INFLUENCE OF RATER'S SEX ON VOICE AND PRONUNCIATION ASSESSMENT

Leo W.A. van Herpt

1.0 INTRODUCTION

This study is part of a project which aims at the development of a reliable and efficient instrument for the perceptual description of voice and pronunciation (V&P) quality. Our approach of this task is based on a procedure described by Osgood and Suci (1955) and involves a multivariate differentiation of the concept V&P in terms of a limited number of semantic scales of known factor composition.
Fundamental problems in this procedure are (1) the selection of a (small) sample of qualifiers of V&P that represents the major dimensions along which the perceptual judgments vary, and (2) the separation of variance attributable to the qualifiers (scales) from subject (listener) and object (speaker) variance.
The present study is directed at the variance problem, especially as to the effects of sex of speakers and listeners.

In an earlier part of the investigation (Blom & van Herpt, 1976; Blom & Koopmans-van Beinum, 1973) a set of bipolar adjectival scales which are applicable to voice characteristics are selected. Factorial studies (Fagel & van Herpt, 1982; Fagel, van Herpt & Boves, 1983) have shown, after extensive testing, that the resulting qualifiers have a reasonably stable structure. The perceptual space appears to be spanned by at least five orthogonal dimensions: I:Voice Appreciation, II:Articulation Quality, III:Voice Quality, IV:Pitch and V:Tempo. There is a possibility that dimension I and III can further be broken down in dimensions which we tentatively named: Ia: Melodiousness, Ib:Evaluation, IIIa:Clarity and IIIb:Subjective Strength.
A methodologically logical next step was to verify the dimensional structure using a larger sample of voices (van Herpt, Fagel & Boves, in prep.). So in the next study the number of speakers was increased from 10 to 72 and a comprised rating form of 14 scales was used.
To cover the domain of possible discriminations in the V&P space we selected fourteen scales (see table 1); two semantic twin scales for each dimension and an extra pair of scales for each of the two dimensions that show a tendency to split up. The scale pairs have been selected as twins on account of their similarity in meaning, in this case because of their closeness in semantic space. E.g. the scales 11:'dragging-brisk' and 12:'slow-quick' are selected as twin scales of the Tempo dimension because of their 'factorial purity', that is to say, because of their high loadings on the Tempo dimension and their low loadings on the other dimensions in combination with a high communality in several factor analyses.
This smaller number of scales in the shortened version enabled us to use a summation method of factor analysis which takes mean scores over judges

instead of the scores of the individual judges as data. The method thus in principle eliminates **subject** variance, assuming it is negligible, consequently the solution is determined by speaker variance only.
In earlier scale-selection experiments the stringing-out method of factor analysis had to be used because the number of variables (scales) was greater than the number of observations (speakers). A drawback of stringing out the data is that listener and speaker variance are inextricable entangled.

Table 1. Scales and dimensions of shortened rating form

Sc.nr	Scale terms[1]		IS[2]	Dimension
01.	eentonig	- melodieus	6.16	
	(monotonous	- melodious)		Ia. Voice Appreciation: Melodiousness
02.	uitdrukkingsloos	- expressief	6.32	
	(expressionless	- expressive)		
13.	lelijk	- mooi	6.26	
	(ugly	- beautiful)		Ib. Voice Appreciation: Evaluation
14.	aangenaam	- onaangenaam)	6.73	
	(unpleasant	- pleasant)		
03	plat	- beschaafd	6.09	
	(broad	- cultured)		II. Articulation Quality
04.	onverzorgd	- verzorgd	5.95	
	(slovenly	- polished)		
05.	dof	- helder	5.92	
	(dull	- clear)		IIIa. Voice Quality: Clarity
06.	hees	- niet hees	5.63	
	(husky	- not husky)		
07.	zwak	- krachtig	5.42	
	(weak	- powerful)		IIIb. Voice Quality: Subjective Strength
08.	zacht	- luid	4.04	
	(soft	- loud)		
09.	schel	- diep	5.04	
	(shrill	- deep)		IV. Pitch
10.	hoog	- laag	4.18	
	(high	- low)		
11.	traag	- vlot	5.63	
	(dragging	- brisk)		V. Tempo
12.	langzaam	- snel	4.69	
	(slow	- quick)		

1) To facilitate readability and statistical treatment all scales are repolarized after the test with the scale term that according to its IS value, is the more desirable one, to the right.
2) Scale values of Ideal Voice & Pronunciation.

Our solutions of the stringing out and of the summation method strongly resemble each other which suggested that subject variance does not have a systematic effect on the correlations between the scales.
However in perception experiments on age and sex (van Herpt & Hoebe, 1985; Boves, Fagel & van Herpt, 1982; van Herpt en Fagel, 1981) indications of subject x object or subject x scale interactions were found. So we have devised a complementary way to consider the validity of the rating instrument. The method, after an idea used by Osgood and Suci (1955:332), involves a rating of the qualifier terms themselves. The subjects are simply asked for their opinion concerning the relations between the scales by having them judge each of the scales against the thirteen remaining attribute scales without presenting any speech.
This procedure of course lacks any **speaker** variance, so the results concern the rating instrument itself (e.g. the twin scales) and the groups of judges. This information must enable us to adjust the rating procedure in such a way that the listener variance is indeed small. Not until then the resulting qualifying structure can be attributed to an underlying organization of scale terms as applied to speakers. This being the case, we also can expect the correlations between perceptual ratings and external acoustic criteria to improve. Hitherto these correlations generally are low, on the perceptual side probably due to listener effects.

2.0 METHOD

2.1 Procedure

Subjects are asked their opinion concerning the correspondence in meaning of different adjectives in the description of the average female respectively male voice.
The method involves a rating of qualifiers on bipolar scales, without realizations of V&P. The qualifiers to be judged are the scale terms (see table 1) of the comprised rating form proposed by Fagel et al.(1983). As said in above-mentioned article "the scale terms in Table 1 and further in this paper are tentative translations of the original Dutch scale terms. We must offer a warning about inevitable differences in connotation which are very important for the measurement result which is to be expected when these English adjectives were to be used." (1983:317)

Each of the fourteen scales has been paired with every other scale, thus generating 91 items (type a). After reversing the polarity of the fourteen stimulus scales each is coupled again with the other scales still in their original orientation, which generates another 91 items (type b). From this collection two test versions are formed. Test A consists of all odd type-a items and all even type-b items; test B of the remaining items. To shuffle the stimulus terms a rotational procedure is used. This left us with only a few successive identical rating scales. These items are moved to the end of the test.

The items are presented in the following form:
1. MELODIOUS - monotonous   broad I--I--I--I--I--I--I--I cultured
2. DULL - clear   shrill I--I--I--I--I--I--I--I deep
3. CULTURED - broad   dragging I--I--I--I--I--I--I--I brisk

The subjects are requested to rate the capitalized word of the first pair on the descriptive continuum between the second pair.
When the capitalized stimulus word is very similar in meaning to one of the scale terms in the left, the stimulus is scored as follows:

GOOD - bad   clean I✓I--I--I--I--I--I--I dirty
or
DISHONEST - honest   good I--I--I--I--I--I--I✓I bad

The smaller the similarity in meaning the closer to the centre of the scale the scoring tick is placed. When neither of both scale terms is applicable or both terms to the same extent the tick comes in the centre of the scale:
I--I--I--I✓I--I--I--I

Both terms of the pair on the left side are given in order to define more accurately the meaning of the stimulus word. Subjects are informed that the terms are meant to be each others opposite in meaning.

Before execution of the main experiment a small investigation was performed to establish if the judgment of one term of the contrastive scale terms was sufficient to determine the rating of the other one too. In this exploration test A and test B both have been answered by twenty female subjects. The correlation between both tests is 0.868. (It is noteworthy that the coefficient is significantly lowered by 7 items only, all of which concern the dimensions Strength and Pitch.) This result justifies the assumption that the left-right polarization of the scales was of little consequence in the judgments, so - for efficiency reasons - we arbitrarely picked test B to use in the present study.
In table 2 the 91 items are given in systematic order and polarized with the scale term judged as more desirable to the right. In the text items are referred to by item number and scale combination, e.g. 52:0511 refers to item 52 in table 2 which consists of scale 05 (dull-clear) and scale 11 - (dragging-brisk).

2.2 Subjects

Raters in the experiment are drawn from the population of male students of Dutch language from the University of Amsterdam and of (mainly) female students from the Training Course of Speech Therapists in Amsterdam. From an earlier investigation (Boves et al.,1982:7) it is known that in the present type of studies samples of female students from these two courses can be considered to be drawn from one population. We expect this to be the case for male subjects too, but were not able to verify it because the speech therapist group consists almost exclusively of female students.
A total of 60 subjects was used, about thirty of each course. Twenty-nine subjects are female (F), twenty-six male (M) and five did not indicate their course, sex and age.
The students are 18 - 41 years of age; mean age of women being 22.3 years, mean age of men 24.3 years. All subjects are native speakers of Dutch. The experiment was carried out in the first year of training so that they may be considered rather naive with respect to speech science.

Table 2 - Test-items in systematic order

Itemcode	Stimulus pair	Nr	Ratingscale
01:0102	monotonous - melodious	/ 02	expressionless-expressive
02:0103		/ 03	broad - cultured
03:0104		/ 04	slovenly - polished
04:0105		/ 05	dull - clear
05:0106		/ 06	husky - not husky
06:0107		/ 07	weak - powerful
07:0108		/ 08	soft - loud
08:0109		/ 09	shrill - deep
09:0110		/ 10	high - low
10:0111		/ 11	dragging - brisk
11:0112		/ 12	slow - quick
12:0113		/ 13	ugly - beautiful
13:0114		/ 14	unpleasant - pleasant
14:0203	expressionless-expressive	/ 03	broad - cultured
15:0204		/ 04	slovenly - polished
16:0205		/ 05	dull - clear
17:0206		/ 06	husky - not husky
18:0207		/ 07	weak - powerful
19:0208		/ 08	soft - loud
20:0209		/ 09	shrill - deep
21:0210		/ 10	high - low
22:0211		/ 11	dragging - brisk
23:0212		/ 12	slow - quick
24:0213		/ 13	ugly - beautiful
25:0214		/ 14	unpleasant - pleasant
26:0304	broad - cultured	/ 04	slovenly - polished
27:0305		/ 05	dull - clear
28:0306		/ 06	husky - not husky
29:0307		/ 07	weak - powerful
30:0308		/ 08	soft - loud
31:0309		/ 09	shrill - deep
32:0310		/ 10	high - low
33:0311		/ 11	dragging - brisk
34:0312		/ 12	slow - quick
35:0313		/ 13	ugly - beautiful
36:0314		/ 14	unpleasant - pleasant
37:0405	slovenly - polished	/ 05	dull - clear
38:0406		/ 06	husky - not husky
39:0407		/ 07	weak - powerful
40:0408		/ 08	soft - loud
41:0409		/ 09	shrill - deep
42:0410		/ 10	high - low
43:0411		/ 11	dragging - brisk
44:0412		/ 12	slow - quick
45:0413		/ 13	ugly - beautiful
46:0414		/ 14	unpleasant - pleasant

Table 2 - (continued)

Itemcode	Stimulus pair	Nr Ratingscale
47:0506	dull - clear	/ 06 husky - not husky
48:0507		/ 07 weak - powerful
49:0508		/ 08 soft - loud
50:0509		/ 09 shrill - deep
51:0510		/ 10 high - low
52:0511		/ 11 dragging - brisk
53:0512		/ 12 slow - quick
54:0513		/ 13 ugly - beautiful
55:0514		/ 14 unpleasant - pleasant
56:0607	husky - not husky	/ 07 weak - powerful
57:0608		/ 08 soft - loud
58:0609		/ 09 shrill - deep
59:0610		/ 10 high - low
60:0611		/ 11 dragging - brisk
61:0612		/ 12 slow - quick
62:0613		/ 13 ugly - beautiful
63:0614		/ 14 unpleasant - pleasant
64:0708	weak - powerful	/ 08 soft - loud
65:0709		/ 09 shrill - deep
66:0710		/ 10 high - low
67:0711		/ 11 dragging - brisk
68:0712		/ 12 slow - quick
69:0713		/ 13 ugly - beautiful
70:0714		/ 14 unpleasant - pleasant
71:0809	soft - loud	/ 09 shrill - deep
72:0810		/ 10 high - low
73:0811		/ 11 dragging - brisk
74:0812		/ 12 slow - quick
75:0813		/ 13 ugly - beautiful
76:0814		/ 14 unpleasant - pleasant
77:0910	shrill - deep	/ 10 high - low
78:0911		/ 11 dragging - brisk
79:0912		/ 12 slow - quick
80:0913		/ 13 ugly - beautiful
81:0914		/ 14 unpleasant - pleasant
82:1011	high - low	/ 11 dragging - brisk
83:1012		/ 12 slow - quick
84:1013		/ 13 ugly - beautiful
85:1014		/ 14 unpleasant - pleasant
86:1112	dragging - brisk	/ 12 slow - quick
87:1113		/ 13 ugly - beautiful
88:1114		/ 14 unpleasant - pleasant
89:1213	slow - quick	/ 13 ugly - beautiful
90:1214		/ 14 unpleasant - pleasant
91:1314	ugly - beautiful	/ 14 unpleasant - pleasant

Of each group fifty percent of the raters is asked to give their ratings bearing in mind the average female voice (♀), the others with the average male voice (♂) in mind. The resulting distribution is given in table 3.

Table 3. Distribution of female (♀) and male (♂) 'voices' over female (M) and male (F) raters.

			'voices' ♂	♀
		n = 60	29	31
Raters	M	26	12	14
	F	29	15	14
	?	5	2	3

2.3 Treatment of data

Subjects gave their opinion concerning the relations between terms on bipolar seven-point scales. The degree to which terms are judged as identical, operationalizes the degree of congruence between the meaning of those qualifiers. The more their ratings on all other scales are identical the more the terms are similar.
To make the scores comparable all ratings are scored as follows.
The scale term closest to the Ideal V&P value is defined as the positive pole. Mean Ideal V&P values, calculated from data from Boves et al.(1982) are given in table 1. All scales are recoded in such a way that they are scored with the positive pole to the right. The value 1 is accorded to the scale position situated on the left extreme and the value 7 to the one on the right extreme.
Next, since the scale midpoint is considered to be the neutral point of relation, the central value 4 is subtracted from all scores. This linear transformation is allowed because the scales are known to be interval scales (Boves, 1984:170; Blom & van Herpt, 1976:40). So a relation value of -3 indicates the maximum degree of correspondence between two negative qualifiers, whereas +3 is the highest possible correlation between a positive and a negative adjective.

In order to be able to determine whether the observed relations between scales are dependent on sex of rater and/or on sex of speaker the data collection is arranged as to sex of 'voice-to-be-judged' and as to sex of rater separately. Further both collections are divided in two subgroups. So the following samples can be compared:

- Sample A1. Male versus female 'voice' for all scores
- Sample A2. Male versus female 'voice' for male raters only
- Sample A3. Male versus female 'voice' for female raters only
- Sample B1. Male versus female raters for all scores
- Sample B2. Male versus female raters for male 'voice' only
- Sample B3. Male versus female raters for female 'voice' only.

Table 4A - T-tests of male (♂) versus female (♀) 'voice' for all (MF), male (M) and female (F) raters

Item-code	Mean MF/♂ n=29	Mean MF/♀ n=31	T-value Sample A1	Sign. P	Mean M/♂ n=12	Mean M/♀ n=14	T-value Sample A2	Sign. P	Mean F/♂ n=15	Mean F/♀ n=14	T-value Sample A3	Sign. P	Item-code
01:0102	-2.310	-2.354	0.179		-1.916	-2.000	0.177		-2.600	-2.785	0.918		01:0102
02:0103	-1.655	-1.322	-1.332	0.25	-1.416	-0.785	-1.630		-1.933	-1.785	-0.449		02:0103
03:0104	-0.689	-0.516	-0.470		-0.416	-0.571	0.257		-0.733	-0.428	-0.579		03:0104
04:0105	-1.310	-1.354	0.155		-1.166	-1.571	0.877		-1.333	-1.357	0.059		04:0105
05:0106	-1.206	-1.000	-0.593		-1.083	-1.071	-0.020		-1.133	-1.142	0.019		05:0106
06:0107	-1.241	-1.032	-0.742		-1.083	-1.142	0.125		-1.266	-1.071	-0.483		06:0107
07:0108	-0.172	-0.193	0.077		0.166	-0.071	0.621		-0.400	-0.357	-0.116		07:0108
08:0109	0.413	0.419	-0.021		0.250	0.642	-0.741		0.400	0.214	-0.582		08:0109
09:0110	0.275	0.258	0.067		0.250	0.428	-0.361		0.066	0.142	-0.326		09:0110
10:0111	-1.137	-1.096	-0.144		-0.666	-1.000	0.806		-1.466	-1.357	-0.283		10:0111
11:0112	-0.793	-0.387	-1.508	0.25	-0.916	-0.571	-0.670		-0.666	-0.285	-1.325		11:0112
12:0113	-2.344	-2.419	0.315		-2.083	-2.357	0.644		-2.533	-2.571	0.147		12:0113
13:0114	-2.413	-2.548	0.575		-2.000	-2.428	0.922		-2.733	-2.714	-0.095		13:0114
14:0203	-1.551	-1.548	-0.010		-1.000	-1.214	0.455		-2.133	-2.000	-0.419		14:0203
15:0204	-0.069	-0.354	1.052	0.25	0.000	-0.142	0.336		-0.133	-0.642	1.205		15:0204
16:0205	-1.793	-1.580	-0.682		-1.500	-1.928	0.975		-1.933	-1.428	-1.038		16:0205
17:0206	-0.275	-0.419	0.493		0.083	-0.357	0.745		-0.533	-0.500	-0.111		17:0206
18:0207	-1.826	-1.612	-0.987	0.25	-1.833	-1.642	-0.428		-1.933	-1.785	-0.487		18:0207
19:0208	-0.793	-0.806	0.046		-0.583	-0.714	0.299		-0.733	-1.000	0.651		19:0208
20:0209	-0.206	-0.322	0.481		-0.083	-0.500	1.211		-0.533	-0.285	-0.730		20:0209
21:0210	0.620	0.709	-0.377		0.916	0.857	0.142		0.266	0.642	-1.310		21:0210
22:0211	-1.793	-1.290	-1.800	0.10	-1.666	-1.428	-0.580		-1.800	-1.071	-1.716	0.10	22:0211
23:0212	-0.724	-1.161	1.341	0.25	-0.250	-1.142	1.716	0.10	-0.800	-1.357	1.306		23:0112
24:0213	-2.172	-2.483	1.284	0.25	-1.416	-2.285	2.054	0.10	-2.733	-2.714	-0.095		24:0213
25:0214	-2.275	-2.193	-0.303		-1.750	-2.071	0.595		-2.600	-2.357	-1.034		25:0214
26:0304	-1.793	-2.032	0.742		-1.000	-1.714	1.299		-2.333	-2.428	0.267		26:0304
27:0305	-1.034	-0.871	-0.555		-0.500	-0.571	0.155		-1.333	-1.142	-0.465		27:0305
28:0306	-1.482	-1.161	-0.957	0.25	-1.000	-1.000	0.000		-1.933	-1.357	-1.161		28:0306
29:0307	-0.396	-0.645	0.765		-0.333	-0.500	0.273		-1.266	-0.928	-0.930		29:0307
30:0308	-0.379	-0.096	-1.091	0.25	-0.333	0.214	-1.120		-0.533	-0.428	-0.355		30:0308
31:0309	-0.655	-0.741	0.296		-0.083	-0.928	1.809	0.10	-0.933	-0.714	-0.581		31:0309
32:0310	-0.206	-0.096	-0.566		-0.166	0.142	-0.921		-0.333	-0.357	0.095		32:0310
33:0311	-0.310	-0.193	-0.377		0.166	-0.500	1.441		-0.400	0.071	-1.090		33:0311
34:0312	0.034	0.129	-0.370		0.500	-0.071	1.397		-0.066	0.357	-1.390		34:0312
35:0313	-2.206	-2.290	0.349		-1.833	-2.071	0.504		-2.533	-2.642	0.579		35:0313
36:0314	-2.172	-2.387	0.736		-1.833	-1.857	0.042		-2.600	-2.857	1.330		36:0314
37:0405	-1.000	-1.129	0.425		-0.666	-1.142	1.011		-1.333	-1.142	-0.439		37:0405
38:0406	-0.137	-0.096	-0.160		0.333	-0.142	1.074		-0.466	-0.071	-1.161		38:0406
39:0407	-0.517	-0.322	-0.775		-0.250	-0.285	1.074		-0.733	-0.428	-0.995		39:0407
40:0408	1.000	1.161	-0.504		0.916	1.428	-0.881		1.066	1.071	-0.012		40:0408
41:0409	-0.448	-0.548	0.402		0.083	-0.571	1.651		-0.933	-0.642	-0.907		41:0409
42:0410	0.344	0.129	0.923		0.333	0.500	-0.488		0.266	-0.214	1.336		42:0410
43:0411	-0.379	-0.258	-0.551		-0.333	-0.142	-0.503		-0.333	-0.428	0.307		43:0411
44:0412	0.551	0.419	0.531		0.583	0.142	1.107		0.600	0.785	-0.529		44:0412
45:0413	-1.724	-1.774	0.156		-1.083	-1.428	0.660		-2.333	-2.142	-0.496		45:0413
46:0414	-1.448	-1.548	0.275		-1.083	-1.285	0.331		-1.800	-1.857	0.120		46:0414
47:0506	-2.241	-2.419	0.638		-1.666	-1.928	0.506		-2.600	-2.928	1.414		47:0506
48:0507	-1.827	-1.871	0.145		-1.166	-1.428	0.476		-2.266	-2.428	0.598		48:0507
49:0508	-1.379	-0.774	-1.894	0.10	-1.083	-0.571	-0.899		-1.400	-1.142	-0.666		49:0508
50:0509	1.137	1.161	-0.086		1.000	1.285	-0.601		1.066	1.000	0.1[illegible]5		50:0509
51:0510	0.448	0.451	-0.009		0.916	1.071	-0.285		-0.200	-0.142	-0.168		51:0510
52:0511	-1.103	-0.806	-1.003	0.25	-0.833	-1.142	0.571		-1.133	-0.642	-1.434		52:0511
53:0512	-0.793	-0.516	-1.059	0.25	-0.750	-0.857	0.314		-0.666	-0.285	-0.886		53:0512
54:0513	-2.103	-1.935	-0.664		-1.750	-1.785	0.082		-2.400	-2.142	-0.821		54:0513
55:0514	-2.275	-2.225	-0.217		-1.833	-2.000	0.411		-2.600	-2.571	-0.122		55:0514
56:0607	-0.655	-1.129	1.144	0.25	-0.250	-0.642	0.613		-1.333	-1.928	1.251		56:0607
57:0608	-1.689	-1.516	-0.495		-1.166	-0.785	-0.559		-1.933	-2.285	1.485		57:0608
58:0609	-0.172	-0.096	-0.269		-0.083	0.357	-0.865		-0.266	-0.500	0.724		58:0609
59:0610	0.482	0.806	-0.838		0.916	1.214	-0.474		0.000	0.428	-0.804		59:0610
60:0611	-0.620	-0.354	-1.168	0.25	-0.500	-0.428	-0.198		-0.800	-0.285	-1.845	0.10	60:0611
61:0612	-0.586	-0.419	-0.649		-0.416	-0.714	0.589		-0.666	-0.142	-2.029	0.10	61:0612
62:0613	-1.482	-1.548	0.213		-1.166	-1.285	0.278		-1.866	-1.642	-0.467		62:0613
63:0614	-0.931	-1.064	0.339		-0.500	-0.785	0.384		-1.266	-1.357	0.201		63:0614
64:0708	-2.206	-1.935	-0.887		-1.916	-1.500	-0.678		-2.333	-2.357	0.088		64:0708
65:0709	0.310	0.322	-0.042		0.666	0.500	0.335		0.000	0.071	-0.207		65:0709
66:0710	-0.689	-0.677	-0.035		-0.250	-0.071	-0.295		-1.133	-1.428	0.897		66:0710
67:0711	-1.517	-1.193	-1.064	0.25	-1.083	-1.214	0.243		-1.666	-1.285	-1.004		67:0711
68:0712	-0.620	-0.322	-0.970	0.25	-0.916	-0.714	-0.380		-0.333	0.000	-0.832		68:0712
69:0713	-1.241	-1.064	-0.592		-1.083	-1.000	-0.180		-1.533	-1.285	-0.552		69:0713
70:0714	-1.793	-1.258	-1.866	0.10	-1.500	-0.857	-0.400		-2.133	-1.714	-1.137		70:0714
71:0809	0.689	0.612	0.245		0.916	1.071	-0.342		0.266	0.214	0.116		71:0809
72:0810	0.827	0.580	0.803		1.500	0.785	1.591		0.133	0.357	-0.530		72:0810
73:0811	-1.034	-0.322	-2.740	0.01	-0.916	-0.785	-0.302		-1.000	0.000	-2.983	0.01	73:0811
74:0812	-0.931	-0.677	-0.987	0.25	-0.916	-0.785	-0.275		-1.000	-0.714	-0.876		74:0812
75:0813	0.069	0.064	0.016		0.333	0.500	-0.423		-0.133	-0.285	0.299		75:0813
76:0814	0.103	0.064	0.097		0.750	0.642	0.181		-0.533	-0.428	-0.185		76:0814
77:0910	-2.482	-2.193	-1.128	0.25	-2.333	-2.071	-0.532		-2.533	-2.357	-0.685		77:0910
78:0911	0.689	0.774	-0.339		0.750	0.928	-0.461		0.733	0.785	-0.158		78:0911
79:0912	0.758	0.838	-0.303		0.916	1.071	-0.366		0.533	0.714	-0.473		79:0912
80:0913	-2.241	-2.193	-0.176		-1.916	-2.071	0.298		-2.533	-2.500	-0.127		80:0913
81:0914	-1.206	-1.225	0.056		-0.750	-0.928	0.259		-1.600	-1.714	0.419		81:0914
82:1011	1.241	0.871	1.367	0.25	1.416	0.928	1.045		1.000	0.928	0.194		82:1011
83:1012	1.103	1.129	-0.092		1.500	1.071	1.049		1.000	1.357	-0.924		83:1012
84:1013	-0.862	-0.548	-0.920		-0.666	-0.071	-1.052		-1.333	-1.214	-0.328		84:1013
85:1014	-1.206	-0.612	-1.654	0.10	-0.833	-0.285	-0.980		-1.800	-0.928	-2.052	0.05	85:1014
86:1112	-2.206	-1.645	-2.413	0.05	-2.416	-2.000	-1.192		-1.933	-1.500	-1.412		86:1112
87:1113	-1.310	-1.354	0.155		-1.166	-1.571	0.877		-1.933	-1.357	0.059		87:1113
88:1114	-1.379	-1.451	0.220		-0.916	-1.357	0.872		-1.666	-1.642	-0.050		88:1114
89:1213	0.379	0.387	-0.024		0.416	0.071	0.787		0.600	0.642	-0.079		89:1213
90:1214	-1.000	-0.483	-1.382	0.25	-0.250	-0.285	0.061		-1.733	-0.642	-2.113	0.05	90:1214
91:1314	-2.758	-2.677	-0.513		-2.583	-2.500	-0.255		-2.866	-2.857	-0.068		91:1314

Table 4B - T-tests of male (M) versus female (F) raters for male (6) and female (9) 'voice' combined and separately.

Item-code	Mean M/♂♀ n=26	Mean F/♂♀ n=29	T-value Sample B1	Sign. P	Mean M/♂ n=12	Mean F/♂ n=15	T-value Sample B2	Sign. P	Mean M/♀ n=14	Mean F/♀ n=14	T-value Sample B3	Sign. P	Item-code
01:0102	-1.961	-2.689	2.988	0.01	-1.916	-2.600	1.784	0.10	-2.000	-2.785	2.473	0.05	01:0102
02:0103	-1.076	-1.862	3.082	0.01	-1.416	-1.933	1.313		-0.785	-1.785	3.121	0.01	02:0103
03:0104	-0.500	-0.586	0.220		-0.416	-0.733	0.605		-0.571	-0.428	-0.240		03:0104
04:0105	-1.384	-1.344	-0.132		-1.166	-1.333	0.329		-1.571	-1.357	-0.607		04:0105
05:0106	-1.076	-1.137	0.165		-1.083	-1.133	0.087		-1.071	-1.142	0.143		05:0106
06:0107	-1.115	-1.172	0.188		-1.083	-1.266	0.374		-1.142	-1.071	-0.186		06:0107
07:0108	0.038	-0.379	1.594		0.166	-0.400	1.586		-0.071	-0.357	0.728		07:0108
08:0109	0.461	0.310	0.506		0.250	0.400	-0.307		0.642	0.214	1.190		08:0109
09:0110	0.346	0.103	0.940		0.250	0.066	0.440		0.428	0.142	0.882		09:0110
10:0111	-0.846	-1.413	2.036	0.05	-0.666	-1.466	2.011	0.05	-1.000	-1.357	0.892		10:0111
11:0112	-0.730	-0.482	-0.847		-0.916	-0.666	-0.569		-0.571	-0.285	-0.771		11:0112
12:0113	-2.230	-2.551	1.340		-2.083	-2.533	1.170		-2.357	-2.571	0.708		12:0113
13:0114	-2.230	-2.724	2.046	0.05	-2.000	-2.733	1.977	0.10	-2.428	-2.714	0.903		13:0114
14:0203	-1.115	-2.069	3.484	0.01	-1.000	-2.133	2.778	0.05	-1.214	-2.000	2.067	0.05	14:0203
15:0204	-0.076	-0.379	1.016		0.000	-0.133	0.272		-0.142	-0.642	1.409		15:0204
16:0205	-1.730	-1.689	-0.124		-1.500	-1.933	0.960		-1.928	-1.428	-1.037		16:0205
17:0206	-0.153	-0.517	1.152		0.083	-0.533	1.076		-0.357	-0.500	0.472		17:0206
18:0207	-1.730	-1.862	0.501		-1.833	-1.933	0.267		-1.642	-1.785	0.382		18:0207
19:0208	-0.653	-0.862	0.708		-0.583	-0.733	0.333		-0.714	-1.000	0.719		19:0208
20:0209	-0.307	-0.413	0.438		-0.083	-0.533	1.086		-0.500	-0.285	-0.826		20:0209
21:0210	0.884	0.448	1.774	0.10	0.916	0.266	1.885	0.10	0.857	0.642	0.606		21:0210
22:0211	-1.538	-1.448	-0.300		-1.666	-1.800	0.333		-1.428	-1.071	-0.822		22:0211
23:0212	-0.730	-1.069	0.992		-0.250	-0.800	1.040		-1.142	-1.357	0.522		23:0212
24:0213	-1.884	-2.724	3.563	0.01	-1.416	-2.733	4.027	0.01	-2.285	-2.714	1.376		24:0213
25:0214	-1.923	-2.482	1.995	0.05	-1.750	-2.600	1.784	0.10	-2.071	-2.357	0.906		25:0214
26:0304	-1.384	-2.379	3.099	0.01	-1.000	-2.333	2.745	0.05	-1.714	-2.428	1.696		26:0304
27:0305	-0.538	-1.241	2.338	0.05	-0.500	-1.333	1.953	0.10	-0.571	-1.142	1.300		27:0305
28:0306	-1.000	-1.655	1.898	0.10	-1.000	-1.933	1.983	0.10	-1.000	-1.357	0.701		28:0306
29:0307	-0.423	-1.103	1.989	0.05	-0.333	-1.266	1.535		-0.500	-0.928	1.220		29:0307
30:0308	-0.038	-0.482	1.597		-0.333	-0.533	0.468		0.214	-0.428	1.777	0.10	30:0308
31:0309	-0.538	-0.827	0.955		-0.083	-0.933	1.945	0.10	-0.928	-0.714	-0.529		31:0309
32:0310	0.000	-0.344	1.679		-0.166	-0.333	0.495		0.142	-0.357	2.037	0.10	32:0310
33:0311	-0.192	-0.172	-0.063		0.166	-0.400	1.086		-0.500	0.071	-1.560		33:0311
34:0312	0.192	0.137	0.215		0.500	-0.066	1.383		-0.071	0.357	-1.439		34:0312
35:0313	-1.961	-2.586	2.601	0.05	-1.833	-2.533	1.954		-2.071	-2.642	1.717	0.10	35:0313
36:0314	-1.846	-2.724	3.078	0.01	-1.833	-2.600	2.385	0.05	-1.857	-2.857	2.097	0.05	36:0314
37:0405	-0.923	-1.241	1.003		-0.666	-1.333	1.522		-1.142	-1.142	0.000		37:0405
38:0406	0.076	-0.275	1.269		0.333	-0.466	1.872	0.10	-0.142	-0.071	-0.201		38:0406
39:0407	-0.269	-0.586	1.162		-0.250	-0.733	1.117		-0.285	-0.428	0.409		39:0407
40:0408	1.192	1.069	0.354		0.916	1.066	-0.259		1.428	1.071	0.875		40:0408
41:0409	-0.269	-0.793	2.042	0.05	0.083	-0.933	2.581	0.05	-0.571	-0.642	0.222		41:0409
42:0410	0.423	0.034	1.559		0.333	0.266	0.235		0.500	-0.214	1.763	0.10	42:0410
43:0411	-0.230	-0.379	0.624		-0.333	-0.333	0.000		-0.142	-0.428	0.803		43:0411
44:0412	0.346	0.689	-1.307		0.583	0.600	-0.054		0.142	0.785	-1.532		44:0412
45:0413	-1.269	-2.241	3.079	0.01	-1.083	-2.333	2.758	0.05	-1.428	-2.142	1.581		45:0413
46:0414	-1.192	-1.827	1.694	0.10	-1.083	-1.800	1.245		-1.285	-1.857	1.131		46:0414
47:0506	-1.807	-2.758	3.512	0.01	-1.666	-2.600	2.193	0.05	-1.928	-2.928	2.888	0.01	47:0506
48:0507	-1.307	-2.344	3.548	0.01	-1.166	-2.266	2.127	0.05	-1.428	-2.428	3.288	0.01	48:0507
49:0508	-0.807	-1.275	1.395		-1.083	-1.400	0.600		-0.571	-1.142	1.339		49:0508
50:0509	1.153	1.034	0.413		1.000	1.066	-0.140		1.285	1.000	0.805		50:0509
51:0510	1.000	-0.172	3.774	0.01	0.916	-0.200	2.115	0.05	1.071	-0.142	3.441	0.01	51:0510
52:0511	-1.000	-0.896	-0.334		-0.833	-1.133	0.575		-1.142	-0.642	-1.409		52:0511
53:0512	-0.807	-0.482	-1.179		-0.750	-0.666	-0.199		-0.857	-0.285	-1.557		53:0512
54:0513	-1.769	-2.275	1.962	0.05	-1.750	-2.400	1.796	0.10	-1.785	-2.142	0.944		54:0513
55:0514	-1.923	-2.586	2.945	0.01	-1.833	-2.600	2.054	0.05	-2.000	-2.571	2.103	0.05	55:0514
56:0607	-0.161	-1.620	3.729	0.01	-0.250	-1.333	1.478		-0.642	-1.928	4.077	0.01	56:0607
57:0608	-0.961	-2.103	3.327	0.01	-1.166	-1.933	1.341		-0.785	-2.285	3.746	0.01	57:0608
58:0609	0.153	-0.379	1.838	0.10	-0.083	-0.266	0.384		0.357	-0.500	2.428	0.05	58:0609
59:0610	1.076	0.206	2.154	0.05	0.916	0.000	1.491		1.214	0.428	1.447		59:0610
60:0611	-0.461	-0.551	0.396		-0.500	-0.800	0.801		-0.428	-0.285	-0.551		60:0611
61:0612	-0.576	-0.413	-0.591		-0.416	-0.666	0.552		-0.714	-0.142	-1.792	0.10	61:0612
62:0613	-1.230	-1.758	1.656	0.10	-1.166	-1.866	1.613		-1.285	-1.642	0.743		62:0613
63:0614	-0.653	-1.310	1.578		-0.500	-1.266	1.000		-0.785	-1.357	1.471		63:0614
64:0708	-1.692	-2.344	2.043	0.05	-1.916	-2.333	0.792		-1.500	-2.357	2.237	0.05	64:0708
65:0709	0.576	0.034	1.867	0.10	0.666	0.000	1.505		0.500	0.071	1.086		65:0709
66:0710	-0.153	-1.275	3.399	0.01	-0.250	-1.133	1.552		-0.071	-1.428	3.712	0.01	66:0710
67:0711	-1.153	-1.482	1.027		-1.083	-1.166	1.158		-1.214	-1.285	0.172		67:0711
68:0712	-0.807	-0.172	-1.962	0.05	-0.916	-0.333	-1.107		-0.714	0.000	-1.194	0.10	68:0712
69:0713	-1.038	-1.413	1.192		-1.083	-1.533	0.922		-1.000	-1.285	0.672		69:0713
70:0714	-1.153	-1.931	2.638	0.05	-1.500	-2.133	1.581		-0.857	-1.714	2.027	0.10	70:0714
71:0809	1.000	0.241	2.424	0.05	0.916	0.266	1.402		1.071	0.214	1.949	0.10	71:0809
72:0810	1.115	0.241	2.817	0.01	1.500	0.133	3.415	0.01	0.785	0.357	0.918		72:0810
73:0811	-0.646	-0.517	-0.459		-0.916	-1.000	0.249		-0.785	0.000	-1.863	0.10	73:0811
74:0812	-0.846	-0.862	0.057		-0.916	-1.000	0.179		-0.785	-0.714	-0.219		74:0812
75:0813	0.423	-0.206	1.957	0.05	0.333	-0.133	0.883		0.500	-0.285	1.990	0.10	75:0813
76:0814	0.692	-0.482	2.905	0.01	0.750	-0.533	1.953	0.10	0.642	-0.428	2.114	0.05	76:0814
77:0910	-2.192	-2.448	0.965		-2.333	-2.533	0.572		-2.071	-2.357	0.703		77:0910
78:0911	0.846	0.758	0.355		0.750	0.733	0.053		0.928	0.785	0.370		78:0911
79:0912	1.000	0.620	1.359		0.916	0.533	0.866		1.071	0.714	0.988		79:0912
80:0913	-2.000	-2.517	1.876	0.10	-1.916	-2.533	1.439		-2.071	-2.500	1.171		80:0913
81:0914	-0.846	-1.655	2.326	0.05	-0.750	-1.600	1.352		-0.928	-1.714	2.272	0.05	81:0914
82:1011	1.153	0.965	0.642		1.416	1.000	0.999		0.928	0.928	0.000		82:1011
83:1012	1.269	1.172	0.346		1.500	1.000	1.481		1.071	1.357	-0.642		83:1012
84:1013	-0.346	-1.275	2.843	0.01	-0.666	-1.333	1.397		-0.071	-1.214	2.526	0.05	84:1013
85:1014	-0.536	-1.379	2.379	0.05	-0.833	-1.800	1.765	0.10	-0.285	-0.928	1.493		85:1014
86:1112	-2.192	-1.724	-2.000	0.05	-2.416	-1.933	-1.677		-2.000	-1.500	-1.393		86:1112
87:1113	-1.384	-1.344	-0.132		-1.166	-1.333	0.329		-1.571	-1.357	-0.607		87:1113
88:1114	-1.153	-1.655	1.460		-0.916	-1.666	-1.339		-1.357	-1.642	0.697		88:1114
89:1213	0.230	0.620	-1.138		0.416	0.600	-0.322		0.071	0.642	-1.403		89:1213
90:1214	-0.269	-1.206	2.389	0.05	-0.250	-1.733	2.302	0.05	-0.285	-0.642	-0.826		90:1214
91:1314	-2.538	-2.862	1.959	0.05	-2.583	-2.866	1.419		-2.500	-2.857	-1.326		91:1314

The means of the ratings for each item of the different samples are given in Table 4A and 4B.

For each item we checked in the six above mentioned comparisons whether observed differences between two sample means are indicative of the fact that the samples come from populations with unequal means. In testing the significance of the differences Students t for small and independent samples is used. T-values and relevant levels of significance are also indicated in table 4A and 4B.

3.0 RESULTS

3.1 Twin scales and dimensions

In order to verify whether each pair of twin scales (scale 1-2, 3-4, etc.) can be considered as really belonging together, all relation values < -1.50 are sorted out. If one of the values is < -1.50 all three values (all, female, and male raters, respectively) are given in table 5. For dimension I and III, which both show a tendency to split up, the relation values are given for both subdimensions if any value is < -1.50.

Table 5 - Correspondence of scales and dimensions according to all raters (MF, n=60), to male raters (M, n=26) and female raters (F, n=29). Relation values < -1.50 and values to match, are inserted in the table (see text 3.1). Minus signs and decimal points are omitted in the numbers.

Dim.	Scale	Rat.	02	03	04	05	06	07	08	09	10	11	12	13	14	Rat.	Sc.
Ia	01	FM	233	148	060	133	110			042				238	248	FM	01
Melod.		M	196	108	050	138	108			046				223	223	M	
		F	269	186	059	134	114			031				255	272	F	
	02	FM		155		168		173		027		153		233	223	FM	02
		M		112		173		173		031		154		188	192	M	
		F		207		169		186		041		145		272	248	F	
II	03	FM			191									225	228	FM	03
Artic.		M			138									196	185	M	
Qual.		F			238									259	272	F	
	04	FM												175	140	FM	04
		M												127	119	M	
		F												224	183	F	
IIIa	05	FM					233	185	103					202	225	FM	05
Clarity		M					181	131	081					177	192	M	
		F					276	234	127					228	259	F	
	06	FM						090	160					152		FM	06
		M						016	096					123		M	
		F						161	210					176		F	
IIIb	07	FM							207						152	FM	07
Subj.		M							169						115	M	
Strength		F							234						193	F	
	08	-														-	08
IV	09	FM									233			222	122	FM	09
Pitch		M									219			200	085	M	
		F									245			251	166	F	
	10	-														-	10
V	11	MF											192		150	MF	11
Tempo		M											219		115	M	
		F											172		165	F	
	12	-														-	12
Ib	13	MF													272	MF	13
Eval.		M													254	M	
		F													286	F	

From these data the following conclusions can be drawn.

1 High correspondences exist within the twin scales, so in all likelihood the two scales of each pair represent the same dimension.

2 The average of the four relation values of the scales 01 and 02 with 13 and 14 is very high (-2.36), indicating a functional equivalence. This is supported by the extent to which both pairs display the same pattern of interrelatedness across other scales. This impression of similarity shows that scale variance alone does not bring about a splitting up of the appreciation dimension, which implies that Ia:Melodiousness and Ib:Evaluation can be considered as one dimension or as subspaces of the same dimension.

3 The Voice Quality dimension (III), on the other hand, does seem to fall apart. The mean relation value of the scales 05 and 06 with 07 and 08 is rather low (-1.35). It is noteworthy that this is not caused by low correlations of all four scale combinations, but by the low degree of interrelatedness of scale 05 and 08 (-1.03) and of 06 and 07 (-0.90), with relation values smaller than -1.50 for their counterparts (48:0507, 57:0608). A partial explanation can be found in different connotations of the same term for female and male and concurrent difference in rating behaviour. Impressionistic analysis of the scales concerned indicates e.g. such a difference in connotation between scale 07:'weak-powerful' and scale 08:'soft-loud': scale 08 lacks the appreciative aspects that 07 has, e.g. 'weak' is related with monotonous, broad, ugly and unpleasant. Female raters indicate stronger appreciative connotations than men do and consider the Strength scales 07 and 08 more suited for the description of the male voice, where male emphasize that these scales are less suitable to describe the female V&P. (Further validation studies on these data by means of factor analyses are being conducted and will be available shortly.)

3.2 Sex of speaker

In the opinion of all raters as a group the relations between the scales are not dependent on sex of speaker (see table 4A.) The only significant exceptions ($p<.05$) are item 73:0811 and 86:1112 which both concern Tempo. When the ratings of female and male judges are considered separately the result is essentially the same. At 5% there are no significant differences for male raters, whereas female raters differentiate between female and male voice on three scale combinations only (73:0811, 85:1014, 90:1214), two of which again concern Tempo.
If the threshold of significance is lowered to 0.25 there are nineteen items in which the mean relation of scales is higher for the male than for the female voice, and eleven of those combinations apply to Tempo. Furthermore, it is striking that all eight combinations of V:Tempo with the Voice Quality dimension (IIIa+IIIb) are at issue. This is caused primarily by the female raters who consider those combinations less appropriate in the description of the female voice. (See Table 6.)

Table 6 - Overall averages of relations between the eight combinations of scale 05, 06, 07 and 08 of dimension III and scale 11 and 12 of dimension V.

			Raters	
			M	F
			-0.80	-0.67
'voices'	♂	-0.90	-0.79	-0.91
	♀	-0.58	-0.83	-0.42

3.3 Sex of rater

An important and striking datum in our results is that, unlike sex of speaker, sex of rater influences the overall judgments considerably. Comparison of the mean scores of female and male raters (table 4B, sample B1) shows that they disagree in almost 50% about the relatedness of scale combinations. Prominent in those differences is that the women almost always indicate a closer relationship between the scales.
Moreover - as is shown in the following paragraphs - there is a nonrandom deviation from the true scores for several scales and dimensions, suggesting some change in factor structure, or at least differences of allocation of concepts within it, due to sex of raters.

3.3.1 Tempo (V)

Considering all 93 significant t-values ($p < .10$) in the three samples of table 4B which compare female and male raters, we meet four items in which the female judges do not indicate an interscale correlation higher than men do. These four exceptions (61:0612, 68:0712, 73:0811 and 86:1112) concern the tempo scales 11 and 12.
In such a case, in which one group scores generally more extreme, it is interesting to have a closer look at the items which the other group judges more extreme, even if a difference is not significant. There are in sample B1 fourteen of those items with negative t-values and twelve of them are again combinations with tempo scales. Of the remaining thirteen tempo items the women consider only two items significantly related (10:0111 and 90:1214), both in connection with Voice Appreciation. Female, unlike male raters are negative in their judgment when men speak slowly.

Summarizing so far, our female raters see consistently a higher degree of relatedness between the scales than men do, except when the Tempo dimension comes into play. 'Dragging-brisk' and 'slow-quick' seem to be more male oriented scales. Men consider the tempo terms suitable qualifiers, whereas women judge them, especially in the description of the female voice, less applicable.

3.3.2 Pitch (IV)
It is obvious that positive correlations can be expected between the scale poles which are judged desirable (resp. undesirable). Nevertheless there turn out to be thirteen scale combinations with a negative relationship (larger than half a scale unit). This is independent of sex of rater or speaker; it is the Pitch dimension which seems to be involved. The 'sociogram' in figure 1 shows the thirteen negative scale relations, from which ten are related to pitch (scales 09 and 10). Moreover, ten of the remaining pitch scales have relation values around zero with other scales.

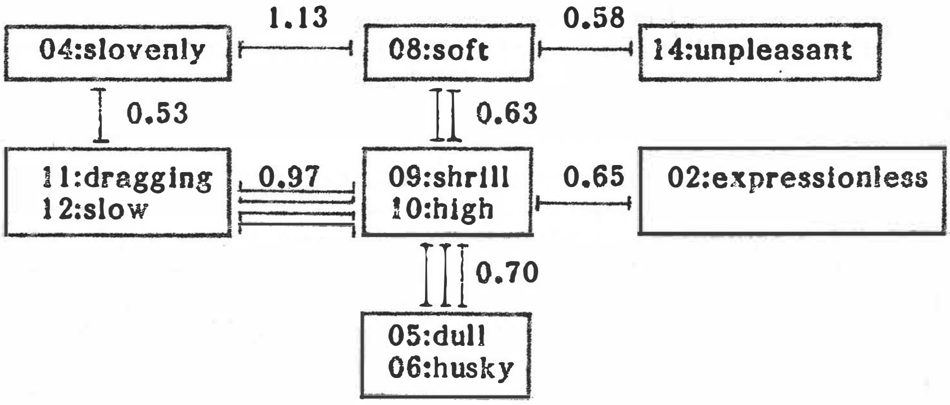

Figure 1. 'Sociogram' of negative relations between scales. Connecting lines indicate direction and number of negative correlations; adjoining the mean relation values.

This is rather puzzling at first sight since all scales, including 09 and 10, are polarized and scored with the scale term closest to the Ideal V&P value to the right. The explanation can be found in the relation values of the pitch versus evaluation scales. Their four combinations (80:0913, 81:0914, 84:1013, 85:1014) are positively correlated.
So it appears that Pitch is unrelated to all scales except 13 and 14, meaning that it has a characteristic evaluative connotation which does not implicate melodiousness. This would make Pitch an attractive and rather pure dimension, but close reading of the data reveals a noticeable number of irregularities.
As noted before the relation value of scale 09 and 10 is very high (-2.33), but their patterns of relatedness across the other scales are quite different. This is also caused by a number of significant differences which exists between the judgments of the two sexes (see table 7).

The most striking of those differences are the following.
- Male raters suggest a relation, for male and female voices, between 'clear' versus 'shrill' and 'high', between 'husky' versus 'low', whereas the female raters do not indicate this relation between dimension IIIa and IV.
- The qualifiers 'weak' and 'soft' of the Strength dimension are associated with 'high' by women, and with 'low' by men.

- There is a significant higher correlation between negative evaluation and the pitch qualifiers 'high' and 'shrill' in the opinion of females than according to males. This is particularly the case as far as the qualifier 'high' is concerned in relation to the male voice. 'High-low' has more to do with the female voice; 'deep' is more positively associated with the male voice. This is supported in data from Boves et al.(1982), which show that the average "own voice judgments" on the scale 'high-low' is much farther from the Ideal for women than it is for men, while on the scale 'shrill-deep' the reverse is the case. The different aspects of the pitch dimension evidently do not have identical meanings for men and women.

Table 7 - Relation values of all scales with pitch scale 09 and 10 (N=60). Between brackets the level of significance if a difference exists between female and male raters. (Decimal points are omitted.)

IV:Pitch	Ia:Melodiousness		II:Artic.Quality		IIIa:Clarity	
	01:monot.	02:expr.	03:broad	04:slov.	05:dull	06:husky
09:shrill 10:high	+42 +27	-27 +67 (10)	-70 -15 (10)	-50 (05) +23	+115 + 45 (01)	-13 (10) +65 (05)

IV:Pitch	IIIb:Strength		V:Tempo		Ib:Evaluation	
	07:weak	08:soft	11:dragg.	12:slow	13:ugly	14:unpleas.
09:shrill 10:high	+32 (10) -68 (01)	+65 (05) +70 (01)	+ 73 +105	+ 80 +112	-222 (10) - 70 (01)	-122 (05) - 90 (05)

3.3.3 Voice Appreciation (I)

Female and male judges assess most aspects of Ib:Evaluation differently. The two sexes disagree significantly about the degree of association in 19 out of the 25 combinations of evaluation scales 13 and 14 with all other scales. In these combinations men consider the relatedness of scales less high, in other words women show a tendency to ascribe more evaluative connotations to the different V&P dimensions, Tempo excepted.

Ia:Melodiousness and Ib:Evaluation have similar patterns of interrelatedness across most other scales. This, together with their high mutual relation values (see 3.1.2), gives the impression that Ia and Ib form part of the same dimension, which we tentatively called Voice Appreciation. Three scales (04, 06 and 09) differentiate between Ia and Ib: a slovenly speaking, husky and shrill voice is neither pleasant nor beautiful, but these characteristics do not affect the Melodiousness of the speakers.

The difference in behaviour of female and male raters holds, as anticipated, for this joined dimension too. Women consider - irrespectively of sex of speaker - the Appreciation factor closer linked with the other scales than men do. Men differentiate in this respect between female and male voices, and indicate relatively stronger appreciative aspects when the female voice is concerned. Raters of both sexes describe 'beautiful' as almost synonymous

with 'expressive', but require a higher level of expressivity from the female voice. In general, the female speaker primarily has to have a higher articulation quality, whereas the male speaker is sooner negatively appreciated when he speaks slowly with a weak and high voice.
The raters agree that there is a clear relation between I:Voice Appreciation (Expressivity excepted) and II:Articulation Quality, but men consider this relationship significantly weaker than women. According to men the connotations of Articulation Quality are mainly restricted to these appreciative aspects, but women describe broad speaking - especially by a man - also as monotonous, dull, husky, weak and shrill.

An otherwise interesting observation is that there are three scales with a low correlation with Voice Appreciation, viz. the psychophysical scales 08:'soft-loud', 10:'high-low' and 12:'slow-quick'. However, the respective twin scales (07, 09 and 11) show considerable correlations with the same dimension. The latter scales all have - according to their Ideal V&P value (see table 1) - a rather clearly defined negative and positive pole. They are what Lemann and Solomon (1952) call 'alpha scales', in contrast to 'beta-scales' which have the positive position between two negative poles. Since the psychophysical scales are of type 'beta', the differences in scale behaviour can at first sight be explained as an artefact of the correlation method. However, the relation values between the twin scales themselves are high (see table 5) which suggests another possibility, namely to distinguish denotative scales who lack the appreciative associations from connotative scales.

4.0 DISCUSSION ON THE SUBJECT OF RATER VARIANCE

This simple experiment which we performed, rendered a lot of information concerning the instrument and the raters. It showed that the judgments are not only based on actual speech characteristics but also on the idiosyncrasies of the listener.
In earlier studies (van Herpt et al., in prep.; Boves, 1984) we did not find substantial correlations between the perceived speech characteristics and supposed acoustic criteria of these attributes. Boves (1984:163) suggests that this might be the result "of an intricate, and probably highly nonlinear weighing of a large number of acoustic parameters", in which case the problem can be attacked from two sides. Other, higher-order, acoustic measures and/or perceptual descriptions on a lower phonetic level must be developed. Our comments in this discussion are about the perceptual side and concern especially listener effects which cloud the relationship between perceptual and acoustic features.

A positive result of the present study is that it strongly suggests that the dimensional structure of V&P is almost independent of sex of speaker. On the other hand there is quite a lot of variance brought about by sex of rater, which suggests that females and males might differ in their qualifying framework of speech description. Osgood, May and Miron (1975:57) report that they have no knowledge of studies in which significant variation in semantic factor structure between men and women are found; although there are, of course, differences based on sex in the meaning of particular concepts. In terms of our study this would mean that raters of both sexes

share a common semantic reference frame and that sex-related differences in meaning of V&P are expressed in differences in allocation of speakers within it. So, our next research goal is to decide whether or not female and male judges use a common semantic framework. To do so it is a necessity to assess the relative amount of variance of each of the three modes. The present study, from which speaker variation is methodologically excluded, explores primarily the listener mode variance.
Variance consists of 'true' variance and error variance. In rating experiments 'true' variance is due to the stimuli, e.g. the speakers. Error variance must be divided in random error or 'noise' and biased error or distortion. Random error is the variation that can be ascribed to the imprecision of the instrument and error that is caused by individual differences and temporal variations in responses of the judges. In contrast with biased error, this type of variance can be diminished or eliminated by standard statistical techniques, e.g. by 'repeating' the measurements. With the scales we used, we reach an effective reliability of 0.90 or higher when about 25 raters are involved (Fagel et al., 1983:322).
Biased error is by definition due to a systematic error that disturbs our analyses. The major problem is that it derives from a latent influence that in many cases is not recognized beforehand.
A systematic error which is obvious in our investigation is style of scale checking, which seems to be sex-related. Men appear to avoid the endpoints of the scales and use more often the intermediate positions; women score more extreme, which in the present case amounts to higher correlations between scales. This difference in scoring behaviour has been found many times (McC.Miller, 1974) and the core of most proposed explanations is that women tend to distort their opinion in the direction of social desirability. Our data point to it that women weigh the appreciative connotations of qualifiers they consider relevant, heavier than men do.
This appreciation bias seems to affect the scaling unit only and not to influence the semantic dimensional frame of the raters. In factor analyses on which we are presently working we'll check whether this supposition is correct. If it is, the bias can be controlled either by assigning equal numbers of men and women to the raters' panel, or by attempting to measure the effect in order to control for it statistically.

But there are more distortions in the scores of raters, such as sex-related correlations between scales.
The judges seem to be liable to halo-effect: a tendency to bias their judgments on the basis of one particular feature. The ratings of specific voice characteristics are - although the twin scales representing the five dimensions are meant to be unrelated - guided by a general impression of the speaker or by a striking quality of the speaker or his speech. This causes the same voice to be evaluated differently in consequence of information on a distinguishing feature such as age or sex. When a voice is identified as that of a male it is judged more in relation to Strength and Tempo dimensions, whereas a female voice is significantly stronger related with Evaluation. These dimensions then serve as points of reference from where the halo radiates to other scales. So, when the correlations between scales from reference dimensions and the other scales are calculated, the sizes of the coefficients vary considerably depending on sex of speaker, i.e. all ratings of female speakers tend to be systematically biased in one direction, those of males in another.
The problem is how to distinguish this bias which obscures the pattern of attributes within the object V&P from true conjunction of positive and

negative qualities. The usual method to prevent or reduce halo-effects when such a complex concept as V&P is rated, is to decomposite the complex in its distinctive elements and have them rated on separate scales. Since this approach is inherent already in the semantic differential technique we used, we tried - on a small scale - two additional procedures.
First, the judgment procedure was changed in such a way that ten voices were judged successively on a single rating scale instead of each voice on all successive scales. This try-out with five listeners did not show a significant shift in mean scores. Similar results are obtained by Boves (1984:14). Secondly, the naive raters of the normal procedure were replaced by (three) trained judges. The interjudge reliability of the experts indicates that a smaller number of raters can then be used. However, the mean scores, i.e. the validity, were hardly affected, which provides another argument for the suitability of naive raters and with that for the generality of the scales. In sum, these procedural manipulations did not effectuate significant changes in the perceptual ratings, so we'll have to try to control the halo-effect statistically. One possibility is to identify the most important sex-distinguishing scales and then investigate the relationship between the other scales with one or more reference scales held fixed.

But judges make many constant errors. Another mechanism producing systematic bias appears anew from our study. Female and male raters don't have the same image of either a man's or a woman's voice. They lay different (degrees of) relations between scales and emphasize different dimensions, but each of the sexes tends to agree in its attribution of differential speech characteristics. Commonly this phenomenon is called stereotyping. The American journalist Walter Lippmann (1922:16) who was the first to use this term in connection with social perception, defines a stereotype as a simple cognition on the basis of which "the real environment (which) is altogether too big, too complex and too fleeting for direct acquaintance" can be handled. Stereotypes can be understood as consensually preconceived conceptions concerning assumed characteristics of an individual on the basis of his group membership. The existence of stereotypical conceptions concerning V&P is supported in several studies (Kramer, 1977; Boves et al.,1982). From these studies it appears that the V&P scores of a man or a woman are distorted in different directions. Our study points to it that this is more strongly influenced by the sex of the rater than by the sex of the speaker. This means that Lippmann's definition must be tightened in that sense that the consensually preconceived conceptions "are shared by the members of a social group whose composition depends on the object under consideration". In the present case the raters do not belong to the same sex group and to study their stereotypes and prejudices concerning the female as well as the male V&P, both groups must be treated separately.
The result of stereotyping resembles the halo-effect in that the perceptions of the rater are transformed in such a way as to agree with this general conception. Raters have, as Lippmann calls it, different "pictures in the head" of V&P, which cause men and women to accentuate different attributes. These stereotypical conceptions can be considered as centers of gravity whose haloes radiate to other features and influence their values. When assuming these two phenomena the main methodological problem is to separate their effects from the natural covariation of positive or negative features. With respect to the halo-effect we mentioned the possibility to control it statistically. Sex role stereotypes which influence the way raters respond to men and women most probably can be controlled experimentally by withholding the raters knowledge of the speaker's sex. For the latter purpose

an experiment with manipulated stimuli is conducted; unfortunately results are not yet available.

We have seen that many scales have a sex-related tendency to be contaminated with appreciative aspects. Women ascribe appreciative connotations to the different V&P dimensions. Men do the same but to a lesser degree, especially with regard to the male voice. So, one way to make the ratings of men and women concerning the female and the male voice more comparable, is to use scales with less emphasis on the appreciation factor. An extra and desirable result would be that more factors of a denotative sort could be expected to appear and that the all including appreciation factor itself will break down. However, it appears to be very difficult to find many specific scales which are orthogonal with respect to appreciation and have their variance (almost) entirely in one dimension. Our analyses (Blom & Koopmans, 1973; Blom & van Herpt 1976; Fagel et al., 1982; Boves, 1984) which started from over 800 adjectives referring to V&P, yielded only three acceptable denotative scales, viz. 'soft-loud', 'high-low' and 'slow-quick'. Given our failure to control the appreciative aspects experimentally, it is indicated to remove the effect of this variable statistically. Partial correlation calculation provides us with a measure of strength of the correlation between the scales while holding the effect of one or more scales in the relation between the other scales constant. Analysis of the partial correlations will enable us to expose spurious correlations, which are among other things caused by halo-effects. E.g. it is conceivable that the correlation between scale 05:'dull-clear' and 07:'weak-powerful' (r=.60) is the result of the fact that scale 07 varies along with evaluative scales 13 and 14 (r=.45) which are also intrinsically related with scale 05 (r=.70). In this case, with Evaluation held constant, 'dull-clear' would no longer vary with 'weak - powerful' and further insight would be gained in the relationship of the Clarity and Strength dimensions. This points to the following solution. When the partial correlation matrices are factorized it is to be expected that, due to the great reduction of variables with a strong appreciative character, the proportion of variance explained by the first factor decreases in favour of the explanatory power of the next factors extracted. The resulting denotative factors then, although minor in terms of explained variance, will be interpretable on a purer phonetic level and as such may play an important role in our perceptual description when comparing subjective judgments with acoustic measures.
Finally, an improvement in the scoring procedure itself must be considered. Our results repeatedly demonstrated deviating behaviour of the denotative scales (08:'weak-soft', 10:'high-low' and 12:'slow-quick') which can but partly be explained by lacking connotations. Especially the low communalities of these three scales (<.50) found by Fagel et al. (1983:320) signify a great quantity of unexplained variance composed of specificity and error. It is unlikely that three different scales each have - apart from their rather pure factor loadings - another variance that typically characterizes them. So, we must assume that the uniqueness consists predominantly of error variance.
This error then can be explained as an artefact of our statistic: Pearson's product moment correlation coefficient which is based on linear relationships. The three denotative scales are beta scales (see 3.3.3) as appears from the fact that they have their scale values of Ideal V&P less than one scale unit from the center of the scale (see table 1). All other scales are of the alpha type, so the relation between both types is bound to be curvilinear and use of a straight line to represent the general pattern of the data

artificially lowers the coefficient of association. There are several ways to prevent this. However, it is complicated by the fact that the artefact is intertwined with rating distortions.
We propose a solution which kills two birds with one stone.
Assume the Ideal V&P value of each scale to be the positive maximum of that scale, divide the longer tail in equal intervals on a scale of e.g. 0-100 and scale the smaller tail with the same unit.
This data treatment is supposed to have several effects.
First, all scales are scores as standard alpha scales. Secondly - when the calculations are done separately for female and male raters in connection with Ideal V&P values according to **raters** of the corresponding sex - this procedure also corrects for sex-related scale checking style. And, thirdly - when also the Ideal V&P of the female and male **speaker** are taken into consideration - stereotypical conceptions concerning V&P of men and women are to a certain extent controlled too. So, this type of data manipulation is the first step to be considered in order to correct several systematic biases.

5.0 CONCLUSION

A major problem in perception experiments is to assess how far listeners' ratings are based on actual differences in speech production and how far the responses are influenced by (systematic error) variables that are not covered by the acoustic criteria against which is validated.
Our data show that voice perception is likely to be affected, among other things, by sex of the perceiver. This does not necessarily mean that female and male raters use different frames of reference. Roughly there is a lot of agreement among all raters concerning the direction of relatedness of scales. But when female and male raters do actually allocate speakers in the same space these allocations are also differentially determined by the sex of the rater. This implies that to increase the validity of perceptual ratings, attention must be paid to general habits, interests, expectations, attitudes, prejudices and stereotypes that are shared by groups of judges.
A consequence is that perception experiments in which sex of rater is not a considered variable are not acceptable or at least must be judged very critically. In quite a lot of publications sex differences of subjects or objects are not mentioned at all. We support Hoogstraten's position (1979:75) that this omission makes any interpretation very precarious. If potential sex differences are not examined, it is very likely that interaction phenomena between attributes of speakers and raters remain concealed. When only sex of speaker is taken into consideration, it is even likely that at least some of the reported sex differences of speakers have to be ascribed to the listeners' sex. And when the use of subjects is limited to one sex or to the other, we generally consider that a bad solution because - apart from chances of overlooking important sex-related differences - it severely limits the applicability of research findings.

In the fore-going we amply stated that the judgment of V&P is not only determined by its objective qualities, but also by rater characteristics. In other words, the listener mode has to be controlled. We proposed a few data treatments in order to accomplish that listener variance is small. In many perception experiments this is wrongly taken for granted. And only, as is explained by Osgood et al.,(1975), when this is the case the resulting factorial structure is attributable to an underlying organization of scale terms as applied to speakers. The speaker mode was controlled in our study methodologically; we employed a design which itself eliminated individual speaker differences. So the resulting factorial structure of the scales cannot be attributed to the particular sample of speakers used. Along the dimensions of this qualifying framework judgments are expected to vary meaningfully, so that all potential voices find expression in differences of allocation. Thus, to be able to make unambiguous interpretations concerning the structure of any mode in this type of investigation it is a necessity to assess the contribution of each of the classification modes to the total amount of variance. And, only when the listener effect and its interactions are indeed relatively small the resulting structure is adequate, otherwise further corrections of the type proposed in the preceding discussion are required. A conclusion must be that in this type of research three mode factor analysis or multidimensional scaling techniques must take the place of the standardly used two dimensional techniques.

ACKNOWLEDGMENT

I would like to thank Louis Pols and Florien Koopmans for reading an earlier version of this article and for their comments on it.

REFERENCES

Blom, J.G. & Herpt, L.W.A. van (1976). The evaluation of jury judgments on pronunciation quality. Proc. Inst. Phonetic Sciences, Univ. of Amsterdam, 4, 31-47.

Blom, J.G. & Koopmans-van Beinum, F.J. (1973). An investigation concerning the judgment criteria for the pronunciation of Dutch. Proc. Inst. Phonetic Sciences, Univ. of Amsterdam, 3, 1-24.

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

Boves, L., Fagel, W.P.F. & Herpt, L.W.A. van (1982). Conceptions of women and men concerning the speech of men and women. (In Dutch) De Nieuwe Taalgids, 75-1, 1-23.

Fagel, W.P.F. & Herpt, L.W.A. van (1982). Analysis of the perceptual qualities of voice and pronunciation. Proc. Inst. Phonetic Sciences, Univ. of Amsterdam, 7, 1-25.

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

Herpt, L.W.A. van & Fagel, W.P.F. (1981). Sex influences in the judgment of voice and pronunciation. (In Dutch) Toegepaste Taalwetenschap in Artikelen, 9, 146-156.

Herpt, L.W.A. van, Fagel, W.P.F. & Boves, L. (in prep.). A rating instrument for voice and pronunciation in Dutch.

Herpt, L.W.A. van & Hoebe, A.P. (1985). Attribution of age from perceived speech. Proc. Inst. Phonetic Sciences, Univ. of Amsterdam, 9, 1-23.

Hoogstraten, J. (1979). De machteloze onderzoeker Boom, Meppel, Amsterdam.

Lemann, T.B. & Solomon, R.L. (1952). Group characteristics as revealed in sociometric patterns and personality ratings. Sociometry, Vol. 15, 7-90.

Lippmann, W. (1922). Public Opinion. New York, Harcourt Brace.

McC.Miller, P. (1974). A note on sex differences on the semantic differential. Brit. J. Soc. Clin. Psychology, Vol.13, 33-36.
Osgood, C.E., May, W.H. & Miron, M.S. (1975). Cross-cultural universals of affective meaning. Urbana, University of Illinois Press.
Osgood, C.E. & Suci, G.J. (1955). Factor Analysis of Meaning. J. of Exp. Psychology, 50, 325-338.