Institute of Phonetic Sciences,
University of Amsterdam,
Proceedings 15 (1991), 109-120.

PERCEPTUAL EVALUATION OF VOICE QUALITY BEFORE AND AFTER RADIOTHERAPY OF PATIENTS WITH EARLY GLOTTIC CANCER AND OF NORMAL SPEAKERS

Irma M. de Leeuw

Abstract

The semantic scaling experiment presented in this paper is part of a pilot study to obtain parameters that can describe voice quality of patients with early glottic cancer before and after radiotherapy and of normal speakers. Perceptual evaluation by untrained listeners is one of the measurements that will be used; the aim of this experiment is to obtain a reliable set of semantic scales that can be used to describe voice quality. Voice samples (read aloud text and sustained /a/) of 4 patients before radiotherapy, of 5 patients 6 months after radiotherapy, and of 5 normal speakers are judged by 24 untrained listeners on 22 scales. Reliability coefficients are calculated and factor analyses are carried out on the reliable scales. It appears that voice quality can be described with 16 scales in a 6-dimensional space for the read aloud text and with 10 scales in a 4-dimensional space for the sustained /a/. Finally, factor scores for the 14 voices are calculated for further research.

1 Introduction

Within the scope of a cooperative research project with the Netherlands Cancer Institute (Antoni van Leeuwenhoek Hospital), the Academic Hospital of the Free University of Amsterdam, and the Institute of Phonetic Sciences of the University of Amsterdam, the present author is charged with aspects of voice quality. Central to the whole project is a study on dose response in radiotherapy of early glottic cancer that will be carried out at the Netherlands Cancer Institute by radiotherapist G. Baris. The purpose of that study is to determine the optimal radiation dose to be delivered to small glottic tumours. Optimal radiation dose should be based upon dose response curves for tumour control and upon complications of the radiation on normal tissue. One of these complications causes decrease of voice quality.

In the present paper a pilot study on perceptual evaluation of voice quality of untrained listeners is described as part of a larger study that involves clinical, acoustic, and perceptual parameters to describe voice quality.

Clinical methods involve tests related to singing fundamental frequency and sound pressure level (phonetogram), to speaking fundamental frequency, to aerodynamic efficiency (phonation quotient, phonation flow) and to vocal fold vibration (evaluation of stroboscopic recordings, electroglottogram). A detailed description of these methods is given in de Leeuw (1990).

Acoustic analyses might involve long-term average spectrum analyses, F0-analyses, and jitter and shimmer measurements; the actual approach still has to be decided.

For the perceptual evaluation of speech characteristics, trained as well as untrained listeners will be used. Perceptual evaluation of trained listeners is most frequently done in clinical settings (Laver,1980). That is the reason why judgements of trained listeners will be used in this project as well; their task will be to provide an analytic description of voice quality. However, in order to obtain more associative judgements, in the present experiment untrained listeners are used.

The aim of the experiment presented in this paper is to construct an instrument which will give a reliable evaluation of voice quality by untrained listeners. The instrument used to obtain judgements of untrained listeners consists of a set of semantic bipolar seven-points scales. In various experiments on normal voices (Fagel & Van Herpt, 1983; Van Herpt, 1986) it appears that by means of 14 scales voice and pronunciation can be described in a 5-dimensional perceptual space: Voice Appreciation (Melodiousness and Evaluation), Articulation Quality, Voice Quality (Clarity and Strength), Pitch, and Tempo.

In a previous experiment (de Leeuw, 1990) these 14 scales have been used to relate perceptual and clinical parameters of voice quality of patients with early glottic cancer before and after radiotherapy and of normal speakers. The scale *not intelligible-intelligible* was added to the 14 scales in order to give a description of pathological voices. One conclusion in that experiment was that other scales should be added in order to get a better description of the abnormality of pathological voices. Another conclusion was that the attempt to relate perceptual and clinical parameters was made too prematurely; first the nature of the perceptual scales had to become more clear .

In order to obtain a reliable set of scales which can describe normal voices as well as pathological voices (i.e. voices of patients with early glottic cancer before and after radiotherapy) a number of scales is chosen that can describe the pathology in voices. These scales are added to the 15 scales that were used in our earlier research. Listeners are asked to give their ratings on the resulting 22 scales. Reliability coeffients are calculated for every scale; finally, the most reliable scales will be used in factor analyses in order to reduce the number of variables. The resulting factors and the factor scores of every voice will be used as variables in further research .

This procedure is performed on two kinds of speech material: read aloud text and sustained /a/. In our earlier research we also used read aloud text. The reason to use the sustained /a/, is that this utterance is used in other measurements, like phonation quotient/flow, phonetogram, and vocal fold vibration as well. In a later stage, perceptual judgements will be related to these clinical data. One question is whether raters give different ratings on the same semantic scales for read aloud text and sustained /a/.

2 Experiment

2.1 Speakers/speech samples

The speakers in this experiment are divided into three groups. Group 1 and 2 consists of the same 5 patients with early unilateral glottic cancer with no impaired cord mobility. Group 1 is the group of patients before radiotherapy, group 2 is the group of patients 6 months after radiotherapy. There are no speech recordings of patient nr 5 before radiotherapy, so group 1 actually consists of 4 speakers, whereas group 2

consists of all 5 speakers. Group 3 is a control group of 5 speakers without any known defects.
The matching between group 1 and 3 includes the following parameters: sex, age, as well as smoking habits. The matching did take place before radiotherapy. A review of the matching criteria is given in table 1.

Table 1. Matching criteria of patients (speakers 1-5) and control speakers (speakers 11-15) (smoking: number of cigarettes per day, according to the speaker; before radiotherapy --> after radiotherapy; the speakers are all male).

speaker	age	smoking	speaker	age	smoking
1	64	25 -> 25	11	64	25
2	62	12 -> 0	12	61	10
3	71	10 -> 0	13	67	10
4	50	5 -> 0	14	44	15
5	60	25 -> 2	15	58	30

The speakers read aloud a text of about 5 minutes and produced a sustained /a/. All the material was recorded using a Philips D6920 MK2 cassette recorder and a Philips N8214 microphone. Fragments of all texts (ca 1 min.) were copied in random speaker order to one reel tape; the sustained /a/ of all speakers were copied three times in a row and added after the fragments of the texts in a different speaker order. This tape was copied to 24 cassette tapes, one for each listener, by using a Revox A 77 tape recorder and a Tandberg TCD 310 cassette recorder.

2.2 Raters/rating procedure

The raters in this experiment are 24 female students (first year Speech Therapy). They are considered to be untrained in listening experiments. They were paid for their participation in this experiment. The raters received written instructions. First they heard examples (one sentence) of every voice in order to get a reference frame. After the examples the 14 fragments of read aloud texts were presented and the raters jugded voice quality on all 22 scales. Finally, the raters had to judge the sustained /a/ of all speakers. The tapes were presented binaurally via a cassette recorder and headphones. The raters listened to the tapes in three groups in a quiet room, separated from each other. Every rater had control over her own cassette recorder, so that she could take as much time as she wanted. On the average, the whole rating procedure (instructions + rating) took about 1 hour.

2.3 Semantic scales

The set of semantic scales consists of 15 scales that have been used in previous experiments (scales 1-15 in table 2). In order to obtain additional scales that could describe the pathology of the voices in this study, the audio tape was judged first of all by 2 experienced listeners according to an extended list of various scales. This list was gathered from different stages in previous research by van Herpt. The terms that could describe the pathology according to the 2 listeners resulted in another 7 scales: *creaky-not creaky*, *panting-not panting*, *tense-relaxed*, *not fluent-fluent*, *speaking with difficulty-speaking without difficulty*, *steady-unsteady*, and *deviating-not deviating* (scales 16-22 in table 2).

The same resulting 22 scales are used both for read aloud text as well as for the sustained /a/, although one can expect that some of the scales are difficult to rate for the sustained /a/ (e.g. *monotonous-melodious*, *slow-quick*). In order to be able to compare the results of the read aloud text and the sustained /a/, and to avoid listeners to confuse two different rating forms, it was decided to use all 22 scales for both types of speech material.

Table 2. Semantic scales as used to rate voice quality of patients with early glottic cancer before and after radiotherapy and of normal speakers.

Scale nr.	Scale	Dutch terms
1	*monotonous -- melodious*	*eentonig -- melodieus*
2	*expressionless -- expressive*	*uitdrukkingsloos -- expressief*
3	*unpleasant -- pleasant*	*onaangenaam -- aangenaam*
4	*ugly -- beautiful*	*lelijk -- mooi*
5	*slovenly -- polished*	*onverzorgd -- verzorgd*
6	*broad -- cultured*	*plat -- beschaafd*
7	*husky -- not husky*	*hees -- niet hees*
8	*dull -- clear*	*dof -- helder*
9	*soft -- loud*	*zacht -- luid*
10	*weak -- powerful*	*zwak -- krachtig*
11	*high -- low*	*hoog -- laag*
12	*shrill -- deep*	*schel -- diep*
13	*slow -- quick*	*langzaam -- snel*
14	*dragging -- brisk*	*traag -- vlot*
15	*not intelligible -- intelligible*	*slecht verstaanbaar -- goed verstaanbaar*
16	*creaky -- not creaky*	*krakerig -- niet krakerig*
17	*panting -- not panting*	*hijgerig -- niet hijgerig*
18	*tense -- relaxed*	*gespannen -- ontspannen*
19	*not fluent -- fluent*	*vloeiend -- niet vloeiend*
20	*speaking with difficulty -- speak -diff.*	*met moeite spreken -- zonder moeite sprek.*
21	*unsteady -- steady*	*onvast -- vast*
22	*deviant -- not deviant*	*afwijkend -- niet afwijkend*

3 Results

3.1.1 Reliability for read aloud text

A reliability coefficient is calculated for all 22 scales: Ru. This is a measure of the reliability of the means of the ratings given by a panel of raters (Asendorpf, 1979; van Bezooijen, 1987; van Erp, 1990). Ru is defined as:

$$Ru = 1 - \frac{MS_w}{MS_b}$$

in which MS_w = Mean Square within objects and MS_b = Mean Square between objects.

The results are given in table 3. It appears that the reliability of all scales is reasona-bly high. Therefore, all 22 scales will be taken for factor analysis in the next section.

Table 3. Reliability coefficient Ru, MS_w, and MS_b for the ratings of 24 raters on 22 scales on the voice quality of **read aloud text** of 14 speakers.

Scale nr.	Scale	Ru	MS_w	MS_b
1	*monotonous-melodious*	.93	2.22	33.13
2	*expressionless-expressive*	.93	2.16	30.48
3	*unpleasant-pleasant*	.96	1.97	51.58
4	*ugly-beautiful*	.97	1.28	46.06
5	*slovenly-polished*	.91	2.32	25.01
6	*broad-cultured*	.93	2.16	30.31
7	*husky-not husky*	.95	2.69	56.23
8	*dull-clear*	.95	1.49	31.48
9	*soft-loud*	.95	1.24	26.57
10	*weak-powerful*	.94	1.94	31.39
11	*high-low*	.96	1.55	44.27
12	*shrill-deep*	.94	1.56	25.71
13	*slow-quick*	.90	1.52	15.53
14	*dragging-brisk*	.90	1.76	17.61
15	*not intelligible-intelligible*	.91	1.96	21.77
16	*creaky-not creaky*	.91	2.64	29.08
17	*panting-not panting*	.90	2.01	19.64
18	*tense-relaxed*	.93	2.43	33.79
19	*not fluent-fluent*	.81	2.99	15.35
20	*speak. +dif. - speak. -dif.*	.95	2.36	48.24
21	*unsteady-steady*	.93	2.28	31.05
22	*deviant-not deviant*	.97	2.08	66.72

3.1.2 Factor analysis for read aloud text

To reduce the number of parameters, a factor analysis is carried out. The correlations of the mean ratings over the 14 voices on read aloud text were tabulated in a correlation matrix. The Principal Component Analysis (PCA) is used to decompose this matrix into factors. The initial factors are rotated to a varimax criterion (Wilkinson, 1989).

A common criterion for determining the number of factors is to retain factors with eigenvalue 'greater than 1'. When this criterion was applied the PCA produced 5 factors.

On the basis of the factor loadings (> .50), the factors are labeled as Abnormality (*dull-clear, ugly-beautiful, unpleasant-pleasant, husky-not husky, tense-relaxed, not fluent-fluent, speaking with difficulty-speaking without difficulty, unsteady-steady, deviant-not deviant*), Melodiousness/Strength (*monotonous-melodious, expressionless-expressive, soft-loud, weak-powerful*), Articulation Quality (*broad-cultured, slovenly-polished*), Tempo (*slow-quick, dragging-brisk*), and Pitch (*high-low, shrill-deep*).The factor Abnormality contains 9 scales, which is a rather high number of scales for a single factor. For that reason a selection is made on the basis of the reliability coefficients; the scales on the factor Abnormality with Ru < .95 will not be taken for further research: *tense-relaxed, not fluent-fluent*, and *unsteady-steady.*

Furthermore, the scales *not intelligible-intelligible*, *creaky-not creaky*, and *panting-not panting* will be left out too, because their factor loadings are below .50.

As a conclusion, the following scales will be taken for the final factor analysis on read aloud text: *monotonous-melodious, expressionless-expressive, unpleasant-pleasant, ugly-beautiful, slovenly-polished, broad-cultured, husky-not husky, dull-clear, soft-loud, weak-powerful, high-low, shrill-deep, slow-quick, dragging-brisk, speaking with difficulty-speaking without difficulty,* and *deviant-not deviant.*

With the remaining 16 scales a new factor analysis is carried out. In order to split the factor Melodiousness/Strength the factor analysis is forced into 6 factors. In this solution the criterium of eigenvalue 'greater than 1' cannot be hold. But, apart from the interpretability of the data, a more objective argument for a 6-dimensional solution is the fact that the percent of total variance explained drops from 9.3% of the sixth factor in a 6-dimensional space to 6.3% of the seventh factor in a 7-dimensional space. The results of the analysis, forced into 6 factors, are given in table 4.

On the basis of the factor loadings, the factors are labeled as Abnormality (*unpleasant-pleasant, ugly-beautiful, husky-not husky, dull-clear, speaking with difficulty-speaking without difficulty*, and *deviant-not deviant*), Melodiousness (*monotonous-melodious, expressionless-expressive*), Pitch (*high-low, shrill-deep*), Articulation Quality (*slovenly-polished, broad-cultured*), Strength (*soft-loud, weak-powerful*), and Tempo (*slow-quick, dragging-brisk*).

Table 4. Significant factor loadings (>.50) and percent of total variance explained after varimax rotation of the 16 scales in each of the 6 factors of **read aloud text**; all loadings are significant at the 1% level.

factor	1	2	3	4	5	6
% of total variance explained after rotation	24.7	11.9	10.7	10.5	9.5	9.3
scale						
monotonous-melodious		.84				
expressionless-expressive		.83				
unpleasant-pleasant	.70					
ugly-beautiful	.81					
slovenly-polished				.79		
broad-cultured				.88		
husky-not husky	.77					
dull-clear	.68					
soft-loud					.87	
weak-powerful					.65	
high-low			.86			
shrill-deep			.78			
slow-quick						.85
dragging-brisk						.83
speak. +dif.- speak. -dif.	.79					
deviant-not deviant	.85					

3.2.1 Reliability for sustained /a/

The same procedure as used in section 3.1.1 for the reliability of the scales is fol-lowed for the ratings on the sustained /a/. The results are given in table 5.

It appears that a number of scales have a low coefficient. The cause of a low reliability coefficient can be a lack of variation between the speakers (a low MS_b), or low agreement between the raters (a high MS_w) or a combination.

The interpretation of the scales with a lack of variation within the raters *monotonous-melodious, expressionless-expressive, broad-cultured, slovenly-polished, slow-quick, dragging-brisk,* is not hard to give: on forehand, the expectation was that raters would have difficulty in judging a sustained /a/ on parameters as tempo, articulation, and melodiousness. Still, for the comparison with the ratings on the read aloud text, these scales were also inserted in the rating procedure on the sustained /a/. These scales are considered to be unreliable. The same counts for the scale *shrill-deep*, although there is no clear interpretation to give for the low variation between the speakers. The scales with low agreement among the raters are *husky-not husky, not intelligible-intelligible, tense-relaxed, not fluent-fluent,* and *unsteady-steady.* The scale *husky-not husky* still has a high reliability coefficient (Ru=.96) because of a high MS_b.

On the basis of these findings the decision is made to take those scales with Ru > .90 for further analysis.

Table 5. Reliability coefficient Ru, MS_w, and MS_b for the ratings of 24 raters on 22 scales on the voice quality of **sustained /a/** of 14 speakers.

Scale nr.	Scale	Ru	MS_w	MS_b
1	*monotonous-melodious*	.31	1.79	2.59
2	*expressionless-expressive*	.74	0.98	3.70
3	*unpleasant-pleasant*	.95	1.88	35.72
4	*ugly-beautiful*	.96	1.49	41.77
5	*slovenly-polished*	.82	1.63	8.96
6	*broad-cultured*	.60	1.39	3.48
7	*husky-not husky*	.96	2.19	49.49
8	*dull-clear*	.96	1.52	36.54
9	*soft-loud*	.97	1.36	42.20
10	*weak-powerful*	.95	1.94	40.66
11	*high-low*	.95	1.76	32.45
12	*shrill-deep*	.83	1.65	9.71
13	*slow-quick*	.50	1.55	3.07
14	*dragging-brisk*	.20	1.21	3.55
15	*not intelligible-intelligible*	.88	2.71	22.45
16	*creaky-not creaky*	.95	1.65	32.37
17	*panting-not panting*	.92	1.73	22.89
18	*tense-relaxed*	.91	2.21	23.52
19	*not fluent-fluent*	.80	2.34	20.82
20	*speak. +dif. - speak. -dif.*	.95	1.79	32.79
21	*unsteady-steady*	.93	2.37	32.73
22	*deviant-not deviant*	.96	1.97	48.84

As a conclusion the following scales will be used for factor analysis in the next section concerning sustained /a/: *unpleasant-pleasant, ugly-beautiful, husky-not husky, dull-clear, soft-loud, weak-powerful, high-low, creaky-not creaky, panting-not pant-ing, tense-relaxed, speaking with difficulty-speaking without difficulty, unsteady-steady,* and *deviant-not deviant.*

3.2.2 Factor analysis for sustained /a/

The same factor analysis as on read aloud text is carried out on the data for sustained /a/. When the criterium eigenvalue 'greater than 1' was applied, the PCA produced 2 factors. With this number of factors no satisfactory interpretation could be made: too many scales were added together on one factor. Because the scales *monotonous-melodious*, *expressionless-expressive*, and *slovenly-polished, broad-cultured,* and *slow-quick, dragging-brisk* (the factors Melodiousness, Articulation Quality, and Tempo found in the previous section on read aloud text) are not taken in the factor analysis on sustained /a/, the expectation is that these 3 factors will be absent for sustained /a/.

Therefore the factor analysis applied to the sustained /a/ material is forced into 3 factors. On the basis of the significant loadings (> .50) of each scale, the factors are labeled as Abnormality (*unpleasant-pleasant, ugly-beautiful, husky-not husky, dull-clear, panting-not panting, tense-relaxed, speaking with difficulty-speaking without difficulty, unsteady-steady, deviant-not deviant*), Strength (*soft-loud, weak-powerful*), and Pitch/Creakiness (*high-low, creaky-not creaky*). As in the results of read aloud text, the factor Abnormality contains 9 scales. Therefore, the same selection as for the read aloud text is made, based on the reliability coefficients; the scales on the factor Abnormality with Ru < .95 will not be taken for further research: *panting-not panting, tense relaxed*, and *unsteady-steady.*

With the remaining 10 scales another factor analysis is carried out. In order to split the factor Pitch/Creakiness, the factor analysis is forced into 4 factors. The results after varimax rotation, are given in table 6.

Table 6. Percentages of total variance explained (after varimax rotation) of the 4 factors (**sustained /a/**) and significant factor loadings (>.50) of the 10 scales in each of the 4 factors; all loadings are significant at the 1% level.

factor	1	2	3	4
% of total variance explained	42,7	17,7	10,9	10,4
scale				
unpleasant-pleasant	.83			
ugly-beautiful	.88			
husky-not husky	.72			
dull-clear	.73			
soft-loud		.90		
weak-powerful		.73		
high-low			.95	
creaky-not creaky				.92
tense-relaxed	.83			
speak. +dif.- speak. -dif.	.89			
deviant-not deviant	.79			

On the basis of the factor loadings the factors are labeled as Abnormality (*un-pleasant-pleasant, ugly-beautiful, husky-not husky, dull-clear, speaking with difficulty-speaking without difficulty, deviant-not deviant*), Strength (*soft-loud, weak-powerful*), Pitch (*high-low*), and Creakiness (*creaky-not creaky*).

3.3 FACTOR SCORES

Finally, factor scores of all 14 voices are calculated both for the read aloud text (on 6 factors) as well as for the sustained /a/ (on 4 factors). The factor scores give the position of each voice on each factor. The scores can be correlated with the results of acoustic and clinical analyses in further research. The scores are given in table 7 and they are represented in figure 1, factor by factor.

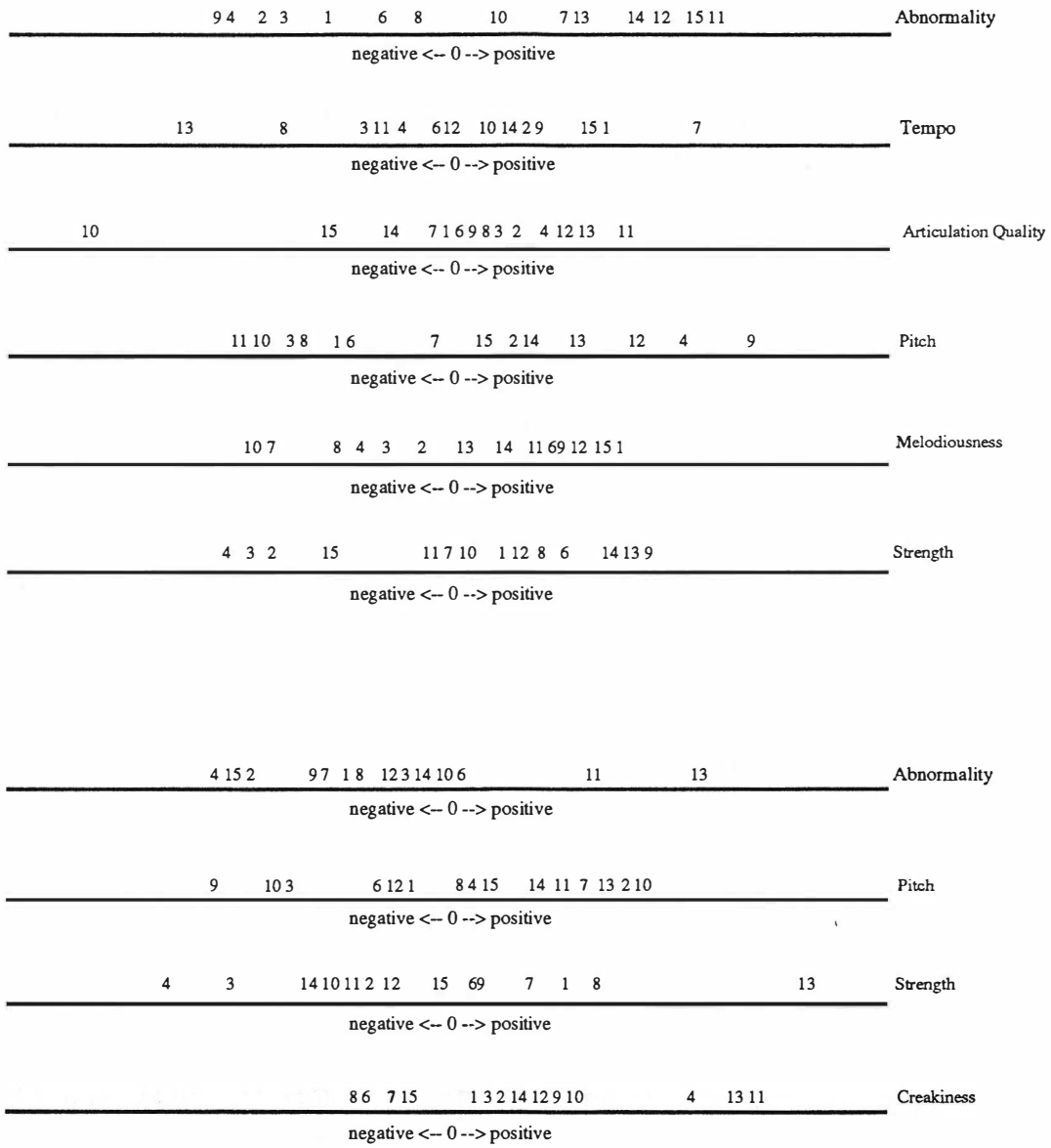

Figure 1. Representations of factor scores of all 14 voices on the 6 factors of **read aloud text** (upper 6) and on the 4 factors of **sustained /a/** (lower 4). Speakers 1-4: patients before radiotherapy; speakers 6-10: patients 6 months after radiotherapy; speaker 11-15: normal speakers.

A negative score on the factor Abnormality means a "husky, dull, unpleasant, ugly, speaking with difficulty and/or deviant" voice; on the Tempo factor it means a "slow and/or dragging" voice and on the Articulation Quality factor it means a "broad and/or slovenly" voice; a negative score on the Pitch factor means a "low and/or deep" voice, on the factor Melodiousness a "expressionless and/or monotonous" voice, and on the factor Strength it means a "weak and/or soft" voice; on the factor Creakiness it means a "creaky" voice. The higher the score on a factor, the more positive a voice is judged by the raters.

Although the scores on the read aloud text and the sustained /a/ have to be considered separately, because they represent two different factor analyses, it appears that there are similarities between the read aloud text and the sustained /a/: for instance, speaker 4 has a very high negative score on the Abnormality and Strength factor for read aloud text as well as for the sustained /a/. In general, the speakers before radiotherapy (speakers 1-4) show much similarity among the two kinds of speech material. Whenever their scores on a factor of the read aloud text are negative, the scores on the sustained /a/ are negative as well. This is not the case for the patients 6 months after radiotherapy (speakers 6-10) and for the control speakers (speakers 11-15). For instance, speaker 15 has a high positive score on the Abnormality factor for read aloud text (his voice is rated as "normal"), but for the sustained /a/ the score is negative.

Apart from the differences among read aloud text and sustained /a/ between the speakers, another difference can be noted: on the read aloud text, speakers 1-4 (patients before radiotherapy) do have a rather high negative score on the factor Abnormality; the scores of the speakers 6-10 (patients after radiotherapy) varies: speaker 9 for instance has a very high negative score in his group (note that speaker 9 and 4 are the same patients); the scores of the control speakers (11-15) are all clearly positive. This tendency seems to hold for the factors Melodiousness, and Strength as well, although there is more variation between the speakers in every group.

Table 7a. Factor scores of the 14 voices on the 6 factors of **read aloud text**; Speakers 1-4: patients before radiotherapy; speakers 6-10: patients 6 months after radiotherapy; speakers 11-15: normal speakers.

factor	Abnor-mality	Tempo	Articul. Quality	Pitch	Melo-diousnes	Strength
speaker 1	-0.61	0.65	-0.04	-0.54	0.64	0.23
speaker 2	-0.91	0.26	0.24	0.27	-0.15	-0.83
speaker 3	-0.84	-0.39	0.13	-0.77	-0.34	-0.94
speaker 4	-1.06	-0.24	0.40	1.04	-0.47	-1.02
speaker 6	-0.39	-0.10	0.02	-0.52	0.43	0.48
speaker 7	0.58	1.09	-0.10	-0.12	-0.86	-0.04
speaker 8	-0.20	-0.81	0.06	-0.75	-0.54	0.35
speaker 9	-1.08	0.32	0.04	1.37	0.43	0.76
speaker 10	0.15	0.16	-1.68	-0.89	-0.92	0.03
speaker 11	1.17	-0.34	0.77	-0.94	0.34	-0.09
speaker 12	0.89	-0.05	0.52	0.82	0.55	0.24
speaker 13	0.59	-1.24	0.53	0.50	0.03	0.73
speaker 14	0.78	0.23	-0.37	0.30	0.24	0.68
speaker 15	1.13	0.58	-0.61	0.12	0.62	-0.61

This tendency cannot be seen for the factors Tempo, Articulation Quality and Pitch: the speakers within each speaker group do vary in their scores.

For the sustained /a/ the speaker groups cannot be differentiated on any factor as clearly as on the factors concerning the read aloud text. For instance, the speakers before radiotherapy (speakers 1-4) do have rather high negative scores on the factor Abnormality; the speakers 6 months after radiotherapy are less "abnormal"; but the normal speakers vary on the factor Abnormality: speaker 13 is judged very "normal" but speaker 15 clearly not.

Table 7b. Factor scores of the 14 voices on the 4 factors of **sustained /a/**; Speakers 1-4: patients before radiotherapy; speakers 6-10: patients 6 months after radiotherapy; speakers 11-15: normal speakers.

factor	Abnor-mality	Pitch	Strength	Creaki-ness
speaker 1	-0.48	-0.19	0.48	0.17
speaker 2	-0.96	0.80	-0.47	0.22
speaker 3	-0.21	-0.77	-1.05	0.19
speaker 4	-1.13	0.08	-1.30	1.12
speaker 6	0.06	-0.37	0.15	-0.43
speaker 7	-0.62	0.55	0.35	-0.29
speaker 8	-0.46	0.06	0.63	-0.46
speaker 9	-0.68	-1.04	0.17	0.42
speaker 10	-0.09	0.85	-0.49	0.48
speaker 11	0.62	0.49	-0.48	1.31
speaker 12	-0.32	-0.25	-0.32	0.33
speaker 13	1.12	0.70	1.55	1.24
speaker 14	-0.14	0.34	-0.61	0.21
speaker 15	-0.99	0.10	-0.04	-0.22

4 Conclusion

The aim of this experiment was to obtain a set of semantic scales that can describe voice quality of patients with early glottic cancer before and after radiotherapy, and of control speakers. The results have shown that for the read aloud text 16 scales can be described in a 6-dimensional perceptual space, representing Abnormality (*unpleasant-pleasant, ugly-beautiful, husky-not husky, dull-clear, speaking with difficulty-speaking without difficulty,* and *deviant-not deviant*), Pitch (*high-low, shrill-deep*), Strength (*soft-loud, weak-powerful*), Articulation Quality (*broad-cultured, slovenly-polished*) and Melodiousness (*monotonous-melodious, expressionless-expressive*).

The results for the sustained /a/ have shown that 10 scales can be described in a 4-dimensional perceptual space respresenting Abnormality (*unpleasant-pleasant, ugly-beautiful, husky-not husky, dull-clear, speaking with difficulty-speaking without difficulty,* and *deviant-not deviant),* Pitch *(high-low),* Strength *(soft-loud, weak-powerful),* and Creakiness *(creaky-not creaky).*

The results in the present experiment agree with the results found in earlier research on read aloud text of normal speakers (Fagel & Van Herpt, 1983; Van Herpt, 1986);

they found 5 factors with 14 (partly different) scales: Voice Appreciation (Melodiousness and Evaluation) contains the scales *monotonous-melodiousness*, *expressionless-expressive*, *unpleasant-pleasant* and *ugly-beautiful*; Articulation Quality contains *slovenly-polished* and *broad-cultured*; Voice Quality (Clarity and Strength) contains *husky-not husky*, *dull-clear*, *soft-loud* and *weak-powerful*; Pitch contains *high-low* and *shrill-deep*, and Tempo *slow-quick* and *dragging-brisk*. The differences with the present results are that the subfactors Melodiousness and Strength have become factors of their own; the subfactors Evaluation and Clarity have come together on one factor (Abnormality), together with the newly added scales *speaking with difficulty-speaking without difficulty,* and *deviant-not deviant.*

In this experiment, with only 14 speakers, the results give a clear insight in how untrained listeners judge voice quality of speakers. It appears that the scale *unintelligible-intelligible* used in previous research (de Leeuw, 1990) to describe pathological voices is not useful. Instead, the scales *speaking with difficulty-speaking without difficulty, deviant-not deviant* give a better description of the abnormality of pathological voices. When the appropiate scales are offered, the listeners will use the same perceptual space for both read aloud text and sustained /a/. With the scales used in this experiment, the same 3 factors are found. For the read aloud text, 3 additional factors are found: Melodiousness, Articulation Quality, and Tempo. For the read aloud text one additional factor is found: Creakiness.

The set of 16 semantic scales for the read aloud text and the set of 10 scales for the sustained /a/ will be the basis for further experiments. The eventually obtained factor scores of every voice sample on the factors will be used in correlational studies between perceptual, acoustic and clinical parameters. The aim is to discriminate between voices of 1) patients before and 2) after radiotherapy and 3) control speakers, and between voices of patients who are treated with different doses of radiation.

References

Asendorpf, J. & Wallbot, H. (1979): "Masse der Beobachteruebereinstimmung - Ein systematischer Vergleich", *Zeitschrift fuer Sozialpsychologie* **10**: 243-252.

Bezooijen, van R. (1987): "Transcription of long-term speech characteristics", *Zeitschrift für Dialektologie und Linguistik* **54**: 111-140.

Blom, J. & Herpt, van L. (1976): "The evaluation of jury judgements on pronunciation quality", *Proceedings of the Institute of Phonetic Sciences, University of Amsterdam* **4**: 31-46.

Boves, L. (1984): *The phonetic basis of perceptual ratings of running speech*, Foris Publications, Dordrecht, Cinnaminson.

Erp, van A.J.M. (1991): *"The phonetic basis of personality ratings, with specific reference to cleft-palate speech,* PTT Research, Leidschendam.

Fagel, W., Herpt van L. & Boves, L. (1983): "Analysis of the perceptual qualities of Dutch speakers' voice and pronunciation", *Speech and Communication* **2**: 315-326.

Hammarberg, B. (1980): *Perceptual and acoustic analysis of dysphonia*, Stockholm.

Herpt, van L. (1986): "Influcence of rater's sex on voice and pronunciation assessment", *Proceedings of the Institute of Phonetic Sciences, University of Amsterdam* **10**: 19-39.

Kim, J. & Mueller, C.W. (1978): *Factor analysis, statistical methods and practical issues*, Sage Publications, Beverly Hills, London.

Laver, J. (1980): *The phonetic description of voice quality*, Cambridge.

Leeuw, de I.M. (1990): "The relation between perceptual and clinical parameters of voice quality of patients with early glottic cancer before and after radiotherapy and of normal speakers", *Proceedings of the Institute of Phonetic Sciences, University of Amsterdam* **14**: 27-38.

Rosenthal, R. (1982): "Conducting judgment studies", In: Scherer, K. & Ekman, P. (eds): *Handbook of methods in nonverbal behavior research*, Cambridge: 287-361.

Wilkinson, L. (1989): *Systat, the system for statistics*, Evaston, IL: Systat Inc.