That norm has not been discovered in
a laboratory and even less can it be
established there or ..... invented.
W.Gs. Hellinga
(on Standard Pronunciation of Dutch)

THE EVALUATION OF JURY JUDGMENTS ON PRONUNCIATION QUALITY

by Jan G. Blom and Leo W.A. van Herpt

When studying the verbal use of language the investigator is often obliged to take qualitative judgments as a starting-point for his evaluation of pronunciation and the way language and voice are used. When using a reasonably large panel it is possible to get fairly reliable quantitative data. In order to do this it is necessary to explore which judgment-criteria are relevant and in which way a representative group of listeners can be composed.
The Institute of Phonetic Sciences of the University of Amsterdam published (1973) the preliminary results of an investigation on the structure of opinions on the specifics of the pronunciation of Dutch (4). So as to determine the dimensionality of the semantic space of judgments on quality of speech production, we then used 7-point scales of successive categories and factor analyzed the scorings. Four independent groups of listeners were involved in the investigation, two of which will be the subject of our present discussion: 100 Dutch listeners from the west of the country and 100 Belgian listeners.
The speech material to be judged consisted of tape recordings of free renderings of a written story by 5 male and 5 female speakers. These 10 speakers were recruited from different social settings and different levels of education and were selected in such a way that they presented

presented a great variety in voice and pronunciation.
For each group 4 orthogonal common factors were found. These two sets of common factors are only partly identical. Factor stability therefore was poor, and consequently it was not possible to construct a test for "Standard Dutch".

More recently we have reconsidered the earlier work and approached it from a different angle.
In the absence of a well-defined circumscription of the pronouncing component of Standard Dutch we have introduced the concept of Relative Speech Appreciation, henceforth to be called RSA.
The RSA of a speaker is operationally defined as his rank within a group of speakers allocated by a jury by means of a global judgment of his speech production. (As a matter of fact the RSA of a person depends on the group in which he is judged. Groups can, however, be compared when they share some of the speakers; in scaling-techniques known as anchoring.)
From the literature on Dutch phonetics and pronunciation we collected some 800 terms referring to special attributes of speech. These terms are impressionistic by nature and most authors do not define their private use of these terms.
Out of this collection a selection was made of pairs of contrasting terms in order to form bipolar scales. Terms concerning linguistic properties as lexis and grammar, and also terms referring to clearly pathological conditions were avoided. In this way we got a list of 85 conceivable scales. Next we tried to avoid clear synonyms, and ended up with 46 scales - of which, at first sight, about 30 refer to features of pronunciation and 30 to features of voice, with an overlap between these two subsets.
A listing of all 46 scales is given in table I. (Because the terms are metaphoric, many of them cannot be translated literally).
The preliminary investigation mentioned above deals only with the subset pertaining to pronunciation. (See table I, column 3).

To get more insight into the meaning of the antonyms, we examined how the properties designated by the individual terms were valued. The results confirmed our suspicions about the usability of a number of scales.

There were a number of scales of which the separate terms were indeed regarded as opposites, which therefore covered a continuum from positive to negative.

Such as:	pleasant	-	unpleasant	(-1.45/1.52)
	krachtig	-	zwak	(-0.77/0.85)
	arm	-	rijk	(-1.00/1.00)

But some of the scales of which one scale term was considered neutral had an opposite term which was felt to be positive or negative, e.g.

cultivated	-	slip-shod	(-0.09/1.26)
velar-r	-	rolling-r	(0.11/0.58)
high	-	low (for ?)	(-0.03/0.47)

Besides there were a number of scales of which both terms got a negative judgement, e.g.

la-di-da	-	vulgar	(1.55/1.52)
humble	-	supercillious	(1.17/1.26)
hard	-	mawkish	(0.91/1.11)

Such combinations can give rise to problems, because listeners may interpret these scales in different ways.

Therefore we decided to make a more careful analysis of our 46 scales. In this analysis we used the data of the Dutch listeners. The first analysis was done on the basis of the law of Categorical Judgement (7). The experimental procedure for this method of successive intervals requires a number of stimuli to be sorted into a number of semantic categories on some attribute continuum. This procedure yields a frequency distribution for each stimulus over several of the categories. The basic consideration in successive interval scaling is whether or not these frequency distributions can be simultaneously converted to a common distribution, allowing unequal means and variances on the same linear continuum.

The means of the converted distributions correspond to the scale values of the stimuli, and the standard deviations to the discriminal dispersions.

Scale values for category boundaries are also obtained by the method of succesive intervals, thus permitting estimates of the sizes of the categories rather than assuming them to be equal. In this way ordinal scales are converted into interval scales (not all our scales proved to be ordinal). Interval scales have no natural origin. For practical reasons we shifted the scale value for the least valued category to the origin. Moreover, we normalized all scales in such a way that the scale value of the most valued category becomes unit. This normalization has no implications for our analyses, as correlations between variables are invariant against shift of origin and change of scale.

We programmed (1,2) the computional method as suggested by Diederich, Messick and Tucker (5) and supplemented it by a procedure to get least square estimated of the category values (3).

Thus we calculated scale values and dispersions for our 10 speakers, scale values for the categories and for the boudaries between the categories for the 46 attribute continua.

This is exemplified in the diagrams of figure 1, the first of which gives an illustration of scale number 1 (pleasant - unpleasant).

This scale satisfies all requirements of a good scale of successive categories:

1. category values are monotonously related to the rank of the categories,
2. the scale has a good discriminative power, as can be concluded from the range of the scale values of the speakers, and
3. the values for each speaker given by the group of listeners are normally distributed on the continuum.

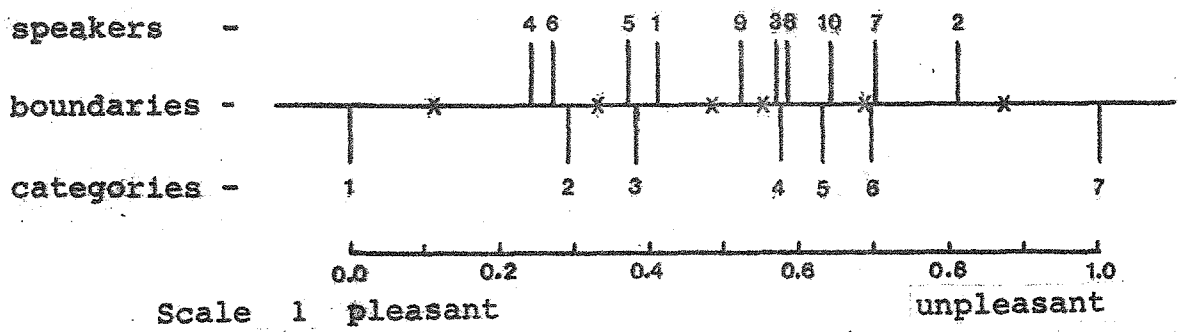

Fig. 1 - Speakers, categories and boudaries on a monotonous scale.

Some scales, however, show a considerable departure from this ideal situation. The most serious deviation is: non-monotonicity, which means that the scale is not ordinal but nominal, in other words the categories are non-successive. (Among others, scales number 16 and 35 show this deficiency; see figure 2).

Non-monotonicity can be caused by the fact that the attribute is not one-dimensional, as may be the case in scale 35 (high - low) where besides the intended dimension 'pitch', a second dimension related to 'sex' seems to exist. Another possible explanation may be ambiguity of terms, which we suppose to occur in scale 16 (controlled - temperamental), where the left side is related to 'controlled use of language' and the right side (temperamental) to 'uncontrolled use of voice'.

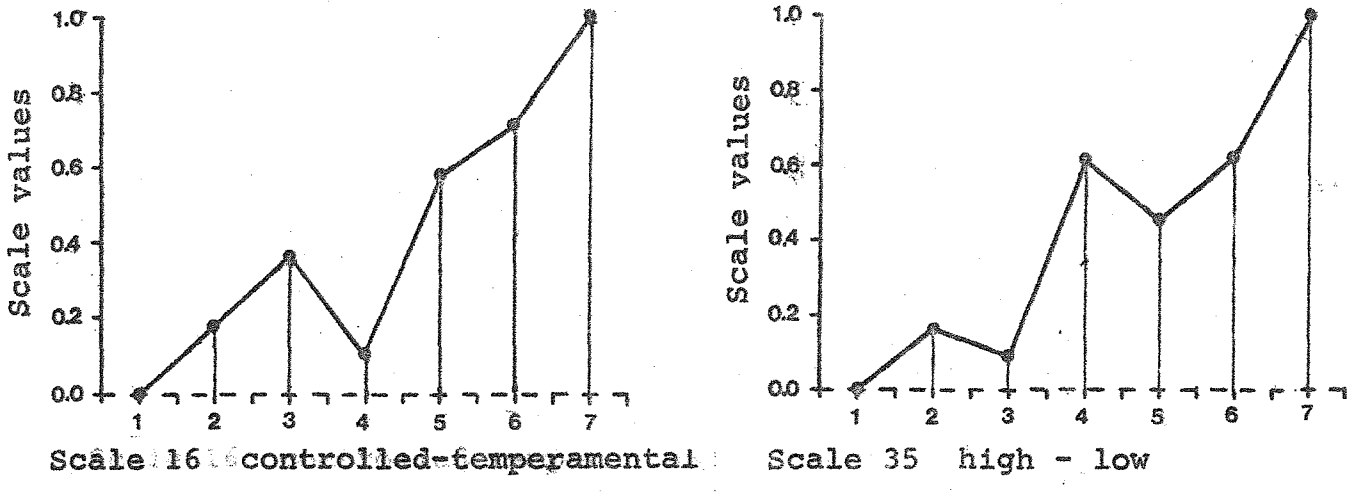

Fig. 2 - Scale values as function of rank numbers of categories.

A second serious deficiency is lack of discriminative power. This is found e.g. in scale 24 (rustic - urban) where the listeners do not agree on the attribute, and in scale 13 - figure 3 (biting - caressing) where the listeners seem incapable of giving any meaning at all to the attribute in relation to normal speech. When this occurs the observed distributions of values on the continuum are error distributions; as is reflected by the deviation from normality. (Although normality is postulated by the scale model we did not use depature form normality as a sole criterion for the rejection of a scale, because the technique seems to be sufficiently robust).

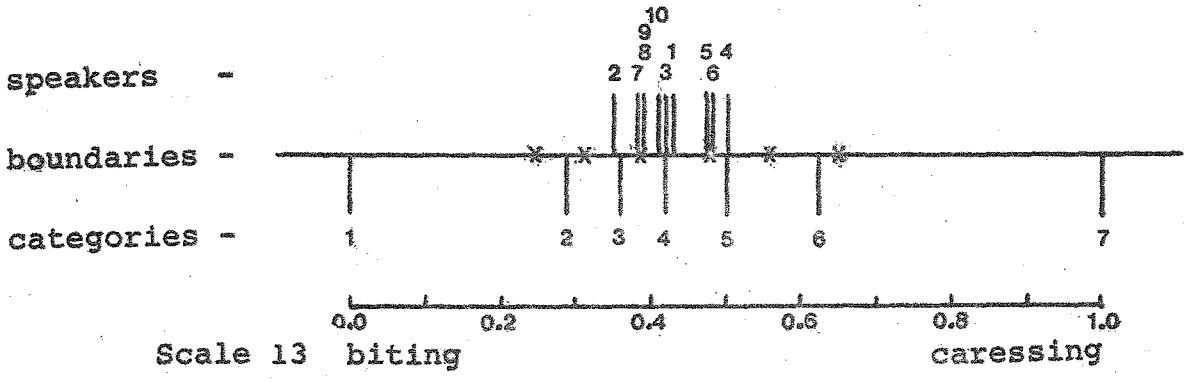

Fig. 3 - Speakers, categories and boundaries on a non discriminating scale.

These considerations resulted in the elimination of 16 scales.

The scales left for further analysis are shown in table I, column 4. The remaining scales are monotonous, but the mapping of the categories on the continuum is often non-linear, which manifests itself in the fact that some categories are highly stretched, while others are highly compressed, especially in some cases where the central category has no semantic value. This leads to the question whether it is allowed to use raw scores on scales of this type in a linear analysis like factor analysis.

From our Dutch listener group we have a collection of 100 observations of 30 variables (scales) of each of our 10 speakers. It would be highly unrealistic to assume that these 30 variables are mutually independent and therefore we tried to reduce our 30,000 data.
An obvious approach is factor analysis (6,8). First, therefore, a brief outline of factor analysis for those who are not familiar with this technique.

The aim of factor analysis is to explain observed relations among numerous variables in terms of simpler relations. This reduction takes the form of creating a smaller set of hypothetical variables, called factors. To find out "what goes with what" among the variables, the variables can be intercorrelated as they vary over the observations. Obviously variables which are highly correlated have much in common, variables which are low correlated are almost independent. The process of factor analysis is designed to replace the intercorrelation matrix by a factor matrix, in which the number of factors is considerably smaller than the number of variables. These factors may be considered as underlying determinants, which can be substituted for the more numerous original variables, and which largely account for the correlations among these variables. Because the factors describe what the original variables have in common, we speak of common factors.
Each variable can be decomposed in a part which it has in common with the other variables and a part which is unique - which means that it is partly specific for that variable and partly composed of error (random fluctuations). The variance accounted for by the common factors is called communality. Only those variables which have a high communality play an important role in a factor model.
It is possible to replace an observation of many variables by a small number of factor scores equal to the number of factors decided upon. This means that an original observation which can be mapped as a point in a multi-dimensional observation space is represented as a point in a less dimensional factor space. The relation between different observations is indicated by their relative position in the factor

space. This relative position is in fact independent of the way in which the factor-axes are chosen. It is common use to rotate the axes in such a way that their meaning becomes interpretable in ordinary language. This process is called rotation to simple structure.

Now we come back to our scales, we first applied factor analysis to the data of the group of male and the group of female speakers separately. Instead of the raw data values obtained in the analysis of the scales were used. Three common factors were extracted and rotated to simple structure. The factors found for the male group proved to be essentially the same as those found for the female group, which implies a high degree of factor stability. These factors can be caracterized as: "voise appreciation", "articulation quality" and "abnormality". Scale 15 (monophthongized - not-monophthongized) proved completely unique, and scales 22 (thick - thin) and 23 (feminine - masculine) behaved totally differently for the two sexes. So we decided to eliminate these three scales (table I, column 5).
We proceeded with a factor analysis on the whole group of speakers. After rotation to simple structure the same three factors were found. This was to be expected because of the high factor stability between the group of male and the group of female speakers.
Factor scores were calculated. Mean factor scores for the speakers are illustrated in figure 4. The differences of opinion about the speakers must be reflected by the distances. This led us to the hypothesis that the RSA is related to the distances between factor score centroids. This proved to be the case.
Taking all 45 distances between the 10 speakers into account, we derived a least square solution to their ranking order, which is 0.95 (Spearman) correlated with the ranking of the speakers obtained from the global judgements by a jury (the operational definition of RSA).

All rank correlations are given in table II.
As one factor alone accounted for some 70% of the total explained variance, we decided to try a one-factor solution. And, although 13

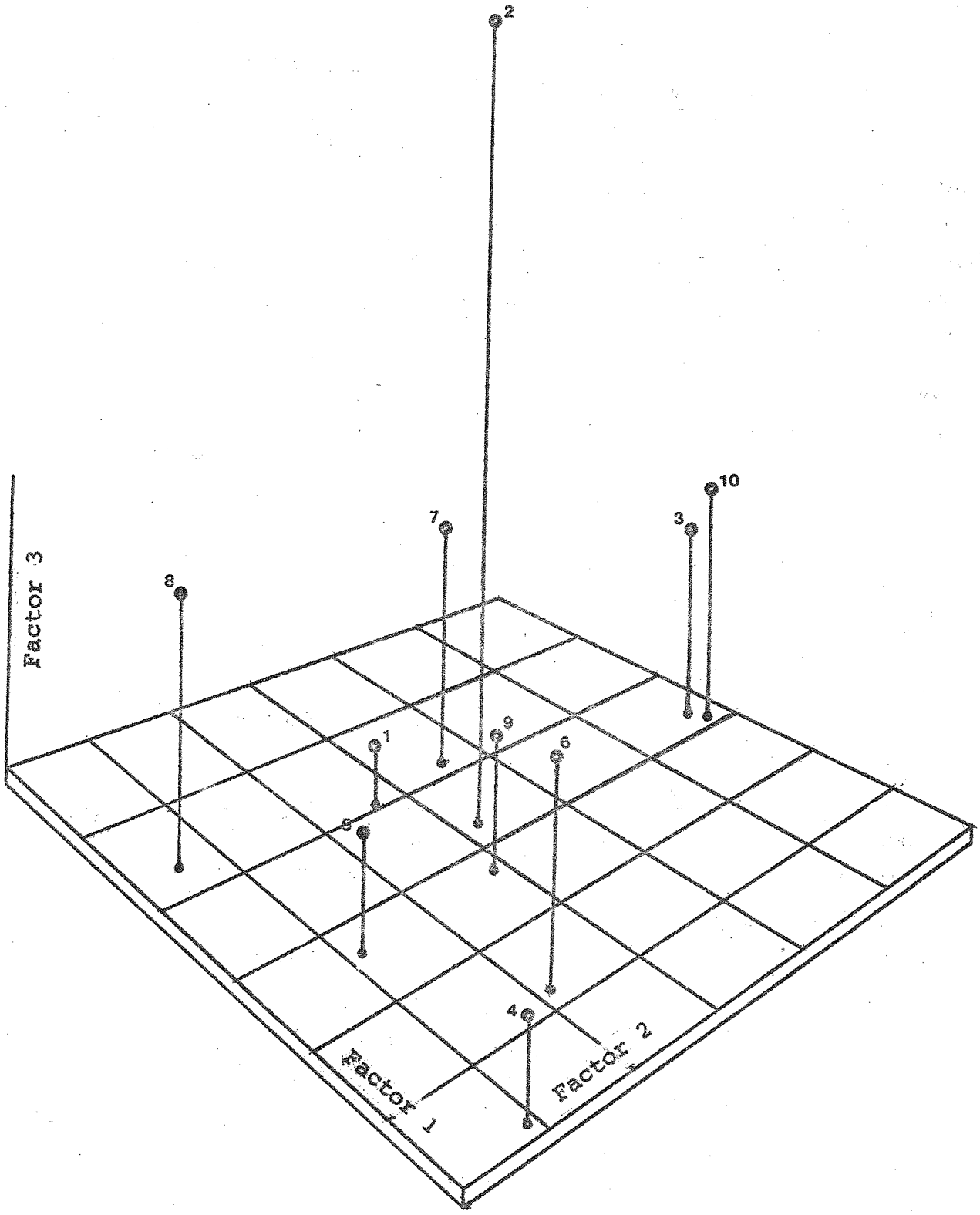

Fig. 4 - 10 Speakers in a 3-dimensional factor space.

variables had low communalities, the rank correlation between mean factor scores and RSA is still 0.94. This finding inspired us to eliminate another 13 variables, retaining those scales, only 14 in number, which have high communalities and, consequently, high loadings. (These remaining scales are shown in table I, column 6). The mean factor scores obtained from the one-factor model of these 14 variables are essentially the same as those obtained with 27 variables. They have exactly the same ranking and their product-moment correlation is 0.998, making it clear that 14 scales are sufficient to predict jury judgements (product-moment correlations are given in table III). The influence of individual attributes on the RSA, as estimated by 14 scales, can be deduced from the profiles for the 10 speakers (see figure 5). These profiles are in agreement with the verbal impressionistic descriptions of the speakers.

Having eliminated all error-producing scales, we now have a set of 14 relevant variables. All these 14 scales are monotonous, so the correlation between the scale values of categories and the rank-numbers of categories must be high. So it was tempting to compare the findings of the last analysis with an analysis on the raw data of the 14 relevant scales.
So we did and the results were these: rankings obtained from raw data are exactly equal to those from scale values. The correlation between the two sets of mean factor scores on which the rankings are based, is 0.999. The implication is that the conversion of category numbers to scale values can be omitted when using appropriate scales. These last two analyses revealed that all 14 scales are about equally important, so we asked ourselves the question whether the unweighted sums of raw scores on the 14 scales can be used as predictors for RSA. The ranking obtained from sums of raw scores proved to be 0.927 (Spearman) correlated with RSA, so the answer to the question is yes.

From our 14 scale tests we have derived three estimates for RSA:

mean factor scores of scale values
mean factor scores of raw data, and
unweighted sums of raw data.

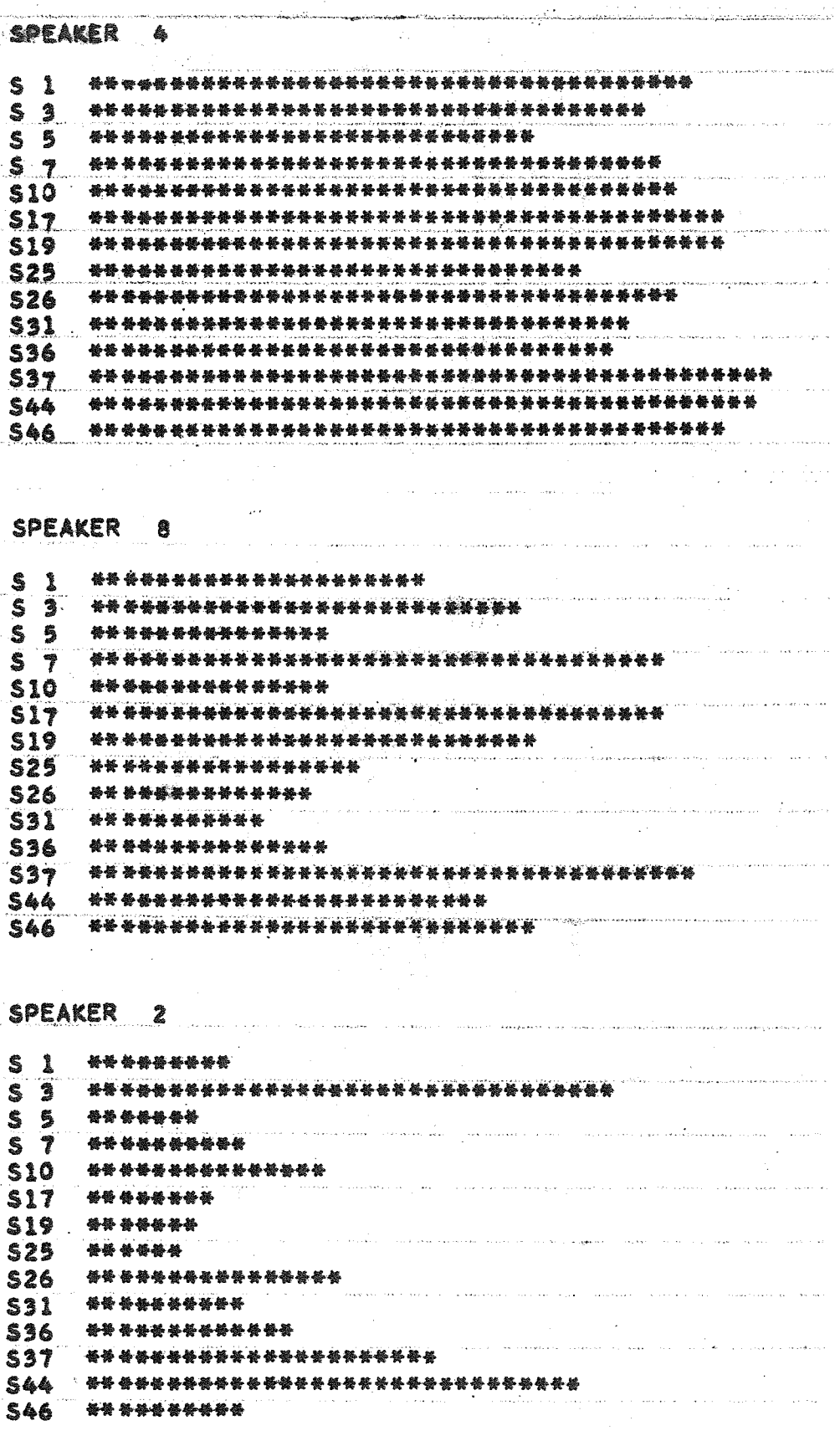

Fig. 5 - Speech appreciation profiles.

These estimates are equivalent, as their product-moment correlations all exceed 0.997 (see table III). The first two estimates can only be calculated with the aid of a computer, the third can be obtained by straightforward handscoring. So, for practical applications the third is the handiest.

Now having constructed a simple test with high concurrent validity, a crucial question is whether the composition of the listener group will bias the results.
To investigate this, we compared the scores by our Dutch listener group with scores of a listener group recruited from Dutch speaking Belgians - judging the same group of speakers. For various reasons it was legitimate to suppose that their scorings might considerably differ from the scorings of the Dutch listeners. This, however, proved not to be the case, as can be deduced from tables II and III in which all rank correlations and the correlations between the estimates are shown together. The concordance as to RSA between three independent listener groups - the jury, the Dutch group and the Belgian group - is very good. Kendall's coefficient of concordance W = 0.957, which is significant at a 0.5% level.

To sum up: as we mentioned at the beginning we found in the literature some 800 terms used for the differentiation of speech and pronunciation. Our work has made it clear, in our opinion, that comparisons made by non-phoneticians are based on a very small set of strongly related attributes.

BIBLIOGRAPHY

1. Blom, J.G. and L.W.A. van Herpt - Schaaltechnieken II,
Schaaltechniek voor Categorische Data,
Programma voor IBM 1130, version 1, level 1.
Amsterdam, 1974, IFA-publication no. 43.1.

2. Blom, J.G. and L.W.A. van Herpt - Schaaltechnieken II,
Schaaltechniek voor Categorische Data,
Listings and flowcharts, version 1, level 1.
Amsterdam, 1974, IFA-publication no. 44.1.

3. Blom, J.G. and L.W.A. van Herpt - A least square solution for category values.
IFA-publication in preparation.

4. Blom, J.G. and F.J. Koopmans-van Beinum - An investigation concerning the judgement criteria for the pronunciation of Dutch I.
Proceedings 3, Institute of Phonetic Sciences,
Amsterdam, 1973.

5. Diederich, G.W., Messick, S.J. and Tucker, L.R. - A general least square solution for successive intervals.
Psychometrika, vol. 22, no. 2, June 1957.

6. Harman, H. - Modern Factoranalysis.
The Univeristy of Chicago Press, Chicago & London,
1967, second rev. edition.

7. Torgerson, W.S. - Theory and Methods of Scaling.
New York, Wiley, 1967.

8. Statistical System (1130-CA-06X)
Publication H20-0341, International Business Machines Corporation, 1967.

TABLE I - SCALES

1 SCALE NUMBER	2 SCALE TERMS		3 ORIG. PRON. SCALES	4 30 SC	5 27 SC	6 14 SC
1	PLEASANT (AANGENAAM)	- UNPLEASANT (ONAANGENAAM)	*	*	*	*
2	PINCHED (GEKNEPEN)	- FULL (VOL)	*			
3	VIGOROUS (KRACHTIG)	- WEAK (ZWAK)	*	*	*	*
4	CONTEMPORARY (HEDENDAAGS)	- OLD-FASHIONED (OUDERWETS)	*	*	*	
5	POOR (ARM)	- RICH (RIJK)		*	*	*
6	SOUND (GAAF)	- MUTILATED (GESCHONDEN)				
7	CARELESSLY ARTIC. (SLORDIG GEART.)	- HYPER-CORRECT (HYPERCORRECT)	*	*	*	*
8	NORTHERN (NOORDELIJK)	- SOUTHERN (ZUIDELIJK)	*			
9	DISTINGUISHED (GEDISTINGEERD)	- COMMON (VOLKS)	*			
10	EXPRESSIVE (EXPRESSIEF)	- EXPRESSIONLESS (UITDRUKKINGSLOOS)	*	*	*	*
11	QUICK (SNEL)	- SLOW (LANGZAAM)	*	*	*	
12	LA-DI-DA (BEKAKT)	- VULGAR (ORDINAIR)	*	*	*	
13	BITING (BIJTEND)	- CARESSING (STRELEND)				
14	POMPOUS (GEWICHTIG)	- PLAYFUL (SPEELS)	*	*	*	
15	MONOPHTHONGIZED (GEMONOFTONGEERD)	- NOT MONOPHTONGIZED (NIET GEMONOFTONG.)		*		
16	CONTROLLED (BEHEERST)	- TEMPERAMENTAL (ONBEHEERST)	*			
17	CULTIVATED (GECULTIVEERD)	- SLIPSHOD (ONVERZORGD)	*	*	*	*
18	SPRIGHTLY (KWIEK)	- WHINING (ZEURIG)	*	*	*	
19	DEVIATING (AFWIJKEND)	- NORMAL (NORMAAL)	*	*	*	*
20	HUMBLE (ONDERDANIG)	- SUPERCILIOUS (UIT DE HOOGTE)	*			
21	EASTERN (OOSTELIJK)	- WESTERN (WESTELIJK)	*			
22	THICK (DIK)	- THIN (DUN)		*		
23	FEMININE (VROUWELIJK)	- MASCULINE (MANNELIJK)		*		

TABLE I (CONTINUED) - SCALES

1 SCALE NUMBER	2 SCALE TERMS		3 ORIG. PRON. SCALES	4 30 SC	5 27 SC	6 14 SC
24	RUSTIC (BOERS)	- URBAN (STADS)	*			
25	NASTY (LELIJK)	- BEAUTIFUL (MOOI)	*	*	*	*
26	STEREOTYPED (STEREOTIEP)	- VARIED (GEVARIEERD)	*	*	*	*
27	PEDESTRIAN (BANAAL)	- SOLEMN (PLECHTIG)				
28	VELAR-R (BROUW-R)	- ROLLING-R (ROLLENDE-R)	*			
29	DIPHTHONGIZED (GEDIFTONGEERD)	- NOT-DIPHTHONGIZED (NIET-GEDIFTONG.)		*	*	
30	DULL (DOF)	- CLEAR (HELDER)		*	*	
31	MELODIOUS (MELODIEUS)	- MONOTONOUS (EENTONIG)	*	*	*	*
32	ARTLESS (ONGEKUNSTELD)	- AFFECTED (GEAFFECTEERD)	*	*	*	
33	SMOOTH FLOWING (VLOEIEND)	- STACCATO (STACCATO)	*	*	*	
34	HARD (HARD)	- MAWKISH (WEEK)				
35	HIGH (HOOG)	- LOW (LAAG)				
36	COLOURLESS (FLETS)	- SONOROUS (KLANKRIJK)		*	*	*
37	BROAD (PLAT)	- CULTURED (BESCHAAFD)	*	*	*	*
38	VEILED (OMFLOERST)	- SHRILL (SCHEL)				
39	SPELLING PRONUNC. (SPELLINGSUITSPR.)	- NATURAL PRONUNC. (NAT. UITSPRAAK)	*			
40	WARM (WARM)	- COLD (KOUD)		*	*	
41	NASAL (NASAAL)	- NON-NASAL (NIET NASAAL)	*	*	*	
42	DRAWN OUT (GEREKT)	- CLIPPED (VERKORT)	*	*	*	
43	CLEAR (HELDER)	- HUSKY (HEES)		*	*	
44	SPIRITLESS (DOODS)	- VIVACIOUS (LEVENDIG)	*	*	*	*
45	GRATING (KRAKERIG)	- SMOOTH (GLAD)				
46	STEADY (VAST)	- UNSTEADY (ONVAST)		*	*	*

TABLE II - CORRELATIONS BETWEEN RANKINGS (SPEARMAN)

		1	2	3	4	5	6	7	8
1.	RSA **) (JURY)	-	.952 *)	.939	.927	.927	.927	.939	.939
2.	MFS - SV - NETH. 27 SCALES - 3 FACT	.952	-	.939	.952	.952	.952	.915	.915
3.	MFS - SV - NETH. 27 SCALES - 1 FACT.	.939	.939	-	.987	.987	.987	.963	.963
4.	MFS -SV - NETH. 14 SCALES - 1 FACT.	.927	.952	.987	-	1.000	1.000	.939	.939
5.	MFS - RD - NETH. 14 SCALES - 1 FACT.	.927	.952	.987	1.000	-	1.000	.939	.939
6.	SUM - RD - NETH. 14 SCALES	.927	.952	.987	1.000	1.000	-	.939	.939
7.	MFS -RD - BELG. 14 SCALES - 1 FACT.	.939	.915	.963	.939	.939	.939	-	1.000
8.	SUM - RD - BELG. 14 SCALES	.939	.915	.963	.939	.939	1.000	.939	-

TABLE III - CORRELATIONS BETWEEN THE THE TEST VALUES USED FOR RANKING (PRODUCT-MOMENT)

		1	2	3	4	5	6
1.	MFS - SV - NETH. **) 27 SCALES - 1 FACT.	-	.998 *)	.998	.998	.964	.967
2.	MFS - SV - NETH. 14 SCALES - 1 FACT.	.998	-	.999	.997	.966	.969
3.	MFS - RD - NETH. 14 SCALES - 1 FACT.	.998	.999	-	.999	.965	.968
4.	SUM - RD - NETH. 14 SCALES	.998	.997	.999	-	.959	.963
5.	MFS - RD - BELG. 14 SCALES - 1 FACT.	.964	.966	.965	.959	-	.999
6.	SUM - RD - BELG. 14 SCALES	.967	.969	.968	.963	.999	-

*) $P(R > 0.9) < 0.001$

**) RSA = RELATIVE SPEECH APPRECIATION
MFS = MEAN FACTOR SCORES
SUM = UNWEIGHTED SUM
SV = SCALE VALUES
RD = RAW DATA