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

ISCA & IEEE Workshop on Spontaneous
Speech Processing and Recognition
Tokyo Institute of Technology, Tokyo, Japan
April 13-16, 2003

ANNOTATION CONVENTIONS AND CORPUS DESIGN IN THE INVESTIGATION OF SPONTANEOUS SPEECH PROSODY IN TAIWANESE

Shu-hui Peng and Mary E. Beckman

National University of Kaohsiung (Taiwan) and Ohio State University (USA)

ABSTRACT

Understanding how intonational phrasing and focal prominence interact with lexically specified tone patterns is one of several problems in the investigation of speech processing in Chinese languages that cannot be addressed fully with read speech alone. This paper explores such problems for Taiwanese, one of the major languages in the southern Min dialect group. It outlines what is known about Taiwanese prosody and describes prosodic annotation conventions currently under development. It illustrates the problems with examples from a corpus of spontaneous speech dialogues and read versions of sentences extracted from the dialogue transcripts.

1. INTRODUCTION

Spontaneous speech recognition of Chinese is challenging because discourse-level factors such as focus of attention affect the realization of both tones and segments. In many Chinese languages, this means that a word can have different lexical tones in different discourse contexts. Taiwanese is a good language to use to illustrate these challenges, for several reasons. First, it has a very complex system of lexical tone changes whereby every content word has at least two alternative tone patterns, depending on its position within an intermediate level of prosodic grouping that is analogous to the accentual phrase in Japanese. Second, it has grammatical particles, such as the genitive marker /e/ which functions like Mandarin *-de* and Japanese *–no*, which often are realized as if they have no intrinsic lexical tone specifications, yet can affect the positionally conditioned lexical tonal alternations on neighboring syllables. Finally, it also has many sentence-level discourse particles, which vary in tone to signal the pragmatic relationship between the sentence and its context. Each of these three phenomena complicates the description of the Taiwanese prosodic system, and none of them can be understood just by examining phrases and sentences read in isolation.

In this paper, we describe a speech database which we designed and recorded in order to compare the prosody of utterances produced in spontaneous task-based dialogues with lexically identical utterances produced by the same speakers reading a list of transcribed sentences extracted from the dialogues. We use example utterances from this database in reviewing what is known about the Taiwanese prosodic system, before outlining the prosodic annotation system that we are developing for the language. Finally, we list the problems and outstanding questions that are being highlighted by developing these conventions and applying them to spontaneous speech.

2. THE SHAPES GAME DATABASE

The Shapes Game Database is a corpus containing two distinct speech styles. The primary data are spontaneous dialogues produced by pairs of Taiwanese native speakers participating in the Shapes Game. This game was designed originally by Y. J. Fon, who has used it with speakers of standard Japanese, American English, Guoyu (Taiwan standard Mandarin), and Putonghua (PRC standard Mandarin). Each participant in a game has a game board, which is a 2x3 grid containing six game pieces, and a stack of extra pieces to substitute for any of the pieces on his or her board. Each game piece is a distinct colored shape, such as a green triangle, a pink triangle, or a pink square. The speakers communicate over headphones and microphone to negotiate the most economical way to make their game boards match, replacing one piece at a time in alternate moves. Points are lost for each move, including each piece moved from the top of the stack to get at a piece that is lower in the stack. The task is meant to elicit natural variation in focus of attention on the modifying adjective versus the head noun versus the phrase as a whole. The database also includes read productions by the same speakers of sentences selected from transcripts of the dialogues.

Three pairs of speakers have been recorded to date. Each participant's utterances were recorded onto separate tracks of a DAT tape. We then made an orthographic transcription of each dialogue using Chinese characters, and created a list of sentences selected from the transcripts for the same participants to read. (Although a sizeable minority of Taiwanese speakers are literate in Romaji, the Church romanization system [1], secular Taiwanese is not normally a written language. However, nearly all younger Taiwanese speakers are bilingual and literate in Guoyu Mandarin, and most Taiwanese words and phrases can be written in a fairly unambiguous way by using the character for a semantically related Mandarin Chinese morpheme to write each component syllable.)

For each spontaneous utterance in the dialogues and for each read utterance from the sentence list, we annotated the surface tone pattern, and are currently in the process of making a more complete annotation of the prosodic structure, using the TW-ToBI annotation conventions (described in Section 4). The utterances transcribed so far can be used to illustrate some of the prosodic phenomena that have been studied in read speech (reviewed in Section 3) and to help us understand what further questions need to be answered before TW-ToBI can be finalized (see Section 5).

3. TAIWANESE PROSODIC STRUCTURE

Prosody in Taiwanese is a hierarchical structure that emphasizes grouping rather than metrical prominence. Prosodic constituents at each level of this hierarchy are marked by structural constraints on the distribution of segments and tones relative to the constituent edges, as well as by phonetic cues such as initial pitch expansion, medial consonant weakening, and final lengthening. This section describes the constraints and cues that mark the edges of the three best-motivated constituent types.

The smallest of these constituents is the syllable. As in all other Chinese languages, syllable structure is very constrained. The obligatory nucleus is a simple vowel, a complex vowel or a syllabic nasal, as in the three words /li^{51}/ 'you', /gua^{51}/ 'I', and /mŋ^{33}/ 'ask' in Fig. 1. As these words show, all three nucleus types can be preceded by an onset consonant. Syllables also can be closed by a coda consonant, from the set of voiceless unaspirated stops /p, t, k, ʔ/ and nasals /m, n, ŋ/, as in the last two syllables of /sã^{55}.kak^{21}.hiŋ^{24}/ 'triangle' in Fig. 2. None of these codas other than /ʔ/ can occur after a syllabic nasal. Another co-occurrence restriction involves the onset. The consonants /b, l, g/, which are a voiced series in contrast to voiceless unaspirated /p, t, k/ and aspirated /p^{h}, t^{h}, k^{h}/, cannot occur before syllabic nasals or nasal vowels, whereas /m, n, ŋ/ cannot occur as onsets before oral vowels.

Figs. 3-4 illustrate a consequence of these two distributional constraints for connected speech. In Fig. 4, the particle /e/ occurs in /tan^{51}.e/ 'wait' and /ts^{h}ẽ^{55}.sik^{21}.e/ 'green'. In both cases, the particle has fused closely with the preceding root, as indicated by the resyllabification of the coda of the syllable before /e/. Resyllabifying coda /n/ nasalizes the following vowel (/tan^{51}.e/ → [ta^{51}.nẽ]) whereas resyllabifying /k/ results in a voiced onset consonant (/ts^{h}ẽ^{55}.sik^{21}.e/ → [ts^{h}ẽ^{33}.si^{21}.ge]). In Fig. 3, similarly, coda /t/ resyllabifies as onset [l] in the phrase /het^{21}.e/ → [he^{53}.le] 'that one'; hence our identification of /l/ as the dental consonant corresponding to /b/ and /g/. Resyllabification seems most characteristic of sequences involving syllables such as the particle /e/, which seem to surface typically with no tonal specification. That is, resyllabification seems to be part of a complex of features of "reduction" which include loss of lexical tonal contrast.

In general, the distribution of lexical tone is another diagnostic for the syllables. That is, while there are a few morphemes such as the particle /e/, which typically surface with what we might call a "neutral tone", adopting the term that describes the unstressed second syllables of Mandarin words such as *dōu-fu* 'tofu' and *dōng-xi* 'thing', there are very few words analogous to these two Mandarin words, where the toneless syllable cannot be identified as a derivational particle. The word /kĩã^{55}.si/ 'scared to death', which minimally contrasts with /kĩã^{55}.si^{51}/ 'scared of dying', is one of the few examples that we can think of. Thus, counting the number of lexical tone specifications can differentiate between a diphthong and a sequence of two syllables, and this is why we can be sure that /gua^{51}/ 'I' is one syllable and not two.

Fig. 5 lists the lexical tone categories, and also describes the tonal alternations (tone sandhi rules) which affect their realization in compound words and phrases. For example, the word /sã^{55}/ 'three', which has the high level tone, is the first morpheme in 'triangle', where it is realized with the mid level tone even in a very careful deliberate read speech style; that is, /sã^{55}.kak^{21}.hiŋ^{24}/ → [sã^{33}.kak^{51}.hiŋ^{24}]. In this case, tone sandhi acts as a prosodic marker of compound word formation. However, sandhi tones do not occur just in compound words. The word /gua^{51}/ 'I' which is realized with its 'base' tone value in the spontaneous speech utterance jts57 in Fig. 1 is realized with its sandhi tone value [gua^{55}] in the read version of this sentence in the same figure and in the spontaneous speech utterance in Fig. 3. In neither case is it plausible to analyze the pronoun as being part of a compound word with the following verb. Rather, here the distribution of sandhi tones indicates non-final position in a prosodic constituent above the syllable, which Chen [2] called the tone sandhi group (abbreviated TSG or TG). As Chen put it, "each TG is punctuated, as it were, by the diagnostic presence of a base tone" [p. 114].

Other phonetic markers of the tone sandhi group that have been found in studies of read speech include a small amount of final lengthening and a somewhat expanded pitch range on the final (base tone) syllable, resulting in a steeper fall for tone 51 and a larger rise for tone 24 [3]. However, these effects are not consistent across speakers and even for speakers who do show them, they are not as large as the effects on duration and pitch range of utterance position. All five speakers in [3] showed longer durations and expanded pitch range on syllables with base tone at the ends of utterances. In the Shapes Game Database, the utterance-level effects documented in [3] are also seen utterance medially at the boundaries of the intonation phrase (IP), a prosodic constituent posited by Chen [2] as well as by subsequent researchers.

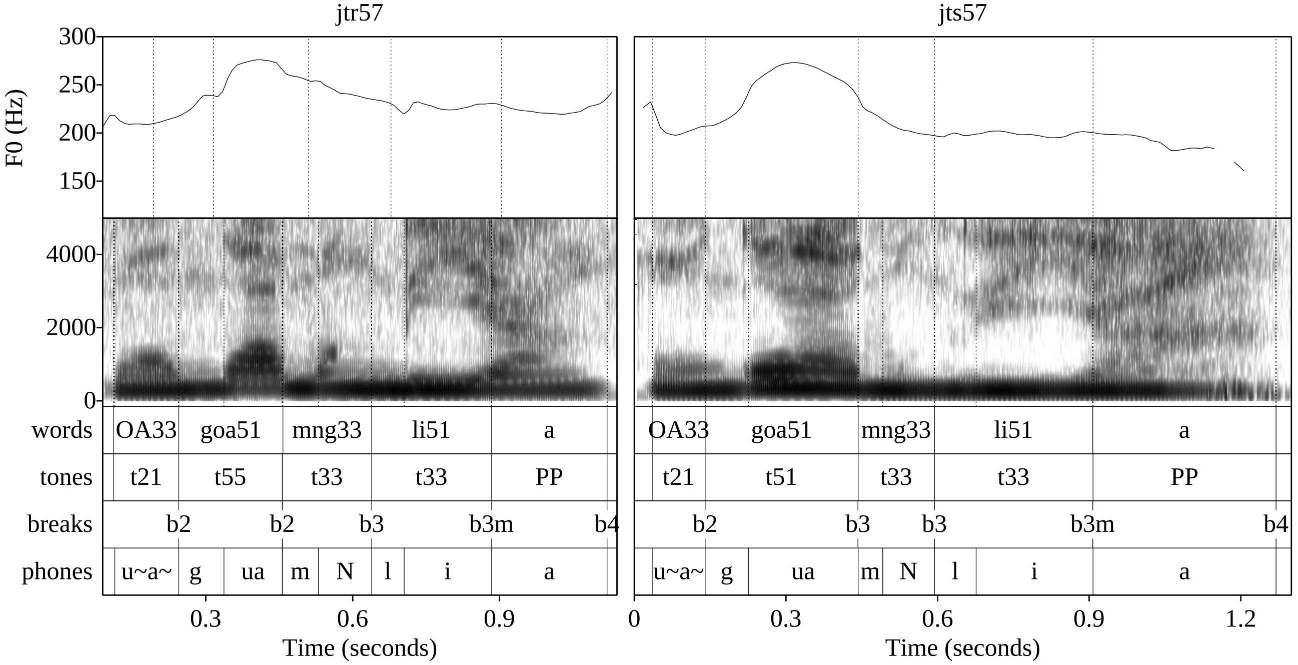


Figure 1. Fundamental frequency contours, spectrograms, and partial TW-ToBI transcriptions of the read version (left) and original spontaneous utterance (right) of sentence jt57 /ũã^{33} ɡua^{51} mŋ̩^{33} li^{51} a/ ‘It’s my turn to ask you.’ (literally: ‘switch I ask you Prt’). The context for the spontaneous speech utterance is that JT’s dialogue partner has monopolized the task exchange with a long sequence of questions about what game pieces she has showing on her board and stack.

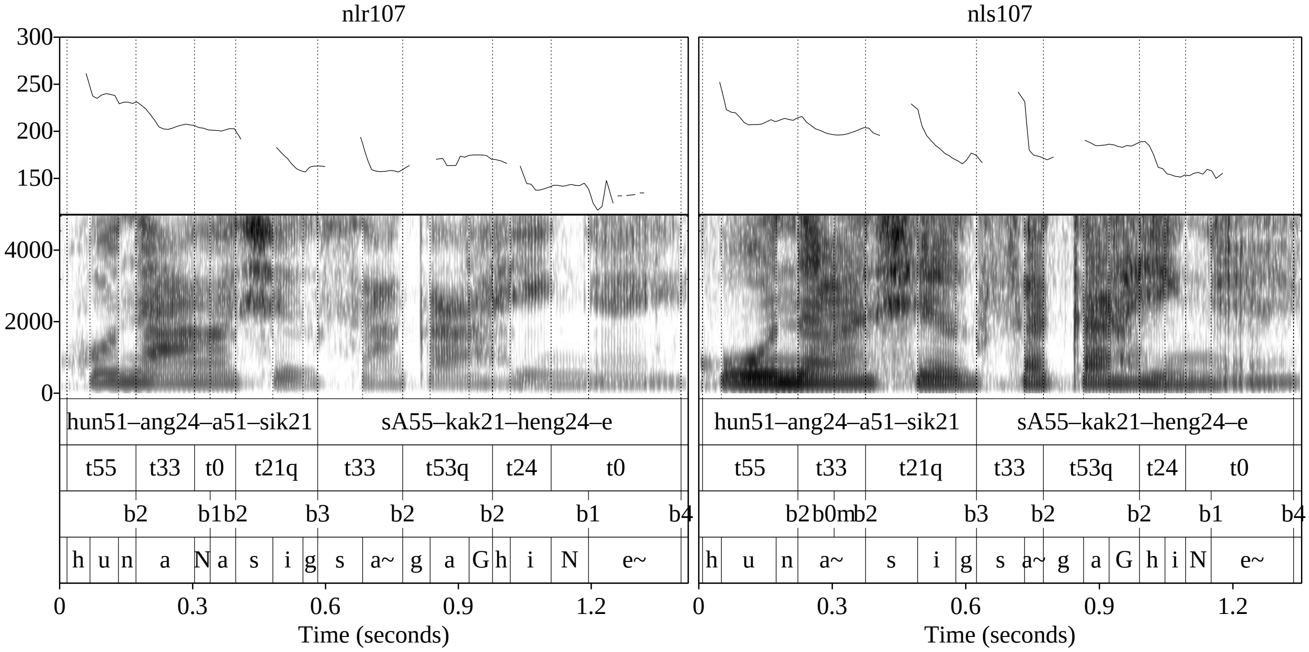


Figure 2. Read version (left) and original spontaneous utterance (right) of nl107 /hun^{51}.aŋ^{24}.a^{51}.sik^{21} sã^{55}.kak^{21}.hiŋ^{24} e/ ‘It’s a pink triangle.’ (literally: ‘pink triangle Prt’). Context: NL is drawing game pieces from her stack to try to find a match for any piece on her partner’s game board.

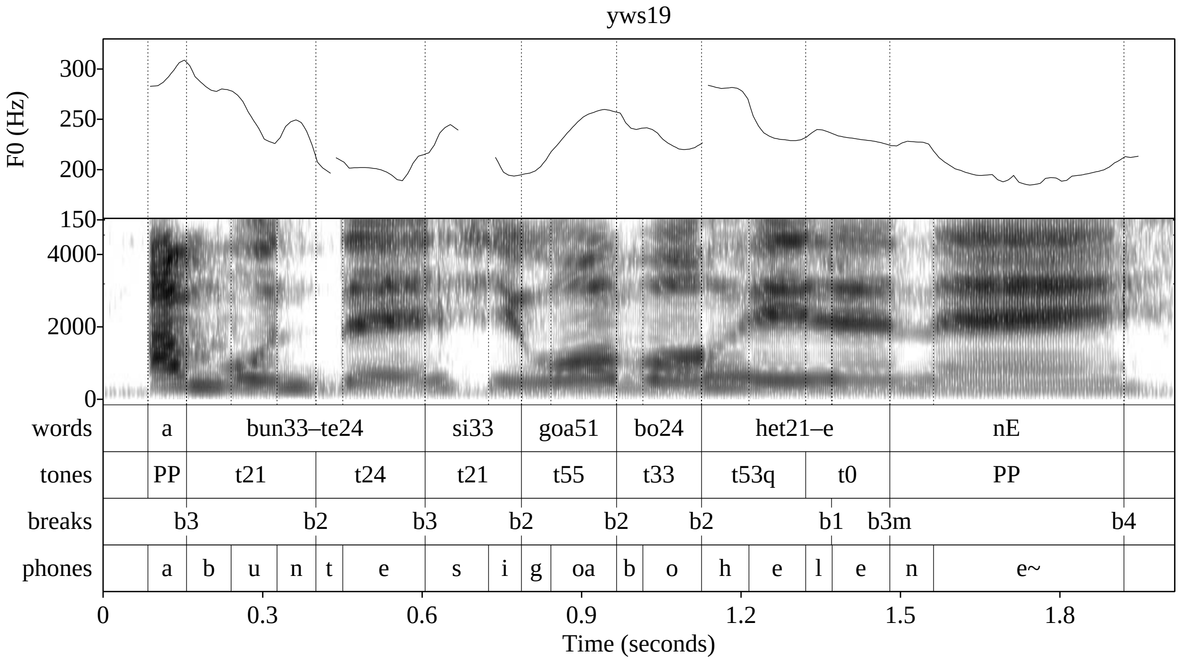


Figure 3. Original spontaneous utterance of yw19 /a bun^{33}.te^{24} si^{33} goa^{51} bo^{24} het^{21} e nẽ/ 'Well, the problem is I don't have that piece.' (literally: 'Prt problem be I not that-Prt Prt'). Context: YW's dialogue partner has misconstrued her seeming lack of cooperation and explained the game rules to her once again.

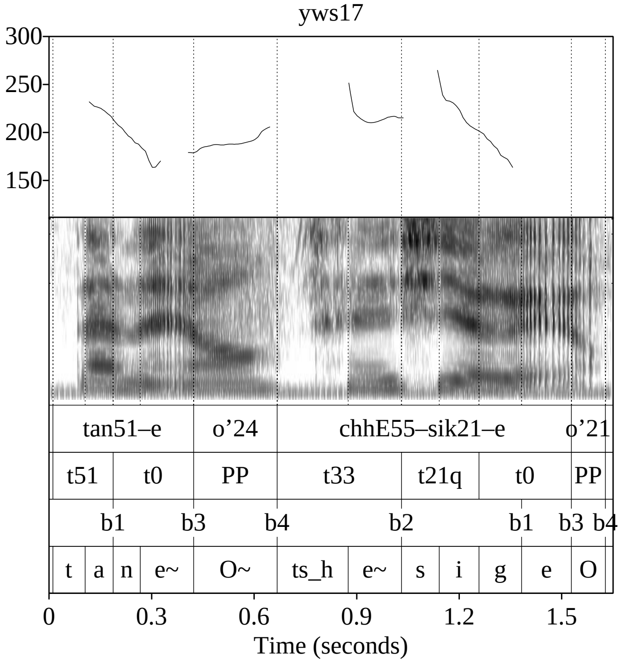


Figure 4. Original spontaneous utterance of yw17 /tan^{51}.e ɔ^{24} ts^{h}ẽ^{55}.sik^{21}.e ɔ^{21}/ 'Wait! You said a green one?' (literally: 'wait-Prt Prt green-Prt Prt'). Context: YW searches her board for a match to the game piece that her dialogue partner has just drawn from his stack and described to her.

For example, later read speech studies show domain-initial "strengthening" effects such as longer VOT values for IP-initial /p^{h}, t^{h}, k^{h}/ [4] and larger electropalatographic contact area in IP-initial dental consonants [5]. In the Shapes Game Databases, the IP also is the domain of more global pitch trends, such as the gradual uptrend on the last TSG in some types of questions, as contrasted to an overall downtrend and marked final lowering in declarative utterances that are final in their discourse segments. This contrast is illustrated by Fig. 7, where the original spontaneous utterance is a question, but the read version is produced with citation form intonation. The same contrast can be seen by looking at the overall pitch trend on the last TSG /sã^{55}-kak^{21}-hiŋ^{24}/ 'triangle' in utterance dhs65 in Fig. 6 (a question) as compared to nls107 in Fig. 2 (a turn-final statement). Utterance yws17 in Fig. 4 shows that a similar uptrend can function as a continuation rise at an utterance-medial IP boundary.

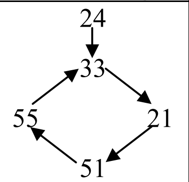
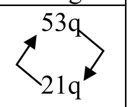
open syllables			checked syllables		
tones		sandhi changes	tones		changes
High	55		High	53q	
Mid	33		Low	21q	
Low	21				
Rising	24				
Falling	51				

Figure 5. Lexical tones and tonal alternations.

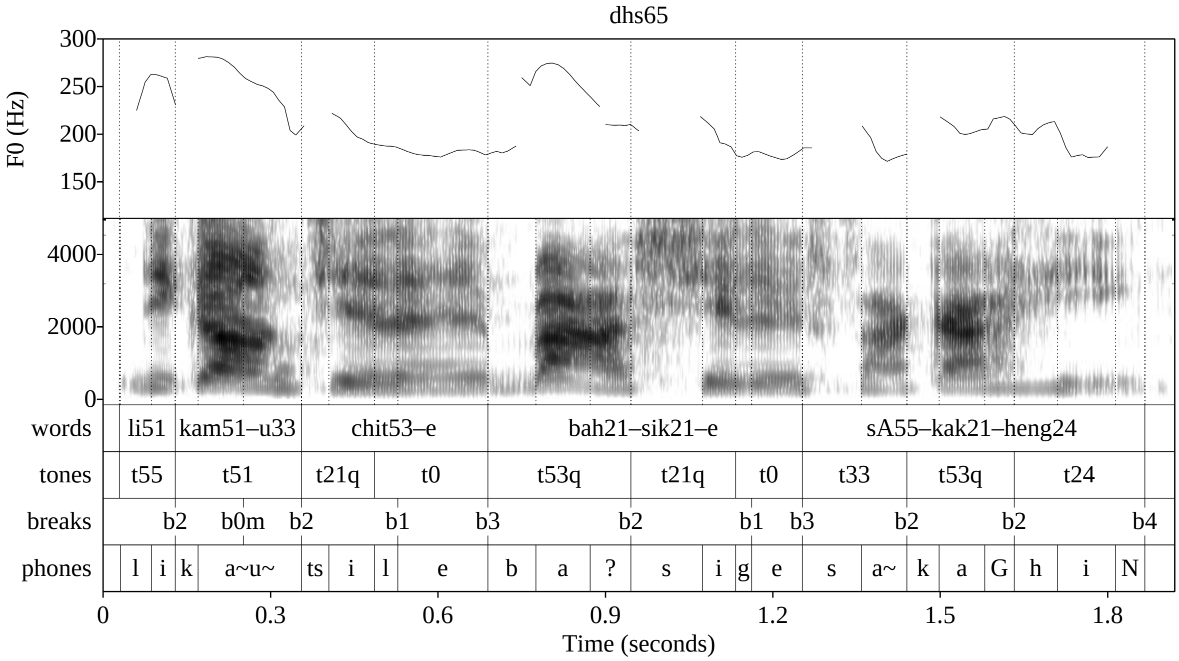


Figure 6. Original spontaneous utterance of dh65 /li^{51}.kam^{51}.u^{33} chit^{53}.e baʔ^{21}.sik^{21}.e sã^{55}.kak^{21}.hiŋ^{24}/ 'Do you have an orange triangle?' (literally: 'you have this-Prt skin-color-Prt triangle.')

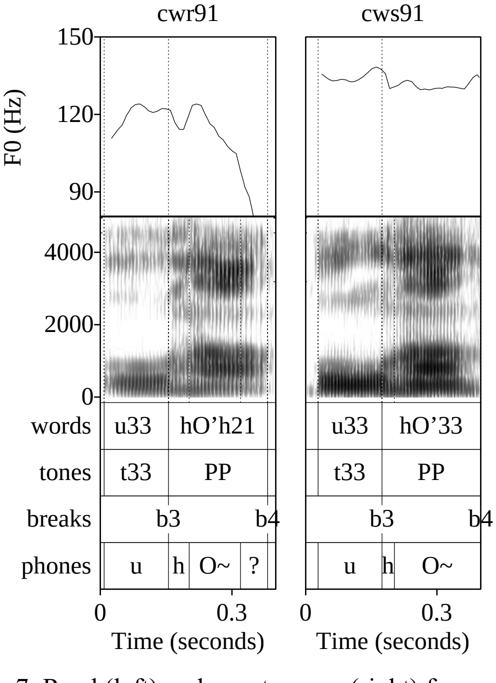


Figure 7. Read (left) and spontaneous (right) for cw91 /u^{33} hɔ̃/ 'You have it, right?' (literally: 'have Prt.')

4. THE TW-ToBI ANNOTATION CONVENTIONS

These three constituents — syllable, TSG, and IP — are the ones recognized currently in the TW-ToBI (Taiwanese Tones and Break Indices) annotation system. This system is intended to provide a way of tagging known prosodic properties of utterances that cannot yet be extracted automatically using standard ASR techniques, such as the intonational phrasing, while not committing prematurely to analyses of less well-studied phenomena.

b4	intonation phrase boundary, either utterance finally (all figures) or medially (e.g., in utterance yws17)
b3	tone sandhi group (TSG) boundary
b3m	percept of TSG boundary without sandhi tone
b2m	base tone without percept of the TSG ending
b2	ordinary "word-internal" syllable boundary
b1	resyllabification (e.g., /het^{21} e/ → [he^{53}.le] in yws19)
b0m	syllable fusion (e.g. /hun^{51}-aŋ^{24}-a-sik^{21}/ → [hun^{55}.aã^{33}.sik^{21}] in nls107)

Figure 8. TW-ToBI break index values.

Currently the system has 6 tiers. The ***phones*** tier is a fairly narrow segmental transcription and the ***words*** tier is an orthographic transcription in Chinese characters. The ***syllable*** tier is analogous to the tier with the same name in the Pan-Mandarin ToBI system (M-ToBI) [6]. Each element on this tier is an "underlying" syllable that might be written with a single Chinese character. The primary tags on this tier are a syllable-by-syllable orthographic transcription in an ASCII version of Romaji. This tier is like the *words* tier in systems for languages with alphabetic writing systems in that elements on this tier define where tags must be placed on the ***breaks*** tier. The inventory of break index values is shown in Fig. 8. The *syllable* tier romanization keeps a record of the base tone, whereas the ***tones*** tier marks the surface tone, which should be different from the base tone if the syllable is not final in the TSG. The *tones* tier also could record non-canonical tone shapes that result when two tone specifications are compressed onto one syllable in syllable fusion, and it tags pragmatic particles as "PP" in lieu of committing prematurely to an analysis of the pitch shapes of these elements. As in the M-ToBI system, each *syllable* is also marked for ***stress***, using one of the values s0 (for a reduced syllable such as the /a^{51}/ in nlr107 in Fig. 2), s1 (for most syllables), and s2 (for a syllable with focal prominence such as /gua^{51}/ in jts57 in Fig. 1). For reasons of space, we have substituted a Romaji transcription for the Chinese characters on the *words* tier and omitted the *stress* and *syllable* tiers in the TW-ToBI transcriptions in Figs. 1-4 and 6-7.

5. OUTSTANDING QUESTIONS

As the above description of the "PP" tag suggests, there are outstanding questions concerning Taiwanese prosody that we hope to address by annotating a sufficiently large and varied corpus of read and spontaneous speech. For example, Figs. 1, 3, 4 and 7 all contain IP-final pragmatic particles. In the first two figures, we did not assign tones to the particles in the *syllable* tier transcription, whereas in the other two figures, we transcribed tones based on Cheng's [7] description of the meanings conveyed by varying the pitch pattern on otherwise homophonous particles. For example, the first particle in Fig. 4 is transcribed as /ɔ^{24}/, whereas the second is /ɔ^{21}/, following Cheng's description of these particles as having a rising pitch shape (to signal "warning") versus a low falling pitch shape (to signal "surprise"). Cheng does not explicitly equate these pitch shapes with the lexical tones, however, and marking these simply as "PP" on the *tones* tier lets us remain open to the possibility that the particles are lexically toneless and get their pitch shapes from IP-final boundary tones similar to those described for Cantonese in [8]. Thus, one goal of our research is to understand what determines the pitch shapes on PP-tagged syllables. Another is to understand why the syllable preceding a PP sometimes surfaces with its base tone value (as in Fig. 7) and sometimes with its sandhi tone value (as in Fig. 1).

More generally, we want to understand all the factors that determine the distribution of base tone versus sandhi tone in natural discourse. One such factor seems to be the use of prosodic phrasing to set off items that are in narrow focus in the discourse context, as illustrated in Fig. 1. In the read version of the sentence, /gua^{51}/ 'I' is grouped into the same TSG with the following verb, a phrasing that seems typical of pronoun subjects. In the original spontaneous utterance, by contrast, there is narrow focus on /gua^{51}/. This focus is marked by a local increase in duration and expansion of the pitch range, and also by making /gua^{51}/ final in its TSG, as indicated by the transcription of the base tone on the *tones* tier and the following b3 on the *breaks* tier.

Another factor that must be understood better is the role of neutral tone, as in Figs. 2 and 4 where the syllable preceding the derivational particle /e/ surfaces with its base tone. Chen [2] accounts for such cases by grouping the particle with following material, whereas our analysis groups the /e/ together with the preceding syllable, to account for the resyllabification of the coda consonant. This analysis is more in keeping with the close juncture that our TW-ToBI transcribers felt when they transcribed b1 rather than b3 before the /e/. It also illustrates the advantage of the ToBI framework design, which separates the marking of juncture on the *breaks* tier from the specification of the pitch categories on the *tones* tier.

6. ACKNOWLEDGMENTS

This work was supported by a Taiwan National Science Council grant to the first author. We also thank H. Pan for her leading role in the development of TW-ToBI.

7. REFERENCES

[1] http://www.coastalfog.net/languages.xiarom.html

[2] M. Y. Chen, "The syntax of Xiamen tone sandhi," *Phonology Yearbook* 4, pp. 109-149, 1987.

[3] S. Peng, "Production and perception of Taiwanese tones in different tonal and prosodic contexts," *Journal of Phonetics* 25, pp. 371-400, 1997.

[4] C. Hsu and S.-A. Jun, "Prosodic strengthening in Taiwanese: Syntagmatic or paradigmatic?" *UCLA Working Papers in Phonetics* 96, pp. 69-89, 1998.

[5] P. A. Keating, T. Cho, C. Fougeron, and C. Hsu, "Domain-initial strengthening in four languages," *Papers in Laboratory Phonology* VI, Cambridge University Press, in press.

[6] S. Peng, M. K. M. Chan, C. Tseng, T. Huang, O. J. Lee, and M. E. Beckman, "Towards a Pan-Mandarin prosodic annotation system," In S.-A. Jun (ed.) *Prosodic models and transcription: Towards prosodic typology*, Oxford University Press, in press.

[7] R. Cheng, "Taiwanese question particles," *Journal of Chinese Linguistics* 5, pp. 153-185, 1977.

[8] W. Y. P. Wong, M. K. M. Chan, and M. E. Beckman, "An Autosegmental-Metrical analysis and prosodic annotation conventions for Cantonese," In S.-A. Jun (ed.) *Prosodic models and transcription: Towards prosodic typology*, Oxford University Press, in press.