ISCA Archive
http://www.isca-speech.org/archive

Speech and Language Technology in Education (SLaTE2011)
Venice, Italy
August 24-26, 2011

Analysis of L2 English Speech Corpus by Automatic Phoneme Alignment

Hajime Tsubaki[1,3], *Mariko Kondo*[2,3]
[1] GITI, [2] SILS & [3]Language and Speech Science Research Laboratories,
Waseda University, Japan
hjm-tsubaki@asagi.waseda.jp, mkondo@waseda.jp

Abstract

This study tested the application of adapted HTK for automatic alignment of speech corpus of Asian speakers' English. The HTK tool with TIMIT has problems in aligning non-native speakers' English. New sets of phoneme sequences for each word were listed to test if an adapted alignment module could accurately analyze pronunciation of Japanese speakers' English. The new sets of phoneme sequences produced better alignment of Japanese accented English and showed that the L2 incorporated new alignment module could perform more accurate automatic alignment of L2 English data. The same methods should be able to be applied to other language data.

Index Terms: automatic alignment, HTK, TIMIT, speech corpus, L2 English

1. Introduction

The role of English as a lingua franca has been well established. It is used as one of the official languages or a second language in many Asian countries. Asia is the largest English speaking community and market, and provides rich variation in pronunciation, lexicon and grammar, which are blended with local languages [4][5]. Therefore it is important to learn about Asian language speakers' English. The Asian English Speech cOrpus Project (AESOP) aims to construct a common shared large scale English speech corpus of Asian language speakers [8]. It currently involves institutions in Japan, Taiwan, Hong Kong, China, Indonesia, Korea, Myanmar, Thailand, and Vietnam, and there are a few more institutions in other countries that have also shown interest in the project. In order to construct the required large scale Asian English speech corpus for AESOP, we set up a common recording platform and shared experimental task set for data collection.

The AESOP recording platform covers typically common phonetic and phonological problems in English as a Foreign Language (EFL) for Asian language speakers. Problems include (1) phonemes, (2) syllable structure, (3) prosody and (4) phonological rules. For example, certain vowels and consonants, and their allophonic variations, are problems for many Asian speakers; other common problems include stress and vowel reduction, assimilation, elision, and syllable structure ([3]).

AESOP-collected corpora are an open resource. Therefore a common, open-ended annotation system is needed that can be shared by all participating institutions. In particular, considering the expansion of the project and increasing number of participating institutions, it is desirable to have communal methods of data analysis using free software and commonly available programs. This is very important because we expect to create a very large speech corpus, but the data collection facilities vary from institution to institution and country to country. Therefore, we have used Hidden Markov Model Toolkit (HTK) modules [10] to align data automatically because it is free and widely available. Another potential advantage of using the HTK modules is that they can provide useful information from the corpus to a wide range of potential users, such as phoneticians, language teachers, speech scientists, and engineers.

However, we have come across a few problems in aligning L2 English utterances. The HTK modules with ARPABET were essentially developed to annotate native American English speakers' utterances and are not necessarily suitable for L2 English utterances. Non-native speakers' English does not always match native English speakers' utterances in terms of segmental and prosodic information. A previous study overcame these problems at the word level by using an acoustic model adaptation with non-native English speech data [7]. In our current study we wanted to work at the sentence level by adding and altering phonemes in order to reflect non-native accented English speech. Therefore, to overcome the previously mentioned problems we used the HTK modules with expanded pronunciation dictionary files to enable them to accurately detect acoustic information of, initially, one Asian language speakers' English, namely Japanese. In this study we tested the application of the adapted HTK for speech corpus of Japanese speakers' English.

If the adaptation is shown to be successful we intend to extend the adaptations to other language speakers' English.

2. Language specific problems of English for Japanese speakers

The AESOP platform covers typical problems in EFL common to many Asian language speakers, such as (1) – (4) in Section 1. There are also some language specific problems typical of Japanese speakers.

2.1. Segmental problems

2.1.1. Consonants

Since Japanese has fewer consonantal phonemes than English, Japanese speakers map unfamiliar English phonemes to close Japanese phonemes. Some English consonants which are phonemically distinct are realized as the same sound because there is no phonemic contrast in Japanese: e.g. in Japanese, both /b/ and /v/ are recognized as /b/, /s/ and /θ/ are recognized as /s/, /l/ and /r/ recognized as /r/, and /z/, /ð/, and /ʒ/ can be allophones of /z/ ([dʑ]) (Table 1(a)). Allophonic variations of some consonants in Japanese are context specific, and only fixed combinations of consonant and vowel are phonotactically permitted. As shown in Table 1(b), in Japanese, /s/ is realized as the allophone [ɕ] before /i/. The /t/ sound typically has two allophones, namely [tɕi] before /i/ and [ts] before /u/ ([ɯ]). Therefore, [ti] and [tsi] are not natural sequences for Japanese speakers. Similarly [tɯ] (/tu/) is not permitted except in recent loanwords. The /h/ has an allophone [ɸ] which is used as a substitute sound for [f] in Japanese, occurs only before /u/ in native words, and therefore the [hu] sequence is unnatural for Japanese speakers and it is mixed

with [ɸɯ]. Similarly, [ɸo] does not occur in native words, so [fo] is often interpreted as [ho].

Table 1. *Mismatch of English and Japanese consonants.*

(a) Phonemic differences

English	b	v	s	θ	z	ð	ʒ	dʒ	l	r
Japanese	b		s		dz				r	

(b) Phonotactic differences

English	si	ʃi	ti	tʃi	tsi	hu	fu	ho	fo
Japanese	si		ti~tɕi			ɸɯ		ho	

2.1.2. Vowels

Japanese has 5 vowels /i, e, a, o, u/, whereas English has many vowels; e.g. General American has 15 vowels, RP has 19 vowels, Australian has 18 or 19, New Zealand has 17-19, and Canada has 14 vowels [9]. It is likely that Japanese speakers ignore phonemic differences of English vowels and categorize English vowels into five equivalents of Japanese vowels.

2.1.3. Syllable structure

Japanese permits fairly simple syllable structure compared with English (Table 2). Japanese allows a maximum of six elements in a syllable, but about 90% of syllable structure is a light open syllable /(C)V/, and only limited consonant cluster types are allowed. The only consonant allowed in a syllable final position in Japanese is either the first element of geminate consonants or a moraic nasal. With the initial consonant of the following syllable, it forms the only possible consonant clusters. English has up to three consonants in an onset and three in a coda.

Table 2. *Syllable structures of English and Japanese.*

English	Japanese
(C)(C)(C)V(C)(C)(C)	(C)(j)V(V)(C)(C)

Table 3 shows loanword adaptation rules in Japanese. Since the rules do not allow any consonant deletion, English consonant clusters are broken up by vowel insertion. Also syllable final consonants are closed by a vowel, except for either a moraic nasal /N/ or a geminate consonant [6].

Table 3. *Vowel insertion in Japanese loanword adaptation. Inserted vowels are underlined.*

	English			Japanese adaptation
(a)	*strange*	/streɪndʒ/	➔	/su̲to̲re(i)Nzi̲/
(b)	*strengths*	/streŋθs/	➔	/su̲to̲reNgu̲su̲/

2.1.4. Stress, rhythm and vowel reduction

English uses stress to express lexical accent, and the unit of speech rhythm is the foot. This rhythmic characteristic affects vowel quality and duration whether or not a vowel is in a stressed syllable. In English, vowels in unstressed syllables are lax vowels and are centralized to [ə], and their durations are shorter. Japanese is a pitch accent language and is mora-timed. Therefore, the presence or absence of accent does not affect vowel quality or duration. This vowel reduction in relation to lexical stress is a difficult feature for Japanese speakers to acquire, and even advanced learners tend to use full vowels in unstressed syllables and so their F1 and F2 values are closer to those of Japanese vowels represented by alphabet letters [2].

2.1.5. Phonological rules

Some phonetic and phonological rules in English cause difficulties for Japanese speakers. Typical English phonological rules are crucial to make Japanese speakers' English sound natural and fluent. These phonological rules include allophonic variations caused by assimilation (*and then* with dental [n̪] and [d̪]; laterally released [t^{l}] in *at last*) and coalescent (*got you* → [gɑtʃuː]), elision (*next day* → [neks deɪ]), syllabification of consonant by /ə/-elision (sudden → [sʌdn̩li]), and epenthesis (*prince* → [prɪnt̩s]). These rules are important in EFL because they affect English perception and also affect production in relation to vowel insertion, syllable structure, and speech rhythm.

3. Automatic Annotation Using HTK

Japanese speakers have acquired Japanese phonology and are likely to use their L1 Japanese phonology when they speak English, so their English pronunciation is different from that of native English speakers. In their English utterances, the problems featured in Section 2 are likely to occur. Japanese speakers may produce incorrect sounds and sound sequences, insert unexpected sounds, or delete sounds.

For the AESOP corpus, we expect to accumulate a large number of English speech data from various Asian language speakers, and so a good universally available automatic annotation system is needed for its analysis.

We tested the HTK modules with the expanded dictionary files using recorded data of 20 Japanese subjects reading an English version of The *North Wind and the Sun* [1]. The subjects were university undergraduate students in Japan and their English level varied from upper elementary to advanced level.

3.1. Phoneme alignment

To obtain phoneme alignment of Japanese accented English speech, we used the Hidden Markov Model Toolkit (HTK) and expanded the associated pronunciation word dictionary files written using ARPABET symbols for phonetic transcriptions; the ARPABET symbols were developed under the Advanced Research Projects Agency (ARPA) Speech Understanding Project (1971-76).

The Hidden HTK is a set of modules for building and manipulating hidden Markov models (HMM), and is mainly used in speech recognition research. The toolkit is used for speech analysis, HMM training, testing and results analysis in acoustic engineering. Phonemes and their durations in speech are recognized and computed by HMM-based acoustic models (Figure 1). The original word dictionary file was built using the TIMIT speech corpus, which is a corpus of English sentence speech read by a large number of native speakers of American English. The corpus is designed for acoustic engineering studies, and for development and evaluation of automatic speech recognition. It contains 16-bit, 16kHz speech waveform files, phonetic symbols, and word transcriptions written with the ARPABET symbols.

We trained an HMM-based acoustic model using HTK and the TIMIT speech corpus [12]. The phonemes are registered in a word dictionary file including pairs of letter strings and phonemes for each word. Generally the word dictionary file that was used is based on the TIMIT speech corpus.

In this research, in order to analyze phoneme alignment, two word dictionary files were employed. One of the word dictionary files used only ARPABET phonetic transcriptions, based on the original native English speaker TIMIT speech corpus, which we call the original alignment module. The

second file used the same ARPABET transcriptions, together with added phonetic transcriptions that reflect Japanese accented English (Table 4), which we call the new alignment module. Since the TIMIT speech corpus is based on native English speakers' data, its phonetic transcription is straightforward and has less variation. For example, using ARPABET the phonetic transcription of the English word *blew* should be "b l uw" for the IPA transcription /blu:/ based on model pronunciation by native English speakers.

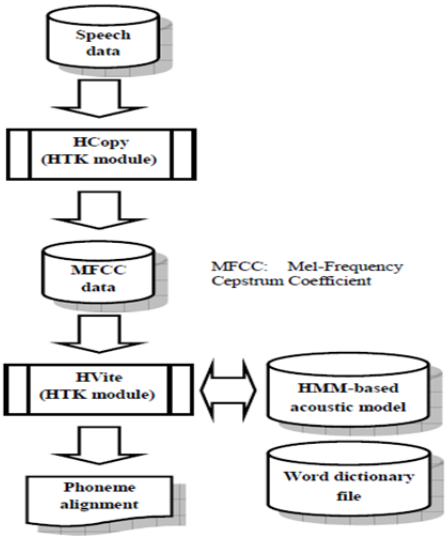

Figure 1: *Phoneme Alignment Process*

The result of the automatic alignment of actual speech data using the original dictionary file, based only on TIMIT, is shown in the upper part of Figure 2. It transcribed the Japanese speaker's utterance of '*...he blew...*' as [hi: blu:]. However, the actual pronunciation of this utterance was [hi: bʊ ɾu:]. The speaker inserted the vowel [ʊ] within the consonant cluster /bl/ and substituted the tap [ɾ] for /l/. The tap [ɾ] is a typical allophone of the Japanese phoneme /r/. The original HTK alignment did not detect these typical features of Japanese speaker's pronunciation and transcribed it as [he blu:]. Hence the alignment was incorrect.

We then repeated the automatic phoneme alignment using the new alignment module with word dictionary files adapted for Japanese speech. The result is shown in the bottom part of Figure 2. The new module listed possible pronunciations of the words by Japanese speakers including [bʊɾu:] and correctly detected the inserted [ʊ] and substitution of [ɾ] for /l/. Both the vowel insertion in consonant clusters and the use of /r/-type allophones for /l/ are very common by Japanese speakers (Section 2).

Table 4. Examples of word dictionary for words *blew* and *stronger* adopted for automatic alignment of *North Wind and the Sun* recording by Japanese speakers

Word '*blew*' /blu:/

Predicted Japanese pronunciation	ARPABET phonetic transcription	
[blu:]	b l uw	English speakers
[bʊlu:]	b uh l uw	Japanese Speakers
[bɹu:]	b r uw	
[bʊɾu:]	b uh r uw	

Word '*stronger*' /stɹɑŋgɚ/

Predicted Japanese pronunciation	ARPABET phonetic transcription	
[stɹɑŋgɚ]	s t r aa ng g er	English speakers
[stɹɒŋgəɹ]	s t r ao ng g er	Japanese Speakers
[stɹɒŋgə]	s t r ao ng g ah	
[stoɾoŋga:]	s t ao r ao ng g aa	
[sʊtoɾoŋga:]	s uh t ao r ao ng g aa	
[stlɑŋgɚ]	s t l aa ng g er	
[stlɒŋgəɹ]	s t l ao ng g er	
[stlɒŋgə]	s t l ao ng g ah	

The alignment module needs to list all possible pronunciations in the dictionary. Predicted pronunciation by Japanese speakers is quite complicated because it has to cover all possible outputs based on all EFL phonetic and phonological factors including (a) vowel insertion, (b) vowel quality, (c) consonant alternation, and (d) syllabification. All words in the new alignment module have more than one predicted pronunciation in the word dictionary file. Some words list many predicted pronunciation patterns by Japanese speakers; e.g. '*that*' = 8, '*traveler*' and '*cloak*' = 15 each, '*succeed*' = 20, 'agreed' = 27, and 'obliged' = 50.

3.2. Results

The text of The *North Wind and the Sun* contains 113 words. Analysis of the 20 speakers' data using the original alignment module found only two words, '*in*' and '*could*' that showed only one pattern. The other 111 words showed more than one pattern of pronunciation by the Japanese speakers. However, the new alignment module was able to detect fine differences in pronunciation, which the original alignment module was unable to detect. For example, there were 5 different pronunciations for the word '*fold*', corresponding to

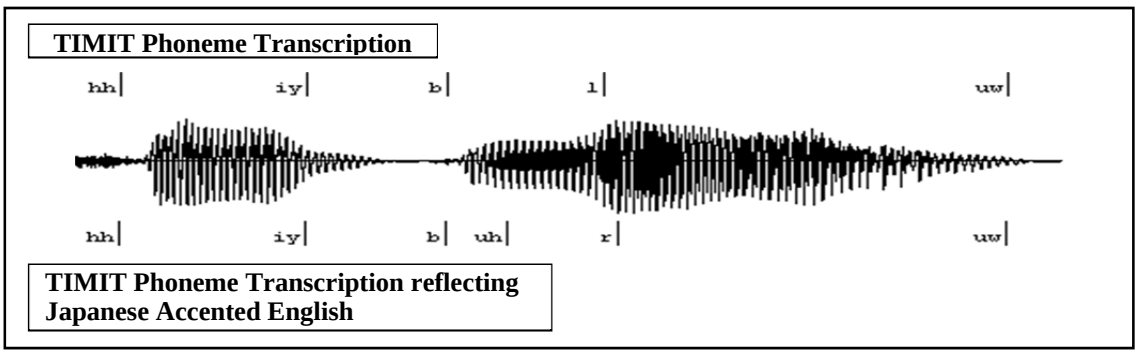

Figure 2: *The original (top) and added (bottom) transcription for automatic phoneme alignment of a word "blew" uttered by a Japanese speaker*

the original TIMIT transcription "f ow l d" (/fould/), also there were 6 different pronunciations for '*cloak*' corresponding to the TIMIT "k l ow k" (/klouk/), and 11 pronunciations for the word 'obliged' (/əblaɪdʒd/) corresponding to the TIMIT "ah b l ay jh d" (Table 5).

Table 5. Examples of alignment mismatch between the original TIMIT-based word dictionary and added word dictionary for Japanese speakers' English utterances.

(a) '*fold*' "f ow l d"

Added ARPABET	No.
f ao l d	5
f ao l d uh	2
f ao r uh d	1
hh ao l d	1
hh ow l d	4

(b) '*cloak*' "k l ow k"

k l aa k	3
k l ao k	22
k l ao k uh	7
k uh l ao k	1
k uh l ao k uh	1
k uh r ao k	3

(c) '*obliged*' "ah b l ay jh d"

Added ARPABET	No.
ah b l iy jh d	1
ah b l iy zh d	3
ah b r ay jh d	1
ah b r iy jh d	2
ah b uh r ay jh d	1
ah b uh r ay jh ih d uh	1
ah b uh r iy zh eh d	1
ao b l ay jh d	2
ao b l iy jh d	1
ao b r ay jh d	1
ao b r iy jh d	1

The new alignment module was able to more accurately align the Japanese speakers' English data than the original. In addition, the new alignment module was able to detect fine differences in pronunciation and therefore better shows features of Japanese speakers' English. These detected features included many of the typical EFL problems discussed in Section 2, including:

i. Sounds listed in Table 1 which are phonemically non-contrastive are difficult to differentiate.
ii. There are patterns in phoneme alteration.
iii. Vowel insertion within consonant clusters was persistently found in some lower level speakers.
iv. Vowel insertion after a word final consonant was found in lower level speakers' data. But they inserted lax vowels [ə], [ɔ], [ʊ] rather than [o] or [u] unlike traditional vowel insertion rules stated in loanword phonology.
v. Apart from inserted vowels as in (iii), Japanese speakers tended to use full vowels in most positions; i.e. vowel reduction did not occur in many samples.

The new alignment module still presents a few problems. It is adapted from the original alignment module that is based on native English speakers' speech data. As a result many common Japanese speakers' allophonic variations such as [ɾ] for /r/, [ɛ] for /s/ or /ʃ/, [dz] for /dʒ/, [tɛ] for /tʃ/, [ɸ] for /f/, and the vowel [ɯ] for /u/ are not included in the word dictionary file and so the new alignment module cannot detect these sounds when speakers produce them. In addition, some other allophones by assimilation processes cannot yet be detected, such as dental [n̪], [t̪] and [d̪], nasally released [t^N] and [d^N] before a homorganic nasal, and laterally released [t^L] and [d^L] before an /l/. These issues are important in EFL because EFL teachers are likely to seek this kind of information when they use the corpus.

Another issue that still needs to be addressed is that there is a large *native* English speaking population in Asia and the phonetics and phonology of their English accents may be very different from those of traditional English speaking countries, such as USA and UK. Therefore it is important to develop the system to cater for data of L2 English and varied native English accents.

Conclusions

The original HTK alignment module is used as a common research tool. It is free and widely available but it is not necessarily suitable for L2 English speech which can be very different from native English speakers' English utterances.

The proposed new alignment method was able to detect pronunciation variations of Japanese speakers' English and consequently was able to perform more accurate automatic alignment. Based on these results it should be possible to use similar methods to develop other language specific alignment systems for data of other Asian language speakers such as Chinese, Thai or Indonesian speakers' English.

Quantitative evaluation of phoneme recognition and time alignment is still necessary to fully evaluate the effectiveness of the system. In addition the method needs to be tested with more data for both Japanese and other language speakers' English. However, the study shows the potential for the HTK with a modified TIMIT speech corpus to be used for automatic L2 alignment.

Acknowledgement

Part of this research was supported by JSPS Grants-in-Aid for Scientific Research, (C) No.22520585, and Sophia University Open Research Center. We would like to express our gratitude to Prof. Yoshinori Sagisaka of GITI, Waseda University for his support with this research.

References

[1] IPA (1999), *Handbook of the International Phonetic Association: a guide to the use of the International Phonetic Alphabet*. Cambridge: Cambridge University Press.
[2] Kondo, M. (2009), 'Is Acquisition of L2 Phonemes Difficult? Production of English Stress by Japanese Speakers'. In *Proc of the 10th Generative Approaches to Second Language Acquisition Conference,* Bowles, M. Ionin, T., Montrul, S., and Tremblay, A. (Eds.), 105-112. Somerville, MA: Cascadilla Proceedings Project.
[3] Meng, H., Tseng, C., Kondo, M., Harrison, A. and Viscelgia, T., (2009), 'Studying L2 Suprasegmental Features in Asian Englishes: A Position Paper', *The Proc. of 2009 INTERSPEECH*, 1715-1718, Brighton, UK.
[4] Nariai, T. & Tanaka, K. (2008), 'A study of pitch patterns of Japanese English analyzed via comparative linguistic features of Japanese and English', 776-779, Proc of Interspeech 2008 Brisbane, Australia.
[5] Piat, M. D. Fohr, I. Illina. (2008), "Foreign accent identification based on prosodic parameters", 759-762, Proc. of Interspeech 2008. Brisbane, Australia.
[6] Shinohara, S. (2004), 'Emergence of universal grammar in foreign word adaptations', *Constraints in Phonological Acquisition*, Kager, R., Pater, J. and Zonneveld, AW. (Eds.), 292-320, CUP.
[7] Tsubota, Y., Kawahara, T., Dantsuji, M. (2002), 'Recognition and verification of English by Japanese students for computer-assisted language learning system', In ICSLP-2002, 1205-1208. Denver, Colorado, USA
[8] Viscelgia, T, Tseng, C, Kondo, M, Meng, H, and Sagisaka, Y, (2009), 'Phonetic Aspects of Content Design in AESOP (Asian English Speech cOrpus Project)', *2009 Oriental COCOSDA,* Beijing, China.
[9] Wells, J.C. (1982), Accents of English, Vol. 1-3, CUP.
[10] http://htk.eng.cam.ac.uk/
[11] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1