Pitch dynamism of English produced by proficient non-native speakers: preliminary results of a corpus-based analysis of second language speech

Juhani Toivanen
Department of Finnish, Saami and Logopedics, University of Oulu

Abstract
A speech database containing spoken English produced by proficient non-native speakers is described. The material was collected in connection with thesis seminars where the speakers, 20 Finnish university students of English, discussed the topics of their MA theses. For each pair, the speech material was digitally recorded with DAT in an anechoic studio to produce a 44.1-kHz, 16-bit CD-format recording. The analysis involves several hierarchically organized tiers. For each word, accentuation is described in terms of four degrees of stress. A nuclear tone approach is used to transcribe pitch direction and the pitch of accented syllables. Pauses with a minimum duration of 100 ms are annotated in terms of duration and type. Finally, for each intonation-unit, a categorization of voice quality and rhythm is chosen. In addition, the speech data is acoustically analyzed in terms of global parameters such as pitch dynamism and range. Some preliminary results are reported in this paper.

Introduction
There is now a growing interest in the phonetic aspects of second language (L2) speech (cf. Hincks 2003) but the existing research mainly focuses on the segmental features. Systematic studies of the prosodic features of L2 speech have been rare. In particular, the corpus-based approach has to date been virtually non-existent. This is understandable, given the well-known problems of investigating non-native speech data: theoretical problems will be inevitable if the researcher wishes to use only those phonological categories that can be comfortably used when describing first language (L1) prosody. The problem revolves around the issue of contrastive paradox or comparative fallacy: the researcher should not uncritically assume that the conceptual and descriptive tools and categories used in the phonological study of one language (L1) can be used in the study of another.

In this paper, a large corpus-based research project on the intonation of the English spoken by Finns is reported. The research is motivated by at least two factors. Firstly, there is no continuing tradition of the study of the prosody of Finnish English speech – in contrast to segmental aspects, which have been investigated much more thoroughly (see Morris-Wilson, 1992 for a detailed discussion of the subject). Secondly, the features typical of Finnish English intonation have been attributed almost solely to the effect of negative transfer or interference. Thus, Hirvonen (1967) categorically concludes that “the English rising intonation causes great learning problems for Finnish students of English”. Hirvonen’s claim is based on the common assumption that, as there are no rising intonation contours in standard spoken Finnish, rising tones in a second language will automatically be a problem for the Finn. One of the eventual aims of the present research project is to assess the all-important role of phonetic interference that is still assumed to be a major factor in second language acquisition.

Speech data
All the material used for the research was produced by 20 volunteers, Finnish students of English philology at Oulu University. The speakers were females (aged between 23 and 29) and were all fourth year students planning their MA (“pro gradu”) research on applied language studies. Prior to the recording sessions, during the thesis seminars, the speakers had formed presenter-opponent pairs to discuss their research plans, and this arrangement was utilized when collecting the data. The conversation of each pair was digitally recorded with DAT in an anechoic studio to produce a 48 kHz, 16-bit recording. The data was collected in 10 consecutive sessions within a period of two months. The data was stored in a PC as wav format files. For each speaker, a total of 5-15 minutes of English speech data was stored (including silent periods etc).
Acoustic analysis

The acoustic analysis was carried out by means of the PRAAT software package (Boersma & Weenink 1998) on an iMac computer. In the pitch analysis an autocorrelation method was used to obtain f0 values in Hertz every 10 milliseconds (Boersma 1993). To remove errors produced by the pitch tracker (e.g. the halving or doubling of individual f0 values), the pitch values for PRAAT were hand-corrected. Pitch tracks were obtained for each turn-constructional unit (TCU) produced by each speaker in the database. The term TCU comes from Conversation Analysis where the idea of turn-taking as an organized activity is one of the core ideas of research. Briefly, a TCU is a unit in conversation after which speaker change is possible. For practical purposes, in this study a TCU was equated with a tone unit or phrase containing a nuclear tone or accent (see the next section).

For each TCU, three acoustic/prosodic parameters were determined: pitch level, pitch range and pitch dynamism (cf. Daly & Warren 2001). Pitch level refers to the average speaking fundamental frequency (f0) value expressed on the linear Hertz scale. The Hertz data was converted to ERB (Equivalent Rectangular Bandwidth), a scale based on the frequency selectivity of the human auditory system. The following formula was used to convert each f0 value from the pitch analysis to ERB: ERB=16.7 LOG (1+f/165.4), f being frequency in Hertz. For each TCU pitch range values were calculated based on the ERB values. Following the procedure suggested by Daly & Warren (2001), pitch dynamism was calculated using the first-differential of the ERB pitch curves, multiplied by 100 in order to obtain a quantitative measure of dynamism expressed in ERB per second.

Analysis of speech melody

The acoustic parameters described in the previous section provide valuable information about the “global” prosodic structure of the speech data, but the parameters do not describe intonation, or speech melody, as such. It is thus clear that a more linguistically oriented analysis is needed to investigate intonation. The acoustic parameters described above can be measured completely independently of the linguistic/phonological structure of the speech data. For example, the parameter describing the rate of change or dynamism of the f0 movement does not indicate where the most dynamic f0 contours are located in a TCU. Thus, measures of mean f0-based values for the speech material do not provide information on the course of pitch at the utterance level, nor do they provide any direct information on the linguistically relevant variation in pitch. To investigate speech melody, auditory analysis/annotation is necessary. The purpose of the linguistic/phonological analysis is to investigate the functionally important parameters that are not directly measurable from the acoustic signal.

To describe these features of the speech data, the corpus is currently being analyzed and annotated utilizing PRAAT and Transcriber, a tool for assisting the manual annotation of speech signals. The aim is to create interlinear transcription, where each word is annotated with prosodic and syntactic information (displayed under the word).

A “British school” type of intonation analysis is carried out. The tone unit (or TCU) is assumed to have a fairly clearly-defined internal structure, containing at least the tonic/nuclear syllable, with optional proclitic and enclitic elements. The nuclear pitch patterns are analyzed acoustically, and f0 movements wider than 3 ST are analyzed as dynamic (“kinetic”) tones, and narrower f0 movements are analyzed as level tones. In the analysis of the data, the following tones are possible: fall, rise-fall, rise, fall-rise, and level tone. Using the framework developed by Brazil (e.g. 1997), the tones recognized in the analysis are thus the following: proclaiming p tone (fall), proclaiming p+ tone (rise-fall), referring r tone (fall-rise), referring r+ tone (rise), and oblique o tone (level).

In the future work, the ToBI framework can be used to complement the traditional “British school” type of nuclear approach. ToBI (Tones and Break Indices) is a framework for developing generally agreed-upon conventions for transcribing the intonation and prosodic structure of spoken utterances in a language variety (see e.g. Beckman & Ayers 1993). Originally developed for English, the system has been used in the description of the prosody of a number of languages, including Dutch, Spanish, Italian, Japanese and French – to date, the ToBI framework has not been extensively applied to Finnish (or Finnish English).
Analysis of prominence, rhythm and voice quality

To annotate prominence, a classification with four degrees of stress is used. A division is made into primary stress, secondary stress, tertiary stress and unstressed. Primary stress refers to the principal pitch prominence in the tone unit, i.e. it is the nuclear accent near the end of the intonation phrase. Secondary stress refers to a subsidiary pitch prominence in a tone unit, i.e. it is a non-nuclear pitch accent. Tertiary stress involves a prominence produced with length and loudness, i.e. it is not prominent enough to be an accent. Unstressed words completely lack any sort of prosodic prominence.

The breaks or pauses are transcribed in terms of phonetic information. In the annotation, a silent pause is defined as a segment characterized by the absence of energy in the speech signal, with a minimum duration of 100 ms. Briefer pauses are not transcribed. Very brief pauses (50-100 ms in duration) typically occur between words, and may be caused by segmental features, such as plosives. A filled pause, with a minimum duration of 100 ms in the annotation, typically contains vocalisation of some kind. Filled pauses may be “vocalizations” (e.g. uh, um) or “vocal noises” (inhalations, exhalations, laughter, sobbing, etc.). The latter category basically represents “respiratory reflexes” and “voice qualifications” as defined by Crystal and Quirk (1964).

Finally, auditory features of voice quality are included in the annotation. For each TCU or tone unit, a categorization of voice quality is chosen. The descriptors include “modal voice”, “falsetto”, “creak”, “whisper”, “tense” and “rough”; these labels are basically those suggested by Laver (1994) for the description of different phonation types. Features of rhythm are also annotated. While the absolute speed of speech can be measured in terms of phonemes/syllables per second, utilizing automatic speech recognition methods, for example, more complex qualities are difficult to quantify. Following Roach et al (1998), the following labels, in addition to “neutral”, are used to describe the rhythm of speech: “fast”, “slow”, “accelerating”, “decelerating”, “clipped”, “drawled”, “precise”, and “slurred”.

All in all, the annotation involves multilevel transcriptions. The first layer is used to transcribe the speech material orthographically. The second layer is used to code the tones in terms of the nuclear tone approach. The third layer represents lexical and sentence stress, distinguishing between four degrees of accentuation. The fourth layer describes the duration and type of each pause in the data. The fifth layer is used to code features of voice quality, and the sixth layer is reserved for features of the rhythm of speech.

Preliminary results

To carry out the analysis at each level described above, in addition to the acoustic measurements, is a highly time-consuming and arduous task. So far, the acoustic analysis and the investigation of tone choice have been carried out for ten speakers. Analysis at the other tiers is still unfinished.

Table 1 presents the results of the analysis of tone choice in phrase final position (for ten speakers), along with the results of the acoustic analysis (for ten speakers).

<table>
<thead>
<tr>
<th>TONE CHOICE</th>
<th>PERCENTAGE</th>
<th>PITCH LEVEL</th>
<th>PITCH RANGE</th>
<th>PITCH DYNAMISM</th>
</tr>
</thead>
<tbody>
<tr>
<td>P (fall)</td>
<td>55</td>
<td>205 Hz</td>
<td>2.6 ERB</td>
<td>7.4 ERB/s</td>
</tr>
<tr>
<td>P+ (rise-fall)</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R (fall-rise)</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R+ (rise)</td>
<td>30</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>O (level)</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

On the basis of these preliminary results, it is possible to call into question some old myths about Finnish English interlanguage intonation. Firstly, although a falling tone was clearly the most common type in phrase final position, endings with rising tones were also common. If the claims based on the notion of phonetic interference were correct, the Finns would use almost exclusively falling tones when speaking English. Secondly, and more importantly perhaps, the values of the continuously variable prosodic parameters in the data are not that different from those reported for native speech corpora (see e.g. Daly & Warren, 2001 and Toivanen, 2001 for detailed discussions of the subject).
the analysis progresses, it will be possible to give a fuller description of the prosodic (phonological and phonetic) features of the English spoken by Finns.

Conclusion
The corpus described here is one of the largest Finnish English speech databases, containing spontaneous speech that is being carefully annotated at several phonetic/phonological prosodic levels. Once fully annotated, the database will, hopefully, be a valuable tool in basic research, enabling a theoretically coherent multi-parametric description of the prosody of spoken Finnish English. The database can provide the researchers with a rich source of evidence that can be utilized when assessing arguments about non-native speakers’ L2 speech skills. Only by analyzing large amounts of authentic speech data with empirical and statistical methods, can the researcher be more confident in questioning old views and stating new hypotheses.

References
Hirvonen P. (1967) On the problems met by Finnish students in learning the rising interrogative intonation in English. Department of Phonetics, University of Turku.