PROMOTING INCREASED PITCH VARIATION IN ORAL PRESENTATIONS WITH TRANSIENT VISUAL FEEDBACK

Rebecca Hincks and **Jens Edlund**

KTH Royal Institute of Technology

This paper investigates learner response to a novel kind of intonation feedback generated from speech analysis. Instead of displays of pitch curves, our feedback is flashing lights that show how much pitch variation the speaker has produced. The variable used to generate the feedback is the standard deviation of fundamental frequency as measured in semitones. Flat speech causes the system to show yellow lights, while more expressive speech that has used pitch to give focus to any part of an utterance generates green lights. Participants in the study were 14 Chinese students of English at intermediate and advanced levels. A group that received visual feedback was compared with a group that received audio feedback. Pitch variation was measured at four stages: in a baseline oral presentation; for the first and second halves of three hours of training; and finally in the production of a new oral presentation. Both groups increased their pitch variation with training, and the effect lasted after the training had ended. The test group showed a significantly higher increase than the control group, indicating that the feedback is effective. These positive results imply that the feedback could be beneficially used in a system for practicing oral presentations.

INTRODUCTION

In the wake of globalization, public speaking is increasingly done in the world's second language, English. One form of public speaking is the oral presentation, a genre often studied and taught in courses such as Academic English, Business English or Technical English. Holding an oral presentation in front of their classes gives students who have reached a certain level of communicative competence the opportunity to practice a task that they in all likelihood will meet in their working life. Teachers are given the opportunity to listen in a focused manner to the spoken production of individual students, and classes are given a chance to learn about various topics from their peers rather than from their teachers. Oral presentations can be assigned a grade, and deserve treatment as a genre in themselves, comparable to traditional written genres.

One aspect of a successful oral presentation is that the speaker has used his or her voice in a way that has facilitated access to the content of the presentation. This involves temporal features, such as speaking at a pace that is appropriate for the audience, and expressive features, such as using pitch and loudness to give aural shape to the information structure of one's intended message. This use of intonation can be a challenge for any novice public speaker, but it is more so for those who are speaking in a second language. This is particularly true for speakers whose native languages have intonational systems that differ greatly from English.

In the research reported on in this paper we have taken steps in the direction of developing a system for practicing oral presentations with feedback provided by speech technology. People who are required to hold a presentation in a second language are inclined to practice the presentations, especially if they are to receive a grade. Because of the widespread use of presentation software, most speakers are in the proximity of computers as they practice. This presents an opportunity for computer-based feedback (Hincks, 2005). Speech recognition could be used to provide a transcript of the presentation, which could be analyzed for the presence of desirable and undesirable linguistic features. Speech recognition and analysis could also be used to give feedback on the speaker's pronunciation and intonation. In order to

achieve the goal of presentation feedback, however, we must find ways to successfully apply speech technology to the production of free, rather than modeled, speech.

Speech analysis for teaching intonation

The term speech technology covers three basic technologies: speech analysis, speech synthesis, and speech recognition. An overview of the use of speech technology for teaching pronunciation was recently published by Levis (2008). Speech recognition can be used to provide feedback on speaker pronunciation at the phonemic level (e.g., Neri, Cucchiarini, & Strik, 2008). Since it has achieved considerable breakthroughs in recent years, it can now also be used to provide immediate transcripts of presentations made with native-accented speech (Hincks, 2009). However, we focus on speech analysis, which can be used for feedback on intonational features.

Speech analysis is a technology that separates a speech signal into its component parts in order to provide information about the frequencies and intensities of the sound. A typical visual display of an analysis has three main components: the speech waveform, showing the intensity, or how loud the sound is; the spectrogram, showing the distribution of the resonant frequencies; and the display of fundamental frequency, in which a broken line, known as the contour, curve or tracing, represents pitch. Speech analysis was once available to end users mainly in the form of expensive software such as *VisiPitch* or *Speechviewer*; however, programs such as *WaveSurfer* (Sjölander & Beskow, 2000) and *Praat* (Boersma, 2001) are now freely available via the Internet.[1]

The display of fundamental frequency has long been used to teach intonation patterns in a second language (Anderson-Hsieh, 1992; De Bot, 1983; Hardison, 2004; Molholt, 1988). A visual display of the pitch contour of a learner utterance can be compared to a teacher model of the utterance, in order to heighten the learner's perception of the importance of appropriate pitch movement and to give immediate feedback on the learner's production. The early work by de Bot (1983) established the effectiveness of giving learners audio-visual feedback on their intonation rather than audio-only. Hardison (2004) showed that training in intonation with real-time visual pitch display not only improved learner production at the supra-segmental level, but also at the segmental level. Commercially available software packages for pronunciation training, such as those produced by Auralog, incorporate speech analysis, and display the user's pitch curve along with a target model.

There are a number of limitations inherent in the way speech analysis is traditionally used for teaching intonation. One is the standard procedure of using a target model with which to compare the learner utterance. This limits the extent to which learners can use the technology on their own, and also the extent to which it can be integrated into training based on naturally occurring, authentic communication. Learners need some training in order to interpret the pitch contour. The admonition to compare with a teacher model may be interpreted by students as a requirement to match the model precisely—a task at which they are bound to fail. Furthermore, the pitch contour represents not only the intonation that is appropriate to the target language but also intonation related to, for example, speaker attitude or regional dialect. While these features in themselves could provide further pedagogical goals for a certain type of student (Chun, 1998), the type of mimicking required to match a contour precisely is probably frustrating and counter-productive. Many learners have pronunciation goals that are more oriented toward comprehensibility than to achieving a native-like accent. As English consolidates its position as the global lingua franca, there are more students whose goals are closer to the former than to the latter (Jenkins, 2000).

Further problems stem from the fact that the fundamental frequency analysis that is used to create the pitch contour is an imperfect technology, with errors ranging from octave errors—the analysis frequently missing by a full octave, something that can be caused both by the nature of fundamental frequency processing and by the processing due to the nature of phonation—to less easily caught errors. Ideally,

maximum and minimum fundamental frequency values should be set for each individual speaker in order to limit misrepresentations in the pitch contour.

The use of speech analysis over long stretches of discourse is problematic. Scrolling windows allow for the continuous display of information, but students must be able to make connections between their speech and the fairly complex visual pitch patterns that are displayed instantaneously and simultaneously. Language students who have the opportunity to receive personal tutoring on their use of intonation in extended discourse may be presented with a series of pitch tracings—something that can only be accomplished off-line, *after* the speech has been produced and recorded. However, pitch contours are by nature quite different from how language in natural contexts is perceived. They constitute a static, post-hoc, abstract representation of some of the acoustic properties of utterances that are already spoken and lost, whereas the acoustics of speech are normally perceived only in the moment: they are transient and direct rather than static and analytical.

We know that giving learners feedback on intonation is valuable, and that it is enabled by the visual representation provided by speech analysis. The standard technique can be advantageously used for practicing phrases in the type of pronunciation training done at elementary levels of language training, but is inadequate for stimulating intonational development over longer stretches of discourse (Chun, 1998) such as those produced by intermediate and advanced learners who make oral presentations.

Pitch variation and movement in native and non-native public speaking

Let us now turn to what is known about the way pitch is used by native and non-native speakers as they speak in public. First-language speech that is directed to a large audience is normally characterized by more pitch variation than conversational speech (Johns-Lewis, 1986). In studies of English and Swedish, high levels of variation correlate with perceptions of speaker liveliness (Hincks, 2005; Traunmüller & Eriksson, 1995) and charisma (Rosenberg & Hirschberg, 2005; Strangert & Gustafson, 2008).

The variable that can be used to represent pitch variation is the normalized standard deviation of fundamental frequency. The standard deviation will decrease with increasing amounts of data, but if the amount of data under analysis is constant, it will reflect differing amounts of variation. In our work we examine the standard deviation of a window of ten seconds of speech at a time. The window moves through the speech as it is processed. If the speaker makes little movement from his or her mean fundamental frequency, the standard deviation will be low. If the speaker has raised or lowered fundamental frequency to give focus to an important word or concept or to indicate a change in topic, the standard deviation will be higher.

Speech that is delivered without pitch variation affects a listener's ability to recall information and is not favored by listeners. This was established by Hahn (2004) who studied listener response to three versions of the same short lecture: delivered with correct placement of primary stress or focus, with incorrect or unnatural focus, and with no focus at all (monotone). She demonstrated that monotonous delivery, as well as delivery with misplaced focus, significantly reduced a listener's ability to recall the content of instructional speech, as compared to speech delivered with natural focus placement. Furthermore, listeners preferred incorrect or unnatural focus to speech with no focus at all.

Intonation has many functions in English, many of which are related to the interaction between speakers in dialogue. In this study, we focus, however, exclusively on intonational functions that are relevant for monologue. Chun (2002) summarizes the functions of intonation in English from the language learning perspective. For monologue, relevant functions would include those "beyond the sentence level for the purpose of achieving continuity and coherence within a discourse, regardless of the length of the discourse" (p. 56). For example, a presenter needs to use intonation to "mark prominence, focus, or newsworthiness of a piece of information in a discourse" and to "mark boundaries in a discourse, e.g. boundaries between sentences, paragraphs, [and] topics" (p.56). Roughly, pitch movement, usually to a

higher level, is used in English to mark focus, and pitch resets, again to a higher level, are used to introduce new topics. If a speaker speaks with little pitch movement, in a near-monotone voice, the speaker is not making use of the potential of intonation to add structural cues to help the audience understand his or her message. The speaker also risks conveying an impression of disengagement in the topic and the audience (Pickering, 2001).

A number of researchers have pointed to the tendency for Asian L1 (first language) individuals to speak in a monotone in English. Speakers of tone languages have particular difficulties using pitch to structure discourse in English. Because in tonal languages "pitch functions to distinguish lexical rather than discourse meaning" (Wennerstrom, 1994, p. 417) they tend to strip pitch movement for discourse purposes from their production of English. Pennington and Ellis (2000) tested how speakers of Cantonese were able to remember English sentences based on prosodic information, and found that even though the subjects were competent in English, the prosodic patterns that disambiguate sentences such as *Is HE driving the bus?* from *Is he DRIVing the bus?* were not easily stored in the subjects' memories. Their conclusion was that speakers of tone languages simply do not make use of prosodic information in English, possibly because for them pitch patterns are something that must be learned arbitrarily as part of a word's lexical representation. In a second study, however, Pennington and Ellis showed that for certain prosodic features, improvement could be achieved when the subject's attention was explicitly drawn to prosodic information.

Many non-native speakers have difficulty using intonation to signal meaning and structure in their discourse. Wennerstrom (1994) studied how non-native speakers used pitch and intensity contrastively to show relationships in discourse. She found that "neither in … oral-reading nor in … free-speech tasks did the L2 (second language) groups approach the degree of pitch increase on new or contrastive information produced by native speakers. Similarly, there was less reduction of pitch and volume on … redundant words in the oral reading on the part of L2 subjects relative to native speakers" (pp. 415-416). This more monotone speech was particularly pronounced for the subjects whose native language was Thai—a tone language like Chinese. Chinese-native teaching assistants use significantly fewer rising tones than native speakers in their instructional discourse (Pickering, 2001) and thereby miss opportunities to ensure mutual understanding and establish common ground with their students. In a specific study of Chinese speakers of English, Wennerstrom (1998) found a significant relationship between the speakers' ability to use intonation to distinguish rhetorical units in oral presentations and their scores on a test of English proficiency. Pickering (2004) applied Brazil's (1986) model of intonational paragraphing to the instructional speech of Chinese-native teaching assistants at an American university. Intonational paragraphing gives structure to English discourse by means of pitch resets at topic changes, and a corresponding series of decreasing peaks until there is a new topic change. By comparing intonational patterns in lab instructions given by native and non-native TAs, she showed that the non-natives lacked the ability to create intonational paragraphs and thereby to facilitate the students' understanding of the instructions. The analysis of prosodic units in Pickering's work was "hampered at the outset by a compression of overall pitch range in the [international teaching assistant] teaching presentations as compared to the pitch ranges found in the [native speaker teaching assistant] data set" (2004, p. 31). The Chinese natives were speaking more monotonously than their native-speaking colleagues.

Learning to speak with more variation

One pedagogic solution to the tendency for Chinese native speakers of English to speak monotonously as they hold oral presentations would be simply to give them feedback when they have used significant pitch movement in any direction. The feedback would be divorced from any connection to the semantic content of the utterance, and would basically be a measure of how non-monotonously they are speaking. While a system of this nature would not be able to tell a learner whether he or she has made pitch movement that is specifically appropriate or native-like, it should stimulate the use of more pitch variation in speakers who underuse the potential of their voices to create focus and contrast in their instructional discourse. It

could be seen as a first step toward more native-like intonation, and furthermore to becoming a better public speaker. In analogy with other learning activities, we could say that such a system aims to teach students to swing the club without necessarily hitting the golf ball perfectly the first time. Importantly, because the system would give feedback on the production of free speech, it would stimulate and provide an environment for the autonomous practice of authentic communication such as the oral presentation.

The use of a computer environment for practicing oral presentations was inspired by the CALL theoretical framework proposed by Levy (1997), who advised that CALL designers give careful consideration to the role they expect the computer to play in the teaching and learning process. Designers should in particular be wary of assigning the role of trusted 'tutor' to a computer program that may deliver incorrect feedback on learner production. Here we see the computer's role as more of a 'tool' than a virtual tutor—a tool that will provide a learning environment capable of responding interactively to learner production, without attempting to provide 'right' or 'wrong' answers to the way the student delivers the presentation.

Like the majority of CALL systems, a presentation practice system would provide environments for skills practice where learners are rewarded for meeting certain targets. Unlike most CALL systems, however, the student input would be freely-generated speech with an authentic communicative intent. Enabling communication with a computer is no simple matter, yet much research points to the supremacy of constructive methodologies when it comes to teaching a second language. Having the computer respond to the prosody of presentation speech rather than its lexical content is one way of having it react to the communicative intent of the speaker. In such a system, the target levels for prosodic variation could be flexible, allowing for instructional scaffolding in response to the initial skills of the learner. By providing an environment for rehearsing a presentation, the system would encourage the use of self-assessment by allowing learners to record themselves as they practice. Many learners are bewildered by advice such as: 'use more variation in your speaking style'; such a system would allow them to test different styles on their own. Finally, like many applications of information and communication technologies in learning situations, the application would stimulate lifelong learning, by being available to users outside traditional classroom settings.

Our study was inspired by four points concluded from previous research:

1. Visualization of pitch movement is beneficial to learners but current techniques have limitations.

2. Public speakers need to use varied pitch movement to structure discourse and engage with their listeners.

3. Second language speakers, especially those of tone languages, are particularly challenged when it comes to the dynamics of English pitch.

4. Learning activities are ideally based on the student's own language, generated with an authentic communicative intent.

These findings generated the following primary research question:

Will on-line visual feedback on the presence and quantity of pitch variation in learner-generated utterances stimulate the development of a speaking style that incorporates greater pitch variation?

Following previous research on technology in pronunciation training (de Bot, 1983; Motohashi-Saigo & Hardison, 2009), comparisons were made between a test group that received visual feedback and a control group that was able to access auditory feedback only. Three hypotheses were tested:

1. Visual feedback will stimulate a greater increase in pitch variation in training utterances as compared to auditory-only feedback.

2. Participants with visual feedback will be able to generalize what they have learned about pitch movement and variation to the production of a new oral presentation.

3. Participants with visual feedback will experience a greater degree of satisfaction with their training experience.

In addition, we conducted a preliminary follow-up test of human perception of the effect of the training, to ensure that the feedback did not stimulate the development of a speaking style that would be perceived as odd or unnatural.

METHOD

Base system

The system we used consists of a base system allowing students to listen to teacher recordings (targets), read transcripts of these recordings, and make their own recordings of their attempts to mimic the targets. Students may also make recordings of free readings.[2] Furthermore, students can browse through targets, make new recordings and listen to their latest recording. The interface keeps track of the students' actions, and some of this information, such as the number of times a student has attempted a target, is continuously presented to the student.

The amount of control the student has over the details in the base system is limited, as it is designed for simplicity of use: file names are assigned automatically, the target files are selected by the teacher and constitute a fixed set of utterances (from the student's perspective; the teacher can change the utterance set), and only the latest recording can be replayed. Sacrificing detailed control allows us to present the student with an interface that is easy to learn and difficult to misuse.

A student session is initiated by informing the software of who the student is. After that, all student actions (listen, record, read transcript, playback) are logged and all student recordings are saved, together with information on the context in which they were recorded.

Pitch analysis

The meter is fed data from an online analysis of the recorded speech signal. The analysis used in these experiments is based on the */nailon/* online prosodic analysis software (Edlund & Heldner, 2006) and the *Snack Sound Toolkit*.[3] As the student speaks, a fundamental frequency estimation is continuously extracted using an incremental version of *getF0/RAPT* (Talkin, 1995). The estimation frequency is transformed from Hz to logarithmic semitones, a move from fundamental frequency (an acoustic measure) to pitch (a perceptual measure). There are several reasons for this transformation. Semitones are perceptually relevant, because they are perceptually equidistant, so that a rise of one semitone from 1 to 2 is perceptually the same as a rise from 4 to 5, whereas a rise from 100 to 200 Hz is perceptually much higher than one from 400 to 500 Hz. This gives us a kind of perceptual speaker normalization, which affords us easy comparison between pitch variation in different speakers. Similarly it allows us to compare the variation of a speaker on different occasions, even if the speaker ends up speaking with a generally higher pitch on one of the occasions. Fundamental frequency distributions in Hz over a single speaker also fit a normal distribution less closely than pitch distributions expressed in semitones (Edlund & Heldner, 2007), making the following steps more reliable.

After the semitone transformation, the next step is a continuous and incremental calculation of the standard deviation of the student's pitch over the last 10 seconds. The result is a measure of the student's recent pitch variation.

Pitch variation feedback

The base system is extended with a component providing online, instantaneous and transient feedback visualizing the degree of pitch variation the student is currently producing. The feedback is presented in a meter that is reminiscent of the amplitude bars used in the equalizers of sound systems: the current amount of variation is indicated by the number of bars that are lit up in a stack of bars, and the highest variation over the past two seconds is indicated by a lingering top bar, as seen in Figure 1. The meter has a short, constant latency of 100 ms.
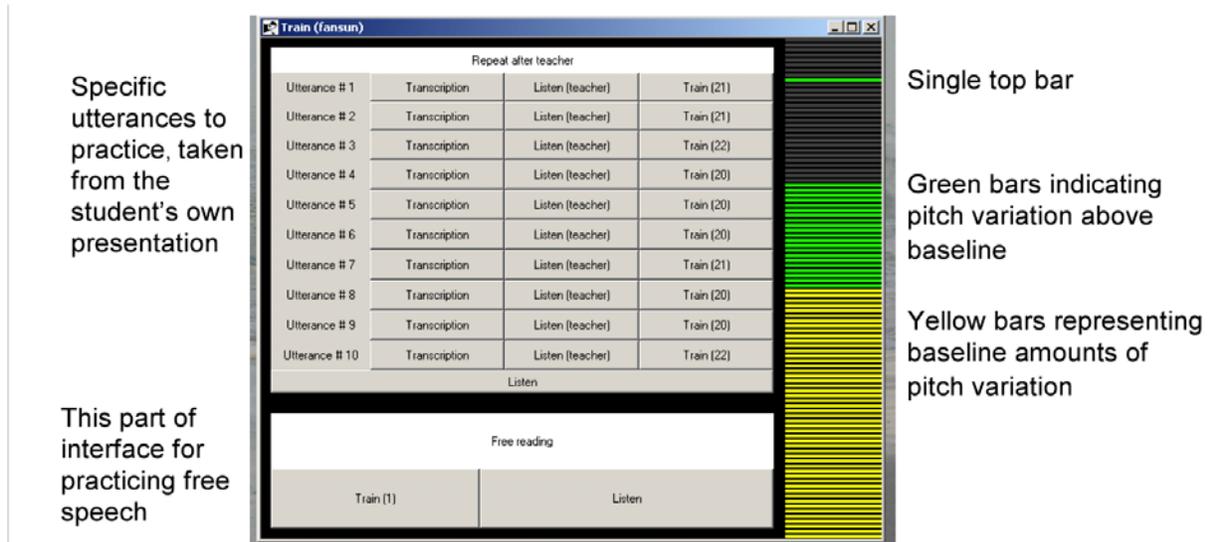

Figure 1. Training interface, showing pitch meter to the right. Green bars indicate that the speaker is speaking with relatively increased pitch variation; the single top bar represents the highest relative variation measured in the preceding two seconds.

The pitch variation fed to the meter is first normalized against a base value, that is, the pitch variation the student produced in the initial session. The meter utilizes a dampening function, making it impossible for students to max the meter out—the more bars that are lit up, the more variation that is needed to light another one. The pitch meter shows yellow bars when the pitch variation is low or similar relative to the student's initial reading, and green bars when it is higher.

The kind of feedback provided by this system is very different from the kind of feedback given by contour visualization. Its transience does not allow for post-analysis together with the student. It is designed to be used independent of expert interpretation, by students working on their own with a computer. Its automaticity potentially allows a maximum amount of time on task, where the immediacy of seeing a light flash when part of an utterance has been stressed by means of a rise in pitch should reinforce positive developments in speaking habits.

Participants

The test group and the control group each consisted of 7 students of engineering, 4 women and 3 men.[4] The participants were recruited from English classes at Sweden's largest technological university, and were exchange students from China, in Sweden for stays of six months to two years. Participants' proficiency in English was judged by means of an internal placement test to be at the upper intermediate to advanced level, with one student at the lower intermediate level. The mean age of the test group was slightly lower than the control group, 22.3 vs. 24.5 years, but both groups had started studying English at an average age of 11. The reported years of English studies were therefore higher for the control group: 12.3 vs. 10.3 years (Table 1). The participants spoke a variety of dialects of Chinese but used Mandarin

with each other and for their studies. They did not speak Swedish and were using English with their teachers and classmates. Four of them had spent four years at an English-language university in Singapore, but none of them had spent extended periods of time in an inner-circle (Kachru, 1985) English L1 country.

Table 1. *Data Regarding Participants and Time on Training Task*

	Test (*n* = 7)		Control (*n* = 7)	
	M	*SD*	*M*	*SD*
Age	22.3	0.90	24.5	1.51
Years studying English	10.3	3.59	12.3	1.98
Minutes of training	181	21	171	29
Repetitions of utterances	32	11	23	7

Procedure and Material

Each participant began the study by giving an oral presentation of about five minutes in length, either for their English classes or for a smaller group of students. Audio recordings were made of the presentations using a small clip-on microphone that recorded directly into a computer. The presentations were also video-recorded, and participants watched the presentations together with one of the researchers, who commented on presentation content, delivery and language. The individualized training material for each subject was prepared from the audio recordings. A set of 10 utterances, each of about 5-10 seconds in length, was extracted from the participants' speech. The utterances were mostly non-consecutive and were chosen on the basis of their potential to provide examples of contrastive pitch movement within the individual utterance. The first researcher recorded her own (native-American English) versions of them, making an effort to use her voice as expressively as possible and making more pitch contrasts than in the original student version. For example, a modeled version of a student's flat utterance could be represented as: "And THIRDly, it will take us a lot of TIME and EFfort to READ each piece of news." Two sample sets of utterances are shown in the Appendix.

The participants were assigned to the control or test groups following the preparation of their individualized training material. Participants were ranked in terms of the global pitch variation in their first presentation, as follows: they were first split into two lists according to gender, and each list was ordered according to initial global pitch variation. Participants were randomly assigned pair-wise from the list to the control or test group, ensuring gender balance as well as balance in initial pitch variation. Four participants who joined the study at a later date were distributed in the same manner.

Participants completed approximately three hours of training in half-hour sessions; some participants chose to occasionally have back-to-back sessions of one hour. The training sessions were spread out over a period of four weeks. The mean training time per group and number of repeated utterances are reported in Table 1. Control participants repeated a fewer number of utterances than did test participants; this is probably due to the fact that the only feedback they could receive was to listen to recordings of their production, which in itself used up some of the training time. Training took place in a quiet and private room at the university language unit, without the presence of the researchers or other onlookers. For the first four or five sessions, participants listened to and repeated the teacher versions of their own utterances. They were instructed to listen and repeat each of their 10 utterances between 20 and 30 times. Test group participants received the visual feedback described above and were encouraged to speak so that the meter showed a maximum amount of green bars. The control group was able to listen to recordings of their production but received no other feedback.

Upon completion of the repetitions, both groups were encouraged to use the system to practice their second oral presentation, which was to be on a different topic than the first presentation. For this practice,

the part of the interface (Figure 1) designated for 'free speech' was used. In these sessions, once again the test participants received visual feedback on their production, while control participants were only able to listen to recordings of their speech. Within 48 hours of completing the training, the participants held another presentation, this time about ten minutes in length, for most of them as part of the examination of their English courses. This presentation was audio recorded.

Questionnaire

Participants also completed a questionnaire (Table 2) about their experience taking part in the training. The questionnaire consisted of 10 statements about the training, to which the participants responded on a 5-point Likert scale where 5 = agreed completely and 1 = disagreed completely. Test participants responded to an additional four statements specifically about the pitch meter.

Perception Tests

In the final stage of the study, listening tests were carried out in order to ensure that the pitch variation feedback had not stimulated the development of an unnatural speaking style. Two native-speaking authorities on intonation rated one minute of speech from each of the first and second presentations. The minute between 3.00 and 4.00 was extracted from the second presentations. Because the first presentations were shorter, it was necessary to use the minute between 2.30 and 3.30 to avoid lexical cues that the presentations were coming to an end. Raters listened to sets of the 12 male files and 16 female files separately, in separately randomized orders. They rated the speech on a five-point scale for four qualities: naturalness, liveliness, pronunciation, and intelligibility.

RESULTS

Pitch variation

We measured development in two ways: over the roughly three hours of training per student, in which case we compared pitch variation in the first and the second half of the training for each of the 10 utterances used for practice, and in generalized form, by comparing pitch variation in two presentations, one before and one after training. Pitch estimations were extracted using the same software used to feed the pitch variation indicator used in training, an incremental version of the *getF0/RAPT* (Talkin, 1995) algorithm. Variation was calculated in a manner consistent with Hincks (2005) by calculating the standard deviation over a moving 10-second window.

In the case of the training data, recordings containing noise only or those that were empty were detected automatically and removed. For each of the 10 utterances included in the training material, the data were split into a first and a second half, and the recordings from the first half were spliced together to create one continuous sound file, as were the recordings from the second half. The averages of the windowed standard deviation of the first and the second half of training were compared. The basic assumption was that speakers from both groups should have a higher pitch variation in the latter half of training than in the first, and Hypothesis 1 states that the test group should show a greater increase than the control group. For the two presentations, the basic assumption was that both groups should show an increased variation after training as compared to before. Hypothesis 2 states that the test group should show a larger increase than the control.

The mean standard deviations for each data set and each of the two groups are shown in Figure 2. The y-axis displays the mean standard deviation per moving 10-second frame of speech in semitones, and the x-axis the four points of measurement: the first presentation, the first half of training, the second half of training, and the second oral presentation. The experimental group shows greater pitch variation across all points of measurement following training. Learning takes place in the first half of training, where the difference between the two groups jumps significantly from nearly no difference to one of more than 2.5

semitones, and is maintained in the second half of training. It is then transferred to the second presentation, which took place without the presence of feedback.
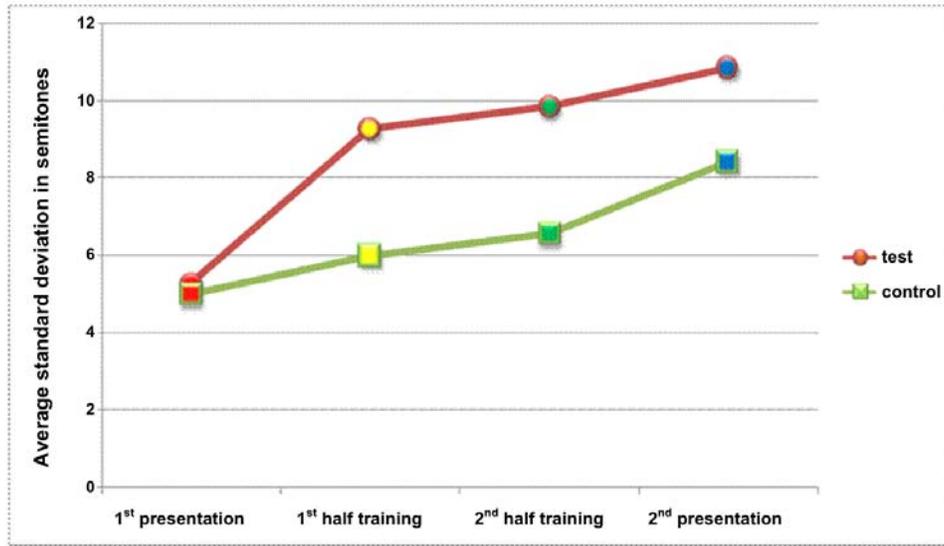
Figure 2. Average pitch variation over 10 seconds of speech for the two experimental conditions during the 1st presentation, the 1st half of the training, the 2nd half of the training, and the 2nd presentation. The test group shows a statistically significant effect of the feedback they were given**.**
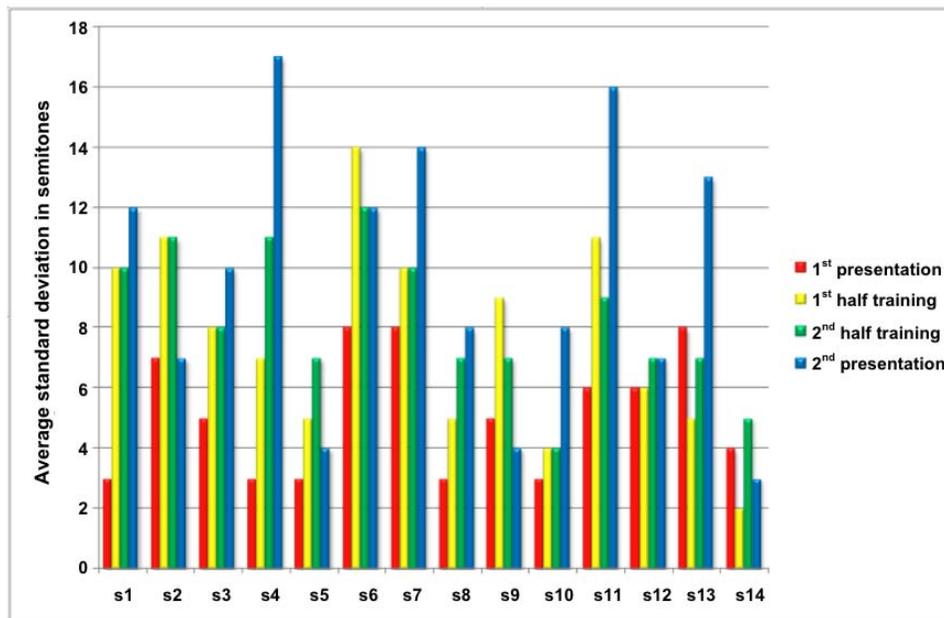

Figure 3. Average standard deviation of pitch over 10 seconds of speech for each of the participants during the 1st presentation, the 1st half of the training, the 2nd half of the training and the 2nd presentation. S1-s7 belong to the test group; s8-14 are the control.

The effect of the feedback method (test group vs. control group) was analyzed using an ANOVA with time of measurement (1st presentation, 1st half of training, 2nd half of training, 2nd presentation) as a within-subjects factor. The sphericity assumption was met, and the main effect of time of measurement was significant ($F = 8.36$, $p < .0005$, $\eta^2 = 0.45$) indicating that the speech of the test group receiving visual feedback increased more in pitch variation than the control group. Between-subject effect for

feedback method was significant ($F = 6.74$, $p = .027$, $\eta^2 = 0.40$). The two first hypotheses are confirmed by these findings. The individual results per speaker are illustrated in Figure 3.

Expert ratings

The results of the preliminary listening test are encouraging. The purpose of rating the one-minute samples from each presentation was to eliminate concerns that the visual feedback could promote the development of an unnatural speaking style if speakers made wild pitch excursions in order to make the green lights flash. The averages of the two ratings are shown in Figure 4. The responses of only two raters provide too little data to allow for statistical analysis, and their inter-rater agreement was only moderate, with a Pearson correlation of 0.41. However, ratings for the test group indicate a positive trend toward a slight improvement in both liveliness and naturalness. The control group was also perceived to increase in liveliness, but was found to worsen in terms of naturalness. Little change is perceived for either group in pronunciation and intelligibility, which were two features our system did not attempt to address. Though we reiterate that this preliminary test can only give indications as to the effect of the training, we believe that the ratings show that we do not need to be concerned that test participants were prompted to put new pitch movement in unnatural places. The feedback thusly did not have a damaging effect on the participants.
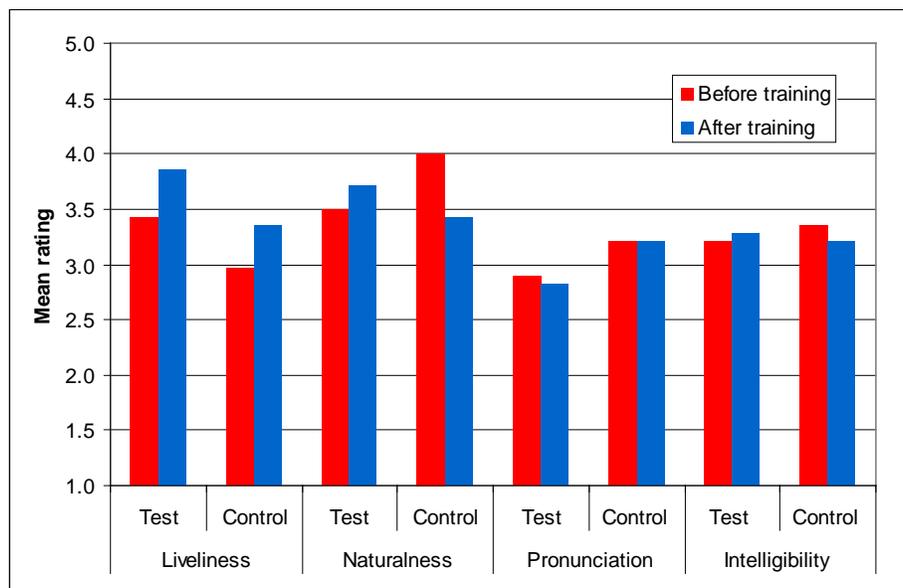

Figure 4. Average of two expert blind ratings of one minute of speech extracted from the first oral presentation (before training) and the second oral presentation (after training).

Questionnaire

Both control and experimental students were satisfied with their training. Table 2 shows the results of the questionnaire. On the 5-point scale, the mean responses to all questions but one were 4 or above. The mean satisfaction for the test group was slightly higher than for the control group: 4.34 vs. 4.29, but a two-sample t-test assuming equal variance of the mean responses to the 10 questionnaire statements common to both groups showed no effect of group, $t(18) = .330$, $p < .05$. Both groups must thus be seen to be equally satisfied with their training, and Hypothesis 3 must be rejected.

Table 2. *Results of Questionnaire Regarding Student Satisfaction with Training*

#	Question	Test M (*n* = 7)	SD	Control M (n = 7)	SD
1	Watching the video of my first presentation showed me what I had to do to improve my pronunciation	4.29	.95	4.57	.53
2	Receiving the teacher's comments on the first presentation helped me improve my pronunciation	4.86	.38	4.71	.49
3	Listening to the teacher version of my utterances helped me improve my pronunciation	4.43	.53	4.71	.76
4	Imitating teacher utterances helped me improve my pronunciation	4.43	.79	4.43	.53
5	Listening to my own new recordings of the utterances helped me improve my pronunciation	4.43	.79	4.00	.58
6	Trying to get the pitch meter to reach a high level as I practiced the utterances helped me improve my pronunciation	4.00	.82	n.a.	
7	Listening to the recording as I practiced my second presentation helped me improve my pronunciation	4.14	.90	3.86	.69
8	Watching the pitch meter as I practiced my second presentation helped me improve my pronunciation	4.29	.76	n.a.	
9	My pronunciation in general has improved because of my participation in this project	4.00	.58	4.29	.76
10	My English intonation has improved because of my participation in this project	4.29	.49	4.14	.69
11	My production of the individual sounds of English has improved because of the project	4.00	1.00	3.86	.69
12	My presentation skills have improved because of my participation in the project	4.43	.53	4.29	.76
13	I would recommend the pitch meter to other people who want to improve their pronunciation	4.71	.49	n.a.	
14	I understood the connection between what I did with my voice and the movement of the pitch meter	4.57	.53	n.a.	
	General satisfaction with training (mean)	**4.35**		**4.29**	

DISCUSSION

The major goal of the work underlying this study has been to stimulate intermediate and advanced learners of English to make use of the expressive potential of English intonation as they speak in public. A basic point of departure has been that English speech intended for a large audience is characterized by a

large amount of pitch variation (Johns-Lewis, 1986). We hypothesized that lights briefly flashing in response to the standard deviation of fundamental frequency would be an effective means of stimulating an increase in pitch variation in monologue. Hypothesis 1, which stated that test participants would increase pitch variation significantly more than control participants during the course of their training, was confirmed by our data. Hypothesis 2, which stated that test participants would be better able to generalize what they had learned to the production of a new presentation, was also confirmed.

Our results are in line with other research that has shown that visual feedback on pronunciation is beneficial to learners (De Bot, 1983; Hardison, 2004; Motohashi-Saigo & Hardison, in press; Neri, Cucchiarini, & Strik, 2008). The visual channel provides information about linguistic features that can be difficult for second language learners to perceive audibly. The first language of our Chinese participants uses pitch movement to distinguish lexical meaning; these learners can therefore experience difficulty in interpreting and producing pitch movement at a discourse level in English (Pennington & Ellis, 2000; Pickering, 2004; Wennerstrom, 1994). Our feedback gave each test participant visual confirmation when they had stretched the resources of their voices beyond their own baseline values. It is possible that some participants had been using other means, particularly intensity, to give focus to their English utterances. The visual feedback rewarded them for using pitch movement only, and could have been a powerful factor in steering them in the direction of an adapted speaking style. While our data were not recorded in a way that would allow for an analysis of the interplay between intensity and pitch as Chinese speakers give focus to English utterances, this would be an interesting area for further research.

Based on the results of the questionnaire, both the experimental and the control participants felt that they had had a rewarding experience by participating in this study. Hypothesis 3, which stated that participants receiving feedback would feel more positively about their pronunciation development than control students, must therefore be rejected. It is likely that all participants were pleased by the extra contact they were able to have with their English teacher, and indeed the questions mentioning the teacher received the highest responses (Table 2). This is perhaps symptomatic of the current hunger for English proficiency found in the Chinese culture. Although many of the participants interacted socially with each other, none was aware of the differences between the control and the test interfaces, and seemed to think that imitating teacher models was central to the study. Test students did spend a slightly higher mean time on their training (Table 1). In a study of this nature, it is difficult to control all possible variables, and because our students were diligent, reliable and interested, we found, as did both Motohashi-Saigo and Hardison (in press) and Wang and Munro (2004), that it was possible to successfully run a study where students could participate according to their own schedules.

A serious potential concern with intonation feedback that is divorced from semantic content is that it could promote the development of an unnatural speaking style. Speech that is produced with unnatural focus is slightly more difficult to comprehend than monotone speech, though it is still preferred to monotone speech (Hahn, 2004). Fortunately, it does not appear to be the case that speakers put new focus in unnatural places. The raters, both of whom are experienced researchers in the field of first-language intonation, judged the naturalness of the intonation of a sample from the test group's second presentation to be at least as good as their first presentation. This would indicate that the feedback does not have a damaging effect. The increased pitch variation that we have measured is more likely to be contributing to an improvement in the speakers' intonation. Since researchers have pointed to the problems associated with flat tones produced by Chinese speakers making oral presentations, such as conveying the impression that the speaker is disengaged with the audience (Pickering, 2001), it could be argued that speakers should be encouraged to use more pitch movement in any direction as a step towards developing more expressive English intonation.

Given greater resources in terms of time and potential participants, it would have been interesting to compare the development of pitch variation with other kinds of feedback. For example, we could also have given completely random feedback to a third group of students to test for a placebo effect, though

that would be ethically questionable. However, the fact that our students were not aware of the differences between the two interfaces would indicate that we do not need to be concerned about a placebo effect. We could also have displayed pitch tracings of the training utterances. It has not been an objective of our study, however, to prove that our method is superior to showing pitch tracings. We simply feel that circumventing the contour visualization process allows for the more autonomous use of speech technology. A natural development in future research will be to have learners practice presentation skills without teacher models.

It is important to point out that we cannot determine from these data that speakers became better presenters as a result of their participation in this study. A successful presentation entails, of course, very many features, and using pitch well is only one of them. Other vocal features that are important are the ability to clearly articulate the sounds of the language, the rate of speech, and the ability to speak with an intensity that is appropriate to the spatial setting. In addition, there are numerous other features regarding the interaction of content, delivery and audience that play a critical role in how the presentation is received. Our presentation data, gathered as they were from real-life classroom settings, are in all likelihood too varied to allow for a study that attempted to find a correlation between pitch variation and, for example, the perceived clarity of a presentation. However, we do wish to further explore perceptions of the speakers beyond the preliminary ratings of one minute of speech per presentation. We also plan to develop feedback gauges for other intonational features, beginning with rate of speech. We see potential to develop language-specific intonation pattern detectors that could respond to, for example, a speaker's tendency to use French intonation patterns when speaking English. Such gauges could form a type of toolbox that students and teachers could use as a resource in the preparation and assessment of oral presentations.

Our study contributes to the field in a number of ways. It is, to the best of our knowledge, the first to rely on a synthesis of online fundamental frequency data in relation to learner production. We have not shown the speakers the absolute fundamental frequency itself, but rather how much it has varied over time as represented by the standard deviation. This variable is known to characterize discourse intended for a large audience (Johns-Lewis, 1986), and is also a variable that listeners can perceive if they are asked to distinguish lively speech from monotone (Hincks, 2005; Traunmüller & Eriksson, 1995). In this paper, we have demonstrated that it is a variable that can effectively stimulate production as well. Furthermore, the variable itself provides a means of measuring, characterizing and comparing speaker intonation. It is important to point out that enormous quantities of data lie behind the values reported in our results. Measurements of fundamental frequency were made 100 times a second, for stretches of speech up to 45 minutes in length, giving tens of thousands of data points per speaker for the training utterances. By converting the Hertz values to the logarithmic semitone scale, we are able to make valid comparisons between speakers with different vocal ranges. This normalization is an aspect that appears to be neglected in commercial pronunciation programs such as Auralog's *Tell Me More* series, where pitch curves of speakers of different mean frequencies can be indiscriminately compared. There is a big difference in the perceptual force of a rise in pitch of 30Hz for a speaker of low mean frequency and one with high mean frequency, for example. These differences are normalized by converting to semitones.

Secondly, our feedback can be used for the production of long stretches of free speech rather than short, system-generated utterances. It is known that intonation must be studied at a higher level than that of the word or phrase in order for speech to achieve proper cohesive force over longer stretches of discourse (Brazil, 1997; Chun, 2002; Levis & Pickering, 2004; Pickering, 2004). By presenting the learners with information about their pitch variation in the previous ten seconds of speech, we are able to incorporate and reflect the vital movement that should occur when a speaker changes topic, for example. In an ideal world, most teachers would have the time to sit with students, examine displays of pitch tracings, and discuss how peaks of the tracings relate to each other with respect to theoretical models such as Brazil's intonational paragraphs (Brazil, 1997; Levis & Pickering, 2004). Our system cannot approach that level

of detail, and in fact cannot make the connection between intonation and its lexical content. However, it can be used by learners on their own, in the production of any content they choose. It also has the potential for future development in the direction of more fine-grained analyses.

A third novel aspect of our feedback is that it is transient and immediate. Our lights flicker and then disappear. This is akin to the way we naturally process speech; not as something that can be captured and studied, but as sound waves that last no longer than the milliseconds it takes to perceive them. It is also more similar to the way we receive auditory and sensory feedback when we produce speech—we only hear and feel what we produce in the very instance we produce it; a moment later it is gone. Though at this point we can only speculate, it would be interesting to test whether transient feedback might be more easily integrated and automatized than higher-level feedback, which is more abstract and may require more cognitive processing and interpretation. The potential difference between transient and enduring feedback has interesting theoretical implications that could be further explored.

This study has focused on Chinese speakers because they are a group where many speakers can be expected to produce relatively monotone speech, and where the chances of achieving measurable development in a short period of time were deemed to be greatest. However, there are all kinds of speaker groups who could benefit from presentation feedback. Like many communicative skills that are taught in advanced language classes, the lessons can apply to native speakers as well. Teachers who produce monotone speech are a problem to students everywhere (Hamilton, 2006). Nervous speakers can also tend to use a compressed speaking range, and could possibly benefit from having practiced delivery with an expanded range. Clinically, monotone speech is associated with depression (Nilsonne, 1987) and can also be a problem that speech therapists need to address with their patients. However, the primary application we envisage here is an aid for practicing, or perhaps even delivering, oral presentations.

It is vital to use one's voice well when speaking in public. It is the channel of communication, and when used poorly, communication can be less than successful. If listeners either stop listening, or fail to perceive what is most important in a speaker's message, then all actors in the situation are in effect wasting time. A speaker delivering a monologue has an even larger responsibility than a writer does, in that the speech is transient and cannot be re-read if the meaning is missed. We require writers to write clearly, and we should require speakers to speak clearly. Clear speech in English requires pitch contrasts to show given and new information and introduce topic changes. We hope to have shown in this paper that stimulating speakers to produce more pitch variation in a practice situation has an effect that can transfer to new situations. People can learn to be better public speakers, and technology should help in the process.

NOTES

1. Available at http://www.speech.kth.se/wavesurfer/ and at http://www.praat.org.

2. Sound files illustrating the repetitions of the utterances are available at http://llt.msu.edu/vol13num3/hincks.zip. There is one file for a test participant and one for a control participant. In the files can be heard first the original utterance taken from the first oral presentation, then a repetition of the utterance made before training started. After these two baseline utterances come seven training utterances, for each the 1st, 5th, 10th, 15th, 20th, 25th, and 30th repetitions. Note that the control student's intonation varies little between the repetitions, while the student who saw feedback tries different ways of speaking.

3. Sjölander 1997-2008, available at http://www.speech.kth.se/snack/.

4. The original groups consisted of 8 participants, 4 men and 4 women. However, two outlying participants who had joined the study without being tested as to their English abilities were removed

because they differed drastically from that of the rest of the group: one turned out to have studied English for only two years and did not have the degree of proficiency to benefit from the training, and the other was about ten years older than the other participants, being a Ph.D. rather than a Master's student.

APPENDIX: Sample Training Utterances

Sample Set 1

1. Good evening, my name is XXXX, I'm here from Nanjing Technological University Singapore, and I'm here for my half-year exchange study

2. And today I'd like to give you a short presentation on RSS, which stands for really simple syndication

3. So what RSS does is that it's an effective web technology that delivers web content to the user

4. Let me first give you an example of how we read news or blog entries in the web now

5. This procedure has been passed on maybe ten years ago and it is still going on nowadays

6. But whether we're aware of it or not, there are some problems we have in this process

7. First of all when we're reading a piece of news, we are not aware of what time it is published

8. And thirdly, it will take us a lot of time and effort to read each piece of news

9. So the old technology is a pull technology, we are pulling data from the website

10. In other words, now we are going to have a push technology: we are going to let the web content push to us directly

Sample Set 2

1. Today I'd like to talk about lithium batteries

2. Portable consumer electronic devices just like laptops, cameras and mobile phones

3. As we know, lithium is a positive ion, which is held in the anode of the electrolyte of the batteries

4. This process enables the electron's flow as an external current.

5. On the other hand, when we apply a voltage to recharge it

6. The lithium ions are driven back to the anode again, and are ready to give us power

7. And this is mainly because of the overheating, which is caused by the defect manufacturing in the battery

8. Because lithium burns violently when it is exposed to moisture

9. Then second, do not try to put out a battery fire with water

10. The right way to do that is to use a chemical-based extinguisher

ACKNOWLEDGEMENTS

ABOUT THE AUTHORS

Rebecca Hincks is Associate Professor of English at the Unit for Language and Communication at KTH. She teaches Technical and Scientific English at different levels, and her primary research interest is in the use of speech technology to develop spoken language skills.

E-mail: hincks@speech.kth.se

Jens Edlund is a researcher at the Centre for Speech Technology at KTH. His main research interest is spoken communication, both between humans and between humans and computers. He currently investigates the specifics of spoken interaction within the Swedish Research Council project #2006-2172 (Vad gör tal till samtal/What makes speech special).

E-mail: edlund@speech.kth.se

REFERENCES

Anderson-Hsieh, J. (1992). Using electronic visual feedback to teach suprasegmentals. *System, 20*(1), 51-62.

Boersma, P. (2001). PRAAT: A system for doing phonetics by computer. *Glot International, 5,* 341-345.

Brazil, D. (1986). *The communicative value of intonation in English*. Birmingham, UK: University of Birmingham, English Language Research.

Brazil, D. (1997). *The communicative value of intonation in English*. Cambridge, UK: Cambridge University Press.

Chun, D. (1998). Signal analysis software for teaching discourse intonation. *Language Learning & Technology, 2*(1), 61-77. Retrieved from http://llt.msu.edu

Chun, D. (2002). *Discourse intonation in L2: From theory and research to practice*. Philadelphia: John Benjamins.

de Bot, K. (1983). Visual feedback of intonation I: Effectiveness and induced practice behavior. *Language and Speech, 26*(4), 331-350.

Edlund, J., & Heldner, M. (2006). */nailon/ - software for online analysis of prosody.* Paper presented at the Interspeech 2006 ICSLP, Pittsburgh PA, USA.

Edlund, J., & Heldner, M. (2007). Underpinning /nailon/ - automatic estimation of pitch range and speaker relative pitch. In C. Müller (Ed.), *Speaker Classification II: Fundamentals, Features, and Methods*, 229-242. Berlin/Heidelberg, Germany: Springer.

Hahn, L. D. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly, 38*(2), 201-223.

Hamilton, A. (2006, August 4). Vocal cords need to be brushed up for the classroom. *The Times Online*. Retrieved from http://www.timesonline.co.uk/tol/news

Hardison, D. (2004). Generalization of computer-assisted prosody training: Quantitative and qualitative findings. *Language Learning & Technology 8*(1), 34-52. Retrieved from http://llt.msu.edu

Hincks, R. (2005). Measures and perceptions of liveliness in student oral presentation speech: A proposal for an automatic feedback mechanism. *System, 33*(4), 575-591.

Hincks, R. (2009). Speaking rate and information content in English lingua franca oral presentations. *English for Specific Purposes*. Advance online publication. doi:10.1016/j.esp.2009.05.004

Jenkins, J. (2000). *The phonology of English as an international language: New models, new norms, new goals*. Oxford: Oxford University Press.

Johns-Lewis, C. (1986). Prosodic differentiation of discourse modes. In C. Johns-Lewis (Ed.), *Intonation in discourse* (pp. 199-220). Breckenham, Kent: Croom Helm.

Kachru, B. B. (1985). Standards, codification and sociolinguistic realism: the English language in the outer circle. In R. Quirk & H. Widdowson (Eds.), *English in the world*. Cambridge, UK: Cambridge University Press.

Levis, J. (2008). Computer technology in teaching and researching pronunciation. *Annual Review of Applied Linguistics 27*, 184-202.

Levis, J., & Pickering, L. (2004). Teaching intonation in discourse using speech visualization technology. *System, 32*, 505-524.

Levy, M. (1997). *Computer-Assisted Language Learning*. Oxford: Clarendon Press.

Molholt, G. (1988). Computer-assisted instruction in pronunciation for Chinese speakers of American English. *TESOL Quarterly, 22*(1), 91-111.

Motohashi-Saigo, M., & Hardison, D. (2009). Acquisition of L2 Japanese geminates: Training with waveform displays. *Language Learning & Technology 13* (2), 29-47. Retrieved from http://llt.msu.edu

Neri, A., Cucchiarini, C., & Strik, H. (2008). The effectiveness of computer-based speech corrective feedback for improving segmental quality in L2 Dutch. *ReCALL, 20*(2), 225-243.

Nilsonne, Å. (1987). *Speech in depression: A methodological study of prosody.* Stockholm: Karolinska Institute.

Pennington, M., & Ellis, N. (2000). Cantonese speakers' memory for English sentences with prosodic cues. *The Modern Language Journal 84*(iii), 372-389.

Pickering, L. (2001). The role of tone choice in improving ITA communication in the classroom. *TESOL Quarterly, 35*(2), 233-255.

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

Rosenberg, A., & Hirschberg, J. (2005). *Acoustic/prosodic and lexical correlates of charismatic speech.* Paper presented at the Interspeech 2005, Lisbon.

Sjölander, K., & Beskow, J. (2000). *WaveSurfer: An open source speech tool.* Paper presented at the International Conference on Spoken Language Processing 2000, Beijing.

Strangert, E., & Gustafson, J. (2008). *Subject ratings, acoustic measurements and synthesis of good-speaker characteristics.* Paper presented at the Interspeech 2008, Brisbane, Australia.

Talkin, D. (1995). A robust algorithm for pitch tracking (RAPT). In W. B. Klejin, & Paliwal, K. K (Ed.), *Speech coding and synthesis* (pp. 495-518). Amsterdam: Elsevier.

Traunmüller, H., & Eriksson, A. (1995). The perceptual evaluation of F_0 excursions in speech as evidenced in liveliness estimations. *Journal of the Acoustical Society of America, 97*(3), 1905-1915.

Wang, X., & Munro, M. (2004). Computer-based training for learning English vowel contrasts. *System, 32*, 539-552.

Wennerstrom, A. (1994). Intonational meaning in English discourse: A study of non-native speakers. *Applied Linguistics, 15*(4), 399-421.

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