Measuring Liveliness in Presentation Speech

Rebecca Hincks

Department of Speech, Music and Hearing
Unit for Language and Communication
Royal Institute of Technology, Stockholm, Sweden
hincks@speech.kth.se

Abstract

This paper proposes that speech analysis be used to quantify prosodic variables in presentation speech, and reports the results of a perception test of speaker liveliness. The test material was taken from a corpus of oral presentations made by 18 Swedish native students of Technical English. Liveliness ratings from a panel of eight judges correlated strongly with normalized standard deviation of F0 and, for female speakers, with mean length of runs, which is the number of syllables between pauses of >250 ms. An application of these findings would be in the development of a feedback mechanism for the prosody of public speaking.

1. Introduction

Advice on speaking in public invariably includes instruction to use a voice that varies in pitch, tempo and intensity. Speakers are also told to practice their talks ahead of time, preferably with a listener who can give feedback on the way the voice has been used. The idea behind the research presented in this paper is that automatic feedback on speaker prosody could be provided by signal analysis software that would quantify intonational variation and measure speaking rate in syllables per second. Such a system would help in particular monotone speakers and speakers who tend toward speaking too quickly for international audiences.

1.1. Motivation

Using one's voice well is not the absolutely most critical aspect of making a good presentation; 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, the message can be lost. Any number of popular manuals on public speaking advise speaking with a lively voice that varies in intonation. 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.

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 [1, 2] 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 [3-5]. Pickering [5] compared the way native and non-native teaching assistants used intonational paragraphing 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" (p.19). One of the contributing problems was an overall narrower pitch range, which made the identification of prosodic units difficult.

1.2. Design of proposed feedback mechanism

The public speaking feedback mechanism (Fig. 1) is conceptualized as a 'speech checker' in analogy with grammar and spell checkers. Within the framework of computer-assisted language learning theory [6], the mechanism would be placed in a supportive rather than a tutorial role. Speaker-dependent speech recognition could provide information on repeated segmental errors and the transcript of the presentation could be analyzed for features that are regarded as hallmarks of successful or problematic presentations. Speech analysis would quantify prosodic variation, measure speaking rate and pause time, and search for repetitive pitch contours that could be negatively perceived by an audience. An appropriate feedback interface for the prosodic analyzer would be a talking head that would listen alertly as long as the prosody showed adequate variation but show lose attention, or even fall asleep, if the speech became monotonous.

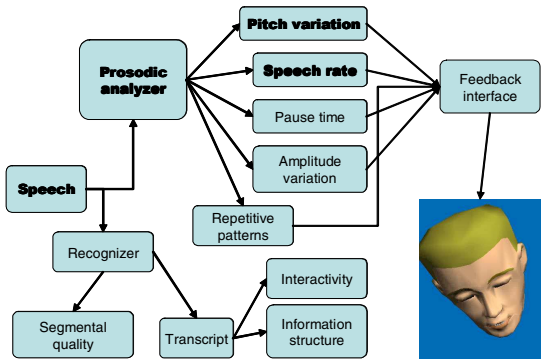

Figure 1. Schematic diagram of feedback mechanism for public speaking. Features in bold are treated in this paper.

1.3. SD F0 in relation to liveliness

The standard deviation of F0 in relation to perceptions of liveliness in short synthesized utterances has been previously studied by [7], and [8] examined pitch variation in different types of discourse. In the research reported here, the primary questions were to 1) determine the levels of SD F0 found in a corpus of natural presentation speech and 2) test the perceptions of liveliness of different levels of pitch variation in natural speech. Results would guide the setting of the parameters for prosodic feedback.

1.4. Temporal variables in relation to liveliness

This paper also briefly investigates the temporal variables that would be of value in a prosodic feedback mechanism. [7] found a positive relationship between rate of speech and perceptions of liveliness, but it is important that speakers, in their attempt to produce lively speech, not begin to speak too quickly. Studies have shown that lecturers have difficulty in adapting their rates of speech even when they know that their audience may not be able to follow them [9]. Quantifying the rates of speech that proficient non-native speakers of English use with each other should help establish norms that might be of benefit to native speakers as well.

2. Method

2.1. Speech material

The speakers in the presentation database were Swedish students of engineering recorded as they completed a graded requirement in their courses in Technical English, where improved presentation skills are an important course goal. The presentations were 7 to 10 minutes long and on varied topics of a general technical nature. An analysis of the pronunciation errors made by speakers in the corpus was presented in [10]. Out of a total of 28 recordings, 18 presentations were chosen for prosodic analysis. The resulting subcorpus consisted of balanced groups of males and females at three levels of English proficiency, two intermediate groups (A and B) and one advanced (C).

2.2. Pitch variation quotient, PVQ

Pitch extraction was performed using the ESPS analysis incorporated in WaveSurfer [11] with pitch boundaries set at 60-400 Hz for males and 120-500 Hz for females. The contours were visually inspected and the location of extraction errors noted for later deletion. For each 10-second segment of speech (up to 60 segments per speaker), the mean and standard deviation of F0 were calculated. To normalize between speakers, the standard deviation for each segment was divided by the mean for each segment, yielding a value termed the PVQ, for pitch variation quotient.

2.3. Perception of liveliness test

Nine ten-second samples of speech per speaker were selected for a perception test of speaker liveliness. The nine samples represented three peak PVQs, three median PVQs, and three low PVQs per speaker. Separate tests for males and females were prepared, giving 81 samples to rate for each test.

Eight university teachers of English (two males, six females) were recruited as raters. They were not aware that it was pitch variation that was being tested. The raters were presented with a randomized set of icons, each representing a speech sample, on a computer screen. They were instructed to move the icons to a continuous scale whose endpoints were marked 'lively' and 'monotone,' and they were allowed to listen and move the icons as many times as they wanted to. The tests took about 45 minutes to complete and the raters did the male and female tests on different days.

2.4. Calculating temporal variables

The transcripts of the 18 presentations were divided into utterance runs between silent pauses >250 ms. The number of syllables per run was counted and these numbers were used to calculate the speaking rate in syllables per second (SPS) and the mean length of runs (MLR).

3. Results

3.1. Pitch variation quotients per segment and speaker

PVQ values for individual ten-second samples of speech ranged from a low of about .06 to a high of .34. Males achieved higher values than females. Figure 2 shows the distribution of the values in the corpus.

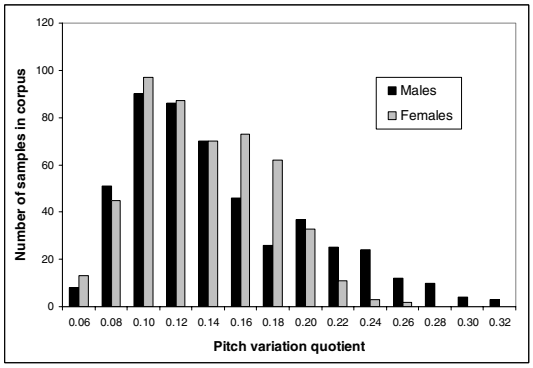

Figure 2. Distribution of PVQ values for ten-second segments in presentation corpus

The mean PVQs per speaker for the whole presentation ranged from a low of .11 to a high of .24. Figure 3 shows these measures in relation to the speaker's proficiency in English as measured by the placement test the students took before they started studying English. There is only a weak correlation (.42) between language ability and pitch variation.

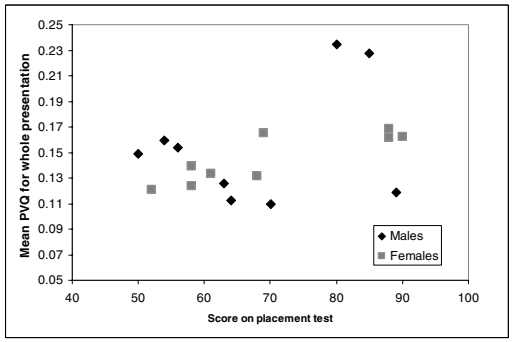

Figure 3. Mean PVQ per speaker in relation to English proficiency

3.2. PVQ in relation to liveliness perception

The liveliness perception test returned a score between 0 and 1000, which is the scale used on the x-axis in Figures 4 and 5. The first figure shows the composite liveliness ratings for the nine speech samples per speaker in relation to the mean of the PVQs of the speech samples. There is a moderate correlation (r = .83, p < .01) between the perception of liveliness and PVQ.

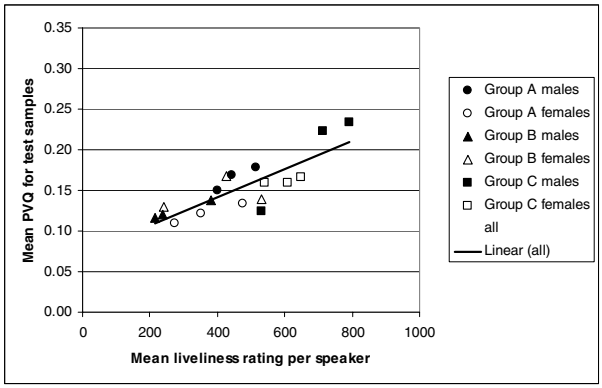

Figure 4. Liveliness ratings in relation to PVQ per speaker

Figure 5 plots PVQ against liveliness ratings per speech sample. The correlations are stronger for the males than for the females. Speech samples from the more proficient speakers of Group C are clustered in the lower right hand quadrant, indicating that their speech could be perceived as lively even when the PVQs were low. In contrast, the less proficient speakers in groups A and B are more likely to be found in the upper left quadrant.

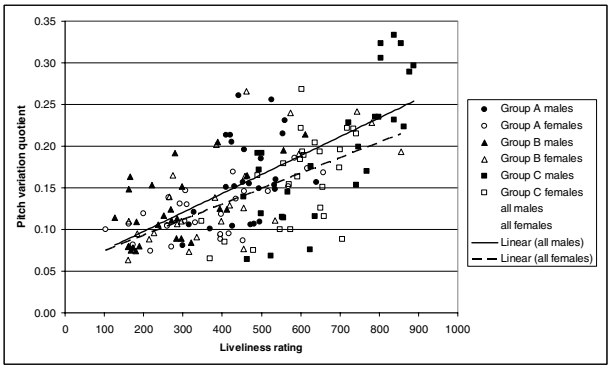

Figure 5. Liveliness ratings in relation to pitch variation quotient per speech sample

3.3. Temporal measures in relation to liveliness perception

Table 1 shows the correlations by sex between liveliness ratings of speech samples and the speaking rate (SPS), mean length of runs (MLR), and pitch variation (PVQ) of those samples. Speaking rate as measured in SPS was not found to correlate with liveliness perceptions for either sex. For female speakers, the correlation between MLR and liveliness is stronger than the correlation between PVQ and liveliness. In stark contrast, the males show an insignificant negative correlation between the two values.

Table 1. Pearson correlation coefficients between prosodic variables and liveliness ratings of 162 speech samples

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

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

4. Discussion

4.1. Potential for feedback

The results presented above are encouraging for the development of an automatic feedback mechanism. When the standard deviations are expressed as a percentage of the mean fundamental frequency (PVQ) and measured over ten-second intervals of natural speech, the values generated in this corpus of about three hours of presentation speech range from a low of .06 to a high of .34. Speech samples with PVQ values less than .15 were generally rated as on the monotone side of a monotone-lively scale. Mean PVQ values for 7-10 minutes of speech had a smaller range, from .11 to .24; here again, speakers with mean values of less than .16 were rated as more monotone than lively. A feedback mechanism could thus be programmed to reward speech where the variation was more than 15% of the mean.

In this corpus of non-native speakers, some of whom were of a near-native proficiency level, we find speaking rates as measured in syllables per second over the whole presentation that range between 2.42 and 3.56. These rates are lower than what is found in spontaneous speech due to the inclusion of pauses, which can be lengthy in a presentation. The effect of different levels has not been perceptually tested, but it would be interesting to pursue this line of investigation to see what range of speaking rates are appropriate when using English internationally.

4.2. Sex-based perceptual differences

Pitch variation was a stronger perceptual cue to liveliness in male speech than in female speech. This is consistent with other work on personality and speech perception, such as that of [12], who found that listeners were sensitive to *variation* in male F0 but to the *means* of female F0. A possible explanation to this, proposed by [13], is that male speakers are expected to be monotone and are therefore rewarded more strongly when they vary their pitch.

Furthermore, the PVQ values were found to be higher for male speech than for female speech. This is consistent with the analysis done by [13], who, by converting published data from many researchers to semitones, showed that contrary to popular perceptions, males had a larger pitch range than females. It should be mentioned here that different values were obtained for the females in this corpus when WaveSurfer's pitch extraction boundaries were set at different levels, and that the optimal boundaries are not yet known. For heuristic reasons, I believe it is preferable to remain working with the Hertz scale rather than using the raw standard deviation from the semitone scale, though this would ensure more valid comparisons between males and females.

Speech rate as measured in syllables per second was not found to be a correlate of liveliness. However, for female

speech samples, the number of syllables between pauses (MLR) was a stronger correlate of liveliness than pitch variation. This is an interesting finding, which should be tested on a database of native speech. Because MLR is strongly correlated with language proficiency [14], it is possible that the raters were influenced by the speakers' perceived proficiency in English when rating female speech.

Hahn [3] has shown that speaking with sentence focus is important for the successful communication of content, and Pickering [5] 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 [15] 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. As in English, new information in an utterance should receive more focus than given information. This focus is achieved primarily by pitch movement in 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. It is likely that for speakers of many L1s, once they are reminded to place focus somewhere, they will know where to do it.

5. Acknowledgements

Many thanks to advisors David House and Björn Granström and to the colleagues who let me into their classrooms and generously gave their time rating liveliness. This research was funded by the Unit for Language and Communication at KTH and supported by the Centre for Speech Technology. A longer version of this article has been accepted for publication in *System*.

6. References

[1] Rowley-Jolivet, E. and S. Carter-Thomas, "Genre awareness and rhetorical appropriacy: Manipulation of information structure in the international conference setting". *English for Specific Purposes*. 242005: p. 41-64.

[2] Ventola, E., C. Shalom, and S. Thompson, eds. *The Language of Conferencing*. 2002, Peter Lang: Frankfurt am Maim.

[3] Hahn, L.D., "Primary Stress and Intelligibility: Research to Motivate the Teaching of Suprasegmentals". *TESOL Quarterly*. 38(2)2004: p. 201-223.

[4] Levis, J. and L. Pickering, "Teaching intonation in discourse using speech visualization technology". *System*. 322004: p. 505-524.

[5] Pickering, L., "The structure and function of intonational paragraphs in native and nonnative speaker instructional discourse". *English for Specific Purposes*. 232004: p. 19-43.

[6] Levy, M., *Computer-Assisted Language Learning*. Oxford: Clarendon Press.1997.

[7] Traunmüller, H. and A. Eriksson, "The perceptual evaluation of F_0 excursions in speech as evidenced in liveliness estimations". *Journal of the Acoustical Society of America*. 97(3)1995: p. 1905-1915.

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

[9] Griffiths, R. and A. Beretta, "A controlled study of temporal variables in NS-NNS lectures". *RELC Journal*. 22(1)1991: p. 1-19.

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

[11] Sjölander, K. and J. Beskow. "WaveSurfer: An open source speech tool". in *ICSLP 2000*. 2000. http://www.speech.kth.se/snack/.

[12] Aronovich, C.D., "The voice of personality: stereotyped judgements and their relation to voice quality and sex of speaker". *Journal of Social Psychology*. 991976: p. 207-220.

[13] Henton, C., "Fact and fiction in the description of female and male pitch". *Language and Communication*. 9(4)1989: p. 299-311.

[14] Kormos, J. and M. Dénes, "Exploring measures and perceptions of fluency in the speech of second language learners". *System*. 322004: p. 145-164.

[15] Wennerstrom, A., "Intonational meaning in English discourse". *Applied Linguistics*. 151994: p. 399-421.