interspeech 2008 incorporating SST 08

What Makes a Good Speaker?
Subject Ratings, Acoustic Measurements and Perceptual Evaluations

Eva Strangert [1], *Joakim Gustafson* [2]

[1] Department of Language Studies, Umeå University, Sweden
[2] CSC, Department of Speech, Music and Hearing, KTH, Stockholm, Sweden
eva.strangert@nord.umu.se, jocke@speech.kth.se

Abstract

This paper deals with subjective qualities and acoustic-prosodic features contributing to the impression of a *good speaker*. Subjects rated a variety of samples of political speech on a number of subjective qualities and acoustic features were extracted from the speech samples. A perceptual evaluation was also conducted with manipulations of F0 dynamics, fluency and speech rate with the sample of the lowest rated speaker as a basis. Subjects' ranking revealed a clear preference for modified versions over the original with F0 dynamics – a wider range – being the most powerful cue.
Index Terms: prosody, speaker skill, subject ratings, acoustic measurements, synthesis evaluation

1. Introduction and outline of the study

The present study starts out from the assumption that prosody has a key role in efficient communication and that being "a good speaker" includes using prosody in an optimal way.

In the public speech domain and in politics, which is the area dealt with here, the attention of the listener is crucial. The speaker should not only have an interesting piece of information to deliver, but also wrap it in a form that gets the message across to the audience in the best possible way. Although being a good speaker may basically be the same in different domains, we can expect a greater variation with a potential for a wider spectrum of argumentative and emotionally colored speech in politics. A rich expressive repertoire, in which no doubt prosody has a major role, is a great advantage in order to be "heard".

Thus, expressiveness is in the foreground here. Wichmann in [11] contributes to our understanding about the relations between prosody (primarily F0) and affective functions. A distinction is made 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 coloring 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.

The current research further builds on previous work on prosody in the specific area of public speech. [8] contains a comparative analysis of speech samples from a news announcer and a well-known politician. Both used prosody very efficiently, but the politician had a greater variety of expressions and also a wide repertoire of argumentative and emotionally colored expressive acts conveyed by prosody. In [7], ratings of speaker qualities were compared to acoustic data in order to characterize American politicians in terms of "charisma". Also, in an extension to this study [1], cross-cultural comparisons were made of American, Palestinian and Swedish ratings of charisma in American English and Palestinian Arabic political speech. Although perception of charisma was partly influenced by cultural factors, some acoustic features gave similar ratings irrespective of the language rated and the nationality of the raters. Such features among others include mean pitch, mean rms intensity and pitch range. That pitch, and pitch variation in particular, is a powerful cue as affective functions are concerned is well attested. Liveliness, for example, has been found to be strongly associated with pitch variation [5, 10].

To study the affective functions of prosody, a suitable methodology is needed. [6] used multiple scales for subjective ratings of emotions. This methodology was also used in [7], where ratings of charisma were combined with acoustic data in order to characterize American charismatic speech and, in addition, in the aforementioned cross-cultural study [1] concerned with charisma. The same methodology was also used in [9], dealing with subjective judgments of Swedish in relation to some (restricted) acoustic data.

In the present study, we build on this previous work in [9], extending the acoustic analyses and, in addition, include a perceptual evaluation of potential cues to the impression of speaker skill. Thus we deal with our central issue in three ways: We have subjects judge a variety of samples of political speech on a number of subjective qualities. We extract acoustic features from the speech samples. And we use synthesis to evaluate our findings.

2. Material

The material consisted of 16 samples of speech, all between 30 and 36 seconds in duration. 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 so as to represent a variety of speakers (more and less skilled ones, according to the first author; eight male and eight female). On the basis of findings of insignificant effects of topic variation reported in [7], the issues covered by the speakers were allowed to vary. The material also included samples of both read and more or less spontaneous speech.

3. Ratings of speaker qualities

Ratings of the speech samples were made by 18 native Swedish students of language and literature (nine female and nine male) via a web interface. 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*, *hesitant*, *monotonous*, *aggressive*, *accusing, agitating*,

objective, trustworthy, *humble*, *expressive*, *powerful*, *involved*, respectively. There was also an overall "good-speaker" rating based on the statement *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*. The qualities chosen for investigation thus concerned properties and states as well as speaker attitudes towards the audience or the message [11]. Findings reported in [7] also influenced the selection.

The samples, normalized for intensity, were repeated with two seconds of silence in-between until the subject had completed the 13 ratings for each 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. In total, 3744 ratings were made (16 speakers x 13 statements x 18 subjects).

To find out about the qualities underlying the good-speaker ratings, these ratings were matched against the ratings of all the other qualities (= statements). Qualities such as *expressive*, *powerful* and *involved* as well as *trustworthy* were found to have a strong positive (significant) correlation (all with $r \geq .89$) with being a good speaker (based on means of all individual ratings for each quality). The same holds for qualities (or attitudes) such as *aggressive*, *accusatory* and *agitating* (with $r \geq .65$), which seem to indicate other expectations of a politician than of speakers in other situations. This assumption is further supported by the rather strong negative correlation with *humble* ($r = -.55$). However and not unexpectedly, the strongest negative (significant) correlations were found for *hesitant, insecure* and *monotonous* ($r = -.86$, $-.87$ and $-.91$, respectively), while *objective* with a low and insignificant correlation appeared as an unnecessary quality. A detailed description of the experiment can be found in [9].

In section 4, acoustic measurements are matched with rating data, the mean ratings of the individual speakers on statement 13 (*good speaker*). These ratings showed a considerable variation – from 3.39 for the speaker rated highest to 0.56 for the lowest rated speaker.

4. Acoustic analysis

A number of acoustic features were extracted from the speech samples and correlated with the mean speaker ratings. The measurements included mean pause duration (= silent interval), mean duration of speech chunks (between pauses), mean pause to mean chunk duration, speech rate (syllables/second, including pauses), articulation rate (syllables/second excluding pauses), F0 range (in semitones, and as a ratio of mean F0 maximum of focused words to mean F0) as well as number of focus positions. Further, minimum, maximum and mean F0 and mean of F0 maximum of focused words were measured separately for the male and female speakers. (Focused words were identified by the first author through listening.)

In the following we concentrate on those features correlating strongly with "being a good speaker". With this restriction, all duration and also all rate measurements will be excluded, as all show insignificant and very weak correlations. That is, from our data we have to conclude that the temporal features extracted do not have a relationship (at least not a simple one) to the overall rating of a speaker along the good/less good dimension. Interestingly, in the cross-cultural comparisons in [1], correlations between charisma ratings and speech rate varied between positive and negative depending on the raters' native language as well as the language rated.

4.1. F0 measures including focus

Table 1 summarizes focus and F0 measures and their correlations with the good-speaker ratings.

Table 1. *F0 measures and their correlations with the good-speaker ratings.* *=p<.05; **=p<.01.*

Feature		max	min	r	p
Focused words (N)		23	3	.47	.07
Mean F0 max of foc words/mean F0		1.85	1.17	.61	.01*
F0 range, 75-25 percentiles (ST)		8.78	2.44	.61	.01*
Mean F0 max of focused words	female	384	219	.87	.005**
	male	233	150	.76	.03*
F0 max	female	516	265	.82	.01*
	male	325	190	.63	.09
F0 mean	female	246	180	.62	.10
	male	164	113	.65	.08

As can be seen, the number of focused words varies considerably and there is a positive correlation of .47 between the good-speaker rating and the number of focused words. Thus, we cannot exclude frequency of focusing to have some importance, although the correlation does not reach significance ($p=.07$) in our material.

Correlations which *are* significant and positive ($p<.05$) include F0 range, measured in semitones and also as a ratio between mean F0 maximum of focused words and mean F0 ($r=.61$ in both cases). As the two measures follow the same trend, we confine ourselves to the semitone calculation in the following. Ranges in semitones between the 25% and 75% points in the F0 distribution vary between 2.44 and 8.78 for individual speakers, and the median is 4.7. These figures may be compared to similarly computed ranges (25% -75% points) extracted from 314 prompted utterances from each of 498 speakers in the Swedish SpeeCon database collected at KTH [2]. These (ordinary, non-professional) speakers had a range of 2-5 semitones, except for a few cases with smaller or greater ranges than the majority. Only about half of our speakers' ranges then fall within the 2-5 semitone interval, while the other half have ranges of a greater magnitude.

There is moreover a significant correlation with F0 maximum for the female speakers ($r=.82$; $p<.05$) and with the mean of F0 maximum of focused words as well ($r=.87$; $p<.01$). For the male speakers, only the correlation with the mean of F0 maximum of focused words shows a correlation ($r=.76$; $p<.05$) with the good-speaker rating. This latter feature is comparable to the pitch range measure (mean HiF0, the highest accented pitch peak) in [1]. Mean HiF0 correlated positively with charisma in American English and Palestinian Arabic when rated by American and Palestinians as well as Swedish subjects. This feature in addition was found to be more important for the Swedish subjects than for the others; data suggested that Swedish subjects found higher pitched speech to be more charismatic than did Americans and Palestinians.

Thus apart from minor differences, F0 dynamics appear to influence the impression of charisma/good speaking across widely different languages. Also, for Swedish F0 dynamics is primarily associated with the extent to which the range is widened upwards; the correlation with F0 minimum is weak and similarly insignificant for the female and male speakers. Neither is mean F0 over the entire speech sample significant, although both female and male speakers' correlations exceed .6 with $p=.08$ and .10, respectively in our study. Biadsy and

co-workers in [1], on the other hand, found significant correlations between mean F0 and charisma ratings, irrespective of the language of the raters, for both American and Palestinian Arabic speech materials.

5. Fluency and speaking style

Some speakers were perceived as altogether fluent while others were more or less disfluent. Accordingly, to make possible a correlation with the good-speaker rating a measure of disfluency was calculated; the number of positions with a slip of the tongue, a repetition or a repair was determined through listening by the first author combined with inspection of the speech wave. This measure showed a strong negative correlation ($r=-.72$; $p<.01$) with "being a good speaker". A similar negative correlation was found also in the cross-cultural comparison in [1] with the exception of the Swedish judgments of American English. We note this difference between Swedish judgments of Swedish and English, respectively, which *may* be ascribed to cultural influences.

Disfluencies occur primarily in speech produced spontaneously as a result of problems with the planning of what to say next. As some of our speakers read from a manuscript and some spoke more freely, we could relate the disfluency scores to the read vs. spontaneous style of speaking. Even though the three most disfluent speakers were speaking spontaneously, there was no obvious relation taking all speakers into account. Neither was there any obvious relation between speaking style and the good-speaker rating.

6. Synthesis

In the acoustic analysis, F0 features, in particular a wide F0 range and high peaked focused words, were found to give high ratings of "good speaker", while the opposite, a smaller range and focused words with lower peaks was given low ratings. Also, the good speakers were to a great extent fluent, while the less good ones had lots of repetitions, repairs etc. These results were elaborated in a resynthesis experiment in which the sample of the speaker with the lowest score (0,56) for "good speaker" was modified in several ways. The assumption was that, relying on our production results, we could improve the perceived skill of speaking.

First however, we describe the selected (male) speaker in some detail. Concerning ratings of speaker qualities, his scores were at the high end of the scale for *insecure*, *hesitant*, *monotonous* and at the low end for *expressive*, *powerful*, *aggressive* and *trustworthy*, while he, on the other hand, got the highest score of all the speakers for *humble*. He is further the second most disfluent speaker with a total of 12 disfluency positions, and regarding F0 he has the smallest range (2,84 ST), F0 maximum (190 Hz), and mean F0 maximum of focused words (150Hz). This speaker also is the slowest with a speech rate (including pauses) of 3.46 syllables per second. Thus, on the basis of the characterization above, this speaker is a natural candidate for the synthesis experiment.

6.1. Hypotheses

The features to be evaluated first of all included the two found to have the highest correlations (positive and negative, respectively) with being a good speaker: F0 dynamics, and fluency. As the selected speaker was extremely slow, we also included speech rate (although the speech rate features overall correlated insignificantly with "good speaker").

We hypothesized that of these features, rate would be the least effective for improvement of speaker skill. Concerning the other two, there is little to base an assumption on regarding their respective perceptual strength, which gives us two alternative hypotheses a) and b). There might also be interactions between the features, giving a third alternative:

a) F0 dynamics > fluency > speech rate

b) fluency > F0 dynamics > speech rate

c) F0 dynamics, fluency and speech rate interact

6.2. Stimulus preparation and experimental setup

To create the experimental stimuli, we used the KTH resynthesis toolkit EXPROS [4] together with the Mbrola diphone synthesis toolkit [3]. This was a three step process: first the EXPROS toolkit was used to automatically generate the data needed to build a new Mbrola diphone database from the original speech sample (36 seconds in length). Then the Mbrola toolkit was used to build a customized Mbrola mini-voice. Finally, EXPROS was used to modify the prosodic features of the original speech sample. The following three manipulations were performed:

- F0 dynamics: The pitch scale was transformed to a semitone scale. The mean pitch was increased by two semitones and the range was expanded, so that the standard deviation was doubled.
- Fluency: Reduction of disfluencies were made by cutting out slips of the tongue and repetitions.
- Speech rate: Speech rate was increased by 5% and long silent hesitation pauses were considerably shortened.

Thus, there were eight stimuli (2 x 2 x 2) including all combinations of original and modified F0 dynamics (orig/mod F0), fluency (orig/mod fluency) and speech rate (orig/mod rate).

The task of the 12 subjects participating in the experiment, all academic teachers or advanced students in areas other than phonetics, was to make a rank ordering of the eight versions. They did so using an interactive computer program implementing a visual sort and rate/rank method.

Each of the eight versions was represented by an icon in random order on the computer screen. The subjects were instructed to rank them from best (1) to worst (8) in reference to the criterion for "a good speaker", that is, a person capable of catching the attention of an audience through her/his way of speaking", according to the definition used here. Before coming up with the ordering they preferred, the subjects could listen to the stimuli as many times as they wished and try different rankings by moving the icons around.

6.3. Results

In judging this kind of complex phenomena, we cannot expect total uniformity between subjects' rankings. Despite this, there was a fair degree of consistency; the correlation between subjects (Kendall's W) was .48 ($p<.001$).

The results support the general assumption that perceived speaker skill can be improved by modifications such as those suggested by our production data. The general trend is that the more modifications, the higher the ranking. There are considerably more high rankings than low for modified versions. This is demonstrated in Figure 1a), which shows the number of judgments for the modified versions of F0 dynamics, fluency and speech rate grouped by rank 1-4 and 5-8 respectively. The results in addition suggest that F0 modifications play the major role, with modifications of fluency and speech rate being second and third. Figure 1b) presents more detailed data showing the judgments for each rank separately. The results pooled across all subjects then come close to an

ordering according to hypothesis a), that is, F0 dynamics > fluency > speech rate in terms of perceptual weight.

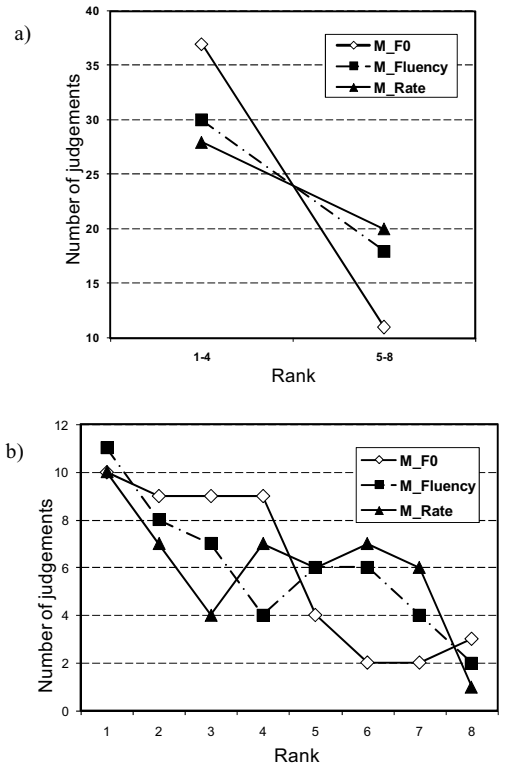

Figure 1: *Judgments for the modified versions of F0 dynamics (M_F0), fluency (M_Fluency) and speech rate /M_Rate) separated between versions ranked 1-4 and 5-8 (a) and 1 through 8 (b).*

An even more detailed analysis reveals interesting differences between subjects. Some had ratings reflecting a completely systematic ordering of the features in accordance with hypothesis a), while others were less "systematic". This is not unexpected, as making judgments about such a phenomenon as speaker skill most reasonably is not a simple task. The features under investigation may be expected to interact in complex ways, but individual experiences and preferences may also play a role. Several of the subjects after the test spontaneously commented on their impressions of the speech stimuli. Some of them, for example, found slips and other disfluencies to be very disturbing, while others looked upon the same phenomena as something natural and more or less unavoidable. Still most of them, according to the general result, favored a modified F0 range and some reported that they very easily could divide the eight versions in two groups (original and modified F0 dynamics, respectively), but that priorities within these groups were much more difficult.

7. Conclusions and future work

In this three-part study, we aimed at uncovering features contributing to the impression of a "good speaker", that is, "a person capable of catching the attention and interest of an audience through her/his way of communicating". Using speech samples from a number of speakers chosen so as to vary in speaking skill, ratings of each individual sample were made on a number of qualities in addition to an overall rating of the skill of speaking. Acoustic measurements of the samples combined with the overall ratings revealed strong correlations between the ratings and some acoustic features in particular; F0 peak height of focused words and F0 range on the one hand turned out to correlate positively with being a good speaker, while on the other hand, the correlation with disfluency was negative. Concerning F0 range we found that of our speakers – politicians used to speak before great audiences – those with the highest overall ratings were considerably more dynamic than speakers in general; half of our speakers had a range exceeding that of the great majority of the "ordinary" speakers analyzed in [2]. Features associated with F0 dynamics also influence ratings of charisma [1, 7] and liveliness [5, 10], which additionally points to the importance of F0 variability for being a good speaker.

The resynthesis experiment gives even more support for such a dependence. We found that by increasing F0 dynamics, eliminating disfluencies and hesitation pauses, and speeding up the speech, the impression of speaker skill improved considerably. Modifying F0 dynamics produced the greatest effects and changes of disfluencies and speech rate, respectively, ranked second and third in terms of perceptual weight.

Combined with more acoustic data, resynthesis evaluations like the one conducted here could shed further light on what makes a speaker a *good* speaker. Also, in addition to showing the perceptual relevance of acoustic cues to speaker skill, the results from the synthesis experiment open up for useful applications, for example speaker training.

8. Acknowledgements

We thank John Lindberg and Roberto Bresin, KTH, for making the evaluation software available for the perceptual ranking. We also thank all subjects for their participation in the perceptual evaluation and in the rating procedure. The work has been supported by funding from the Swedish Research Council.

9. References

[1] Biadsy, F., Rosenberg, A., Carlson, R., Hirschberg, J. and Strangert, E., "A Cross-Cultural Comparison of American, Palestinian, and Swedish Perception of Charismatic Speech", To appear in Proc. Speech Prosody, Campinas Brazil, 2008.

[2] Carlson. R., Elenius, K. and Swerts, M.. "Perceptual Judgments of Pitch Range", Proc. Speech Prosody 2004, Nara, Japan, 689-692, 2004.

[3] Dutoit, T., Bataille, F., Pagel, V., Pierret, N., Van der Vreken, O. "The MBROLA Project: Towards a Set of High-Quality Speech Synthesizers Free of Use for Non-Commercial Purposes", Proc. ICSLP 96, Philadelphia, 1996.

[4] Gustafson, J. and Edlund, J., (in press) "expros: a toolkit for exploratory experimentation with prosody in customized diphone voices" To be published in Proceedings of the 4th IEEE Workshop on Perception and Interactive Technologies for Speech-Based Systems. Kloster Irsee, Germany.

[5] Hincks, R., "Computer Support for Learners of Spoken English", Diss. Speech and Music Communication, KTH, 2005.

[6] Liscombe, J., Venditti, J. and Hirschberg, J., "Classifying subject ratings of emotional speech using acoustic ratings". Proc. Eurospeech 2003, Geneva, Switzerland, 2003.

[7] Rosenberg, A. and Hirschberg, J., "Acoustic/prosodic and lexical correlates of charismatic speech", Proc. Interspeech 2005, 513-516, 2005.

[8] Strangert, E., "Prosody in public speech: Analyses of a news announcement and a political interview", Proc. Interspeech 2005, Lisboa, Portugal, 3401-3404, 2005.

[9] Strangert, E., "What makes a good speaker? Subjective ratings and acoustic measurements", Proc. Fonetik 2007, KTH, Sweden, 29-32, 2007.

[10] Traunmüller, H. and Eriksson, A., "The perceptual evaluationof F_0 excursions in speech as evidenced in liveliness estimations", JASA, 97 (3):1905-1915, 1995.

[11] Wichmann, A., "Attitudinal intonation in the inferential process", Proc. Speech Prosody 2002, Aix-en-Provence, 2002.