CHINESE TIMT: A TIMIT-LIKE CORPUS OF STANDARD CHINESE

Jiahong Yuan[1], Hongwei Ding[2], Sishi Liao[2], Yuqing Zhan[2], and Mark Liberman[1]

[1]Linguistic Data Consortium, University of Pennsylvania
[2]Institute of Cross-Linguistic Processing and Cognition, Shanghai Jiao Tong University

ABSTRACT

This paper describes an effort to build a TIMIT-like corpus in Standard Chinese, which is part of our "Global TIMIT" project. Three steps are involved and detailed in the paper: selection of sentences; speaker recruitment and recording; and phonetic segmentation. The corpus consists of 6000 sentences read by 50 speakers (25 females and 25 males). Phonetic segmentation obtained from forced alignment is provided, which has 93.2% agreement (of phone boundaries) within 20 ms compared to manual segmentation on 50 randomly selected sentences. Statistics on the number of tokens and mean duration of phones and tones in the corpus are also reported. Males have shorter phones/tones but more and longer utterance internal silences than females, demonstrating that males in this dataset speak faster but pause more frequently and longer.

Index Terms— TIMIT, Forced alignment, Maximum coverage, Standard Chinese

1. INTRODUCTION

Since it was created three decades ago, the TIMIT speech corpus has been widely used in speech science and speech technology development [1-3]. The great success of TIMIT prompted the ongoing effort at the Linguistic Data Consortium to create "Global TIMIT" – a series of TIMIT-like corpora in a number of languages [4].

The original TIMIT dataset contains a total of 6300 sentence tokens, 10 sentences spoken by each of 630 speakers from eight major dialect regions of the United States. The sentence prompts include 2 dialect "Shibboleth" sentences (SA), 450 phonetically-compact sentences (SX), and 1890 phonetically-diverse sentences (SI). The dialect "Shibboleth" and phonetically-compact sentences were elaborately designed whereas the phonetically-diverse sentences were selected from existing text sources.

The design of "Global TIMIT" adopts a scheme different from that of the original TIMIT. Instead of having 630 speakers and 10 sentences per speaker, the new design has 50 speakers and 120 sentences per speaker. This makes the corpus size comparable to the original TIMIT but requires much less time and effort for recruiting and recording. Among the 120 sentences read by a speaker, 20 are "Calibration" sentences, read by all speakers; 40 are "Shared" sentences, read by 10 speakers; and 60 are "Unique" sentences, read by only one speaker. The total number of sentence types is, therefore, 20 + 40*(50/10) + 60*50 = 3220. The design is summarized in Table 1.

Table 1: *The design of "Global TIMIT".*

Sentence Type	#Sentences	#Speakers /Sentence	Total	#Sentences /Speaker
Calibration	20	50	1000	20
Shared	200	10	2000	40
Unique	3000	1	3000	60
Total	3220		6000	120

The creation of a TIMIT-like corpus consists of three steps: design or selection of sentences; speaker recruitment and recording; and phonetic transcription and segmentation. This paper describes our effort to build Chinese TIMIT in these steps.

2. SENTENCE SELECTION

2.1. Candidate sentences

All sentences were selected from the corpus of Chinese Gigaword Fifth Edition [5], which is a comprehensive archive of newswire text data from Chinese news sources. 5000 candidate sentences were selected from the corpus by the following steps: 1. Extract sentences that are 10-20 characters long, excluding those containing characters that are not on the list of the 3500 most frequently used Chinese characters (现代汉语常用字表); 2. Manually go through the list of extracted sentences in a random order, to remove those with uncommon words (e.g., person or place names) or inappropriate meaning (e.g., politically sensitive viewpoints), and also to segment the sentences into words. This was done until a pool of 5000 candidate sentences was generated, which contain approximately 6600 unique words and 2200 unique characters.

Calibration, Shared, and Unique sentences were selected from the candidate pool using computer algorithms. A pronouncing dictionary was made for sentence selection and

phonetic segmentation. The dictionary and the sentence selection procedure are described in the following selections.

2.2. Pronouncing dictionary

The pronouncing dictionary only transcribes the canonical pronunciation of a word as appeared in the dataset. Only a few words have more than one pronunciation, for which all pronunciations were listed. *Hanyu Pinyin* was used to transcribe the pronunciation, including initials, finals, and tone. A final in Mandarin Chinese may consist of one or more vowels (or vowels and glides, depending on the adopted phonological analysis), with or without a nasal coda. Because /o/ and /uo/ occur in complementary distribution and the acoustic difference between the two finals is negligible [6], they were treated as the same final. /i/ has three pronunciation variants, often transcribed as [ɿ] (when appearing after an alveolar fricative/affricate), [ʅ] (when appearing after a retroflex fricative/affricate), and [i] (in all other contexts). The three variants were treated as different finals, /i/ for [i], /ii/ for [ɿ], and /iii/ for [ʅ]. In total, there were 21 initials and 36 finals. Tones were marked on the finals, including Tone1 through Tone4, and Tone0 for the neutral tone. The phonetic labels are listed in Table 2.

Table 2: *Phonetic labels (in Pinyin).*

Initials	b, p, m, f, d, t, n, l, g, k, h, j, q, x, zh, ch, sh, r, z, c, s
Finals	a, ai, an, ang, ao
	e, ei, en, eng, er
	i, ii, iii, ia, ian, iang, iao, ie, in, ing, iong, iu
	ong, ou
	u, ua, uai, uan, uang, ui, un, uo
	v, van, ve, vn *
Tones	1, 2, 3, 4, 0
Silence	sil

* "v" represents "ü" in *Pinyin*, "ii" is for [ɿ], and "iii" is for [ʅ].

2.3. Selecting sentences

Twenty Calibration sentences were selected from the candidate pool to cover the maximum number of (tone-independent) syllable types in the language. This problem is known to be NP-Hard, but it can be approximately solved using greedy approximation [7]:

Greedy Approximation:

```
1: covered set is empty
2: Repeat
3:   Pick the sentence with the maximum number of
       syllable types not in the covered set
4:   Add syllable types in the chosen sentence
       into the covered set
5: Until 20 sentences are selected
```

As illustrated in Figure 1, we randomized the candidate sentences before the selection, and repeated the procedure 1000 times to obtain 1000 sets of 20 sentences. The set that contains the most number of tone-independent syllable types was used as Calibration sentences.

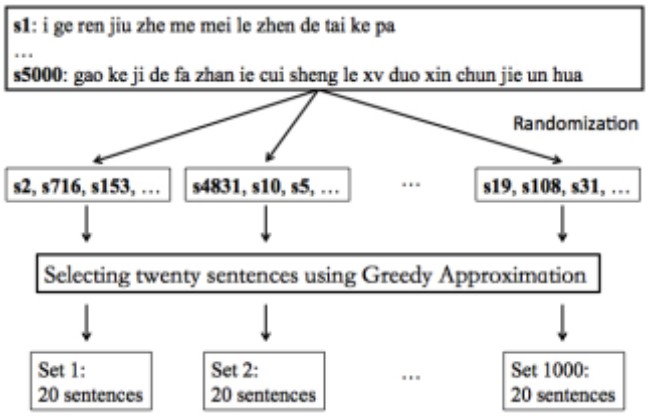

Figure 1: *Procedure for selecting Calibration sentences*.

Shared sentences were selected to cover the maximum number of tones and (within-word) tonal combinations. We need five sets of Shared sentences: each set has 40 sentences and will be read by 10 speakers. The first 20 sentences were selected to have at least five occurrences for each of the mono- and bi- tones. The second 20 sentences were selected to cover the maximum number of three- and four- tone combinations. The procedure was similar to that used for selecting Calibration sentences.

Unique sentences were randomly selected from the remaining sentences in the candidate pool. 50 sets of 60 sentences were selected, each to be read by one speaker only.

3. SPEAKER RECRUITMENT AND RECORDING

50 college students at Shanghai Jiao Tong University, 25 females and 25 males, were recruited to read the sentences. All of them speak Standard Chinese.

As a criterion to determine whether a subject speaks Standard Chinese, his/her spoken Mandarin proficiency assessed by *Putonghua Shuiping Ceshi* (which is the national standard Mandarin proficiency test) was used. There are seven levels of proficiency assessed by the test, which are, from highest to lowest: Class 1 Level 1, Class 1 Level 2, Class 2 Level 1, Class 2 Level 2, Class 3 Level 1, Class 3 Level 2, and Failed. In order to qualify for teaching K-12, one must pass Class 2 Level 2. The speakers recruited for the experiment all achieved Class 2 Level 1 or better on *Putonghua Shuiping Ceshi*.

The recording was made in a sound-treated recording booth at Shanghai Jiao Tong University, using the SpeechRecorder Software [8]. The sentences were displayed on a computer screen for subjects to read, one at a time, controlled by the person who monitored the recording.

A total of 6000 utterances were recorded, 120 utterances for each speaker.

4. PHONETIC SEGMENTATION

4.1. Forced Alignment

HMM/GMM-based forced alignment was applied to obtain phonetic segmentation. In prior work [9,10], we demonstrated that employing explicit phone boundary models within the HMM framework could significantly improve forced alignment accuracy for both English and Mandarin Chinese. The phone boundary models were a special 1-state HMM (as shown in Figure 2), in which the state cannot repeat itself:

Figure 2: *Special 1-state HMM for phone boundaries with transition probabilities* $a_{01} = a_{12} = 1$.

Therefore, a boundary can have one and only one state occurrence, i.e., aligned with only one frame. The special 1-state phone boundary HMMs were combined with standard monophone HMMs. Given a phonetic transcription, phone boundaries were inserted between phones. For example, "*sil i g e sil*" becomes "*sil sil_i i i_g g g_e e e_sil sil*". The boundary states were tied through decision-tree based clustering, similar to triphone state tying developed in speech recognition.

We started with the acoustic models trained on Hub4 Mandarin Broadcast News Speech [11], and retrained the models by combining the Broadcast News Speech data and our recordings (Training on the combined data sets had better results than training on Chinese TIMIT data only). Tone-independent models were employed. The acoustic features were the standard 39 PLPs extracted with 25 ms Hamming window and 10 ms frame rate. Initials, monophthong finals (/a, e, i, ii, iii, u, v/), and silence were 3-state HMMs, all other finals (including diphthongs, triphthongs, and nasal-coda finals) were 5-state HMMs. Each state had 2 Gaussian mixture components with diagonal covariance matrices. The system was built using the HTK Toolkit [12].

4.2. Evaluation of segmentation accuracy

To evaluate segmentation accuracy, 50 randomly selected sentences were manually corrected by three of the authors. Excluding the boundaries between silence and a stop or an affricate, where the boundary cannot be determined because of the stop closure, there are 1431 boundaries in the 50 sentences. 93.2% of the boundaries (1333 boundaries) have an agreement of within 20 ms between forced alignment and manual segmentation, which is on par with state-of-the-art results in terms of accuracy of automatic phonetic segmentation.

5. STATISTICS OF THE CORPUS

5.1. Statistics of phones

Based on the phonetic segmentation of the corpus, we calculated the total number of occurrences of every phone and its mean duration. The results are listed in Table 3, in which males and females are calculated separately.

Table 3: *Number of tokens and mean duration of phones in the corpus*.

Phone	#tokens (all)	Male		Female	
		#	duration (sec.)	#	duration (sec.)
/b/	3827	1928	0.0699	1899	0.0714
/p/	969	489	0.1062	480	0.1159
/m/	3558	1805	0.0714	1753	0.0685
/f/	2383	1207	0.0925	1176	0.0964
/d/	8849	4423	0.0547	4426	0.0559
/t/	3527	1769	0.1004	1758	0.109
/n/	1871	916	0.0666	955	0.0707
/l/	4774	2374	0.0537	2400	0.0542
/g/	4307	2158	0.0709	2149	0.0726
/k/	1978	974	0.1111	1004	0.1208
/h/	3818	1917	0.0961	1901	0.1016
/j/	6370	3126	0.0881	3244	0.0916
/q/	2860	1423	0.1178	1437	0.1243
/x/	4585	2225	0.1058	2360	0.1127
/zh/	5868	2969	0.083	2899	0.0875
/ch/	2731	1384	0.1151	1347	0.1228
/sh/	6821	3446	0.1081	3375	0.12
/r/	2097	1053	0.0733	1044	0.0721
/z/	2980	1493	0.0828	1487	0.0867
/c/	1421	712	0.1234	709	0.1287
/s/	1306	651	0.1176	655	0.1251
/a/	3182	1600	0.1037	1582	0.1099
/e/	8730	4423	0.0765	4307	0.0814
/i/	7449	3709	0.1018	3740	0.1156
/ii/	1314	672	0.0843	642	0.0871
/iii/	4614	2310	0.0808	2304	0.0834
/u/	5324	2703	0.0924	2621	0.0974
/v/	1944	943	0.105	1001	0.1066
/ai/	3807	1899	0.1187	1908	0.1289
/ao/	2497	1278	0.1266	1219	0.1333

/ei/	1368	686	0.1066	682	0.119
/er/	291	141	0.1788	150	0.1896
/ia/	1036	532	0.1394	504	0.1531
/iao/	1879	942	0.1398	937	0.1507
/ie/	1915	977	0.1271	938	0.1388
/iu/	2281	1148	0.1428	1133	0.1482
/ou/	2015	1004	0.1219	1011	0.121
/ua/	431	220	0.1505	211	0.1594
/uai/	289	136	0.1711	153	0.185
/ui/	2715	1329	0.116	1386	0.1255
/uo/	4088	2034	0.1241	2054	0.1251
/ve/	925	455	0.1146	470	0.1229
/an/	2986	1478	0.1317	1508	0.1443
/ang/	2802	1431	0.1358	1371	0.1417
/en/	4040	2045	0.1115	1995	0.1185
/eng/	2665	1316	0.1274	1349	0.1325
/ian/	3769	1861	0.1443	1908	0.1551
/iang/	1809	875	0.1504	934	0.1621
/in/	1961	958	0.1295	1003	0.14
/ing/	3148	1559	0.1369	1589	0.143
/iong/	249	119	0.1863	130	0.1972
/ong/	3200	1598	0.1363	1602	0.1399
/uan/	1080	543	0.1455	537	0.1572
/uang/	943	465	0.1707	478	0.1759
/un/	737	361	0.1438	376	0.1536
/van/	766	376	0.1815	390	0.1873
/vn/	475	215	0.1561	260	0.1558
Pause (all)	1730	976	0.2389	754	0.2073
Pause (Calibration)	326	179	0.2609	147	0.2227

Interestingly, we can see from the table that males have a shorter duration across phones than females. Paired-saples t-test shows that the difference is statistically significant ($p < 0.001$). This result suggests that males speak faster than females. On the other hand, however, males made more pauses (976 vs. 754) and longer pauses (0.2389 sec. vs. 0.2073 sec.) than females in the corpus (Utterance internal silences that are longer than 50 ms were counted as pauses). Because textual factors such as sentence length and syntactic complexity affect pause production, we also calculated pauses in the Calibration sentences only to remove the effects of those factors on the difference between males and females (they read the same sentences). The result is listed at the end of Table 3. For the Calibration sentences only, still, males made more pauses (179 vs. 147) and longer pauses (0.2609 sec. vs. 0.2227 sec.) than females.

5.2. Statistics of tones

The number of tokens and mean duration of tones (entire syllables) are listed in Table 4 and shown in Figure 3. We can see that Tone0 is the shortest; Tone1 and Tone2 are longer than Tone3 and Tone4. And again, males have a shorter duration on every tone than females.

Table 4: *Number of tokens and mean duration of tones in the corpus*.

Tone	#tokens (all)	Male		Female	
		#	duration (sec.)	#	duration (sec.)
T1	18674	9371	0.2027	9303	0.2153
T2	17882	8948	0.2047	8934	0.2153
T3	16408	8194	0.1875	8214	0.1968
T4	29158	14513	0.1899	14645	0.2031
T0	6602	3315	0.1347	3287	0.141

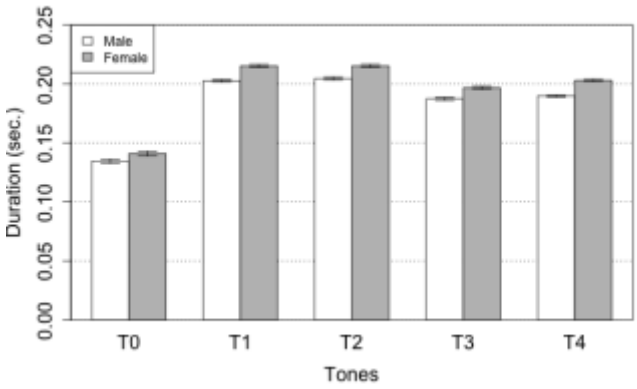

Figure 3: *Mean duration of tones in the corpus*.

6. CONCLUSION

In this paper, we detailed the development of a TIMIT-like corpus in Standard Chinese. A simple analysis of the corpus shows that males speak faster but pause more frequently and longer than females. This result is consistent with our previous investigation of this topic based on telephone conversations and monologue speech [13, 14].

Along with Chinese TIMIT, we have also created an L2 English TIMIT, for which the same 50 speakers read "easy" sentences selected from the original TIMIT. We plan to extend the effort to L2 Chinese and L1 English, to make a basis for four-way comparison between L1 and L2 and between Chinese and English.

6. REFERENCES

[1] V. Zue, "Speech Database Development," Final Technical Report submitted to the Defense Advanced Research Projects Agency (for Contract #00039-85-C-0341, June 1985 - June 1987), 1988.

[2] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Communication* 9(4), pp. 351-356, 1990.

[3] Garofolo, J., *et al.*, *TIMIT Acoustic-Phonetic Continuous Speech Corpus* (LDC93S1), Linguistic Data Consortium, 1993.

[4] N. Chanchaochai, J. Yuan, J. Wright, C. Cieri, and M. Liberman, "Global TIMIT: Towards Creating TIMIT-analogous Speech Corpora," manuscript.

[5] Parker, R., *et al.*, *Chinese Gigaword Fifth Edition* (LDC2011T13), Linguistic Data Consortium, 2011.

[6] J. Yuan, "The spectral dynamics of vowels in Mandarin Chinese," *Proceedings of Interspeech 2013*, pp. 1193-1197, 2013.

[7] U. Feige, "A Threshold of ln *n* for Approximating Set Cover," *J. of the ACM* 45(5), pp. 634–652, 1998.

[8] C. Draxler and K. Jänsch, "SpeechRecorder - a Universal Platform Independent Multi-Channel Audio Recording Software," *Proceedings of LREC*, pp. 559-562, 2004.

[9] J. Yuan, N. Ryant, M. Liberman, A. Stolcke, V. Mitra, and W. Wang, "Automatic phonetic segmentation using boundary models", *Proceedings of Interspeech 2013*, pp. 2306-2310, 2013.

[10] J. Yuan, N. Ryant, and M. Liberman, "Automatic phonetic segmentation in Mandarin Chinese: Boundary models, glottal features and tone," *Proceedings of ICASSP 2014*, pp. 2539-2543, 2014.

[11] Huang, S., *et al.*, *1997 Mandarin Broadcast News Speech (HUB4-NE)* (LDC98S73), Linguistic Data Consortium, 1998.

[12] Young, S., *et al.*, *The HTK Book*, Web Download. http://htk.eng.cam.ac.uk

[13] J. Yuan, M. Liberman, and C. Cieri, "Towards an integrated understanding of speaking rate in conversation," *Proceedings of Interspeech 2006*, pp. 541-544, 2006.

[14] J. Yuan, X. Xu, W. Lai, and M. Liberman, "Pauses and Pause Fillers in Mandarin Monologue Speech: The Effects of Sex and Proficiency," *Proceedings of Speech Prosody 2016*, pp. 1167-1170, 2016.