Phone Boundary Annotation in Conversational Speech

Yi-Fen Liu[1], Shu-Chuan Tseng[2], J.-S. Roger Jang[3]

[1]Institute of Information Systems and Applications, National Tsing Hua University
101, Section 2, Kuang-Fu Road, Hsinchu, Taiwan
[2]Institute of Linguistics, Academia Sinica
128, Section 2, Academia Road, Taipei, Taiwan
[3]Department of Computer Science & Information Engineering, National Taiwan University
1, Section 4, Roosevelt Road, Taipei, Taiwan
E-mail: yifenliu@gmail.com, tsengsc@gate.sinica.edu.tw, jang@csie.ntu.edu.tw

Abstract

Phone-aligned spoken corpora are indispensable language resources for quantitative linguistic analyses and automatic speech systems. However, producing this type of data resources is not an easy task due to high costs of time and man power as well as difficulties of applying valid annotation criteria and achieving reliable inter-labeler's consistency. Among different types of spoken corpora, conversational speech that is often filled with extreme reduction and varying pronunciation variants is particularly challenging. By adopting a combined verification procedure, we obtained reasonably good annotation results. Preliminary phone boundaries that were automatically generated by a phone aligner were provided to human labelers for verifying. Instead of making use of the visualization of acoustic cues, the labelers should solely rely on their perceptual judgments to locate a position that best separates two adjacent phones. Impressionistic judgments in cases of reduction and segment deletion were helpful and necessary, as they balanced subtle nuance caused by differences in perception.

Keywords: Inter-pausing unit, reduced words, forced alignment, human verification

1. Introduction

Spoken corpora with signal-aligned phone boundary annotation are indispensable language resources for quantitative speech analyses and automatic speech systems. Well-annotated spoken corpora in various languages have proved to be useful for studying pronunciation variants and developing automatic speech systems, for instance the Switchboard Corpus, the Buckeye Corpus, the Spoken Dutch Corpus, and the Corpus of Spontaneous Japanese (Greenberg et al., 1996; Pitt et al., 2005; Raymond et al., 2002; Oostdjik, 2000; Maekawa, 2003). Moreover, knowledge about speech variability has successfully improved recognition error rates of ASR systems (Liu, 2004; Tsai et al., 2007). This type of phone-aligned data was also used for evaluating the performance of automatic segmentation systems (Wang et al., 2008).

However, the problem with producing phone-aligned spoken corpora lies not only in the high cost of time and man power, if the labeling is purely done by hand. It is also difficult to assess the applicability and granularity of annotation criteria and to achieve high inter-labeler's consistency. In the case of automatic processing such as applying a recognizer or an aligner to produce boundary alignment, the equivalence to human perceptual judgment is often doubtful. In particular, the thresholds for recognizing segment deletion based on acoustic features, which are adopted by system developers, may not always correspond to those set and shared by human (Wester et al., 2001).

This is the main reason why we adopted a combined verification approach to construct phone boundary annotation for a conversational speech corpus. Phone boundaries that were automatically generated by an aligner were verified by human labelers. The labelers did not exactly label the phone boundary, but to give judgments whether the separation of two adjacent phones is perceptually acceptable for them. New boundaries were only assigned, when the labelers disagreed with the original locations and were in the opinion that the new ones can better separate the two adjacent phones.

The organization of this paper is as follows. Section 2 describes the dataset and the annotation procedure. The principles and results of the verification experiment are presented in section 3, followed by a general discussion and conclusion.

2. Automatic forced alignment

2.1 Dataset

The 42-hour Taiwan Mandarin Conversational Corpus (the TMC Corpus)[1] has been constructed at the Institute of Linguistics, Academia Sinica (Tseng, 2013) accounting for three different corpus scenario settings (free conversation, task-oriented and map task dialogues). The TMC Corpus contains 500K orthographically transcribed Chinese words. Speaker turn boundaries are annotated in PRAAT format by the transcribers (Boersma & Weenink, 2012). A dataset of 3.5 hours of speech extracted from the TMC Corpus was used for the present study. It consists of 702 long speaker turns produced by 7 male and 9 female speakers. Boundaries of syllables and individual instances listed in Table 1 are annotated in PRAAT, which are used later as cues for segmenting the original speech data into

[1] Details about the TMC Corpus please refer to the official website http://mmc.sinica.edu.tw.

inter-pausing units.

Types	Total
Ordinary syllables	52,314
Silence	1,465
Speech-like paralinguistic sounds: laughing while speaking	1,408
Noise-like paralinguistic sounds: inhaling, breathing, coughing, etc.	3,108
Syllable fragments	601
Foreign words	57
Fillers (with one or more than one syllables)	428
Discourse particles (originated from Mandarin Chinese or other Chinese dialects)	1,410
Total	60,791

Table 1: Data overview.

2.2 Phone boundary annotation: The procedure

As discussed in the introduction session, both automatic generation and human labeling have advantages and disadvantages. The procedure we suggest in this paper tries to reduce the cost and to raise the level of agreement in the sense of human perception. The procedure is illustrated in Figure 1. Since we chose long speaker turns, a pre-segmentation of the sound files into smaller units was necessary. Silent pauses, speech-like and noise-like paralinguistic sounds that were marked in our dataset were used as segmentation cues to obtain inter-pausing units (IPU) (Bigi & Hirst, 2012; Schuppler et al., 2011).

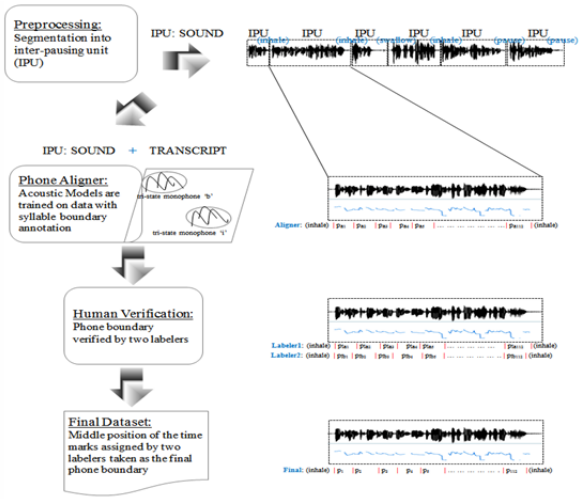

Figure 1: A combined annotation procedure.

IPUs that did not contain Mandarin Chinese at all, but dialects spoken in Taiwan, were excluded from our final dataset. As a result, 5,276 IPUs were used for the phone boundary verification experiment. The average length of IPUs is 10.4 syllables, approximately equivalent to seven words in Mandarin Chinese. Making use of the sound and the transcript, we first ran the automatic phone aligner to obtain an initial version of phone boundaries. Subsequently, the boundaries were verified by human labelers. After a satisfactory rate of boundary deviation between the labelers was achieved, the final dataset was automatically derived from the verified boundaries with no human intervention.

2.3 Phone set

Acoustic models trained and decoded in continuous speech recognition systems developed for Mandarin Chinese are often based on INITIAL and FINAL components (Tsai et al., 2007). INITIAL and FINAL correspond to onset and rhyme in phonological terms. That is, a FINAL may consist of a nucleus and a coda consonant, implying a higher level than the level of phones. If the acoustic information is learned or trained from this level, it would not be possible to account for pronunciation variant in extremely reduced words such as those with a deleted nucleus, but the coda is present. Thus, we decided to use *phone* as the decoding unit. Previous corpus-based study on syllable contraction in Mandarin conversational speech showed that the reduction of multi-syllabic sequences resembles a spectrum of reduced phonetic representations with syllable merger as its target form (Tseng, 2005). Applying the general contraction rules, a considerable number of pronunciation variants involving (at least) disyllabic words that are reduced to different degrees should be predictable based on the canonical form. Ranging from consonant deletion, nuclei merging, to syllable merger, these pronunciation variants that are often found in conversational speech, cannot be properly taken into account, if only INITIAL and FINAL are distinguished for syllables.

Conventionally, a Mandarin Chinese syllable has the form of CGVX with C denoting a consonant onset, G a glide, V a vowel, and X a consonant coda (Duanmu, 2000; Ho, 1996). A set of 22 consonants presented in terms of the place of articulation in Table 2 can occupy the onset position, except for /ŋ/. Only the nasals /n/ and /ŋ/ can appear in coda position. Our inventory set also contains two glides /j/ and /w/, and 15 vowels including monophthongs and diphthongs /i, ɨ, u, ʉ, y, a, o, ə, e, ɚ, ai, ei, au, ou, ye/.

Bilabial		p	p^h	m	
Labiodental	f				
Dent-alveolar		t	t^h	n	l
Alveolar	s	ts	ts^h		
Retroflex	ʂ	tʂ	$tʂ^h$		ʐ
Alveolo-palatal	ɕ	tɕ	$tɕ^h$		
Velar	x	k	k^h	ŋ	

Table 2: Mandarin consonants.

2.4 Training a phone aligner

To construct the phone aligner, we first built acoustic models for individual phones found in ordinary syllables. For the other instance types listed in Table 1, we built acoustic models for each of them, except for fillers and discourse particles, because their phonetic representations can be categorized into distinct types systematically. For fillers, we accounted for three different acoustic models

by considering the nasality and the length in syllables: two filler types for monosyllabic instances, one with and one without a nasal coda, and one type for multi-syllabic fillers irrespective of the presence of a final coda. Four acoustic models were trained for discourse particles. One was applied to high frequency particles; one to particles originated from a Chinese dialect (Southern Min) that is dominantly spoken in Taiwan; and two to Mandarin Chinese particles that were grouped together according to the degree of similarity in syllable structure.

As a result, 51 tri-states acoustic models were used for training our aligner with the HTK toolkit (Young et al., 2006). The models, facilitated by the already annotated syllable boundaries, were trained at a frame shift of 5ms and a window length of 15ms, where for each frame, 13 MFCCs (the mel-scaled cepstral coefficients C0-C12) and their first and second order derivatives (39 features) were calculated. For the current phone aligner, fundamental frequency is not taken into account, although pitch-related features may play an important role in understanding and processing Chinese lexical tones. With a frame shift of 5ms and no skipping three emitting states of models, the phones will be assigned a minimum length of 15ms within manually pre-determined syllables. For the verification experiment, the results of applying 5,276 IPUs of data to our phone aligner were converted to a PRAAT readable format (Boersma & Weenink, 2012) with the transcript information on one tier and the aligned phone boundaries on the other, as illustrated in Figure 2.

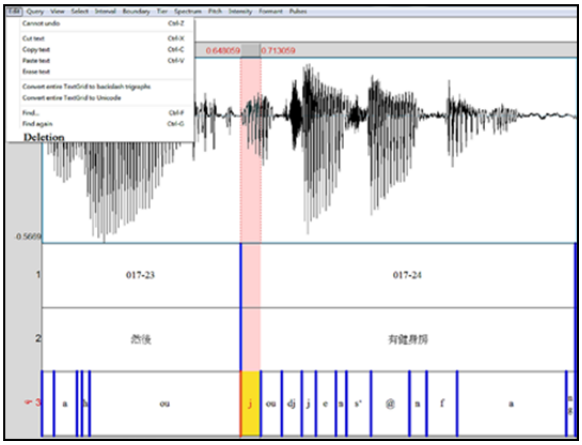

Figure 2: PRAAT format for human verification.

3. Human verification of boundaries

The 3.5-hour speech dataset contains in total 133,339 phones. We used 10% for the training phase and 90% for the evaluation phase. The summary in terms of phone categories is given in Table 3. Six labelers with prior experiences with speech data annotation were grouped into three pairs (A, B, and C). We ran a training phase to acquaint our labelers with the verification criteria. The training data was divided into three subsets. Each subset was verified by two labelers in the same group, i.e. all data were examined by two labelers. Mutual discussion among labelers was allowed in the training phase to achieve a common consensus. The threshold we set for a satisfactory inter-labeler's consistency was that 85% of the verified phone boundaries should not deviate more than 20ms within each pair of labelers. The labelers did not start with the actual experiment until they achieved the consistency threshold. In the evaluation phase, no discussion was allowed. The entire experiment period including the training and evaluation phases last three months.

Phone categories	Occurrences
Monophthongs	39,097
Nasals	17,951
Glides	15,369
Plosives	14,685
Diphthongs	13,218
Retroflex	9,021
Affricates	7,351
Filler+para. sounds	7,067
Approximant	3,095
Fricatives	2,995
Lateral	2,080
Particles	1,410
Total	133,339

Table 3: Phones in the dataset

3.1 Phone boundary verification

3.1.1. Separation of two adjacent phones

The essential, but also the most difficult, issue for labeling phone boundaries in natural speech is the transition from one phone to the other. No concrete rule was instructed to the labelers how they should deal with the inter-segmental transition. They should only rely on their perceptual judgments to decide on a location that best separates two adjacent phones. Thus, visualized acoustic cues such as pitch, spectrogram and intensity information were not provided to the labelers. They made their decisions based on the listening only. On the one hand, the perceptual judgment of the labelers served as additional modification to the acoustically based alignment result. On the other hand, as our data have been verified by two labelers, the perceptual discrepancy between the labelers is regarded as a kind of balance of individual perceptual differences.

3.1.2. Allophones

When words were correctly transcribed in the text, but the labelers with a sensitive hearing perceived allophonic differences, they should ignore the allophonic variation, as it does not affect the word meaning. On the one hand, it was not the goal of this study to do narrow transcription. On the other hand, it would be difficult to keep up a high inter-labeler consistency rate, if allophonic differences in the dataset are to be considered consistently.

3.1.3. Marking deletion

What should the labelers do, if phones that should appear in the canonical form are not heard in the speech? In this case, the words were correctly transcribed, but some of the phones were audibly not present. This is a known, difficult problem labelers are often encountered with,

when they are to determine phone boundaries in natural, continuous speech such as conversational speech. To preserve all phone labels generated by the forced aligner for later automatic processing and result analysis, as well as to reduce the effort of locating a boundary for non-existing phones, we designed a function in PRAAT to help labelers mark deletion. If a phone was absent for labelers, they were instructed to use the **Deletion** function that automatically minimize absent phones to an interval of 5ms, as shown in the very bottom of the menu in Figure 2. Using this function, the deletion is annotated without spending unnecessary time for determining the extremely difficult boundaries.

The examples shown in Figure 3a-c are three variants of the disyllabic word *ranhou* /ʐan xou/ ('then'). Figure 3a shows an instance of *ranhou* with all phones present. In Figure 3b, the coda nasal /n/ was marked as deleted by the labelers, whereas Figure 3c shows both the nasal coda of the first syllable and the onset of the second syllable are not present.

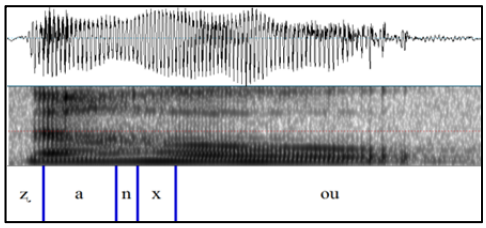

Figure 3a: No deletion in *ranhou* /ʐan xou/.

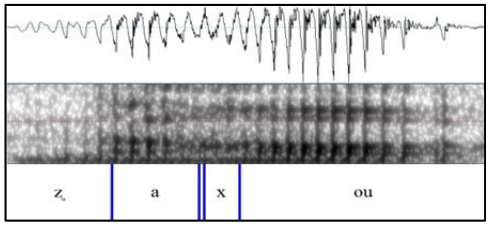

Figure 3b: Deletion of /n/ in *ranhou* /ʐan xou/.

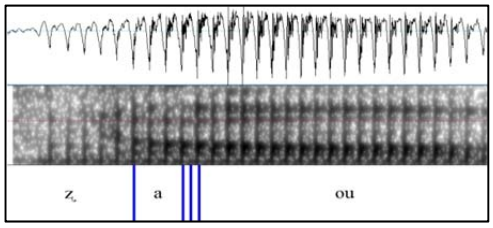

Figure 3c: Deletion of /n/ and /x/ in *ranhou* /ʐan xou/.

3.1.4. Transcription errors

In the case of transcription errors, that is, when the content of the speech did not exactly match with that transcribed in the text, the labelers had to add marks to the phones of the entire sequence. Speech errors, homograph errors, and non-Mandarin words were dealt with in the same way. In total, they made up about 1% of the entire dataset. The evaluation of the experiment result did not account for these phones.

3.2 Experiment results

3.2.1. Verification results

The verification results of three labeler pairs are shown in Table 4. In each of the three pairs, more than 90% of the phone boundaries assigned by the two labelers deviated from each other with an interval less than 20ms. In the calculation of deviation, we compared the ending boundary of each phone. The deviation in the evaluation phase was reduced compared to that in the training phase. Furthermore, to closely observe the verification results, Figure 4 summarizes the results in terms of boundary deviation at a time step of 5ms. The results are similarly distributed across the three pairs, suggesting a successful training phase and a good inter-labelers' consistency.

Labelers	Training set		Evaluation set	
	phones	< 20ms	phones	< 20ms
Pair A	4,832	87.75%	47,057	92.95%
Pair B	4,541	90.71%	44,497	92.16%
Pair C	4,609	86.90%	26,665	93.68%
Total	13,982	88.43%	118,219	92.82%

Table 4: Labelers' consistency.

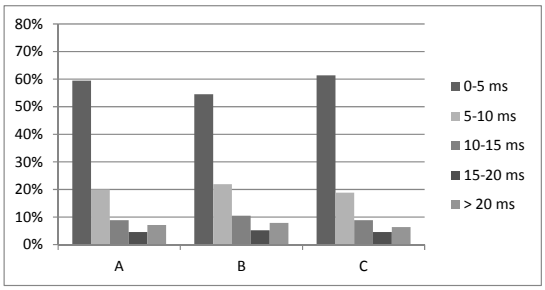

Figure 4: Phone boundary deviation.

3.2.2. Phones with severe deviation

Table 5 lists the results of phone categories that caused deviations larger than 20ms. Glides /j, w/, monophthongs /a, ə/, and the lateral /l/ were more problematic than the other phone categories. These phone categories caused a large boundary discrepancy between the labellers in more than 10% of the phones. Contrastively, fricatives, plosives, and affricates were less problematic with regard to boundary annotation. One possible reason for this result may lie in different transition types of phone categories. While the transition in sonorants is continuous and non-abrupt, fricatives, plosives, and affricates are more clearly presented in acoustic configuration that leads to a more successful alignment of the phone aligner.

Phone categories	Phones (>20ms deviation) No. (% in dataset)
Glides	1,821 (11.85%)
Monophthongs	4,162 (10.65%)
Lateral	219 (10.53%)
Approximant	240 (7.75%)
Diphthongs	912 (6.9%)
Nasals	1,184 (6.6%)
Retroflex	547 (6.06%)

Particles	84 (5.96%)
Plosives	646 (4.4%)
Affricates	267 (3.63%)
Fricatives	83 (2.77%)
Filler+para. sounds	93 (1.32%)
Total	10,258 (7.8%)

Table 5: Phones annotation with a large deviation.

3.2.3. Phones marked as deleted

Table 6 lists the occurrences of phone categories that were marked deleted by at least one labeler, with information about the percentages in the overall data. As a result, 3,690 phones in our dataset were marked as *not present* (deleted), making up 2.8% of the overall data. Among the 3,690 phones, only about 20% were not audible for both labelers, suggesting that the thresholds each person sets for phone deletion vary individually. It was more likely for the approximant /x/, the lateral /l/, and the glides /j, w/ to be marked deleted. For the nasal /n/, 445 out of the 497 deleted occurrences are located in coda position. Only 52 occurrences of the deleted /n/ are in onset position. Among 536 deleted monophthongs, 214 were Schwa.

Phone categories	Deletion No. (% in dataset)
Glides	881 (5.73%)
Nasals	578 (3.22%)
Monophthongs	536 (1.37%)
Plosives	441 (3.00%)
Approximant	313 (10.11%)
Retroflex	301 (3.34%)
Affricates	206 (2.80%)
Fillers+para. sounds	160 (2.26%)
Lateral	113 (5.43%)
Fricatives	94 (3.14%)
Diphthongs	65 (0.49%)
Particles	2 (0.14%)
Total	3,690 (2.8%)

Table 6: Phone omission.

3.3 Final dataset

For producing the final dataset, the middle position of the boundaries assigned by the labelers was calculated and taken as the final phone boundary with no further human intervention.

4. Discussion

Utilizing our phone-aligned dataset, a wide range of research works concerning conversational speech is possible. For instance, deleted phones in reduced spoken words are normally judged and identified by phonetically trained experts, often with a focus on phone sequencing instead of the contextual information such as the perception of spoken words. As we have marked deleted phones in the experiment, we are able to identify what kinds of words are more likely to be reduced by observing words in which at least one phone was marked deleted. Reduced words with omitted segments normally also cause obvious difficulties for automatic speech systems to recognize or to align. Besides, location relative to the position in the IPUs also provides contextual information about the surface form of words and possibly also the discourse structure associated with these words.

Table 7 lists ten words that have the highest percentage rates of deletion in our dataset. In the TMC Corpus, these words are among the top 100 most frequent words (Tseng, 2013). As shown in Table 7, different preferences for IPU positions are found. It is more likely for *yinwei* (because) and *ranhou* (then) to be reduced in IPU initial position. *Dui* tends to be reduced in both IPU initial and medial positions, but less likely in IPU final position. The rest of words are more likely to be reduced, when they are in IPU medial position. Concerning the deleted phones, we found some systematic distributions.

When a word contains glides, it is usually the glide that is deleted, e.g. *wo*, *women*, *jiu*, and *dui*. In the case of disyllabic words, it is usually the onset of the second syllable that is deleted, e.g. *ranhou*, *xianzai*, and *shihou*. But for disyllabic words with a plural suffix, the preference for segment deletion differs. The Schwa and the coda of the second syllable (the suffix) tend to be deleted and the bilabial nasal onset of the suffix tends to be preserved. For monosyllabic words like *dui* and *de*, though both with a plosive onset consonant, the reduction forms are quite different. *De* is a structural particle that is usually preceded by a head, while *dui* is a preposition that is usually followed by a head, or a predicate that has a function of acknowledgement. The frequently reduced segment in *de* is the onset (the onset is in the juncture between the head and *de*), but the glide in *dui*. These preliminary observations with regard to deletion point out the interrelationship between reduction degree, IPU position (i.e. position in a spoken discourse), and the phonological property of segments. To closely study reduced forms in different IPU positions, we are currently conducting a free phone decoding experiment to automatically generate phone sequences for individual words and then select the prototypical pronunciation variants. Phonological and phonetic rules of segment deletion will be compared in a systematic way.

	IPU initial	IPU medial	IPU final	No. in dataset	Total deletion
tamen (they)	10	77 (47%)	6	165	56%
ranhou (then)	93 (28%)	66 (20%)	8	328	51%
yinwei (because)	91 (32%)	39 (14%)	12	287	49%
jiu (then, so)	9	63 (28%)	5	226	34%
shihou (time)	0	69 (30%)	8	229	34%
xianzai (now)	14	78 (18%)	6	433	23%
dui (for, yes)	48 (9%)	48 (9%)	6	509	20%
de (structure part.)	0	104 (17%)	5	598	18%
women (we)	49	184 (13%)	7	1,460	16%
wo (I)	11	61 (5%)	8	1,362	6%
Total	325	789	71	5,597	21%

Table 7: Deletion in different prosodic positions.

5. Conclusion

While expert opinions on fine details may play a decisive role in phonetic analysis, it is also important to make available large-scale well-annotated spoken language resources. Phonetic representations of spoken words may

largely deviate from their canonical forms in conversational speech. By adopting a combined procedure of automatic phone alignment and perceptual verification, we constructed a phone-aligned dataset of conversational speech with relatively low cost of human power, but relatively high annotation quality.

6. Acknowledgements

This work was supported by the Taiwan International Graduate Program and the Institute of Linguistics of Academia Sinica granted to the first author as well as the NSC project 100-2410-H-001-093 granted to the second author. The authors would like to sincerely thank the team members who have prepared the corpus along the years and those who participated in the verification experiment. Without their hard work, we would not be able to obtain the results presented in this paper.

7. References

Bigi, B., Hirst, D. (2012). Speech Phonetization Alignment and Syllabification (SPPAS): a tool for automatic analysis of speech prosody. In *Proceedings of Speech Prosody*, pp. 19-22.

Boersma, P., Weenink, D. (2012). Praat: doing phonetics by computer. Software package, www.praat.org.

Duanmu, S. (2000). *The phonology of standard Chinese*. New York: Oxford University Press.

Greenberg, S., Hollenback, J., Ellis, D. (1996), Insights into Spoken Language Gleaned from Phonetic Transcription of the Switchboard Corpus. In *Proceedings of the International Conference on Spoken Language Processing* (ICSLP-96), pp. S24-27.

Ho, D.-a. (1996). *Some Concepts and Methodology of Phonology*. Da-An Press. (in Chinese)

Liu, Y., Fung, P. (2004). State-dependent Phonetic Tied Mixtures with Pronunciation Modeling for Spontaneous Speech Recognition. *Journal of IEEE Transactions on Speech and Audio Processing*, 12(4), pp. 351-364.

Maekawa, K. (2003). Corpus of Spontaneous Japanese : Its design and evaluation. In *Proceedings of Spontaneous Speech Processing and Recognition* (SPSS), 4, pp. 7-12.

Oostdjik, N. (2000). The Speech Dutch Corpus. Overview and First Evaluation. In *Proceedings of the International Conference on Language Rescourses and Evaluation* (LREC), pp. 887-894.

Pitt, M. A., Johnson, K., Hume, E., Kiesling, S., Raymond W. (2005). The Buckeye Corpus of Conversational Speech : Labeling Conventions and a Test of Transcriber Reliability. *Speech Communication*, 45, pp. 89-95.

Raymond, W. D., Pitt, M., Johnson, K. Hume, E. (2002). An Analysis of Transcription Consistency in Spontaneous Speech from the Buckeye Corpus. In *Proceedings of the International Conference on Spoken Language Processing* (ICSLP-02), pp. 1125-1128.

Schuppler, B., Ernestus, M., Scharenborg, O., Boves, L. (2011). Acoustic Reduction in Conversational Dutch: A Quantitative Analysis based on Automatically Generated Segmental Transcriptions. *Journal of Phonetics*, 39, pp. 96-109.

Tsai, M.-Y., Chou, F.-C., Lee, L.-S. (2007). Pronunciation Modeling with Reduced Confusion for Mandarin Chinese Using a Three-Stage Framework. *Journal of IEEE Transactions on Speech and Audio Processing*, 15(2), pp. 661-675.

Tseng, S.-C. (2005). Contracted Syllables in Mandarin: Evidence from Spontaneous Conversations. *Language and Linguistics*, 6(1), pp. 153-180.

Tseng, S.-C. (2013). Lexical Coverage in Taiwan Mandarin Conversation. *International Journal of Computational Linguistics and Chinese Language Processing*, 18(1), pp. 1-18.

Wang, H.-M., Kuo, J.-W., Lo, H.-Y. (2008). Towards A Phoneme Labeled Mandarin Chinese Speech Corpus. In *Proceedings of International Conference on Speech Databases and Assessments* (ICSDA).

Wester, M., Kessens, J. M., Cucchiarini, C., Strik, H. (2001). Obtaining Phonetic Transcriptions: A Comparison between Expert Listeners and a Continuous Speech Recognizer. *Language and Speech*, 44(3), pp 377-403.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., et al. (2006). Technical Report: The HTK book (version 3.4), Cambridge University, Engineering Department.