Corpus Approach to Phonetic Investigation— Methods, Quantitative Evidence and Findings of Mandarin Speech Prosody

Chiu-yu Tseng
Phonetics Lab, Institute of Linguistics, Academia Sinica, Taiwan
cytling@sinica.edu.tw

Zhao-yu Su
Phonetics Lab, Institute of Linguistics, Academia Sinica, Taiwan
morison@gate.sinica.edu.tw

Abstract

This paper reports how by developing a corpus approach we could bring new dimensions to phonetic research. To illustrate we show how Mandarin Chinese fluent speech prosody is discourse prosody in addition simply tones and intonation, and how higher level discourse information can be accounted for through corpus analyses of speech data. Methodological issues include sample size, domain and speech unit and their implications as well as quantitative analysis and significance of obtained results. Acoustic phonetic correlate discussed is F0 contour patterns across speech flow. Our study demonstrates how through careful design of data collection and corresponding quantitative analyses, corpus phonetics could bring new kind of data, new research methods and critical evidences to phonetic investigation that traditional phonetic approach could not accomplish.

1. Introduction

Traditional phonetic research has always adopted a bottom-up perspective, both in data collection and data analysis. Units of phonetic investigation are usually miniature fragments of speech, issues of investigation are usually phonetic and/or acoustic characteristics of such fragmentary units elicited in isolation, sample size never the major concern, research orientation always ultimate abstraction towards constructing phonological theories. By default phonetic units from bottom up are segments, syllables, words, phrase and stop at simple sentences.

When speech prosody became a central issue for technology development, especially in speech synthesis and unlimited TTS, the default unit of investigation remained simple sentence and/or phrase still and the default issue sentence intonation at rest. The fact that there is no operating definition for intonation of complex sentence has not brought much attention, nor has it ever been an issue as to how to define intonation group phonetically. Extended to tone languages such as Mandarin Chinese, although the acoustic side became more complicated because tones and intonation are both supra-segmental, unit of prosodic investigation remained at phrase or simple sentence. As a result, Mandarin prosody became tones and intonation by default, and Mandarin intonation strings form Mandarin fluent speech prosody. Nevertheless, technology development has revealed not only Mandarin intonation strings did not produce speech prosody on the one hand; intonations from fluent speech also exhibited too many variations for the simple definition to accommodate. Consequently, there were at least two major problems to resolve: one was to account for what Mandarin speech prosody was; another was to deal with intonation variations in continuous speech.

From the viewpoint of prosody, that is the inherent supra-segmental features of human speech, we note first that Mandarin tones are lexically defined; its unit the syllable, therefore tones simply lexical prosody. Secondly, we note that intonation is syntax defined; its unit types of simple sentence, therefore intonation is simply syntactic prosody. Since fluent continuous speech is most significant characterized by narration, lecturing or story telling instead of simple sentences elicited in isolation, it is in fact spoken discourse we should address rather than tones and/or intonation produced or elicited in isolation. Note also though discourse is composed of sentences, the sentences within a discourse are NOT unrelated. Rather, semantic coherence and cohesion are obligatory; sentences in a discourse are ASSOCIATED. Since syntax does NOT contain information above individual sentence; nor does syntax govern and constrain between-sentence association, the question then is clear: some form of prosodic expressions from higher level discourse information must be involved in fluent speech prosody in addition to tones and intonation; some kind of prosodic ASSOCIATION should and must exist. It is feasible to assume that speech prosody is NOT strings of unrelated intonations, but rather strings of structured intonations. The question now is: how and where to find and account for discourse information above sentences in prosody. Inspired by data-driven approaches commonly adopted by engineering approaches, we also assumed that observing and describing limited size of speech samples from bottom upward would most likely not resolve the variation problem, either.

In the following sections, we will show how by (1.) collecting large amount of speech data to construct speech corpora, (2.) gathering relatively large pieces of spoken discourses instead of elicited simple sentences, and (3.) adopting a more holistic top-down perspective to dissect the speech corpora, it is not only possible to account for higher level discourse information present in fluent continuous speech prosody, but also possible to provide quantitative evidences to remedy problems of insufficient sample size and out-of-hand variations. Analysis of F0 contour patters across speech will be used to support the hypothesis and illustrate our points.

There are three research problems to address: (1.) identify where additional prosodic information is located in the speech signals, (2.) separate discourse prosody from Mandarin tones and intonation in prosody analysis, and (3.) account for discourse prosody through quantitative analysis.

We consider the following prerequisites critical to prosody investigation. (1) Only fluent continuous speech of spoken discourses should be used so that the associative relationships between and among sentences and phrases are available in the speech data. (2) Top-down rather than bottom-up perspective of segmenting speech data is preferred so that prosodic associations would be maintained. (3) Speech units above IU should be available in the analysis so that behavior of individual intonation would be analyzed with proper context rather than independent from each other. Furthermore, methods of quantitative analysis and predictions should also accommodate contribution of different size of prosodic units under consideration.

2. Prosodic Phrase Grouping (Tseng, 2004; 2005)

Tseng's previous corpus studies of syllable duration patterns across read discourses have demonstrated that multiple-phrase speech paragraphs is a prosodic unit above phrases in discourse; the phrases within speech paragraph are actually structured into three discourse relative positions to yield higher-level association and indicate how and where speech paragraphs begin, continue and ends. She has since put forth a multiple-phrase prosody hierarchy called Prosody Phrase Grouping (PG) [13-21] and stated how PG represented prosodic organization that specified a higher prosodic node above phrases and groups them by three PG positions PG-initial, -medial and –final. The PG hierarchical framework specifies the organization whereby lower-level units are subject to higher-level constraints. Each layer in the framework contributes differently to output prosody, but cumulatively make up ultimate global prosody output. Therefore, the dynamics of speech prosody is characterized by a package of globally associated multiple phrases rather than unrelated strings of IUs. The simple prosody hierarchy states explicitly that by adding a higher PG level/node above phrases/IU, the respective prosodic roles of phrases PG groups can be defined. Compared to other attempts of automatic prosodic segmentation for continuous speech that proposed the classification of phrases into eight phrase types, [7, 8.] the major difference lies in the sufficiency of only three PG relative positions to capture and explain cross-phrase association in relation to higher-level discourse information; whereas the eight types remain arbitrary numbers that still assume phrases as independent, unrelated prosodic units without any relationship to each other.

The concept of phrase grouping is not just specific to Mandarin. It has been well accepted that utterances are phrased into larger constituents; together they (utterances and larger constituents) are hierarchically organized into various domains at different levels of prosodic organization.[9-11] Unfortunately this hierarchical organization is often ignored, as the necessary distinction between syntactic prosody (intonation) and discourse prosody often goes un-clarified. Tseng's PG framework not only specifies phrase as immediate subordinate units, but also by default specifies phrases at the same layer as subjacent sister constituents. By the same logic, PGs can further be extended as immediate constituents of a yet higher node discourse. Figure 1 is a schematic illustration of the framework that also includes the node *Discourse* above PGs.

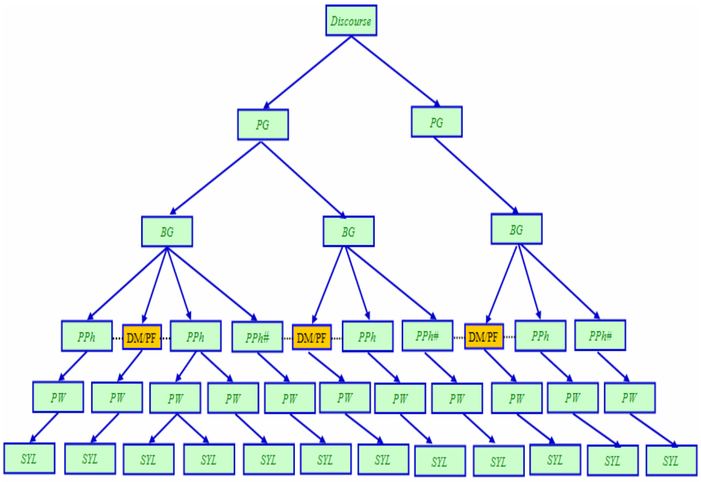


Figure 1: A schematic representation of how PGs form spoken discourse and where DM (Discourse Marker) and PF (Prosodic Filler) are additional associative linkers.

The 6-layer framework is from Tseng et al 2005[20] and based on the perceived units located within different levels of boundary breaks across the speech flow. The units used were perceived prosodic entities. The boundaries (not shown in Figure 1 to keep the illustration less complicated), annotated using a ToBI-based self-designed labeling system,[12] marked small to large boundaries with a set of 5 break indices (BI), B1 to B5, purposely making no reference to either lexical or syntactic properties in order to be able to study possible gaps between these different linguistic levels and units. Looking at Figure 1 from bottom up, the layered nodes are syllables (SYL), prosodic words (PW), prosodic phrases (PPh) or utterances, breath groups (BG), prosodic phrase groups (PG and Discourse. Optional discourse markers (DM) and prosodic fillers (PF) between phrases are linkers and transitions within and across PGs, whereby DMs function as attention callers and PFs as parenthetical speech units. These constituents are, respectively, associated with break indices B1 to B5. B1's denote syllable boundaries and may not correspond to silent pauses; B2's, perceived minor breaks between PWs; B3's, breaks between PPhs; B4's, points when the speaker takes in a full breath upon running out of breath, and also breaks at the BG layer; and B5's, perceived trailing-to-a-final-ends that occur followed by the longest break. In the framework, an IU is usually a PPh. When a speech paragraph is relatively shorter and does not exceed the speaker's breathing cycle, the BG and PG layers collapse into one PG layer, another feature of layered organization. Both BGs and/or PGs can be immediate subjacent units of a discourse.

The most significant features of the PG framework are how it explains and accounts for variations in intonation across the speech flow and higher-level contributions to final output prosody. Tseng et al has shown that a modified linear regression analysis corresponding to the PG hierarchy successfully proved how layered contributions cumulatively accounted for output speech rhythm[13-16,18,19]. These quantitative evidences confirm the existence of cross-phrase prosodic associations in fluent continuous speech, and explain how higher-level discourse information is realized in cross-phrase associations. Evidences of cross-phrase templates for syllable duration patterns were derived, as were intensity distribution patterns and boundary breaks that both individually and collectively accounted for systematic as well as layered contributions to output rhythm, intensity and boundary breaks.[14-16,18,19]

In the following discussion we will present analysis of F0 contour patterns to further illustrate discourse prosody, emphasizing on the corpus approach developed. The framework also accounts for why discourse information dwells not in individual IUs but in cross-phrase associations between and among them. The dynamics of speech prosody is in fact systematic.

2. 1. Speech Melody: Global F0 Patterns of PG

A cadence template of perceived normalized prosodic melody is presented in Figure 2, the trajectory denotes a minimal 3-phrase PG, preceded and followed by B4 or B5 where phrases within are separated by B3's. Note how the melody of a PG is featured by the relative patterns of how each position causes the respective phrase which corresponds to an IU to begin, hold and end[14,15,17,18].

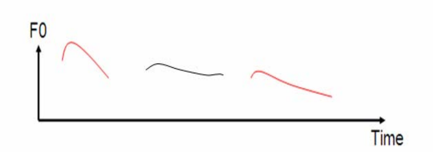


Figure 2. Schematic illustration of the global trajectory of perceived F0 contours of a 3-PPh PG. Within-PG units are PPh's and separated by boundary breaks B3's.

The following experiments illustrate how to analyze a small corpus of speech prosody and the quantitative methods used.

2. 1. 1. Speech data

Mandarin Chinese speech data from Sinica COSPRO 08 were used[22,23]. A 30-syllable, 3-phrase complex sentence representing a short PG was constructed as a carrier paragraph with target single syllables embedded in three PG positions, namely, "△是一個常見的字，一般人常把△字掛在嘴邊，講話時動不動就會提到△. (Translation: △ is a frequently used word, people often use the word △ in their speech, and make mention of △ from time to time quite frequently.)". △ denotes the target syllable. The PG-initial, -medial and -final phrases in the carrier PG consisted of 8, 11 and 11 syllables, respectively. Note that (1) the speech paragraph is designed to remove as much lexical and semantic focus as possible and renders canonical global PG patterns, (2) the target syllables were embedded into the 1st, 6th and last syllable of the first, second and third phrases, thereby occurring at the initial, medial and final locations of the PG-initial, -medial and -final positions respectively; and (3) in spoken discourse, a multiple-phrase PG usually exceeds three phrases indicating the PG-medial phrases are often more than one. Furthermore, when compared to the reading of text passages, although such a 3-phrase complex sentence contains relatively minimal speech prosody we believe it would still contain discourse information and at the same time offer repetitions of syllables in a uniform context for tone-prosody investigations. Speech data from a male (M054C) and a female (F054C) native speaker of Mandarin Chinese spoken in Taiwan were recorded in sound proof chambers. Both were instructed to read 1,300 speech paragraphs at their normal speaking rate with natural focus into microphones. The speaking rates are 289 and 308 ms/syllable for M054C and F054C, respectively. 60 files from F054C with target syllables of tone 1 were analyzed to illustrate PG effects. Analyses and predictions of F0 values were performed via parameters of the Fujisaki model (*Ap*, *Aa*) .[1-3, 24,25]

2.1.2. Speech Data Annotation

The speech data were manually labeled by independent transcribers for perceived boundaries and breaks (pauses), using a 5-step break labeling system corresponding to Figure 1. Pair-wise consistency was obtained from the transcribers.

2.1.3. Higher-level Discourse Information in Prosody Analysis

The goal of the following two experiments is to look for phrase components and accent components that also contain additional higher-level information from the PG hierarchy. The Fujisaki model operates on IU to derive F0 curve tendency of both the syllables and the phrase.[1-3,24,25] Therefore, the three phrases are first analyzed independently then compared in relation to their relative PG positions. Accent components (Aa) and phrase components (Ap) are first separated by a lowpass-filter[4-6] then calculated independently, whereby (Aa) predicts more drastic local F0 variations over time and (Ap) predicts smoother global F0 variations over time. The steps involved are first, analyzing these two components at the PPh level, that is, F0 curve tendency of individual phrases. Next, the same two components are analyzed in relation to higher-level PG information, that is, PPh's are classified by the three PG positions and analyzed respectively. Following that, a comparison of whether differences exist among the three PG positions is made. Lastly, we add contributions from the PPh level and the PG level to derive cumulative predictions and these predictions are then compared with speech data to test the validity.

2.1.3.1 Experiment 1

The aim of this experiment is to investigate (1) whether patterns of Ap could be derived from speech data, (2) whether there is evidence of interaction between Ap predictions from the PPh level and Ap predictions from higher-level PG positions, and (3) whether the evidence found could predict pitch allocation in the speech flow. Two levels of the PG framework are examined. According to the definition of PG hierarchy, all three PPh's at the PPh level are subjacent subordinate constituents of PG which are sister constituents to each other; each PPh is still an independent IU without any higher-level PG information. At the immediate upper PG level, each PPh is then assigned a PG role in relation to the three PG positions. Thus, at the PPh level, each of the three phrases is assumed as an independent prosodic unit. The magnitude of Ap's is generalized and assigned to predict the Ap within, while ignoring higher-level PG information. Next, at the PG level, the PG effects are considered where different values of Ap are assigned to predict phrase components according to where each of the three PPh's is located in PG-positions. Finally, prediction accuracy between PPh's with and without PG effects are compared with the original speech data for validity.

First, speech data are analyzed to provide prediction references. Ap values are extracted from the speech data and their characteristics examined. The respective range and distribution of extracted Ap values in each PG-position from the speech data are illustrated in Figure 2 and Table 1. Next, the characteristics of distribution in each PG-position are generalized and used for subsequent Ap predictions. Using a step-wise regression technique, a linear model is developed and modified for Mandarin Chinese to predict Ap. The hierarchical PG organization of prosody levels (the aforementioned system of boundaries and units) is used to classify Ap at the levels of the framework. Moving from the PPh level upwards to the PG level, we examine how much was contributed by the PG level. All of the data are analyzed using DataDesk™ from Data Description, Inc. Two benchmark values are used to evaluate how close predicted values are when compared with values derived from original speech data. The first benchmark is percentage of sum-squared errors at the lower PPh layer. The PG framework assumes that errors at a lower level are due to lack of information from higher levels. Therefore, residual errors (RE), defined as the percentage of sum-squared residues (the difference between prediction and original value) over sum-squared values of original speech data, are then included into the immediate higher-level for further predictions. If predictions improve from a lower level upward, the difference between two subjacent levels are considered as contributions from the immediate higher level.

Table 1. Range of values of Ap from phrases produced by female speaker

F054C in three PG-related positions are presented.

PG Position	Ap range
-Initial	0.959~0.499
-Medial	0.615~0.04
-Final	0.678~0.093

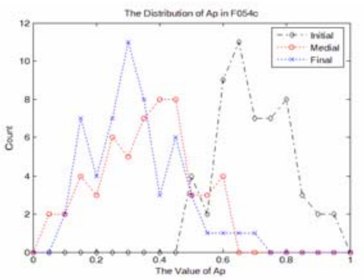


Fig. 3. A schematic representation of the distribution of the Ap's of speaker F054C where the horizontal axis represents values of Ap and the vertical axis represents number of Ap occurrence.

- Results

Table 2 illustrates the coefficients of Aps from PPhs in a PG. At the PPh level, when each PPh is treated as independent prosodic unit, the expected cell mean is at 0.4595. However, at the PG level, where the PPh's were classified by the three PG positions, namely, PG-initial, -medial and -final, the expected cell mean with PG effects are 0.6984, 0.3536 and 0.3265, respectively. In contrast to PG-initial PPh, the Ap of PG-final PPh is shortened. The coefficients reflect a clear distinction between PG-initial and PG-final prosodic phrases.

Table 2. The expected cell mean of predictions with and without the PG effect. The top row shows the expected cell mean value when PG effects are ignored at the PPh level. The bottom row displays the expected cell mean values when PG effect is considered at the PG level, with the three relative positions, namely, PG-I(nitial), PG-M(edial) and PG-F(inal).

without PG effects at PPh level:	0.4595		
with PG effect at PG level:	PG-I	PG-M	PG-F
	0.6984	0.3536	0.3265

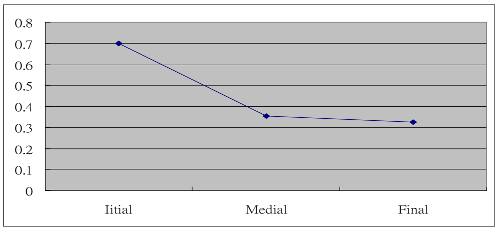
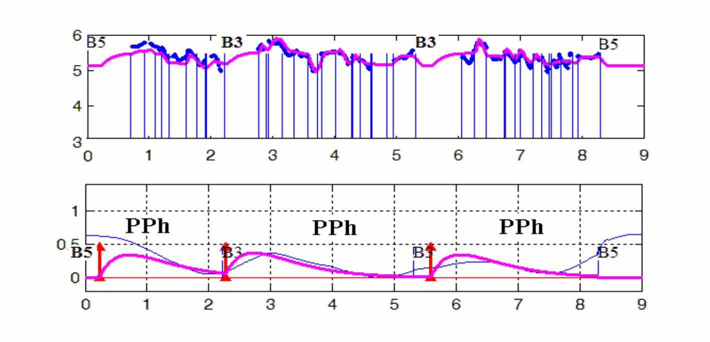


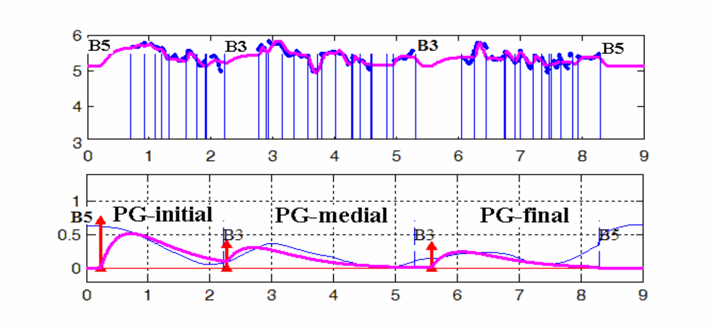
Fig. 4. is a schematic representation of the patterns of phrases after PG effect is taken into consideration. Note how the PG-initial and PG-final groups possess the sharpest distinction.

When each IU (PPh in our framework) is analyzed independently, results revealed that correct predictions were only 40.15% and 59.85% were errors. After considering PG effects one level upward of the prosodic hierarchy, predictions were improved by 24.84%. Cumulative perdition accuracy was 65%. Ap adjustments with respect to PG positions provide further evidence of how prosodic units and layers function as constraints on the Ap in the speech flow and how higher-level prosodic units may be constrained by factors that differ from those constraining lower-level units. If higher-level information is ignored, inputs of prediction would be insufficient.

Finally, by adding up the predictions of the PG layer, we are able to derive a prediction of F0 curve allocation for all three phrases. Comparisons between predictions with and without PG effects are then made with the original speech data. Figures 4 and 5 show these comparisons. The final cumulative predictions indicate that patterns of F0 allocation in Mandarin speech flow cannot be adjusted by the PPh level alone. Input from the PG level must be included. Moreover, these results are also evidence demonstrating that the PPh is constrained and governed by higher-level information (PG). As illustrated in Figure 5, the distinction between PG-initial and PG-final is most obvious. If PG effect is neglected, the accuracy will diminis



(a)



(b)

Figure 5. Comparisons of Ap predictions without PG-effect (a) and with PG-effects (b) to the original speech data. The darker line in the upper panels shows F0 plotting of 3 phrases, while the lighter line indicates 3 predicted F0 curves; vertical lines denote syllable boundaries. In the lower panels, the thin line shows comparisons of lowpassed F0 curve, while the thicker line indicates predicted phrase components. Each arrow on the lower panels denote an Ap.; their heights represent Ap values. In each panel, the vertical axis represents logarithm value of F0 curve; while the horizontal axis represents temporal code.

In summary, the PPh layer only constitutes around 40% of the prosody output while higher-level discourse information at the PG layer puts in an additional 25%. Together, the PPh and PG layer make up a total of 65% of prosody output. Note however, that since the PG layer is higher in the prosody hierarchy and commands all phrases under it, its effect is not to be ignored. Without it, there would be no discourse prosody. By definition of the PG hierarchy, the remaining 35% of contributions should come from the lower syllabic (tonal) and word (both lexical and prosodic) levels. Working upwards in the prosody hierarchy, tonal information certainly is not the most significant contributor of fluent speech prosody.

2.1.3.2. Experiment 2

We assume that accent components (Aa, in the Fujisaki model) are also governed by the PG hierarchy as specified by our PG framework. Hence, the SYL, PW and PPh levels in the PG framework should all contribute to output prosody, respectively. The aim of this second experiment is to investigate the contributions of the SYL to PPh prosodic levels from an analysis of Aa. A similar regression technique is used to calculate contributions from each prosodic level to the final output in terms of magnitude of Aa from the SYL, PW and PPh levels.

At the syllable layer, the method adopted is to approach the F0 curve of each syllable by one accent component. In other words, each syllable is connected to one Aa, which makes us unable to extract SYL Aa accurately at the current stage. Nevertheless, the SYL, PW and PPh level models are postulated as follows:

The SYL Layer Model:

$$\mathrm{Aa} = \mathrm{constant} + \mathrm{SYL} + \mathit{Delta}\ 1 \qquad (1)$$

SYL in the above represents syllable type. Factors considered include 23 syllable categories (excluding target syllables), and 5 tones (4 lexical tones and 1 neutral tone).

The PW Layer Model:

$$Delta1 = f(PWLength, PWSequence) + Delta2 \qquad (2)$$

Each syllable is labeled with a set of vector values; for example, (3, 2) denotes that the unit under consideration is the second syllable in a 3-syllable PW. The coefficient of each entry is then calculated using linear regression techniques identical to those of the preceding layer.

The PPh Layer Model:

$$Delta2 = f(PPhLength, PPhSequence) + Delta3 \qquad (3)$$

Each syllable was labeled with a set of vector values; for example, (8, 4) denotes that the unit under consideration is the fourth syllable in an 8-syllable PPh. The coefficient of each entry is calculated using linear regression techniques identical to those of the preceding layer

- Results

Table 3. Cumulative accuracy of Aa predictions from SYL, PW and PPh levels.

Prosodic level	Contribution	Cumulative accuracy
SYL	19.89%	19.89%
PW	1.1%	20.99%
PPh	5.07%	25.16%

Table 3 shows contributions and cumulative prediction accuracy at each prosodic level from Aa analyses.

If the factors considered include only 5 tones without syllable categories, the accuracy of Aa prediction is about 12.5%. When syllable categories are included, the cumulative accuracy is improved to a cumulative 19.89%. From the SYL layer upwards to the PW level, cumulative prediction is improved to 20.99%. Finally at the PPh level, the cumulative accuracy of Aa prediction is 25.16%.

3. Discussion

From the quantitative evidences of the tow experiments, we have demonstrated that how a hierarchical prosody framework accommodating more than one sentence/phrase could (1.) capture cross-phrase melodic associations, and (2.) explain why tones and independent intonation contours are not the only contributors to variations of F0 contour patterns across speech flow and (3.) account for why discourse information is crucial. The above experiments showed once a phrase becomes a subordinate constituent of a higher node as PPh becomes a PG constitute, it is no longer an independent IU. The higher node PG requires each PPh it groups to adjust by PG relative positions to form discourse association, and hence the intonations vary. Note however, that the variations are systematic; the PG-initial and –final positions specify two respective PPhs to retain intonation contours differing in relative starting point, slope, with boundary effects and boundary breaks, thereby yielding the basic canonical cross phrase melody of continuous speech, whereas the medial phrases are specified to hold flat to signal continuation rather than termination. Note also though both the PG-initial and -final PPh's may exhibit declination, the relative degrees and slope of declination are different, while final-lengthening-and-weakening occurs only at the PG-final PPh. Pair-wise contrast between the PG-initial and -final phrases is significant.[17.] It should be evident by now that phrases in continuous speech must be considered in relation to one another instead of individually one at a time; intonation variations are in fact systematic and predictable. Furthermore, note the selected 3-PPh complex sentence presented in the present study represents a relative unmarked representation of a canonical default prototype of PG while our collected speech corpora (as in COSPRO [22,23]) include multiple-phrase paragraph of up to 12 PPh's. In other words, depending on the speaking rate, up to 10 PPh's could be tolerated and accommodated between a PG-initial and -final PPh; all 10 of them with relatively flatter intonation to signal continuation. The melodic canonical form from PG also presents a base form for other add-ons such stress, focus, and emphasis.

Furthermore, we argue that though global melody and rhythm may differ from one language to another, higher level discourse prosody is not language-specific. Any attempt at prosody organization and modeling should incorporate language-specific patterns of duration allocation and intensity distribution in addition to F0 contours, but maintain the discourse coherence and association.

4. Conclusion

From the quantitative evidences of F0 analysis, we have shown from corpus analysis corresponding to a hierarchical prosody framework that both Mandarin speech melody and speech prosody are not simply about tones and intonation. The F0 contour patterns reveal crucial higher level discourse information across speech flow. Each layer of the PG hierarchy contributes to output F0 and cumulatively adds up to the final prosody output.[13-21] We must caution here that speech prosody is not merely patterned cross-phrase F0 variations, but should also include cross-phrase duration patterns for speech rhythm as well as cross-phrase intensity distribution and boundary effects with boundary breaks to cover all of the acoustic correlates as we have. We believe the results could enhance both speech synthesis and recognition. Last but not least, none of the above augments could hold without the quantitative evidences from corpus analysis. Therefore, we conclude that corpus phonetics offers new frontiers to linguistic research as well as technology enhancement.

References

1. Fujisaki, H., Ohno, S., Tomita, O., 1996. Automatic parameter extraction of fundamental frequency contours of speech based on a generative model, Proceedings of 1996 International Conference on Signal Processing, vol. 1, pp. 729-732.
2. Fujisaki, H., Ohno, S., Wang, C., 1998. A command-response model for F_0 contour generation in multilingual speech synthesis, Proceedings of the 3rd ESCA/COCOSDA International Workshop on Speech Synthesis, pp. 299-304.
3. Fujisaki, H., Ohno, S., Gu, W., 2004. Physiological and physical mechanisms for fundamental frequency control in some tone languages and a command-response model for generation of the F_0 contour, Proceedings of International Symposium on Tonal Aspects of Languages with Emphasis on Tone Language, pp. 61-64.
4. Mixdorff, H., Fujisaki, H., 1997. Automated Quantitative Analysis of F_0 Contours of Utterances from a German ToBI-Labeled Speech Database. In: Proceedings of the '97 Eurospeech, vol.1, pp. 187-190.
5. Mixdorff, H., 2000. A Novel Approach to the Fully Automatic Extraction of Fujisaki Model Parameters. Proceedings of ICASSP 2000, vol. 3, pp. 1281-1284.
6. Mixdorff, H., Hu, Y. and Chen, G., 2003. Towards the Automatic Extraction of Fujisaki Model Parameters for Mandarin. In Proceedings of Eurospeech 2003.
7. Singer., H. and Nakai., M. 1993. Accent Phrase Segmentation Using Transition Probabilities Between Pitch Pattern Templates, *EUROSPEECH'93*, pp. 1767-1770
8. Nakai., M., Singer., H., Sagisaka., Y. and Shimodaira., H. 1995. Automatic prosodic segmentation by *F0* clustering using superpositional modeling, *ICASSP95*, 624–627.
9. Shattuck-Hufnagel, S., Turk, A., 1996. A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguist Research*, 25(2):

193.

10. Gussenhoven, C. 1997. Types of Focus in English? In Daniel Buring, Matthew Gordon and Chungming Lee (eds.) *Topic and Focus: Intonation and Meaning: Theoretical and Crosslinguistic Perspectives*. Dordrecht: Kluwer.
11. Selkirk, E. 2000. The interaction of constraints on prosodic phrasing. In Merle Horne (ed.) *Prosody: Theory and Experiment*, Dordrecht: Kluwer. 231-262.
12. Tseng, C. and Chou, F. 1999. "A prosodic labeling system for Mandarin speech database" *Proceedings of the 14th International Congress of Phonetic Science,* (Aug. 1-7, 1999), San Francisco, California, 2379-2382.
13. Tseng, C. and Lee, Y. 2004. "Speech rate and prosody units: Evidence of interaction from Mandarin Chinese" *Proceedings of the International Conference on Speech Prosody 2004*, (Mar. 23-26, 2004), Nara, Japan, 251-254.
14. Tseng, C., Pin, S., Lee, Y., 2004. Speech prosody: issues, approaches and implications. in Fant, G., H. Fujisaki, J. Cao and Y. Xu Eds. *From Traditional Phonology to Mandarin Speech Processing, Foreign Language Teaching and Research Process*, 417-438.
15. Pin, S., Lee, Y., Chen, Y., Wang, H. and Tseng, C. 2004. "Mandarin TTS system with an integrated prosody model," *Proceedings of the 4th International Symposium on Chinese Spoken Language Processing*, (Dec. 15-18, 2004), Hong Kong , 169-172
16. Tseng, C. and Lee, Y. (2004). "Intensity in relation to prosody organization ," *Proceedings of the 4th International Symposium on Chinese Spoken Language Processing*, (Dec. 2004), Hong Kong , 217-220
17. Tseng, Chiu-yu and Pin, Shao-huang (2004). "Modeling prosody of Mandarin Chinese fluent speech via phrase grouping" Proceedings of *Speech and Language Systems for Human Communication* (SPLASH-2004/Oriental-COCO SDA2004), (Nov. 17-19, 2004), New Delhi, India, 53-57.
18. Tseng, C. Pin, S. and Lee, Y., Wang, H. and Chen, Y. 2005. "Fluent Speech Prosody: Framework and Modeling", *Speech Communication (Special Issue on Quantitative Prosody Modeling for Natural Speech Description and Generation)*, Vol. 46:3-4, 284-309.
19. Tseng, C. and Fu. B. 2005. "Duration, Intensity and Pause Predictions in Relation to Prosody Organization," *Proceedings of Interspeech, 2005,* (September 4-8 ,2005) , Lisbon, Portugal, 1405-1408
20. Tseng, C., Chang, C. and Su, Zh. 2005. "Investigation F0 Reset and Range in relation to Fluent Speech Prosody Hierarchy", *Technical Acoustics,* Vol. 24, 279-284.

21. Tseng, C., Su, Zh., Chang, C. and Tai, C. 2006. Prosodic filers and discourse markers—Discourse prosody and text prediction. *TAL 2006 (The Second International Symposium on Tonal Aspects of Languages)* April 27-29, 2006, La Rochelle, France.
22. Tseng, C., Cheng, Y. and Chang, C. 2005. “Sinica COSPRO and Toolkit—Corpora and Platform of Mandarin Chinese Fluent Speech” *Proceedings of Oriental COCOSDA 2005*,(Dec. 6-8, 2005), Jakarata, Indonesia, 23-28
23. Tseng, Sinica Mandarin Continuous Speech Prosody Corpora (COSPRO http://www.myet.com/COSPRO)
24. 鄭秋豫、李岳凌、蔡蓮紅、鄭雲卿 （排印中）“兩岸口語語流韻律初探—以音強及音節時程分佈為例” *首屆海峽兩岸現代漢語問題學術研討會論文集*. 上海商務印書館
25. Tseng, C. 2006. ”Higher Level Organization and Discourse Prosody“, Invited keynote paper, *TAL 2006 (The Second International*
26. *Symposium on Tonal Aspects of Languages)*, April 27-29, 2006, La Rochelle, France. 23-34