AESOP (Asian English Speech Corpus Project) and TWNAESOP

Chiu-yu Tseng
Institute of Linguistics
Academia Sinica, Taipei, Taiwan
cytling@gate.sinica.edu.tw

Tanya Visceglia
Department of Applied English
Ming Chuan University
orlandotaipei@hotmail.com

Abstract

AESOP (Asian English Speech Corpus Project) is a multi-national research project, which aims to collect and compare L2 English speech corpora from as many Asian countries as possible, in order to derive a set of core properties common to all varieties of Asian English, as well as to discover features that are particular to individual L2 dialects. Collaborators include linguists, speech scientists, psychologists and educators from Japan, Taiwan, Hong Kong, China, Thailand, India, Indonesia, Korea, the Philippines, Vietnam and Mongolia. This project is primarily motivated by the need for advanced speech technology development to improve computer-assisted language learning (CALL) applications, as well as to enhance the performance of Internet and mobile interface implementations catering to Asian L2 speaker populations, which have grown to outnumber ENL English speakers. The goal of AESOP is to build up an open-resource Asian L2 English speech corpus consortium, in which collected corpora will be open resource, available to the research community at large. Each regional research team will use the same recording setup, platform and core data design, but each team is also free to design supplementary materials and collect additional data to address language-specific features. A common, open-ended annotation system will also be developed. AESOP welcomes new collaborators, and it is hoped that in the future, the collected speech corpora will represent all varieties of L2 English spoken in Asia. In this talk, we address our design for core data collection, focusing on spoken-language tasks designed to elicit production of a comprehensive range of English suprasegmental characteristics. We will discuss both read and spontaneous speech tasks, which support analysis at the lexical, phrasal and discourse levels. Tasks which support systematic analysis of L2 suprasegmental features have been structured using a set of textual environments which have been carefully selected to illustrate how prosodic features can convey linguistic information at many levels, including lexical stress, syntax, focus, illocutionary force and information structure. The TWNAESOP research team was formed to collect and contribute Taiwan L2 English to the AESOP consortium and to develop a systematic understanding of L2 Taiwan English. We expect that analyses of such corpora will yield research findings with important implications for both language education and speech technology.

1. Introduction

The blending of English with local languages and dialects in countries and regions such as Greater China, India, Malaysia and the Philippines has given rise to a wide variety of world Englishes, which exhibit rich variation in pronunciation, lexicon and grammar. English is also being studied and spoken as a second language in more countries than ever before. Thus, a comprehensive understanding of the variation present in the dialects of English spoken in the world today is a fundamental issue for the development of English language education as well as speech science and technology. Asia is home to the largest number of English learners and speakers in the world; it has been claimed that combining native and non-native speakers, India now has more people who speak or understand English than any other country in the world (Crystal, 2004). Following India is the People's Republic of China (Zhao and Campbell, 1995). Thus, research in Asian English dialects from a multidisciplinary perspective is urgently needed to address issues in communication, learning and technology.

Research on the influence of a speaker's native language phonological system on the development of second-language phonology has primarily focused on the speaker's ability to perceive and produce segmental (single-sound) contrasts (for review see Flege, 1995). However, accent-rating studies have found that prosody (the intonation and rhythm of speech) also makes a significant contribution to the perception of a non-native accent (Anderson-Hsieh, J. Johnson, R. & Koehler, K., 1992; Munro, M., 1995; Tajima, K., Port, R. & Dalby, J., 1997). In addition, research demonstrates that suprasegmental features play a significant role in shaping second-language production. Jian (2004) found that Taiwan English is influenced by the rhythm of Taiwan Mandarin; thus native Mandarin speakers are significantly less likely than L1 English speakers to reduce vowels in English unstressed syllables. Influence of L1 rhythm is common among other varieties of Asian English, such as Japanese English (Kondo, Y., Kitagawa A. & Nakano, M. 2008). In addition, F0 analysis of Taiwan English found that non-native speakers do not differentiate discourse positions and/or illocutionary force using utterance-initial global pitch setting in the way that native speakers do (Visceglia & Fodor, 2006). This study also found that L1 Mandarin speakers tend to confine their English illocutionary prosody, such as question rises and statement falls, to the utterance-final syllable, whereas L1 English speakers usually anchor their rise or fall to the last pitch accent in an utterance.

Comparisons of native and non-native discourse-level prosody in English have found that non-native speakers demonstrate sporadic use of prosodic markers related to discourse structure (Wennerstrom, 1998). These markers include a continuation rise at phrase boundaries to link related constituents and use of paratone (an expansion of pitch range to signal topic shift). It has also been observed that non-native speakers produce a significantly narrower pitch range than native speakers do (Mennen, 1998; Pickering, 2004). Quantitative

analyses of Japanese English found that Japanese English was slower in speaking rate and shorter in sentence length than L1 English is (Kondo et al, 2008). The phonological characteristics of L2 English have been found to exhibit many levels of acoustic variation; thus, materials for the current research are designed to investigate the acoustic characteristics of L2 Asian English at the word, phrase, sentence and discourse levels.

2. Corpus Design

The data design presented in the following section addresses all of the above-mentioned issues, and specifically targets elicitation of a wide range of suprasegmental characteristics, particularly those of communicative and spontaneous speech, which have both been underrepresented in the literature. A core data design for the AESOP community has been developed based on documented L2 Asian English suprasegmental speech characteristics, and TWNAESOP is in the process of collecting data to determine whether these characteristics can be found in Taiwan Mandarin. We predict that the following features are likely to occur in L2 speech: (1) tonal borrowing at the lexical level; (2) L1-specific differences in rhythm and timing; (3) differences in realization of phrase boundary phenomena such as declarative falls and interrogative rises and (4) differences in the shape, timing and location of pitch accents, which are used to create broad and narrow focus in English and (5) differences in production of discourse-level prosody, such as pitch downstepping and resetting within and across information units.

The full set of experimental materials includes six read speech and three spontaneous speech tasks, all of which can be completed in approximately one hour of recording time. AESOP's compact, yet comprehensive set of core experimental tasks, its recording platform, and recording protocol manual (which includes guidelines for recording setup, hardware specifications, and a detailed set of recording instructions for both the proctor and the speaker) were developed in a collaborative effort by AESOP members in Taiwan, Japan and Hong Kong. Subsequently, they were made available online as an open resource for all AESOP members.

2.1 Target Words

We have developed a list of target word candidates for each possible stress type present in English 2-, 3- and 4-syllable words, which have been excerpted from the CMU Dictionary database (http://www.speech.cs.cmu.edu/cgi-bin/cmudict). Words of five syllables or more have been excluded to avoid the possible confounds of secondary and tertiary stress[1].

[1]Comprehensive lists of each token type have been made available to all collaborators.

Selection of target words from this candidate list was based on lexical familiarity (piloted), overall frequency and stress type (based on the analyses of the CMU dictionary) and semantic versatility (to facilitate construction of experimental sentences). We chose to include words representing a range of stress patterns and syllabicities, based on our prediction that the realization of lexical stress will differ between L1 and L2 English speakers. Our list of target words is categorized according to syllabicity and stress type; it includes tokens of 2-4 syllable words with initial, medial or final stress.

English is a stress-timed language, one consequence of which is that stressed syllables in individual words tend to be louder, higher in pitch and longer in duration than unstressed syllables are. Moreover, the vowels in English unstressed syllables are often reduced to schwa. Mandarin, in contrast, marks the distinction between stressed and unstressed syllables with reduction of syllable duration, rather than with vowel reduction and intensity decay (Lin, 1983; Cao, 1986). Japanese is a pitch accent language in which the presence of stress or accent does not affect vowel duration or quality. Thus, L1 Japanese and Mandarin speakers often have trouble realizing stress assignment on multi-syllabic words and use inappropriate cues to differentiate stressed and unstressed syllables (Kondo et al., 2008).

2.2 Task 1: Production of Target Words in Carrier Sentences

Each target word in the set was embedded in a fixed, neutral context for baseline comparisons of inherent duration and formant values with tokens of those words appearing in other experimental conditions. Speakers read a list of identical carrier sentences: “I said the word XX five times”. Each of those sentences contains one target word appearing in a broad-focused position two syllables removed from any phrase boundary.

2.3 Task 2: Production of Target Words at Phrase Boundaries

Previous research in L2 prosody suggests that L2 speakers realize prosodic phrase boundaries differently from L1 speakers (Visceglia, 2006; Wennerstrom, 1998). To further investigate this phenomenon, we embedded target words in four prosodic boundary positions: (1) the final fall of a wh-question, (2) the final rise of a yes-no question, (3) the continuation rise found in multiple-clause sentences, and (4) the final fall in declarative sentences. To realize phrasal and sentential prosodic boundaries, L1 English speakers usually anchor the nuclear (most prominent) pitch accent to the last prominent syllable in an intonation phrase, from which they begin their rise or fall to a phrase boundary. Our design elicits L2 productions of the acoustic features associated with phrase and sentence boundaries by placing each of the target words at a boundary position (Example: target word: overnight, boundary type: yes-no question. Experimental sentence: “Can packages be shipped overnight?”)

2.4 Task 3: Production of Target Words in Contrastive Stress Positions

Differences have also been found between L1 and L2 English speakers' production of the pitch accent used to mark narrow focus in English (McGory, 1997). In order to elicit these data from speakers, we have placed each of the target words in a narrow-focus contrastive context (Example: Target word: overnight Experimental sentence: We have to finish the project *overnight*, not over the weekend).

2.5 Task 4: Production of Stressed and Unstressed Function Words

English is a stress-timed language, another consequence of which is that function words, such as pronouns, prepositions and auxiliary verbs, generally carry a minimal semantic load and are therefore not acoustically prominent. L1 speakers often reduce the vowels in function words and may even delete them in spontaneous speech. We designed one set of sentences with same function words appearing in stressed and unstressed positions to investigate whether L2 speakers would appropriately reduce the unstressed tokens (Example: "I can [reduced kən] run faster than you can [canonical kæn].").

2.6 Task 5: Production of Prosodic Disambiguation

There is evidence to suggest that L1 English speakers use prosody to disambiguate different syntactic structures in identical phonetic strings (Price, Ostendorf, Shattuck-Hufnagel & Fong, 1991). The strongest use of prosodic cues for this purpose has been found in differentiation of early and late closure sentences such as the following: (1) When you learn // gradually you worry more. (2) When you learn gradually // you worry more. Our materials include a small set of syntactically ambiguous sentences for the purpose of investigating whether L2 speakers produce the prosodic cues used to differentiate clause boundary locations between such sentence types.

2.7 Task 6: Reading Passage "The North Wind"

Following the tasks described above, each speaker will read Aesop's fable "The North Wind" aloud. This passage is recommended by the IPA for the purpose of eliciting all phonemic contrasts in English. Another phonetically-balanced passage of the same length in the speaker's L1 will be included in this task in order to compare L1 and L2 global speech rate, pitch range and production of prosodic cues associated with information structure.

2.8 Task 7: Picture Description Task

This task presents participants with an illustration of a man standing at the entrance of a supermarket holding a shopping list, preparing to do his grocery shopping. Participants are required to study the illustration and respond to a series of questions, which guide them to describe different aspects of the scene. Each question is presented individually on a computer screen, and no time limit is imposed. Participants are permitted to continue looking at the picture while they answer questions. The purpose of this task is to elicit suprasegmental characteristics as they occur in spontaneous (unscripted) speech, including: lexical stress, phrase and utterance-level intonation contours used to mark continuation/finality or illocutionary force (e.g. question/statement), and the features associated with long-range prosodic planning of larger discourse units, such as pitch reset between topics and pitch downstepping within topics.

Words appearing on the man's shopping list have been deliberately chosen to represent a range of syllabicities and stress types in order to investigate L2 speakers' production of lexical stress, as well as the possibility of interaction between location of pitch accent and realization of phrase boundaries. Target words include: watermelon (4 syllables, initial stress); orange juice (left headed N-N compound); red wine (right headed Adj.-N compound); noodles (2 syllables, initial stress) and strawberries (3 syllables, initial stress). The questions participants answer following picture viewing were each designed to elicit particular prosodic features:

Question 1: "What does the man plan to buy?" This is designed to elicit continuation rise between the items on the shopping list and a final fall at the end of the utterance.
Sample Answer: "The man wants to buy watermelon, orange juice, red wine, noodles and strawberries".
Question 2: "At the supermarket, what will the man do first, second, third, fourth and last?" This is designed to elicit topic-initial pitch setting, pitch downstep within the intonation unit, and production of intermediate and final phrase boundaries.
Sample Answer: "First, he will go to Aisle 1 to get watermelon and strawberries, second he

will go to Aisle 2 to get red wine, third, he will go to Aisle 3 to get noodles, then he'll go to Aisle 4 to get orange juice, and finally he will go to the cashier to pay."

Question 3: "What do you think the man will do after he leaves the supermarket?" This is designed to elicit production of a paratone, i.e. pitch resetting associated with change in discourse topic.

Sample Answer: "After he leaves the supermarket, the man will go home and put his food away. Then, he will make dinner for himself and his family."

It should be noted here that speaker anxiety may still prevent production of more detailed responses. That is to say, participants may choose to utter the briefest possible response in order to decrease their chances of making a positive error. Therefore, we have designed an additional dialogue experiment, which more directly elicits production of longer, more detailed responses.

2.9 Task 8: Computer-Prompted Dialogue

Our computer-prompted dialogue task embeds suprasegmental features in an interactive discourse in order to elicit a range of sentence types and target words embedded in various discourse positions. Dialogue, unlike picture description, includes prosodic cues for turn-taking, prosodic marking of new and given information, and initiation of new topics. Moreover, picture description has the inherent limitation most responses being generated in the form of declarative sentences. The discourse requirements of the interactive dialogue task we have designed, in contrast, elicit a greater range of illocutionary prosody, including: wh-question, yes-no question; either/or question and imperative intonation.

Additional features have been built in to investigate whether L2 speakers are able to reduce/delete/link unstressed syllables/words in a target-like manner, as well as to investigate the possibility of tone borrowing on letters of the alphabet and numbers. This task also elicits prosodic features related to representation of information structure, such as pitch accents used to mark broad and narrow (prominent and contrastive) focus within sentences, pitch setting over longer units of discourse, prosodic marking of parenthetical information and production of intonation in post-focused positions.

At the beginning of this task, participants are presented with an audio and visual display of the following instructions: "You are a reservation agent for EVA Airlines. Help this customer reserve a flight from Taipei to New York." Participants will then receive a series of audio and visual prompts which move the transaction forward. In the course of this interaction, the participant, acting as a travel agent, is required to solicit information from the customer, confirm details such as dates, spelling of names and credit card numbers, and give the customer sequences of information and instructions.

2.10 Task 9: Elicitation of Letter and Number Strings

When L1 English speakers are asked to spell out names or other words in alphabetic letter strings, they use intonation as a grouping strategy. Similarly, when asked to produce number strings such as telephone and credit card numbers, they tend to use fixed prosodic configurations (Aylett, 2004). Alphabetic letter and number strings are important considerations of man-machine interface, yet little data have yet been reported on L2 speakers' production of such strings. Modelling these patterns are of primary importance to the development of speech technology, as most computer interfaces require speakers to spell their names and addresses, or to provide their phone, identification or credit card numbers. We have designed a series of prompts, which require speakers to spell the name and address of their sponsoring institution and to repeat a series of number strings appearing on a computer screen. This task is designed specifically to target L2 speakers' prosodic groupings of the English alphabetic letter and number strings that are solicited most frequently by man-machine interface systems.

3. Conclusion

The experimental tasks described above represent our core data set, designed to elicit a comprehensive inventory of English suprasegmental features in a concentrated and easily implemented set of materials. Speech database collection using this task set will provide specific information on the greatest number of phonetic features with the least amount of data collection effort. These experiments are included in the phonetic database design of AESOP, which will serve as a cross-linguistic core resource to increase our understanding of the ways in which L2 spontaneous speech differs from read speech, as well as the ways in which L2 Asian English differs from L1 English. These findings could also inform and help improve computer-assisted language learning and other ICT tool development tailored to the Asian English speaking population. Other research interests represented by the AESOP international collaboration project are open at this stage. TWNAESOP has collected data from over 330 speakers from eight academic institutions to date. We welcome feedback and participation from L2 researchers in all fields.

Acknowledgements

The AESOP consortium is initiated by Professor Yoshinori Sagisaka of Waseda University. Other collaborators include Professor Michiko Nakano of Waseda University, Dr. Chai Wutiwiwatchai of the National Electronics and Computer Technology Center in Thailand, Professor Sudaporn Luksaneeyanawin, Professor Tavicha Phadvibulaya and Professor Kulaporn Hiranburana of the Chulalongkorn University, Dr. Wai-Kit Lo, Dr. Pauline Lee and Alissa Harrison of The Chinese University of Hong Kong, Dr. Lan Wang of the CAS-CUHK Shenzhen Institute of Advanced Integration Technologies, and Dr. Sakriani Sakti and Dr. Dawa Idomuco from ATR.

References

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

Aylett, M. (2004). Merging data driven and rule based prosodic models for unit selection TTS. Proceedings of 5th ISCA Speech Synthesis Workshop, Pittsburgh, PA, USA.

Cao, J.S. (1986). 普通話輕聲音節特性分析，《應用聲學》，1986 年 4 期.

Crystal, D. "Subcontinent Raises Its Voice" Guardian Weekly: Friday 19 November 2004

Flege, J. E. (1995). "Second-language speech learning: Theory, findings and problems." In *Speech Perception and Linguistic Experience: Issues in Cross-language Research*. Strange, W. (Ed.), Timonium, MD: York Press, 233-277.

Jian, H. L. (2004). "On the syllable timing in Taiwan English" Proceedings of Speech Prosody 2004. Nara, Japan. International Speech Communication Association.

Kondo, Y., Kitagawa A., & Nakano, M. (2008). "Second language speech: Subjective evaluation and objective measures" Proceedings of 2nd International Workshop on Language and Speech Science 2008. Waseda University, Tokyo, Japan.

Lin, T. (1983) 探討北京話輕音性質的初步實驗《語言學論叢》第 10 輯，北京：商務印書館。

McGory, J. (1997). "The Acquisition of Intonation Patterns in English by Native Speakers of Korean and Mandarin." Unpublished doctoral dissertation, Ohio State Dissertations in Linguistics. Department of Linguistics: The Ohio State University

Mennen, I. (1998). "Can language learners ever acquire the intonation of a second language?" Proceedings of the ESCA workshop on speech technology in language learning (pp. 17–20). Marholmen, Sweden:International Speech Communication Association.

Munro, M. (1995). "Nonsegmental factors in foreign accent," SSLA (17), 17-34.

Pickering, L. (2004). "The structure and function of intonational paragraphs in native and nonnative speaker instructional discourse" *English for Specific Purposes* (23) 19-43.

Price, P., Ostendorf, M., Shattuck-Hufnagel, S., & Fong, C. (1991). The use of prosody in syntactic disambiguation. *Journal of the Acoustical Society of America, 90*, 2956-2970.

Tajima, K., Port, R., & Dalby, J. (1997). “Effects of temporal correction on intelligibility of foreign-accented English,” Journal of Phonetics (25) 1-24.

Visceglia, T., & Fodor, J. D. (2006). “Fundamental frequency in Mandarin and English: Comparing first- and second-language speakers”. In Interfaces in Multilingualism, Lleó, Conxita (ed.), 27–59.Wennerstrom, A. 1998. "Intonation as cohesion in academic discourse: A study of Chinese speakers of English" *Studies in Second Language Acquisition* (20), 1-25.

Zhao, Y., & Campbell, K. P. (1995). "English in China". World Englishes 14 (3): 377–390.