Prosody Labeling and Modeling for Mandarin Spontaneous Speech

Yu-Lun Chou[#], Chen-Yu Chiang[#], Yih-Ru Wang[#], Hsiu-Min Yu[*], Sin-Horng Chen[#]

[#]*Dept. of Communication Engineering, NCTU, Taiwan,* [*]*Language Center, Chung Hua University, Taiwan*

***Abstract*—**An unsupervised joint prosody labeling and modeling (PLM) method for exploring the prosody of spontaneous Mandarin speech is proposed. It is designed to automatically label a speech corpus and construct prosodic models simultaneously. Experimental results on a large dialog corpus confirmed its effectiveness. Many meaningful characteristics of spontaneous-speech prosody were investigated from the parameters of the well-trained prosodic models. The prosodic feature patterns of high-level constituents of the postulated prosody hierarchy were derived. An analysis of disfluencies related to the labeling results was also discussed. Those findings would provide rich prosodic information for various speech processing applications.

I. Introduction

In recent years, prosodic information are widely used in spontaneous speech processing for the detection of disfluencies and sentence-like boundaries [1-2], the segmentation of dialog acts [3], the improvement of automatic speech recognition (ASR) [4], etc. But the research progress is still very limited because of the following two difficulties. One is the need of preparing a large spontaneous speech corpus with prosody tags being properly labeled. Another is the lack of a sophisticated prosody modeling method.

In this paper, an unsupervised joint prosody labeling and modeling (PLM) method for spontaneous Mandarin speech is proposed. It is an extended version of the previous PLM method [5] proposed for read Mandarin speech. It labels a speech corpus with two types of prosody tags, the break types of inter-syllable junctures and the prosodic states of syllables, and builds eight prosodic models automatically. These two types of prosody tags are used to construct a prosody hierarchy of Mandarin speech, while these eight prosodic models are used to describe the relationships of acoustic prosodic features, prosody tags of utterances, and the linguistic features of the associated texts.

The remaining of the paper is organized as follows. Section II presents the proposed PLM method. Section III introduces the speech corpus used in this study. Section IV discusses the experimental results. Some conclusions are given in the last section.

II. The Proposed Method

A prosody hierarchy [5,6] shown in Fig. 1 is adopted for the speech prosody study. It employs four layers to describe the prosodic constituents for the normal fluent speech part, while it uses two layers for the other particular sound parts like disfluencies, particles, uncertain pronunciation, and over-lengthening due to hesitation. The layers are syllable (SYL), prosodic word (PW), prosodic phrase (PPh), and breath group/prosodic phrase group (BG/PG) for normal speech; and SYL and particular prosodic constituent (PPC) for particular sound. Two kinds of prosody tags are employed to structure the prosody hierarchy. One is a set of ten break types of syllable juncture, $B_n \in$ {B0, B1, B2-1, B2-2, B2-3, B3, B4, BPI, BP, BPO}, used to delimit these prosody layers. Here, B0 and B1 represent an intra-PW boundary with adjacent syllables being tightly and normal coupled; B2-1, B2-2 and B2-3 are divided from B2 and defined as a PW boundary with obvious F0 reset, perceived short pause and pre-boundary lengthening; B3 and B4 represent major breaks with median and long pause durations, respectively; BPI, BP, and BPO represent entrance, intermediate, and exit of PPC, respectively. The prosodic state is used to characterize the variation of a prosodic feature of a syllable in a prosodic constituent. In this study, three separate types of prosodic state are used, respectively, for the three prosodic features of syllable log-F0 contour $\mathbf{sp}_n$, syllable duration sd_n, and syllable energy level se_n.

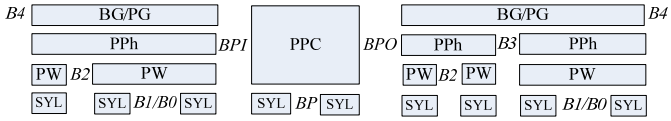

Fig. 1. The prosody hierarchy for Mandarin spontaneous speech.

Some inter-syllable acoustic features are extracted to characterize the break type of syllable juncture, including pause duration pd_n, energy-dip level ed_n, normalized pitch jump pj_n, and normalized duration lengthening factor dl_n. Besides, some linguistic features related to speech prosody are also extracted from the associated texts. They include lexical tones t_n, base-syllable type s_n, final type f_n of syllable n, and some high-level features $\mathbf{l}_n$ around syllable n for normal speech; and syllable-like classes pr_n for particular sound.

The proposed PLM method is formulated as a parametric optimization problem to find the best of prosodic tag sequence $\mathbf{T}=\{ B_n , p_n , q_n , r_n \mid n=1\sim N\}$ given with the acoustic feature sequence $\mathbf{A}=\{\mathbf{sp}_n, sd_n, se_n, pd_n, ed_n, pj_n, dl_n \mid n=1\sim N\}$ of the input speech utterance and the linguistic feature sequence $\mathbf{L}=\{\mathbf{l}_n, t_n, s_n, f_n, pr_n \mid n=1\sim N\}$ of the associated text:

$$B_1^{N*},p_1^{N*},q_1^{N*},r_1^{N*}=\mathbf{T}^*=\arg\max_{\mathbf{T}} P(\mathbf{T}|\mathbf{A},\mathbf{L})=\arg\max_{\mathbf{T}} P(\mathbf{T},\mathbf{A}|\mathbf{L}) \quad (1)$$

where

$$\begin{aligned} P(\mathbf{T},\mathbf{A}|\mathbf{L}) &= P(B_1^N,p_1^N q_1^N,r_1^N,\mathbf{sp}_1^N,sd_1^N,se_1^N,pd_1^N,ed_1^N,pj_1^N,dl_1^N|\mathbf{L}) \\ &\approx \prod_{n=1}^{N} P(\mathbf{sp}_n|p_n,B_{n-1}^n,\mathbf{L})P(sd_n|q_n,B_{n-1}^n,\mathbf{L})P(se_n|r_n,B_{n-1}^n,\mathbf{L}) \\ &\quad P(p_1)P(q_1)P(r_1)\prod_{n=2}^{N} P(p_n|p_{n-1},B_{n-1})P(q_n|q_{n-1},B_{n-1})P(r_n|r_{n-1},B_{n-1}) \\ &\quad \prod_{n=1}^{N-1}\left(P(pd_n,ed_n,pj_n,dl_n|B_n,\mathbf{l}_n)P(B_n|\mathbf{l}_n)\right) \end{aligned} \quad (2)$$

$P(\mathbf{sp}_n|p_n,B_{n-1}^n,\mathbf{L})$, $P(sd_n|q_n,B_{n-1}^n,\mathbf{L})$, and $P(se_n|r_n,B_{n-1}^n,\mathbf{L})$ are, respectively, syllable pitch contour, duration and energy-level models; $P(p_n|p_{n-1},B_{n-1})$, $P(q_n|q_{n-1},B_{n-1})$ and $P(r_n|r_{n-1},B_{n-1})$ represent pitch, duration and energy prosodic state transition models; $P(pd_n,ed_n,pj_n,dl_n|B_n,\mathbf{l}_n)$ is the break-acoustics model describing the relationship of various intersyllable acoustic features with break and linguistic features; $P(B_n|\mathbf{l}_n)$ represents the break-syntax model that build the relationship between break type and linguistic features.

The three syllable prosodic feature models are then elaborated to consider some major affecting factors that control their variations:

$$P(\mathbf{sp}_n|p_n,B_{n-1}^n,\mathbf{L}) \approx \begin{cases} N(\mathbf{sp}_n;\boldsymbol{\beta}_{t_{n-1}^{n+1},B_{n-1}^n}+\boldsymbol{\beta}_{p_n}+\boldsymbol{\mu},\mathbf{R}) & \text{for normal speech} \\ N(\mathbf{sp}_n;\boldsymbol{\beta}_{pr_n}+\boldsymbol{\beta}'_{p_n}+\boldsymbol{\mu},\mathbf{R}') & \text{for particular sound} \end{cases} \quad (3)$$

$$P(sd_n|q_n,B_{n-1}^n,\mathbf{L}) \approx \begin{cases} N(sd_n;\gamma_{t_{n-1}^{n+1},B_{n-1}^n}+\gamma_{s_n}+\gamma_{q_n}+\mu_d,R_d) & \text{for normal speech} \\ N(sd_n;\gamma_{pr_n}+\gamma'_{p_n}+\mu_d,R'_d) & \text{for particular sound} \end{cases} \quad (4)$$

$$P(se_n|r_n,B_{n-1}^n,\mathbf{L}) \approx \begin{cases} N(se_n;\alpha_{t_{n-1}^{n+1},B_{n-1}^n}+\alpha_{f_n}+\alpha_{r_n}+\mu_e,R_e) & \text{for normal speech} \\ N(se_n;\alpha_{pr_n}+\alpha'_{r_n}+\mu_e,R'_e) & \text{for particular sound} \end{cases} \quad (5)$$

where β_x , γ_x and α_x represent the affecting patterns (APs) of affecting factor x for syllable pitch, duration and energy models, respectively; $\boldsymbol{\mu}$ / μ_d / μ_e and $\mathbf{R}$ / R_d / R_e denote respectively the global means and the covariances of residuals. In the practical realization, tone with coarticulation APs, i.e. $\{\beta_x$, γ_x , $\alpha_x|x=_{t_{n-1}^{n+1},B_{n-1}^n}\}$, are obtained from decision trees for each current tone (t_n) constructed by a question set.

A special training procedure is designed. It first employs a sequential optimization to label prosodic tags and determine model parameters for the normal speech part of the unlabeled corpus. Then, the global means of normal speech models are used to assist in labeling prosodic tags and determine model parameters for the particular sound part. The reasons of adopting the special training procedure are stated as follows. First, it can prevent the training from being dominated by the particular syllable-like sounds which have wilder prosody variability. Second, using the same global means of normal-speech prosodic models in the training of particular sound part makes their APs have the same bases to compare.

III. Experimental Database

The database used consists of eight dialogue sessions selected from the Mandarin Conversational Dialogue Corpus (MCDC) [7] collected by the Institute of Linguistics of Academia Sinica, Taiwan. Its total length is about ten hours (121,242 syllables). It consists of 3,501 dialogue turns. The eight dialogue sessions were uttered by nine female and seven male speakers, and transcribed into Chinese texts with some other tags including discourse marker (DM), particles, and pauses by professional linguist annotators. Some important spontaneous speech phenomena were also annotated. They include disfluencies, particular pronunciation, discourse-related items, and sociolinguistic phenomena.

The preprocessing of the corpus included forced-alignment into syllable sequences, pitch detection, frame- and syllable-wise speaker normalizations of pitch and duration/energy levels, and representation of syllable pitch contour by four coefficients of orthogonal transformation.

IV. Experimental Results

The experiment was conducted on the MCDC corpus. The number of prosodic states was empirically set to be 20 including 16 for normal speech and 4 for PPC. The training took 59 iterations to reach a convergence.

A. *Analyses on Model Parameters*

Table I displays the total residual errors (TREs) which are the percentage of sum-squared residue of log-F0/syllable duration/syllable energy level over the observed sum-squared counterparts with respect to the use of different combinations of affecting factors. As shown in the table, TRE is reduced as more APs were used. The lower-level APs (i.e., tone with coarticulation, base-syllable, final type, and particular syllable-like types) accounted for 9.4%/16.3%/13.2% of prosodic variation in pitch/duration/energy level for the normal speech part, and 11.9%/3.4%,/8.6% for the particular sound part. The high-level prosodic constituents contributed another 76.5%/82.0%/84.3% and 39.8%/82%/47.2% for these two parts. Obviously, the contributions of low-level APs are relatively small.

TABLE I: TREs (%) of The Syllable Pitch Contour, Duration and Energy Level Model w.r.t. The Use of Different Combinations of APs for (A) normal speech part and (B) patICular sound part.

	APs	Pitch	Duration	Energy
(A)	+Tone with coarticulation	90.6	94.0	94.3
	+Base syllable/final		83.7	86.8
	+Prosodic state	14.1	1.7	2.5
	APs	**Pitch**	**Duration**	**Energy**
(B)	+Particle Class	88.1	96.6	91.4
	+Prosodic State	48.3	18.6	44.2

Fig.2 shows the syllable duration APs of 5 tones, 82 reduced base-syllable types, and 24 particular syllable-like classes. As expected, Tone 3 is shorter and Tone 5 is shortened seriously. Most APs of particular syllable-like classes are lengthened, while only three (particles of "GE", "O", and uncertain pronunciation) are shortened.

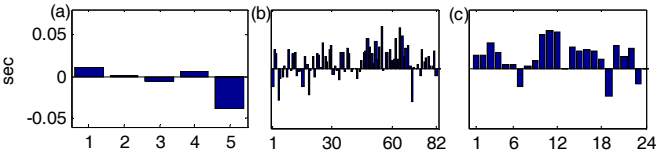

Fig. 2: The syllable duration APs of (a) tone, (b) base syllable, and (c) particular syllable-like type.

Fig.3 displays the prosodic state APs for normal speech (State 1 to 16), particular sound (State 17 to 19), and over-lengthening syllable (State 20). As shown in Fig. 3(b), the

over-lengthening syllable owing to hesitation has very large duration AP. From Fig. 3(c), State 17, corresponding to deep pronunciation of particle filler, has very small energy AP.

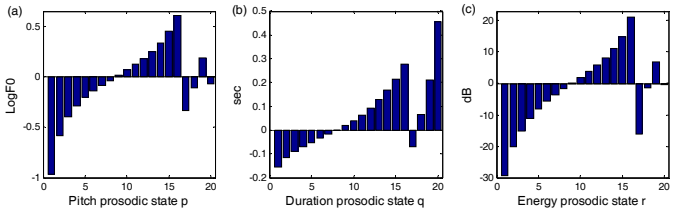

Fig. 3: The syllable (a) pitch, (b) duration and (b) energy APs of 16 normal-speech prosodic states and 4 particular syllable-like prosodic states.

Fig. 4 displays the distributions of 4 inter-syllable acoustic features for the seven break types of normal speech. Generally, break types of higher level were generally associated with longer pause duration, lower energy dip, larger normalized pitch jump, and larger normalized lengthening factor. These findings conform to our knowledge about break types.

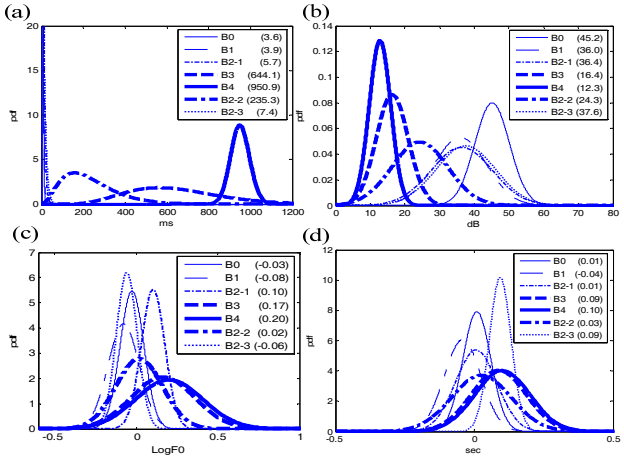

Fig. 4: The *pdf*s of (a) pause duration, (b) energy dip, (c) normalized pitch jump and (d) normalized lengthening factor for these 7 break types. Numbers in () denote the mean values.

From the prosodic state transition model of pitch $P(p_n|p_{n-1},B_{n-1})$, we found that B4 and B3 have significant pitch resets across the syllable boundary, and B2-1 also has obvious pitch resets. For B0 and B1, the general high-to-low, nearby-state transitions showed that the syllable log-F0 level declined slowly within PWs. For B2-2 and B2-3, no apparent pitch reset exists.

From $P(q_n|q_{n-1},B_{n-1})$, we found that the significant high-to-low transitions of B3 and B4 showed that PPhs and BG/PGs usually begin with lower states and ended with higher states to manifest the significant duration lengthening effect before major break junctures. B2-3 also had large high-to-low transition to imply the pre-boundary lengthening of minor break junctures.

The break-syntax model $P(B_n|\mathbf{l}_n)$ was built by the decision tree method using a linguistic question set. From the resulting decision tree, we found that most intra-word junctures were labeled as non-break B0 and B1. For inter-word junctures, their labelings were complicated. They were likely labeled as B2-1, B3, or B4 if the following word is a conjunction; as B2-1 and B1/B0 if it preceded or followed a pronoun.

B. Analysis on Speech Flow

1) Normal Speech

Table II lists the average lengths of PW, PPh, and BG/PG. Both the average lengths of PW and PPh were comparable to those of read speech [5], while the average length of BG/PG was much shorter. Two reasons caused BG/PGs to be shorter. One is that many speech turns are very short in this dialogue corpus. Another is due to the use of PPC in the postulated prosody hierarchy.

TABLE II: AVERAGE LENGTHS OF PW、PPH AND BG/PG. (UNIT:SYLLABLE)

PW		PPh		BG/PG	
mean	std	mean	Std	mean	std
2.55	1.92	5.23	5.89	6.27	8.74

To explore the general prosodic feature patterns of PW, PPh, and BG/PG for normal speech, we first extract the normalized prosodic feature by eliminating influences of tone, base-syllable type, final type, and the global means. For pitch feature, we obtain

$$\mathbf{pm}_n = \mathbf{sp}_n - \boldsymbol{\beta}_{t_{n-1}^{n+1},B_{n-1}^n} - \boldsymbol{\mu} \tag{6}$$

Sequences of $\mathbf{pm}_n(1)$ delimited by B2/B3/B4 at both sides are regarded as prosodic patterns formed by integrating the log-F0 level patterns of PW, PPh, and BG/PG. A superposition model is therefore defined by

$$\mathbf{pm}_n(1)=\mathbf{pm}_n^r(1)+\boldsymbol{\beta}_{PW_n}(1)+\boldsymbol{\beta}_{PPh_n}(1)+\boldsymbol{\beta}_{BG/PG_n}(1) \tag{7}$$

where $\mathbf{pm}_n^r$ is the residual; $\boldsymbol{\beta}_x$ represents APs of affecting factor x; $PW_n=(i,j)$, $PPh_n=(i,j)$, and $BG/PG_n=(i,j)$ denote that syllable n is located at the j-th place of an i-syllable PW, PPh, and BG/PG, respectively. A sequential optimization procedure based on the MMSE criterion is adopted to train the model. The patterns of PW, PPh, and BG/PG for duration and energy are similarly calculated.

Fig. 5 displays syllable duration patterns of PW, PPh and BG/PG. From Fig. 5, we found that the last syllables of all PPh and PW patterns were lengthened significantly, while those of most BG/PG patterns were shortened. These finding are similar to those of read speech [5,6]. For log-F0 level patterns, both PW and PPh are of falling patterns. We also find that BG/PGs have flat patterns with small dynamic range. Lastly, the energy patterns of these three prosody constituents are very similar to those of log-F0 patterns.

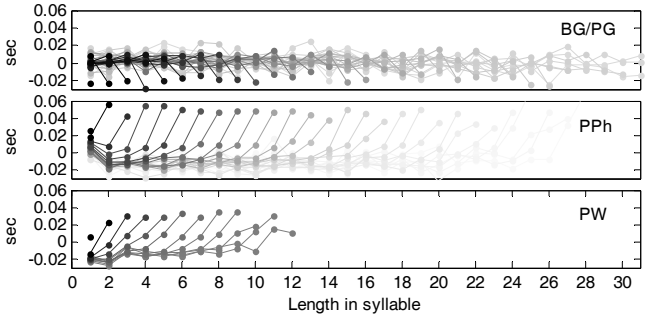

Fig. 5: The duration patterns of BG/PG, PPh and PW.

A typical prosody labeling example is given in Fig. 6. It can be found from the figure that the utterance is divided into

many PWs of 1-4 syllables. Most PWs are delimited by B2-1 with pitch reset. The insertion of the first B2-3 is due to hesitation. A major break B3 is set at the end of the first sentence. The speaker produces a DM (NE GE) with flat pitch and then followed by a PW "ji-long-yi-lan" with pitch accent to emphasize it. This example shows that our method functions quite well for automatic prosody labeling.

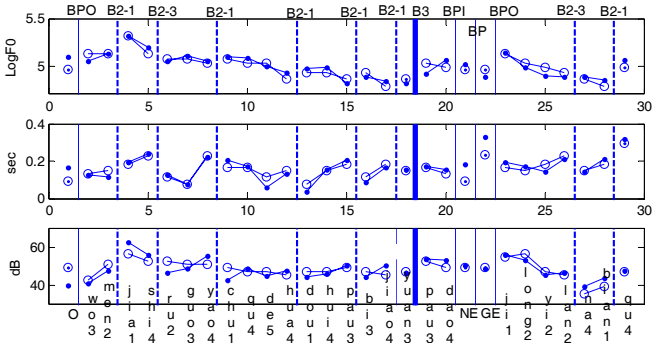

Fig. 6: A prosody labeling example: it shows the observed (solid diamond) and prosodic state+global mean (open circle) of syllable log-F0 level (Upper), duration (middle) and energy level (lower). The utterance is "O(O) wo3-men2(our) jia1-shi4(family is) ru2-guo3-yao4(if we want to) chu1-qu4-de5-hua4(have a trip), dou1-hui4-pau3(always go) bi3-jiao4(quite) yuan3(far) pau3-dao4(go to), NE GE(NE GE) ji1-long2-yi2-lan2(Keenlong and Yilan) na4-bian1-qu4(there).".

2) Edit Disfluencies

The structure of edit disfluencies can be expressed by

(reparandum) * [editing term] correction

Here, '*' indicates an interrupt point (IP). We now analysed three major types of IP: repetition, repair and restart. Their counts are 1379, 362, and 764. Fig. 7 displays the distribution of break tag labeling results. We found that all the three types of IP are more likely to be labeled as minor or major breaks than regular syllable junctures. Moreover, both repair and restart have more major breaks and B2-2 to show that they are likely to have long pause duration. On the contrary, most repetitions with B0/B1 are pragmatic repetitions.

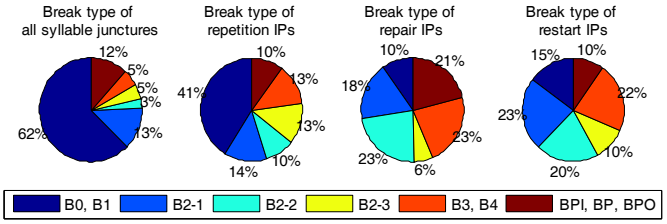

Fig. 7. IP corresponds to distribution of the break type

Fig. 8 displays the normalized prosodic feature patterns of reparandum and correction for repetition, repair, and restart IPs with different lengths. From Fig 8(a), the beginning pitch levels of corrections for the three types of IP are likely to be reset to the beginning pitch levels of reparandums. From Fig. 8(b), the pre-IP lengthenings of reparandum are reset and shortened for the beginning syllable of correction. From Fig.8(c), the beginning energy levels of correction for both repair and restart are likely to be reset to higher levels than those of reparandum. These findings match well with those of Tseng [8].

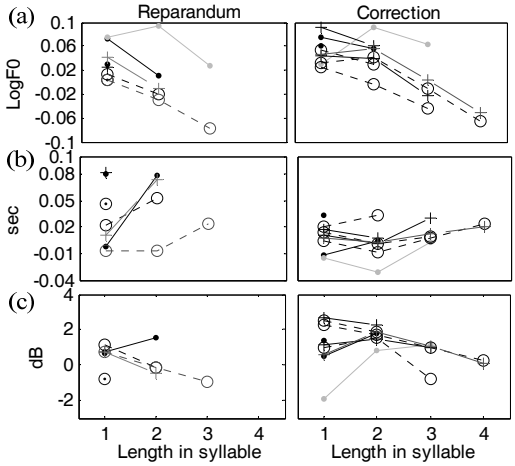

Fig. 8: (a) Log-F0, (b) duration and (c) energy patterns of reparandum and correction for repetition IP (solid line, solid diamond), repair IP (dotted line, open circle), and restart IP (dashed line, plus marker) with different lengths.

V. Conclusion

In this paper, an unsupervised joint prosody labeling and modeling (PLM) method for spontaneous Mandarin speech has been discussed. Experimental results on the MCDC corpus showed that rich and meaningful prosodic information can be explored from the well-trained prosodic models as well as from the automatically-labeled prosody tags. We believe that those findings should be beneficial to other spontaneous-speech. applications.

Acknowledgment

This work was supported by NSC under contract NSC98-2221-E-009-075-MY3. The authors want to thank Dr. S.-C. Tseng of Academia Sinica for providing the MCDC Corpus.

References

[1] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, "Enriching Speech Recognition with Automatic Detection of Sentence Boundaries and Disfluencies," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1526-1540, 2006.

[2] C. K. Lin, and L. S. Lee, "Improved Features and Models for Detecting Edit Disfluencies in Transcribing Spontaneous Mandarin Speech," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 17, no. 7, pp. 1263-1278, 2009.

[3] J. Kolar, E. Shriberg, and Y. Liu, "On speaker-specific prosodic models for automatic dialog act segmentation of multiparty meetings," in *Proc. of Interspeech 2006*, pp. 2014-2017.

[4] S. Ananthakrishnan and S. Narayanan, "Unsupervised Adaptation of Categorical Prosody Models for Prosody Labeling and Speech Recognition," *IEEE Transactions on Audio, Speech, and Language Processing,* vol.17, no.1, pp.138-149, 2009.

[5] C. Y. Chiang, S. H. Chen, H. M. Yu, and Y. R. Wnag, "Unsupervised Joint Prosody Labeling and Modeling for Mandarin Speech" *J. Acoust. Soc. Am.*, vol. 125, no. 2, pp. 1164-1183, 2009.

[6] C.-Y. Tseng, S.-H. Pin, Y.-L. Lee, H.-M. Wang and Y.-C. Chen, "Fluent speech prosody: framework and modeling," Speech Commun., vol.46, Issues 3-4, Special Issue on Quantitative Prosody modeling for Natural Speech Description and Generation, pp. 284-309, 2005.

[7] S. C. Tseng, "Processing spoken mandarin corpora," *Traitement Automatique des Langues*, vol. 45, no. 2, pp. 89–108, 2004.

[8] S. C. Tseng, "Repairs in Mandarin Conversation," *Journal of Chinese Linguistics*, vol. 34, no.1, pp. 80-120, 2006