Investigating Consonant Reduction in Mandarin Chinese with Improved Forced Alignment

Jiahong Yuan, Mark Liberman

Linguistic Data Consortium, University of Pennsylvania

jiahong@ldc.upenn.edu, myl@ldc.upenn.edu

Abstract

Phonetic reduction has been an important topic in linguistics research. It also presents a great challenge for forced alignment, a technique widely used for automatic phonetic segmentation. In this study, we employed skip-state HMMs to improve forced alignment quality and to make forced alignment applicable to the investigation of phonetic reduction and deletion. With skip-state HMMs, forced alignment accuracy at 10 ms agreement was improved from 73.3% to 75.6% on a corpus of Mandarin Chinese broadcast news speech. Our analysis based on the improved forced alignment of Mandarin broadcast news speech – verified by hand segmentation of a random sample of cases – shows that: 1. The durations of frication and aspiration are additive in the production of plosives and affricates; 2. Plosives are more likely to be deleted than affricates; and 3. Plosives and affricates in higher-frequency words and at word-medial position are more likely to be reduced.

Index Terms: forced alignment, skip-state HMMs, reduction, deletion, plosives, affricates

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. The most common approach for automatic phonetic segmentation is "forced alignment", based on two inputs: a speech audio waveform and a transcription at the phone- or word-level. In the case of word-level transcription, the words are first mapped into a phone sequence using a pronouncing dictionary, or grapheme to phoneme rules. In standard Hidden Markov Model (HMM) based forced alignment [1], each phone is a HMM that typically has three left-to-right non-skipping states, as shown in Figure 1.

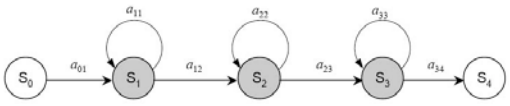

Figure 1: *HMM with three non-skipping states*.

Many studies have attempted to improve forced alignment accuracy [2-5]. In our previous work [6-7], we employed explicit phone boundary models within the HMM framework. The phone boundary models were a special 1-state HMM (as shown in Figure 2), in which the state cannot repeat itself:

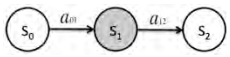

Figure 2: *Special 1-state HMM for phone boundaries with transition probabilities $a_{01} = a_{12} = 1$.*

Our results demonstrated that using special 1-state HMMs for phone boundaries could significantly improve forced alignment accuracy on both English TIMIT (~25% relative error reduction) and Mandarin Hub-4 Broadcast News Speech (~40% relative error reduction).

A persistent challenge in forced alignment and speech technology is phonetic reduction [8]. Phonetic reduction is pervasive in natural speech [9], simple word-to-phoneme mapping (either by using a pronouncing dictionary or grapheme to phoneme rules) may not always generate phone sequences that contain the correct pronunciation. Figure 3 shows three examples of the phoneme /j/ (which is /tɕ/ in IPA, an alveolo-palatal affricate) from the same speaker in the corpus of Mandarin HUB-4 Broadcast News Speech (LDC98S73). From both the waveforms and spectrograms we can see that the first example is a full phonetic realization of the phoneme, which contains a complete closure followed by a portion of frication noise. The second example only contains an incomplete closure but no frication. The third example does not show any consonantal features, suggesting that the phoneme is deleted.

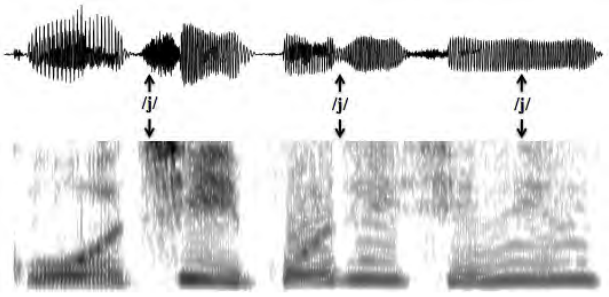

Figure 3: *Examples of variation in the phonetic realization of /j/: full, reduction, and deletion*.

A common approach to address the problem of severe reduction (e.g., change of phone class) and deletion in speech recognition is through pronunciation modeling, for example, to include multiple pronunciations for a word in the lexicon. However, such a lexicon (or more sophisticated pronunciation models) is difficult to build, especially given that most reduction processes seem gradient rather than categorical [10]. In addition, phonetic deletion poses another challenge in pronunciation modeling. In many cases coarticulation and phonetic transitions remain even a phone is "deleted", as we

can see from the third example of /j/ in Figure 3. In order to be able to model the phonetic transitions to and from a "deleted" phone, the "deleted" phone must be preserved at some level and tractable.

Another, less common, approach to the problem of phonetic reduction and deletion in speech recognition is to employ skip-state HMMs [11]. Figure 4 illustrates a 3-state HMM in which every state can be skipped. If all the states are skipped, the result will be a phone with zero duration – a phone that is deleted in the surface form but still preserved in the lexicon or pronunciation model.

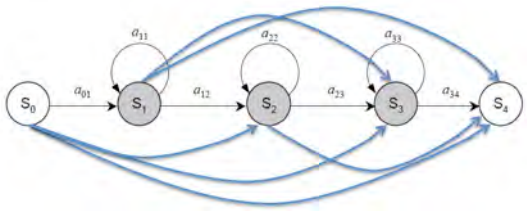

Figure 4: *HMM with skip-state transitions*.

In this study, we aim to employ skip-state HMMs to improve forced alignment quality and to make forced alignment applicable to the investigation of phonetic reduction and deletion. The study builds on our previous work on forced alignment in Mandarin Chinese [7], which applied explicit phone boundary models, glottal features and tone information, and achieved 93.1% agreement within 20 ms and 73.3% within 10 ms without boundary correction.

With the improved forced alignment, we investigate the reduction in terms of duration of plosives and affricates in Mandarin broadcast news speech. Phonetic reduction has been an important topic in the literature. A number of linguistic factors have been found to contribute to the phonetic reduction in English, including word frequency and predictability [12], phonological neighborhood density [13], and morphological category [14]. Previous research on the phonetic reduction in Mandarin Chinese has mostly focused on syllable contraction (i.e. reduction of two or more syllables into one) [15] and tonal reduction [16].

There are four types of plosives and affricates in Mandarin Chinese: unaspirated stops, aspirated stops, unaspirated affricates, and aspirated affricates. The plosives and affricates can only appear in syllable initial position in the language. As listed in Table 1, these consonants contrast on two acoustic dimensions in production: aspiration and frication. How do the two dimensions relate to each other in determining the duration of the consonants? How do word frequency and word boundary affect the phonetic reduction of the consonants? We attempt to answer these questions.

Table 1. *Plosives and affricates in Mandarin Chinese (in Pinyin and IPA).*

Unaspirated stops: b[p], d[t], g[k]	-A, -F
Aspirated stops: p[pʰ], t[tʰ], k[kʰ]	+A, -F
Unaspirated affricates: z[ts], zh[tʂ], j[tɕ]	-A, +F
Aspirated affricates: [tsʰ], ch[tʂʰ], q[tɕʰ]	+A, +F

(A = aspiration; F = frication)

2. Data

The 1997 Mandarin Broadcast News Speech (HUB4-NE, LDC98S73) corpus was used. We extracted "utterances" (which are between-pause units that were manually time-stamped) 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). 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.

3. Forced alignment with skip-state HMMs

In [7], consonants were non-skipping 3-state HMMs (as shown in Figure 1) and phone boundaries were special 1-state HMMs (as shown in Figure 2). The minimum duration of a consonant determined by the forced alignment system built in [7] is 40 milliseconds - 30 ms from the consonant HMM plus 5 ms from the preceding boundary HMM (i.e., half of the boundary duration) and 5ms from the following boundary HMM. Apparently the non-skipping 3-state HMMs cannot handle severe reduction and deletion in natural speech. In this study, we model the phonetic reduction and deletion by employing skip-state HMMs as shown in Figure 4. The minimum duration of a consonant from forced alignment using the new acoustic models is 10 ms – 0 ms from the consonant HMM plus 5 ms from the preceding boundary HMM and 5ms from the following boundary HMM.

As in [7], we use the CALLHOME Mandarin Chinese Lexicon (LDC96L15), the standard 39 PLP features [17], augmented with 39 MFCCs [18] extracted from band-limited (0-2000Hz) glottal waveforms that are derived from glottal inverse filtering [19]. The features are extracted with 25ms Hamming window and 10ms frame rate. Building on the acoustic models in [7], we replaced the state transition matrices of all consonants with the one represented in Figure 4. The initial values of the transition probabilities were set to be the same across the consonants. All parameters in the new acoustic models were re-estimated using the same training data as in [7], and tested on the same test set. The HTK toolkit is used for the experiment (http://htk.eng.cam.ac.uk/).

Table 2 lists the results of forced alignment using the new acoustic models, as well as the results reported in [7].

Table 2. *Forced alignment accuracies of using non-skipping vs. skip-state HMMs.*

Consonant models	20 ms Agr.	10 ms Agr.
Non-skipping HMMs [7]	93.1%	73.3%
Skip-state HMMs	93.3%	75.6%

From Table 2 we can see that employing skip-state HMMs for consonants only slightly improved forced alignment accuracy at 20 ms agreement. However, it had a greater impact on forced alignment accuracy with smaller tolerance. The accuracy at 10 ms agreement was improved from 73.3% to 75.6%, a relative error reduction of 8.6%.

Figure 5 draws the relative error reduction, with respect to 10 ms agreement, on different types of consonants, counting both the preceding and following boundaries of the consonants. The greatest improvement shows on unaspirated stops (/b, d, g/), which have more than 20% relative error reduction. Unaspirated affricates (/z, zh, j/), aspirated stops (/p, t, k/), aspirated affricates (/c, ch, q/) and glides (/r, l/) have moderate improvement, with a relative error reduction of 5%-10%. The fricatives (/f, h, x, sh, s/) have little improvement. The forced alignment accuracy on the nasal initials (/m, n/), on the other hand, decreases. The negative impact of skip-state HMMs on nasal initials in Mandarin Chinese is consistent with the observation that the nasal initials are usually not deletable in the language. For example, the word "wo.men" (我们, "we") in Mandarin Chinese can be reduced to "wo.m" but not "wo.en" or "wo.n".

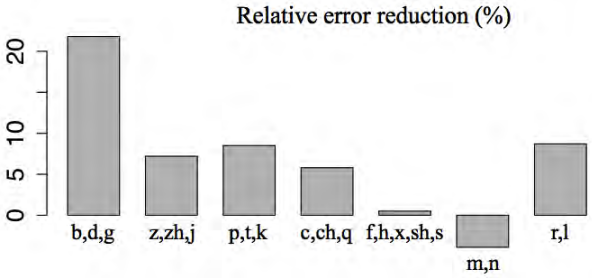

Figure 5: *Relative error reduction on 10-ms agreement with skip-state HMMs.*

4. Plosives and affricates in Mandarin broadcast news speech

In this section we present a case study of using the improved forced alignment to investigate the temporal reduction of plosives and affricates in Mandarin broadcast news speech. We run forced alignment on the training data, and analyze the durational characteristics of the totally 45,870 plosives and affricates in the data set.

4.1. Durational characteristics

The mean durations of the four types of plosives and affricates are listed in Table 3. From the table we can see that the inherent durations of the four consonant types, on the ascending order, are: unaspirated stops (~50 ms), unaspirated affricates (~65 ms), aspirated stops (~85 ms), and aspirated affricates (~100 ms). We can also see that the two dimensions – aspiration and frication – in the production of these consonants are additive in terms of segment duration: the base duration (unaspirated stops) is ~50 ms; frication takes ~15 ms and aspiration takes ~35 ms.

4.2. Reduction and deletion

Figure 6 shows the duration distributions (cumulative percentages) of the four consonant types. We can see that inherently longer plosives and affricatives are less likely to be shorter than a given duration, if the duration is 30 ms or longer, than inherently shorter plosives and fricatives. This result suggests that the four types of plosives and fricatives have similar patterns of reduction (and strengthening) in terms of duration.

Table 3. *Mean durations of the four types of plosives and affricates.*

Consonant type	Aspira-tion	Frica-tion	Duration (ms)
Unaspirated Stops /b,d,g/	-	-	50.2 base
Unaspirated Affricates /z, zh, j/	-	+	65.7 ≈ base + 15 (F)
Aspirated Stops /p, t, k/	+	-	85.4 ≈ base + 35 (A)
Aspirated Affricates /c, ch, q/	+	+	98.1 ≈ base + 15 (F) + 35 (A)

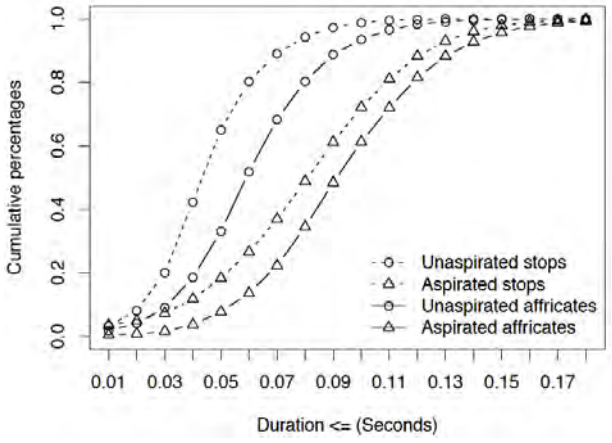

Figure 6: *Duration distributions of the four types of plosives and affricates.*

At 10 ms and 20 ms (which represents a severe reduction or deletion), however, the cumulative percentages are not correlated with the inherent durations of the consonant types. From Table 4, we can see that the aspirated stops have higher cumulative percentages than the unaspirated affricates at 10 ms (3.5% vs. 1.9%) and 20 ms (4.9% vs. 4.1%), although the inherent duration of the aspirated stops is longer than the unaspirated affricates (and therefore less likely to reduce if the correlation holds). This result suggests that stops are more likely to be deleted than affricates in Mandarin broadcast news speech. It also suggests that reduction and deletion may result from different phonetic processes, rather than a continuum of the same process.

Table 4. *Cumulative percentages at 10,20 and 30 ms.*

Consonant type	<= 10 ms	<= 20 ms	<= 30 ms
Unaspirated Stops /b,d,g/	3.4%	8.2%	20.1%
Unaspirated Affricates /z, zh, j/	1.9%	4.1%	9.0%
Aspirated Stops /p, t, k/	**3.5%**	**4.9%**	7.3%
Aspirated Affricates /c, ch, q/	0.5%	0.8%	1.6%

4.3. Effects of word boundary and word frequency

Finally, we investigate the effects of word boundary and word frequency on the reduction of plosives and affricates. Because the four consonant types have different inherent durations but similar patterns of reduction, a relative duration threshold is used to determine whether a consonant is reduced in the corpus. For every consonant type we use the duration corresponding to its ~20% cumulative percentage as the threshold of reduction: 30 ms for unaspirated stops, 40 ms for unaspirated affricates, 50 ms for aspirated stops, and 60 ms for aspirated affricates.

Figure 7 shows the effects of word boundary and word frequency on the reduction rates of plosives and affricates in Mandarin broadcast news speech. The word frequencies are from the CALLHOME Mandarin Chinese Lexicon, which are based on the frequency counts of approximately 3.5 million words in the *Xinhua* newswire. We separate the words into four frequency bins (0-100, 100-1000, 1000-10000, and more than 10000). From Figure 7 we can see that word-medial plosives and affricates are more likely to be reduced than word-initial ones, and that plosives and affricates in higher-frequency words are more likely to be reduced.

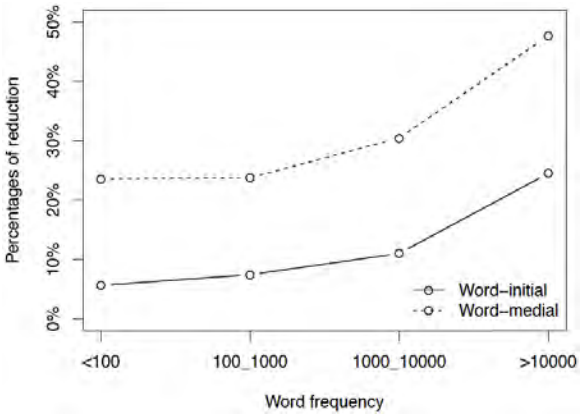

Figure 7: *Effects of word boundary and word frequency on the reduction rates of plosives and affricates.*

5. Conclusions

The use of forced alignment in phonetics and linguistics research has been rapidly growing in the last few years. Forced alignment is not only an indispensible tool for automatic phonetic segmentation, it can also be used as a new method of phonetic analysis. In [20] and [21], for example, we applied forced alignment to the quantification of /l/ darkness and to automatic detection of "g-dropping". In this study, we demonstrated that with the employment of skip-state HMMs forced alignment can be used to investigate temporal reduction and deletion. Our results showed that, in Mandarin broadcast news speech, word-medial plosives and affricates are more likely to be reduced than word-initial ones, and that plosives and affricates in higher-frequency words are more likely to be reduced. We also found that the durations of frication and aspiration are additive in the production of plosives and affricates, and that plosives are more likely to be deleted than affricates.

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

6. References

[1] 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.

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

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

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

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

[6] 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.

[7] Yuan, J., Ryant N. and Liberman, M., "Automatic phonetic segmentation in Mandarin Chinese: boundary models, glottal features and tone," *Proceedings of ICASSP 2014*, pp. 6689-6693, 2014.

[8] Furui, S., Nakamura, M., Ichiba, T., Iwano, K., "Why Is the Recognition of Spontaneous Speech so Hard?" *Proceedings of TSD 2005*, pp. 9-22, 2005.

[9] Johnson, K., "Massive reduction in conversational American English," In Yoneyama, K. and Maekawa, K. (eds.), *Spontaneous Speech: Data and Analysis,* pp. 29-54, 2004.

[10] Ernestus, M. and Warner, N., "An introduction to reduced pronunciation variants," *Journal of Phonetics*, 39, pp. 253-260, 2011.

[11] Liu, D., Nguyen, L., Matsoukas, S., Davenport, J., Kubala, F. and Schwartz, R., "Improvements in spontaneous speech recognition," in *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

[12] Jurafsky, D., Bell, A., Gregory, M. and Raymond, W., "Probabilistic Relations between Words: Evidence from Reduction in Lexical Production," In Bybee, J. and Hopper P. (eds.), *Frequency and the emergence of linguistic structure*, pp. 229-254, 2001.

[13] Gahl, S., Yao, Y. and Johnson, K., "Why Reduce? Phonological Neighborhood Density and Phonetic Reduction in Spontaneous Speech," *Journal of Memory and Language*, 66, pp. 789-806, 2012.

[14] Guy, G., "Explanation in variable phonology: An exponential model of morphological constraints," *Language Variation and Change*, pp. 1-22, 1991.

[15] Tseng, S., "Syllable Contractions in a Mandarin Conversational Dialogue Corpus," *International Journal of Corpus Linguistics*, 10, pp. 63-83, 2005.

[16] Chen, Y. and Xu, Y., "Production of weak elements in speech: Evidence from neutral tone in Standard Chinese," *Phonetica*, 63, pp. 47-75, 2006.

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

[18] 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.

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

[20] Yuan, J. and Liberman, M., "Investigating /l/ variation in English through forced alignment," *Proceedings of Interspeech 2009*, pp. 2215-2218, 2009.

[21] Yuan, J. and Liberman, M., "Automatic detection of 'g-dropping' in American English using forced alignment," *Proceedings of 2011 IEEE Automatic Speech Recognition and Understanding Workshop*, pp. 490-493. 2011.