Standard deviation of $F_0$ in student monologue

Rebecca Hincks
Centre for Speech Technology, Dept. of Speech, Music and Hearing, KTH, Stockholm, Sweden
Unit for Language and Communication, KTH, Stockholm, Sweden

Abstract
Twenty ten-minute oral presentations made by Swedish students speaking English have been analyzed with respect to the standard deviation of $F_0$ over long stretches of speech. Values have been normalized by division with the mean. Results show a strong correlation between proficiency in English and pitch variation for male speakers but not for females. The results also identify monotone and disfluent speakers.

Introduction
Hincks (2003) presented ideas for how language technologies could one day give support to oral production in ways analogous to current spelling and grammar checkers. I proposed that speakers preparing to make an oral presentation might rehearse not only with a stopwatch in hand but also with a recording and analysis device that could give feedback about, for example, prosody. General advice about public speaking is to use a lively voice that varies in intonation and rhythm. This is not always easy for speakers to accomplish, particularly when they are presenting in a foreign language. In a recent study, Pickering (2004) showed that non-native teaching assistants at an American university were unable to manipulate their prosody to create intonational paragraphs and that they used narrower pitch ranges than their native-speaking counterparts.

The study of intonation in discourse usually involves a fine-grained analysis of the co-occurrence of semantic content and prosodic cues. This work tells us a lot about how we use our voices to express meaning, but it does require a human presence. For the purposes of computer-assisted language learning, we would like to find fully automatic ways of giving valuable feedback. The question underlying the research reported on in this paper is whether one can use the standard deviation of $F_0$—a value easily derived from CTT’s WaveSurfer, for example—as a measure of speaker liveliness. My hypothesis was that higher standard deviations in a speaker’s speech would indicate livelier voices and thus a speaking style that was perceived as pleasant and interesting to listen to.

Previous studies of prosody in monologue, for example Bruce and Touati (1992) have looked at pitch range rather than standard deviation; see also Carlson et al. (2004). The study of pitch range is useful for looking at individual phrases, but to find out how much variation is present in the speech over long periods of time, the standard deviation is a better measure. In this paper I look at the standard deviations of $F_0$ for both 10-second segments—covering roughly five phrases, depending on the speaker—and up to a full ten minutes of speech.

It is important to acknowledge from the beginning that using an appropriate amount of pitch variation does not in itself make an appealing presentation. Rhythm and intensity should also be varied in the production of lively speech. Speakers should work to establish contact with their listeners and be aware of their body language. Most important of course is the content of the presentation: it should be well-structured, appropriate for the audience and confidently mastered by the speaker. This study thus focuses on only one aspect of the delivery of a presentation.

Data collection and analysis
Students from four different courses in Technical English at KTH were recorded as they made the oral presentations that were a graded course requirement. The recordings were made with a MiniDisc recorder and a small clip-on microphone. Other data that were collected include information about the students’ exposure to English, and for three of the four classes, written feedback on the presentation from teachers and classmates.

From a total database of 35 recordings, twenty files were chosen to meet the following requirements: all native speakers of Swedish, equal numbers of males and females, and representing a matched range of ability in English as shown by the students’ scores on the required placement test in English that all had taken before the courses started.
A maximum of ten minutes of each presentation was divided into 30-second sound files for handling and analysis. Extremely long pauses of 10 seconds or more and interruptions in the presentation were edited out. I chose 10-second long segments of speech as the unit for final analysis because it was a long enough time unit to still include quite a bit of speech even if the speaker had made a couple of pauses. The speech was segmented without regard for phrase or word boundaries.

Pitch extraction was performed using WaveSurfer with the pitch window set at 25-400 Hz for males and 25-500 for females. Each pitch analysis was visually inspected to mark the location of pitch points that were derived from sources in the classroom such as laughs, coughs, chair scrapings, and in some cases, vibrations from audiovisual equipment. When the pitch data then was transferred to a spreadsheet program, where each cell corresponded to 10 ms of speech, the cells corresponding to non-linguistic events could be edited out.

PDQ: pitch dynamism quotient
Means and standard deviations of F0 for each 10-second segment of speech were calculated. Normalization between speakers was performed by dividing the standard deviation by the mean. For purposes of discussion, we can call the value so obtained, where the standard deviation is expressed as a percentage of the mean, the pitch dynamism quotient, or PDQ.

Results and Discussion
Individual profiles
Individual PDQ values for 10-second segments ranged from a low of 0.09 to a high of 0.53. Figures 1 and 2 show examples of the development of PDQ for the course of two presentations, both by males, where the first figure represents a very strong presentation, and the second a weaker one.

The written teacher evaluation for the presentation in Figure 1 judges it to be “well-structured,” “confident,” “easy to follow,” “very coherent,” and the speech as “well-modulated” and with “varied intonation.” The graph shows a rhythmic development where up to four points at a time form inclines and declines. The pitch variation reaches a maximum two thirds of the way through the presentation, and settles down to end after the recommended ten minutes.

Speaker 63NW3, shown in Figure 2, has a more erratic development. His teacher comments point to his hesitation, lack of thematic development, the fact that he spoke for only seven minutes, and the general impression that he was not very well prepared for his talk. His graph is correspondingly less cohesive, especially in comparison to the graph in Figure 1.

Global values
Figure 3 shows the mean PDQs for each entire presentation, by speaker. Speakers are shown in order of ability in English, where the first digits in the identifying code show their result on the 100-pt placement test in English, and the final digit the course they were attending. The weaker students are thus on the left of the graph, and the stronger on the right. Males are in light grey, and females in dark grey. The mean PDQ value for all males was 0.20; the mean value for all females was 0.25.
In Figure 3 we can note that male speakers with a better command of English have higher total PDQ values than weaker students. This would be consistent with observations made by Mennen (1998) and Pickering (2004). The correlation between test result and global PDQ for male speakers is 0.47, indicating the capacity of more proficient male speakers to use more variation in F₀ as they speak in public. In contrast, the females show no correlation between global PDQ and language ability. It may be that the more Swedish-like intonation of the weaker female speakers is causing their PDQ values to be high. For example, Swedish females often employ a speaking style using high-rising terminals that are sometimes transferred to their English, and this could contribute to higher PDQ values.

Outliers

The two outliers in Figure 3 are 52VJ2 and 89EH4. The latter, a male with the second highest English score in the whole database, produces the lowest average PDQ of the group, only 0.15. Does the total PDQ value indicate a problem with his presentation? In her written feedback, his teacher commented that while his language was “fluent and easy to follow,” his delivery was “a little deadpan,” that “more animated facial expressions would be good,” and that the presentation would be improved by “showing more enthusiasm.” His profile is shown in Figure 4; note the low values and flat curve.

Speaker 52VJ2, a female, is in the opposite position to 89EH4. She is one of the weaker students, yet shows by far the highest mean PDQ value of the whole group, 0.34. Her individual profile is shown in Figure 5. Though her teacher’s feedback is not available as data, her presentation was clearly unsuccessful. She is uncertain of her subject matter and of the pronunciation of key words in the area. She uses an extremely high amount of filled pauses and re-starts, and sounds painfully uncomfortable in the position of speaking in public. 52VJ2’s high PDQ value disproves the initial hypothesis that better speakers would use more variation in their speech.

Conclusions and future work

This paper has looked at the standard deviation of F₀ from two different perspectives: as a global score based on processing up to ten minutes of speech, and as a series of values derived at 10-second intervals throughout the presentation. The global scores and graphic profiles have been examined in relation to speaker proficiency in English and teacher feedback to see how observations about intonation relate to the value I have termed the PDQ, for pitch dynam-
mism quotient. In the future I plan to test whether there is a relationship between the scores derived for 10-second segments and listener perceptions of liveliness and confidence.

An interesting finding so far has been the correlation between general language proficiency and pitch variation for male speakers. The moderate correlation of 0.47 strengthens to 0.92 if speaker 89EH4’s results are removed; this could perhaps be justified on the basis of his teacher’s comments that he clearly could have used his voice better. Other studies have noted differences in pitch range between male natives and non-natives, for example the six native speakers of Mandarin and six Americans in Pickering’s (2004) study. However, to my knowledge none has made within-group comparisons of second language users of varying proficiency. The question then arises why these results are not at all shown for females. It would be interesting to run my analysis on English speakers of other native languages to see whether similar results hold cross-linguistically. I also plan to analyze presentations made by native speakers of English to see what sorts of values they obtain. Finally, it would be very interesting to compare a single speaker making the same presentation in two different languages.

The envisioned application of this work is the speech-checker described in Hincks (2003). That paper described a number of functions such a checker could have; this paper focuses on feedback on prosody, specifically pitch variation. The initial hypothesis that speakers should try to achieve as much variation as possible was quickly disproven by seeing how disfluent passages could reach high scores. It seems more likely that there is an optimum range of values around 0.25 where speech is lively but not disfluent. It is also likely that for the purposes of an eventual feedback mechanism, this range should be adapted for speaker proficiency, allowing higher acceptable values for the more able speakers. This could be done by adding information about rate of speech, which has been used as a measure of second language ability (Hincks 2001, Townshend et al. 1998). I hope that perception tests will establish the boundaries of a range of satisfactory PDQ values, in order to be able to give precise feedback about problematic passages in the presentation. However, it seems that one can safely say that for the two outliers analyzed in this paper, the global scores are enough to give warning signals: to the high scorer: you need to practice more; and to the low scorer: you need to liven up your delivery.

Acknowledgements

This study has been funded by the Unit for Language and Communication at KTH and carried out at the Centre for Speech Technology, a competence center at KTH supported by VINNOVA (The Swedish Agency for Innovation Systems), KTH and supporting Swedish companies and organizations. My special thanks to teachers Sandra Brunsberg, Beyza Björkman and Donya Ekstrand, and to the many students who generously allowed me to record and study their speech. Further thanks to my doctoral advisors David House and Björn Granström, and to Kåre Sjölander for help with WaveSurfer.

References