# Building and Analysis of Asian English Speech Corpus : Japanese Speakers' Phonemic Recognition of English Consonants

Kondo, Mariko
Tsubaki, Hajime
Konishi, Takayuki
Sagisaka, Yoshinori

(Citation)
Learner Corpus Studies in Asia and the World,2:103-114

(Issue Date)
2014-05-31

(Resource Type)
departmental bulletin paper

(Version)
Version of Record

(JaLCDOI)
https://doi.org/10.24546/81006693

(URL)
https://hdl.handle.net/20.500.14094/81006693

# Building and Analysis of Asian English Speech Corpus: Japanese Speakers' Phonemic Recognition of English Consonants

Mariko KONDO
Hajime TSUBAKI
Takayuki KONISHI
Yoshinori SAGISAKA
*Waseda University*

## Abstract

Asia has the largest English speaking population, so it is important to understand the characteristics of Asian English for speech technology research and for effective EFL teaching. In this paper we will explain the construction of the Asian English Speech cOrpus Project (AESOP), and discuss results from our analysis of consonantal variation in 94 Japanese speakers' read English speech of "the North Wind and the Sun".

Many segments (6,620) deviated from the model pronunciation (American English). Analyzing consonant variations confirmed the Japanese speakers' phonemic interpretation of English consonants; i.e. they produced more variants for consonants which are not phonemes in Japanese. For example, they substituted /r/-type consonants for /l/ more than /l/-type consonants for /r/, and also commonly substituted [b] for /v/, but did not substitute [v] for /b/. Therefore, these results confirm that Japanese speakers use Japanese phonemes to interpret English sounds.

## Keywords

Asian Englishes, L2 English, Phonological acquisition, Segmental variation

## I Introduction

English has the largest speaking population in the world. It is used as a lingua franca and there are more non-native (L2) speakers of English than first language (L1) speakers. McArthur (2001) reported that there were 375 million English speakers in traditional English speaking countries (the Inner Circle countries), another 375 million English speakers in countries where English is used as a second language (the

Outer Circle countries), and a further 750-1,000 million English speakers in countries where English is a foreign language and the majority of the population are not native English speakers (the Expanding Circle countries) (Karchu 1985). However, Bolton (2004) has estimated much higher numbers of Outer Circle and Expanding Circle country speakers. He estimates that in Asia alone, there are over 600 million speakers, including over 300 million in India and over 200 million in China. These figures indicate that Asia has the largest English speaking population, many of whom are L2 English speakers, more than the Inner Circle countries (Crystal 2003). The number is increasing because many Expanding Circle countries, in Asia and other regions, have been adopting English as a business-language or a language of higher education. Therefore, the majority of English speakers are non-native speakers who do not speak like English speakers in the Inner Circle, and so there are more opportunities to speak to non-native English speakers and listen to their non-native English. Therefore, it is important to study characteristics of non-native English to enhance communication.

The large number of English language learners in Asia and Asian countries are important markets for English teaching and learning, and their materials. These non-native English speakers use speech technology such as telephone booking systems and are more likely to be users of automatic translators, English learning software and CALL systems than L1 English speakers. Therefore, it is essential to understand Asian language speakers' English, such as their sound systems, phoneme inventories, and phonetic and phonological features, not only for linguistic studies but also for the practical need of developing technology.

In order to understand important linguistic features of Asian language speakers' English, a project called *Asian English Speech cOrpus Project (AESOP)* was launched in 2008. The main objective of the AESOP project is to build an L2 English language speech corpus of Asian language speakers to advance research to improve understanding of Asian accented English. Such a corpus will help research in phonetics and phonology, speech science and technology, second language teaching and learning, and other linguistic research areas. Despite the big diversity of English spoken in the world, Asian Englishes have not been considered as varieties of English. They are considered as foreign accents and have been treated as mistakes or wrong pronunciations, and so have not received much focus of attention as pronunciation variants.

By understanding their systems and acquisition processes, we can apply this knowledge to develop appropriate English teaching materials, CALL systems and speech technology. Another objective of the project is to study the linguistic and socio-cultural diversity of Asian Englishes to help establish a standard to correctly assess English language levels of Asian English speakers.

The AESOP corpus consists of core English speech data collected using the same tasks at institutions in Japan, Taiwan, Hong Kong, China, India, Indonesia, Myanmar,

Mongolia, Nepal, Thailand, and Vietnam. In addition to the core data, all the institutions can add their own further data. Anyone can access data collected by any of the institutions once they have collected and provided speech data of a minimum 20 speakers.

The following sections describe the design of the data collection tasks and automatic data annotation method, and present some results on segmental analysis of Japanese speakers' AESOP data.

## II Design of the AESOP Corpus

A common recording platform was created and data collection procedures were developed to highlight phonetic and phonological characteristics of the languages spoken in the participating countries. Linguistic features are a typological issue rather than a geographical issue, and so a geographical region such as "Asia" is not necessarily important. However, there are many phonological characteristics of L2 English which are shared by speakers of different Asian languages. So common key phonological parameters were selected and featured in the eight tasks to collect reading (Tasks 1-6 and 7) and semi-spontaneous (Task 8) speeches.

Second language difficulty occurs at both the segmental and suprasegmental levels. Studies have shown that segmental accuracy is important for correct word recognition, but they have also shown that suprasegmental features are more important for the intelligibility of speech and that native speakers are more sensitive to suprasegmental accuracy than segmental accuracy in terms of speaking proficiency (Anderson-Hsieh et al. 1992, Gut 2003). Suprasegmental features are also important because they convey crucial linguistic information for communication. For example, suprasegmental features carry (i) syntactic information such as structural ambiguity, grammatical emphasis on words, phrase boundaries, and old-new information, (ii) semantic information such as lexical meaning, lexical ambiguity, and pragmatic information, and (iii) paralinguistic information such as feeling and emotion; all of these are very important in communication. However, suprasegmental features in L2 speech have not been studied as much as segmental features (Jilka, 2007). Therefore, the AESOP corpus is intended to study suprasegmental features as well as segmental features. The test words were selected from the CMU dictionary database ("CMU Dictionary Database", 2011) by taking into consideration segmental and prosodic aspects as well as word familiarity, and the test words are used repeatedly in different tasks.

The tasks have been designed to highlight phonological features of English which can be commonly found in Asian language speakers.

Task 1 features target words, chosen from the CMU dictionary database, in carrier sentences. The test words were chosen based on (i) the number of syllables: from 2 to 4 syllables, (ii) stress position and patterns, (iii) lexical familiarity and word frequency,

(iv) semantic versatility, so that the same test words could be used in different tasks, (v) segmental variation in pronunciation, (vi) vowel quality, (vii) lexical stress such as location and acoustic manifestation, and (viii) timing control in smaller phonological units like a segment, syllable and word.

Task 2 examines the prosodic realization of the target words at boundaries. The target words are the same ones used in Task 1, but are presented in phrases in order to observe how speakers demonstrate speech acts using intonation: (i) declarative statements, (ii) yes – no questions, (iii) WH questions, and (iv) two-clause sentences.

Task 3 uses the same target words in narrow-focus sentences to test if speakers use phrase accent to highlight important information: (i) indicating new information and (ii) contrasting words or information.

Task 4 investigates vowel reduction in stressed and unstressed syllables by using function words, such as prepositions and auxiliaries. It also observes differences in vowel quality of these function words with or without a following syntactic boundary. Many Asian languages have a large set of vowels, but an important characteristic of vowels in the English spoken in most Asian countries is a lack of vowel reduction in relation to stress (Chen and Robb 2000, Chen 2006, Zhang et al. 2008). Vowel reduction has important linguistic functions in English; it provides significant clues in communication including the part-of-speech of a word, syntactic meaning and syntactic structure. Therefore, it is important to determine how speakers change vowel quality in relation to stress and grammatical functions.

Task 5 tests if speakers use intonation to signal syntactic structure of syntactically ambiguous sentences.

Task 6 is a text reading task in which speakers read the Aesop fable "*The North Wind and the Sun*". This fable is widely used in phonetic research of English because it contains all English phonemes. So it is a good text to examine segmental pronunciation.

Task 7 is a role play in which the speaker plays the role of a reservation agent at an airline company and helps a customer to reserve a flight ticket over the phone. This is not a free conversation task and the dialogue is controlled. The speaker hears the customer's voice through the headset and can also read what the customer says on the computer monitor. Fixing the dialogue makes it easier to compare and assess the segmental and suprasegmental characteristics.

Task 8 is a picture description task designed to accumulate semi-spontaneous speech. The speaker sees a picture of a man holding a shopping list in front of shelves with various goods in a supermarket. The speaker is asked to answer various questions based on the picture. The speaker hears the questions through headphones and can also see the questions on the monitor at the same time. There are no fixed answers.

The full lists of test words and phrases, and a theoretical explanation of the construction of test words and phrases are discussed in Meng et al. (2009), Visceglia et

al. (2009) and Kondo (2012).

All prompts are provided on a computer monitor and speech data is recorded directly on a computer with a Sennheiser PC 166 Headset. The recordings are done in an anechoic chamber if available or in a quiet room if it is not available.

## III Annotation

The total amount of data collected by all the institutions will be very large, and it is difficult and time consuming to label L2 speech sounds with varied L1 accents. Therefore, we have been testing an automatic annotation method to label native Japanese speakers' data of "*the North Wind and the Sun*". We chose the data of *the North Wind and the Sun* because it contains all English phonemes and therefore it is ideal data to examine the accuracy of the auto-annotation.

We have been testing the Hidden Markov Model Toolkit (HTK) ("The Hidden Markov Model Toolkit", 2011) with associated pronunciation word dictionary files based on the TIMIT speech corpus ("TIMIT Acoustic-Phonetic Continuous Speech", 2011) written ARPABET symbols (Keating et al. 1994, The CMU Pronouncing Dictionary, 2011). The HTK is a set of modules for building and manipulating hidden Markov models (HMM), and is mainly used in speech recognition research. 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 contains a large data-set of English sentence speech read by native American English speakers.

First we trained an HMM-based acoustic model using HTK and the TIMIT speech corpus. The phonemes are registered in a word dictionary file including pairs of letter strings and phonemes for each word. However, since the TIMIT speech corpus only contains read speech data of native American English speakers it cannot annotate Japanese accented English perfectly because Japanese speakers produce different phonemes and sequences. Therefore, a new dictionary was created which contains phonemes and sequence of phonemes reflecting Japanese accented English, and this was added to the modules. As a result, the modified annotation system has two dictionary files (Figure 1): dictionary file A is based on the TIMIT speech corpus with native English speakers' pronunciation, and dictionary file B contains both the original ARPABET transcriptions from the TIMIT corpus and also added phonetic transcriptions that reflect Japanese accented English. For example, word dictionary file A contains a transcription of the English word *blew* as "b l uw" (ARPABET transcription) for /blu:/ (IPA transcription), which is based on the model pronunciation by native English speakers. This is a typical transcription of the pronunciation of the word *blew* produced by native English speakers, i.e. [blu:] (ARPABET "b l uw"). However, Japanese speakers produce more varieties of pronunciation: e.g. "b uh l uw" (/bulu:/) by inserting a vowel /u/ between the /b/ and /l/, "b r uw" (/bru:/) by substituting

[ɹ] for the /l/, "b uh r uw" (/bʊruː/) by both a substitution of /r/ for /l/ and an insertion of the vowel /u/, as well as the original TIMIT-based model pronunciation "b l uw" /bluː/. Word dictionary file B contains all of these transcriptions, including the original TIMIT transcription. The two dictionaries have enabled the HTK modules to perform auto-annotation of Japanese speakers' English with greater accuracy. However, there were some parts which were not segmented accurately, so the automatic annotation was checked and corrected manually for more accurate segmentation. One of the most difficult aspects in manual annotation of L2 English speech is the annotation of vowels. For example, Japanese speakers produce many vowel variants which are difficult to label with standard English phonemes. However, the auto-annotation method enables the labelling of Japanese accented vowels with fixed criteria, and therefore it is more reliable than human labelling. The details of the automatic annotation method are described in Tsubaki and Kondo (2011).

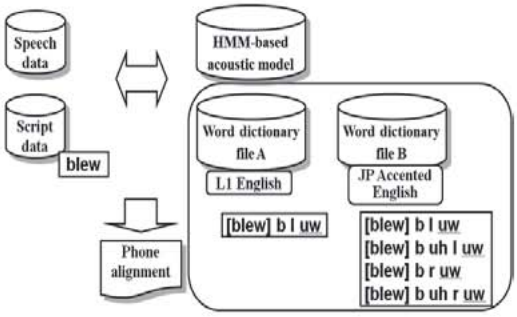


Fig. 1 Two word dictionaries A and B for automatic alignment testing

## IV Analysis of Japanese AESOP "The North Wind and the Sun"

Automatic annotation was performed on 94 Japanese speakers' read speech data of *the North Wind and the Sun*. Their pronunciation was examined with reference to General American pronunciation which is the model pronunciation used for the TIMIT database. Segmental variation of their pronunciation was analyzed in relation to speaker fluency levels. The speakers' English fluency levels were evaluated on a 9-point scale (1 to 5 in 0.5 increments) by eight English language teachers (4 native English, and 4 native Japanese speakers with very high English proficiency) who are currently teaching or have taught English at Japanese universities. The fluency level was judged by overall auditory impression of segmental and suprasegmental fluency and accuracy: 1 = very poor (almost incomprehensible) ~ 2 = poor (difficult but just be able to follow) ~ 3 = average ~ 4 = good (fluent with some segmental or prosodic mistakes) ~ 5 = native like. Most speakers' levels were distributed between 2 to 4: the average score was 2.98

and the median was 2.97 (Figure 2).

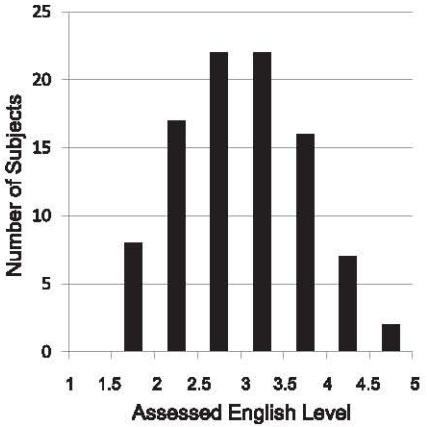


Fig. 2 English fluency levels of 94 Japanese speakers assessed by eight English teachers (4 native English speakers and 4 Japanese speakers)

Figure 3 shows the number of segmental variants identified in each speaker, plotted in relation to their assessed English fluency level; there were a total of 6,620 variants detected in the 94 speakers. Since General American accent was chosen as the reference pronunciation, any non-American phonemes were judged as variants, i.e. different from the model pronunciation, even if they are not necessarily wrong. Yet, there were still more variants found in less fluent speakers, with a moderate negative correlation between the number of variants and fluency level. The Pearson's correlation coefficient (R) between English fluency and the number of pronunciation changes, including epenthesized and deleted sounds, was -0.631 ($N$=94, $p < .001$; Figure 3). As shown in Figure 4, there were twice as many changes identified with vowels as with consonants: 4,478 examples compared with 2,142 examples, respectively. Annotation of vowels is not easy. It is also the case that vowel quality varies depending on regional accent and so typical vowel phoneme in non-American accents may have been identified as variants rather than as correct phonemes. Therefore, we have focused our analysis on the consonants.

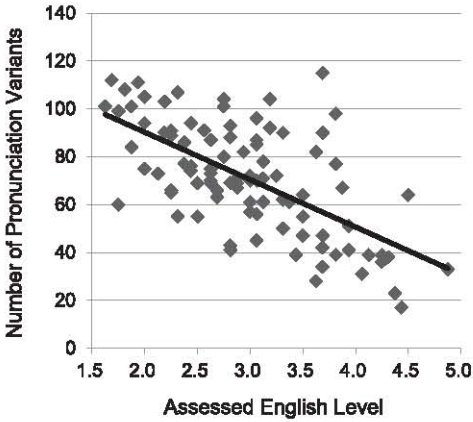


Fig. 3 Number of pronunciation changes by individual speakers in relation to fluency level. The line shows the best fit correlation coefficient. ($R^2$ = 0.398, y = -19.807x+129.87)

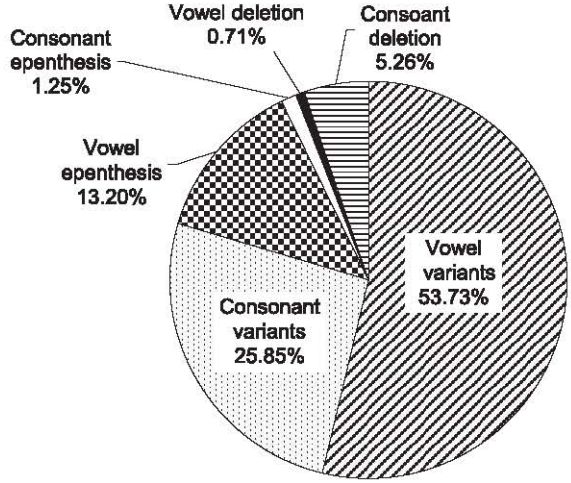


Fig. 4 Pronunciation changes in 94 Japanese speakers' English (Total = 6,620)

The consonantal variants presented in Figure 5 reflect typical problems discussed in EFL phonology of L1 Japanese speakers. Variants were produced for the phoneme alternations or the substitution of /l/ (418 examples), /ð/ (267 examples), /θ/ (191 examples), /r/ (124 examples), /h/ (50 examples), /z/ (41 examples), /v/ (39 examples), /f/

(20 examples) and /dʒ/ (12 examples).

All these consonants are phonemes in English, but Japanese does not have the /l/, /ð/ and /θ/ phonemes and so they are confused with /r/, /z/ and /s/ respectively. In our study, the typical variant of /l/ was /r/ ([ɹ], [ɾ], [ɻ]), and the variant of /ð/ was [dz] which is the most common realization of /z/ in Japanese; /θ/ was mostly substituted by [s]. Japanese speakers substitute Japanese phonemes for English phonemes which do not exist in Japanese.

The alternations of [f] for /h/ and [h] for /f/ were common. Japanese /h/ has phonotactic constraints: the [h] occurs only before /e/, /a/ and /o/. Also, Japanese /h/ has the allophone [ç] before /i/ and the allophone [ɸ], which is commonly substituted for /f/, before /u/. In other words, [h], [ç] and [ɸ] are in complementary distribution in Japanese, but both /h/ and /f/ occur before all vowels in English. This affects the recognition of the /h/ and /f/ in relation to the following vowel. Phonotactic constraints also affected /w/ realization. The consonant /w/ occurs in Japanese and the pronunciation of /w/ is not particularly difficult for Japanese speakers. However, in modern Japanese the /w/ only occurs only before /a/ and so the sequences of /w/ with other vowels are not easy to differentiate from /u+V/ sequences.

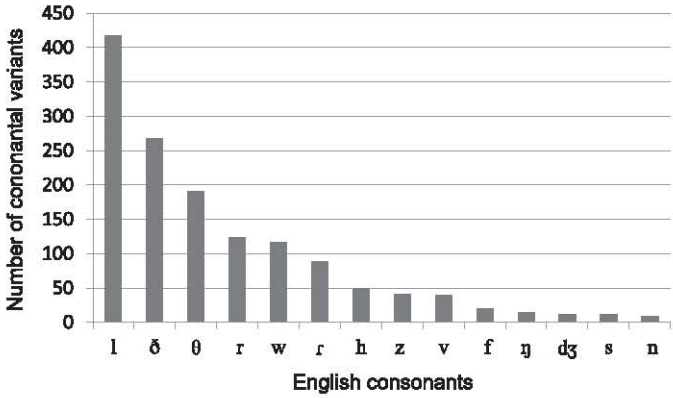


Fig. 5 Consonantal variants in Japanese speakers' English speech (Total = 1,404).
Note: Since the model pronunciation is the General American accent, the alveolar tap [ɾ] occurs as an allophone of intervocalic /t/ and /d/ in the TIMIT corpus, and other realizations of intervocalic /t/ and /d/ are categorized as variants.

In Japanese /l/ - /r/ variants and /b/ - /v/ variants are quite common. In our study we found three times as many variants of /l/ than of /r/. There were more than three times as many variants of /l/ than of /r/. Most /l/-variants were /r/-like sounds, whereas the most common variant of /r/ was the alveolar tap [ɾ] which is a common allophone of

Japanese /r/. The alveolar tap [ɾ] is also an allophone of English /r/ in some accents, and therefore it is not an incorrect pronunciation. With regard to /b/ - /v/ variants, all the variants of /v/ were [b], but there were no examples of [v] substituting for /b/.

These results confirm the typical characteristics of Japanese speakers' English pronunciation. In addition, the data also indicates that Japanese speakers are likely to substitute /r/ for /l/ and /b/ for /v/, but not /l/ for /r/ or /v/ for /b/. Therefore, these data suggest that pronunciation training for Japanese speakers should focus on the pronunciation of /l/ and /v/ rather than on /r/ and /b/.

## VI Conclusions

The AESOP corpus showed that many pronunciation variants occurred for consonants such as /l/, /ð/, /v/, and /θ/ that are not phonemes in Japanese. Also, Japanese phonotactics played an important role in the production of consonantal variants. Our subjects were confused by pairs of English consonants, which are separate phonemes in English but in Japanese they are complementary distributed allophones of a single phoneme, e.g. /h/ and /f/.

Phonetic realization of Japanese phonemes also affects the production of English sounds. For example, the most common variant of English /z/ by Japanese speakers was the affricate [dz] instead of the fricative [z], indicating that Japanese speakers do not differentiate between the fricative [z] and the affricate [dz].

While Japanese speakers had difficulty in producing /l/ and /v/, they had much less difficulty producing /r/ and no difficulty producing /b/. This suggests that pronunciation teaching should pay more attention to /l/ rather than the pronunciation of /r/, and to the pronunciation of /v/ rather than /b/.

Finally, we showed that automatic annotation using HTK with a modified TIMIT dictionary can be used for segmentation of the read speech of L2 English corpus, but it is still advisable to manually check annotation of segmental boundaries. We need to investigate further to improve the accuracy of annotation.

## Acknowledgement

AESOP is an international project and the data collecting materials and platform were developed jointly by teams from Waseda University, Academia Sinica in Taiwan and the Chinese University Hong Kong. Especially we would like to thank Professor Chiu-yu Tseng (Academia Sinica), Mr. Chi-Feng Huang (Academia Sinica), Professor Helen Meng (CUHK), Dr. Wai-Kit Lo (CUHK), Professor Tanya Visceglia (National Yang Ming University, Taiwan).

## References


Anderson-Hsieh, J., Johnson, R. & Koehler, K. (1992). The relationship between native speaker judgments of non-native pronunciation and deviance in segmentals, prosody, and syllable structure, *Language Learning, 42*(4), 529-555.

ARPABET: http://www.speech.cs.cmu.edu/cgi-bin/cmudict

Bolton, K. (2004). World Englishes. In A. Davies & C. Elder (Eds.), *Handbook of applied linguistics* (pp.367-396). Oxford, UK: Blackwell.

Chen, Y. (2006). Production of tense-lax contrast by Mandarin speakers of English. *Folia Phoniatrica Logopaedica, 58*, 240-249.

Chen, Y. & Robb, M. (2000). Acoustic features of vowel production in Mandarin speakers of English, *Proceedings of the 6th ICSLP*. 2, Beijing, China, 16-20 October 2000, 587-590.

The CMU Pronouncing Dictionary: http://www.speech.cs.cmu.edu/cgi-bin/cmudict

Crystal, D. (2003). *English as a global language*. Cambridge, UK: Cambridge University Press.

Gut, U. (2003). Non-native speech rhythm in German. *Proceedings of 15th ICPhS, Barcelona, Spain, 3-9 August 2003*, 2437-2440.

Jilka, M. (2007). Different manifestations and perceptions of foreign accent in intonation. In J. Trouvain & U. Gut (Eds.), *Non-native prosody: Phonetic description and teaching practice* (pp. 77-96). Berlin, Germany: Mouton de Gruyter.

Kachru, B. B. (1985). Standards, codification and sociolinguistic realism: The English language in the outer circle. In R. Quirk, & H.G. Widdowson (Eds.), *English in the world: Teaching and learning the language and literatures* (pp. 11-30). Cambridge, UK: Cambridge University Press.

Keating, P., Byrd, D., Flemming, E., & Todaka, Y. (1994). Phonetic analysis of word and segment variation using the TIMIT corpus of American English. *Speech Communication, 14*, 131-142.

Kondo, M. (2012). Design and analysis of Asian English speech corpus - How to elicit L1 phonology in L2 English data-. In Y. Tono, Y., Kawaguchi, Y. & Minegishi, M. (Eds.). Tokyo University of Foreign Studies Studies in Linguistics Vol. IV: *Developmental and crosslinguistic perspectives in learner corpus research* (pp. 251-278). Amsterdam/ Philadelphia, The Netherlands/USA: John Benjamins.

McArthur, T. (2001). World English and world Englishes: Trends, tensions, varieties, and standards. *Language Teaching*, January, 1–20.

Meng, H., Tseng, C., Kondo, M., Harrison, A. & Visceglia, T. (2009). Studying L2 suprasegmental features in Asian Englishes: A position paper. *Proceedings of 2009 INTERSPEECH*, 1715-1718.

The Hidden Markv Model Toolkit (HTK): http://htk.eng.cam.ac.uk/

TIMIT Acoustic-Phonetic Continuous Speech: http://catalog.ldc.upenn.edu/LDC93S1W

Tsubaki, H. & Kondo, M. (2011). Analysis of L2 English speech corpus by automatic phoneme alignment. *Proceedings* of *SLaTE 2011,* Venezia, Italy.

Visceglia, T, Tseng, C, Kondo, M, Meng, H, & Sagisaka, Y. (2009). Phonetic aspects of content design in AESOP (Asian English Speech cOrpus Project). *2009 Oriental COCOSDA*, Beijing, China, 10-12 August, 2009, 52-57.

Zhang, Y., Nissen, S. & Francis, A. (2008). Acoustic characteristics of English lexical stress produced by native Mandarin speakers. *Journal of the Acoustical Society of America*, Vol. 123, Issue 6, 4498-4513.