ALL FOR A REASON—
PROSODIC REDUCTION IN CONTINUOUS SPEECH

Helen Kai-yun Chen, Yen-Hsing Chen, Chiu-yu Tseng

Phonetics Lab, Institute of Linguistics, Academia Sinica, Taiwan
cytling@sinica.edu.tw

ABSTRACT

This study focuses on a particular instance of prosodic reduction, namely *parentheticals*, in continuous speech. Provided by perception-based definition on top of traditional syntactic and semantic definitions, instances of *parenthetical* construction plus its hosting *frame* as bearing units have been identified from selected continuous speech. Based on methodologies of acoustic analyses as well as calculation of *contrast degree* and *emphasis weighting*, the exploratory study suggests that the named construction is neither an insertion nor simply a linear integration into its hosting *frame*; instead it belongs to part of perceived prosodic highlight attributed information planning within the discourse context. The present findings on parenthetical correlated prosodic reductions thus shed lights on how parenthetical construction should be viewed as an *integrated* part of information attributed discourse planning at higher levels, which eventually contributes to global context prosody.

Index Terms— prosodic reduction, parenthetical construction, continuous speech, context prosody, discourse association and coherence

1. INTRODUCTION

To capture features belonging uniquely to continuous speech, it is essential that we identify what contributes most to the understanding towards discourse association and coherent speech production at discourse levels. As have shown, discourse coherence could be better accounted for when we consider closely how speech is perceived in *context*: not only its structure and meaning, but mostly by the *prosodic* manifestation. In particular, specifications of discourse association could be more faithfully represented by: a. prosody-oriented boundaries and breaks in a hierarchical relationship (e.g. [1], [2]), and b. perception-based prosodic highlight allocations in correlation with information planning (e.g. [3]). On the one hand, the hierarchical framework can better illustrate relative cross-phrase discourse associations beyond linear 'beads-on-the-string' integration. On the other hand, discourse prosody features in the **coarsely-graded** nature (i.e. [4]) that directly mirrors the deployment of perceived prominence for information planning. Crucially, discourse association in its varied surface realizations can be translated and converged into relative compositions of 'ups' and 'downs' in limited numbers of patterns from speech context -- prosodic patterns in speech context hence exist for specific reasons.

Focusing specifically on the 'up' parts from speech signals, it has been recently identified that more often they are associated with information *projection*, mainly advance prompting of information [3], [5]. It is further suggested that alternations and compensations between perceived prominence prompted indexes for focal information and information *projection* thus play the role in illustrating features from discourse-level *context prosody* [3], [5]. Having been concentrating on the 'up' parts of continuous speeches, currently we turn our attention to the 'down' parts, namely perceivable prosodic *reduction*. Specifically we explore *parentheticals* as one of the levels carrying perceivable *reduction* in continuous speech. The examination of prosodic *reduction*, most of all, is inspired by the belief that perceptually their reduced nature goes hand-in-hand with prosodic highlight in the speech context. It is held that together prosodic highlight and reduction are responsible for bringing out of speech signals perceivable saliency, hence one of the sources for speech expressiveness. Here exploration of perception-based *parentheticals* aims at contributing to a full account of features uniquely belonging to speech and eventually better interpretations for **context prosody**.

Parenthetical construction in literature has been discussed from various viewpoints, including theoretical syntax and morpho-syntactic approaches, also by its prosodic realizations, as well as its meanings and functions from interactional perspectives. In syntactic approach *parenthetical* has traditionally been treated as a construction that is 'linearly integrated in another linguistic structure' [6] but bears no direct relationship nor contributes to the latter [7:179]. One of the challenges to a precise definition for *parenthetical* is due to that it does not correspond to one single morpho-syntactic class [8]. Consequently, in the relevant literature it has been treated as an inserted sequence that is structurally independent from its host (e.g. [9]). Moreover, *parentheticals* feature in acoustic realizations including: lower F0 and/or compressed F0 range, weaker

intensity, faster speaking rate and sometimes bounded by pre-/post-pause [6], [10], [11]. As for meaning and function, it has been illustrated that *parenthetical* plays a role in interaction and function to provide supplementary information that contributes to metacommunication [12].

However, given that a syntax-solely approach has been quite vague about defining *parentheticals* and most of the acoustic features being descriptive, here we aim at offering an alternative yet systematic account for *parentheticals* in continuous speech. The identification of parentheticals has been mainly *perception*-oriented, while taking into consideration at the same time the syntactic/semantic completion by both *parenthetical* itself and its hosting *frame*. Specifically, we explore into how *parentheticals* as an instance of prosodic *reduction* interact with discourse prosody. Incorporating methodologies of acoustic analyses and calculation of *contrast degree* as well as *emphasis density*, the current results foreground the particular case of *parentheticals* as an integrated part of the hosting *frame*. Neither insertion nor a linear integration, it is suggested that *parentheticals* is part of overall information planning motivated by perceived emphasis allocations in continuous speeches. Eventually the exploration on prosodic *reduction* contributes to and facilitates understanding toward coherent discourse processing and the co-construction of global **context prosody**.

2. SPEECH DATA AND ANNOTATION

2.1. Speech data and preprocessing

The present data is a selection from university classroom lecture [13], delivered by a male professor in form of continuous speech. The total time of the corpus is about 2.5 hours, equaling to approximately 33980 syllables. The speech data was first preprocessed by using HTK Toolkit to force-align preliminary segmentations and the output was then manually checked by trained transcribers. Afterwards the data have undergone labor-intensive annotations in separate layers for prosody-related and perception-based information, including discourse-prosodic units/boundaries, perceived prosodic highlight and the identification for instances of *parentheticals*. The annotation schemes will be introduced in the following.

2.2. Data annotations

2.2.1. Annotations for discourse-prosodic unit (DPU)

Following the hierarchical prosodic phrase grouping (HPG) framework from [1], [14] and [15], discourse-prosodic units (DPU) in five levels were annotated for the current data. The boundaries of five levels are marked from B1 to B5, corresponding respectively to *syllable* (**SYL**), *prosodic word* (**PW**), *prosodic phrase* (**PPh**), *breath group* (**BG**, a physio-linguistic unit constrained by change of breath while speaking continuously) and *multiple phrase speech paragraph* (**PG**). By default the boundary breaks, prosodic units and their relationship within the HPG framework could be accounted for by:

SYL/B1< PW/B2< PPh/B3< BG/B4< PG/B5 [1].

2.2.2. Annotations for perceived prosodic highlight

The data were manually tagged by trained annotators, in a separate layer, into a string of perceived emphasis/non-emphasis tokens (ETs). The annotation is according to four degrees of relative prominence strength by perception judgement, each defined as:

- E0 -- reduced pitch, lowered volume, and/or contracted segments
- E1 -- normal pitch, normal volume and clearly produced segments
- E2 -- raised pitch, louder volume and irrespective of the speaker's tone of voice
- E3 -- higher raised pitch, louder volume and with the speaker's change of tone of voice

By this scheme of annotation we emphasize the fact that degrees of prominences can be consistently perceived by *only* limited levels for contrastiveness.

2.2.3. Annotations for parenthetical and frame

Parenthetical construction (also **PAR**) together with its bearing unit *frame* (**FRA**) was annotated in other layers. The identification for **PAR** is basically perception-based, defined as a construction that is disruptive to the current speech production and is perceived distinctively by discernible acoustic features from the context. Meaning/function completion can be established by both **PAR/FRA**. The relationship between **PAR** and its **FRA** is further assumed to co-construct a complete information-bearing unit, in which **FRA** equals roughly to an utterance. Viewed as nested within **FRA**, each **PAR** is anchored by one **FRA** divided up into two parts: **FRA-(A)**, which immediately precedes **PAR** and functions to *project* the following **FRA-(B)** (i.e. [12]). **FRA-(B)** follows right after **PAR** and sometimes may back-trace part of the content from **FRA-(A)** (i.e. [6]). Note that we do not rely on a single morpho-syntactic category to identify **PAR** and its length can range in different sizes from one prosodic word in minimal up to several prosodic phrases.

3. METHODOLOGY

To explore *parentheticals* as instances of prosodic reduction in continuous speech, we first concentrate on the acoustic profiles of the *parenthetical* (**PAR**)-*frame* (**FRA**) construction. Then calculations of *contrast degree* and *emphasis density* have been conducted, following the methodologies below.

3.1. Acoustic features

To offer a glimpse of the acoustic realizations of **PAR-FRA** construction, we first compute the mean values of major acoustic features, including: F0, F0 range, duration and

intensity throughout the construction. Specifically, we extract values of these features by the DPU *prosodic word* (**PW**) and *prosodic phrase* (**PPh**) located immediately prior to and after the initiation and ending boundaries of **PAR**, and also throughout **PAR-FRA**.

First of all, we use SAP toolbox [16] to extract F0 value from the adjacent **PW**s/**PPh**s by the boundaries of **PAR**. To obtain F0 range, we then simply subtract F0 minimum from the maximum values. As for duration, we first normalize the length of every phoneme to remove the intrinsic duration differences, and then average the phoneme duration to obtain speaking rate from adjacent PWs/PPhs at **PAR** boundaries. Finally intensity (dB) is extracted from the same units also by using SAP toolbox.

3.2. Contrast degree

To verify the relationship between **PAR**/**FRA**, we further calculate *contrast degree* (**CD**), following similar rational from previous study [17]. Here **CD** is defined as the differences in the average acoustic values derived from *prosodic phrase* (**PPh**) unit by both **PAR** and **FRA** at boundaries. **CD** calculation, moreover, is based on the average acoustic values including F0, intensity and duration derived from previous calculations. So CD values are obtained, following:

$$CD_{f,S} = \frac{\sum_{n \in S} M_{R,f,n} - M_{L,f,n}}{|S|} \qquad (1)$$

in which the subscript letters to the mean value M are defined as: *L*/*R* indicating the **PPh**s before/after a boundary; *f* denoting the acoustic feature type; *n* for the index locating a **PPh** in a domain S, which can be either **PAR** or **FRA**.

3.3. Emphasis density

Emphasis density (**ED**) is an ad-hoc estimation based on levels of perceived prosodic highlight annotated for the current data. Following similar rational from [1], we arbitrarily assigned weighting scores to the emphasis degree tags: [1 1 2 3] for [E0 E1 E2 E3]. Afterwards we modify the score assignment to [-1 1 2 3] for [E0 E1 E2 E3], with the goal to foreground the incorporation of *reduction* in the annotation of perceived prosodic highlight. Following (2), we calculate **ED** by the *prosodic word* **PW** unit from each *prosodic phrase* **PPh** in both **PAR** and **FRA**:

$$\mathrm{ED}_i = \frac{Score_{i-1} + Score_i + Score_{i+1}}{3} \qquad (2)$$

For **PW** of index *i* in any PPh, Score_i is calculated simply by counting the weighting score corresponding to that **PW**. Note here since we further consider the emphasis levels annotated for both pre-/post- **PW**s to the current one, so that the average score ED_i reflects not merely each PW per se, but together the average scores from **PW**s in the neighborhood.

4. RESULTS AND DISCUSSIONS

4.1. Tokens of parenthetical-frame construction

As explicated in 2.2.3, the *parenthetical-frame* construction actually can be indicated via a tripartite structure in sequence as: **FRA**-(**A**)/**PAR**/**FRA**-(**B**). Based on the annotation scheme, 80 sets in this sequence have been identified. We preliminarily estimate the length of each **PAR** annotated, plus examine the boundary distribution by the boundaries of **PAR**. The results are summarized in Tables 1 and 2.

Table 1: *Summary of length of parenthetical PAR*

DPU	Min. (#)	Max. (#)	Ave. (#)
SYL (B1)	3	49	14.6
PW (B2)	1	20	6.1
PPh (B3)	0	5	1.7

Table 2: *Boundary type distribution by PAR*

Boundary Type	boundary 1 FRA-(A)|PAR	boundary 2 PAR|FRA-(B)
B2	15 (19%)	13 (16%)
B3	65 (81%)	66 (82%)
B4	0	1

4.1.1. Discussion

Table 1 demonstrates that the length of **PAR** is highly various: by *prosodic word* (**PW**), it can be of the size ranging from 1 up to 20 **PW**s. The discrepancy in its length reinforces why it is a challenge to capture parentheticals based solely on one single morpho-syntactic class. Turning to Table 2, the summary reflects that majority of the boundaries (both pre/post) to **PAR** fall by *prosodic phrase* (**PPh**/**B3)** boundaries. In the following analyses and calculations, therefore, **PPh** will serve as the base unit to account for **PAR-FRA**.

4.2. Acoustic features

To profile the **PAR-FRA** construction by the tripartite structure, we first calculate the mean values of major acoustic features. Following the methodology from 3.1, we concentrate on acoustic features including F0, F0 range, intensity and duration. Values of acoustic features were extracted from **PW**s and **PPh**s located by the boundaries of **PAR**. We further calculate the average values of acoustic features throughout **PAR-FRA** (here translated to the **F**rame-(A)/**P**arenthetical/**F**rame-(B), **F-P-F** sequence). The results are summarized in Fig. 1.

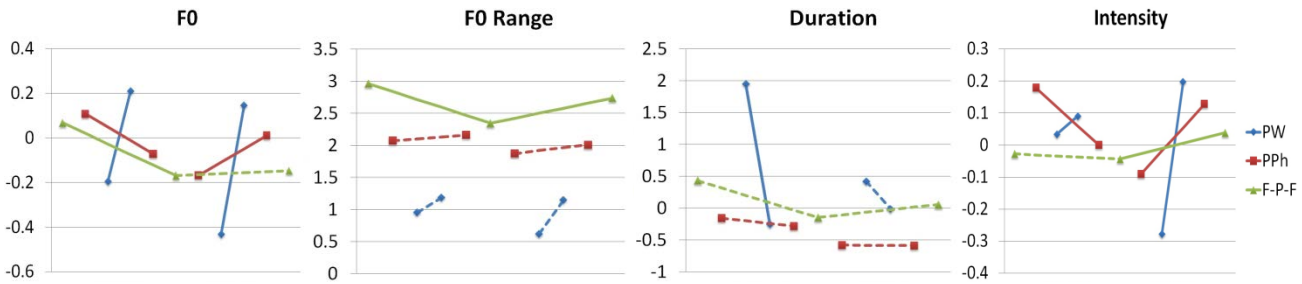

Figure 1: *Results of mean values for four acoustic features (solid line denotes significant T-test results yield).*

4.2.1. Discussion

Focusing on the **F-P-F** sequence, a shared observation across all panels from Fig. 1 is a trend of 'valley-shape' acoustic realizations: e.g. the average values derived from **PAR** are generally lower and form the valley, comparing to the higher values in average from both parts of **FRA**. This implies that **PAR** is perceived as in relatively lower pitch, more compressed F0 range, slower speaking rate and weaker intensity, which in general is consistent with the observations reported in the literature.

Also from Fig. 1 results of mean acoustic values by *prosodic word* (PW) and *prosodic phrase* (PPh) units at the boundaries of **PAR** are presented. It is demonstrated clearly that results by lower-level **PW** unit can only capture the low-high/ slow-fast/ weak-strong patterns in parallel from both **PAR** boundaries. The results by **PPh**, on the other hand, seem more approaching the 'valley-shape' realization as observed across **F-P-F**. Such observation suggests that a mere comparison of the prosodic realization around the initiation and completion of *parentheticals* by lower-level discourse-units such as **PW** can only capture partial features to the whole construction. Not just **PAR** alone, we need to take its hosting *frame* together as a whole to better account for its acoustic realizations.

One place to note is that, although the 'valley-shape' prosodic realization has been identified, the results do not render significant statistical tests for each acoustic feature. We wonder if this is due to the influence of higher-level discourse effects. Thus we attempt categorizing instances of **PAR-FRA** according to three relative positions by the *breath group unit* **BG**, namely **BG**-initial, -medial and -final. Results for F0 feature are summarized in Fig. 2. As can be seen, those **PAR-FRA** located at BG-medial position present the clearest 'valley-shape' realizations, all significant by statistical test. As for instances at BG-initial and -final, obviously they are under stronger influence of overall pitch contour output from higher discourse level, so the instances at initial positions reflects higher F0 in average and those at final positions been affected by the final lowering towards BG endings. This in turn explains why the results of F0 values from Fig 1. are not significant across **PAR-FRA**, since the average value has been neutralized as result of the location effect by discourse prosody from higher levels. Finally note that further examination for intensity and duration by removing discourse effect did not yield as significant results. In all, by the comparison of acoustic values it preliminarily establishes **PAR-FRA** as an integrated planning unit. Next we will turn to the calculation of *contrast degree* based on the current results.

4.3. Contrast degree

To further clarify the relationship between **PAR** and **FRA**, here we calculate *contrast degree* (CD) to provide additional evidences. Adhering to the methodology from 3.2, CD calculation is by **PPh** located at pre-/post- boundaries of **PAR**. Here we additionally calculate CD from boundaries of PPh units that do not contain any instance of **PAR-FRA**, with the assumption that **PAR-FRA** construction can be distinguished acoustically from larger context. Note that we consider only PPhs located at medial positions (M-PPh) by higher-level discourse units, e.g. excluding those at initial and final positions. The results are summarized in Fig 3.

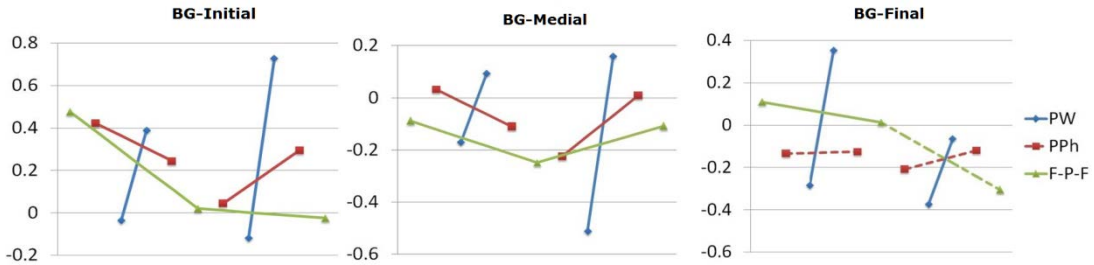

Figure 2: *Results of mean F0 values from **PAR-FRA** at different **BG** position.*

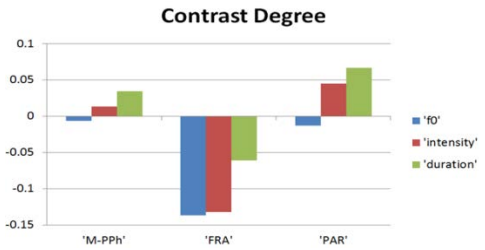

Figure 3: *Results of Contrast degree calculation.*

4.3.1. Discussion

As shown in Fig. 3, the most noticeable distinction lies in between **FRA** and those medial-PPhs (M-PPh) without **PAR-FRA**. Statistical tests reflect that significant differences can be found in CD of F0 and intensity (both h = 1, P <.05). As the hosting unit, therefore, **FRA** should be considered as a construction independent from the larger and general speech context. Interestingly, CD results by the **PAR** construction itself are quite similar to medial-PPhs in all respects. Since we assume that **PAR** is a nested structure within **FRA**, the CD results further substantiate that **PAR** in its respective *frame* **FRA** has been planned and realized as an integrated part of the host planning unit. Finally, we also test the CD derived from **PAR** and **FRA** respectively, the results indicate that significant differences can only be found in intensity, and F0 difference is only approaching significance (P=0.1).

4.4. Emphasis density

The calculation of *emphasis density* is mainly for illustrating perceived emphasis-attributed information planning across **PAR-FRA**. Here the density calculation has been employed on the tripartite structure **FRA**-(A)/**PAR**/**FRA**-(B). Following the methodology from 3.3, density scores were assigned arbitrarily according to levels of perceived prominences. Results are summarized in Fig 4.

4.4.1. Discussion

In Fig. 4, throughout **FRA**-(A) the results demonstrate a general tendency of *heavy-to-light* emphasis density distribution, whereas across **FRA**-(B) it presents rather smooth distribution with occasional local humps. As for **PAR**, no fix pattern seems to be identified out of the general distribution and only sporadically do we find some humps and also slight falling-rising pattern at times, depending on the length of the construction. Here we assume that the local humps may indicate the allocation of focal information in contributing to the heavier emphasis density distribution.

Since the results from Fig 4 is not presented as straightforward as expected, we attempt further score modification by assigning a negative score to the emphasis level of *reduction* E0. It turns out to be more valuable to foreground *reduction*: the results of emphasis density from **PAR** further stand out. So the valley-shape can be better observed across the construction of different length, although occasional local humps remain. As for **FRA**-(**A**), the general *heavy-to-light* information distribution is further strengthened, whereas the contour by **FRA**-(**B**) in general reveals a slight rising tendency. Hence via modeling the perceived emphasis-attributed information planning by taking reduction into account, the finding better illustrates the relationship between **FRA** and **PAR**: together **PAR-FRA** form a complete construction in terms of information planning, as the falling-rising density distribution withholds. Most of all from this trend of information distribution, it justifies why *parenthetical* is best considered nested within *frame* and together **PAR-FRA** treated as one instance of prosodic reduction in continuous speech.

5. GENERAL DISCUSSION, SUMMARY AND FUTURE WORK

Concentrating on *parenthetical* construction, this study takes the initial step towards prosodic *reduction* in continuous speech also in relation to higher level information planning and deployment. Defined mainly by perception, 80 instances of *parentheticals* **PAR** together with their bearing units *frame* **FRA** have been identified. The present analyses have been built on the assumption that together **PAR-FRA** co-constructs a unit within speech context, contributing to the information planning as a whole. Via systematic examinations the results reinforce that singling out **PAR** alone can only yield partial acoustic descriptions in isolation that is no more than local units in adjacency. Instead by considering **PAR-FRA** as a co-constructed unit, findings through acoustic analyses and emphasis density calculations yield a solid acoustic and information-oriented pattern throughout the whole construction. Further analyses by removing upper-level discourse effect strengthen the consistent 'valley-shape' pattern identified across **PAR-FRA**. Calculations of

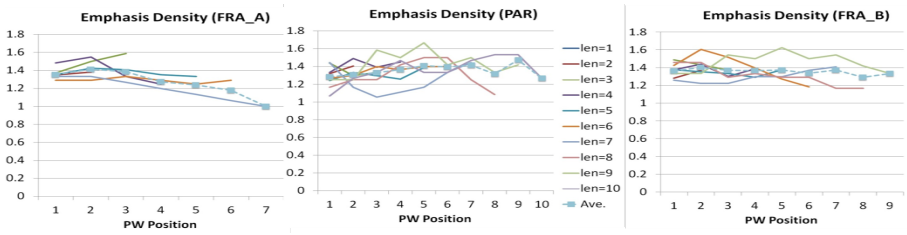

Figure 4: *Emphasis density of across F-P-F, with the weighting score [1 1 2 3] for [E0 E1 E2 E3].*

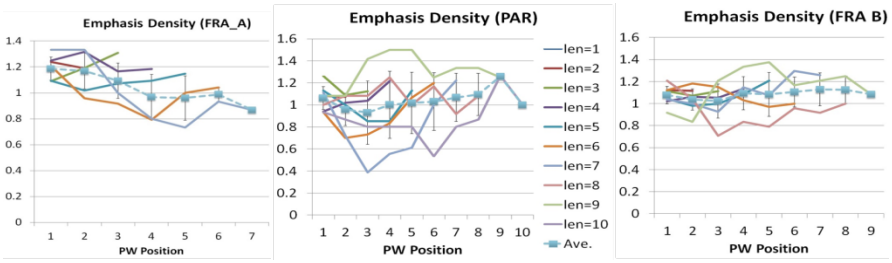

Figure 5: *Emphasis density across F-P-F, with the weighting score [-1 1 2 3] for [E0 E1 E2 E3].*

contrast degree otherwise offer better accounts for how **PAR** to be viewed as nested within **FRA**. In the end, *parentheticals* as an instance of prosodic reduction is best understood as operating within higher level discourse units; together **PAR-FRA** form a perceivable information planning unit within the speech context beyond mere syntactic planning.

To summarize, this study contributes to explorations towards prosodic *reduction* in continuous speech. It is within our understanding that *parentheticals* could pose only as one of the instances of prosodic reduction and speakers may as well resort to other means of reduction for various functions during speech production. Nevertheless, it is crucial to identify and pinpoint how prosodic reduction interweaves with its counterpart from the speech signals, namely perceived prosodic highlight, in order to sustain the expressiveness and coherence from speech production: it is held that neither of them is the result from randomization and their behavior patterns systematically. Above all, it is the allocation and compensation between perceived emphasis and reduction in such patterned ways that constitute the fundamentals to global context prosody, which may appear to be highly varied on the surface. For future studies we plan to investigate further into prosodic reduction in continuous speeches of different genres, as well as identification of other cases of prosodic reduction and how they interact with perceived emphases in patterns, with the goal to further our understanding toward the nature of prosodic variations in continuous speech.

6. REFERENCES

[1] C. Tseng. "An F0 analysis of discourse construction and global information in realized narrative prosody," *Language and Linguistics*, 11. 2, pp. 183-218. 2010.

[2] C. Tseng. "Corpus Phonetic Investigations of Discourse Prosody and Higher Level Information," *Language and Linguistics*, 9. 3, pp. 659-719. 2008.

[3] H. Chen, W. Fang, and C. Tseng. "Advance Prosodic Indexing - Acoustic realization of prompted information projection in continuous speeches and discourses." *ISCSLP 2016 - The 10th International Symposium on Chinese Spoken Language Processing*, Tianjin, China. pp. 242-246. Oct. 17-20, 2016.

[4] H. Chen, W. Fang, and C. Tseng. "The Convergence of Perceived Prosodic Highlight for Discourse Prosody - A Cross-Speech Genre Analysis." *Speech Prosody 2016*. Boston, USA. pp. 654-658. 2016.

[5] H. Chen, W. Fang, and C. Tseng, "Prosodic prompts of information content in speech - A cross genre comparison of prominence as key, projector and projections," *IACL 2016 – the 24th Annual Conference of International Association of Chinese Linguistics*, Beijing, China, July 17-19, 2016.

[6] N. Dehé. *Parentheticals in spoken English: The syntax-prosody relation*. Cambridge University Press, London, 2014.

[7] N. Burton-Roberts. "Parentheticals." *Encyclopedia of Language and Linguistics*, E. Brown Ed. Elsevier Science, pp.179-182. 2006.

[8] L. Grenoble. "Parentheticals in Russian," *Journal of Pragmatics*, 36.11, pp. 1953-1974. 2004.

[9] H. Bussmann, *Routledge Dictionary of Language and Linguistics*, Routledge, London. 1996.

[10] M. Payá, "Prosody and pragmatics in parenthetical insertion in Catalan" *Catalan Journal of Linguistics*, 2, pp. 207-227. 2003.

[11] H. Mazeland. "Parenthetical sequences," *Journal of Pragmatics*, 39.10, pp. 1816-1869. 2007.

[12] O. Duvallon and S. Routarinne, "Parenthesis as a resource in the grammar of conversation." *Syntax and lexis in conversation: Studies on the use of linguistic resources in talk-in-interaction*, A. Hakulinen, and M. Selting, Ed. John Benjamins, Amsterdam. pp. 45-74. 2005.

[13] C. Tseng and Z. Su, "Spontaneous Mandarin Speech Prosody—the NTU DSP Lecture Corpus." Proceeding of *Oriental COCOSDA 2008* Kyoto, Japan. pp.171-174. 2008.

[14] C. Tseng, S. Pin, Y. Lee, H. Wang, and Y. Chen. "Fluent speech prosody: Framework and modeling," *Speech Communication*, 46. 3-4, pp. 284-309, 2008.

[15] C. Tseng and C. Su. "Discourse prosody and context – Global F0 and tempo modulations," Proceedings of *INTERSPEECH 2008 – 9th Annual Conference of the International Speech Communication Association*, Brisbane, Australia, pp.1200-1203, 2008.

[16] J.-S. Jang, "Speech and Audio Processing (SAP) Toolbox", retrieved from http://mirlab.org/jang/matlab/toolbox/sap.

[17] C. Tseng and C.-H. Chang. "Pause or No Pause? –Prosodic Phrase Boundaries Revisited." *Tsinghua Science and Technology* 13.4, pp.500-509. 2008.