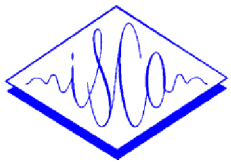
ISCA Archive
http://www.isca-speech.org/archive



Speech and Language Technology in Education (SLaTE2009)
Wroxall Abbey Estate, Warwickshire, UK
September 3-5, 2009

Using speech technology to promote increased pitch variation in oral presentations

Rebecca Hincks[1], *Jens Edlund*[2]

Unit for Language and Communication, CSC, KTH, Sweden[1]
Centre for Speech Technology, CSC, KTH, Sweden[2]
hincks@speech.kth.se, edlund@speech.kth.se

Abstract

This paper reports on an experimental study comparing two groups of seven Chinese students of English who practiced oral presentations with computer feedback. Both groups imitated teacher models and could listen to recordings of their own production. The test group was also shown flashing lights that responded to the standard deviation of the fundamental frequency over the previous two seconds. The speech of the test group increased significantly more in pitch variation than the control group. These positive results suggest that this novel type of feedback could be used in training systems for speakers who have a tendency to speak in a monotone when making oral presentations.

1. Introduction

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 [1]. 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.

1.1. Speech analysis for teaching intonation

The display of fundamental frequency has long been used to teach intonation patterns in a second language. 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 [2] established the effectiveness of giving learners audio-visual feedback on their intonation rather than audio-only. 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 [3], 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 [4].

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 [3] such as those produced by intermediate and advanced learners who make oral presentations.

1.2. 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 [5]. In studies of English and Swedish, high levels of variation correlate with perceptions of speaker liveliness [1, 6] and charisma [7, 8].

The variable that can be used to represent pitch variation is the normalized standard deviation of F0. 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 F0, the standard deviation will be low. If the speaker has raised or lowered F0 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 [9]. A number of researchers have pointed to the tendency for Asian L1 individuals to speak in a monotone in English [10, 11]. 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, they tend to strip pitch movement for discourse purposes from their production of English.

1.3. 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.

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

1. Public speakers need to use varied pitch movement to structure discourse and engage with their listeners
2. Second language speakers, especially those of tone languages, are particularly challenged when it comes to the dynamics of English pitch

These points 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? Comparisons were made between a test group that received visual feedback and a control group that was able to access auditory feedback only. Two 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.

2. Method

2.1. Base system and pitch analysis

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. 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 pitch 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 [12] and the Snack sound toolkit. As the student speaks, a fundamental frequency estimation is continuously extracted using an incremental version of getF0/RAPT [13]. The estimation frequency is transformed from Hz to logarithmic semitones. This gives us a kind of perceptual speaker normalization, which affords us easy comparison between pitch variation in different speakers.

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.

For the test students, the base system was 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. The meter has a short, constant latency of 100 ms.

2.2. Experiment

The test group and the control group each consisted of seven students of engineering, four women and three men each. The participants were recruited from English classes at KTH, 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.

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 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 researcher recorded her own (native-American speaking) 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."

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 spread out over a period of four weeks. 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. Aside from the visual feedback, all conditions were the same for the two groups.

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 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. The feedback was not present for this session. The presentation was audio recorded.

3. Results

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 [13] algorithm. Variation was calculated in a manner consistent with [1] by calculating the standard deviation over a moving 10-second window.

In the case of the training data, recordings containing noise only or that were empty were detected automatically and re-moved. 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 mean standard deviations for each data set and each of the two groups are shown in Figure 1. 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,

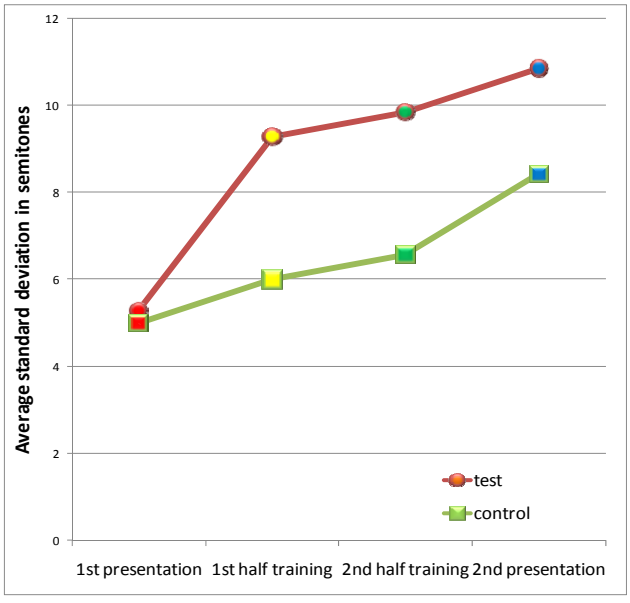


Figure 1. Average pitch variation over 10 seconds of speech for the two experimental conditions during the 1st presentation, the 1st half of the training, the 2nd half of the training and the 2nd presentation. The test group shows a statistically significant effect of the feedback they were given.

the second half of training, and the second oral presentation. The experimental group shows a greater increase in pitch variation across all points of measurement following training. Improvement is most dramatic 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. The gap between the two groups narrows somewhat in the production of the second presentation.

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 hypotheses are confirmed by these findings.

4. Discussion and conclusions

Our results are in line with other research that has shown that visual feedback on pronunciation is beneficial to learners. 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 [10, 11]. 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.

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. We have begun to explore the perceptions of the speakers, reported in forthcoming research. 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.

5. Acknowledgements

This paper is an abbreviated version of an article accepted for publication in *Language Learning and Technology* (October 2009). We acknowledge the contributions made by anonymous reviewers and the LLT editors. The technology used in the research was developed in part within the Swedish Research Council project #2006-2172 (What makes speech special?).

6. References

[1] R. Hincks, "Measures and perceptions of liveliness in student oral presentation speech: a proposal for an automatic feedback mechanism," *System*, vol. 33, pp. 575-591, 2005b.

[2] K. De Bot, "Visual feedback of intonation I: Effectiveness and induced practice behavior," *Language and Speech*, vol. 26, pp. 331-350, 1983.

[3] D. Chun, "Signal Analysis Software for Teaching Discourse Intonation," *Language Learning and Technology*, vol. 2, pp. 61-77, 1998.

[4] J. Jenkins, *The Phonology of English as an International Language: New Models, New Norms, New Goals*. Oxford: Oxford University Press, 2000.

[5] C. Johns-Lewis, "Prosodic differentiation of discourse modes," in *Intonation in Discourse*, C. Johns-Lewis, Ed. Breckenham, Kent: Croom Helm, 1986, pp. 199-220.

[6] H. Traunmüller and A. Eriksson, "The perceptual evaluation of F_0 excursions in speech as evidenced in liveliness estimations," *Journal of the Acoustical Society of America*, vol. 97, pp. 1905-1915, 1995.

[7] A. Rosenberg and J. Hirschberg, "Acoustic/Prosodic and Lexical Correlates of Charismatic Speech," presented at Interspeech 2005, Lisbon, 2005.

[8] E. Strangert and J. Gustafson, "Subject ratings, acoustic measurements and synthesis of good-speaker characteristics," presented at Interspeech 2008, Brisbane, Australia, 2008.

[9] L. D. Hahn, "Primary Stress and Intelligibility: Research to Motivate the Teaching of Suprasegmentals," *TESOL Quarterly*, vol. 38, pp. 201-223, 2004.

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

[11] A. Wennerstrom, "Intonational meaning in English discourse: A Study of Non-Native Speakers " *Applied Linguistics*, vol. 15, pp. 399-421, 1994.

[12] J. Edlund and M. Heldner, "/nailon/ -- Software for Online Analysis of Prosody," *Proceedings of Interspeech 2006* 2006.

[13] D. Talkin, " A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, W. B. Klejin, & Paliwal, K. K, Ed.: Elsevier, 1995, pp. 495-518.