Rich Prosodic Information Exploration on Spontaneous Mandarin Speech

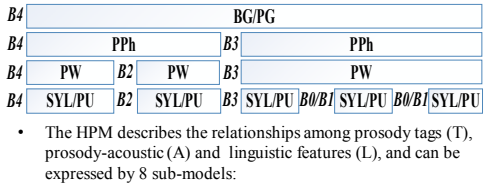
Cheng-Hsien Lin [1], Chung-Long You[1], Chen-Yu Chiang[2], Yih-Ru Wang[1], Sin-Horng Chen[1]
[1]Dept. of Electrical and Computer Engineering, National Chiao Tung University, Taiwan
[2]Dept. of Communication Engineering, National Taipei University, Taiwan

Summary

- The joint prosody labeling and modeling algorithm for Mandarin read speech is extended for spontaneous speech by additionally considering the affecting patterns:
  1. Particular Unit (PU) type: particle, marker, uncertain and foreign speech
  2. Part of speech for normal speech part
  3. Break types of syllable boundaries
  4. Syllable contraction and lengthening
- A hierarchical prosodic model (HPM) is constructed for mandarin spontaneous speech using MCDC corpus
- The prosodic characteristics and disfluency events of spontaneous Mandarin speech are explored by investigating the parameters of HPM

The Hierarchical Prosody Modeling

- The four-layer prosodic structure consisting syllable/particular unit (SYL/PU), prosodic word (PW), prosodic phrase (PPh), and breath group/prosodic phrase group (BG/PG) is adopted
- Two types of tags are employed to represent the prosody structure: **BREAK type** (syllable/PU juncture) and **PROSODY STATE** (inside syllable/PU)



- The HPM describes the relationships among prosody tags (T), prosody-acoustic (A) and linguistic features (L), and can be expressed by 8 sub-models:

$$P(\mathbf{T}|\mathbf{A},\mathbf{L},\Lambda) \quad \text{1} \quad \text{2}$$
$$\approx \prod_{n=1}^{N} \left[p(\mathbf{sp}_n|B_{n-1}^{n}, p_n, t_{n-1}^{n+1}, pos_n)\, p(sd_n|B_{n-1}^{n}, q_n, t_n, s_n, pos_n, cl_n) \right.$$
$$\left. p(se_n|B_{n-1}^{n}, r_n, t_n, f_n, pos_n) \right] \quad \text{3}$$
$$P(p_1)P(q_1)P(r_1) \prod_{n=2}^{N} \left[P(p_n|p_{n-1}, B_{n-1}) P(q_n|q_{n-1}, B_{n-1}) P(r_n|r_{n-1}, B_{n-1}) \right.$$
$$\left. \text{4} \quad \text{5} \quad \text{6} \right.$$
$$\left. \text{7} \right.$$
$$\prod_{n=1}^{N-1} \left[P(pd_n, ed_n, pj_n, dl_n|B_n, \mathbf{l}_n) P(B_n|\mathbf{l}_n) \right] \quad \text{8}$$

- Prosody tags include Breaks (B) and prosody states of pitch contour (p), duration (q), and energy (r) for syllable/PU

$$\mathrm{T} = \{B_n, p_n, q_n, r_n \mid n = 1 \cdots N\}$$

- (**1,2,3,7**) Prosody-acoustic feature related sub-models, representing within syllable/PU features: pitch (sp), duration(sd), energy(se), and between syllable/PU features: pause duration(pd), energy dip(ed), pitch jump(pj) and duration lengthening(dl)

$$\mathbf{A} = \{sp_n, sd_n, se_n, pd_n, ed_n, pj_n, dl_n | n = 1 \cdots N\}$$

- (**4,5,6**) Prosody-state sub-models of pitch, duration and energy
- (**8**) Break-syntax sub-model
- The linguistic feature set (L) includes: reduced linguistic feature(l), tone(t), base-syllable(s), final(f), part-of-speech(pos), and contraction/lengthening tag(cl)

$$\mathbf{L} = \{l_n, t_n, s_n, f_n, pos_n, cl_n | n = 1 \cdots N\}$$

- sp, sd, and se are further expressed:

$$p(\mathbf{sp}_n|B_{n-1}^{n}, p_n, t_{n-1}^{n+1}, pos_n) = \begin{cases} N(\mathbf{sp}_n; \beta_{t_n} + \beta^{f}_{B_{n-1}, t_{n-1}} + \beta^{b}_{B_{n-1}, t_{n-1}} + \beta_{pos_n} + \beta_{B_{n-1}, B_n} + \beta_{p_n} + \beta_{\mathbf{sp}}, \mathbf{R}^{r}_{\mathbf{sp}}), & \text{for SYL} \\ N(\mathbf{sp}_n; \beta_{pu_n} + \beta_{B_{n-1}, B_n} + \beta_{p_n} + \beta_{\mathbf{pu\_sp}}, \mathbf{R}^{r}_{\mathbf{pu\_sp}}), & \text{for PU} \end{cases}$$

$$p(sd_n|B_{n-1}^{n}, q_n, t_n, s_n, pos_n, cl_n) = \begin{cases} N(sd_n; \gamma_{cl_n} + \gamma_{t_n} + \gamma_{s_n} + \gamma_{pos_n} + \gamma_{B_{n-1}, B_n} + \gamma_{q_n} + \mu_{sd}, R^{r}_{sd}), & \text{for SYL} \\ N(sd_n; \gamma_{cl_n} + \gamma_{pu_n} + \gamma_{B_{n-1}, B_n} + \gamma_{q_n} + \mu_{pu\_sd}, R^{r}_{pu\_sd}), & \text{for PU} \end{cases}$$

$$p(se_n|B_{n-1}^{n}, r_n, t_n, f_n, pos_n) = \begin{cases} N(se_n; \alpha_{t_n} + \alpha_{f_n} + \alpha_{pos_n} + \alpha_{B_{n-1}, B_n} + \alpha_{r_n} + \mu_{se}, R^{r}_{se}), & \text{for SYL} \\ N(se_n; \alpha_{pu_n} + \alpha_{r_n} + \alpha_{B_{n-1}, B_n} + \mu_{pu\_se}, R^{r}_{pu\_se}), & \text{for PU} \end{cases}$$

- β_x γ_x α_x are the affecting patterns (APs) of affecting factors x for syllable/PU pitch contour, duration , and energy models
  $\boldsymbol{\mu}_x(\mu_x)$ $\mathbf{R}_x(R_x)$ denote the global mean vector and covariance matrix of modeling residual
- The unsupervised PLM algorithm is employed to simultaneously train the HPM and label the corpus with prosodic tags.

$$\mathbf{T}^{*}, \Lambda^{*} = \arg\max_{\mathbf{T},\Lambda} P(\mathbf{T}|\mathbf{A},\mathbf{L},\Lambda)$$

Database and Preprocess

- ***Database***

- MCDC 8-hour dialogues from 16 speakers
- Texts with phonetic and linguistic tags are transcribed by linguist experts
- Spontaneous speech characteristics: repetition/restart/repair, pause, pronunciation variation and sociolinguistic phenomena are also annotated

- ***Syllable alignment***

- HMM forced alignment with manual error correction

- ***Syllable acoustic features***

- Utterance level mean-and-variance normalization for duration and energy
- Speaker level mean-and-variance normalization for F0 values
- Syllable contour is represented by four Legendre Polynomial coefficients

Modeling results of HPM

- The total residual errors (TREs) of modeling for normal syllable

TRE: 100% * Var(orginal_feature-APx) / Var(orginal_feature)

Pitch		Duration		Energy	
AF	TRE	AF	TRE	AF	TRE
-		+Contra.& Lengthen. (cl_n)	81.2%	-	
+Tone (t_n)	90.7%	+Tone	78.9%	+Tone	94.4%
+Coarticul-ation (tp_n)	84.1%	+Syllable Type (s_n)	68.6%	+Final (f_n)	85.3%
+POS (pos_n)	77.2%	+POS	65.2%	+POS	82.4%
+Break (B_n)	71.7%	+Break	56.0%	+Break	80.1%
+Prosodic State (p_n)	8.0%	+Prosodic State (q_n)	3.6%	+Prosodic State (r_n)	2.3%

- TREs of modeling for PU

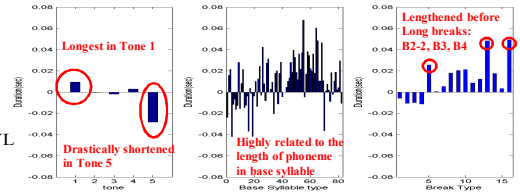
Pitch		Duration		Energy	
AF	TRE	AF	TRE	AF	TRE
-		+Contra.& Lengthen.	84.4%	-	
+PU ID (pu_n)	83.3%	+PU ID	72.2%	+PU ID	86.5%
+Break	76.3%	+Break	64.6%	+Break	81.4%
+Prosodic State	9.5%	+Prosodic State	2.9%	+Prosodic State	44.2%

- SYL/PU **duration APs** of **contraction**, **lengthening** and normal speech

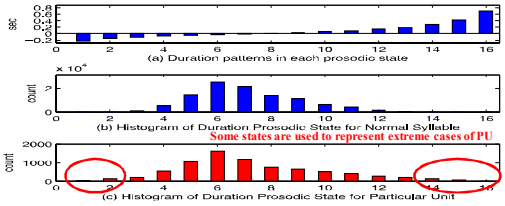
Largely elongated

	Contraction	Lengthening	Normal
SYL duration (*ms*)	-28	156	9
PU duration (*ms*)	-44	143	-19

- SYL **duration APs** of **tone**, **base-syllable**, and **adjacent breaks**



- SYL **duration APs** of **prosody states** and the histograms of normal syllable and PU



Analysis on Disfluency Events

- The utterance examples of **Repetition**

Type	Start Interrupt End Sentence
1	^(對)*(對)*(對)*(對)我有聽說所以像我先生講以前有一種說法 ^(dui)*(dui)*(dui)#wo you ting shuo suo yi xiang wo xian sheng jiang yi qian you yi zhong shuo fa
2	因為^(他)*(他)#舉了一個例子 yin wei^(ta)*(ta)#ju le yi ge li zi
3	待一整天對^(類似)*(類似)#亞歷山大那種 dai yi zheng tian dui^(lei si)*(lei si)#ya li shan da na zhong
4	吹涼風就覺得^(很舒服)*(很舒服)#所以這個 chui liang feng jiu jue de^(hen shu fu)*(hen shu fu)#suo yi zhe ge

- Break labeling results of repetition at different boundaries

Dominated by Non-pause breaks

Interrupt boundary							
Type	B0	B1	B2-1	B2-2	B2-3	B3	B4
1	28%	43%	9%	3%	15%	<1%	
2	7%	18%	25%	16%	9%	20%	4%
3&4	4%	11%	27%	15%	15%	19%	8%

Dominated by Short-to-long pause breaks

End boundary							
Type	B0	B1	B2-1	B2-2	B2-3	B3	B4
1	5%	5%	41%	8%	3%	29%	9%
2	19%	23%	37%	5%	11%	3%	1%
3&4	17%	29%	27%	6%	12%	6%	3%

Short stops for all three types

Utterance beginning

Start boundary								
Type	B0	B1	B2-1	B2-2	B2-3	B3	B4	Bs
1	2%	7%	14%	6%	8%	11%	6%	46%
2	8%	25%	22%	9%	3%	14%	5%	17%
3&4	9%	18%	23%	5%	5%	14%	6%	19%

Pitch resets

- The utterance examples of **Restart** and **Repair**

Restart	^(我父)*(我父親)#也沒有EI ^(wo fu)*(wo fu qin)#ye mei you EI
Repair	四十分鐘左右^(就到)*(就回到)#台北 si shi fen zhong zuo you^(jiu dao)*(jiu hui dao)#tai bei

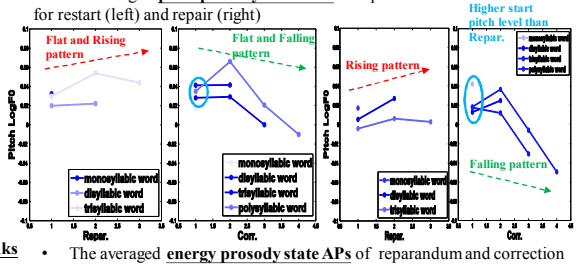
- Break labeling results of restart and repair at different boundaries

Restart							
Boundary	*B0*	*B1*	*B2-1*	*B2-2*	*B2-3*	*B3*	*B4*
Start	10%	29%	24%	8%	10%	14%	5%
Interrupt	2%	9%	34%	25%	9%	16%	5%
End	15%	37%	20%	4%	15%	7%	2%

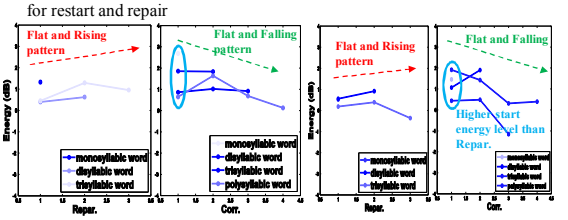
Stops exist in the speech flow

Repair							
Boundary	*B0*	*B1*	*B2-1*	*B2-2*	*B2-3*	*B3*	*B4*
Start	13%	30%	24%	7%	13%	10%	3%
Interrupt	3%	11%	29%	27%	10%	18%	2%
End	13%	36%	16%	5%	18%	10%	1%

- The averaged **pitch prosody state APs** of reparandum and correction for restart (left) and repair (right)



- The averaged **energy prosody state APs** of reparandum and correction for restart and repair



- The averaged **duration prosody state APs** of reparandum and correction for restart and repair