ANALYSIS OF THE PERCEPTUAL QUALITIES OF VOICE AND PRONUNCIATION*

by Wil P.F. Fagel and Leo W.A. van Herpt

1. INTRODUCTION

When we examine the speech product of any speaker acoustically, we will be confronted with a very complex signal, changing continuously in time, mainly as a function of linguistic meaning. In spite of all this, we can distinguish and recognize different speakers of one language easily by ear, even when all speakers read out the same text, thus ruling out all cues due to lexical and grammatical variation. It even appears that we can attach various verbal labels in a very consistent way to the speech products of different people. Among these labels there are many based upon voice quality and features of articulation, for example aesthetical and social-evaluative labels (for an extensive survey on this subject see: Scherer & Giles, 1979).
In so far the attribution of such labels is based solely upon auditory information, the acoustic signal must contain certain features which are responsible for eliciting the perceptual processes involved. These features are essentially measurable, though it is not easy to find a compact descriptive system that can be used for an efficient examination of the relation between acoustic and perceptual characteristics of voice and pronunciation. For this part of our investigation we refer to the work done by Koopmans-van Beinum (1980) at our institute and by Boves at the Instute of Phonetics in Nijmegen (Vieregge & Nuytens, 1978; Boves, 1981). We will limit ourselves here to the perceptual side of the study.

*This research is supported by the Netherlands Organization for the Advancement of Pure Research (ZWO), project nr. 17-21-13.

If we want to analyse the relation between acoustic and perceptual features in an adequate way, it is also important to have a reliable and economical instrument for measuring the relevant perceptual characteristics of voice and pronunciation. In an attempt to construct such an instrument for running speech produced by non-pathological native speakers of Dutch, a series of experiments was set up at the Institute of Phonetic Sciences in Amsterdam (Blom & Koopmans-van Beinum, 1973; Blom & Van Herpt, 1976). We will first discuss these experiments and their results, and report about our follow-up investigation afterwards.

2. PRECEDING RESEARCH

In Dutch, as in other languages, there are many ways to characterize voice and pronunciation in a subjective manner, that is, there are hundreds of adjectives which can be used – and indeed are being used– to describe speech characteristics bearing specifically on voice and pronunciation. Although a subjective description does not always imply statistical unreliability, one cannot a priori assume a commonly used term to mean the same thing for every one.
But even if we would reject 90% of these terms as being too unreliable for our purposes, we would still be left with an inefficient number of adjectives to describe voice and pronunciation. Inefficient not only because of their number, but also because of the redundancy in the information they yield. It is easy to see that in judging someone's voice with a number of terms there may be many interrelations between these terms.
However, by quantifying these relations we might get insight into the structure of the present perceptual system, which may help us to straighten out the tangle of perceptual attributes in a rather objective way. This is actually what we have been trying to do.
First, some 800 terms referring to special attributes of speech were collected. By pairing contrasting items from this collection a list of bipolar seven-point rating scales was composed. Out of this list 46 scales were selected for the rating procedure.

These 46 scales were used by 200 listeners to judge the experimental stimuli. These stimuli consisted of tape recordings of a written story freely retold by five male and five female speakers, who were recruited from different social settings and different levels of education.
By this measurement method, which is actually a form of the Semantic Differential Technique (Osgood, Suci & Tannenbaum, 1957), we can obtain a lot of information in a relatively quick and easy manner. The semantic differential approach has been proved a useful method for characterizing the perceptual correlates of complex physical stimuli, including acoustic ones. Examples can be found at Von Bismarck (1974), Voiers (1964), Solomon (1958), Uldall (1960), Takahashi & Koike (1975) and many others.

The structure of the ratings obtained was studied by means of factor analysis. The scorings of regionally different subsets of listeners were factorized separately. In all cases four independent factors emerged. However, these factors were only partly identical, so it was concluded that the various groups of listeners might have a differing frame of judgment (Blom & Koopmans-van Beinum, 1973). After this conclusion however, it was decided to reconsider the data. Each of the scales from the rating form was tested separately on a number of criteria, like monotonicity, linearity and discriminative power. This eventually resulted in the elimination of 19 scales. The remaining 27 scales were submitted to factor analysis again, which revealed three common orthogonal factors. These factors accounted for 47.7% of the total item variance and could be characterized as "voice appreciation", "articulation quality" and "abnormality" (see Table 1). The first factor extracted accounted for 65.4% of the total explained variance, the second factor for 24.6% and the third factor for 10.0% . Mean factor scores for each of the 10 speakers were calculated and these are illustrated in Figure 1.
This factor structure appeared to be highly stable over speakers (male/female), and listeners (Dutch listeners from the west of the Netherlands and Dutch speaking listeners from Belgium).

Table 1 - Varimax rotated factors resulting from judgments of speech samples (retold stories) from 10 speakers on 27 bipolar rating scales. Loadings ≥ .45 have been outlined below.

Dutch scale terms	English equivalents*	Factor loadings F-1	F-2	F-3
expressief/uitdrukkingsloos	expressive/expressionless	-.77	.07	.13
melodieus/eentonig	melodious/monotonous	-.76	.04	.23
doods/levendig	spiritless/vivacious	.76	.04	-.05
flets/klankrijk	colourless/sonorous	.74	-.13	-.29
kwiek/zeurig	sprightly/whining	-.69	.01	-.19
stereotiep/gevarieerd	stereotyped/varied	.65	-.10	-.09
lelijk/mooi	ugly/beautiful	.63	-.35	-.40
aangenaam/onaangenaam	pleasant/unpleasant	-.58	.23	.37
arm/rijk	poor/rich	.58	-.38	-.28
warm/koud	warm/cold	-.55	.02	.13
krachtig/zwak	powerful/weak	-.46	.22	.20
hedendaags/ouderwets	contemporary/old-fashioned	-.44	.07	.04
dof/helder	dull/clear	.44	.02	-.38
afwijkend/normaal	deviating/normal	.42	-.37	-.39
vast/onvast	steady/unsteady	-.41	.33	.39
nasaal/niet nasaal	nasal/non-nasal	.28	-.25	-.03
bekakt/ordinair	la-di-da/vulgar	-.07	.83	.06
plat/beschaafd	broad/cultured	.32	-.81	-.15
gekultiveerd/onverzorgd	cultivated/slipshod	-.24	.78	.37
ongekunsteld/geaffekteerd	artless/affected	-.12	-.71	-.06
slordig geartic./hyperkorrekt	carelessly artic./hyper-correct	.14	-.70	-.36
gediftongeerd/niet gediftongeerd	diphthongized/not diphthongized	.20	-.52	.04
gewichtig-speels	pompous/playful	.30	.45	.10
snel/langzaam	quick/slow	-.11	-.11	-.66
helder/hees	clear/husky	-.23	.04	.64
gerekt/verkort	drawn out/clipped	-.09	.15	.45
vloeiend/staccato	smooth flowing/staccato	-.34	.11	.41

* Translations are an approximation of the original scale terms. We must warn for the 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.

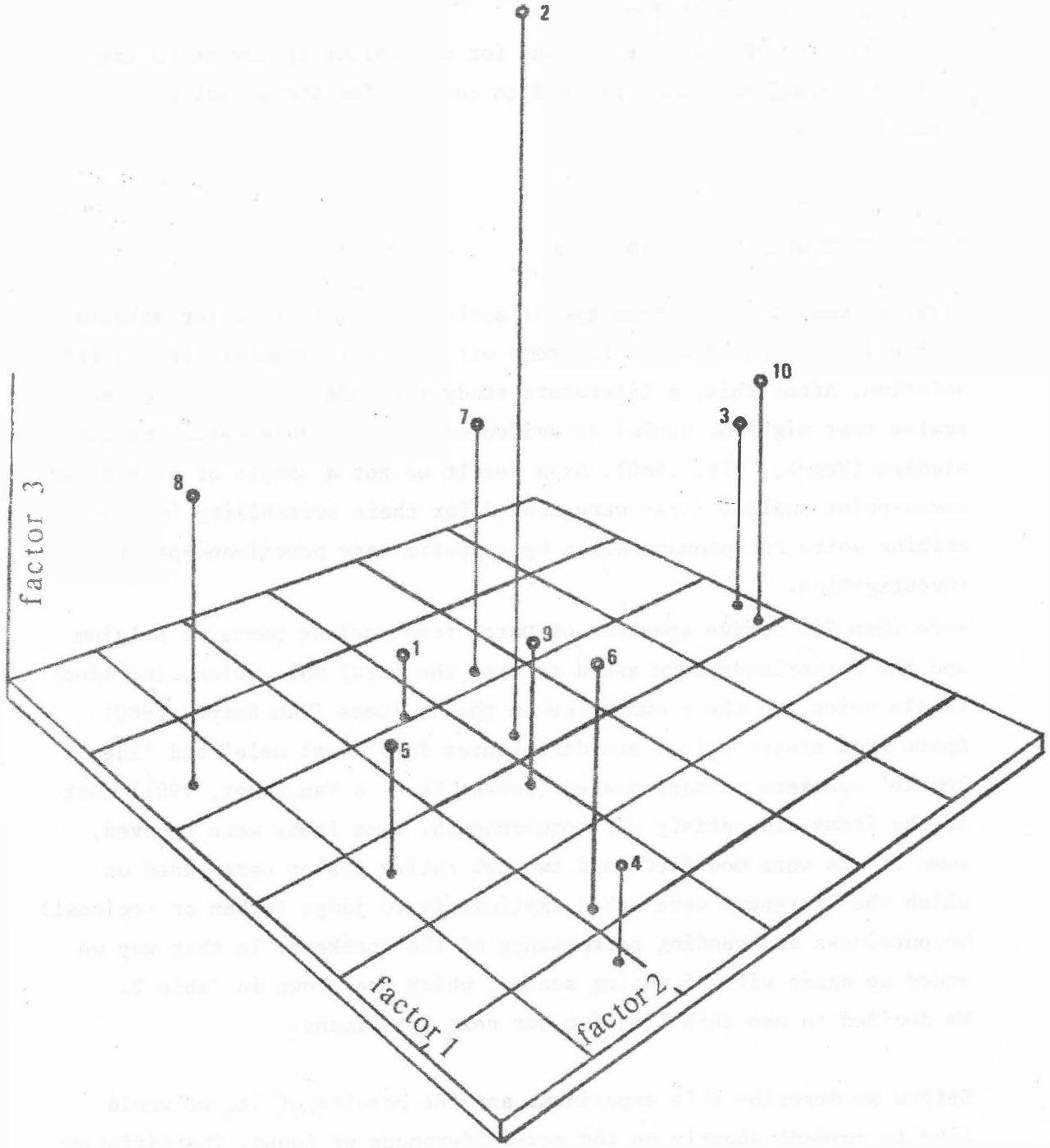

Fig. 1 - Positions of the 10 speakers judged (retold stories) in the 3-dimensional factor space described in Table 1. The dimensions have been labelled 'Voice Appreciation' (factor 1), 'Articulation Quality' (factor 2), and 'Abnormality (factor 3). (Adapted from Blom & Van Herpt, 1976. Slightly corrected).

It should be emphasized, however, that the ratings of the listeners in this experiment might have been influenced by lexical and grammatical variations, since the stimuli consisted of freely retold stories, as remarked before.
Therefore, one of the main reasons for our follow-up investigation to be discussed next was the need to control for the effect of these factors.

3. DEVELOPMENT OF A NEW RATING FORM

First we made a choice from the 27 scales of our last factor solution (Table 1), giving priority to items with highest communalities in this solution. After this, a literature study was made to look for other scales that might be useful as evidenced by comparable speech rating studies (Fagel, 1979, 1980). As a result we got a sample of 35 bipolar seven-point scales. These were tested for their suitability in describing voice and pronunciation by a preliminary pencil-and-paper investigation.
More than 200 native speakers of Dutch from various parts of Belgium and the Netherlands were asked to rate the ideal male voice, the ideal female voice and their own voice on the 35 items (Van Herpt, 1980). Apart from stereotypical sex-differences for 'ideal male' and 'ideal female' speakers on many scales (Boves, Fagel & Van Herpt, 1981) most of the items did satisfy our requirements. Some items were removed, some others were modified, and two new rating scales were added on which the listeners were asked explicitely to judge (urban or regional) accentedness and reading performance of the speakers. In this way we ended up again with 35 rating scales, which are shown in Table 2.
We decided to use this list for our next experiment.

Before we describe this experiment and the results of it, we would like to comment shortly on the sex-differences we found. The different criteria for what should be considered 'ideal' for male voices vs. female voices on many of the continua defined by the rating scales, will probably have an effect on the listeners' scoring of actual female and male speakers on the rating scales involved.

Table 2 - List of rating scales used in the follow-up experiment. N.B.: See note below Table 1.

Dutch scale terms	English equivalents
aangenaam/onaangenaam	pleasant/unpleasant
dof/helder	dull/clear
vriendelijk/kortaf	friendly/curt
luid/zacht	loud/soft
hees/niet hees	husky/not husky
stereotiep/gevarieerd	stereotyped/varied
aktief/passief	active/passive
ongekunsteld/geaffekteerd	artless/affected
flets/klankrijk	colourless/sonorous
slordig/precies	careless/precise
hoog/laag (voor een man/vrouw)	high/low (for a man/woman)
stevig/slap	firm/slack
krakerig/niet krakerig	creaky/not creaky
traag/vlot	dragging/brisk
melodieus/eentonig	melodious/monotonous
krachtig/zwak	powerful/weak
lelijk/mooi	ugly/beautiful
hortend/vloeiend	jerking/smooth flowing
doods/levendig	spiritless/vivacious
gerond/hoekig	rounded/angular
diep/schel	deep/shrill
arm/rijk	poor/rich
gespannen/ontspannen	tense/relaxed
expressief/uitdrukkingsloos	expressive/expressionless
vast/onvast	steady/unsteady
plat/beschaafd	broad/cultured
zelfverzekerd/weifelend	self-confident/wavering
zeurig/opgewekt	whining/cheerful
verzorgd/onverzorgd	polished/slovenly
afwijkend/normaal	deviating/normal
snel/langzaam	quick/slow
opgewonden/rustig	agitated/calm
duidelijk/onduidelijk	distinct/indistinct
In welke mate vindt u dat deze persoon met een (regionaal of stedelijk) aksent spreekt?	To which extent do you think this person speaks with a (regional or urban) accent?
sterk aksent/geen aksent	accentedness (high/low degree)
voorleesprestatie (goed/slecht)	reading performance (good/bad)

For example, on the continuum "powerful↔weak" the 'ideal male voice' is placed more to the "powerful" side than the 'ideal female voice'. This means that any possible acoustic parameter that shows a high positive correlation with perceptual "powerfulness" may be sex-dependent as well, in the sense that a male voice will need a much higher value on such a physical variable to be rated as "very powerful" than a female voice.
We will have to take this fact into account as soon as we are going to relate physical parameters to perceptual parameters that are based on sex-differentiating rating scales.

4. SPEECH MATERIAL

The same ten speakers from the previous experiment also supplied the speech material for our follow-up investigation. This time, however, the stimuli consisted of uniform texts, read aloud by the speakers. This way we hoped to control for the effect of speakers semantics upon listeners ratings.
To these 10 original speakers one male speaker was added for greater comparability of our results with those of a similar experiment done by Boves at Nijmegen. In this experiment 6 male speakers were each rated twice on the same scales while reading two different texts. By inserting one of our speakers in Boves' experiment and adopting one of his speakers for our investigation we got an overlap of two speakers. In this paper however we will limit ourselves to the results of our own experiment.
The 11 speakers were rated by different groups of listeners in one of two orders of presentation. This was done to control for possible speaker-dependent sequence effects upon listeners ratings.

5. RATING PROCEDURE

The rating procedure was essentially the same as in the previous experiments (Blom & Koopmans-van Beinum, 1973). The major changes

have already been pointed out:

1) several new rating scales;
2) speakers reading a text aloud instead of retelling a story;
3) one extra speaker to be judged.

Before the actual listening sessions took place, the subjects were asked to express their opinion about the Ideal Male Voice, the Ideal Female Voice and their Own Voice on the rating scales. This was primarily done to make the judges acquainted with the rating forms. It had the additional advantage that listener groups could be checked on norm differences and differences in 'self image'.
The listener-judges were also allowed, prior to the experiment proper, to hear brief samples of all the speakers to be rated, "as a means of experiencing the range and diversities of speech qualities involved, and of establishing a reference frame in terms of which to make their ratings" (Voiers, 1976).
In the experiment the speech samples of the 11 speakers were presented to the listeners one by one for at most 4 minutes, during which the complete rating form had to be filled in.

6. LISTENERS

The rating experiment was carried out with the following 8 groups of listeners, most of them consisting of students from training courses of Speech Therapists:

1) 24 students of Dutch Language from the University of Amsterdam.
2) 21 students from the Training Course for Speech Therapists (TCST) in Amsterdam; 2nd year of training.
3) 17 students from the TCST in Utrecht; 1st year of training.
4) 37 students from the TCST of the Katholieke Vlaamse Hogeschool in Antwerp (Belgium); 3rd year of training.
5) 49 students from the TCST in Hoensbroek; 2nd year of training.
6) 32 students from the TCST in Eindhoven; 1st year of training.
7) 31 students from the TCST in Eindhoven; 2nd year of training.
8) 24 students from the TCST in Eindhoven; 3rd year of training.

This makes a total of 235 listeners, all native speakers of Dutch themselves.
Though it is desirable to have data from a sample of listeners with a greater dispersion socially as well as regionally, lack of time forced us to work provisionally with the data obtained from the above groups.

7. RESULTS

7.1 Dimensionality of perceptual judgments.

First we calculated the correlations between all 35 scales for each listener group separately. Each resulting matrix of correlations was subsequently factored by the method of principal factoring with iteration. Only factors with an associated eigenvalue (characteristic root) of more than 1.00 were extracted. The initial factor matrices were rotated to a Varimax criterion. This resulted in highly similar factor solutions for all groups, independent of speaker order (Fagel, 1981).
Thereafter the 235 listener-judges were treated as one group and the above factoring procedure was repeated for this group. The Varimax-rotated solution of this analysis (Table 3) reflects the general structure we found in the partial analyses mentioned above.
Five factors accounted for 62,1% of the total item variance. Of this total explained variance, a proportion of 72,3% is accounted for by the first extracted factor; proportions of 10,4%, 7,4%, 6,1% and 3,8% are accounted for by the second to fifth extracted factors respectively.
We might conclude that we reached a highly stable factor structure in which the first factor is strongly dominating.

We will now take a closer look at these factors, comparing them with the perceptual dimensions resulting from the previous experiment (Table 1). For reasons of convenience we will in future refer to this study as the "Retell" experiment, and to our follow-up investigations as the "Read" experiment.

Table 3 - Varimax rotated factors resulting from judgments of speech samples (text read aloud) from 11 speakers on 35 bipolar rating scales. Loadings ≥ .45 have been outlined below.

	F-1	F-2	F-3	F-4	F-5
spiritless/vivacious	.80	.30	.26	.01	-.19
expressive/expressionless	-.80	-.32	-.23	.06	.18
melodious/monotonous	-.77	-.33	-.26	.04	.17
whining/cheerful	.74	.35	.18	-.06	-.20
ugly/beautiful	.72	.36	.33	-.24	.12
friendly/curt	-.71	-.21	-.08	.14	-.07
stereotyped/varied	.70	.24	.23	-.01	-.10
poor/rich	.70	.38	.34	-.21	-.02
colourless/sonorous	.70	.28	.45	-.01	-.07
pleasant/unpleasant	-.69	-.29	-.32	.25	-.08
active/passive	-.62	-.25	-.34	-.05	.30
rounded/angular	-.53	-.31	-.06	.31	-.07
tense/relaxed	.49	.33	.30	-.36	.17
deviating/normal	.45	.39	.40	-.16	.14
artless/affected	-.38	.23	-.02	.25	-.11
broad/cultured	.22	.79	.09	-.12	-.03
polished/slovenly	-.30	-.75	-.27	.08	-.04
careless/precise	.23	.68	.34	.03	.09
good/bad reading performance	-.45	-.63	-.20	.06	.00
distinct/indistinct	-.35	-.59	-.39	.07	-.07
accentedness (high/low degree)	.20	.58	.05	-.18	.05
self-confident/wavering	-.26	-.56	-.32	.17	.19
jerking/smooth-flowing	.42	.50	.23	-.23	-.04
husky/not husky	.26	.12	.66	-.01	-.10
dull/clear	.44	.10	.63	.24	.06
powerful/weak	-.37	-.33	-.63	.22	.20
firm/slack	-.39	-.37	-.60	.18	.17
loud/soft	-.03	-.17	-.55	-.06	.16
steady/unsteady	-.36	-.45	-.49	.29	.01
creaky/not creaky	.37	.26	.44	-.16	.18
deep/shrill	-.26	-.24	-.01	.76	-.09
high/low for a (wo)man	-.00	.09	-.05	-.67	.05
quick/slow	-.07	.04	.02	-.06	.69
dragging/brisk	.35	.16	.21	-.00	-.57
agitated/calm	.15	.33	.20	-.30	.51

The first factor from the Retell experiment also accounted for some 70% of the total explained variance. A comparison of the first Read factor with this first Retell factor shows them to be highly similar. On both factors the items 'spiritless↔vivacious', 'expressive↔expressionless' and 'melodious↔monotonous' have the highest loadings. Originally, the first factor was labelled 'Voice Appreciation'. Indeed, it has a strong evaluative character, as evidenced by high loadings of items like 'ugly↔beautiful' and 'pleasant↔unpleasant'. In the first place, however, this factor seems to be a factor of voice dynamics, of melodiousness in speaking, the perceived value of which largely defines speech evaluation. Perceived melodiousness also seems to relate strongly to certain paralinguistic features, as evidenced by high loadings of items like 'whining↔cheerful', 'friendly↔curt' and 'tense↔relaxed'.
The factor emerging next in the Read experiment also shows a close resemblance to the second factor resulting from the Retell experiment. The highest loading items have to do with preciseness of speech, intelligibility, non-standardness and might indeed be labelled 'Articulation Quality'. (Most of the original scales loading high on this factor have been modified for the Read experiment. This was done to avoid the occurrence of so-called Beta scales (Lemann & Solomon, 1952) that run from one negative extreme through a positive central area to another negative extreme.)
Perceived articulation quality seems to be the factor that determines the impression of self-assurance most, as can be concluded from the high loadings of the paralinguistic scale 'self-confident↔wavering'. Though it is clear that perceived reading ability is highly dependent on a good score for articulation quality, we certainly cannot neglect the impact of the first factor at this point. Intuitively, it is also very plausible that melodiousness of speech is important for a favourable impression of reading performance.
The third Retell factor has been labelled 'Abnormality'. This interpretation was probably brought about by the conspicuous

high score of one rather husky and fast talking speaker.
The results of the Read experiment suggest that this factor can be split in two, or maybe three, components: one Voice Quality factor, strongly associated with perceived clarity or brightness as well as with subjective strength, one Pitch factor and one Tempo factor. Though the last two factors are minor factors in terms of explained variance, they are neatly interpretable and may play an important role in our perceptual description when comparing subjective judgments with acoustic measures.
A lot of other things can be said referring to the factor structure we found, but we will end this part of the discussion by concluding that we seem to have established a reasonably stable frame of reference for the perception of voice and pronunciation with the help of 35 rating scales.

7.2 Reduction of variables.

In the next step of our analysis we tried to select a limited set of rating scales by means of which the dimensional structure of our data can be described as well, without much loss of information. This was primarily done with the intention to obtain, ultimately, a handy and efficient rating procedure for future use on more extensive samples of speakers.
There are a number of criteria that can be used for selecting scales. One of them is the interrater reliability. Therefore we calculated the so-called 'effective reliability' (Rosenthal, 1973) of each scale according to the formula

$$\frac{n\bar{r}}{1 + (n-1)\bar{r}}$$

in which $\bar{r}$ is the mean correlation between raters and n is the number of raters. As can be seen, the value of this coefficient is dependent on the number of raters involved.

Table 4 - Mean correlations between the 8 rater groups ($\bar{r}_8$) and communalities (h^2) per scale.

scale	$\bar{r}_8$	h^2
pleasant/unpleasant	.89	.72
dull/clear	.91	.67
friendly/curt	.88	.58
loud/soft	.83	.36
husky/not husky	.94	.57
stereotyped/varied	.89	.62
active/passive	.87	.65
artless/affected	.68	.27
colourless/sonorous	.90	.78
careless/precise	.93	.64
high/low for a (wo)man	.92	.46
firm/slack	.88	.70
creaky/not creaky	.91	.46
dragging/brisk	.81	.51
melodious/monotonous	.92	.80
powerful/weak	.86	.73
ugly/beautiful	.92	.83
jerking/smooth flowing	.89	.53
spiritless/vivacious	.91	.84
rounded/angular	.91	.48
deep/shrill	.95	.71
poor/rich	.95	.79
tense/relaxed	.91	.59
expressive/expressionless	.93	.82
steady/unsteady	.93	.66
broad/cultured	.92	.69
self-confident/wavering	.92	.54
whining/cheerful	.92	.75
polished/slovenly	.93	.74
deviating/normal	.89	.56
quick/slow	.88	.49
agitated/calm	.92	.53
distinct/indistinct	.91	.63
accentedness (high/low degree)	.87	.42
good/bad reading performance	.95	.65

For all scales an effective reliability of .90 or higher resulted when at least 25 raters were involved, except for the scales 'loud↔soft' and 'quick↔slow' that require about 30 raters to reach an effective reliability of .90.

We also checked the agreement between listener groups in their judgments. For each of the 8 groups mean values for the 11 speakers

were calculated and subsequently the rater groups were inter-correlated on these values. Mean correlations for each scale are shown in Table 4.
Also shown in this Table are the communalities of each item resulting from the factor analysis reported above. These values have also been used as a criterion for scale selection. (Variables with high communalities were preferred above variables with low communalities).
Finally, we analysed the judgments on each scale on the basis of Thurstone's Law of Categorical Judgment. Using the computational method described by Blom & Van Herpt (1976), we checked whether the 7 categories of our rating scales could be considered about equal in length. In other words, we verified whether it would be reasonable to assume the raw scores on these scales to be values on interval scales, and, therefore, to be insensitive to any linear transformation.
To compare the 35 scales on this criterion, Osgood's measure of interval equality was calculated (Osgood et. al., 1957, p. 152), as well as the Edwards-Thurstone measure for goodness-of-fit (Torgerson, 1958). Although these coefficients turned out to be very high for all scales (> .98), even small differences can provide a basis for selection.

7.3 Concise description of perceptual space

Using the above criteria, we made a choice of 12 rating scales. The correlations between the scores on these scales were factored again, using the same factoring method as before. This time, however, we forced the analysis to a 5 factor solution. (The definition of an eigenvalue ≥ 1.00 as a criterion of the number of factors to be extracted is rather arbitrary anyway).
The Varimax rotated solution of this analysis is shown in Table 5. This factor structure appears to be highly similar indeed to the structure we found with 35 variables. The five factors account for 66.1% of the total item variance. The proportions of this total

Table 5 - Varimax rotated factor solution for 12 selected scales.

scale	F-1	F-2	F-3	F-4	F-5
spiritless/vivacious	.84	.22	.21	.04	.18
expressive/expressionless	-.83	-.24	-.18	.04	-.14
ugly/beautiful	.59	.33	.42	-.29	-.04
polished/slovenly	-.31	-.80	-.25	.12	.02
broad/cultured	.23	.77	.09	-.19	.09
husky/not husky	.14	.12	.74	.03	-.06
dull/clear	.27	.11	.71	.26	.08
powerful/weak	-.43	-.31	-.45	.16	-.18
deep/shrill	-.19	-.18	.01	.83	.06
high/low for a (wo)man	-.07	.08	-.13	-.69	-.06
dragging/brisk	.28	.13	.08	.05	.68
agitated/calm	.09	.29	.28	-.33	-.42

explained variance accounted for by the first to fifth factors extracted are respectively 54.9%, 21.7%, 11.9%, 7.1% and 4.5%. Apparently, the proportion of variance explained by the first factor has decreased in favour of the explanatory power of the next factors extracted. This may be due to the great reduction of variables with a strong evaluative character. Indeed, a factor analysis on the same set of scales except 'ugly↔beautiful' and 'powerful↔weak' showed the expected factor structure in which the trend of a decreasing explanatory power of the first factor is even more visible. (The proportions were respectively 45.1%, 23.5%, 17.3%, 7.8% and 6.3% of the total explained variance, which took up 69.1% of the total variance).

Although we have not yet decided on the number of scales to be included in our definite rating form and some scales may still be replaced by others on the basis of new information, we will now take

a closer look at the factorial structure of the 12 rating scales shown in Table 5, and, more specifically, at the projections of our stimuli-speakers on the dimensions of factor space.

7.4 Speakers' positions in perceptual space.

The fact that one factor structure consistently emerges from our data does not necessarily imply that all speakers will have the same values on the co-ordinates involved in this system for all listener groups in all conditions.
However, the high correlations of mean scale values between the listener groups of the Read experiment (see Table 4) exclude great differences between these groups as to factor scores for the speakers. More interesting will be a comparison of the factor scores of the same speakers under the different speech conditions, the Retell and the Read condition. It must be emphasized, however, that differences between these speakers in perceptual space can be due not only to the changed speech conditions itself, but also to the differences in the rating form and to the fact that norms may have changed in the more than ten years lying between the two rating experiments.
Mean factor scores were calculated for our 11 speakers in the Read condition on the 5 factors shown in Table 5. These factor scores are shown in Table 6, together with the factor scores for the 10 speakers that have also been judged in the Retell condition. The Retell factor scores are based on the solution shown in Table 1.
Since speaker 11 had not been judged in the Retell condition, we could only intercorrelate the factor scores of ten speakers. The resulting correlation matrix is shown in Table 7.

Table 6 - Mean factor scores for speakers on Retell and Read factors. These factors are described in Table 1 and Table 5. Speaker 11 has been judged in de Read condition only.

Sp	Mean factor scores							
	factor 1		factor 2		factor 3		factor 4	factor 5
	RETELL	READ	RETELL	READ	RETELL	READ	READ	READ
1	-.529	-.067	.226	.367	.601	.474	.255	-.583
2	-.251	-.003	-.223	-.235	-1.864	-1.604	.190	.579
3	-.151	-.244	-1.377	-.583	.172	.016	-.511	-.356
4	1.169	.902	.537	.546	.488	.761	-.481	-.124
5	.187	.409	.915	.409	.374	.204	-.149	.525
6	.917	1.009	.431	.610	.019	.373	.281	.441
7	-.598	-.711	-.337	-.046	-.002	.180	.458	-.017
8	-.729	-.938	1.116	-.043	-.160	-.428	-.438	.252
9	.118	-.088	.068	.083	.405	.299	.818	-.152
10	-.135	-.203	-1.356	-1.506	-.034	.237	.910	.055
11	--	-.072	--	.397	--	-.512	-1.333	-.619

Table 7 - Correlations between Retell and Read factors, based on mean factor scores of 10 speakers.

		RETELL			READ				
		F-1	F-2	F-3	F-1	F-2	F-3	F-4	F-5
RETELL	F-1	---	.19	.24	.93***	.42	.44	-.14	.17
	F-2		---	-.15	.23	.80**	.07	-.35	.35
	F-3			---	.17	.27	.94***	-.06	-.60*
READ	F-1				---	.48	.36	-.10	.21
	F-2					---	.26	-.36	.09
	F-3						---	.06	-.53
	F-4							---	-.02
	F-5								---

significant at * $p \leq .05$, ** $p \leq .005$, *** $p \leq .001$

It is surprising to see the high significant correlations between the 3 most prominent factors of the two experiments, in spite of all differences existing between these studies.
Especially the voice dynamics and voice quality factors (F-1 and F-3) score high. The articulation quality factors (F-2) resulting from the two studies correlate somewhat lower. There are at least three plausible explanations for this:

1) Judgments on articulation quality may be most sensitive to norm changes. We found some evidence for this in our preliminary investigation (see also section 3), where Ideal Male Voice and Ideal Female Voice have been judged on 35 scales. As the same concepts had been rated 10 years before on the 46 Retell scales, mean judgments on corresponding scales could be compared. The greatest changes in norms were found on the articulation scales, notably on the scale 'broad↔cultured'. These changes in norms tended towards a more tolerant judgment about what should be considered 'ideal' on these scales (Van Herpt, 1980).

2) Most of the original rating scales for articulation quality have been altered from Beta scales to Alpha scales. This means that scales with two negative extremes were changed into scales with one negative extreme and one positive or at least neutral extreme (e.g. 'carelessly articulating↔hyper-correct' was changed into 'careless↔precise'). Therefore it is possible that the 'positive' pole of the second Retell factor is in fact slightly negative, while the second Read factor clearly has a positive pole.
This might explain the great difference for speaker 8 on this factor (a speaker that was rated as 'hyper-correct', 'la-di-da' and 'cultivated' in the Retell experiment).

3) Last, but certainly not least, the differences between the two speech conditions may be most noticeable in the articulation dimension. Besides the effects of differences in speech style and con-

tent, it is obvious upon hearing our material that some speakers make an extra effort in the Read situation to speak 'properly'. Apparently, the fact that these speakers don't have to concentrate on what they say in the Read situation gives them the opportunity to concentrate on how to say it. This probably explains the 'upward shift' of speaker 3 in the second dimension.

The perceptual space defined by the 3 most prominent factors of the Read experiment is visualized in Figure 2. It is interesting, of course, to compare this representation with that in Figure 1. Once more, however, we want to emphasize that differences in speakers' positions cannot be attributed to differences in speech conditions only.

8. CONCLUDING REMARKS

Now that we seem to have established a number of relevant perceptual parameters, we would like to know which acoustic characteristics are responsible for variations along those subjective dimensions and to which extent they are responsible.
Though we have started, by means of Multiple Regression Analysis, to assess the relations between our perceptual data and some acoustic data derived from long-time-average spectra, analyses of fundamental frequency distributions and measurements of 'acoustic system contrast' (ASC; see Koopmans-van Beinum, 1980), a number of things still have to be sorted out before such an investigation can be carried out thoroughly.
One of the most important questions to be answered is: to which extent is the perceptual structure we found dependent upon the specific sample of stimuli-speakers we used?
Once we have decided on a limited set of apparently reliable rating scales, by means of which the hypothesized perceptual structure can be described, it will be more easy to study the generality of this structure over a large sample of speakers.

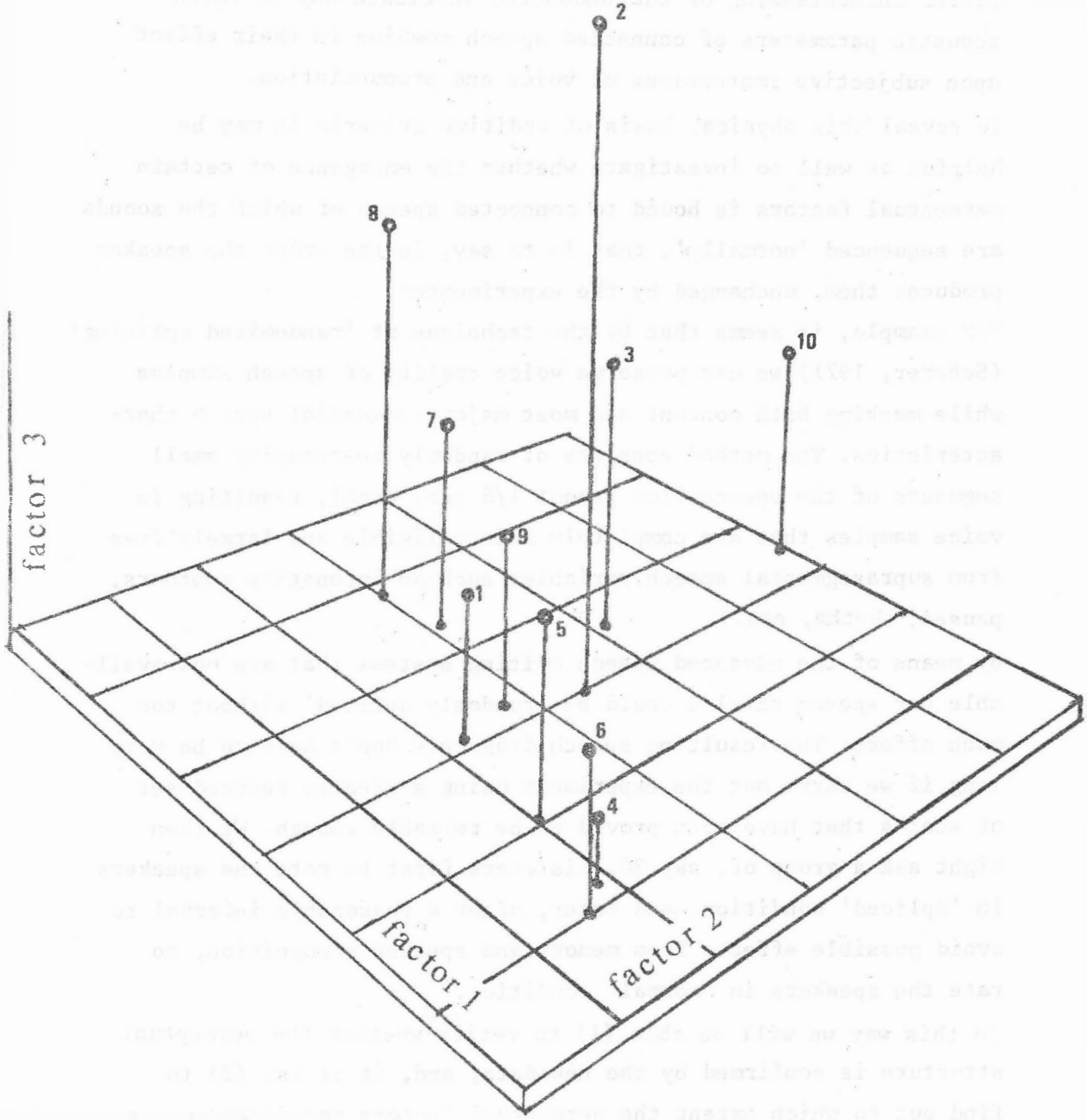

Fig. 2 - Positions of 10 speakers judged by 235 listeners (text read aloud) in the 3-dimensional factor space defined by the first 3 factors of Table 5.

If rating experiments on such a large sample of speakers are to confirm the existence of the judgmental criteria we found, then it will be expedient to obtain acoustic data from as many of these speakers as possible. A lot of observations — acoustic as well as perceptual — will be necessary for a better understanding of the undoubted intricate way in which acoustic parameters of connected speech combine in their effect upon subjective impressions of voice and pronunciation.

To reveal this physical basis of auditive criteria it may be helpful as well to investigate whether the emergence of certain perceptual factors is bound to connected speech of which the sounds are sequenced 'normally', that is to say, in the order the speaker produces them, unchanged by the experimenter.

For example, it seems that by the technique of 'randomized splicing' (Scherer, 1971) we can preserve voice quality of speech samples while masking both content and most major sequential speech characteristics. The method consists of randomly rearranging small segments of the speech flow (about 1/8 sec. each), resulting in voice samples that are completely unintelligible and largely free from suprasegmental speech variables such as intonation contours, pauses, rhythm, etc.).

By means of the advanced speech editing systems that are now available our speech samples could be 'randomly spliced' without too much effort. The resulting speech fragments don't have to be very long if we carry out the experiment using a greatly reduced set of scales that have been proved to be reliable enough. We then might ask a group of, say 30, listeners first to rate the speakers in 'spliced' condition, and later, after a reasonable interval to avoid possible effects from memory and speaker recognition, to rate the speakers in 'normal' condition.

In this way we will be able (1) to verify whether the perceptual structure is confirmed by the new data, and, if it is, (2) to find out to which extent the perceptual factors are dependent on sequential speech characteristics (intonation contours, pauses, rhythm, etc.).

Information on this dependency can be used for a more directed search of the relevant acoustic characteristics underlying subjective judgments on voice and pronunciation.

ACKNOWLEDGEMENTS

We would like to thank Heleen Deighton-van Witsen, Florina Koopmans-van Beinum and Loe Boves for reading an earlier version of this article and for their useful comments on it.

REFERENCES

Bismarck, G. von. (1974). Timbre of steady sounds: a factorial investigation of its verbal attributes. Acustica, 30, 146-159.

Blom, J.G. & van Herpt, L.W.A. (1976). The evaluation of jury judgments on pronunciation quality. Proceedings from the Institute of 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. Proceedings from the Institute of Phonetic Sciences, Univ. of Amsterdam, 3, 1-24.

Boves, L. (1981). On measuring and modelling subglottal pressure signals. Proceedings from the Institute of Phonetics, Univ. of Nijmegen, 5, 41-57.

Boves, L., Fagel, W.P.F. & van Herpt, L.W.A. (1982). Opvattingen van vrouwen en mannen over de spraak van mannen en vrouwen. De Nieuwe Taalgids, 75(1), 1-23.

Fagel, W.P.F. (1979). On a classification scheme for speech rating studies. Proceedings from the Institute for Phonetic Sciences, Univ. of Amsterdam, 5, 103-115.

Fagel, W.P.F. (1980). Literatuuronderzoek uitspraakbeoordeling. Rapport nr. 61, Instituut voor Fonetische Wetenschappen, Univ. van Amsterdam.

Fagel, W.P.F. (1981). De beoordeling van stem en uitspraak. Tussenrapport. Rapport nr. 71, Instituut voor Fonetische Wetenschappen, Univ. van Amsterdam.

Herpt, L.W.A. van. (1980). De beoordeling van de kwaliteit van de uitspraak van het Nederlands. Voortgangsrapport ZWO-projekt nr. 17-21-13.

Koopmans-van Beinum, F,J. (1980). Vowel Contrast Reduction. An acoustic and perceptual study of Dutch vowels in various speech conditions. Amsterdam: Academische Pers B.V.

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

Osgood, C.E., Suci, G.T. & Tannenbaum, P.H. (1957). The Measurement of Meaning. Urbana: University of Illinois Press.

Rosenthal, R. (1973). Estimating effective reliabilities in studies that employ judges' ratings. Journal of Clinical Psychology, 29, 342-345.

Scherer, K.R. (1971). Randomized splicing: a simple technique for masking speech content. Journal of Experimental Research in Personality, 5, 155-159.

Scherer, K.R. & Giles, H., eds. (1979). Social Markers in Speech. Cambridge: University Press, Paris: Editions de la Maison des Sciences de l'Homme.

Solomon, L.N. (1958). Semantic approach to the perception of complex sounds. Journal of the Acoust. Soc. of America, 30, 421-425.

Takahashi, H. & Koike, Y. (1975). Some perceptual dimensions and acoustical correlates of pathologic voices. Acta Oto-Laryngologica, Suppl. 338, 1-24.

Torgerson, W.S. (1958). Theory and Methods of Scaling. New York: Wiley.

Uldall, E. (1960). Attitudinal meanings conveyed by intonational contours. Language and Speech, 3, 223-234.

Vieregge, W.H. & Nuytens, F.Th.G. (1978). Perceptie en evaluatie van stemmen als functie van verschillende vormen van de glottispuls. Subsidie-aanvrage voor ZWO-projekt nr. 17-21-10.

Voiers, W.D. (1964). Perceptual bases of speaker identity. Journal of the Acoust. Soc. of America, 36, 1065-1073.

Voiers, W.D. (1976). Methods of predicting user acceptance of voice communication systems. Final Report, Dynastat Inc., Contr. no. DCA 100-74-C-0056, D-76-001-4.