Available online at www.sciencedirect.com
SCIENCE @ DIRECT®

System 33 (2005) 575–591

SYSTEM

www.elsevier.com/locate/system

Measures and perceptions of liveliness in student oral presentation speech: A proposal for an automatic feedback mechanism

Rebecca Hincks

Department of Speech, Music and Hearing, The Royal Institute of Technology (KTH), SE100 44 Stockholm, Sweden

Received 19 January 2005; received in revised form 18 April 2005; accepted 24 April 2005

Abstract

This paper analyzes prosodic variables in a corpus of eighteen oral presentations made by students of Technical English, all of whom were native speakers of Swedish. The focus is on the extent to which speakers were able to use their voices in a lively manner, and the hypothesis tested is that speakers who had high pitch variation as they spoke would be perceived as livelier speakers. A metric (termed PVQ), derived from the standard deviation in fundamental frequency, is proposed as a measure of pitch variation. Composite listener ratings of liveliness for nine 10-s samples of speech per speaker correlate strongly ($r = .83$, $n = 18$, $p < .01$) with the PVQ metric. Liveliness ratings for individual 10-s samples of speech show moderate but significant ($n = 81$, $p < .01$) correlations: $r = .70$ for males and $r = .64$ for females. The paper also investigates rate of speech and fluency variables in this corpus of L2 English. An application for this research is in presentation skills training, where computer feedback could be provided for speaking rate and the extent to which speakers have been able to use their voices in an engaging manner.
© 2005 Elsevier Ltd. All rights reserved.

Keywords: Prosody; Intonation; Pitch; Speech rate; Oral presentation skills; CALL; Speech analysis; English for specific purposes; Fluency; Learner corpora

E-mail address: hincks@speech.kth.se.

0346-251X/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.system.2005.04.002

1. Introduction

Any number of popular manuals on public speaking (e.g., Lamerton, 2001; Grandstaff, 2004) advise speaking with a lively voice that varies in intonation. The word 'lively' means, if one consults Merriam-Webster (online), 'briskly alert and energetic' and 'imparting spirit and vivacity.' Sinclair (1995) associates liveliness primarily with enthusiasm. According to the manuals, a lively voice is achieved by consciously modifying the three prosodic dimensions of loudness, pitch and tempo. Intonational modification helps the audience understand the content of the message. By pausing before moving to a new point, for example, and then raising pitch as one starts to speak, a speaker helps listeners orient themselves in the flow of information. An important side effect of helping listeners in this way is the maintenance of listener focus on the message, so that their attention does not wander. Lively speakers should also avoid following one intonational pattern utterance after utterance, and include a visual dimension, with the contribution of facial and body gestures.

Using one's voice well is not the absolutely most critical aspect of making a good presentation; obviously it is also important that the content is well-structured, appropriate to the audience, and clearly explained. Yet if the speaker does not use his or her voice in a way that facilitates access to the content, much of the message can be lost. It is surprising that the area has attracted little academic interest; a plausible reason for this is that it is only recently that technological development has allowed smooth processing of large amounts of recorded speech.

The quantity of self-help books on presentation skills on the market today[1] is testimony both to the demands put on oral communication in contemporary workplaces and to people's lack of preparation to meet these demands. Not every graduate has had the opportunity to take a course in speaking skills. Speakers who turn to self-help manuals are told to practice on their own or with a friend. In his chapter entitled "Improving academic and medical presentations" one expert (Grandstaff, 2004) gives this advice:

> "If you are not sure whether you spoke in a monotone, record yourself and listen for how much variety you use in your voice as well as whether you are speaking faster or slower than speakers you enjoy. Ask a friend or colleague to listen to the tape and give suggestions about how you can add interest and variety to your voice. Practice by varying your pitch, pace and volume. Make the variety fit what you are saying. Emphasize key words. Pause to add impact and to allow time for people to take in what you have said. Increase or lower your volume slightly to draw attention to key words or phrases." (p. 237)

This paper suggests that a computer could fill the role of friend or colleague and give automatic, objective and valuable feedback on speaker prosody. Speech analysis software, which has been used for the past 25 years to help second language learners visualize the ways in which their intonation deviates from a target model (Bot and Mailfert, 1982; Molholt, 1988; Anderson-Hsieh, 1992; Öster, 1998; Hardison, 2004), can also be used to gather raw data about pitch variation, speaking rate and pausing. The question underlying the research reported here is whether such data can be used to characterize speaker liveliness. If it could, then a potential new application for speech analysis software would be as a feedback

[1] In November 2004, Stockholm's largest bookstore had three times as many shelves full of books on advice for public speaking as on writing.

or evaluative mechanism for public speaking. The paper focuses exclusively on the variables that can be collected and processed automatically and online; that is, without reference to propositional content. In other words, I will not address, at this point in technological development, Grandstaff's advice to 'make the variety fit what you are saying.' However, I feel that many speakers would still benefit by the kind of feedback I am proposing.

Public speaking difficulties are magnified for second language users, who are operating under a heavy cognitive load of planning lexical content and its articulation at the same time as they may lack confidence and familiarity with the potentialities of spoken academic English. A number of recent works (Ventola et al., 2002; Rowley-Jolivet and Carter-Thomas, 2005) have pointed to the lexical, syntactical and pragmatic difficulties faced by non-native speakers who are presenting or teaching in English. An increasing number of studies have addressed the important issue of non-native prosody in instructional speech (Hahn, 2004; Levis and Pickering, 2004; Pickering, 2004). Pickering (2004) compared the way native and non-native teaching assistants used intonational paragraphing (Brazil, 1997) in the presentation of laboratory instructions. The non-native speakers showed "a considerably weaker control of intonational structure and a disturbance in prosodic composition that materially affects the comprehensibility of the discourse for native speaker hearers" (Pickering, 2004, p. 19). One of the contributing problems was an overall narrower pitch range, which made the identification of prosodic units difficult. International teaching assistants are responsible for a large amount of undergraduate instruction at many North American universities, and their communication difficulties are a serious problem. Yet, as Pickering notes,

> "little may be done to. . .address areas of linguistic competence such as pitch range or pause structure, as they are often perceived to be less crucial for functional competence than lexical or syntactic marking strategies... However, prosodic cues contribute independently to the structure of the discourse, and they cannot be circumvented without a reduction in comprehensibility" (Pickering, 2004, p. 39).

Ideally, an instructor or speaker who faces such challenges would be given individualized expert instruction in how to improve his or her prosody in relation to the content of the message. Since this kind of coaching is probably not realistic in most university settings, an alternative would be participation in a course in speaking and presentation skills. Pickering suggests that such courses could benefit from the introduction of exercises designed for theater training as a potential remedy for the problem of restricted pitch range (ibid p. 39), and Levis and Pickering (2004) suggest using speech analysis for visualization of pitch contours at the discourse level.

As Levis and Pickering point out, speech visualization has too often been used to practice intonation only at the level of phrases or short utterances. When a learner does this, he or she is imitating the pitch contour of a specific speaker. This entails adopting that speaker's attitude and dialect. Such imitation might be helpful to beginning learners, but can be of limited benefit in the long run. Jenkins (2000) claims that speakers of international English do not need the kind of intonational training designed to make them sound like natives, though speakers must master the placement of focus[2] in an utterance. This finding

[2] Both Jenkins and Hahn use the terms 'nuclear stress' and 'primary stress' for what other scholars call 'focus'. Following Wennerstrom (2001) and others I reserve the terms 'stressed' and 'unstressed' to describe the relationships of syllables to each other at the lexical level, and 'focus' to describe relationships of syllables at the utterance level.

was corroborated by Hahn (2004). The system described here would not be able to determine whether the placement of focus was correct or not, but it would help point to whether the speaker was expressing focus at all. In Hahn's research, subjects were presented with non-native instructional speech in three conditions: delivered with normal sentence focus, abnormal sentence focus, and without focus (monotone). Students comprehended and recalled more information from the correctly focused delivery, and in evaluative comments, were most critical of the focus-less delivery. Interestingly, 30% of the students who listened to the focus-less delivery thought that the speaker spoke too fast, though speaking rate and pausing were tightly controlled and no subjects who had heard deliveries with focus commented on speaking rate.

There is a further, practical reason that speech analysis has been used to visualize only short utterances: if the pitch contour of a very long utterance is shown on one computer screen, many important details and movements are lost by being compressed. Therefore, this paper advocates looking at the distribution of the pitch data only, without visualization. The proposed feedback mechanism processes large amounts of pitch data in terms of the standard deviation of the fundamental frequency in order to detect the degree to which the speaker was varying his or her pitch over long stretches of discourse.

Ultimately, one can envision a feedback mechanism for presentation speech that incorporates speaker-dependent speech recognition to recognize, and process at some level, the linguistic content of the presentation or lecture. Using natural language processing, the instant transcript could be textually analyzed for features that have been deemed appropriate for the speaking genre in question. The recent attention paid to spoken academic English has helped our understanding of the lexical and syntactic properties of successful monologue (Camiciottoli, 2003; Simpson et al., 2003; Morell, 2004; Rowley-Jolivet and Carter-Thomas, 2005). Like grammar checkers that can be set to flag certain stylistic features of written texts, this feedback mechanism could check for the presence of personal pronouns (a positive feature in monologue (Morell, 2004)) or passive constructions (a negative feature, indicating difficulties in adapting the information structure of a written text to a format suitable for spoken discourse (Rowley-Jolivet and Carter-Thomas, 2005)). Speech recognition can also be used to give feedback on pronunciation at the segmental level (Eskenazi, 1999; Neri et al., 2002; Hincks, 2003a). However, speech recognition for the purposes described here would need to be adapted to the individual speaker's voice, a time-consuming process that does not lend itself to classroom applications (Coniam, 1999).

Fig. 1 illustrates how an automatic feedback mechanism could consist of two parallel processing operations, one conducted by the recognizer, and the other by speech analysis. This paper discusses the features in bold text, that is, pitch variation and speech rate. An appropriate and friendly feedback interface would be an animated face that could respond alertly to lively speech but would lose attention, perhaps even fall asleep, if the prosody failed to show any characteristics of liveliness. In the more distant future, a feedback mechanism could also incorporate a camera and software for processing speaker gaze, facial expression and body language.

An application of this kind places the computer in a supportive rather than a tutorial role (e.g., Levy, 1997). The system would not presume to correct the user, but merely act as a tool for quantifying the amount of prosodic variation. This allows the computer to do what computers have been proven to do well, which is to facilitate and support human communication, and avoids the pitfalls associated with the artificial intelligence required for tutorial systems.

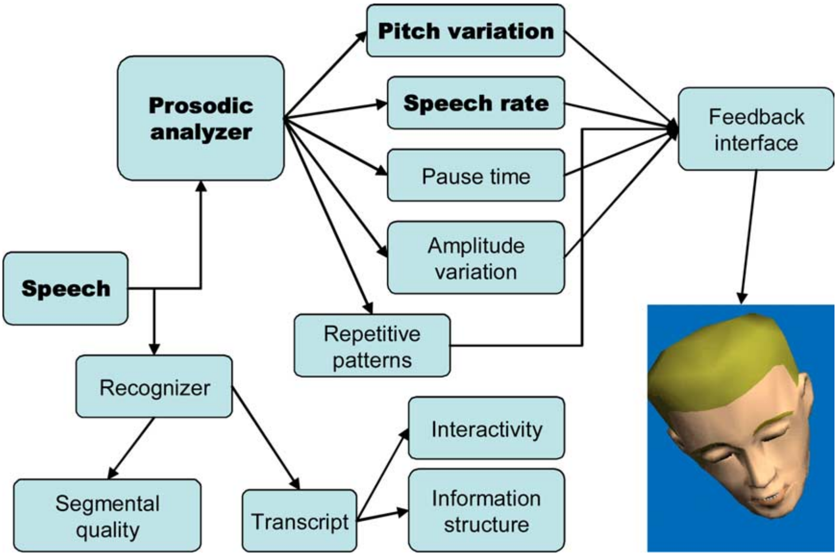

Fig. 1. Schematic design of an automatic feedback mechanism for public speaking. Features in bold are treated in this paper.

This investigation thus asks whether a metric created from data automatically derived from speech analysis software can be used to describe the degree of speaker liveliness. The hypothesis tested is that the higher the standard deviation in fundamental frequency, the more a speech sample will be perceived as lively. This has been shown in earlier work using synthesis of a single utterance (Traunmüller and Eriksson, 1995), but has not been tested on naturally occurring speech.

2. Method

2.1. Material

The material under investigation comes from a corpus consisting of 35 ten-minute oral presentations made by undergraduate engineering students, from four different courses and three different proficiency levels of Technical English.[3] Many of these students expect to find employment in companies whose official language is English, and find instruction in English presentation skills to be a valuable part of their education. All students had taken an in-house placement test to determine their English proficiency, and all had studied English for at least six years in the regular Swedish school system, where great emphasis is placed on oral competence (Oscarson, 1995; Erickson, 2004). The students were audio recorded, with their written permission, in the classroom as they fulfilled a major, graded

[3] An analysis of the lexis and of the pronunciation errors found in part of this corpus was published as Hincks (2003b). That study found a very low frequency (less than 0.05% of words) of pronunciation errors that were likely to either impede intelligibility or be negatively perceived.

requirement in their courses in Technical English. The equipment used was a MiniDisc recorder and a small clip-on microphone.

The prosodic analysis has been performed on a sub-corpus of 18 recordings. The criteria for inclusion in this sub-corpus were the student's sex, first language, and score on the in-house placement test in English. The goal was to have gender-balanced groups of six intermediate, six upper-intermediate, and six advanced students, all native speakers of Swedish. Because there were only eleven recordings of females (reflecting the gender balance at a college of engineering), the males were chosen to match in score the nine females who met the score requirements. The mean age of the students was 25.5, SD 2.6. The in-house placement test had a maximum score of 100 and measured ability in writing, grammar and vocabulary. Group A had scores from 50 to 60, Group B from 61 to 70, and Group C from 80 to 90. For the purposes of comparison with other student groups, it is useful to provide estimations of student competence in the Council of Europe's Common Framework of Reference (Council, 2001). The students in groups A and B could probably be placed in Council group B2, and the students in group C in the upper ranges of Council group C1.

The presentations were transcribed using regular English orthography, and the sound files digitally stored using 16 kHz sampling. In preparation for prosodic analysis, interruptions in the presentations due to equipment problems, pauses of 10 s or more, or question-and-answer segments were edited away.

2.2. Pitch extraction and variation

To enable smoother handling of the large quantity of data, the 7–10-min long recordings were divided into 30-s segments stored as separate sound files. Each file was then processed using the speech analysis function of the program WaveSurfer (Sjölander and Beskow, 2000). WaveSurfer was configured to search for pitch (fundamental frequency) at between 60 and 400 Hz for the male voices and between 120 and 500 Hz for the female voices. The pitch contour produced for each file was visually inspected for evidence of miss-readings, and the location of these errors noted. WaveSurfer extracts a pitch value for every 10 ms of speech; that is, 100 values per second. These values were imported into a spreadsheet program for further analysis.

The next steps in the analysis were to delete from the spreadsheet program all values of zero (from unvoiced segments or silence), and all values that corresponded to errors in the pitch extraction (as evidenced by the visual inspection). Then, for each 10 s of speech, the mean and standard deviations of the pitch were calculated. Ten seconds of speech was chosen as a good unit for data analysis because it was enough time to guarantee the inclusion of a fair amount of speech at normal pausing rates.

The raw standard deviation of the pitch is unsuitable in itself as a measure of pitch variation (Traunmüller and Eriksson, 1995). This is because of the differences among speakers, and particularly between sexes, of the mean pitch level. The higher the frequency of our voices, the larger the standard deviation will be in normal speech, and this would give an 'unfair' advantage to female speakers over male. Therefore, in order to make valid comparisons among speakers, the standard deviation is expressed as a percentage of the mean. For example, a standard deviation of 21 and a mean frequency of 115 (a male voice) yields a value of 0.183; a standard deviation of 36 and a mean frequency of 195 (a female voice) yields a value of 0.185. To simplify expression, I have termed this value the PVQ, for pitch variation quotient.

With this method, the PVQ was calculated for every 10 s of speech for up to 10 min of speech for each of the 18 students. This yielded a database of 986 values for the entire subcorpus.

2.3. Perception test

A perception test was prepared to test the hypothesis that speech with higher PVQ would be perceived as livelier speech. Nine 10-s samples of speech per speaker, representing the speaker's three lowest PVQs, three mean PVQs, and three peak PVQs, were selected as test files. If any of these nine files contained a pause that was longer than 4 s, it was substituted by another for two reasons: one, that the PVQ value was less stable when it represented less speech, and two, that judges should not be rating a speech sample that consisted of nearly 50% silence. Each speaker was thus represented by one and a half minutes of speech in nine separate test files, giving 81 test files to rate for each sex.

Separate tests were prepared for male and female speakers using Visor from the Spruce package (Granqvist, 2003). Respondents were presented with a randomized collection of test file icons on a computer screen and were instructed to listen to each file and then move it to an undivided scale whose endpoints were 'lively' and 'monotone'. They could listen to the files as many times as they wanted to and move them as many times as they wished. They were instructed to disregard impressions of speaker proficiency and focus on qualities of engagement and liveliness, but no judge was aware that it was pitch variation that was being tested. The tests took between one half hour and one and one half hours to complete, and respondents completed the male and female tests on different days, in randomized order. The respondents were eight (two male) university teachers of English, who were natives of Britain (2), Sweden (2), USA, Brazil, Germany and Turkey. None of the teachers were specialists in teaching pronunciation, but three of them had extensive experience in teaching presentation skills.

2.4. Speaking rate

Speaking rate is often expressed in words per minute (WPM). This is an imprecise measurement for a number of reasons (Griffiths, 1990, 1991; Griffiths and Beretta, 1991; Kormos and Dénes, 2004). In contrast, expressing speaking rate in syllables per second (SPS) gives a number of advantages. First, it provides a fair comparison between speakers who use long words versus those who use shorter words. Second, it allows cross-linguistic comparisons between languages with different average word length. Third, it provides a more local measurement so that variations in speaking rate can be tracked. Finally, to calculate WPM one needs a transcript of the event. Since syllables can be characterized as bursts of acoustic energy corresponding to the syllabic nucleus, their number can be counted, or at least reliably estimated, on the basis of the speech waveform without knowing what has been said. For this research, however, a manual rather than automatic method was used for calculating SPS.

Another relevant variable, particularly when analyzing L2 speech, is mean length of runs (MLR). This is the number of syllables the speaker has uttered between pauses. In this paper, as in Kormos and Dénes (2004), a pause is defined as a silent interval longer than 250 ms, or one quarter of a second.

Speaking rates in WPM and SPS, as well as MLR were calculated for the entire presentation of each of the 18 speakers.

3. Results

There are thus a number of variables to take into consideration. For each speaker, there is a value representing proficiency in English, a value for mean pitch variation throughout the presentation, values for mean speech rate and mean length of runs, and finally the mean of the nine liveliness ratings per speaker. These values can be used to characterize the speakers and their presentations. In addition, the liveliness ratings for 162 individual 10-s samples of speech can be seen in relation to PVQ, speech rate, and mean length of run within that sample.

3.1. Pitch variation results

PVQ values for all individual 10-s samples in the corpus are shown in Fig. 2. For both males and females, the lowest values in the corpus are about 0.06 and the values follow each other nearly exactly up to 0.157, where they diverge, with the male values higher than the female. The maximum values for males are above 0.30, while female values reach just above 0.25.

Fig. 3 is an example of the PVQ development over the whole presentation of two male speakers from group C. The speech of speaker 14, whose values vary greatly with 0.23 as the mean, was described by his teacher in written comment on the presentation as being

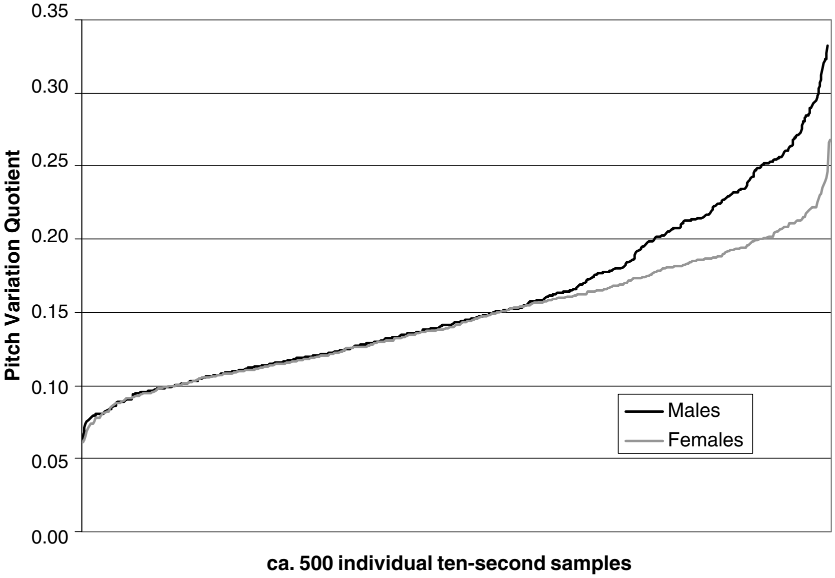

Fig. 2. Distribution of all PVQ values for all 10-s samples in presentation corpus. 492 samples for males and 494 for females.

"well-modulated" and with "varied intonation." In contrast, the speech of speaker 17, whose values range mostly between 0.10 and 0.15, was described as being "a little deadpan."

The relationship between the speakers' proficiency in English and mean pitch variation in the presentation as a whole is shown in Fig. 4, which plots mean PVQ per speaker against the speaker's score on the in-house placement test. For female speakers, there is a significant correlation between pitch variation and proficiency ($r = .83$, $n = 9$, $p < .01$), but for the males the relationship is not significant.

3.2. Perception test results

3.2.1. Inter-rater reliability

A reliability analysis performed on the results of the perception test gave high values for Cronbach's alpha: .98 and .95 for the composite judgments of male and female speakers, respectively, and .94 and .90 for male and female speech samples.

Many of the judges commented that they found the liveliness rating task easier for the male speech than for the female speech, and this is shown in the results of the perception test. Table 1 shows the correlations between each judge's liveliness ratings per speaker (the mean of the ratings of the nine samples per speaker) and the means of the PVQs of the speech that was rated. The table also shows the correlations between the liveliness ratings per speaker and the speaker's score on the English proficiency test. The ratings of male speakers reach a higher level of correlation with PVQ than the ratings of female speakers do. The differences between perceptions of males and females are even more striking when it comes to the correlations with the student's score on the English proficiency test. The judges appear to have succeeded with the instruction to ignore questions of proficiency when rating the male speakers, since for all judges correlations with proficiency are lower

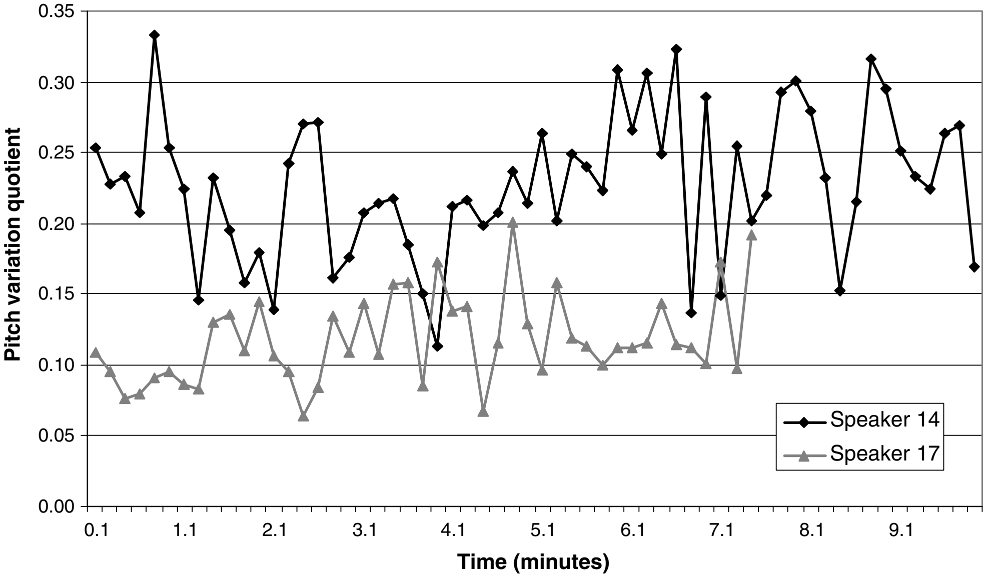

Fig. 3. The development of PVQ over time for two males from Group C.

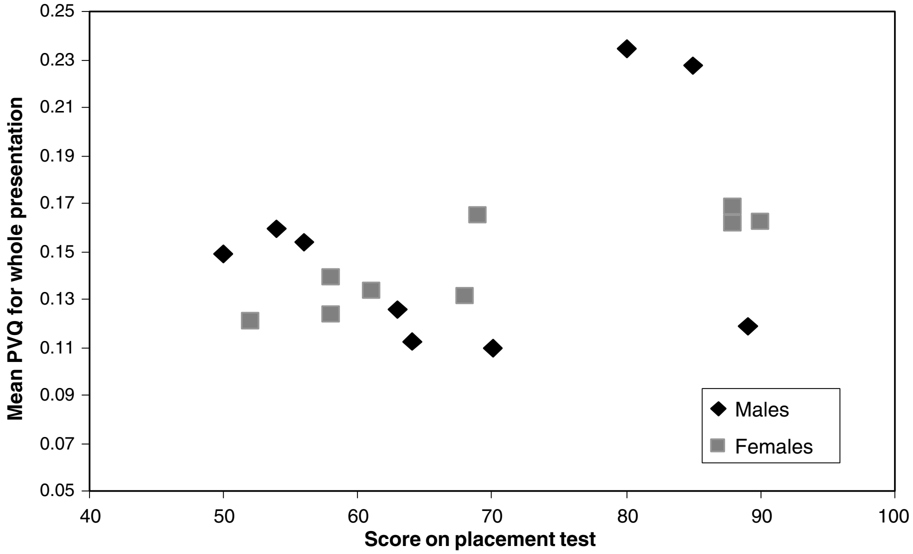

Fig. 4. Mean pitch variation quotient per speaker in relation to proficiency in English as measured by in-house placement test.

Table 1
Pearson correlations between liveliness ratings of speakers and both PVQ and English proficiency as measured by placement test, by judge

Judge	Pitch variation		English proficiency	
	Males	Females	Males	Females
1	.86**	.77**	.47	.87**
2	.95**	.75**	.40	.84**
3	.88**	.38	.52	.57
4	.91**	.68*	.41	.72*
5	.87**	.61*	.63*	.71*
6	.88**	.69*	.59*	.80*
7	.68*	.59*	.32	.56
8	.75*	.67*	.60*	.83**

$n = 9$ males, 9 females.
* $p < .05$, one tail.
** $p < .01$, one tail.

than with pitch variation. The reverse is true for the ratings of female speakers, where for seven of the judges, correlations with proficiency are higher than with pitch variation. This could be partly due to the fact that for the females in this corpus there was also a stronger relationship between PVQ and English proficiency, as shown in Fig. 4.

3.2.2. Perceptions of speakers

Fig. 5 plots means of the PVQs per speaker against the means of the liveliness ratings of all judges per speaker. The program used to produce the perception test uses a visual scale that transforms the placement of an icon on a screen to values between 0 and 1000, which is the scale used here on the x-axis. Males are shown with filled symbols and females with

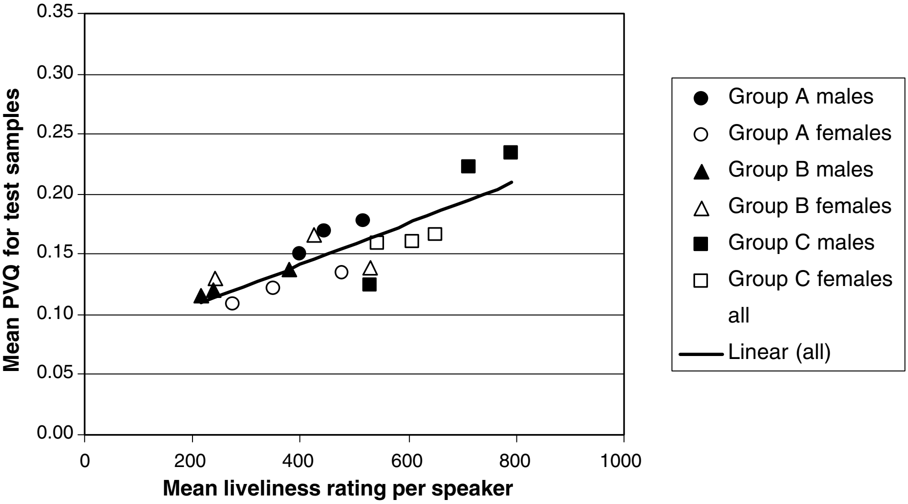

Fig. 5. Mean pitch variation quotient for all test samples per speaker plotted against mean liveliness rating per speaker.

unfilled symbols, with different shapes for the different groups. The correlation between the values is significant ($n = 18$, $p < .01$) and strong: $r = .83$, indicating that the PVQ metric is a reliable indicator of speaker liveliness.

3.2.3. Perceptions of speech samples

The composite of the liveliness ratings for the nine speech samples per speaker gives a more reliable characterization of speakers than what can be discerned from the perception of liveliness in a single 10-s sample of speech, but even at this level the correlations are also significant ($n = 81$, $p < .01$) for both sexes. The mean liveliness ratings for all individual test files are plotted against PVQ in Fig. 6.[4] Once again correlations for males are higher than for females: .70 for the males and .64 for the females. Generally, speech samples with low PVQ occupy the lower left quadrant and samples with high PVQ occupy the upper right quadrant, most noticeably a group of six samples from males of Group C. The lower right quadrant contains mostly squares (the most proficient group), indicating that judges perceived them as lively even when their PVQs were not high. In contrast, speakers from groups A and B occupy the upper left quadrant, indicating that high pitch variation was not always perceived as liveliness for these less fluent speakers.

3.3. Temporal measurements

Table 2 shows means and standard deviations per student proficiency level for three different temporal measures: mean length of runs (MLR), words per minute (WPM), and syllables per second (SPS). The advanced students, Group C, spoke more quickly and produced more speech between pauses than the two intermediate groups. The mean

[4] One of the female test files with an outlying position was re-examined, was found to have violated the pause length criterion, and its results were removed. Another two female test files were discovered to be duplicates and the results for one of them removed.

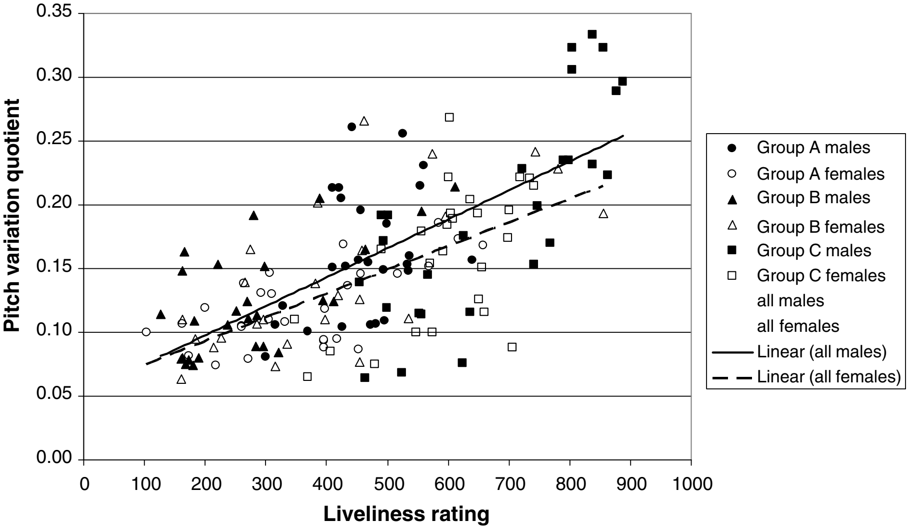

Fig. 6. Pitch variation quotient for 160 10-s samples of speech versus mean liveliness rating from panel of eight judges.

temporal measures correlate perfectly with each group's mean score on the written placement test, supporting earlier findings on the relationship between speaker proficiency and speaking rate (Towell et al., 1996; Cucchiarini et al., 2000; Kormos and Dénes, 2004).

The mean results for MLR, WPM and SPS presented in Table 2 have been calculated on nearly an hour of speech per student group. In addition, local values were calculated for each of the 160 ten-second test samples used in the perception test, in order to examine the extent to which the liveliness ratings correlated with rate of speech. Table 3 shows correlations between the liveliness rating and two temporal measures as well as with PVQ. These results reveal interesting differences between the raters' perceptions of liveliness in male and female speakers. While both temporal measures had no correlation with the perception of liveliness in male speakers, the correlation between MLR and liveliness in the female test samples was .72, higher than the correlation of liveliness with PVQ, which for females was only .64 for individual samples. Plain speaking rate as measured in SPS had no effect on liveliness judgments for either sex.

4. Discussion

The PVQ variable is a very good though imperfect correlate of the perception of liveliness in student presentation speech. The following sections look at the factors that could contribute to moderating the correlations between perceived liveliness and PVQ.

Table 2
Mean values of temporal measures for each student group

Group	*n*	Placement test mean (SD)	MLR mean (SD)	WPM mean (SD)	SPS mean (SD)
A	6	54.7 (3.3)	7.10 (1.34)	115 (11)	2.75 (.25)
B	6	65.8 (3.7)	8.02 (1.11)	118 (11)	2.86 (.26)
C	6	86.7 (3.7)	9.80 (2.12)	128 (18)	3.14 (.36)

Table 3
Pearson correlation coefficients between three prosodic variables and liveliness rating for individual speech samples

	SPS	MLR	PVQ
Males	.06	−.06	.70**
Females	.16	.72**	.64**

$n = 81$ male samples, 79 female samples.
** $p < .01$.

4.1. Causes of high pitch variation

Many of the high PVQ files were examples of speakers successfully using pitch contrasts to structure the content of their presentation when introducing items in a series or moving from an old topic to a new one. High PVQ was also evident when speakers chose to illustrate points on a blackboard, and of course when they showed extra enthusiasm for their topic. Samples of this nature generally received correspondingly high liveliness ratings. On the other hand, some of the samples with high PVQ were due to speakers using large pitch variation for other reasons. For example, pitch resets caused by disfluencies or nervousness could lead to high PVQ values. A few of the speakers, most of them female, had a habit of speaking with high rises at the end of an utterance,[5] which would contribute to a high PVQ though the speech could well be interpreted as uncertain. All of the speakers bore traces of Swedish prosodic patterns to one degree or another, though this was most evident in the lower-proficiency speakers. Because Swedish has distinctive word accents signaled by pitch movement within one word, the presence of these patterns could contribute to a high variation in pitch that was not necessarily rated as lively by the judges. Furthermore, some of the samples with high PVQs contained Swedish names uttered as they should be intoned in Swedish (with large pitch movement); neither would these files receive high liveliness ratings. It is likely that some of the high PVQ files were perceived as being accented rather than lively, though experimental tests would be required to confirm this observation.

4.2. Causes of high liveliness ratings

The five test files that received mean liveliness ratings of above 600 and yet had PVQ below 0.13, found in the lower right quadrant of Fig. 6, all came from the highly proficient speakers of Group C. Four of these files had been selected for the test as examples of that speaker's lowest PVQs; in giving them relatively high ratings, the judges may have been responding to a sort of 'halo effect' where they rated a file highly because they had rated other files from that speaker highly.

4.3. Sex differences

Below the PVQ value 0.157, males and females in the corpus show a nearly identical distribution of values (Fig. 2). Above this point, the values diverge, with males showing

[5] Unlike North American 'uptalk,' however, these high rises found in some Swedish speakers do not convey the impression that a declarative is a question. The rise is more like a plateau.

more pitch variation than females. Interestingly, this point appears again in Fig. 6, where all the PVQ values are plotted against the mean of the liveliness ratings given to them by the panel of eight judges. The intersection of the mid-value 500 with the midpoint between the regression lines of the males and females is very close to the same point at which the male and female values diverge. The judges are indicating that PVQ values higher than 0.157 are perceived as being on the upper end of the lively-monotone scale, and the database indicates that males had a better ability than females to achieve these higher values in their speech. Furthermore, Fig. 6 shows that the highest male PVQ values received correspondingly high liveliness ratings, while the highest female PVQ values did not receive the highest liveliness ratings. The data shown in Table 3 also indicate that males and females may differ in the production and perception of liveliness. For 81 test files, male liveliness ratings show no correlation whatsoever with the fluency of the speech, as measured in mean length of runs, or how many syllables the speakers uttered between pauses >250 ms. Female liveliness ratings, on the other hand, correlate moderately with MLR, more strongly than with the PVQ variable. In other words, the raters may have judged the female speakers on how fluent they were, but not the males.

These possible differences in the perception and production of liveliness in male and female speech are an unexpected result. Traunmüller and Eriksson (1995) found no effect of speaker identity on liveliness perceptions in their study of synthetic speech. However, studies of natural speech have drawn different conclusions. Aronovich (1976) and Henton (1989) concluded that we expect male speech to lack variability and female speech to have a lot of variability; therefore, when males do use a lot of variability, its effect is very salient. The database in this study consists of six speakers per proficiency group but only three of each sex. If different conclusions need to be drawn for the males and females, the groups in this investigation have become too small to say anything conclusive. It is possible that, finding the rating task difficult for females, the judges sorted the females according to perceived proficiency in English.

4.4. Other variables

This paper has looked at variables of pitch and tempo, but speakers have a third means available for varying intonation, and that is intensity, or loudness. It may be that some of the speakers who were rated as lively though their pitch variation was low were using amplitude variation effectively. However, in order to gather reliable measurements of loudness, speakers must be recorded under much stricter conditions, for example, using a head-mounted microphone whose distance from the mouth is kept constant.

The results presented above would be simpler to interpret if the database was of native speech where issues of fluency and foreign accent are moot. Researchers with access to such a database are encouraged to test these methods on that speech. It is likely that the correlations between pitch variation and perceived liveliness would be even greater in such investigations.

5. Conclusion and pedagogical implications

To conclude, I would like to draw some preliminary conclusions regarding the pitch variation levels that correspond to lively versus monotone speech. The automatic feedback mechanism described in Fig. 1 could be configured to respond negatively to PVQ values that were under 0.15 for lengthy periods of time. Speakers could be encouraged to hold mean values between 0.15 and 0.25, and be rewarded for the occasional peak above 0.25. The level of liveliness could be adapted to the speaking genre; while one level would be suitable for an evangelist, another is clearly more appropriate to an academic conference presentation. Furthermore, people's perceptions of what is appropriate and pleasing may be individually and culturally determined to one extent or another.

In terms of speaking rate, a reasonable approach could be for speakers of a given proficiency to aspire to the speaking rate of the next proficiency level. The intermediate (groups A and B) speakers could aspire to reach speaking rates above three syllables per second. Clearly there is a maximum speaking rate above which comprehension becomes impaired, particularly when English is being used in international settings (Camiciottoli, 2005). Some native speakers may be unaware of what kinds of speaking rates may be inappropriate when a majority of the audience consists of non-natives who, though fluent speakers, may not be able to process content at the same speed as natives (Griffiths and Beretta, 1991). These speakers may benefit from feedback telling them to slow down their delivery. Furthermore, since problems with monotonous speech delivery are not restricted to non-native speakers, it is likely that both native and non-native speakers would find the pitch variation feedback useful.

Hahn (2004) has shown that speaking with sentence focus is important for the successful communication of content, and Pickering (2004) has shown that non-natives have a harder time than natives in modifying their pitch at the discourse level. Pickering's non-native subjects were native speakers of Mandarin; other research (Wennerstrom, 1994) has shown that native speakers of a European language do not exhibit the same difficulties as speakers of Asian languages in using pitch to structure discourse in English. Swedish is a language with a close genetic relationship to English, and can be characterized, like English, as a stress-timed rather than a syllable-timed language. As in English, new information in an utterance should receive more focus than given information. This focus is achieved primarily by pitch movement through the focused word, usually accompanied by lengthening of the focused syllable and an increase in intensity. The Swedish speakers in this study should thus not be experiencing negative transfer from their L1 when it comes to providing focus in the appropriate parts of an utterance or in making pitch resets when introducing a new topic, and there was no evidence in the corpus that they had these problems. Yet still some speakers were more monotone than others. It is likely that affect – nervousness – is a large reason for this. It stands to reason that speakers who are nervous presenting in their own language are even more nervous presenting in a second language. These speakers need simply to be reminded not to forget about intonation as they speak – hence the admonitions in the public speaking manuals. A major benefit of an automatic feedback mechanism would be simply to help speakers notice problems and to track their own improvement. When a speaker has learned to let pitch movement loose on the important parts of his or her message, the result should be livelier speech and better communication.

Acknowledgements

Many thanks to the colleagues who generously contributed their time evaluating the test files, the students who allowed me to record them, and David House and Björn Granström. This work was jointly funded by the Unit for Language and Communication and the Centre for Speech Technology at KTH. Portions of this paper were presented at the InSTILL/ICALL Conference in Venice, June 2004.

References

Anderson-Hsieh, J., 1992. Interpreting visual feedback on suprasegmentals in computer assisted pronunciation instruction. CALICO Journal 11 (4), 5–21.

Aronovich, C.D., 1976. The voice of personality: stereotyped judgements and their relation to voice quality and sex of speaker. Journal of Social Psychology 99, 207–220.

Bot, K.d., Mailfert, K., 1982. The teaching of intonation: fundamental research and classroom applications. TESOL Quarterly 16 (1), 71–77.

Brazil, D., 1997. The Communicative Value of Intonation in English. Cambridge University Press, Cambridge.

Camiciottoli, B., 2003. Interactive discourse structuring in L2 guest lectures: some insights from a comparative corpus-based study. Journal of English for Academic Purposes 3 (1), 39–54.

Camiciottoli, B., 2005. Adjusting a business lecture for an international audience: a case study. English for Specific Purposes 24, 183–199.

Coniam, D., 1999. Voice recognition software accuracy with second language speakers of English. System 27 (1), 49–64.

Council, E., 2001. Common European Framework of Reference for Languages. Cambridge University Press, Cambridge.

Cucchiarini, C., Strik, H., Boves, L., 2000. Different aspects of expert pronunciation quality ratings and their relation to scores produced by speech recognition algorithms. Speech Communication 30, 109–119.

Erickson, G., 2004. English: here and there and everywhere. En undersökning av ungdomars kunskaper i och uppfattningar om engelska i åtta europeiska länder. Stockholm, Skolverket: 90.

Eskenazi, M., 1999. Using automatic speech processing for foreign language pronunciation tutoring: Some issues and a prototype. Language Learning and Technology 2 (2), 62–76.

Grandstaff, D., 2004. Speaking as a Professional. W.W. Norton & Co.

Granqvist, S., 2003. Computer methods for voice analysis. Department of Speech, Music and Hearing. Stockholm, KTH. PhD. thesis.

Griffiths, R., 1990. Speech rate and NNS comprehension: a preliminary study in time-benefit analysis. Language Learning 40 (3), 311–336.

Griffiths, R., 1991. Pausological research in an L2 context: a rationale and review of selected studies. Applied Linguistics 12 (4), 345–364.

Griffiths, R., Beretta, A., 1991. A controlled study of temporal variables in NS-NNS lectures. RELC Journal 22 (1), 1–19.

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

Hardison, D., 2004. Generalization of computer-assisted prosody training: quantitative and qualitative findings. Language Learning and Technology 8 (1), 34–52.

Henton, C., 1989. Fact and fiction in the description of female and male pitch. Language and Communication 9 (4), 299–311.

Hincks, R., 2003a. Speech technologies for pronunciation feedback and evaluation. ReCALL 15 (1), 3–20.

Hincks, R., 2003b. Pronouncing the Academic Word List: Features of L2 student oral presentations. In: 15th International Congress of Phonetic Sciences, Barcelona, ICPhS Organizing Committee.

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

Kormos, J., Dénes, M., 2004. Exploring measures and perceptions of fluency in the speech of second language learners. System 32, 145–164.

Lamerton, J., 2001. Collins Complete Guide to Public Speaking. HarperCollins.

Levis, J., Pickering, L., 2004. Teaching intonation in discourse using speech visualization technology. System 32, 505–524.

Levy, M., 1997. Computer-Assisted Language Learning. Clarendon Press, Oxford.

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

Morell, T., 2004. Interactive lecture discourse for university EFL students. English for Specific Purposes 23, 325–338.

Neri, A., Cucchiarini, C., Strik, H., Boves, L., 2002. The pedagogy-technology interface in computer assisted pronunciation training. Computer-Assisted Language Learning 15 (5).

Oscarson, M., 1995. A national evaluation programme in the Swedish compulsory school: assessment of achievement in foreign languages. System 23 (3), 295–306.

Öster, A.-M., 1998. Spoken L2 teaching with contrastive visual and auditory feedback. In: Proceedings of ICSLP, Sydney.

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

Rowley-Jolivet, E., Carter-Thomas, S., 2005. Genre awareness and rhetorical appropriacy: manipulation of information structure in the international conference setting. English for Specific Purposes 24, 41–64.

Sinclair, J. (Ed.), 1995. Collins COBUILD English Language Dictionary. HarperCollins, London.

Simpson, R.C., Lee, D.Y.W., Leicher, S., 2003. MICASE Manuel: The Michigan Corpus of Spoken Academic English. English Language Institute, The University of Michigan, Ann Arbor, MI, USA.

Sjölander, K., Beskow, J., 2000. WaveSurfer: An open source speech tool. ICSLP 2000, Available from: <http://www.speech.kth.se/snack>/.

Towell, R., Hawkins, R., Bazergui, N., 1996. The development of fluency in advanced learners of French. Applied Linguistics 17 (1), 84–119.

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. Applied Linguistics 15, 399–421.

Wennerstrom, A., 2001. The Music of Everyday Speech. Oxford University Press, New York.

Ventola, E., Shalom, C., Thompson, S. (Eds.), 2002. The Language of Conferencing. Peter Lang, Frankfurt am Maim.