EVALUATION OF PROSODIC CHARACTERISTICS IN RETOLD STORIES IN DUTCH BY MEANS OF SEMANTIC SCALES

Monique E. van Donzel and Florien J. Koopmans-van Beinum
University of Amsterdam, Institute of Phonetic Sciences/IFOTT
Herengracht 338, 1016 CG Amsterdam, The Netherlands
tel: +31 20 525 2183, fax: +31 20 525 2197, email: vandonzel@fon.let.uva.nl

ABSTRACT

This paper describes an experiment in which listeners were asked to evaluate various prosodic aspects in retold stories in Dutch, using semantic scales. The aim was to see what features on prosodic level listeners prefer when listening to a retold story in Dutch, and if 'good' and 'bad' speakers can be distinguished in this respect. Results from a factor analysis show that listeners use *Voice appreciation*, *Dynamics*, and *Articulation quality* as main cues in evaluating the retold stories.

1. INTRODUCTION

People tell stories. When listening to someone telling a story, one usually has an opinion on the quality of the telling, in terms of coherence and content of the story, but also on the acoustic-prosodic properties of the speech signal itself, for instance voice or articulation quality, or the amount in which their way of presenting is pleasant to listen to. Listeners should therefore be able to indicate whether a speaker is 'good' or 'bad' in performing his/her task. In this paper we will focus on the *acoustic-prosodic* properties of speech, rather than on aspects involving the way in which the content of a story is rendered.

The experiment reported on in this paper is part of a larger project on the acoustic determinants of focusing in discourse. Within this project, earlier experiments on pausing strategies [1] and speaking rate [2] revealed substantial speaker differences. Furthermore, the discourses they produced (spontaneous speech retelling a read story) differed of course in discourse structure, both on a global (phrasing) and on a local (focal structure) level. These findings led us to investigate the question of how listeners perceive these spontaneously retold stories prosodically. In other words, how do listeners evaluate the way in which the retelling of a story is realized on an acoustic-prosodic level? Or, formulated in a more general frame, what is the perceptual structure behind listening to a retold story, apart from the content?

In order to derive this information, we used semantic scales [3]. This means is very suitable to obtain judgements from listeners, both on 'introspective opinions' (how they would like the ideal speaker to sound) and on perceptual scores (how they evaluate a specific speaker). These data will in a later stage be related in greater detail to additional acoustic measurements, such as speaking rate, intonational characteristics, temporal aspects, and on a more linguistic level to discourse structure.

2. METHODS

2.1. Speakers, stimuli, and listeners

Four male and four female native speakers of Dutch were selected as speakers. They were all students or staff members of the Institute of Phonetic Sciences. The speakers were asked to read aloud a short story in Dutch ('A Triumph' by S. Carmiggelt [4]). After a short break they were asked to retell the same story in their own words, as detailed as possible. During the retelling of the story a listener was present to create a more natural story telling situation. This resulted in eight spontaneously retold versions of the same story (hereafter 'retold version'). All recordings were made in an anechoic room on DAT-tape. The retold versions were stored as digitized audio files (sample rate 48 kHz, 16-bit precision).

Twenty-three listeners (18 female, 5 male) participated in the listening experiment. They were explicitly not students or staff members of our own Institute to make sure that the listeners did not know the speakers personally. The listeners were paid for their participation. The experiment was performed in the language laboratory of the Faculty of Arts of Leiden University.

3. LISTENING EXPERIMENT

3.1. Materials and procedure

The eight retold versions were put on individual audio tapes, in two different orders to account for listing effects. Each listener was given his/her own tape, to enable the listener to work in his/her own tempo as accurately as possible. Each tape contained a retold version to be used for practice, followed by the retold versions of the eight speakers, either in order A (12 listeners), or in order B (11 listeners).

Listeners were first asked to perform a 'paper-and-pen' task, in which they had to indicate how the 'ideal speaker' should sound according to them, using 30 7-point semantic scales. These scales were selected from a set used by Boves [5], but included the 14 scales of the eRelative Speech Appreciation profilei determined by

Fagel et al. [6]. As a second part of the task they were asked to judge the practice version and the eight retold versions for various prosodic aspects, using the same 30 7-point semantic scales.
Furthermore, after having evaluated the 30 scales for one specific speaker, they had to give an overall judgement of the prosodic aspects of the retelling task on a 10-point scale for that particular speaker. Then the next speaker was evaluated.
There was a separate answer sheet for the practice version and for each of the eight retold versions. Listeners took approximately 45 minutes to fulfill the task.

4. RESULTS

4.1. Overall judgement

First of all, we will look at the overall judgement of the retelling task as given by the listeners on a 10-point scale, for the eight 'real' speakers. The mean overall judgement scores and standard deviations are given for each of the speakers in Table 1.

Table 1. Mean overall judgement scores and standard deviations for the eight speakers (Female or Male).

Sex	F	M	F	M	F	M	F	M
Spkr	1	2	3	4	5	6	7	8
Mean	6.6	5.8	7.9	6.4	6.9	4.9	7.3	6.6
sd	1.2	1.7	.6	1.0	1.0	1.3	1.2	1.3

The data clearly show that speaker 3 is evaluated as the best speaker in the retelling task. Her score is the highest overall, with the lowest standard deviation, indicating that listeners agree fairly well. As for the other speakers, we see that only one speaker is evaluated as 'insufficient' (below 5.5, speaker 6). Furthermore, standard deviations are fairly high, which means that listeners scored on a wide range. If we look at the score for speaker 2 for instance (5.8), we may conclude from the sd of 1.7 that some listeners evaluated his retelling as fairly good, while others judged it as rather bad.

4.2. Scale judgements

The next step was to test the reliability of the semantic scales used in this experiment. Since we are interested in the most reliable scales, we computed a reliability value for each of the 30 scales, Cronbach's α (($MS_{(between)}$ - $MS_{(residual)}$) / $MS_{(between)}$). A minimum value of .80 is generally assumed to indicate reliability. For the majority of the scales α exceeded .80, which indicates that they are reliable. Seven scales had a value below .80, and were excluded from further analyses. Five of these seven all involved aspects on voice quality proper, such as 'artless-affected', 'creaky-not creaky', 'rough-smooth', 'tense-relaxed', and 'deviating-normal'. The other two unreliable scales were 'pleasant-unpleasant' and 'friendly-curt'. Apparently, these scales are not very useful for the listeners in evaluating the prosodic characteristics of the retold stories. Pearson's pairwise comparison showed that these last two 'general' scales (epleasanti and efriendlyi) correlated with the overall judgement, thus the higher the overall judgement the more the judgement was epleasanti or efriendlyi. The overall judgement correlates higher with the epleasanti scale than with the efriendlyi scale (-0.90 and -0.70 resp.). In the rest of the paper, only the 23 reliable scales will be considered.

The remaining 23 scales were used in a Principle Components Analysis, to decompose the correlation matrix into (varimax rotated) factors. The number of factors is determined by the criterion 'eigenvalue > 1'. Table 2 at the end of this paper shows the factors extracted in the factor analysis, and the corresponding scales for each factor. Cells with loadings higher than .55 are presented in gray. The five factors together explained 74% of the total variance.

Table 2 clearly shows that not all scales load on only one factor. We see however a very clear clustering in the groups of scales. The scales that load on the first factor (which explains 48% of the variance) all represent instances of *appreciation of voice variability characteristics*. The second factor (explaining 9%) represents *dynamics*. The third dimension (explaining 6%) concerns the pronunciation or *articulation quality*, whereas the fourth and the fifth (both 5%) seem to account for *pitch* and *voice abnormality* aspects respectively. The factors extracted in our analysis are in accordance with earlier studies on the evaluation of voice and pronunciation characteristics for Dutch [5,7].
On the basis of these results, we may conclude that the first three factors are most important. The fourth factor consist of only two scales; the fifth of only one scale corresponding to 'abnormality'.

4.2.1. Ideal speaker

We want to know how the listeners think the 'ideal speaker' should sound. Table 3 (see next page) shows the listeners' judgements on the various 7-point scales for the 'pen-and-paper' task. For a clear legibility, we present all positive aspects on the right hand side of each scale. Low scores indicate that the preferred aspect is on the left side of the scale, high scores indicate that it is on the right side. Low standard deviations furthermore indicate that agreement among the listeners is rather high, whereas high standard deviations show that listeners do not agree very much. The standard deviation is always between 0.5 and 1.3 scale judgement.

The data for the ideal speaker show that listeners have a very clear picture in mind of how a speaker should sound when retelling a story in Dutch. The listeners used both extremes of the scales very clearly. Some aspects

however are judged more extremely than others, such as varied, sonorous, beautiful, vivacious, and cheerful. Standard deviations are generally rather low for these scales (<1). Furthermore, the scales loading on factor 1 have been judged more extremely than those loading on factor 2. Factor 3 has rather extreme scores again.

4.2.2. Real speakers and speaker differences

For the eight 'real' speakers scores averaged over listeners were all between 3.0 and 5.0, and are thus not as extreme as for the ideal speaker. Standard deviations range between 1.1 and 1.7, and thus show that both extremes of the scales were used. Due to space limitations, we will not present the data for each speaker separately for each scale.

Table 3. Mean scores and standard deviations for the ideal speaker on the 23 reliable scales. Scales mainly loading on one factor (see Table 2) are separated by a dashed line. The positive aspect is presented on the right hand side of each scale.

Semantic scale	Mean score	sd
stereotyped-varied	5.7	.9
passive-active	4.9	.8
colourless-sonorous	5.9	.8
monotonous-melodious	5.2	.7
ugly-beautiful	5.5	1.0
spiritless-vivacious	6.5	.5
poor-rich	5.2	.9
expressionless-expressive	5.3	.8
whining-cheerful	6.0	.8
soft-loud	3.5	.8
slack-firm	4.5	.7
dragging-brisk	5.8	.7
weak-powerful	4.6	.8
unsteady-steady	4.1	1.0
wavering-selfconfident	4.5	.5
slow-quick	3.9	.8
careless-precise	5.0	.9
broad-cultured	5.0	1.1
slovenly-polished	4.3	1.2
indistinct-distinct	5.2	.9
shrill-deep	3.7	.7
agitated-calm	4.1	1.1
husky-not husky	5.5	1.3

In what way do the actual speakers, as evaluated by the listeners, correlate with the ideal speaker on each of the 23 scales? Pearson's correlation for pairwise comparison is highest for speaker 7 (0.91), followed by speaker 3 (0.85), and high but negative for speaker 6 (-0.80) (see Table 4). These figures are roughly as expected on the basis of the overall judgements (see Table 1) compared to the judgements given for the ideal speaker.
Comparison *between* speakers revealed that speaker 3 correlates with speakers 1 (0.84), 5 (0.88), and 7 (0.77). These 4 speakers (all female) are also the four best speakers in the overall judgement task. Correlations between speaker 6 (the lowest overall judgement) and speaker 2 are high (0.71).

Furthermore, we want to know how the actual speakers correlate with the ideal speaker *per factor*, to see if some kind of 'speaker profile' can be determined. Table 4 shows the Pearson's correlations for pairwise comparison between the ideal speaker and the individual eight 'real' speakers (*Overall*), and for each of the three most important factors *Voice appreciation*, *Dynamics*, and *Articulation quality*.

The data from Table 4 show that not all factors are equally important for all speakers. For instance speaker 4 scores highly (0.95) only on the 'Articulation'-factor, whereas for speaker 1 the 'Dynamics'-factor is least important. Speakers 3, 5, and 7 score highly on all factors.
A closer look at the three factors separately shows that for the 'Voice appreciation'-factor speakers are more or less divided into two groups: correlations with the ideal speaker are very high for speakers 1, 3, 5, and 7, and highly negative for speakers 2, 6, and 8. Correlations for speaker 4 are moderate.
For the 'Dynamics'-factor there is much variation between the different speakers. This factor has high positive correlations for speakers 3, 5, and 7, but high negative correlations for speakers 2 and 6. This means that for this factor these last two speakers are far from ideal.
The 'Articulation'-factor shows that correlations are positive for all speakers, except for speaker 6. This speaker scores negatively on all factors, and is clearly evaluated as the worst speaker, and resembles the ideal speaker in no way.
In future experiments, where other acoustic aspects will also be included, we expect to find a relation between these acoustic aspects and the evaluation scores for different factors. Intonational phenomena will then be related to scores for scales as 'varied', 'melodious', 'vivacious', and 'expressive' (Factor 1). Durational aspects will be related to scales loading on the 'Dynamics'-factor. Spectral and possibly also durational aspects are expected to relate to scales loading on the 'Articulation'-factor.

Table 4. Correlations between the ideal and the 'real' speakers, broken down per factor.

	Ideal speaker			
Real speakers	Voice appreciation	Dynamics	Articulation	Overall
1	0.96	0.39	0.98	0.78
2	-0.86	-0.76	0.59	-0.39
3	0.97	0.88	0.97	0.85
4	0.37	0.29	0.95	0.29
5	0.97	0.93	0.99	0.85

6	-0.98	-0.88	-0.78	-0.80
7	0.96	0.96	0.99	0.91
8	-0.84	0.71	0.95	0.30

5. CONCLUSION

The results from the listening experiment show first of all that listeners have a clear picture of how the 'ideal speaker' of a spontaneous story should sound. For most prosodic aspects, presented as semantic scales, they agree fairly well. Secondly, they have specific judgements about the acoustic realization of retold stories in Dutch. Some speakers are clearly evaluated as 'better' than others, not only in the overall judgement, but also on the separate semantic scales.

Listeners clearly make a distinction between 'good' and 'bad' speakers. They have a picture in mind of how the ideal speaker should sound, and the closer a real speaker comes to this picture, the better he/she is evaluated. This means that the voice of a speaker retelling a story is an important aspect, and should be taken into account. *Voice variability characteristics* are most important in this respect (Factor 1), followed in importance by *Dynamics* (Factor 2) and *Articulation quality* (Factor 3). This ordering may be different for one speaker to the next, as can be seen in the 'speaker profile' in Table 4.

The results from the present experiment can to some extent be related to specific acoustic aspects, such as pausing strategies, but obviously other aspects need to be taken into account as well. More data on speaking rate will be included at a later stage, as well as detailed analyses of (production and perception of) intonational phenomena (cf. [8,9]). The findings can be of use to improve synthetic speech, especially where spontaneously sounding output is required. Furthermore, reading machines for the blind could benefit from good naturally sounding synthetic speech output. Depending of course on the content and the type of the material to be synthesized, a 'spontaneous' sounding voice can be more appropriate.

6. ACKNOWLEDGMENTS

The authors would like to thank Vincent van Heuven of the Phonetics Laboratory, Leiden University, for creating the opportunity to use the language laboratory. Thanks also to Louis Pols for careful reading of and useful comments on earlier versions of this paper.

7. REFERENCES

[1] M.E. van Donzel and F.J. Koopmans-van Beinum, "Pausing strategies in discourse in Dutch", *Proceedings ICSLP96*, Philadelphia, vol. 2, pp. 1029-1032, 1996.

[2] F.J. Koopmans-van Beinum and M.E. van Donzel, "Relationship between discourse structure and dynamic speech rate", *Proceedings ICSLP96*, Philadelphia, vol. 3, pp. 1724-1727, 1996.

[3] C.E. Osgood, G.J. Suci and P.H. Tannenbaum, *The measurement of meaning*. University of Illinois Press, Urbana, 1957.

[4] S. Carmiggelt, *Fluiten in het donker*, ABC Boeken, Amsterdam, 1966.

[5] L. Boves, *The phonetic basis of perceptual ratings of running speech*, Foris Publications, Dordrecht, 1984.

[6] W.P.F. Fagel, L.W.A. van Herpt and L. Boves, "Analysis of the perceptual qualities of Dutch speakers' voice and pronunciation", *Speech Communication* 2, pp. 315-326, 1983.

[7] M. Tielen, *Male and female speech*, Doct. diss. University of Amsterdam, 1992.

[8] M.E. van Donzel and F.J. Koopmans-van Beinum, "Pitch accents, boundary tones, and information structure in spontaneous discourse in Dutch", to appear in *Proceedings of the ESCA Workshop on Intonation*, Athens, 18-20 September, 1997.

[9] M.E. van Donzel, "Perception of discourse boundaries and prominence in spontaneous Dutch speech", to appear in Working Papers 46, Department of Linguistics and Phonetics, Lund University, 1997.

Table 2. Results of the factor analysis for the eight speakers and the selected 23 scales in five dimensions (% variance explained between brackets). Scales are grouped per factor. Cells with loadings higher than |.55| are presented in gray.

Semantic scale	Factor 1 (48%) evoice characteristicsi	Factor 2 (9%) 'dynamics'	Factor 3 (6%) 'articulation quality'	Factor 4 (5%) 'pitch'	Factor 5 (5%) 'voice abnormality'
stereotyped-varied	.765	-.041	.248	-.082	-.027
active-passive	-.700	.448	-.073	.117	-.105
colourless-sonorous	.765	-.274	.159	-.098	.304
melodious-monotonous	-.841	.133	-.193	.130	-.075
ugly-beautiful	.675	-.131	.493	.127	.010
spiritless-vivacious	.732	-.398	.268	-.245	-.040
poor-rich	.702	-.286	.387	.071	-.034
expressive-expressionless	-.809	.278	-.190	.156	-.028
whining-cheerful	.688	-.451	.160	-.230	.030

loud-soft	-.192	.656	.121	.073	-.534
firm-slack	-.494	.624	-.253	.030	-.034
dragging-brisk	.500	-.564	.285	-.334	-.009
powerful-weak	-.470	.647	-.351	.156	-.022
steady-unsteady	-.162	.681	-.391	-.094	.032
selfconfident-wavering	-.313	.570	-.472	.020	.163
quick-slow	-.273	.622	-.162	.476	.188
careless-precise	.294	-.127	.768	.003	.022
broad-cultured	.117	-.087	.819	.014	.088
polished-slovenly	-.262	.225	-.827	.060	-.081
distinct-indistinct	-.314	.346	-.612	.026	.007
deep-shrill	.022	.043	-.026	-.825	.245
agitated-calm	-.289	.223	.044	.744	.253
husky-not husky	.095	.060	.133	-.009	.851