meaning of levels for the prosodic tags.

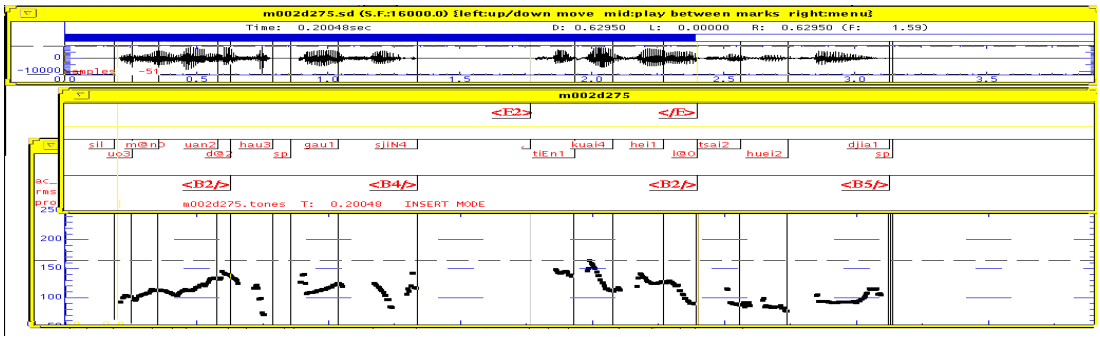

Figure 1. An example of the segmental and prosodic transcription

For the prosodic transcription, the standard for the labeling of break indices was evaluated. The major function of the break indices is to segment the speech flow into smaller units in the hope to form a hierarchical structure of prosody. We proposed a top-down spotting procedure for the labeling of break indices. At first we spotted all the breath groups (B4) in an entire paragraph, then search the prosodic groups (B5) among these boundaries. The second step was to spot the prosodic changes within a breath group. The change that accompanied a short pause was marked as B3; the others were marked as B2. The last step was to spot the reduced syllabic boundaries (B0) that accompanied the reduced syllables. The unmarked boundaries were normal syllabic boundaries (B1). The details are described below:

1. B4 and B5:

B4 is used to indicate the boundary of breath group that was originally proposed by Lieberman [8]. This boundary is a physiological effect that is caused by breathing exemplified by decrease at the levels of pitch and energy. To detect the reset of pitch and energy is no difficult task. However, it is less distinct to detect B5. In writing, a paragraph can be identified not by length but by a distinct format that involves specific spacing at the beginning, leaving off the remaining of a line, and beginning a new line with specific spacing again. The same phenomenon occurs in reading out paragraphs. Our question is: what would the cue for the marking of such a boundary be? In our observation, it is marked by the lengthening of the pause between the two breath groups. In our definition, this "paragraphing" in the reading out process is termed prosodic group, a unit in speaking that is equal or larger than a breath group. In our experiments, the transcribers were asked to spot the prosodic group according to their perception not by measuring the duration of pause. Our purpose is to find the correlation between perception and the prosodic parameters.

2. B2, B3:

After marking the B4 and B5, a paragraph was segmented into many breath groups. The transcribers were asked to detect irregular boundaries within a breath group. The perceived boundaries may be caused by sudden changes in pitch, duration and energy, or it may be caused by the insertion of short pause. The boundaries that are perceived by the pause are marked as B3, and the others are marked as B2.

3. B0:

In our design, we also intend to spot the reduced syllables in contraction, a phenomenon that occurs frequently in spontaneous speech. (However, our transcription showed that the collected read speech corpus almost does not contain such examples. One reason could be the somewhat careful speech style of our informants.)

Table 2 is the comparison of the break indices labeled by two transcribers. Statistical analyses of the pauses is shown in Table 3. The left panel of Table 2 represents independent labeling results of the proposed criteria; the right panel represents the labeling results of the same set of data after the transcribers compared notes of criteria used. We find while consistency between transcribers increases after discussion, the types of less identifiable categories still maintains. Most of the inconsistency occurred in B1 vs. B2 and B4 vs. B5. A total of 204 boundaries were labeled as B1 by transcriber A, but labeled as B2 by transcriber B. Furthermore, 48 boundaries were labeled as B5 by transcriber A, but labeled as B4 by transcriber B. This could mean that transcriber A is more sensitive to global prosodic changes and transcriber B is more sensitive to finer prosodic changes. From the statistical analysis in Table 3, it is evident that consistent use of labeling criteria can be found within each transcriber, whereas their respective chosen criteria may not the same.

A B	B0	B1	B2	B3	B4	B5
B0	0	0	0	0	0	0
	na	na	na	na	na	na
B1	0	2041	114	16	2	4
	0%	93.8%	5.2%	0.7%	0.1%	0.2%
B2	0	205	394	87	2	0
	0%	29.8%	57.3%	12.6%	0.3%	0.0%
B3	0	14	80	187	45	5
	0%	4.2%	24.2%	56.5%	13.6%	1.5%
B4	0	0	1	67	163	108
	0%	0.0%	0.3%	19.8%	48.1%	31.9%
B5	0	1	0	1	7	103
	0%	0.9%	0%	0.9%	6.3%	92.0%

A B	B0	B1	B2	B3	B4	B5
B0	0	0	0	0	0	0
	na	na	na	na	na	na
B1	0	2162	83	1	0	1
	0%	96.2%	3.7%	0%	0%	0%
B2	0	204	422	30	0	0
	0%	31.1%	64.3%	4.6%	0%	0%
B3	0	5	45	330	36	3
	0%	1.2%	10.7%	78.8%	8.6%	0.7%
B4	0	0	1	46	124	48
	0%	0%	0.5%	21.0%	56.6%	21.9%
B5	0	1	0	0	2	103
	0%	0.9%	0%	0%	1.9%	97.2%

Table 2. The break indices labeled by two transcribers (A and B) before (the left) and after (the right) the exchange of notes for labeling.

	BI	B0	B1	B2	B3	B4	B5
A	Mean	0	1.6	12.5	160.4	452.3	747.5
	Std	0	0.4	0.9	7.8	15.5	14.0
B	Mean	0	3.2	14.3	143.9	541.7	757.5
	Std	0	0.8	1.4	7.4	13.5	19.7

	BI	B0	B1	B2	B3	B4	B5
A	Mean	0	1.9	16.2	243.7	623.0	793.0
	Std	0	0.4	1.4	8.5	17.9	16.4
B	Mean	0	1.5	11.2	232.3	658.3	841.7
	Std	0	0.4	0.7	7.6	15.3	31.1

Table 3. The Mean and Std (in ms) for the pause of different break indices before (the left) and after (the right) the exchange of notes for labeling.

4. CONCLUSION

We have proposed a prosodic labeling system for Mandarin in this paper. The system have combined the spirits of ToBI and JSML and modified to suit the prosodic properties of Mandarin. It was shown in the experiments that the results of two transcribers were consistent in a reasonable level. Combined with the proposed ASCII version of IPA symbols (SAMPA-T) [2], We believe that the proposed system can be adopted as the standardized version of machine-readable phonetic transcription system for Mandarin Chinese.

REFERENCES

[1] Barry, W. J. and Fourcin, A. J. 1992. Levels of labelling. *Computer Speech and Language*, 6, 1-14

[2] Tseng, C. Y. and Chou, F. C. 1998. Machine readable phonetic transcription system for Chinese dialects spoken in Taiwan. *Proceedings of the 1st Oriental COCOSDA Workshop*, 179-183

[3] Nespor, M. and Vogel, I. 1986. *Prosodic Phonology*

[4] Roach, P., Knowles, G., Varadi, T. and Arnfield, S. 1993. MARSEC: a machine-readable spoken English corpus. *Journal of the International Phonetic Association*, 23(2), 47-53

[5] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. and Hirschberg, J. 1992. ToBI: a standard for labeling English prosody. *Proceedings of the 2nd International Conference on Spoken Language Processing*, 867-870

[6] Tseng, C. Y. 1997. Prosodic group: suprasegmental characteristics of Mandarin connected speech from a speech database. 6th *International Conference on Chinese Linguistics*

[7] Sun Microsystems Inc., Java Speech Markup Language Specification, 1997, Internet WWW page, at URL:<http://java.sun.com/products/java-media/speech/forDevelopers/JSML/index.html>

[8] Lieberman, P. 1967. *Intonation, Perception, and Language*, MIT Press