Processing the prosody of oral presentations

Rebecca HINCKS
Unit for Language and Communication
Centre for Speech Technology, Department of Speech, Music and Hearing, KTH
Lindstedtsvägen 24
Stockholm, Sweden, 100 44
hincks@speech.kth.se

Abstract

Standard advice to people preparing to speak in public is to use a “lively” voice. A lively voice is described as one that varies in intonation, rhythm and loudness: qualities that can be analyzed using speech analysis software. This paper reports on a study analyzing pitch variation as a measure of speaker liveliness. A potential application of this approach for analysis would be for rehearsing or assessing the prosody of oral presentations. While public speaking can be intimidating even to native speakers, second language users are especially challenged, particularly when it comes to using their voices in a prosodically engaging manner.

The material is a database of audio recordings of twenty 10-minute student oral presentations, where all speakers were college-age Swedes studying Technical English. The speech has been processed using the analysis software WaveSurfer for pitch extraction. Speaker liveliness has been measured as the standard deviation from the mean fundamental frequency over 10-second periods of speech. The standard deviations have been normalized (by division with the mean frequency) to obtain a value termed the pitch dynamism quotient (PDQ). Mean values (for ten minutes of speech) of PDQ per speaker range from a low of 0.11 to a high of 0.235. Individual values for 10-second segments range from lows of 0.06 to highs of 0.36.

1 Introduction

Most of the research into using speech technology for language learning has focused on how technologies can be used to help beginning learners. The best technologies available to us today allow only limited feedback appropriate to a person in the initial stages of acquiring a language. In Sweden, as in other northern European countries, English is a required subject for all students for at least seven years of schooling. Beginners who are ten years old might find it fun to work or play with a computer program that teaches and assesses the sounds of English, but the large majority of these children will go on to attain acceptable pronunciation without any particular intervention or training. True beginners of English can be found in the immigrant population (Hincks 2003[a]) but most beginners in Sweden do not need speech technology to support their acquisition of the sounds of English.

This is not, however, to say that all Swedish adults pronounce English perfectly. They can have occasional problems with specific phonemes, with stress placement in multi-syllabic words, (Hincks 2003b) or with transferring the intonational patterns that have been humorously portrayed by the Swedish Chef character on the children’s program *Sesame Street.* Furthermore, in terms of their spoken English in general, they can be too reliant on the simple vocabulary familiar to them from movies and television, and hesitant to use more sophisticated synonyms when the situation calls for it, for example when making a formal oral presentation. While their English may not always be perfect, most Swedes are confident and comfortable using the language, which may be a benefit of the communicative approach that has been a part of Sweden’s language learning pedagogy for many years.

In what way, then, can this sort of second language user benefit from computer support to develop her spoken language? If we allow ourselves to think into the future, we might envision a time when our word processors contain not only spelling and grammar checkers but also speech checkers. Imagine a system in your PDA that discreetly told you after a presentation just what words you were having trouble with or where your prosody risked putting your audience to sleep. If you’d rather know about these problems before you made your presentation, imagine the speech checker as a friend in your computer who would listen as you practice your presentation and then give you a little feedback. For example, good speaker-dependent dictation systems could give you a transcript of your monologue, allowing your checker to give some feedback as to the appropriateness of the vocabulary you’ve chosen in relation to the intended audience and situation. A pronunciation-minder could tell you that you seem to be improving your command of a particular phoneme. The grammar checker could remind you to be more vigilant about agreement in the third person singular. And finally, your prosody could be processed through a speech analysis program to give you feedback about the delivery of your message. Were you perhaps speaking too quickly? Did you pause from time to time to let your words sink in? Did your words flow smoothly or did you stumble and hesitate your way through the presentation? Did your voice show enthusiasm and energy or did you yourself seem bored by the topic? This last question is the one I attempt to answer in this paper, comparing the pitch dynamism of three Swedish natives as they make ten-minute oral presentations in their courses in Technical English.

2 Corpus of student monologue

The recordings analyzed in this study are a subset of a larger corpus consisting of 35 oral presentations made by students at KTH, the Royal Institute of Technology, during the academic year 2002-03. Twenty-eight of these presentations have been transcribed, creating a corpus of approximately 32,500 words. The students were attending one of four different courses in Technical English, at three different levels, and had been put in the courses on the basis of a placement test. All recordings were made in the classroom as students fulfilled one of the course requirements. The equipment used was a Sony MiniDisc digital recorder and a small clip-on microphone. From this larger corpus, twenty files were selected for prosodic analysis. The criteria for selection were: the students should be native speakers of Swedish, there should be an equal number of females and males, and they should range in language ability from lower intermediate to advanced. Because there were recordings of only eleven females, the males were selected to match ten females on the basis of placement test scores. The mean age of these twenty students is 25.75, SD. 2.75.

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. Ten second segments of speech were chosen 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 (Sjölander and Beskow, 2000) with the pitch window set at 60-400 Hz for males and 75-600 for females. It should be noted that changing these parameters has an effect on the final values obtained; in Hincks (2004a) results with a lower pitch boundary of 25 Hz are reported. Each pitch analysis was visually inspected to mark the location of pitch points that were clearly erroneous or derived from sources in the classroom such as laughs, coughs, or chair scrapings. When the pitch data then was transferred to a spreadsheet program, where each cell represented 10 ms of speech, the erroneous cells and those corresponding to non-linguistic events could be edited out.

2.1 PDQ: Pitch Dynamism Quotient

Means and standard deviations of F_0 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. Other writers (see Traunmüller and Eriksson 1994) refer to this value as the frequency modulation factor.

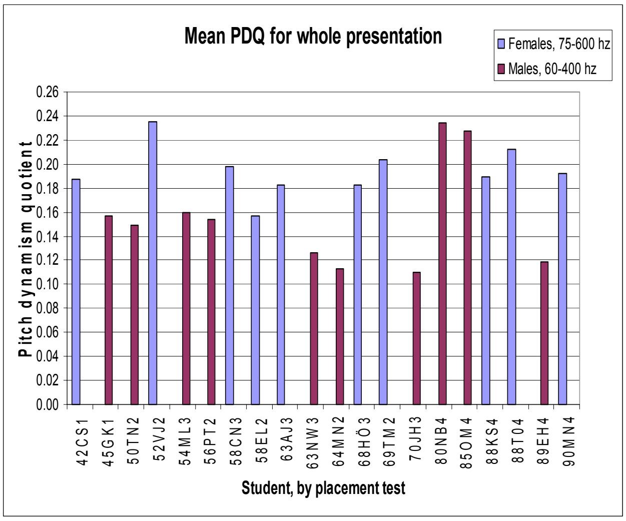

Figure 1. Mean pitch dynamism quotient for up to ten minutes of speech for each student

3 Results

Figure 1 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 are 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. Females are in light grey, and males in dark grey. In courses 1 (lower intermediate), 2 and 3 (both intermediate), females consistently achieve higher PDQ values than men. This is not the case in course 4 (advanced).

The teacher in the advanced course gave detailed written feedback on the students' presentations, containing in most cases specific comments about prosody. Figure 2 shows the variation over time of PDQ for three of these students. Each point represents the PDQ value for 10 seconds of speech, and just as in the larger corpus, values on this chart range from 0.06 to nearly 0.36.

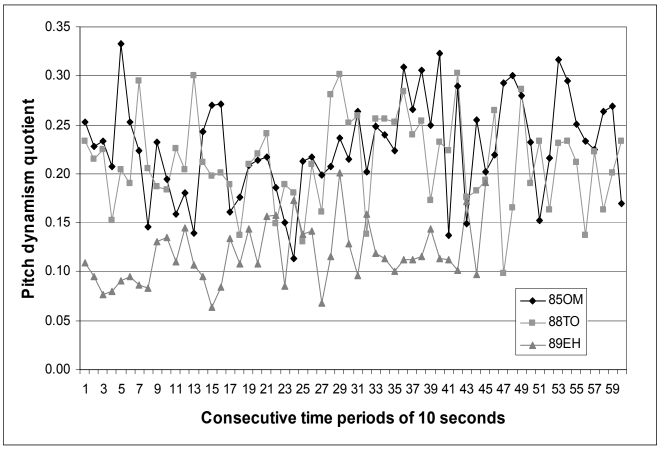

Figure 2. PDQ over time for 3 advanced students

Students 85OM and 88TO have received positive feedback on their prosody. 85OM was told he had "very good projection and modulation," "accurate projection and varied intonation," and "confident delivery, easy to hear and follow." Student 88TO was told she spoke clearly, with a good accent, and that her presentation was "well-rehearsed" and "professional." Student 89EH, however, whose PDQ line moves along the bottom of the graph and who has a mean PDQ value of just over half that of 85OM, was told that his delivery was "a little deadpan" and that his presentation would be improved by showing "more enthusiasm." A typical pitch contour for 89EH in the window of WaveSurfer (Figure 3) appears as a series of nearly flat lines, while a typical contour for his classmate 85OM's rises and falls (Figure 4).

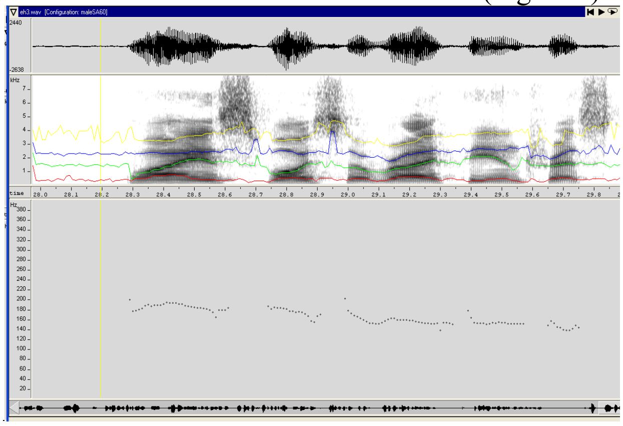

Figure 3. Waveform, spectrogram and pitch contour for a typical utterance by 89EH. Utterance: "Why is voice over IP interesting?"

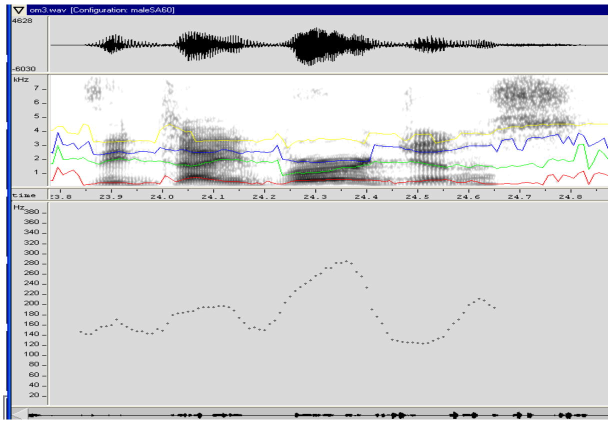

Figure 4. Waveform, spectrogram and pitch contour for a typical utterance by 85OM. Utterance: "the divergence"

4 Discussion and Conclusion

This preliminary analysis of advanced language users, linking one researcher's and one teacher's perceptions of what is appropriate intonation for a short oral presentation, can form the basis for a hypothesis that PDQ values in the neighbourhood of 0.10 characterize nearly monotonous speech, while PDQ values around 0.25 characterize lively and engaging speech. How then could these results be applied to language learning?

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. She encouraged the use of theater exercises to train non-native teaching assistants to "increase confidence" and "explore their voice range." Another way of doing this would be for the teaching assistants to record themselves as they teach and then receive automatic feedback from a pitch extraction program. Pickering's subjects were speakers of Mandarin, who face greater intrinsic challenges mastering English intonation than do speakers of Swedish. All the same, anyone making a presentation or teaching in a second language is likely to be less confident than when using a native language. This lack of confidence can manifest itself in a narrowed pitch range (Mennen 1998). Practicing a presentation with automatic feedback could be one way of raising awareness of the problem and encouraging a more effective use of one's voice; even native speakers could benefit from this kind of feedback. In this context it would be very interesting to ask speakers to perform a presentation twice, once in English and then again in their native language, and compare the results

Some of the high PDQ values obtained by the less proficient speakers in the study raise questions about the contribution of the speakers' emotional state and of native language intonational patterns. Here it can be noted that the high value obtained by speaker 52VJ, a less proficient speaker (see Figure 1), is from a very disfluent and painfully performed presentation, where frequent restarts and hesitation over words account for the large standard deviation, rather than an enthusiastic use of focal accent. 52VJ's values, like those of other students, are even higher when the pitch parameters are lowered to accept frequencies down to 25 Hz, because of the growl-like creaky vocalizations that accompany her mental search for the words to express her thoughts. A possible way of distinguishing high PDQs that come from disfluencies from high PDQs that come from lively speech would be to look at the discrepancies between the values derived from different pitch extraction settings in WaveSurfer. In other words, a possible pitch processor could make two or more passes to look at how much of the standard deviation came from low or high frequencies. A high proportion of low frequency sounds could be a sign of disfluency.

The contribution of native language intonational patterns is perhaps a trickier research question. Perception tests are being planned to gather ratings about accentedness, confidence and liveliness to determine how these characteristics correlate with PDQ. Another aspect to take into consideration is rate of speech, which has shown to be more strongly correlated than pitch variation with perceptions of liveliness (Traunmüller and Eriksson, 1995).

Finally, it is important to acknowledge 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 all 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.

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. 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, to Kåre Sjölander for help with WaveSurfer, and to Preben Wik for building PDQ-izers.

References

R. Hincks. 2003a. Speech technologies for pronunciation feedback and evaluation. *ReCall* 15(1):3-20.

R. Hincks. 2003b. Pronouncing the Academic Word List: Features of student oral presentations. *Proceedings of the 15th International Congress of Phonetic Sciences*, Universidad Autonòma, Barcelona 1545-1548.

R. Hincks 2004a. Standard Deviation of F_0 in student monologue. *Proceedings of Fonetik 2004,* Dept. of Linguistics, Stockholm University, 132-135.

I. Mennen, 1998. Can language learners ever acquire the intonation of a second language? *STiLL 98 Proceedings* Dept of Speech, Music and Hearing, KTH, 17-20.

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

K. Sjölander and J. Beskow 2000. WaveSurfer : an open source speech tool. *Proc of ICSLP 2000.*

H. Traunmüller, and A. Eriksson, 1994. The frequency range of the voice fundamental in the speech of male and female adults, Manuscript, Department of Linguistics, University of Stockholm, (Accessed May 8 2004) http://www.ling.su.se/staff/hartmut/aktupub.htm

H. Traunmüller, and A. Eriksson, 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.