J. Acoust. Soc. Jpn. (E) 20, 3 (1999)

Machine readable phonetic transcription system for Chinese dialects spoken in Taiwan

Chiu-yu Tseng and Fu-chiang Chou

Institute of Linguistics (Preparatory Office), Academia Sinica, Nankang Taipei 115, Taiwan, Republic of China

(*Received 4 August 1998*)

Since existing ASCII versions of phonetic transcription systems appear to aim at transcribing European languages only, they prove to be insufficient to accommodate syllable based tonal languages such as Chinese. An ASCII encoding of phonetic transcription system for speech database of Chinese was designed. Major characteristics of the proposed design are: 1. Though based on the principles of the International Phonetic Alphabet (IPA), the current design includes two levels of transcription, namely, segmental and prosodic. The consequence is a more elaborate system than the IPA or its equivalent. 2. The proposed system specifically aims to transcribe three major Chinese dialects spoken in Taiwan, namely, Mandarin, Taiwanese and Hakka. The consequence is a more language-dependent system rather than a general system as the IPA.

Keywords: Chinese, Tone, Transcription, Prosody

PACS number: 43. 72. −q

1. INTRODUCTION

A well-transcribed speech database is crucial to both basic speech research and the development of related speech technologies. When constructing a transcription system, there are many levels of transcription that may be considered and included in the system. For example, orthographic transcription, citation-phonemic representation, broad phonetic transcription, narrows phonetic transcription, acoustic-phonetic transcription, prosodic transcription and even representation of non-linguistic but significant phenomena.[1] The nature of our proposed system is an ASCII version of broad phonetic transcription system with additional prosodic labels.

The citation-phonemic level of transcription contains the output phoneme string derived from the orthographic form (by lexical access, by letter-to-sound rules, or both).[2] There are various alternatives to represent phonemes with symbols. 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. There exist many alphabetic equivalent systems for speech databases such as ARPABET and KLATTBET (for American English), Edinburgh's Machine Readable Phonetic Alphabet (for British English), SAMPA[3,4] (for European languages) and OGI's WORLDBET (for all languages).[5] Since we set the prerequisite of our system for three specific Chinese dialects, namely, the three major Chinese dialects spoken in Taiwan, it seems feasible to avoid possible notational complexity by keeping all symbols distinct across the intended target languages as did OGI's WORLD BET. The other reason is that each phonetic symbol in WORLDBET is represented as two ASCII characters. In order to maintain the two-character format, an extra space is required after those symbols that have only one character. While this process may be convenient to the computer, it could be somewhat less convenient to human transcribers. Therefore, we chose to adopt the SAMPA system

originally designed for major European languages, and adjusted it to a language-specific set of alphabetic phoneme symbols for the target Chinese dialects spoken in Taiwan.

Though based on the IPA, the proposed system chose not to provide symbols to represent information as would the diacritics of the IPA. The result is a less sophisticated system for segmental symbols; something we define as a broad phonetic system. That is, though the system aimed at enabling a transcriber to represent tokens of speech from our informants, the transcription is nonetheless restricted to phonetic categories rather than labeling finer phonetic details. Consequently, the transcriber would somewhat be forced to label the speech tokens with a set of phonetic symbols available without the tools to represent IPA diacritic information. Our rationale is that since orthographic text is readily available for our working speech data, we have all the information necessary to determine what the intended speech should be. The task of our transcriber is to listen to the collected speech data and make direct reference between the intended citation string from text to the actual utterance produced. In some sense, the transcription can be seen as part transcription and part mapping to text. Moreover, the designed broad phonetic representation at this stage also does not possess the capability to label common phenomena of running speech, such as place assimilation, consonant deletion and vowel reduction. We are fully aware that once running speech is incorporated into our speech database, we will need to expand the system to accommodate running speech transcription in the future. We will also need to include tonal notations in our system in the future since all three Chinese dialects possess lexical tones at the syllabic level.

Leaving tonal representations 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, and the units segmented unavoidably tend to tilt towards the particular chosen theoretical framework. 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 F_0 peaks. The former theoretical orientation, *i.e.*, the use of boundaries, resulted in approaches that used intonation categories proposed by Ref. 6) 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.[7] 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[8] is the most well-known system of this kind, and apparently works 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. We chose to adopt a ToBI-like framework for 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.

The paper is organized as followed: Section 2 describes the reported broad phonetic transcription.; Section 3 the prosodic transcription. The experiments are described in Section 4.

2. BROAD PHONETIC TRANSCRIPTION

To establish a broad phonetic transcription system for a specific language, one needs to first define the phoneme set and corresponding symbols. The most likely choice would be the IPA symbols, which has long been established as the system that is sufficiently equipped to represent sound systems of known languages of the world. Unfortunately, there is no uniformly implemented coding standard for the IPA symbols until efforts such as ISO 10646/Unicode is accepted internationally. The reported system is designed to function with the SAMPA guidelines, but tailored to suit the target languages mentioned.

SAMPA is a phonetic transcription coding that uses normal ASCII characters as replacements for IPA symbols. Mapping of the IPA symbols onto ASCII codes was set up, using up to range 37–126, 7-bit printable ASCII characters only. Originally designed as a phonemic/broad-phonetic transcription system for European languages, both SAMPA

and subsequently proposed X-SAMPA (Extended SAMPA) represent an international collaborative basis for a standard machine-readable encoding of phonetic notations. Guidelines for coding (mapping) of languages to be transcribed are also available. The reported system was designed by following these guidelines, especially maintaining the feature that the resulted transcription is uniquely parsable as specified in SAMPA. As SAMPA specified, we also observed the IPA transcription convention of leaving no space between a string of successive symbols unless syllable boundaries occur. However, we note that the basic form of SAMPA was designed as a phonemic or broad-phonetic transcription system at the segmental level. That leaves the tonal and prosodic notations inadequate for our purpose. We therefore designed a parallel set of prosodic notations on separate levels of representation, which we believe enhances the function and capacity significant for the proposed system. More details of the design of the prosodic aspects of the proposed system will be discussed in later sections of this paper. As we intend to expand the system the future, we also hope that our current system will be adopted as the standardized international ASCII system for the three Chinese target languages.

The principles of SAMPA are quite simple : all IPA symbols that coincide with lower-case letters of a regular English alphabet keyboard remain the same ; all other symbols are re-coded within the ASCII range 37-126. In the process of developing our system, we began by considering Mandarin at first, then expanded it by adding extra symbols for Taiwanese and Hakka. The mapping is described as follows :

Consonants :

We began our design with Mandarin, the official Chinese spoken language, and then expanded it to include Taiwanese and Hakka. As a result, the order of target languages affects the outcome of the system. The consonants of Mandarin Chinese are traditionally considered to comprise 18 obstruents (6 plosives, 6 affricates and 6 fricatives) and 3 sonorants (2 nasals, 1 liquid). Unlike English in which the obstruents are classified in pairs differing in both voicing and aspiration, all Mandarin obstruents are voiceless differing only in aspiration, except for one voiced retroflex fricative. In order to accommodate the SAMPA principle to adapt the keyboard and to use one symbol for minimal pairs, we adopt the "b" "p" notation to represent the voiceless bilabial pair in Mandarin. Voicing is therefore not represented. All other notations for the remaining obstruents follow the same principle. This somewhat arbitrary use of notation is maintained when the system is expanded to include Taiwanese and Hakka where voicing contrasts also exist. That is, though the voiced counterparts do exist for bilabials and velars in Taiwanese and Hakka, making it a three-way contrast instead, "p" and "b" still represent the voiceless bilabial pair which differs in aspiration while an upper case "B" is used to represent the voiced bilabial. The same rule also applies to velar "G" and postalveolar affricate "DZ".

Another kind of contrast that exists in fricatives and affricates in Mandarin is "retroflex". The notation "`" is used to represent "retroflex". For the voiceless alveolo-palatal fricative which in IPA is a curly-tailed c, we propose it to be transcribed as "z\" ; its aspirated counterpart "s\". This use of notation "\" follows the principles of extension in SAMPA. However, the original contrast of voicing between "s" and "z" is replaced to represent aspiration in our system. Again, this principle is constant throughout our system. As for the only voiced consonant in Mandarin, namely, the voiced retroflex fricative, the upper case "Z`" is used to denote voicing. This is also constant throughout our system. Table 1 shows the mapping of consonants of all three Chinese dialects.

Vowels :

There are eight vowels in Mandarin. Seven of them already exist in SAMPA. The only not-so-common one is an apical vowel. The notation "U" is used to represent the apical vowel. This vowel occurs only after 3 fricatives (s, s`, Z`) and 4 affricates (dz, dz`, ts, ts`), and does not occur with other consonants. When the consonant is retroflexed, this vowel is also retroflexed and is represented by adding retroflex after the apical vowel as in "U`". As for Taiwanese and Hakka, three more notations were added. The notation "~" represents nasalization of the preceding vowel ; the notation "?" glottal stop ; and the notation "}" the preceding stop consonants unreleased as in "p}, t}, k}." Other features in Taiwanese include the syllabic bilabials and velar nasals. The syllabification is represented by adding a "^" after these nasal consonant as in "m^, n^" respectively. Table 2 shows the mapping of vowels of all three

Table 1 The mapping between IPA and SAMPA-T for consonants. (T mean character in Taiwanese, H means Hakka)

IPA	SAMPA-T	examples		
		character	syllable	meaning
p	b	爆	bau4	to explode
b	B	肉(T)	Ba?4	meat
p^h	p	泡	pau4	bubble
t	d	倒	dau4	to pour
t^h	t	套	tau4	cover over
k	g	告	gau4	to tell
g	G	阮(T)	Gun2	we
k^h	k	銬	kau4	handcuff
f	f	斧	fu3	ax
x	h	虎	hu3	tiger
v	v	衛(H)	vi5	to guard
ʂ	s	速	su4	quick
ʂ	s`	樹	s`u4	tree
ɕ	s\	細	s\i4	thin, fine
ʐ	Z`	入	Z`u4	to enter
dz	DZ	子(T)	DZi2	small pieces
tɕ	dz\	雞	dz\i1	chicken
$tɕ^h$	ts\	七	ts\i1	seven
ts	dz	租	dzu1	to rent
tʂ	dz`	豬	dz`u1	pig
ts^h	ts	粗	tsu1	rough, big
$tʂ^h$	ts`	出	ts`u1	to exit
m	m	木	mu4	wood
n	n	怒	nu4	anger
ŋ	N	迎(T)	Nia5	to meet
ɳ	J	你(H)	Ji2	you
l	l	錄	lu4	to record

Table 2 The mapping between IPA and SAMPA-T for vowels and some special endings. (T mean character in Taiwanese, H means Hakka. SE_1: finals with a vowel-nasalizing ending, SE_2: finals with a stop ending, SE_3: final with a glottal stop ending)

IPA	SAMPA-T	examples		
		character	syllable	meaning
i	i	椅	i3	chair
u	u	五	u3	five
y	y	雨	y3	rain
a	a	啞	ia3	mute
o	o	我	uo3	I
ɛ	E	也	iE3	also, too
ə	@	餓	@4	hungry
ɚ	@`	二	@`4	two
ɯ	U	絲	sU1	silk
	U`	詩	s`U`1	poem
ai	ai	百	bai3	hundred
ei	ei	北	bei3	north
au	au	咬	iau3	to bite
ou	ou	有	iou3	to have
n̂g	n^	黃(T)	n^5	yellow
m̄	m^	梅(T)	m^5	plum
ŋ	N	洋(T)	iaN7	ocean
SE_1	~	贏(T)	ia~7	to win
SE_2	}	盒(T)	ap}8	box
		力(T)	lat}8	force
		六(T)	lak}8	six
SE_3	?	蠟(T)	la?8	wax

Chinese dialects.

The system proposed above can be used for two kinds of labeling. The basic form is to form a citation-phonemic string in a syllabic unit to represent the pronunciation of each syllable. The other form is to use the symbols for the broad phonetic transcription in segmental labeling. Additionally, we designed a coding system for syllables used in these dialects. Each syllable is divided into three sub-units, namely, initial, medial and final. Each sub-unit has a unique code (I_c, M_c, F_c) and the code of a syllable (Syl_c) can be calculated according to this formula:

$$Syl_c = I_c \cdot 1000 + M_c \cdot 10 + F_c$$

With this scheme, each syllable can be represented by an unsigned short integer (0～32, 767), a basic unit in computer. This coding is straightforward and quite self-explanatory. The contextual information can also be extracted from the code directly. This is very useful for the programming in speech research. Table 3 is the coding for each sub-unit.

For tonal representations, we adopt the Chinese conventional use of numerical system in the phonemic level which may appear to be arbitrary for

Table 3 Coding for initial sub-units, medial sub-units and final sub-units.

initial	#	g	G	k	b	B	p	d	t	ts
	0	1	2	3	4	5	6	7	8	9
sub-	ts`	dz	DZ	dz`	dz\	ts\	s\	s	s`	Z`
units	10	11	12	13	14	15	16	17	18	19
	v	f	h	m	n	J	N	l		
	20	21	22	23	24	25	26	27		
	i	iu	iE	ieu	io	ioi	iou	ia	iau	y
	0	1	2	3	4	5	6	7	8	10
medial	yE	ya	U	U`	u	ui	uE	uei	u@	uo
sub-	11	12	20	21	30	31	32	33	34	35
units	ua	uai	E	ei	eu	@	@`	o	oi	ou
	36	37	40	41	42	50	51	60	61	62
	oE	oa	oai	O	a	ai	au	m^	n^	N^
	63	64	65	66	70	71	72	90	91	92
final	#	m	n	N	p}	t}	k}	h}	h~	~
sub-	0	1	2	3	4	5	6	7	8	9
units										

users who are unfamiliar with tones of Chinese. But until more dialects are included, the numerical system seems most easily acceptable and therefore should be adequate for the time being.

3. PROSODIC TRANSCRIPTION

The prosodic transcription of our system is represented on a separate level following ToBI-like notations. ToBI (for Tones and Breaks Indices) is a system for transcribing intonation patterns and other aspects of the prosody of English utterances. It was devised by a group of speech scientists from various disciplines (electrical engineering, psychology, linguistics, *etc.*) in pursuit of diverse research purposes. Various technological goals that included sharing prosodically transcribed databases across research sites also desired a commonly agreed-upon standard of prosodic elements. 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.[9]

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, both of which involve deliberate manipulation of fundamental frequency patterns and therefore both of which are pitch events, is not only more complex but also not well understood, yet. It is thus more difficult to construct notations like the tone tier in ToBI for these languages. We proposed to label the speech data in more detail at the prosodic domain while leaving the tonal aspects for the time being. Our reason was again the fact that text for our speech data was readily available for reference. When the break indices of an utterance is labeled, the volume, rate, pitch level and pitch range of each prosodic unit can be calculated and labeled in an automatic way according to the physical signals. The perceived changes of above parameters can be manually added on when the transcriber has noted the significant changes in the utterances. Another subjective added-on parameter is the level of emphasis. This level of information is perceptual in nature and may or may not correspond directly to the physical signals. A transcriber is asked to decide the emphasis level of the speech tokens based on subjective perception.

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)[10] 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

Table 4 The meaning of levels for the prosodic tags.

Tags \ Level	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 slow	low	normal	high	very high	
PITCH	very slow	low	normal	high	very high	
RANGE	very small	small	normal	large	very large	

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\i4 〈/EMP〉 h@n3 hau3 〈BREAK level=4/〉

The example shows that "dz\in1 tien1" is a minor prosodic phrase and "tien1 ts\i4" 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 4.

4. 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 Fig. 1.

The proposed segmental system appears sufficient for transcribing speech variations. Major variation of actual speech comes from the fact that a large number of Mandarin speakers in Taiwan are bilingual. They are native speakers of Taiwanese who acquired Mandarin to the point that Mandarin could be their stronger language. The carry-over of their Taiwanese phonology to Mandarin could be quite pronounced; their Mandarin speech contains segmental and suprasegmental variations that Mandarin phonology would not accommodate. One of our aims was to be able to transcribe these variations that definitely occurred in our speech data and subsequently collapsing them in the development of our labeling tools for Mandarin. Our system was designed to represent the presence and absence of the features that correspond to these variations. For example, there are three contrastive pairs of fricative in Mandarin that are distinguished by the presence or absence of retroflex. But for native speakers of the Taiwanese where no retroflex occur, retroflex usually does not occur when they speak Mandarin. This phenomenon can be easily distinguished by the presence of symbol "ˋ" which represents the feature retroflex, for examples, "dz" *vs.* "dzˋ", "ts" *vs.* "tsˋ", and "s" *vs.* "sˋ". Another similar example is the phenomenon of rounding. Native speakers of Taiwanese tend to pronounce the "uo" or "ou" combi-

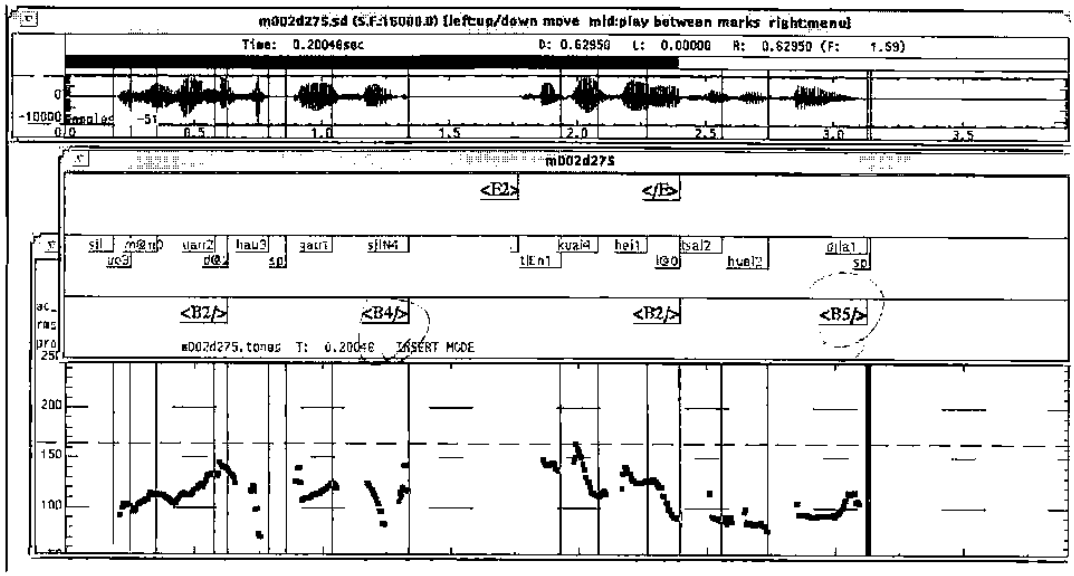


Fig. 1 An example of the segmental and prosodic transcription.

nation with less rounding on the "u" part. To transcribe this variation would be to drop the "u" in the transcription. For examples, intended "uo" could be realized as "o", intended "iou" realized as "io". While conventional syllable based coding system uses all but intended syllables as coding units, in the case of Mandarin a total of 408 syllables only, such realized variations in actual speech could not be accounted for. Our proposed transcription system makes it possible to include actual speech variations to be accounted for in the recognition process. For example, the syllable "sua1" is not a legitimate syllable in Mandarin phonology, but native speakers of Taiwanese pronounce it as an un-retroflexed substitution of "s`ua1" (brush). That is, the syllable "sua1" could in fact be an actual realization of an intended "s`ua1" (brush). Our coding scheme provided different codes for these variations, for example 18360 for "s`ua1" and 17360 for "sua1." These codes are internal representation of Mandarin syllables in the computer and therefore will not cause any extra burden on the human users. This ability to represent phonetic variation in actual speech is very useful for speech recognition and synthesis. For example, in order for the computer to better recognize Mandarin speech, we are able to collapse the actual realization of speech with their intended target to boost the robustness of our recognition program.

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.[11] 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 5 is the comparison of the break indices labeled by two transcribers. Statistical analyses of the pauses are shown in Table 6. The left panel of Table 5 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

Table 5 The break indices labeled by two transcribers (A and B) before (the left) and after (the right) the exchange of notes for labeling.

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%
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%	0.3%	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 6 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.

	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

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 6, it is evident that consistent use of labeling criteria can be found within each transcriber, whereas their respective chosen criteria may not the same.

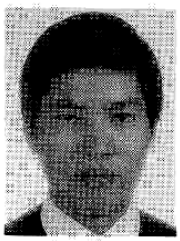
5. CONCLUSION

In this paper, we proposed a machine-readable broad phonetic transcription system for three major Chinese dialects spoken in Taiwan, namely, Mandarin, Taiwanese and Hakka. In the course of collecting speech data and developing tools at the same time, one of our major goals was to construct a well transcribed speech databases for speech research of these languages. We have deigned a set of ASCII symbols following the guideline of SAMPA to represent the phonemes used in these languages and can be used for broad-phonetic transcription. The system also included transcription capacity at the prosodic level, combining the spirits of ToBI and JSML and modified to suit these tonal languages. Though the system is designed to transcribe three Chinese dialects as specified, we are currently testing it on our phonetic and prosody oriented speech database for Mandarin Chinese, and will further test it on Taiwanese and Hakka. We believe that the proposed system can be adopted as the standardized ASCII version of machine-readable phonetic transcription system for Chinese.

REFERENCES

1) D. Gibbon, R. Moore, and R. Winski, *Handbook of Standards and Resources for Spoken Language Systems* (Mouton de Gruyter, Berlin, 1997).
2) W. J. Barry and A. J. Fourcin, "Levels of labelling," Comput. Speech Lang. **6**, 1-14 (1992).
3) J. C. Wells, "Computer-coded phonemic notation of individual languages of the European Community," J. Int. Phonet. Assoc. **19**, 32-54 (1989).
4) J. C. Wells, *Computer-coding the IPA : a Proposed Extension of SAMPA*, Internet WWW page, at URL :<http://www.phon.ucl.ac.uk/home/sampa/x-sampa.htm> (1998).
5) J. L. Hieronymus, "ASCII phonetic symbols for the world's languages : Worldbet," AT&T Bell Labs. Tech. Memo (1994).
6) M. Nespor and I. Vogel, *Prosodic Phonology* (Foris, Dordrecht, 1986).
7) P. Roach, G. Knowles, T. Varadi, and S. Arnfield, "MARSEC: A machine-readable spoken English corpus," J. Int. Phonet. Assoc. **23**(2), 47-53 (1993).
8) K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI : A standard for labeling English prosody," Proc. ICSLP 92, 867-870 (1992).
9) C. Tseng "Prosodic group : Suprasegmental characteristics of Mandarin connected speech from a speech database," 6th Int. Conf. Chinese Linguistics (1997).
10) Sun Microsystems Inc., *Java Speech Markup Language Specification*, Internet WWW page, at URL : <http://java.sun.com/products/java-media/speech/forDevelopers/JSML/index.html> (1997).
11) P. Lieberman, *Intonation, Perception, and Language* (MIT Press, Cambridge, Mass., 1967).

Chiu-yu Tseng received Ph.D. degree in linguistics from Brown University, Providence, RI, in 1982. She is currently a Research Fellow at the Institute of Linguistics (Preparatory Office), Academia Sinica, Taipei, Taiwan, R. O. C. Her research areas include the phonetic aspects of the synthesis and recognition of Mandarin Chinese, acoustic phonetic analysis of Mandarin Chinese and Taiwanese. Her current research interests are the design of speech corpus and investigation of prosody in Mandarin Chinese.



Fu-chiang Chou received the B.S. and Ph.D. degrees in electrical engineering from National Taiwan University, Taipei, Taiwan, Republic of China, in 1989 and 1999, respectively. He is currently a Postdoctoral Fellow at the Institute of Linguistics (Preparatory Office), Academia Sinica, Taipei, Taiwan, R.O. C. His current research interests are in the area of digital speech processing with special interests on the problems of automatic labeling of speech corpus and Mandarin speech synthesis.