AUTOMATIC PHONETIC SEGMENTATION IN MANDARIN CHINESE: BOUNDARY MODELS, GLOTTAL FEATURES AND TONE

Jiahong Yuan, Neville Ryant, Mark Liberman

Linguistic Data Consortium, University of Pennsylvania

ABSTRACT

We conducted experiments on forced alignment in Mandarin Chinese. A corpus of 7,849 utterances was created for the purpose of the study. Systems differing in their use of explicit phone boundary models, glottal features, and tone information were trained and evaluated on the corpus. Results showed that employing special one-state phone boundary HMM models significantly improved forced alignment accuracy, even when no manual phonetic segmentation was available for training. Spectral features extracted from glottal waveforms (by performing glottal inverse filtering from the speech waveforms) also improved forced alignment accuracy. Tone dependent models only slightly outperformed tone independent models. The best system achieved 93.1% agreement (of phone boundaries) within 20 ms compared to manual segmentation without boundary correction.

Index Terms— Forced alignment, boundary model, glottal features, tone, Mandarin Chinese

1. INTRODUCTION

The ability to use large speech corpora for research in many fields such as phonetics, sociolinguistics, and psychology, depends on the availability of phonetic segmentation and transcriptions. Automatic phonetic segmentation in English has been widely investigated and has achieved a level of accuracy comparable to human labelers. In this paper, we conducted experiments on automatic phonetic segmentation in Mandarin Chinese, which has been less studied [1,2].

The most common approach for automatic phonetic segmentation is to build a Hidden Markov Model (HMM) based forced aligner [3-8]. In this approach, each phone is a HMM that has typically 3-5 states. The speech signal is analyzed as a successive set of frames. The alignment of frames with phones is determined by finding the most likely sequence of hidden states (which are constrained by the known sequence of phones) given the observed data and the acoustic model represented by the HMMs. The phone boundaries are simply derived from the alignment of phone states with frames. This approach is different from the manual phonetic segmentation process, in which the acoustic landmarks at phone boundaries [9], e.g., an abrupt spectral change, are used to determine the location of a boundary. To utilize the spectral characteristics of phone boundaries, some researchers have applied local boundary refinement to HMM-based forced alignment. For example, [10] used energy changes in different frequency bands, [2] trained contextual-dependent boundary GMMs, [11] trained support vector machine (SVM) classifiers to differentiate boundaries from non-boundary positions, and [12, 13] employed neural network to refine phone boundaries. [14] described a non-HMM system for phone alignment based on discriminative learning. In their system a set of base functions were learned to measure the confidence for an alignment. [15] proposed several modifications to an HMM-based system, including the use of energy-based features and distinctive phonetic features, and the use of observation-dependent state transition probabilities.

In prior work [16], we demonstrated that employing explicit phone boundary models within the HMM framework could significantly improve forced alignment accuracy on the TIMIT corpus, a standard data set for training and evaluating English forced alignment systems. In that work, manual phonetic segmentation was used for training models. Special one-state phone boundary HMMs (i.e., a boundary can have one and only one state occurrence) were trained using frames extracted at the manually labeled phone boundaries, one frame for each boundary. It remains unclear whether (or to what degree) explicit phone boundary models help if no manual phonetic segmentation is available for training, and whether the special one-state HMM typology for modeling phone boundaries is beneficial for other languages. In this study, we addressed this question by investigating the use of phone boundary models on forced alignment in Mandarin Chinese. Unlike in English, no standard data set is available for study and evaluation of phonetic segmentation in Mandarin. We built a corpus for our study, for which only the test data but not the training data were manually segmented at the phonetic level.

Mandarin Chinese is a tone language. Many studies have demonstrated the benefit of incorporating tones in automatic speech recognition in Mandarin Chinese [17-24]. The goal of automatic phonetic segmentation is, however, different from that of speech recognition. In this study, we investigated whether tone information is useful in terms of improving forced alignment accuracy. To our knowledge,

this question has not been addressed in the literature. There are two major approaches on how to incorporate tones in automatic speech recognition: embedded tone modeling and explicit tone modeling [25]. In embedded tone modeling, tones were treated as a property of other units (e.g., vowels, finals, etc.); and tonal acoustic features are appended to the spectral features at each frame. In explicit tone modeling, tones are independently recognized and then combined with phone recognition. In this study, we adopted the first approach and compared the performance on forced alignment between tone-dependent and tone-independent models. In our experiments appending F_0 features to the spectral feature vector at the frame level resulted in slightly lower forced alignment accuracy, most probably due to problems of pitch tracking and normalization. For the purpose of this study, spectral features extracted from glottal waveforms (by performing glottal inverse filtering from the acoustic speech waveforms, detailed in Section 3.1.3) were used in place of F_0s. The motivation is that the glottal features represent the characteristics of vocal fold vibration, which contains tone information. It was reported that, for example, the difference between the first and second harmonics (which is commonly used as a glottal feature) was correlated with tones in Mandarin Chinese [26], and also spectral cues could be used by native speakers in tone perception [27]. On the other hand, glottal features may also contain information about supraglottal characteristics of speech sounds due to nonlinear coupling between the glottal source and the vocal tract filter [28]. For example, [29] demonstrated that vocal fold vibration was most destabilized when F0 crossed F1 in vowel production; and [30] demonstrated that glottal waveforms were different during both the close and open phases between the four vowels /i, e, a, u/. Therefore, glottal features could help forced alignment (and speech recognition) in both tone-dependent and tone-independent models.

In this study, we trained forced alignment systems differing in their use of phone boundary models, glottal features, and tone, and compared the performance of the systems. The experiments were conducted on a corpus described in Section 2.

2. CORPUS

The 1997 Mandarin Broad News Speech (LDC98S73) corpus was used. We extracted the "utterances" (the between-pause units that are time-stamped in the transcripts) from the corpus and listened to all utterances to exclude those with background noise and music. Utterances from speakers whose names were not tagged in the corpus or from speakers with accented speech were also excluded. The final dataset consisted of 7,849 utterances from 20 speakers. We randomly selected 300 utterances from six speakers (50 utterances for each speaker), three male and three female, to compose a test set. The remaining 7,549 utterances were used for training.

The 300 test utterances were manually labeled and segmented into initials and finals in *Pinyin* (a Roman alphabet system for transcribing Chinese characters). While researchers have disagreed on the vowel phonemes in Mandarin Chinese (see discussion in [31]), the inventories of initials and finals in the language are largely straightforward. 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 [32], 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 37 finals. Tones were marked on the finals, including Tone1 through Tone4, and Tone0 for the neutral tone. The phonetic labels are listed in Table 1.

Excluding boundaries between silence and a stop or an affricate (for which the boundary location cannot be determined because of the silent closure at the consonant onset), the test set contained 6,666 boundaries.

Table 1. *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, ung, uo v, van, ve, vn *
Tones	1, 2, 3, 4, 0
Silence	sil

* "v" represents "ü" in *Pinyin*.

3. SYSTEMS AND EVALUATION

3.1. Systems

HMM-based forced alignment systems were trained with the CALLHOME Mandarin Chinese Lexicon (LDC96L15) using the HTK toolkit [33]. All systems employed the standard 39 Perceptual Linear Prediction (PLP, [34]) features extracted with 25ms Hamming window and 10ms frame rate; the features were augmented with glottal features for some systems (as detailed below). 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 systems differed in their use of explicit phone boundary models, tone information, and glottal features.

3.1.1. Special one-state model for phone boundaries

The phone boundary models were a special 1-state HMM (as shown in Figure 1), in which the state cannot repeat itself. Therefore, a boundary can have one and only one state occurrence, i.e., aligned with only one frame.

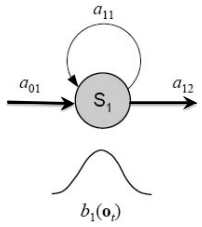

Figure 1*: Special 1-state HMM for phone boundaries with transition probabilities $a_{11} = 0$ and $a_{12} = 1$.*

The special 1-state phone boundary HMMs were combined with monophone HMMs. Given a phonetic transcription (which was manually labeled on the test set and derived from the lexicon and word transcription on the training set), phone boundaries were inserted between phones for both training and testing purposes. For example, "*sil i g e sil*" became "*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 in speech recognition.

3.1.2. Tone information

Tones were treated as a property of the finals. Tone-dependent final models (for which the same final with different tones had different HMMs) were trained to compare with tone-independent final models (for which the same final with different tones shared the same HMM).

3.1.3. Glottal features

Glottal waveforms were derived from the acoustic speech waveforms by performing glottal inverse filtering with the IAIF method developed by Alku and colleagues [35,36]. Mel-frequency cepstral coefficients (MFCCs, [37]) were extracted from band-limited glottal waveforms, with 20 band-pass filters ranging from 0 to 2000 Hz (Window size and frame rate were the same as used for PLPs). For the systems using glottal features, 26 glottal MFCCs (13 static coefficients and 13 delta coefficients) were appended to the PLP feature vector at each frame.

3.2. System evaluation

The accuracy of automatic segmentation is generally measured in terms of what percentage of the automatically labeled boundaries are within a given time threshold (tolerance) of the manually labeled boundaries. 20 ms has been most widely used as a tolerance for measuring phone segmentation quality. In the following section the agreement percentages for 20ms tolerance are reported.

Systematic errors generated by HMM-based forced alignment systems can be corrected using statistical models learned from comparing forced aligned and manually labeled boundaries in the training data [16]. However, because manual phonetic segmentation is not available for the training data used in this study, in the following section forced alignment results are evaluated against manual phonetic segmentation in the test set without boundary correction.

4. RESULTS

Table 2 lists the accuracies (20 ms agreement percentages) of the forced alignment systems that either employ (shown as +) or not employ (shown as -) boundary models, glottal features, and tone.

As a reference, we also calculated the overall mean time difference between forced alignment and manual segmentation for all boundaries in the test set, and then corrected the boundaries by this difference. The accuracies after boundary correction are listed in parentheses in Table 2. It is interesting to note that although the boundary correction procedure significantly increased the accuracies for the systems not using boundary models, it did not change the accuracies for the systems using boundary models. This suggests that there is little system bias in forced alignment for the systems using boundary models.

Table 2. *Forced alignment accuracies of different systems.*

Tone	Glottal features	Boundary models	Accuracy	Mean accuracy
-	-	-	0.859 (0.894)	- Boundary: 0.874 (0.898) + Boundary: 0.925 (0.925) - Glottal: 0.895 (0.906) + Glottal: 0.904 (0.916) - Tone: 0.896 (0.913) + Tone: 0.903 (0.909)
-	+	-	0.871 (0.906)	
+	-	-	0.878 (0.887)	
+	+	-	0.888 (0.903)	
-	-	+	0.923 (0.923)	
-	+	+	0.929 (0.929)	
+	-	+	0.918 (0.918)	
+	+	+	0.928 (0.927)	

4.1. Boundary models

From Table 2 we can see that the use of phone boundary models significantly improved forced alignment accuracy.

On average the accuracy increased from 0.874 to 0.925, representing a relative error reduction over 40%.

The improvement to individual systems is illustrated in Figure 2. It shows that employing phone boundary models consistently improved forced alignment accuracy, for both tone dependent and tone independent systems, and for systems using or not using the glottal features.

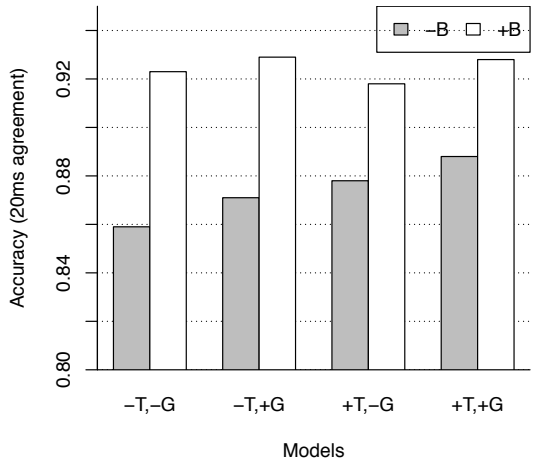

Figure 2: *Comparison between systems using (+B) and not using (-B) phone boundary models. –T: tone independent models; +T: tone dependent models; -G: not using glottal features; +G: using glottal features.*

4.2. Glottal features

From Table 2, and as illustrated in Figure 3, the use of glottal features also improved forced alignment accuracy across the systems, although to a lesser degree. On average, the accuracy increased from 0.895 to 0.904, representing a relative error reduction of 8.6%.

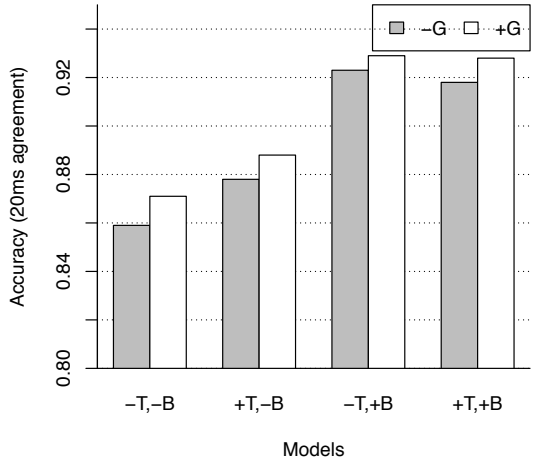

Figure 3: *Comparison between systems using (+G) and not using (-G) glottal features. –T: tone independent models; +T: tone dependent models; -B: not using boundary models; +B: using boundary models.*

4.3. Tone

Unlike phone boundary models and glottal features, tone dependent models had mixed effects. As shown in Figure 4, tone dependent models only outperformed tone independent models when phone boundary models were not used. When phone boundary models were used, tone dependent models slightly underperformed tone independent models. This may be due to the limited amount of data used for training. To overcome the data sparseness problem, we retrained tone dependent models by state tying - the states of the same final in different tones were pooled and clustered. After state tying, tone dependent models did outperform tone independent models, but the difference remained very small. The results are listed in Table 3.

Table 3. *Force alignment accuracies of tone independent system (-), tone dependent system (+), and tone dependent system with state tying (+*).*

Tone	Glottal features	Boundary models	Accuracy
-	+	+	0.929 (0.929)
+	+	+	0.928 (0.927)
+*	+	+	**0.931** (0.931)

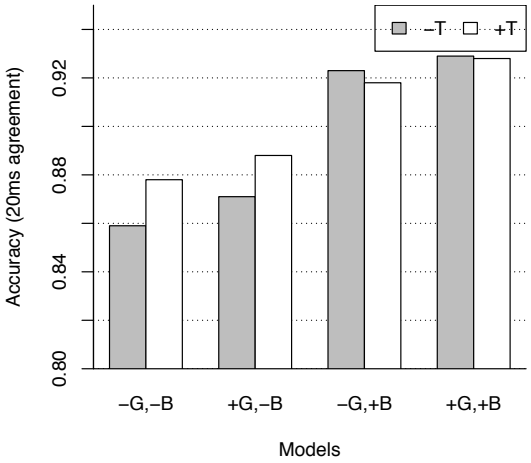

Figure 4: *Comparison between tone independent (-T) and tone dependent (+T) systems. -G: not using glottal features; +G: using glottal features; -B: not using boundary models; +B: using boundary models.*

From Table 2 and 3, we conclude that the best system employed phone boundary models, tone dependent models with state tying, and glottal features. The system achieved 93.1% agreement (of phone boundaries) within 20 ms compared to manual segmentation on the test set without boundary correction.

5. ACKNOWLEDGMENTS

This work was supported in part by NSF grant IIS-0964556.

6. REFERENCES

[1] Wang, L., Zhao, Y., Chu, M., Zhou, J. and Cao, Z., "Refining segmental boundaries for TTS using fine contextual-dependent boundary models," *Proceedings of ICASSP 2004*, pp. 641-644, 2004.

[2] Lin, C., Jang, J. and Chen, K., "Automatic segmentation and labeling for Mandarin Chinese speech corpora for concatenation-based TTS," *Computational Linguistics and Chinese Language Processing*, 10, pp. 145-166, 2005.

[3] Leung, H. and Zue, V.W., "A procedure for automatic alignment of phonetic transcription with continuous speech," *Proceedings of ICASSP 1984*, pp. 73-76, 1984.

[4] Brugnara, F., Falavigna, D. and Omologo, M., "Automatic segmentation and labeling of speech based on hidden Markov models," *Speech Communication,* 12, pp. 357–370, 1993.

[5] Ljolje, A., Hirschberg, J. and van Santen, J, "Automatic speech segmentation for concatenative inventory selection," in J. van Santen, R. Sproat, J. Olive and J. Hirschberg (ed.), *Progress in Speech Synthesis*, Springer Verlag, New York, pp. 305-311, 1997.

[6] Wightman, C. and Talkin, D., "The Aligner: Text to speech alignment using Markov Models," in J. van Santen, R. Sproat, J. Olive and J. Hirschberg (ed.), *Progress in Speech Synthesis*, Springer Verlag, New York, pp. 313-323, 1997.

[7] Hosom, J.P., *Automatic Time Alignment of Phonemes Using Acoustic-Phonetic Information*. PhD thesis, Oregon Graduate Institute of Science and Technology, 2000.

[8] Toledano, D.T., Gomez, L.A.H. and Grande, L.V., "Automatic phoneme segmentation," *IEEE Trans. Speech and Audio Proc.*, 11, pp. 617-625, 2003.

[9] Stevens, K., "Toward a model for lexical access based on acoustic landmarks and distinctive features," *J. Acoust. Soc. Am.*, 111, pp. 1872-1891, 2002.

[10] Kim, Y.-J. and Conkie, A., "Automatic segmentation combining an HMM-based approach and spectral boundary correction," *Proceedings of ICSLP 2002*, pp.145-148, 2002.

[11] Lo, H.-Y. and Wang, H-M., "Phonetic boundary refinement using support vector machine," *Proceedings of ICASSP 2007*, pp. 933-936, 2007.

[12] Toledano, D.T., "Neural network boundary refining for automatic speech segmentation," *Proceedings of ICASSP 2000*, pp.3438-3441, 2000.

[13] Lee, K.-S., "MLP-based phone boundary refining for a TTS database," *IEEE Trans. Audio, Speech, and Language Proc.*, 14, pp. 981-989, 2006.

[14] Keshet, J., Shalev-Shwartz, S., Singer, Y. and Chazan, D., "Phoneme alignment based on discriminative learning." *Proceedings of Interspeech 2005*, pp. 2961–2964, 2005.

[15] Hosom, J.P., "Speaker-independent phoneme alignment using transition-dependent states," *Speech Communication*, 51, pp. 352-368, 2009.

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

[17] Chen, C.J., Gopinath, R.A., Monkowski, M.D., Picheny, M.A. and Shen, K., "New methods in continuous Mandarin speech recognition," *Proceedings of Eurospeech 1997*, pp. 1543-1546, 1997.

[18] Chang, E., Zhou, J., Di, S., Huang, C., Lee, K.F., "Large vocabulary Mandarin speech recognition with different approaches in modeling tones," *Proceedings of Interspeech 2000*, pp. 983-986, 2000.

[19] Huang, H. and Seide, F., "Pitch tracking and tone features for Mandarin speech recognition*," Proceedings of ICASSP 2000*, pp. 1523-1526, 2000.

[20] Zhou, J., Tian, Y., Shi, Y., Huang, C. and Chang, E., "Tone articulation modeling for Mandarin spontaneous speech recognition," *Proceedings of ICASSP 2004*, pp. 997-1000, 2004.

[21] Lei, C., Siu, M., Hwang, M., Ostendorf, M. and Lee, T., "Improved tone modeling for Mandarin broadcast news speech recognition," *Proceedings of Interspeech 2006*, pp. 1237-1240, 2006.

[22] Ni, C., Liu, W. and Xu, B. "Improved large vocabulary Mandarin speech recognition using prosodic and lexical information in maximum entropy framework," in *Proceedings CCPR,* 2009.

[23] Qian, Y. and Soong, F., "A Multi-Space Distribution (MSD) and two-stream tone modeling approach to Mandarin speech recognition*," Speech Communication*, pp. 1169-1179, 2009.

[24] Chao, H., Yang, Z. and Liu, W., "Improved tone modeling by exploiting articulatory features for Mandarin speech recognition," *Proceedings of ICASSP 2012*, pp. 4741-4744, 2012.

[25] Lee, T., Lau, W., Wong, Y.W. and Ching, P.C., "Using tone information in Cantonese continuous speech recognition," *ACM Transactions on Asian Language Information Processing*, 1, pp. 83-102, 2002.

[26] Keating, P. and Esposito, C., "Linguistic voice quality," UCLA Working Papers in Phonetics, 105, pp. 85-91, 2007.

[27] Kong, Y. and Zeng, F., "Temporal and spectral cues in Mandarin tone recongitino," *J. Acoust. Soc. Am.*, 120, pp. 2830-2840, 2006.

[28] Titze, I., "Nonlinear source-filter coupling in phonation: Theory, "*J. Acoust. Soc. Am.*, 123, pp. 2733-2749, 2008.

[29] Titze, I., "Nonlinear source-filter coupling in phonation: Vocal exercises, "*J. Acoust. Soc. Am.*, 123, pp. 1902-1915, 2008.

[30] Lulich, S., Zanartu, M., Mehta, D. and Hillman, R., "Source-filter interaction in the opposite direction: subglottal coupling and the influence of vocal fold mechanics on vowel spectral during the closed phase," in *Proceedings of Meetings on Acoustics 2009*, 6, 2009.

[31] Duanmu, S., *The Phonology of Standard Chinese*, Oxford: Oxford University Press, pp. 35-40, 2000.

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

[33] The Hidden Markov Model Toolkit (HTK): http://htk.eng.cam.ac.uk/

[34] Hermansky, H., "Perceptual linear predictive (PLP) analysis of speech*," J. Acoust. Soc. Am.*, 87, pp. 1738-1752, 1990.

[35] Alku, P., "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Communication*, 11, pp. 109-118, 1992.

[36] Iterative adaptive inverse filtering code (for Matlab): http://users.tkk.fi/~traitio/research.html

[37] Davis, S. and Mermelstein, P., "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28, 357-366, 1980.