

Transient visual feedback on pitch variation for Chinese speakers of English

Rebecca Hincks¹ and Jens Edlund²

¹Unit for Language and Communication, KTH, Stockholm

²Centre for Speech Technology, KTH, Stockholm

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.

Introduction

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).

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.

A number of researchers have pointed to the tendency for Asian L1 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, 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.

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 or in ... free-speech tasks did the L2 groups approach the degree of pitch increase on new or contrastive information produced by native speakers." (p. 416). This more monotone speech was particularly pronounced for the subjects whose native language was Thai, like Chinese a tone language. 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 Chi-

nese-native teaching assistants at an American university. 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 intonational 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.). The Chinese natives were speaking more monotonously than their native-speaking colleagues.

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. 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 three 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.
3. 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, 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.

Method

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 (Edlund & Heldner, 2006) and the Snack sound toolkit. 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. 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 100ms.

The test group and the control group each consisted of 7 students of engineering, 4 women and 3 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. The participants spoke a variety of dialects of Chinese but used Mandarin with each other and for their studies in China. They did not speak Swedish and were using English with their teachers and classmates.

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 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; 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. 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 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.

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 (Talkin, 1995) algorithm. Variation was calculated in a manner consistent with Hincks (2005) by calcu-

lating 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.

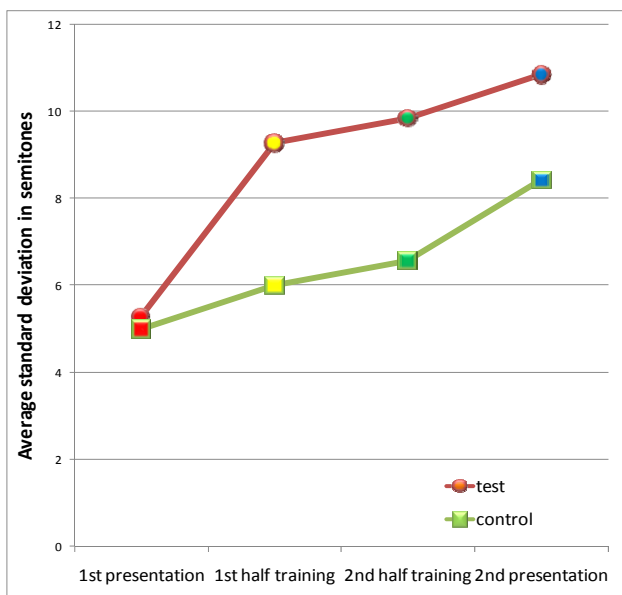


Figure 1. Average pitch variation over 10 seconds of speech for the two experimental conditions during the 1st presentation, the 1st half of the training, the 2nd half of the training and the 2nd presentation. The test group shows a statistically significant effect of the feedback they were given.

The 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, 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.

Discussion

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 (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.

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 have displayed pitch tracings of the training utterances to a third group of participants. 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 develop-

ment 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 explore perceptions of the speakers. 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. 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 (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. 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, 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. 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.

Acknowledgements

This paper is an abbreviated version of an article to be published in *Language Learning and Technology* in October 2009. The technology used in the research was developed in part within the Swedish Research Council project #2006-2172 (Vad gör tal till samtal).

References

Brazil, D. (1986). *The Communicative Value of Intonation in English*. Birmingham UK:

- University of Birmingham, English Language Research
- Edlund, J., & Heldner, M. (2006). /nailon/ -- Software for Online Analysis of Prosody. *Proceedings of Interspeech 2006*
- Hahn, L. D. (2004). Primary Stress and Intelligibility: Research to Motivate the Teaching of Suprasegmentals. *TESOL Quarterly*, 38(2), 201-223.
- 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.
- Johns-Lewis, C. (1986). Prosodic differentiation of discourse modes. In C. Johns-Lewis (Ed.), *Intonation in Discourse* (pp. 199-220). Breckenham, Kent: Croom Helm.
- Levis, J., & Pickering, L. (2004). Teaching intonation in discourse using speech visualization technology. *System*, 32, 505-524.
- 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 non-native 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.
- 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. Kleijn, & Paliwal, K. K (Ed.), *Speech Coding and Synthesis* (pp. 495-518): 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.
- 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.