Recognizing Mandarin Chinese Fluent Speech Using Prosody Information—an Initial Investigation

Chiu-yu Tseng

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

Abstract

The aim of the present paper is to demonstrate how prosody information could be used to recognize Mandarin Chinese fluent speech and what the recognized results imply. By applying our hierarchical prosody framework for fluent speech [1, 2] that specifies boundary breaks and boundary information across phrases and group phrases into speech paragraphs, we were able to develop software that automatically segment speech flow by boundary breaks and label the boundaries systematically. That is, the recognized results are identified speech paragraphs and various levels of prosodic units within each such paragraph. These recognized prosodic units are not unrelated speech units but rather, sister constituents that entail higher-up syntactic as well semantic relationships that cumulatively make up speech paragraphs in fluent continuous speech. Note how this top-down approach differs from most bottom-up approaches. The former offers information from higher up linguistic association whereas the latter treats identified Chinese syllables as discrete unrelated units or lexical words at most, leaving structural information that combines these syllables into linguistically significant units unaddressed. We believe using top-down prosody information may very well offer new breaking ground in fluent speech recognition.

1. Introduction

Research on speech recognition of Mandarin Chinese has always adopted bottom-up approaches and geared towards recognizing syllables and tones and very little afterwards, and units of recognition have remained small [for example 3, 4 for more recent references]. The general consensus of choosing syllable as unit of speech recognition comes from two widely accepted misunderstandings. One is that Mandarin is a mono-syllabic language since the orthography is syllable based, and two a syllable could be a lexical word. Incidentally syllable is also the unit of lexical tones. But in fact, both assumptions are at best over-simplifications. In speech form, as opposed to in writing, Mandarin is not mono-syllabic and lexical words are not necessarily units of fluent continuous speech. The bottom-up approach itself is not altogether falsifiable since considerable success had been achieved over time. Nevertheless, what is misleading with the syllable-and-tone oriented approach is that firstly it is falsifiable to assume that all syllables in fluent Chinese speech are produced in full phonetic and tonal details, and secondly the approach somewhat reduces recognizing speech to recognizing mono- as well as poly-syllabic words only without addressing further structural (syntactic and/or semantic) information involved, and/or prosody information that is an inherent of speech. For example, recognizing all the syllables correctly in a phrase such as “下雨天留客天留我不留” without boundary (and in this case also prosodic) information would at best yield an ambiguous phrase with two possible readings while boundary information is required in speaking form to disambiguate them. “下雨--天留客--,天留--我不留.--” “Raining—heaven keeps visiting guests--, heaven keeps—I won’t” means ‘The rain may keep the visiting guests from leaving, but I the host won’t keep them’ whereas “下雨天--,留客天,--留我不留.--” “Rainy days—keep visiting guest days--, keep me not keep” means “Rainy days are days that keep your guest from leaving, so are you going to ask me (the guest) to stay or not?”

Our earlier perceptual investigations of boundary breaks in fluent continuous speech have shown that speech units are almost never mono-syllabic, lexical words are often not speech units and boundaries between lexical words may not always exist. [5] Labeling results indicated that speaking units are mostly di- and tri-syllabic in slower speech (mean syllable duration 200msec). Furthermore, we have also been able to establish a system of boundary breaks across phrases with the pauses themselves and corresponding pre-boundary information. We therefore argue that boundaries are important prosodic information, they are hierarchical, and they are necessary components of fluent speech [6]. These boundary breaks are also indicators of higher up structural information that concatenates speech units into meaningful phrases and paragraphs. Together with prosodic units at various levels, cross-phrase templates and cadences can be derived from speech corpora that justify and predicts fluent speech prosody [2].

The current hypothesis is if we apply what we have found in fluent speech prosody to speech recognition, we should be able to construct software that segment speech flow into speech units by boundary breaks, and label the breaks as specified by our prosody framework. Subsequently, the recognition results would be prosodic units that represent cross-phrase relationship instead of unrelated speech units. In the following sections, we will present our initial experiments to recognize fluent speech using boundary breaks and boundary information.

2. Speech Data

Mandarin Chinese speech data from Sinica COSPRO Database [7] were used. The speech data consisted of readings of 26 paragraphs (11592 syllables in total) of text ranging from 85 to 981 characters per paragraph by two speakers. 1

female (F051P) and 1 male (M051P) radio announcers, both under 35 years of age, read the text at a normal speaking rate of 200 ms/syllable. Segmental identities were first automatically labeled using the HTK toolkit and SAMPA-T notation, then hand tagged by trained transcribers for perceived boundary breaks using the Sinica COSPRO Toolkit [7]. All labeling was also spot-checked by trained transcribers. Segmental intensities were first derived using an ESPS toolkit. Table 1 summarizes derived speech features of the two speakers. Figure 1 shows distribution of PPh (Prosodic Phrase) length by syllable numbers from the two speakers.

	$\mu_{Duration}$	$\sigma_{Duration}$	$\mu_{Intensity}$	$\sigma_{Intensity}$	μ_{Pause}	σ_{Pause}
F051P	200	65	1298	680	37	106
M051P	190	60	897	350	45	138

Table 1 *Speech features in F051P and M051P*

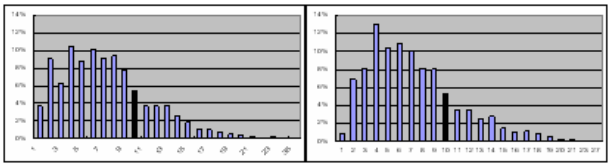

Figure 1 *Distribution of PPh (Prosodic Phrase) Length in syllable numbers from speakers F051P and M051P*

3. Methods of Analysis

We used normalized prosody information to build statistical models for each level of boundary break. Figure 2 shows prosody information based on Gaussian Mixture Model (GMM) to construct Break models. In this section, we will explain methods used in the system. Boundary break B1 denotes normal syllabic boundary (SY), B2 Prosodic Word (PW) boundary, B3 Prosodic Phrase (PPh) boundary, B4 Breath Group (BG) boundary and B5 Prosodic Phrase Group (BG) boundary [1, 2]. In other words, speech unites between two successively recognized B1's would be a syllable, between two successive B2's a PW, B3's a PPh, B4's a BG and B5's a PG. Note that models were built for each boundary break.

3.1.

To eliminate the variation between the speakers, each set of data was normalized with the mean and standard deviation of the entire class. The rationale of normalization is that boundary breaks would affect both pre- and post-boundary speech signals, but most obviously in the last 3 pre-boundary syllables. Our normalization spans to 12-syllable phrases; phrases longer were considered as 12-syllable phrases whereby the last 3 pre-boundary syllables were normalized. Therefore, the normalization is as follows:

$$Y_{nor(i)} = (Y_{(i)} - \mu_Y) / \sigma_Y$$

$$Y_{nor} = \{ Y_{nor(1)}, Y_{nor(2),} \ldots Y_{nor(n)} \}$$

3.2. Hierarchical Regression Model

A layered, hierarchical regression model corresponding to our prosody framework was built from bottom up, namely, the SY layer, the PW layer, the PPh layer, and the BG layer, to account for the respective as well as cumulative contribution of prosody information to the final output.

Syllable Layer :

Ynor = Const + CCt + CVt + Ton
+ PCt + PVt + PTt + FCt + FVt + FTt
+ 2-way factors of each factor above
+ 3-way factors of each syllable
+ Delta 1

PW Layer:
Delta 1 = f(PW length, PW sequence) + Delta 2

PPh Layer:
Delta 2 = f(PPh length, PPh sequence) + Delta 3

BG Layer:
Delta 3 = f(PPh IMF, PPh length, PPh sequence)
+ Delta 4

We used the same linear regression technique to build models for three acoustic modules, namely, Duration, Pause and Intensity Modules. Figure 3 shows Duration, Pause and Intensity patterns of different prosodic units for one speaker (F051P). Table 2 shows evaluations based on predictions of each prosody layer in duration, intensity and pause for both speakers (F051P and M051P).

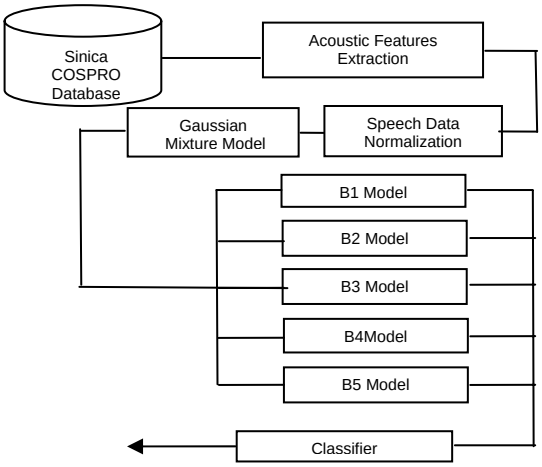

Figure 2 Schematic diagram of fluent speech recognition using prosody information where B's denote different levels of boundary breaks.

F051P		SY	PW	PP	BG
Duration	T.R.E.	46%	44%	39%	36%
	r	0.734	0.748	0.782	0.799
Intensity	T.R.E.	63%	62%	56%	54%
	r	0.611	0.613	0.662	0.682
Pause	T.R.E.	58%	54%	40%	32%
	r	0.649	0.681	0.799	0.827

M051P		SY	PW	PP	BG
Duration	T.R.E.	48%	44%	36%	33%
	r	0.718	0.747	0.805	0.822
Intensity	T.R.E.	56%	55%	51%	48%
	r	0.666	0.669	0.701	0.718
Pause	T.R.E.	50%	47%	34%	27%
	r	0.707	0.731	0.835	0.858

Table 2 *Prediction Evaluations in F051P and M051P*

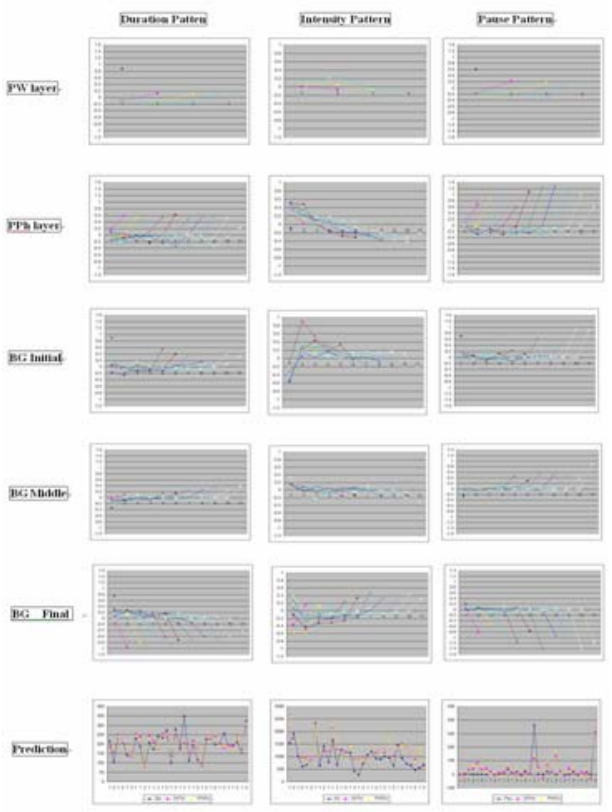

Figure 3 *Duration, Pause and Intensity patterns of different prosodic unit in speaker F051P*
X-axis: specific position with one syllable
Y-axis: corresponding regression coefficient of one syllable at the specific position
Row: Different linguistic levels and prediction
Column: Feature pattern

3.3. Gaussian Mixture Model

According to the above hierarchical regression model, we calculated Duration, Pause and Intensity patterns of F051P and M051P. Due to space limit of the present paper we could only present derived acoustic patterns from one speaker F051P as shown in Figure 3. However, the distinction of normalized acoustic patterns between speakers F051P and M051P was not apparent.

We used GMM to classify normalized prosody information. The GMM method has been widely used in the classification of speech recognition. We incorporated the results of linear regression of every prosodic layer, namely, the SY layer, PW layer, PPh layer, and BG layer, into GMM to analyze the speech data. Prosody information was trained to produce Break models, comprising B1, B2, B3, B4, and B5 model, by GMM. Figure 4 shows data dimension of GMM. Data dimension of GMM included $NP_{(x)}$, $ND_{(x-1)}$, $ND_{(x)}$, $NI_{(x)}$, $NI_{(x+1)}$, $NF_{(x)}$ and $NF_{(x+1)}$.

$NP_{(x)}$: Normalized Pause before Beak.
$ND_{(x-1)}$: Normalized Duration before Break.
$ND_{(x)}$: Normalized Duration after Break.
$NI_{(x)}$: Normalized Intensity before Break.
$NI_{(x+1)}$: Normalized Intensity after Break.
$NF_{(x)}$: Normalized Frequency before Break.
$NF_{(x+1)}$: Normalized Frequency after Break.

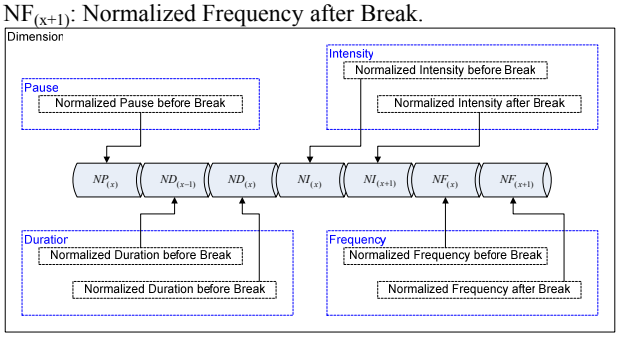

Figure 4 *Data Dimension of GMM*

4. Results and Discussion

Results of inside and outside tests for F051P and M051P were presented and discussed below.

4.1. Inside Test

Results of both precision rate and recall rate from inside test were obtained from speakers F051P and M051P and listed in Table 3, where PB1 denotes the precision number of B1 and RB1 denotes the recall number of B1. For inside test, training data sets were chosen from all speech data. Overall recognition accuracy of F051P and M051P were 94.1% and 94.8%, respectively. The predictions of boundary breaks were very good except for B2. However, the precision rate of F051P and M051P in B2 was the lowest, and the result indicated that B2's were misjudged as B1's in most errors.

F051P	PB1	PB2	PB3	PB4	PB5	Recall
RB1	6218	419	12	0	0	93.5178
RB2	234	3112	1	0	0	92.9788
RB3	5	1	1240	3	0	99.2794
RB4	0	0	1	209	0	99.5238
RB5	0	0	3	0	134	97.8102
Precision	96.2986	88.1087	98.6476	98.5849	100	

M051P	PB1	PB2	PB3	PB4	PB5	Recall
RB1	6246	414	3	0	0	93.7416
RB2	169	3154	4	0	0	94.8001
RB3	7	2	1194	2	2	98.9229
RB4	0	0	0	270	0	100
RB5	0	0	0	0	130	100
Precision	97.2594	88.3473	99.4172	99.2647	98.4848	

Table *3 Result of Inside Test in F051P and M051P*

One major reason why correctly recognizing PW boundary break B2 was less accurate could be that it is a lower-level boundary in the prosody hierarchy and may very well be speech signals with less significant information. Observations of speech data from our corpora showed that these boundaries are not at all likely to occur before or after any focus in the speech flow. In other words, B2 is a relatively less significant boundary break since no keyword would occur in its near neighborhood. So in spite of the relatively poor recognition rate, we have learned significant facts about fluent speech. When speech flows the signals are mixture of clearly and fully produced units such as keywords and/or focus and blurry signals such as rapid co-articulation where variations are to be expected. The question is: communication may very well be comprised largely from the focal points in the signals only and blurry portions could be skipped. We believe this notion merits further and future exploration. An immediate follow-up study would be to adjust weighting assignment of the models.

Future investigations of the occurrence and function of keywords in speech flow, and future development of keyword spotting in recognition system would both be desirable directions to incorporate and integrate.

4.2. Outside Test

Results of both precision rate and recall rate from outside test were obtained from speakers F051P and M051P and listed in Table 4. For outside test, training and testing data set were chosen by randomly dividing 75% of the speech data for training and the remaining 25% for testing. Overall recognition accuracy of F051P and M051P were 61.1% and 62.4%, respectively.

F051P	PB1	PB2	PB3	PB4	PB5	Recall
RB1	2090	490	63	0	0	79.0768
RB2	919	344	88	0	0	25.4626
RB3	35	32	371	79	7	70.8015
RB4	0	0	43	13	2	22.4138
RB5	0	0	39	8	13	21.6667
Precision	68.6597	39.7229	61.4238	13	59.0909	

M051P	PB1	PB2	PB3	PB4	PB5	Recall
RB1	2104	502	46	0	0	79.3363
RB2	859	446	116	0	0	31.3863
RB3	9	27	306	37	3	80.1047
RB4	0	0	88	32	4	25.8065
RB5	0	0	38	18	14	20
Precision	70.7941	45.7436	51.5152	36.7816	66.6667	

Table4. *Result of Outside Test in F051P and M051P*

The results of recognition from outside test were not as satisfactory. Besides, the recognition accuracy of B2 and B4 were very low for all Break levels. For B2 we believe our rationale for inside test as discussed in Section 4.1. also holds here for outside test. As for B4, analysis of the statistics of prosody information indicated that when there were insufficient training data, no significant difference in prosody information among Breaks could be derived. Take F051P for example, there was no significance difference in the duration of B3, B4 and B5; in addition, no significant difference was observed for the intensity in B4 and B5Based on the above the analysis. Future investigations would definitely include more speech data and more speakers.

5. Conclusions

In summary, results from the inside test is encouraging while the outside test suggests further improvement. However, we believe our initial investigation reported above at least showed a positive direction. Note that 1. we have shown that by systematically segmenting speech flow into prosodic units, our approach emphasized and captured the associative relationships between and among speech signals crucial to continuous speech. 2. We have shown how boundary breaks and boundary information could be essential to speech recognition. 3. We moved away from recognizing speech into phonetic units only, thereby suggesting how speech recognition could incorporate information expressed through prosody to further facilitate higher level structural information. 4. We provide evidence as to how far look-ahead and forecast span in continuous speech by locating boundary breaks between speech paragraphs. 5. Our rationale also suggests the notion of weighting or a mixture of focus-and-blur that exists in both human speech communication and speech comprehension. It should be common knowledge that in human speech communication not all sounds are produced in full phonetic details; whereas in human speech comprehension not all phonetic details need to be processed fully, either. Whatever approach developed for speech recognition should benefit from addressing these facts and somehow capture them as human so effortless do. Our initial attempt shows a first step towards this direction. We believe the idea could be further integrated to and with any bottom-up approaches, and further applied to other languages as well. Immediate future directions will focus on testing feasible alternatives such as adjusting weight assignments of prosody information for every Break level, and looking for bottom-up recognition systems for possible integration.

References

[1] Tseng, Chiu-yu, Pin, Shao-huang and Lee, Yeh-lin (2004). "Speech prosody: Issues, approaches and implications" in *From Traditional Phonology to Modern Speech Processing* (語音學與言語處理前沿), edited by Fant, G., Fujisaki, H., Cao, J. and Xu, Y., Foreign Language Teaching and Research Press (外語教學與研究出版社), 417-437, Beijing, China.

[2] Chiu-yu Tseng, ShaoHuang Pin and Yeh-lin Lee, Hsin-min Wang and Yong-cheng Chen(2005). "Fluent Speech Prosody: Framework and Modeling", *Speech Communication (Special Issue on Speech Prosody)*, Vol. 46:3-4, 284-309.

[3] Zhou, Jian-lai, Ye Tian, Yu Shi, Chao Huang and Eric Chang "Tone Articulation Modeling for Mandarin Spontaneous Speech Recognition", ICASP 2004, Jeju, Korea 997-1000.

[4] j Surendran, Dino, Gina-Anne Levow and Yi Xu, "Tone Recognition in Mandarin Using Focus", Interspeech 2005, 3301-3304. Lisbon, Portugal.

[5] Chiu-yu Tseng and Bau-Ling Fu (2005). "Duration, Intensity and Pause Predictions in Relation to Prosody Organization" *Interspeech 2005*, (September 4-8, 2005), Lisbon, Portugal, 1405-1408.

[6] Tseng, Chiu-yu (2003). "Towards the organization of Mandarin speech prosody: Units, boundaries and their characteristics" *Proceedings of the 15th International Congress of Phonetic Science* (ICPhS-2003), (Aug. 3-9, 2003), Barcelona, Spain, 599-602.

[7] Sinica COSPRO and Toolkit http://www.myet.com.COSPRO