What makes a good speaker? Subjective ratings and acoustic measurements

Eva Strangert
Department of Comparative Literature and Scandinavian Languages, Umeå University

Abstract

The paper deals with qualities contributing to the impression of a "good speaker" – a speaker capable of catching the attention of an audience through her/his way of speaking. Subjective ratings of speaker qualities were correlated with acoustic analyses of samples of speech produced in Swedish parliament debates. Raters reliably differentiated between more and less skilled speakers and reached good agreement on qualities contributing to their impression. Acoustic measurements reveal substantial differences with regard to F0 and duration features for speakers rated high and low, respectively, on speaker skill.

Introduction

In the public speech domain and in politics which is the area dealt with here, the attention of the listener is crucial. In addition to having an interesting piece of information to deliver, the speaker should wrap it in a form that makes the message go through. Thus, although it may be difficult to separate *what* is said from *how* it is said, the focus here is on the latter. The assumption is that prosody has a key role in efficient communication and that being "a good speaker" means using prosody in an optimal way.

Within a broader project, described in Strangert (2006), the present study then has its focus on the contribution of prosody to successful interaction with an audience. In politics such interaction is crucial. A rich expressive repertoire, in which no doubt prosody has a major role, is a great advantage in order to be "heard". However, although there may be a wider spectrum of argumentative and emotionally coloured expressive speech in politics, there is reason to believe that good-speaker characteristics – and less good ones – *basically* do not differ much between different public domains.

Background

The study builds on previous work on public speech based on Swedish. Strangert (1993), observed a considerably more dynamic style of speech by a news announcer as compared to an ordinary speaker. A recent study (Strangert 2005) compared samples of speech of a news announcer and a well-known politician. Both used prosody very efficiently. However, the politician had a greater variety of expressions and also a wide repertoire of argumentative and emotionally coloured expressive acts conveyed by prosody.

Braga & Marques (2004) analyzed political debate in order to shed light on acoustic correlates of speaker attributes like "convincing", "powerful" etc. and Mozziconacci (2002) and Wichmann (2002) both deal with the relations between prosody (F0 in particular) and affective functions. Hincks (2005), with focus on oral-presentation skill in the class-room, investigated acoustic correlates of "liveliness".

Wichmann (2002) makes a distinction between "ways of saying" (properties or states relating to the speaker) and "ways of behaving" (attitudes to the listener). "Ways of saying" includes first, how the speaker use prosody per se (stress and emphasis, pausing etc.) and second, the emotional colouring of speech (e.g. "happy", "sad") as well as states such as "excited", and "powerful". Examples of "ways of behaving" are attitudes such as "arrogant" and "pleading". In addition, the speaker may use other argumentative and rhetorical means. All these communicative functions of prosody make it a complex and powerful means for interaction.

To study the affective functions of prosody, listeners' impressions have to be classified appropriately. As human speech very often conveys several states, attitudes and emotions at the same time, a suitable methodology is needed. Liscombe et al. (2003) used multiple and continuous scales for subjective ratings of emotions. The same methodology was used by Rosenberg & Hirschberg (2005) who combined the ratings with acoustic analyses in order to

characterize charismatic speech. The present study has the same general approach combining ratings of speaker qualities and acoustic data.

Data and rating procedure

The analyzed material consisted of 16 samples of speech, all between 30 and 35 seconds in length. The samples were all from debates in the Swedish parliament (Riksdagen) between parliament members and government ministers. They were recordings (audio and video) from the Riksdagen archive made publicly available on the web. The material was chosen to represent a variety of speakers (more and less skilled ones, according to the author; eight male and eight female). On the basis of findings of insignificant effects of topic variation reported by Rosenberg & Hirschberg, the issues covered by the speakers were allowed to vary.

The experiment was run via a web interface. 18 native Swedish students of language and literature (nine female and nine male) were paid a small amount for their participation. While listening to the samples the subjects gave their opinion on 13 statements about the speaker on a five-point scale with "no, absolutely not" (coded as 0) and "yes, absolutely" (coded as 4) as endpoints. The statements had the form *The speaker is* …. followed by *insecure* (osäker), *hesitant* (trevande), *monotonous* (monoton), *aggressive* (aggressiv), *accusing* (anklagande), *agitating* (agitatorisk), *objective* (saklig), *trustworthy* (förtroendeingivande), *humble* (ödmjuk), *expressive* (uttrycksfull), *powerful* (kraftfull), *involved* (engagerad), respectively. There was also an overall "good-speaker" rating formulated as *The speaker is all in all a good speaker, a person capable of catching the attention of an audience through her/his way of speaking*.

Although the distinction is not always clear the qualities chosen in the statements concerned both properties and states (statements 1-3 and 10-12) as well as speaker attitudes towards the audience or the message (4-9). This choice of qualities was guided by the distinction between properties or states of the speaker and speaker attitudes as discussed by Wichmann (2002). The results from a survey (Strangert, 2006) in which a group of subjects named qualities they associated with good-speaking behaviour also influenced the selection as well as findings reported by Rosenberg & Hirschberg.

The samples, normalized for intensity, were repeated with two seconds of silence in-between until the subject had completed the 13 ratings for a specific speaker. Each of the 18 subjects heard the samples in a unique random order. The statements also, with one exception, occurred in random order for each of the samples of speech. The good-speaker statement giving the overall characterization of the speaker always occurred in the last position (as statement 13).

To familiarize the subjects with the test situation, they started the experiment by rating a sample additional to the 16 in the experiment. In total 3744 ratings were made (16 speakers x 13 statements x 18 subjects).

Analysis of subjective ratings

Considering the great amount of subjectivity in rating speaker characteristics, overall agreement cannot be expected to be high. Calculating Fleiss' kappa for multiple observers (Fleiss, 1971), the exact agreement across all speakers and statements gave $\kappa = .11$, with a range between .03 and .22 for individual speakers. Permitting a more relaxed criterion for agreement using a weighting statistic (to reduce influence of extreme disagreement) would give a higher kappa value. Such a relaxed criterion (with quadratic weighting) was used by Rosenberg & Hirschberg who report a mean (Cohen's) kappa value of .21.

However, a more detailed analysis reveals greater amounts of consistency. First, to find out about the qualities underlying the good-speaker ratings, these ratings was matched against the ratings of all the other qualities (= statements). The mean correlation coefficients based on exact agreement between raters for all speakers are shown in Table 1. The table in addition shows correlations between the means of all ratings (coded as 0, 1, 2, 3, 4) for all subjects. (These means for each statement and for each speaker separately are shown in Table 2, where values at the extreme ends (close to 0 or 4) represent cases were agreement is particularly high).

There are both similarities and differences between the two correlation measures in Table 1. However, focus will be on the correlations between rating means (b), as in this study we are less interested in the exact correspondence between subjects than the means across all subjects for each statement.

Not unexpectedly, qualities such as *expressive*, *powerful* and *involved* are positively correlated with being a good speaker. However,

also qualities (or attitudes) such as *aggressive*, *accusatory* and *agitating* show a high positive correlation. This may appear unexpected as, in other situations (e.g. a class-room) these attitudes would be considered negative and therefore most reasonably not adding to the impression of a good speaker.

Table 1. Mean correlation coefficients between good-speaker ratings (statement 13) and ratings of other qualities across speakers. Correlations based on ratings of exact agreement (a) and rating means (b), se explanation in text. Significance indicated for (b).

Qualities	(a)	(b)	
1. insecure	-.414	-.866	**
2. hesitant	-.272	-.856	**
3. monotonous	-.381	-.913	**
4. aggressive	.121	.738	**
5. accusatory	.004	646	**
6. agitating	.132	869	**
7. objective	.374	313	
8. trustworthy	.593	.915	**
9. humble	.298	-.550	*
10. expressive	.533	.956	**
11. powerful	.483	947	**
12. involved	.389	.885	**

Thus, the result here probably indicates other expectations of a politician – expectations in terms of activation, whether positive or negative – than of other speakers. This assumption is supported by the rather high negative correlation with *humble*. In addition, there is a high correlation for *trustworthy*, while being *objective* appears to be an unnecessary quality.

The ranking of the individual speakers on statement 13 (*good speaker*) appears from Table 2. The speaker rated highest had a mean of 3.39, while the one rated lowest had a mean of 0.56.

Acoustic analysis

The analysis presented here is confined to basic acoustic data for the two speakers (both male) rated highest (RH) and lowest (RL), respectively, on statement 13. Measurements included F0 (mean, standard deviation, minimum, maximum, range) and duration features (mean pause duration, total pause duration, total speech duration, pause-to-speech duration, speech rate, articulation rate), see Table 3.

Table 3. F0 and duration data for the two speakers rated highest (RH) and lowest (RL) on the good-speaker rating (statement 13).

Speaker	RH	RL
F0 (Hz)		
mean	138	113
standard deviation	28	12
minimum	75	77
maximum	233	173
range	158	96
Duration (sec)		
total speech sample	33.93	36.62
total pause	4.97	9.05
mean pause*	.41	.57
pause-to-speech ratio	.15	.25
speech rate (syll/sec)	4.27	3.69
articulation rate (syll/sec)**	5.01	4.90

* N = 12 and 16, respectively, for RH and RL
** Calculated after subtraction of pause durations

The speakers differ with respect to almost all features covered here. RH has a considerably greater mean, standard deviation and range of

Table 2. Means of subject ratings of 0 (no, absolutely not) - 4 (yes, absolutely) per speaker and quality.

Speaker	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1. insecure	0.67	2.83	2.44	0.72	0.67	0.56	1.11	3.00	0.67	0.94	1.61	1.11	1.06	0.44	0.89	1.33
2. hesitant	1.11	3.50	3.00	0.94	1.11	0.67	1.06	3.33	0.89	1.28	1.83	1.67	1.06	1.00	1.39	1.28
3. monotonous	1.89	2.72	2.11	1.33	1.17	0.83	0.83	3.39	1.72	0.94	3.61	1.83	2.33	0.89	1.56	1.56
4. aggressive	1.28	0.56	1.50	2.06	2.83	3.06	3.17	0.44	1.61	2.28	1.44	2.33	1.28	3.28	0.83	1.83
5. accusatory	1.39	1.39	2.28	2.56	2.94	3.28	3.39	1.00	1.72	2.00	1.22	3.00	1.44	1.78	1.89	3.06
6. agitating	1.94	1.44	1.61	2.33	2.72	2.78	2.61	1.11	2.39	2.28	1.44	2.50	2.22	3.22	1.94	2.61
7. objective	2.44	2.06	2.22	2.06	2.56	2.78	2.22	1.89	2.83	2.83	3.11	2.44	3.00	2.56	3.00	1.89
8. trustworthy	2.22	0.89	1.44	2.00	2.39	3.00	1.67	0.89	2.39	2.39	1.17	2.17	1.89	2.06	2.67	2.17
9. humble	1.72	1.89	1.83	1.94	0.94	1.44	0.50	2.56	1.56	1.72	1.56	1.11	1.56	1.06	2.00	1.67
10. expressive	1.83	1.00	1.83	2.56	2.89	3.44	3.06	0.78	2.56	2.83	0.89	2.28	1.83	3.17	2.39	2.67
11. powerful	2.33	0.44	1.50	1.94	3.06	3.33	2.78	0.39	2.50	2.89	1.06	2.28	2.00	3.61	2.17	2.11
12. involved	2.11	1.50	2.67	3.11	3.33	3.89	3.50	1.17	2.56	3.11	1.33	2.56	1.89	3.67	2.61	2.89
13. good speaker	2.22	0.61	1.67	2.17	2.72	3.39	2.67	0.56	2.72	2.61	0.83	2.17	2.06	2.72	2.50	2.39

F0. As the speakers have about the same base level value, the greater range of RH is explained by the higher F0 maximum. Based on the restricted data, it would be premature to conclude that a higher mean F0 makes a better speaker than a lower mean. However, Hirschberg & Rosenberg observed a positive correlation between mean F0 and charisma ratings. They also found a greater standard deviation among the more charismatic speakers just as Hincks (2005) and Traunmüller & Eriksson (1995) observed a positive relation between F0 variation and "liveliness". A similar greater spread (standard deviation and range) for RH, as compared to RL, most reasonably is related to the good-speaking quality *expressiveness* as well as to features such as *involved* and *agitating*, see Table 1 for correlations across all speakers and Table 2 for individual data. The more varied F0 of RH was also clearly perceived by the raters (Table 2).

RH and RL, further, has about the same articulation rate. However, taken pause time into consideration, differences in speech rate and pause-to-speech ratio show RL to be the slower speaker. This, in addition to the more frequent pausing (and several repeats and repairs), is the most reasonable explanation for the impression of RL as *hesitant,* and probably also as a less skilled speaker. It is worth noting here that Hirschberg and Rosenberg found that the higher the rate, the higher the charisma rating.

Conclusions and future work

This paper reports ongoing work on the prosody of public speech. Speech samples from 16 politicians were rated on a number of qualities assumed to contribute (positively or negatively) to speaker skill. Raters also gave an overall judgement of each speaker, indicating to which extent they considered her/him to be a "good speaker".

The raters reached considerable agreement on the good-speaker ratings, and correlations between these and ratings of the other qualities give clear indications of the grounds for differentiating between speakers. The results indicate that a good speaker is *trustworth*y, *expressive*, *powerful* and *involved* and that being *insecure*, *hesitant* and *monotonous* leads to the opposite impression. Though this is not quite unexpected, the results that *aggressive*, *accusatory* and *agitating* are positively correlated with speaker skill are more remarkable. A probable explanation might be that these quailties – negative in many other contexts – are allowed in politics.

Comparing the two most extreme speakers, in terms of overall skill revealed considerable differences. The highest rated speaker had a higher mean and a more varying F0 compared to the lowest rated speaker, who was also rated as *monotonous*. The highest rated speaker also had a faster speech rate and less and shorter pauses, while the other speaker's slower rate may explain the impression of *hesitance*.

Future work will include more detailed as well as expanded analyses of the collected material. The research plans also include considering rhetorical aspects, as the choice of linguistic form no doubt can influence speaker ratings.

Acknowledgements

Thierry Deschamps is gratefully acknowledged for technical and data support and Julia Hirschberg and Andrew Rosenberg for sharing their experiences of speaker ratings. The work is supported by the Swedish Research Council.

References

Braga D & Marques M A (2004). The pragmatics of prosodic features in the political debate. *Proc. Speech Prosody 2004*: 321-324.

Fleiss J L (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76 (5): 378-382.

Hincks R (2005).*Computer support for learners of spoken English*. Diss. Speech and Music Communication, KTH.

Liscombe J, Venditti J & Hirschberg J (2003). Classifying subject ratings of emotional speech using acoustic features. *Proc. Eurospeech 2003*: 725-728.

Mozziconacci S (2002). Prosody and emotions. *Proc. Speech Prosody 2002*: 1-9.

Rosenberg A & Hirschberg J (2005) Acoustic/prosodic and lexical correlates of charismatic speech. *Proc. Interspeech 2005*: 513-516.

Strangert E (1993). Speaking style and pausing. *PHONUM*, 2: 121-137. Dept. of Phonetics, Umeå University.

Strangert E (2005). Prosody in public speech: Analyses of a news announcement and a political interview. *Proc. Interspeech 2005*: 3401-3404.

Strangert E & Deschamps T (2006). The prosody of public speech – A description of a project. *Working Papers* 52: 121-124. General Linguistics and Phonetics, Lund University.

Traunmüller H & Eriksson A (1995). The perceptual evaluation of F_0 excursions in speech as evidenced in liveliness estimations. *JASA*, 97 (3): 1905-1915.

Wichmann A (2002). Attitudinal intonation and the inferential process. *Proc. Speech Prosody 2002*: 11-22.