Spontaneous Speech Database Design for Suprasegmental Characteristics in Asian L2 English

Tanya Visceglia[1], Chiu-yu Tseng[2],
1. Department of Applied English Ming Chuan University, Taipei, 111
2. Institute of Linguistics, Academia Sinica, Taipei 115

Abstract

This research is part of the ongoing multinational collaboration "Asian English Speech cOrpus Project" (AESOP), whose aim is to build up a speech corpus representing the varieties of English spoken in Asia. The present paper describes tasks designed to elicit production of a comprehensive range of English segmental and suprasegmental characteristics in the form of spontaneous speech. Segmental and suprasegmental properties of spontaneous speech have been shown to differ significantly from those of read speech; nevertheless, much of the data used to develop man-machine speech communication systems has, for the most part, been based on read speech. Experiments designed to elicit a full range of L2 English segmental and prosodic features in spontaneous speech could efficiently collect an inventory of these features in a database, which could be used for further phonetic studies as well as modeling and ICT tool development tailored to the Asian L2 English-speaking population.

Key words: L2 English, Asia, speech corpus, spontaneous speech, computer-prompted dialogue, picture description, discourse prosody, suprasegmental features

1. Introduction

This research is part of the ongoing multinational collaboration "Asian English Speech cOrpus Project" (AESOP), whose aim is to build up a speech corpus representing the varieties of English spoken in Asia. AESOP is an international consortium of linguists, speech scientists, psychologists and educators from Japan, Taiwan, Hong Kong, China, Thailand, Indonesia and Mongolia. Its primary aim is to collect and compare Asian English speech corpora from the countries listed above 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 varieties. AESOP-collected corpora will be an open resource, available to the research community at large.

As English continues to grow in importance as a language for international communication throughout the world, the face of English itself is continuously changing. 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. Following India is the People's Republic of China [1][2]. Thus research in Asian English dialects from a multidisciplinary perspective is urgently needed to address issues in communication, learning and technology. Research in linguistics can catalogue and analyze the range of variation present in Asian English dialects; research in speech science can implement linguistic findings into the development of ICT tools and environments tailored to the requirements of Asian speaker populations.

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 3]. However, accent-rating studies have found that prosody (the intonation and rhythm of speech) also make a significant contribution to the perception of a non-native accent [4][5][6]. In addition, a substantial body of research exists to demonstrate that suprasegmental phenomena play a significant role in shaping second-language production [7]. It was found that the timing of Taiwan English is influenced by the syllable timing of Taiwan Mandarin; thus native Mandarin speakers are significantly less likely than L1 English speakers to reduce vowels in English unstressed syllables. This property is common among other varieties of Asian English that interact with syllable- or mora-timed languages, such as Thai, Hong Kong Cantonese, and Japanese English [8].

F0 analysis of Taiwan English found that non-native speakers do not perform utterance-initial global pitch setting in the way that native speakers do[9]. Moreover, non-native speakers confined the timing of their 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 found that non-native speakers demonstrate sporadic use of prosodic markers related to discourse structure [10]. These markers include high

pitch at phrase boundaries to link related constituents and paratone, which is 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 [11][12]. Quantitative analyses of Japanese English found that Japanese English is slower in speaking rate and shorter in sentence length than L1 English [8]. The phonological characteristics of Asian English have been found to exhibit phonetic variation at many levels; 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.

Recent studies have demonstrated that the segmental and suprasegmental properties of spontaneous speech differ significantly from those of read speech. At the segmental level, the spectral distribution of vowels in Japanese is significantly reduced in spontaneous speech [13]. At the prosodic level, Swedish data indicate a steeper F0 declination and stronger pitch resetting in read speech. In UK English, speakers tend to position stress differently and mark boundaries in different positions between read and spontaneous speech. Both English and Dutch spontaneous speech contain more pauses than read speech does; in Dutch, those pauses have a stronger tendency to be realized with pre-final lengthening. Furthermore, falling boundary tunes tend to occur more frequently in read speech. It is reasonable to predict similar differences between L2 read and spontaneous speech; however, very little data exist comparing the two speech styles in the L2 population [14]. Thus, it remains largely unknown whether read and spontaneous L2 speech exhibit similar properties, particularly with respect to suprasegmental features. Experiments designed to elicit a comprehensive inventory of suprasegmental features could efficiently collect a L2 spontaneous speech database, which could be used for acoustic phonetic research as well as for modeling and ICT tool development tailored to the L2 population. Our proposal for such a database content design will be described in the sections to follow.

2. Experiment 1--Picture Description Task

The Picture Description 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 (A reproduction of the picture appears in Appendix A). Participants are required to study the illustration, then respond to a series of questions, which guide them to describe different aspects of the scene. The purpose of this task is to elicit segmental and suprasegmental characteristics as they occur in spontaneous (unscripted) speech, including: lexical stress, phrase and utterance-level intonation contours used to mark continuation/finality as well as 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.

2.1 Procedure

Participants are asked to study an illustration of a man standing at the entrance of a supermarket with a shopping list in his hand. After participants have familiarized themselves with the content of the picture, they will then answer a series of questions. Each question will be presented individually on a computer screen, and no time limit will be imposed for answering the questions. Participants are permitted to continue looking at the picture while they answer questions.

2.2 Materials

In the picture, we can see the individual aisles of a supermarket, which are clearly labeled and have products in them: Aisle 1: fruit and vegetables; Aisle 2: beer and wine; Aisle 3: rice and noodles ; Aisle 4: juice and water; Cashier. Words appearing on the man's shopping list have been deliberately chosen to represent a range of phonemes, 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 plans 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, next he'll go to Aisle 4 to get orange juice, and last 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 a paratone, i.e. pitch resetting that is 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."

However, it should be noted here that speaker anxiety may still prevent production of more detailed responses. Therefore, we have designed an additional dialogue experiment, which more strongly elicits production of longer, more detailed responses by providing speakers with more content support in the form of prompts.

3. Experiment 2 -- Computer-Prompted Dialogue

The 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 of mostly generating responses in the form of declarative sentences. The discourse requirements of the interactive dialogue task we have designed, in contrast, will elicit a greater range of sentence types, 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 will also elicit prosodic features related to representation of information structure, such as pitch accents used to mark broad and narrow (nuclear and contrastive) focus within sentences, pitch setting over longer units of discourse, prosodic marking of parenthetical information and intonation on lexical items appearing in post-focused positions.

3.1 Procedure

Participants will be 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."
The participant 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 including times, dates, spelling of names and credit card numbers, and give instructions and information to the customer. A full transcription of this dialogue appears in Appendix B.

4. Experiment 3 -- 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 groupings. When asked to produce number strings such as telephone and credit card numbers, L1 speakers also use fixed prosodic configurations [15]. In other words, alphabetic letter and number strings are important considerations of man-machine interface, yet little data have yet been reported on L2 English speakers' production of these 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 questions, which require speakers to spell the name and address of their sponsoring institution and to repeat a series of number strings that will appear on a screen, in order to capture L2 speakers' prosodic groupings of English alphabetic letter and number strings in a variety of configurations.

5. Predictions

Based on previous research cited in Section 1 and on our pilot observations, we predict that collected data will demonstrate the following differences between L1 and L2 English prosody:

For Experiment 1, we predict the following:

(1) Lexical stress: In many syllable-timed and mora-timed Asian languages, the distinction between stressed and unstressed syllables is marked by reduction of syllable duration and intensity rather than by vowel reduction. Thus, L1 speakers of syllable-timed and mora-timed languages are predicted to have trouble with stress assignment in English multi-syllabic words and use inappropriate cues to differentiate stressed and unstressed syllables.

(2) Phrase and utterance-level intonation: In order to realize phrase or utterance 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 the end of an utterance. We predict that L2 speakers will confine their final rise or fall to the final syllable of a phrase or utterance.

(3) Pitch reset and downstepping within topics: We predict that L2 speakers will divide their discourse into smaller intonation phrases and produce more F0 resets and fewer levels of downstepping than L1 speakers would.

For Experiment 2, we predict the following:

(1) Questions, statements and imperatives: We predict that illocutionary intonation will be confined to utterance-final syllables.

(2) Reduction and linking of function/unstressed syllables or words: We predict that L2 speakers will not reduce or link function words in a target-like manner. Instead, they will produce these words with lexical stress and with underlying vowel quality.

(3) Broad and narrow focus, parenthetical information and post-focused information: We predict that L1 speakers of tone languages will use identical pitch patterns to mark nuclear and contrastive focus; whereas L1 English speakers will use different shapes of pitch accent to realize nuclear and contrastive stress. Moreover, L1 speakers of tone languages will continue to produce pitch accents on parenthetical and post-focused information, whereas L1 speakers never do.

For Experiment 3, we predict the following:

(1) Individual alphabetic letters and numbers: We predict that L1 tone language speakers will associate individual alphabetic letters and numbers with fixed pitch patterns, which are borrowed from their L1 tone inventory, whereas L1 English speakers will group them using phrase intonation.

(2) Alphabetic letter and number strings: We predict that pitch patterns of individual alphabetic letters and numbers will remain fixed irrespective of phrase position for L2 speakers, whereas L1 English speakers will use phrase-level prosody to configure letter and number strings.

6. Conclusion

The experiments described above represent our initial efforts to elicit a comprehensive inventory of the segmental and suprasegmental features of spontaneous speech in a concentrated and easily implementable set of materials. Spontaneous speech database collection using this type of task 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 also includes a series of read speech tasks [16]. This kind of database could 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 modeling and 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. We welcome feedback and participation from L2 researchers in all fields.

Acknowledgements

We extend our sincere thanks to Professors Helen Meng and Mariko Kondo for their insightful comments and suggestions, and to Leonie Reyneke, the artist who contributed the illustration used in the picture description task.

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.

Appendix A: Picture Description Illustration

Appendix B: Text of computer-prompted dialogue

Introduction: You are a reservation agent for EVA Airlines. Help this customer reserve a flight from Taipei to New York.
Customer: Hello. I'd like to reserve a ticket from Taipei to JFK airport in New York.
Prompt: Ask the customer: "When would you like to travel?"
When would you like to travel?
Customer: November twenty-second.
Prompt: Ask the customer: "Did you say the twenty-second or the twenty-seventh?"
Did you say the twenty-second, or the twenty-seventh?

Customer: The twenty-second.
Prompt: Ask the customer: "Would you like a window seat or an aisle seat?" Would you like a window seat or an aisle seat?
Customer: An aisle seat, please.
Prompt: Ask the customer: "Would you like a special dinner?"
Would you like a special dinner?
Customer: No, thank you.
Prompt: Ask the customer: "When would you like to reserve your returning flight?"
When would you like to reserve your returning flight?
Customer: I'm not sure when I'll be returning. I'll call from New York to reserve the date.
Prompt: Tell the customer:
1. Your flight, BR 317, will depart from CKS airport at 11:15 AM on November 22nd.
2. You will arrive at Narita Airport at 2:50 PM.
3. You will transfer to Flight 809 to New York JFK Airport, which departs at 7:08 PM from Gate 13F.
4. You will land at JFK airport at 4:30 PM on November 21st.
Your flight, BR 317 will depart from CKS airport on November 22nd at 11:15 AM. It will arrive at Narita Airport at 2:50 PM, where you will transfer to Flight 201 to JFK, which departs at 7:08 pm from Gate 13F. You will land at JFK airport at 4:30 PM on November 21st.
Customer: Did you say the flight from Narita to New York leaves from Gate 30 F?
Prompt: Tell the customer: "The flight leaves from Gate 13 F, not Gate 30 F.
That flight leaves from Gate 13 F, not Gate 30 F.
Customer: Oh, sorry. Got it. Gate 13 F.
Prompt: Ask the customer "May I have your name?"
Customer: My name is Lucy Hasegawa-Johnson L-U-C-Y H-A-S-E-G-A-W-A J-O-H-N-S-O-N.
Prompt: Repeat the spelling of the customer's name L-U-C-Y H-A-S-E-G-A-W-A J-O-H-N-S-O-N
Prompt: Ask the customer "May I have your credit card number?"
Customer: It's VISA number 5924-8013-6702-3516. Expiration date 09/ 2012
Prompt: Repeat the customer's credit card number: 5924-8013-6702-3516
Expiration date 09/ 2012
Customer: That's right.
Prompt: Ask the customer "May I have your billing address?"
Customer: 1425 Lakeshore Drive, Apartment 47B, Chicago, Illinois 60195
Prompt: Repeat the customer's address: 1425 Lakeshore Drive, Apartment 47B, Chicago, Illinois 60195
Customer: Yes.
Prompt: Ask the customer "May I have your contact phone number?"
Customer: 609-472-1358
Prompt: Repeat the customer's phone number: 609-472-1358
Customer: Yes.
Prompt: Ask the customer: "Is there anything else I can help you with this morning. Is there anything else I can help you with this morning?
Customer: No, thank you. That's all I need today.
Prompt: Say "Goodbye and thank you for calling EVA Airlines".
Goodbye, and thank you for calling EVA Airlines.

References

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

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

[3]. 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.

[4]. 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

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

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

[7]. Jian, H.L. 2004. "On the syllable timing in Taiwan English" In Proceedings of Speech Prosody 2004 Nara, Japan.International Speech Communication Association.

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

[9]. Visceglia, Tanya & Janet Dean Fodor 2006. "Fundamental frequency in Mandarin and English: Comparing first- and second-language speakers". In Interfaces in Multilingualism, Lleó, Conxita (ed.), 27–59.

[10]. Wennerstrom, A. 1998. "Intonation as cohesion in academic discourse: a study of Chinese speakers of English" *Studies in Second Language Acquisition* (20), 1-25.

[11]. Mennen, I. 1998. "Can language learners ever acquire the intonation of a second language?" In Proceedings of the ESCA workshop on

speech technology in language learning (pp. 17–20). Marholmen, Sweden:International Speech Communication Association.

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

[13]. Furui, S. Nakamura, M. Ichiba, T. A[nd Iwano K. “Why is the recognition of spontaneous speech so hard?” in Text, speech and dialogue : Václav Matoušek, Pavel Mautner, Tomáš Pavelka (eds.), Springer-Verlag Berlin Heidelberg 2005

[14]. Strik, H. Cucchiarini, C. and Binnenpoorte, D. 2000. "L2 pronunciation quality in read and spontaneous speech", In *ICSLP-2000*, vol.3, 582-585.

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

[16]. Visceglia, T. Tseng, C. Kondo, M. Meng, H. and Sagisaka, Y. “Phonetic Aspects of Content Design in AESOP (Asian English Speech cOrpus Project)” Oriental-COCOSDA 2009, Aug. 10-12, 2009, Urumuqi, Xinjiang Autonomous Region, China.