Using Forced Alignment for Phonetics Research

Jiahong Yuan*, Wei Lai, Chris Cieri, and Mark Liberman

University of Pennsylvania, Philadelphia, USA
jiahong@ldc.upenn.edu, weilai@sas.upenn.edu, ccieri@ldc.upenn.edu, myl@ldc.upenn.edu

Abstract. Forced alignment has been at the core of speech recognition technology since the 1970s, and was first used in phonetics research in the 1990s. Progress in digital multimedia, networking and mass storage is creating enormous and growing volumes of transcribed speech, which forced alignment can turn into vast phonetic databases. However, speech science has so far taken relatively little advantage of this opportunity, because it requires tools and methods that are now difficult for most speech researchers to access, and are incompletely developed and tested for many applications. But these technologies are leading the study of human speech into a revolutionary new era: a movement from the study of small, private, and mostly artificial datasets to the analysis of published collections of natural speech that are thousands or even millions of times larger. In this chapter, we illustrate some of the ways that forced alignment can be used as a tool in speech science, and discuss directions for improvement.

Keywords: Forced alignment · Corpus phonetics · Phonetic segmentation

1 Introduction

In the last twenty-five years, an enormous and growing body of digital speech has become available: archived broadcasts of news reports, interviews, speeches, and debates; oral histories; court recordings; podcasts; audiobooks; and so on. A small fraction of this material – still comprising many thousands of hours – has been collected and published in the form of corpora for speech technology research. These very-large-scale bodies of data make it possible to use natural speech in developing and testing hypotheses across many types of individual, social, regional, temporal, and contextual variation, as well as across languages. However, in contrast to speech technology research, speech science has so far taken relatively little advantage of this opportunity. This is partly because most researchers lack the knowledge and skills required to access the needed tools and methods, and partly because the tools and methods themselves are incomplete and untested.

Given only digital audio, we can study the distribution of speech and silence segments, or of purely acoustic-phonetic features such as fundamental frequency. But for most kinds of speech science, we need to know which words were said when, and how they were pronounced – and this entails the availability of phonetic segmentation and transcription. Relatively few speech corpora come with such annotations, because manual phonetic segmentation is time-consuming, expensive and inconsistent, with much less than perfect inter-annotator agreement (Godfrey et al. 1992; Leung and Zue 1984; Cucchiarini 1993).

Automatic phonetic segmentation is, therefore, necessary for corpus-based phonetics research. Luckily, automatic phonetic segmentation is the essential result of forced alignment, a technique developed for training automatic speech recognition systems (Jelinek 1976) and for extracting acoustic units for speech synthesis systems (Wightman and Talkin 1997).

This task normally requires two inputs: recorded audio and a conventional (orthographic) transcription. The transcribed words are mapped into a phone sequence or a lattice of possible phone sequences, by using a pronouncing dictionary and/or grapheme to phoneme rules. Phone boundaries are determined by comparing the observed speech signal and pre-trained, Hidden Markov Model (HMM) based acoustic models. Typically every phone in the acoustic models is represented as an HMM that consists of three left-to-right non-skipping states (as shown in Figure 1): the beginning (s_1), middle (s_2), and ending (s_3) parts of the phone, plus empty start (s_0) and end states (s_4) for entering and exiting the phone. From the training data, an acoustic model (e.g., a Gaussian Mixture Model) is built for each state (except s_0 and s_4), as well as the transition probabilities between pairs of states (Figure 1). The speech signal is analyzed as a successive

set of frames (e.g., every 10 ms). 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 derived from transcription) given the observed data and the acoustic models represented by the HMMs. The reported performances of state-of-the-art HMM-based forced alignment systems range from 80%-93% agreement (of all boundaries) within 20 ms compared to manual segmentation (Hosom 2009; Yuan et al. 2013) on the TIMIT corpus (Garofolo et al. 1993). Human labelers have an average agreement of 93% within 20 ms, with a maximum of 96% within 20 ms for highly-trained specialists (Hosom 2000).

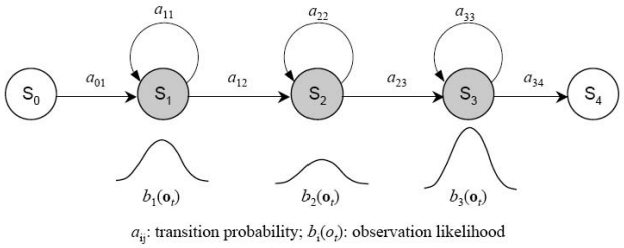

Fig. 1. Hidden Markov Model with three non-skipping states.

With the availability of automatic speech recognition toolkits such as HTK and Kaldi, forced alignment techniques are more easily accessible to speech researchers. In recent years, automatic speech analysis with the use of forced alignment has started to be developed in phonetic and sociolinguistic research, for example, automatic measurement of vowel formants (Evanini et al. 2009; Labov et al. 2013), voice onset time (Sonderegger and Keshet 2012), and speech variation in general (Fox 2006).

This paper describes the use of forced alignment in corpus-based phonetics research. We illustrate the use of forced alignment in phonetics in the case of two very different sorts of speech data: recordings of a Mandarin proficiency test, and collections of Mandarin broadcast news. In this context, we discuss three sorts of research work: using forced alignment as a method to produce phonetic segmentation; using forced alignment as a method for investigating allophonic variation; and efforts to improve the performance of forced alignment itself.

2 Using forced alignment for phonetic segmentation

In Yuan et al. (2016), we investigate the use of pauses and pause fillers (such as 嗯, 呃) in Mandarin Chinese. We focus on two factors: speaker sex and proficiency. Our analysis is based on 13 hours of monologue speech from 267 speakers.

2.1 Corpus

Putonghua Shuiping Ceshi (PSC) is the national standard Mandarin proficiency test in China. The test consists of four parts: The first two parts involve reading 100 monosyllabic and 50 disyllabic words; the third part requires reading an article of 300 characters, randomly selected from a pool of 60 articles; and the last part entails speaking freely on a given topic for three minutes. The four parts are graded separately and numerically and the total score out of 100 points is converted to one of six categorical proficiency levels, ranging from high to low: 一级甲等 (Class 1 Level 1), 一级乙等 (Class 1 Level 2), 二级甲等 (Class 2 Level 1), 二级乙等 (Class 2 Level 2), 三级甲等 (Class 3 Level 1), and 三级乙等 (Class 3 Level 2). In order to qualify for teaching K-12, one must pass 二级乙等 (Class 2 Level 2).

Our dataset consists of recordings of college students at Beijing Normal University who took the PSC test in 2011. We used the spoken monologues (the last part of the test) from 267 speakers, 178 female and 89 male, which contain approximately 13 hours of speech. The proficiency levels of the speakers range across four levels, from 一级乙等 to 三级甲等 (hereafter, L1 to L4).

2.2 Transcription and forced alignment

The spoken monologues were first transcribed by a professional transcriptionist, then proofed for errors and pause fillers, which were ignored in the first pass, were added. The pause fillers were categorized into two types via transcription: one without nasalization (transcribed as *e*) and one with nasalization (transcribed as *en*). We then employed forced alignment to determine the boundaries of the transcribed words, including pause fillers.

Pauses are usually not transcribed. To automatically identify pauses in speech using forced alignment, a special HMM called a "tee-model" can be inserted at word boundaries. A "tee-model" has a direct transition from the entry to the exit node. Therefore, it can either be aligned to a true pause if there is a silence in speech or completely skipped if there is no silence.

Through forced alignment with a "tee-model" for identifying inter-word pauses, we located 26,885 pauses in the 267 monologues in our dataset. The dataset also contains 100,212 words and 3,058 pause fillers, of which 1,192 are *e* and 1,866 are *en*.

2.3 Effect of speaker sex and proficiency level on pauses and pause fillers

The total numbers of pause fillers, pauses, and words for males and females and for different proficiency levels are listed in the top part of Table 1.

Table 1. Frequencies and relative frequencies of pauses and pause fillers.

	Female	Male	L1	L2	L3	L4
# e	671	521	384	403	352	53
# en	1287	579	568	655	478	165
# pauses	17828	9057	7231	11165	6591	1898
# words	68331	31881	29514	42461	22695	5542
e/(e+en)	***0.343***	***0.474***	0.403	0.381	0.424	0.243
e/words	***0.010***	***0.016***	0.013	0.010	0.016	0.010
en/words	0.019	0.018	0.019	0.015	0.021	0.030
(e+en)/words	0.029	0.035	0.032	0.025	0.037	0.039
pauses/words	***0.261***	***0.284***	***0.245***	***0.263***	***0.290***	***0.343***

For each speaker we compute five relative frequencies:

1. *e*/(*e*+*en*): the proportion of *e* in pause fillers;
2. *e/words*: the number of *e* per word;
3. *en/words*: the number of *en* per word;
4. *(e+en)/words*: the number of pause fillers per word;
5. *pauses/words:* the number of pauses (including all silent intervals) per word.

Mixed-effects logistic regression models (Bates et al. 2015) were used to assess the effects of sex and proficiency level on the relative frequencies of pauses and pause fillers, in which speaker was treated as a random factor. The results are shown in the bottom part of Table 1, where the mean values of the five relative frequency measures are listed, with bold-italic numbers representing statistical significance at $p < .05$. We can see that males use more *e* than females, but there is no difference between them on the frequency of *en*. Therefore, the proportion of nasal-final pause fillers is higher in female than in male speakers, as was found in the studies of Germanic languages (Wieling et al. 2016). Proficiency does not appear to affect the frequency of either *e* or *en*. With respect to the use of pauses, both sex and proficiency are a significant factor. Males use more pauses than females, and less proficient speakers also use more pauses.

3 Using forced alignment for investigating speech variation

In Yuan and Liberman (2015), we employed skip-state HMMs to adapt forced alignment to the investigation of phonetic reduction and deletion. With the improved forced alignment, we investigate the reduction of plosives and affricates in terms of duration in Mandarin broadcast news speech.

3.1 Corpus

The 1997 Mandarin Broadcast News Speech (HUB4-NE, LDC98S73) corpus was used (Huang et al. 1998). We extracted "utterances" (defined as between-pause units that were manually time-stamped) from the corpus and listened to each 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.

3.2 Forced alignment with skip-state HMMs

Phonetic reduction is pervasive in natural speech (Johnson 2004). It is not only an important topic in linguistics research, but also presents a great challenge in forced alignment and other speech technologies. Figure 2 shows three examples of the phoneme /j/ (which is /tɕ/ in IPA, an alveolo-palatal affricate) from the same speaker in the corpus. 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.

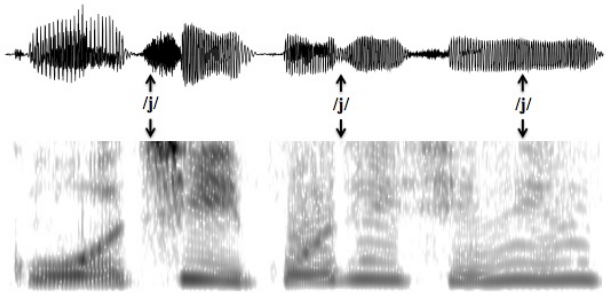

Fig. 2. Examples of variation in the phonetic realization of /j/: full, reduction, and deletion.

Apparently the non-skipping 3-state HMMs (Figure 1) cannot handle severe reduction and deletion in natural speech. We modeled the phonetic reduction and deletion by employing skip-state HMMs (as shown in Figure 3), 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. In many cases coarticulation and phonetic transitions remain even when a phone is "deleted", as we can see from the third example of /j/ in Figure 2.

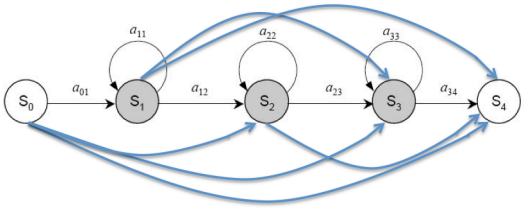

Fig. 3. HMM with skip-state transitions.

3.3 Reduction of plosives and affricates in Mandarin broadcast news speech

Resulting from forced alignment with skip-state HMMs for plosives and affricates, the mean durations of the four types of plosives and affricates are listed in Table 2. From the table we can see that the inherent durations of the four consonant types, in 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 adds ~15 ms and aspiration adds ~35 ms.

Table 2. Mean durations of the four types of plosives and affricates.

Consonant type	Aspiration	Frication	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 4 shows the duration distributions (cumulative percentages) of the four consonant types. We can see that for any given duration of 30 ms or longer, an inherently longer plosive/affricate is unlikely to be shorter than that duration than an inherently shorter one. This result suggests that the four types of plosives and fricatives have similar patterns of reduction (and strengthening) in terms of duration. 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. 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.

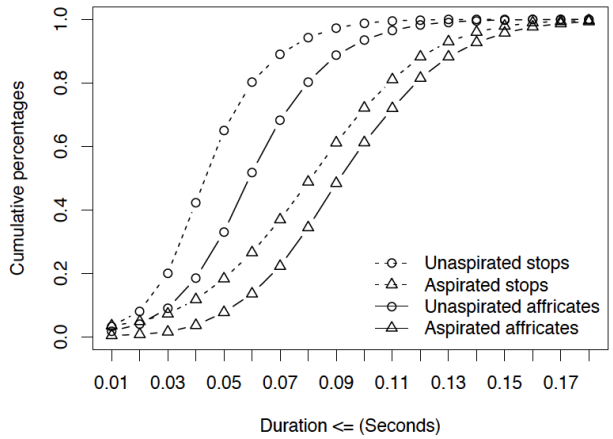

Fig. 4. Duration distributions of the four types of plosives and affricates.

4 Improving forced alignment for phonetic research

4.1 Phone boundary models for forced alignment

A main drawback of the HMM-based forced alignment for phonetic segmentation is that phone boundaries are not represented in the model. The boundaries are simply derived from the alignment of phone states with frames. This is different from the manual phonetic segmentation process, in which the acoustic landmarks at phone boundaries (Stevens 2002), e.g., an abrupt spectral change, are used to determine the location of a boundary. In our effort to overcome this drawback and improve forced alignment (Yuan et al. 2013; Yuan et al. 2014), we employed explicit phone boundary models within the HMM framework. The idea is to treat phones and phone boundaries as independent HMMs. A boundary is determined by the alignment of its own state with frames. The phone boundary models were a special 1-state HMM (as shown in Figure 5), in which the state cannot repeat itself:

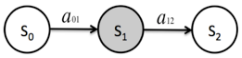

Fig. 5. Special 1-state HMM for phone boundaries with transition probabilities $a_{01} = a_{12} = 1$.

The special 1-state phone boundary HMMs were combined with 3-state phone HMMs. Given a phonetic transcription, phone boundaries were inserted between phones. 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 (Young et al. 1994).

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) (Yuan et al. 2013) and Mandarin Hub-4 Broadcast News Speech (~40% relative error reduction) (Yuan et al. 2014).

4.2 Phone boundary models for automatic scoring of Mandarin proficiency

It is well known that some phonetic contrasts are more difficult in language learning. The retroflex consonants (/zh, ch, sh, r/) in Mandarin Chinese, for example, are difficult to learn for many speakers whose first language does not have retroflex sounds. The pronunciation of these consonants is a prominent cue for native speakers to perceive accent. Phone boundaries may also contain useful information about a speaker's language proficiency. The timing of voicing in stop consonants, which is measured by voice onset time (VOT), is a boundary-bound phonetic feature that has been extensively studied in linguistics (Lisker and Abramson 1964; Cho and Ladefoged 1999). The VOT of stops varies across languages. Individuals who learn an L2 later in life often fail to produce consonants with authentic VOT values in the L2 (Flege 1991).

Having both phone and phone boundary models in forced alignment, we can compare the "goodness" of different phones and phone boundaries in automatic scoring of Mandarin proficiency. Following the method in (Witt and Young 2000), we computed a goodness-of-pronunciation score for every phone and phone boundary in the *Putonghua Shuiping Ceshi* corpus. The idea is to find the posterior probability of a phone p given its acoustic segment $O^{(p)}$, $P(p|O^{(p)})$, which can be approximated by the likelihood of $O^{(p)}$ corresponding to phone p, divided by the maximum likelihood of $O^{(p)}$:

$$\mathrm{GOP}(p) = \log \frac{p(o^{(p)} \mid p)}{\max_{q \in Q} p(o^{(p)} \mid q)}$$

where Q is the set of all phone and boundary models trained on "standard" Mandarin speech. The acoustic segment boundaries of $O^{(p)}$ and the corresponding likelihood (the numerator) was determined by forced alignment. To compute the maximum likelihood of $O^{(p)}$ (the denominator), all utterances were recognized using the acoustic models and an unconstrained phone and boundary loop. The likelihood of $O^{(p)}$ corresponding to the best hypothesis within its boundaries (it may contain more than one phones or boundaries) was used to approximate its maximum likelihood. The goodness of pronunciation scores are expected to have a positive correlation with human scores: A lower goodness of pronunciation score suggests that the phone or boundary fits the "standard" models less well hence should receive a lower proficiency score.

For every speaker in the dataset, we calculated his/her mean goodness of pronunciation score on every phoneme. The phone boundaries were grouped into two types: within-syllable (i.e., boundaries between an initial and a final) and cross-syllable (i.e., boundaries between a final and an initial), and a mean goodness of pronunciation score was calculated for each type. For each phone and boundary type, we then computed the correlation between all speakers' mean goodness of pronunciation scores and their proficiency scores. The results are listed in Table 3.

Table 3. Correlations between goodness of pronunciation and proficiency scores.

Phone or boundary	Correlation (Pearson's r)	Phone or boundary	Correlation (Pearson's r)
within-syl	**0.472**	g	0.157
cross-syl	**0.445**	r	0.144
iii	**0.422**	b	0.141
sh	**0.383**	uan	0.126
zh	**0.327**	m	0.125
s	0.277	iao	0.120
a	0.271	iu	0.114
ch	**0.269**	ai	0.114
ian	0.256	ei	0.112
i	0.245	n	0.111
ing	0.238	eng	0.110
d	0.225	en	0.102
h	0.225	ie	0.100
an	0.224	k	0.060
l	0.214	ong	0.054
z	0.210	uo	0.052
q	0.202	ao	0.045
t	0.194	iang	0.041
j	0.192	u	0.036
f	0.190	ang	0.029
in	0.182	v	0.019
x	0.179	ii	0.007
ui	0.174	***e***	***-0.004***

*The correlations lower than 0.12 are not significant.

We can see from Table 3 that the correlation varies greatly across phonemes. The two boundary types have the highest correlations, suggesting that phone boundaries are more helpful than phonemes in automatic proficiency scoring. Within-syllable boundaries work better than cross-syllable boundaries. Among the phonemes, the retroflex consonants, /zh, ch, sh/, and the vowel following these consonants, /iii/, are better than the others. The vowel /e/ is the only phoneme that has a negative correlation, although the correlation is not significant. /e/ appears in the possessive particle "的" (de0) in Mandarin Chinese, which is the most frequent word in the language. In our dataset, there are 23,501 /e/ tokens, 15,919 (64.7%) of the tokens were from the word "的" (de0).

5 Conclusion

In this article, we have illustrated the integration of forced alignment, a technique developed in automatic speech recognition, into corpus-based phonetics research. We have discussed three aspects of this research: forced alignment as a tool for phonetic segmentation, forced alignment as a method for investigating speech variation, and efforts to improve the technique of forced alignment itself. The integration of techniques from speech technology is helping the field of phonetics to enter a new era: a movement from the study of small, mostly artificial datasets to the analysis of published corpora of natural speech that are thousands of times larger.

Much remains to be done. In particular, we need to do a better job of bridging the gap between standard orthographic transcriptions and phonetic representations. Because natural speech is so highly variable, 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. Moreover, orthographic transcriptions are often inaccurate or incomplete, typically omitting most disfluencies and self-corrections. Future research needs to do a better job of modeling pronunciation variation (e.g., deletion, reduction, and insertion), disfluencies and imperfect transcription in forced alignment.

References

Bates D., Maechler M., Bolker, B., and Walker S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67: 1-48.

Cho, T. and Ladefoged, P. 1999. Variation and Universals in VOT: Evidence from 18 Languages. *Journal of Phonetics* 27: 207-229.

Cucchiarini, C. 1993. *Phonetic transcription: a methodological and empirical study*, PhD thesis, University of Nijmegen.

Evanini, K., Isard, S., and Liberman. M. 2009. Automatic formant extraction for sociolinguistic analysis of large corpora. *Interspeech 2009*: 1655-1658.

Flege, J. 1991. Age of learning affects the authenticity of voice-onset time (VOT) in stop consonants produced in a second language. *Journal of the Acoustical Society of America* 89: 395-411.

Fox, M.A.M. 2006. *Usage-based effects in Latin American Spanish syllable-final/s/lenition*. Doctoral dissertation, University of Pennsylvania.

Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., Dahlgren, N., and Zue, V. 1993. *TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1*. Web Download. Philadelphia: Linguistic Data Consortium.

Godfrey, J.J., Holliman, E.C. and McDaniel, J. 1992. SWITCHBOARD: Telephone speech corpus for research and development. *Proceedings of ICASSP 1992*: 517-520.

Hosom, J.P. 2009. Speaker-independent phoneme alignment using transition-dependent states. *Speech Communication*, 51: 352-368.

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

Huang, S., Liu, J., Wu, X., Wu, L., Yan, Y., and Qin, A. 1997. *Mandarin Broadcast News Speech (HUB4-NE) LDC98S73*. Web Download. Philadelphia: Linguistic Data Consortium.

Jelinek, F. 1976. Continuous speech recognition by statistical methods. *Proceedings of the IEEE* 64: 532-556.

Johnson, K. 2004. Massive reduction in conversational American English. In Yoneyama, K. and Maekawa, K. (eds.), *Spontaneous Speech: Data and Analysis*. *Proceedings of the 1st Session of the 10th International Symposium*: 29-54.

Labov, W., Rosenfelder, I., and Fruehwald, J. 2013. One hundred years of sound change in Philadelphia: Linear incrementation, reversal, and reanalysis. *Language*, 89: 30-65.

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

Lisker, L. and Abramson, A. 1964. A Cross-Language Study of Voicing in Initial Stops: Acoustical Measurements. *Word* 20: 384-422.

Sonderegger, M. and Keshet, J. 2012. Automatic measurement of voice onset time using discriminative structured prediction. *The Journal of the Acoustical Society of America* 132: 3965-3979.

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

Wieling, M., Grieve, J., Bouma, G., Fruehwald, J., Coleman, J., and Liberman, M. 2016. Variation and change in the use of hesitation markers in Germanic languages. *Language Dynamics and Change*: 199-234.

Wightman, C. and Talkin, D. 1997. 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*, 313-323, Springer Verlag, New York.

Witt, S. and Young, S. 2000. Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication* 30: 95-108.

Young, S. J., Odell, J.J. and Woodland, P.C. 1994. Tree-based state tying for high accuracy acoustic modeling. *Proceedings of the ARPA Workshop on Human Language Technology*: 307–312.

Yuan, J. and Liberman, M. 2015. Investigating consonant reduction in Mandarin Chinese with improved forced alignment. *Proceedings of Interspeech 2015*: 2675-2678.

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

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

Yuan, J., Xu, X., Lai, W., and Liberman, M. 2016. Pauses and pause fillers in Mandarin Monologue Speech: the effects of sex and proficiency. *Proceedings of Speech Prosody 2016*: 1167-1170.