3003

L BOVES ■ THE PHONETIC BASIS OF PER-
CEPTUAL RATINGS OF RUNNING SPEECH

**THE PHONETIC BASIS OF PERCEPTUAL RATINGS
OF RUNNING SPEECH**

The phonetic basis of perceptual ratings of running speech

Proefschrift

ter verkrijging van de graad van
doctor in de letteren
aan de Katholieke Universiteit te Nijmegen,
op gezag van de Rector Magnificus
Prof. Dr. J.H.G.I. Giesbers
volgens besluit van het College van Dekanen
in het openbaar te verdedigen
op 14 juni 1984 des namiddags te 3.00 uur precies

door

LODEWIJK WILLEM JOHAN BOVES
geboren te Maastricht

FORIS PUBLICATIONS
Dordrecht - Holland/Cinnaminson - U.S.A.

Promotores: Prof. Dr. W.H. Vieregge
Prof. Dr. E.Th.G. Nuijens

*To the people of the Netherlands
and those abroad who have contributed to our prosperity*

Contents

Preface	XI
1 Research into non-verbal aspects of speech	1
1.0 Introduction: form and content	1
1.1 Previous research into non-verbal aspects of speech	2
1.2 Methodological problems in the description of vocal aspects of speech	3
1.3 Vocal aspects of speech in modern linguistics	4
1.4 Aims of the present research	6
1.5 Operationalization of the research goals	7
1.6 Flavours: methodology versus content	9
2 The construction of a scaling instrument	11
2.0 Introduction	11
2.1 Selection of the objects to be judged	12
2.2 The administration of a semantic differential	14
2.3 The selection of the scales	15
2.3.1 Factorial composition and relevance of the scales	16
2.3.2 Stability of meaning and linearity of scales	16
2.3.3 Choosing from inventories of scales	17
2.3.4 The status of the rating scales	21
2.4 Testing the scales for linearity and relevance	22
2.5 Conclusion	27
3 The evaluation experiment	29
3.0 Introduction	29
3.1 Method	29
3.1.1 The preparation of the stimuli	29
3.1.2 The collection of the ratings	32
3.2 Analysis of the individual scales	33
3.2.1 Introduction of the scaling model	34
3.2.2 Results	38
3.2.2.1 The linearity of the scales	38
3.2.2.2 The goodness-of-fit of the model	40
3.2.3 The reliability of the ratings	41

3.2.4	Conclusions from the scale analysis	44
3.3	The effect of the texts on the ratings	45
3.3.1	Conclusions	48
3.4	Factor analysis	49
3.4.1	Introductory remarks	49
3.4.2	The dimensionality of the judgement space	51
3.4.2.1	Discussion	53
3.4.3	The construction of an efficient scaling instrument	54
3.4.4	The position of the talkers in the judgement space	55
3.5	Conclusions	58
4	Separation of the contributions of phonation and articulation to the radiated speech signal	61
4.0	Introduction	61
4.1	Registrations of subglottal pressure	68
4.1.1	Techniques for recording subglottal pressure	71
4.2	Making the simultaneous registrations	73
4.3	Establishing criterion instants based on physiological registrations	78
4.3.1	The photoglottogram and the moment of the maximum glottal flow	81
4.3.2	The P_{sg} recordings and the moment of glottal opening	83
4.3.3	Conclusions	86
4.4	Inverse filtering techniques	88
4.5	Inverse filtering of vowels from running speech produced by several speakers	90
4.5.1	Speech material	90
4.5.2	Processing of the inverse filtering results	91
4.5.3	Results	94
4.5.3.1	Speaker characteristics	94
4.5.3.2	Stability of results upon repeated measurements	98
4.5.3.3	Glottal pulse differences between stressed and unstressed syllables	98
4.5.3.4	The influence of supraglottal load	99
4.5.4	Discussion	102
4.6	Conclusions	105
5	Development of acoustic measures to explain perceptual ratings	107
5.0	Introduction	107
5.1	Acoustic measures of tempo	108
5.2	Acoustic measures of pitch	111
5.2.1	Measures of central tendency	112
5.2.2	Measures of variability	113
5.2.3	Pitch perturbation measures	113
5.2.4	The voiced-unvoiced ratio	115

5.2.5	The pitch data	115
5.3	Long-term average spectra	117
5.3.1	The generation of LTAS	118
5.3.2	The representation of LTAS	119
5.3.3	The influence of fundamental frequency on LTAS	121
5.3.4	LTAS as measures to characterize speakers	122
5.3.5	LTAS as representation of the glottal flow spectrum	124
5.4	Measures based on formant analyses	128
5.5	Summary and conclusions	134
6	Relations between acoustic measures and perceptual ratings	135
6.0	Introduction	135
6.1	Simple correlations	136
6.1.1	Voice dynamics	136
6.1.2	Articulation quality	137
6.1.3	The paralinguistic factor in the Nijmegen experiment	141
6.1.4	Potency	143
6.1.5	Voice quality	144
6.1.6	Tempo	148
6.1.7	Discussion and conclusion	149
6.2	Multiple regression analysis	150
6.2.1	Voice dynamics	154
6.2.2	Articulation appreciation	155
6.2.3	The sociolinguistic factor	155
6.2.4	The potency factor	156
6.2.5	Voice quality	156
6.2.6	Tempo	157
6.3	General discussion: method and results	157
6.4	Correlations versus experimental research	160
6.5	Summary and conclusions	162
7	The status quo and perspectives	165
7.0	Introduction	165
7.1	The state-of-the-art in acoustic descriptions of speech signals	167
7.2	The state-of-the-art in perceptual descriptions of running speech	169
7.3	The relations between acoustic and perceptual descriptions of running speech	171
Appendix I Development of an accurate closed glottis interval technique for inverse filtering		173
A1.0	Introduction	173
A1.1	The closed glottis interval approach to inverse filtering	173
A1.2	Determination of the CGI from the speech	174

X

A1.3	Estimating the CGI from the EGG	177
A1.4	The implementation of a closed glottis inverse filtering technique using the EGG	178
A1.5	Accuracy of formant estimates obtained from the CGI	180
References		183
Samenvatting		189
Curriculum vitae		197

Preface

This book is a dissertation in phonetics, an interdisciplinary science which draws from linguistics, physics, engineering, sociology, and psychology, to name just the most obvious disciplines. From the Table of Contents of this book one can see that most, if not all, of the sciences just mentioned have played a role in the work reported here. From this the reader may draw the conclusion that the work cannot have been done by a single person, operating in isolation. This is why the first person plural is used throughout, instead of the singular which might seem to be more appropriate since there is only one author.

Nothing would be more natural, then than to expect an extensive list of persons whose general or specific, but in any case generous contributions to this work have to be acknowledged. However, for reasons nobody has been able to explain, the promotion regulation of this University forbids to acknowledge the contributions of faculty members. Expressing gratitude to other people, apparently including those who are on the faculty of other universities, is allowed. Since I have received help from many people from all over the country this puts me in the awkward situation that it is permitted to name some, but not others, perhaps the ones whose contributions have been greatest. Therefore, I decided to refrain from naming anybody. Careful reading of the text and some study of the list of references will reveal the names of those who have provided most support.

Yet I would like to name one person, not to thank him, for he was not of very great help; on the contrary, the fact that Jan Blom died when the research reported here had only just begun has aggravated my task considerably. If he had lived to see the results, chapter 3 of the present book, and perhaps also some other chapters, would have looked very different. Much of chapter 3 pertains to work he intended to take care of.

Besides people there are also organisations and institutions which contribute to the work of individuals. My work would not have been possible if the money that freed me during three years from all my teaching and administrative duties had not been made available. Thus I should like to acknowledge at least one contribution, using the standard phrase: This research was supported by the Foundation for Linguistic Research, which is funded by the Netherlands Organisation for the advancement of pure research, Z.W.O.

Research into non-verbal aspects of speech

1.0 INTRODUCTION FORM AND CONTENT

Speech is often called man's most natural and efficient mode of communication. Whether or not this is true, it can hardly be denied that it is a very treacherous mode, since the speech signal contains a lot of information which is not so much related to the linguistic content of the verbal message but rather to the identity and physical and psychological status of the speaker. For the greater part this non-verbal (para- or extralinguistic) information is conveyed unintentionally and more often than not the speaker is absolutely unaware of its presence and its effects. Also listeners do not normally pay explicit attention to this kind of information which certainly is not to say that they do not normally perceive it and use it to their advantage (or to their detriment). Non-verbal information may be used to the advantage of the listener if he or she recognizes the speaking person correctly or infers the speaker's state and intention correctly from how he or she sounds. Non-verbal information is used to the listener's detriment if the speaker succeeds in making a false impression of friendliness, honesty or trustworthiness.

That the impact of a message does not only depend on its content but also very much on the form in which it is presented has been known for more than 2000 years and most probably much longer. At least since Greek antiquity the art or science of rhetoric exists and has been practiced and studied extensively by many people who aim at making a public career, an undertaking which is aided very much if as many people as possible get a favourable impression of the aspiring person. Training and study of rhetoric has resulted in an extensive list of speech characteristics which are thought to have specific effects on the listeners. A scan of the literature reveals, however, that the rhetoric tradition has concentrated its efforts on the argumentative structuring of speeches. This state of affairs reflects the close relationship of rhetoric and logic. The delivery aspect of a speech has not been neglected by the rhetoric tradition, but here the precepts and advice to prospective career-makers are nowhere nearly as precise and specific as in the domain of argumentation theory. Thus it seems honest to say that classical rhetoric was born a science as far as the linguistic, logic argumentation aspect is concerned and born an art with respect to the delivery aspect.

It should be noted that at least since Aristotle it has been realized that apart from the content of a speech and its structuring, and the way in which a speech is delivered, there is a third aspect which influences the impact of a rendition. This latter aspect is what was called the *ethos* of the speaker by Aristotle. The most general translation of this term is the credibility and honesty of the speaker as the source of the message in question. The *ethos* aspect is, of course, a very complicated one if only because it depends heavily on the situation. In dealing with people one has known and seen behaving for quite some time, prior experience may be much more important in estimating the trustworthiness of a message than the content of the message itself and the way in which it is conveyed. In initial interactions with strangers non-verbal aspects of communication may become extremely important. Thus it is clear that in our present-day mass-communication society a speaker's (perceived) *ethos* is not on the same level as the argumentative structure of his message and its style of delivery, but that the former tends to be one of the products of the latter two. It may safely be assumed that the rhetoric tradition has had a considerable influence on the formation of stereotypical ideas about the relation between the formal characteristics of a message and the personality characteristics of the person who conveys the message. Despite of their vagueness, prescripts concerning delivery may have been more important here than the rules of argumentation.

The single most characteristic feature of the rhetoric tradition is that it is advisory and prescriptive; in no way is it an empirical science. Empirical research into the effects of the verbal and non-verbal features of a message did not start until the rise of the science of mass-communication. Much like rhetoric this science has developed along two quite independent lines, one concentrating on the verbal (content, formulations, structuring) aspects of a message, the other dealing with the non-verbal aspects. The present study is exclusively concerned with the non-verbal aspects of communication, and not necessarily mass-communication. It can be maintained in general that many of the results of the research into the non-verbal aspects of mass-communication hold equally well in person-to-person communication especially in initial interaction between strangers. There have been contributions to the field from such diverse disciplines as social psychology, clinical psychology and psychiatry, personality theory, and - more recently - sociolinguistics.

1.1 PREVIOUS RESEARCH INTO NON-VERBAL ASPECTS OF SPEECH

There is no need for reviewing the research into the non-verbal aspects of communication if only since review papers are beginning to appear in which earlier reviews are reviewed together with more recent original research (Scherer, 1979). At this point it suffices to mention the two most important conclusions. The first conclusion, which one is obliged to draw from a survey

of the literature, is that the research efforts have resulted in little more than a confirmation of the existence of vocal stereotypes, i.e., stereotypical and often incorrect ideas about the relation between non-verbal (= vocal) aspects of a person's speech and her or his personality. The second conclusion, which has been reiterated by Scherer (e.g. Scherer, 1978), is that the first conclusion seems hardly warranted in view of the fundamental methodological shortcomings of the research under review.

Scherer proposes a research model which is meant to alleviate the methodological problems, or at least to point them out explicitly. To honor its originator it is called the Brunswikian Lens model. In this model, the details of which need not concern us here, a strict separation of levels of description is advocated. Specifically it is emphasized that three levels must be distinguished, viz. an 'objective' acoustic description of the speech, a 'perceptual' description of the speech, which is still meant to be quite objective, and a third level concerned with the attribution of personality characteristics. In most respects this proposal is not revolutionary. It parallels the commonly accepted notion of separation of levels in linguistics and phonology. Only in the research practice of the past, the indiscriminate intermixing of levels seems to have been the rule rather than the exception.

1.2 METHODOLOGICAL PROBLEMS IN THE DESCRIPTION OF VOCAL ASPECTS OF SPEECH

The description of the non-verbal aspects of speech is one of the fields where technical and methodological problems abound and solutions are rare. There even seems to be a lack of agreement on the collection of features which rightly belong to the class of non-verbal aspects of speech. In such a situation it is not surprising that terminological clashes abound. Although it is the major aim of this study to contribute to the improvement of the description of the non-verbal aspects of speech, our contribution is concerned with technical problems rather than with the construction of a theoretically sound and comprehensive description system. Therefore we can - and will - afford the luxury of taking a very pragmatic approach to the problem of defining the non-verbal aspects of speech and staking out the field. The pragmatic 'definition' of the non-verbal (or vocal, a term which is considered as synonymous) aspects of speech for the aim of our work is very much facilitated by the fact that it deals exclusively with texts read aloud. In such a situation the non-verbal aspects of a message can be said to be all information which is added to the text in and by the process of reading aloud. This additional information pertains to the identity of the talker (for those who know him or her personally) or to her/his sex, approximate age, approximate regional and social origin and approximate physical and mental states (for the listeners to whom the reader is a stranger). Vocal information is carried by characteristics of the speech signal which are more or less

constantly present, in contrast to the signal features that carry the verbal (linguistic, content) information and which need to vary constantly if any non-trivial message is to be conveyed. The aspects of the speech signal we are aiming at can (informally and without any claim as to completeness and precision) be described as the average speaking pitch, pitch variations, average loudness and loudness variations, average tempo and tempo variations, voice quality and precision of articulation. The majority of those features will be recognized as belonging to what is often called the suprasegmental or paralinguistic properties of speech.

The term 'paralinguistic features' suggests at least in part an explanation why the description of the properties in question suffers from so many problems and perplexities. The mere fact that we have to do with phenomena that are at the same time linguistic in nature (i.e., behave according to rules which are essentially arbitrary and used to convey meanings which are essentially discrete) and non-linguistic in nature (i.e., vary continuously as a function of e.g. the physiological state of arousal of the speaker reflecting conditions which vary continuously rather than discretely) makes their study very difficult and frustrating. Linguistically inclined scientists are hampered by the non-rule based continuous variations of the phenomena, whereas researchers geared towards the empirical correlational methodology of the social sciences have difficulty in coming to grips with the discrete linguistic part of the phenomena.

1.3 VOCAL ASPECTS OF SPEECH IN MODERN LINGUISTICS

Perhaps it is time now to interrupt this fairly general and abstract discussion by the presentation of some research examples which show how vocal aspects of speech play their role in modern linguistic research and at the same time bring to light the problems caused by the difficulties encountered in describing and manipulating non-verbal speech characteristics. The first case chosen for illustration is the much debated topic of the 'inherent norm' versus 'imposed value' hypotheses in sociolinguistics. The case centers around the explanation of the prescientific knowledge and the empirical scientific fact that people tend to display more positive attitudes towards some languages or language varieties than towards others (Nuijten, 1962; Brown & Lambert, 1976; Giles & Powesland, 1975). Brown and Lambert presented recordings of 20 adult male French-Canadians reading a short standard passage to 26 American students, none of whom knowing any French. The listeners were asked (among other things) to rate the social status of the speakers. It appeared that 47% of the variance in the ratings could be accounted for by the real status differences. This result compared favourably with the outcomes of a similar experiment in which the raters were French-Canadians (67% of the rater variance explained by real status differences). Brown and Lambert suggest that the relatively accurate perfor-

mance of the American raters shows the existence of cross-culturally similar voice quality differences between members of the lower and upper class.

Giles and Powesland, on the other hand, describe a series of experiments in which the reactions towards different language varieties, e.g. a dialect and the standard language or a low and a high prestige dialect were investigated using the matched-guise technique. This means that the same bi- or multilingual speakers produced the speech stimuli in all language varieties under study. It was hoped that in this way all properties of the speech signals not related to the language or dialect differences under study were kept constant. In these experiments it was invariably found that the expected differences in the evaluation of the language varieties did not show up if the raters were unable to recognize the languages or dialects and therefore could not identify the high prestige variety as such. This is the case when, for instance, Welsh students are asked to judge talkers using the (high prestige) dialect of Athens and the (low prestige) dialect spoken in the isle of Krete. From these findings it is inferred that there are no real audible differences between languages except those which are a direct result of the linguistic differences and that all differences in evaluative reactions are due to culture specific, learned attitudes which are triggered by the recognition of a certain language or dialect.

Clearly the methodological differences between the studies of Brown and Lambert on the one hand and those of Giles and his associates on the other are more than large enough to preclude any comparison. Also the outcomes of neither suffice to disprove the alternative hypothesis (nor to prove the hypothesis adhered to, for that matter). Brown and coworkers would be much more persuasive if they were able to prove the existence of systematic differences in the paralinguistic properties of the speech of high and low class talkers. Although they have attempted to obtain a description of the French-Canadian stimuli, they do not bring those descriptions to bear. Instead, they limit themselves to making remarks about having the intuitive impression that between-class differences are indeed present on the level of paralinguistic features. Most probably this course of action is taken since Brown et al. did not succeed in obtaining a reliable and accurate description of the paralinguistic properties of their stimuli. This assumption is completely in line with the fact that the group around Brown took refuge to a research strategy in which the stimuli under judgement were synthesized utterances, the prosodic characteristics of which were varied systematically (Brown, Strong & Rencher, 1974).

Giles and his coworkers are in an equally vulnerable position. First of all the lack of a description of the paralinguistic features of the stimuli used in their matched-guise experiments leaves no alternative option than to hope and trust that they were indeed constant over the language varieties. The fact that no effect of the varieties was found on the level of evaluation suggests that they did reach that goal. It may be objected, however, that bi- or tridialectal speakers do not constitute a random and representative sample of the speakers of the language varieties under study. Since we do not have

descriptions of the paralinguistic features of dialects and languages, it is not possible to rule out the possibility that the talkers have transferred those characteristics from one language variety to the others, where they may be less appropriate.

The possibility should seriously be considered that the results of the two research groups are by no means as contradictory as the hypotheses promoted suggest. To reconcile the seeming contradictions or to falsify one (or both) hypotheses would, however, call for experiments in which all relevant properties of the speech stimuli, including the vocal characteristics, are known and controlled systematically. It is quite obvious that this ideal cannot be reached except via the way of developing adequate techniques for dealing with the vocal characteristics of speech.

A quite different example of the importance of the non-verbal characteristics of speech, one which brings us back to the framework of personality perception, is taken from the field of speech technology. An American manufacturer of motor cars decided to equip the luxury models with a speech synthesis module for advising the driver as to the state of his vehicle. Since the synthesis module operates on the basis of Linear Prediction analysis-resynthesis (refer to chapter 4 for an explanation of this technique) a suitable talker had to be chosen. 'Because surveys indicate that luxury-car drivers prefer a male synthesized voice (....) a commanding yet warm male speaking voice' was sought (Finkelstein, 1983). From this example it appears that it may even become of economic importance to know what kind of voice is most appropriate in a given situation and -equally preponderant- to know exactly how that kind of voice can be produced.

1.4 AIMS OF THE PRESENT RESEARCH

It has been said before that in the description of vocal characteristics of speech three levels must be distinguished: acoustic and perceptual descriptions should not be confused and both should be distinguished from the level of attribution of personality characteristics. The first two levels are easily recognized as two subdisciplines which are often distinguished within the phonetic sciences, viz. acoustic and perceptual phonetics. The relation between the levels overlaps with the field of psychoacoustics. The level of personality attribution is less easily interpreted in phonetic or linguistic terms, but it certainly has very much to do with the pragmatic aspects of language behaviour. It has also been said that on all levels of description a large number of technical and methodological problems await solution.

Our study is concerned with the development of techniques for obtaining reliable, valid and cost-effective descriptions of some vocal aspects of speech on all three levels mentioned in the previous paragraph and with establishing links between these levels. At first sight a detailed perceptual description of

speech samples seems to be somewhat superfluous if one has an 'objective' description of the speech on the acoustic level. Also, perceptual descriptions are sometimes considered as very much less reliable than acoustic measurements. Yet, we think that both acoustic and perceptual descriptions are indispensable. Not only will it appear from this study that one can obtain highly reliable perceptual ratings of vocal speech parameters, but it seems also to be impossible to establish the validity of acoustic measurements when bypassing the low-level perceptual processes. This is, of course, not to say that no special precautions have to be taken in order to secure reliable perceptual descriptions of vocal characteristics of speech. Consequently chapters 2 and 3 of the present book are mainly concerned with the description and development of instruments and techniques for collecting and processing perceptual ratings of speech.

1.5 OPERATIONALIZATION OF THE RESEARCH GOALS

Despite of a deliberately pragmatic approach to the problem of the definition of vocal parameters we must, of course, make some decisions as to which features are to be measured both at the acoustic and perceptual level. Similar decisions are also necessary on the level of attribution, but there an independent strategy had to be followed. The choice of features, and therewith the development of measurement instruments, has been guided by two quite different and independent bodies of theory and empirical facts. The first is formed by the outcomes of previous research into the relation between vocal characteristics of speech and person perception. Even if the majority of the older studies has employed questionable methodologies leading to similarly questionable results a review allows one to make an inventory of features which have been found to relate in some way or another to personality characteristics. Since we will explicitly refrain from attempts to link the results of our descriptions with the true personality characteristics of our talker-subjects, the findings from research leading to the disappointing conclusion that everything was based on vocal stereotypes are as good as any other. In passing, it should be noted that the study of vocal stereotypes is as important as research into the relations between true personality and vocal characteristics, because it is those stereotypes which operate when we try to deduce the personality and psychological or physical state of a stranger from his or her voice. It is from an overview of this research that the list of features summed up earlier originated. To be a little more specific, increased pitch and increased pitch variations are known to induce the impression of a person who lacks self-confidence and who is tense. High loudness and a large dynamic range are associated with powerful and extrovert personalities, as is a fast speaking rate. At this stage it is not important that the relations mentioned (and all remaining ones not cited here) may only hold in specific cultures or only for adult males but not females. What counts is that a certain feature has been reported to have an effect on personality judgements.

The second body of theory and knowledge which has guided our work consisted of linguistic and phonetic studies aiming at the construction of a comprehensive framework for defining and describing vocal properties of speech. A recent attempt to integrate existing knowledge into a full-fledged description system is Laver (1980). Although features connected to intonation are lacking, more recent extended versions of his system do, at least in part, alleviate this shortcoming. Laver's work is important for two reasons. Firstly, he brings to bear phonetic theory in order to structure and systematize the rather large collection of vocal features. Secondly, he attempts not to mix different levels of description. On the contrary, while aiming at a description on the perceptual level, he carefully details the acoustic and articulatory manifestations or correlates of the perceptual features. The parallel with Scherer's emphasis on separation of levels will be obvious.

Readily available recent reviews of the literature on the definition and description of vocal features (besides Laver's 1980 book see also Crystal, 1975) and on the relations between various levels of description (physiological, acoustic, perceptual, linguistic) relieves us from the obligation to repeat such a review here. From the data collected in these reviews it appears that we can safely assume that on all these levels vocal phenomena relevant to person perception do exist and that there must be systematic relations between the levels.

Within the fields of physiological, acoustic and perceptual phonetics a further subdivision should be recognized, which is very often made but equally often left implicit, viz. the distinction between phonation and articulation. It is customary in phonetics to concentrate on articulation and to leave phonation to the attention of the more medically (or socially) inclined sciences. This division of labour is obviously adequate if phonetics is conceived of as tightly linked to structural linguistics, since phonation types hardly ever act as distinctive features in the sense of classical phonology. (It should be mentioned, of course, that some authors (e.g. Ladefoged 1973; Ladefoged & Maddieson, 1983) do recognize so called laryngeal features, but the use of these features is limited to a few languages and their status is not unchallenged.) This corresponds with the idea that phonation belongs more to the realm of continuously varying phenomena than to the one of discretely varying objects (or should one rather say: described as discretely varying) which linguistics traditionally deals with.

Within the framework of research into the attribution of personality characteristics based on speech, features related to phonation play a role which is at least as important as that of articulatory features. Since the acceptable range of variation within articulatory features may be much more restricted, due to their linguistic function, phonatory features might even be the more important ones here. Thus there are at least two reasons why much effort will have to be invested into the development of procedures for measuring phonatory features i.e., their importance in forming the speaker's image and their relative neglect in existing phonetic work.

Conceptually it seems fairly easy to separate phonation and articulation, by defining phonation as everything going on at the level of the larynx and articulation as everything regarding the actions of and in the supra-laryngeal vocal tract. When it comes to 'objective' reliable measurements, however, things prove to be less easy. On the level of physiological phonetics the larynx and the supra-laryngeal structures can indeed be treated separately (at least in essence), but then the larynx proves to be very difficult to access without disturbing the process of speech production to a degree where the measurement results run the risk of completely losing their validity. On the acoustic level the contributions of articulation and phonation are intertwined in an extremely intricate way and it will appear very difficult to effect a separation.

Such a separation of contributions is certainly not easier on the level of perception. The importance of separating articulation and phonation on the level of the acoustic signal (in order to arrive at an accurate and explicit description of the distinguishing features of phonation) is reflected by the fact that a complete chapter of the present book is devoted to this topic (viz. chapter 4). The separation of phonation and articulation also plays a major role in the development of the rating instrument described in chapter 2.

Despite the difficulties encountered in trying to measure physiological activity of the larynx during normal speech production, we have carried out such measurements as part of our efforts in developing valid and reliable techniques for isolating the phonatory component on the acoustic level. These laryngeal measurements are the only physiological measurements carried out in the course of the work reported here. Yet, the position will be taken that physiological signals, recorded simultaneously with the acoustic speech wave, offer a wealth of information not easily obtainable from acoustic or perceptual measurements. In some cases additional physiological information even seems to be indispensable when it comes to interpreting the results of acoustic analyses. Also, physiological signals can be very helpful for guiding the processing of acoustic speech signals.

1.6 FLAVOURS: METHODOLOGY VERSUS CONTENT

One of the most prominent features of the present book is the emphasis which is laid on methodology and techniques. One might even say that at some places the subject matter which is of eventual interest, viz. an improved understanding of the way in which vocal speech characteristics induce the attribution of certain personality characteristics, has deliberately been sacrificed in favour of a more detailed analysis of measurement methods. The premeditated choice to favour method above content can easily be motivated. First of all, reliable results of an investigation into the relation between speech characteristics and (real or perceived) personality cannot be expected unless there are reliable and valid methods to describe and measure the phenomena under study. But equally importantly, the present book is prima-

rily a study in instrumental phonetics. As such it reflects the opinion that, especially if one works in an environment of linguists and social scientists, it is an important task of the instrumental phonetician to provide his fellow-investigators with the tools needed in order to carry out their material research in a way which stands up to current requirements with respect to methodology.

In summary, then, this book deals with vocal aspects of speech and their impact on person perception. Since previous efforts in that field have produced unsatisfactory results mainly because of inadequate methodology, the development of improved methods for describing and measuring vocal speech phenomena and for the processing of the results of measurement and description is heavily stressed. In no case, however, is a technique developed or explained which is not actually applied in small-scale studies of the way in which rather small groups of talkers are judged by groups of listeners. Somewhat paradoxically, the groups of raters have always been fairly large. The paradox is resolved, however, if it is recognized that the size of the rater groups was dictated by exactly the methodological inclination of the work.

In more concrete terms, chapter 2 deals with the development of a preliminary instrument for obtaining perceptual ratings of vocal speech characteristics and a limited number of personality characteristics. A number of properties of a rating instrument can only be determined from its actual use. These properties are dealt with in chapter 3. Scaling analysis and factor analysis of the ratings of large groups of judges eventually resulted in a small and versatile scaling instrument that covers the relevant dimensions of the vocal speech characteristics. Chapter 4 is devoted to the development of a reliable technique for separating the contributions of phonation and articulation on the level of the acoustic speech signal. Chapter 5 deals with the development and description of acoustic measurement techniques for a large number of vocal parameters which appeared to play a role on the level of speech perception. In chapter 6 an attempt is made to establish links between the acoustic and perceptual (and attributional) levels. Chapter 7, finally, attempts to interpret and integrate the findings and to identify promising opportunities for further research.

The construction of a scaling instrument

2.0 INTRODUCTION

If we are to explain perceptual, subjective ratings of speech (and therewith of talkers) by means of acoustic measures, we clearly need both ratings and acoustic measurements. This chapter deals with the considerations and preparations for the development of an instrument with which the perceptual ratings might be procured.

If perceptual ratings are to be related to acoustic measurements they must not be too vague and general. On the contrary, we aim explicitly at the separation of articulation and phonation and the consequences of the perception of those speech characteristics for the rating of the psychological and social status of the talkers. One way to obtain ratings of articulation and phonation is to use phoneticians or speech scientists who are specifically trained for the job. The alternative approach is to employ essentially naive judges. The former approach undoubtedly has the advantage of securing ratings that are much more detailed and perhaps also much more precise. The latter approach seems to have the advantage of a much higher ecological validity: what we are really interested in is to know how naive listeners perceive speech and in what way their perception of somebody's speech influences the perception of his or her personality.

For a number of reasons we decided to try and acquire ratings of 'almost naive' judges. The most important of those reasons must be mentioned and explained here, since it has numerous and pervasive consequences. Using 'almost naive' subjects allowed a very close cooperation with a number of colleagues at the Institute of Phonetic Sciences at the University of Amsterdam (to wit Wil Fagel, Leo van Herpt and Jan Blom) who were working on the construction of an efficient scaling instrument for exactly that purpose, i.e., to obtain reliable ratings of the quality of phonation and articulation from essentially untrained raters. Thus the results reported in this (and in the next) chapter derive from the efforts of a team of which the present author was something of an outside member.

The work on the scaling instrument was started in the late sixties (Blom & Koopmans-van Beinum, 1973). From the very beginning of that long-term project it was decided to use the technique known as the semantic differential (Osgood, Suci & Tannenbaum 1957). A semantic differential is a very

attractive tool, since there is quite extensive experience in its application for the rating of an extremely wide range of concepts and objects. The technique is also capable of acquiring a large amount of ratings very quickly.

It must be admitted, though, that the use of batteries of bipolar scales to assess the opinion of judges is not approved of in all scientific circles. In a footnote (that is very clear for one part and highly cryptical for the other) Brown, Strong & Rencher (1975) complain that psychology journals refused to accept their paper reporting on research in which a number of scales were used that were not standard Osgoodian. Now it is not clear what exactly are standard Osgoodian scales, if only for the simple reason that Osgood et al. (1957) refrain from canonizing specific scales. Therefore it is not possible to decide if Brown and his colleagues have been the victims of a general suspicion harboured by experimental psychologists against research techniques based on factor analysis or if they instead have not met the minimal prerequisites for constructing a measurement instrument that can yield reproducible results. To prevent a comparable, and possibly justifiable suspicion against our results, we will summarize the basic requirements a scaling instrument should fulfil and show that the instrument we have developed meets these requirements to a reasonable degree

Chapter 3 in Osgood et al. (1957) opens with an extensive treatment of the 'Construction and Administration of a Semantic Differential'. The topics dealt with are the selection of the objects or concepts that are to be judged, the way in which the actual judgement experiment must or can be carried out and the selection of the scales the instrument is composed of. We will treat these topics below.

2.1 SELECTION OF THE OBJECTS TO BE JUDGED

It goes without saying that the selection of the objects or concepts is determined primarily by the aim of the research. Problems can only arise if practical considerations force one to sampling a population. In such a situation the sampling must fulfil the usual requirements. the objects or concepts chosen must be representative of the population and relevant with regard to the purpose of the research. The population that we are interested in eventually is 'the speech of humans when reading a text'. Because we know very little about the distribution of that population, random sampling is not possible; we have to take recourse to what is called 'good judgement' instead. Exercising 'good judgement' means selecting objects which span the ranges of interest in the population, which have a stable and unitary meaning to the judges and which are familiar to all of them.

The objects to be judged are recordings of a number of speakers reading a text. In order to be able to assess the influence of the content of the text we

have used two texts, differing widely in content, one summarizing a touristic bicycle tour in Denmark, the other reporting on the state of affairs in Iran after the overthrow of the regime of the Shah. The speakers who read these texts differ -to our intuitive impression- considerably both in phonation and articulation quality. Therefore we may safely state that our samples are in a sense representative of and relevant to the population of interest.

The requirement that the samples have a stable and unitary meaning to the judges seems to be somewhat problematic. We want to use a semantic differential to obtain judgements on articulation, phonation and more general aspects of speech samples. As long as all these things can be taken to be more or less independent aspects or dimensions of one concept, viz. 'the speech of this speaker' everything is all right. To the degree, however, to which those aspects are not independent or scales do not refer unambiguously to one single aspect, it is possible that raters vacillate between different meanings of the scale adjectives or allow their judgements to be influenced by their perception of other aspects in different ways for different stimuli (the so called 'Halo-effect'). We will return to this subject when dealing with the selection of the scales.

Although the pronunciation of all the speakers in our experiment is fairly close to what is called 'Algemeen Beschaafd Nederlands' (General Cultivated Dutch, a concept which is more or less equivalent to Received Pronunciation for English) the majority of the speakers can nevertheless be recognized as coming from the Southern part of the country. Thus there may be some differences in the familiarity with the speech samples between groups of listeners originating from different parts of the country. It can be maintained as well, however, that these differences are more related to social attitudes towards speech of 'strangers' than to the degree of familiarity.

It seems worthwhile and necessary to end this section on the selection of the stimuli with a remark on their number. The collection of the rating scores was done as part of the work aiming at the construction of an efficient and reliable measuring instrument. Since neither the number of judgement dimensions nor the behaviour of the individual rating scales was known exactly, it was decided to include as many scales as possible in the preliminary version of the instrument. As will be explained shortly, we ended up with a rating form consisting of 35 scales. According to Osgood et al. (1957:80) such a test will take three to five minutes, so if we do not want to hurry our listeners we cannot have more than twelve stimuli (= readings) judged in a session lasting one hour. Combined with the fact that it proved impossible to have large groups of listeners return for a second or third session this made us decide not to surpass the number of twelve stimuli to be judged. With each speaker reading two texts this confines the number of speakers to a mere six. Although sociolinguists claim (Labov, 1972) that five informants suffice to represent a social group, we do not feel that six speakers are sufficient to

cover the total range of all relevant aspects of articulation and phonation quality. Therefore the present study can only generate hypotheses that should be tested using techniques that are more efficient than our preliminary semantic differential and allow for the judgement of a much larger number of stimuli.

2.2 THE ADMINISTRATION OF A SEMANTIC DIFFERENTIAL

Of several methods of administering the semantic differential described in Osgood et al. (1957:80) only what is called the Form II technique is applicable in our work. Against this method -in which all scales are printed on a single sheet of paper- some objections can be, and have actually been, raised, the most important one being that it may cause the judges to try to be 'consistent' in their scores on similar looking scales. There is another method of administering a semantic differential, not mentioned in Osgood et al. but developed in order to alleviate the said objections (Nuijten, personal communication), in which there is one scale per page in a scoring booklet. This technique too has its drawbacks. It will lead more easily to subjects omitting scales because they inadvertently turn two or more pages of their booklet with scales at once. An additional drawback of this method is that it takes much more time to prepare the scoring booklets and to process the scores.

In order to assess the influence of the method of presentation the following experiment was carried out. We asked a group of 39 subjects, students of Dutch at an institute of higher education in Utrecht, to rate the 35 scales in a preliminary version of the instrument for three concepts viz. the ideal male voice, the ideal female voice, and their own voice. These subjects were each given three booklets consisting of 35 pages, one scale on every page. Mean values for the three concepts were computed on each scale and compared with the scores of a group of 55 first-year students of Dutch at the University of Leyden, who got all scales on a single sheet of paper and the same instruction (i.e., to score the scales for ideal male and female and own voice). Using a t-test only three differences between the groups appeared to be significant at the 1% level, viz. the scale 'careless $\leftrightarrow$ precise articulation' in the scores for the ideal male voice and the scales 'tense $\leftrightarrow$ relaxed' and 'distinct $\leftrightarrow$ indistinct' for the ideal female voice. This number of significant differences is about what one must expect on the basis of pure chance when making 105 independent tests. Therefore, and because differences might also be explained as real differences between groups, we decided that the enormous amount of extra work entailed by the 'one scale per page' format would not be justified.

Although the order of the scales did not prove to be a factor of importance in previous research (Voiers, 1964), we decided to put the conjecture to at least some form of test. In order to do so we gave half the subjects who got the booklet form in the experiment described above the scales in one order and

the other half got the scales in the reverse order. A comparison of the scores of both subgroups did not reveal a single significant difference. From this we drew the conclusion that it was not necessary to produce a number of different scoring forms but that one form having all scales in a fixed order would suffice.

We have used seven point scales in all our experiments. The meaning of the scale positions were explained to the judges in the instruction that preceded each set of scoring forms. The instruction was essentially a translation of the example given in Osgood et al. (1957).

2.3 THE SELECTION OF THE SCALES

The considerations regarding the selection of the scales are very difficult to explain without reference to the theory of meaning from which the semantic differential has emerged. This theory, then, regards the meaning of a concept as a point in a multidimensional Euclidean space. The exact dimensionality of the semantic space is not known; supposedly it is to some extent dependent on the nature of the concepts or objects to be rated. Nevertheless in most experiments three dominant dimensions show up, without being sufficient to explain all variance. From this it must be concluded that these three major dimensions do not exhaust the dimensionality of the semantic space, or in other words, that the meaning of objects can differ along more orthogonal dimensions than those that have come to be called 'evaluation', 'potency' and 'activity'.

Not all dimensions of the semantic space appear to be equally important in terms of the percentage of variance explained. It is usual for the evaluation dimension to explain about the same amount of variance as the potency and activity dimensions together; the potency dimension seems to be as important as the activity dimension. Together the three dominant dimensions tend to account for between 60% and 80% of the total variance.

The said three dimensions have emerged from factor analyses on the scores of many groups of subjects rating a number of objects or concepts on a large number of bipolar scales, with polar adjectives of clearly opposite meaning. Additional factors will only appear if one uses scales that form clusters of high mutual correlation but that do not correlate highly with the scales related to one of the dominant dimensions.

Since the factor analysis model (Guilford, 1954) is ideally suited to look for independent dimensions in a data aggregate, it seems to be the natural model to process the scores obtained with a semantic differential. Both the underlying semantic model that assumes the 'meaning' of objects to be a point in a multidimensional space and the factor analysis model for the processing of large amounts of correlation data demand that the scales have a number of

specific properties, the most important ones being factorial composition, relevance to the objects under analysis, stability of meaning and linearity. All these features have played a role in our decisions to accept or reject scales found in the literature and in constructing new scales.

2.3.1 Factorial composition and relevance of scales

Given the objects to be rated, one can almost always conceive of scales that do not have an established relation with one of the dominant factors (i.e., evaluation, activity and potency), but which are nevertheless very relevant in the sense that the objects differ in an interesting way with respect to the attribute described by the scale. These scales are said to be of unknown factorial composition, i.e., it is not known with which factor these scales are associated. In our study of judgements of speech stimuli we are interested in articulation, phonation and paralinguistic features, aspects that are perhaps multidimensional of their own. In such a situation it is only natural to try to devise special scales to probe these properties. Osgood et al explicitly approve of this procedure under the condition that the new scales are carefully chosen and combined with enough reference scales of known factorial composition to allow for the interpretation of the results.

Relevance to the objects under study is also a criterion for choosing scales from the well established collections. Osgood et al give a very illuminating example of what is meant with relevance. Both the scales 'beautiful $\leftrightarrow$ ugly' and 'fair $\leftrightarrow$ unfair' are representatives of the evaluation factor, but that does not mean that they are always equally appropriate. When rating politicians for their opinions a scale like 'fair $\leftrightarrow$ unfair' seems to be extremely relevant whereas a scale like 'beautiful $\leftrightarrow$ ugly' obviously is not. If, on the other hand, not politicians are to be judged but objects of creative art the reverse will be true.

2.3.2 Stability of meaning and linearity of scales

Many adjectives have a metaphorical meaning along with their 'literal' meaning. In which meaning a polar adjective on a semantic differential scale will be interpreted does not only depend on the meaning of the term that it is paired with (and that may have a similar metaphorical meaning) but also to a large degree on the concepts being judged. Raters are likely to interpret the adjectives in a meaning relevant to the objects. If the set of objects under study is not homogeneous, it may occur that a scale is used in one sense for some objects and in another sense for others. An example of such an effect was found in the study of Blom & van Herpt (1976-a). Their subjects rated the speech of five women and five men on a.o. the scale 'high pitch $\leftrightarrow$ low pitch'. It appeared that part of the subjects more or less consistently interpreted the scale in the sense of high absolute or low absolute pitch of the voice irrespective of the sex of the speaker (leading to the women clustering at the

'high' end of the scale and the men at the 'low' end), whereas the other part of the listeners took the meaning in a relative sense, viz. as 'a high (low) pitch for a female' when the speaker was a woman and similarly for the men. The latter interpretation leads to men and women obtaining scores at every point of the scale about equally often. Instability of meaning within a set of objects is less likely to occur if the objects are more homogeneous. Raters can be prevented from choosing the 'wrong' interpretation of a scale by giving specific instructions, or by adding modifiers to the scale adjectives, like 'low (high) pitch for a male/female'.

The mathematical model underlying factor analysis requires that the correlations between the rating scales be linear. Obviously, this requirement will be fulfilled if the scales are linear themselves. A bipolar scale labelled with adjectives of opposite meaning can nevertheless be highly nonlinear; in actual fact such scales can even be non-monotonic. An example, again taken from Blom & van Herpt (1976-a) will clarify this. One of their scales read 'sloppy pronunciation $\leftrightarrow$ hypercorrect pronunciation'; the adjectives have clearly opposite meanings but it is equally true that they both have a negative connotation. Such scales are very detrimental to analysis techniques that presuppose linear correlations between scales. The example hinted at by Osgood et al. (1957:79) is perhaps even more dramatic. The scale 'rugged $\leftrightarrow$ delicate' can behave as a positive - neutral - negative continuum or as a negative - positive or a positive - negative continuum depending on the object which is judged.

The requirements dealt with up to now are necessary to make a good measuring instrument but they are not sufficient. One aspect the characteristics explained so far have in common is that they can be treated a priori, either by test or by 'good judgement'. Scales are also required to be reliable, sensitive (i.e., to have a high discriminating power) and valid. Unfortunately it is impossible to establish the degree to which these additional requirements are met a priori. The sensitivity of a scale is highly dependent on the dissimilarity of the objects with respect to the attribute at issue; the same is true for the reliability whereas the validity can only be assessed post hoc by comparing scale values assigned to objects with independent measurements. We will deal with these properties of the scales at length in chapter 3, when we are in the position to discuss actual scores.

2.3.3 *Choosing from inventories of scales*

Since we wanted to comply with the position taken by Osgood et al. that it is an acceptable procedure to make a semantic differential by combining specially designed scales with reference scales of known factorial composition, we had two inventories of scales to choose from. The first one is the fairly closed set of scales of which it is guaranteed that they have high loadings on either the evaluation, the potency or the activity dimension.

These scales can be said to be 'standard Osgoodian'. The other inventory is an open set of scales that have been used in previous investigations of subjective ratings of speech and to which one can, of course, add new ones. We had identified three aspects on which we would like to obtain information, viz. the articulation and the phonation quality of speakers and the ratings given to them by listeners on socio- and paralinguistic scales. These aspects are not necessarily mutually independent. On the contrary, one would assume that the ratings on socio- and paralinguistic scales are determined to a large extent by the listeners' perception of phonation and articulation quality.

Although one might be inclined to think that it should be easy to distinguish between phonation and articulation it proves to be very difficult to disentangle the two contributions. This is as true on the level of human perception as on the level of acoustic analysis.

A survey of the literature on perceptual ratings of speech quality and on the attribution of personality traits on the basis of perceived speech quality resulted in a large number of references. Our first criterion for eliminating publications was the explicitness with which the functioning of the rating scales used was treated. The idea underlying this decision was that it seemed pointless to select scales from which it could in no way be established to what extent they fulfilled the formal requirements when used in earlier research. This criterion reduced the number of relevant publications considerably. As a matter of fact we took as our point of departure the papers by Uldall (1960), Voiers (1964), von Bismarck (1974), Brown, Strong & Rencher (1975), Takahashi & Koike (1975), Blom & van Herpt (1976-a) and Fritzell, Hammarberg & Wedin (1977). More specifically, we took the scales that proved to function well in Blom and van Herpt and tried to add to them scales for those aspects that we felt were not yet covered sufficiently.

Blom and van Herpt identified a set of 14 scales that were sufficient to describe a 'Relative Speech Appreciation profile' for ten speakers rated by 100 listeners. Since our study has employed a similar approach, it is worthwhile to explain the steps taken by Blom & Koopmans-van Beinum (1973) and Blom & van Herpt (1976-a) to construct their rating scales in some detail. They started out with some 800 terms used in the Dutch literature on phonetics and speech science to describe aspects of articulation and phonation. These terms were pairwise combined to form bipolar scales. Omitting all terms which refer to grammar and lexis or to obviously pathological conditions 85 scales could be constructed. Discarding scales that were clearly synonymous with others resulted in an eventual set of 46 scales. This set could be divided into three heavily overlapping subsets of about equal size according to whether the scale attribute bears on phonation, on articulation or on both.

Next, all adjectives used to label the bipolar scales were presented to subjects who were instructed to rate them with respect to their desirability for speech. It appeared that a number of scales were labelled with one desirable

(positive) adjective and an opposing one rated as undesirable (negative). Other scales, however, appeared to pair a neutral adjective (a quality not especially desirable nor undesirable) with either a positive or a negative one. A number of scales appeared to pair two adjectives which both were rated as undesirable. Furthermore, the scores of a group consisting of 100 raters judging recordings of five females and five males were analysed using the Law of Categorical Judgement (Torgerson, 1958). This procedure was used to obtain estimates of a number of formal properties of the scales. (A similar procedure was suggested by Osgood et al. (1957:146)). The scale analysis resulted in the identification of 16 scales that violated at least one of the requirements for being used in a linear procedure like factor analysis. Dropping these scales Blom & van Herpt (1976-a) were left with a set of 30 scales.

A factor analysis resulted in three factors, labelled 'voice appreciation', 'articulation quality' and 'abnormality', the first factor accounting for about 70% of the explained variance. This overwhelming contribution motivated the decision to try a one factor solution, from which 14 scales appeared to have high communalities. For both the three and the one factor solution factor scores for the ten speakers were computed and converted to positions on a one-dimensional scale. For the one factor solution factor scores are, of course, positions along one dimension; for the three factor solution a least squares approximation to the distances between the speakers in the three dimensional space was found. The rankings of the speakers as produced by both solutions were compared with each other and with a ranking arrived at independently in an experiment in which subjects gave global judgements of the speech samples. The Spearman correlation between the one and the three factor solution amounted to .94, that between the three factor solution and the ranking based on global judgements was .95 and the correlation between the one factor solution and the global judgement was again .94. From these results Blom and van Herpt concluded that the 13 scales that had low communalities in the one factor solution could be eliminated too and that the remaining 14 scales suffice to measure what they called 'Speech Appreciation'. These 14 scales have formed the basis for our semantic differential. They are marked with an asterisk in Table 2.1.

Somewhat surprisingly, two of those 14 scales appeared to be of the type with 'undesirable' adjectives at both poles. We found it necessary to change one of the adjectives in those scales in order to be sure that they are positive ↔ negative. We have changed 'careless ↔ hypercorrect' into 'careless ↔ precise' and 'cultivated ↔ slipshod' into 'polished ↔ slipshod'.

Taking into account the motivation for the selection of the eventual 14 scales, viz. their high communalities in a one factor solution, it is not surprising to find that the majority bears on both phonation and articulation, if they are not outright evaluation scales (recall that the one factor was called Speech Appreciation). Because one of our objects is to separate articulation and phonation it was necessary to look for additional scales that relate to either phonation or articulation. Additional scales were also neces-

Table 2.1. Scales comprising the preliminary form of the semantic differential. Column I contains the Dutch adjectives, printed on the scoring forms; the approximate English translations are given in column IV. Column II indicates the purpose the scales are supposed to serve. Scales which have an asterisk in column II form the set of 14 scales making up the 'Voice Appreciation Profile' in Blom & van Herpt. Column III indicates the origin of the scales.

	I	II	III	IV		
1	aangenaam	onaangenaam	PA ¹ *	V/vB/01/B1 ²	pleasant	unpleasant
2	dof	helder	P	V/vB/TK/B1	dull	clear
3	vriendelijk	kortaf	A S	B	friendly	curt
4	luid	zacht	P O	V/vB	loud	soft
5	hees	niet hees	P O	B1	husky	not husky
6	stereotiep	gevariëerd	PA *	V/B1	stereotyped	varied
7	aktief	passief	A S	V/TK/B/03	active	passive
8	ongekunsteld	geaffecteerd	A S	B1	artless	affected
9	flets	klankrijk	P *	B1	colourless	sonorous
10	slordig	precies	A *	B1	careless	precise
11	hoog	laag	P O	V/vB/B1	high pitch	low pitch
12	stevig	slap	P	02	firm	slack
13	krakerig	niet krakerig	P	B1	creaky	not creaky
14	traag	vlot	AO/S	03	dragging	brisk
15	melodius	eentonig	PA *	B1	melodious	monotonous
16	krachtig	zwak	PA S *	V/vB/TK/B1/02	powerful	weak
17	lelijk	mooi	PA *	V/TK/B1/01	ugly	beautiful
18	ruw	glad	P	V/vB	rough	smooth
19	doods	levendig	PA *	vB/B1	spiritless	vivacious
20	regionaal	niet regionaal	A	B1	regional	non regional
21	gerond	hoekig	P	vB/03	rounded	angular
22	diep	schel	P		deep	shrill
23	arm	rijk	PA *	01/B1	poor	rich
24	gespannen	ontspannen	PA S	V/vB/U/F	tense	relaxed
25	expressief	uitdrukingsloos	A S *	B1	expressive	expressionless
26	vast	onvast	P *	V/F/B1	steady	unsteady
27	plat	beschaafd	A S *	B1	broad	cultured
28	zelfverzekerd	weifelend	A S	V/vB	selfconfident	wavering
29	zeurig	opgewekt	A S	B1	whining	cheerful
30	verzorgd	onverzorgd	A *	B1	polished	slovenly
31	afwijkend	normaal	PA *	B1	deviating	normal
32	snel	langzaam	A	V/TK/03/B1	quick	slow
33	vol	iel	P		full	thin
34	opgewonden	rustig	A	U/B	agitated	calm
35	duidelijk	onduidelijk	A		distinct	indistinct

1. A: articulation scale, O: objective physical property, P: phonation scale, S: socio- or paralinguistic scale.
2. B: Brown et al., 1974; B1: Blom & van Herpt, 1974; F: Fritzell et al., 1977; 01-03: Standard reference scales for evaluation, potency and activity factors, respectively; TK: Takahashi & Koike, 1975; U: Uldall, 1960; vB: von Bismarck, 1974.

sary to comply with the requirement that enough reference scales of known factorial composition are included and to investigate social evaluation. Finally, we wanted to include a number of scales which could serve as references for the comparison of subjective ratings with acoustic or physical measurements.

As stated before, the set of 46 scales of Blom and his associates served as the first supply of additional scales. We considered all scales that were eliminated due to low communalities or low discriminating power as useful candidates, whenever similar scales had proved to perform well in the experiments of a.o. Voiers (1964) and von Bismarck (1974). From the scales

marked with an asterisk in table 2.1 only (1), (17) and perhaps also (23) can be taken as reference scales with guaranteed high loadings on the evaluative factor in the standard Osgoodian semantic differential, whereas only (16) represents the potency dimension. The activity dimension is not represented at all. In order to include a set of reference scales for the activity factor we have added the scales 'active $\leftrightarrow$ passive', 'quick $\leftrightarrow$ slow', and 'dragging $\leftrightarrow$ brisk'; to complete the set of potency scales we have added 'firm $\leftrightarrow$ slack'.

As references for acoustic and physical measurements we have introduced the scales 'loud $\leftrightarrow$ soft', 'quick $\leftrightarrow$ slow' and 'high $\leftrightarrow$ low'. The latter two scales had been eliminated from the original set of Blom et al. because the former did not discriminate between speakers and the latter was two-dimensional (refer to the discussion in 2.3.2). The multidimensional character of the scale 'high $\leftrightarrow$ low' was dealt with by adding the modifiers 'for a man' or 'for a woman', depending, of course, on the sex of the speaker under judgement. Note that, at least for the time being, the scale 'quick $\leftrightarrow$ slow' is supposed to serve two goals. The scale shares this property with many others.

Additional scales to measure aspects of phonation are taken from Voiers and von Bismarck or they are newly constructed; additional articulation scales are constructed by combining adjectives used by Blom et al. in a new way. Scales meant to measure social evaluation are partly taken from Blom et al., partly from Uldall (1960) and partly constructed from the adjectives used by Blom and associates. Table 2.1 summarizes the first version of our semantic differential; it also contains information on the origin of the scales and on the purposes they are meant to serve in our research. Column I of this table lists the Dutch adjectives which appeared on the scoring forms. The English translations in column IV are only approximate.

2.3.4 *The status of the rating scales*

Ratings procured by means of a semantic differential are most probably associative rather than analytical. Raters do not get any special training, nor very specific instructions with respect to the meaning the experimenter wants them to attach to the scale terms. On the contrary, much effort is (or at least should be) spent in ascertaining that all scale adjectives are among the very frequently used words in the language and yet have an unambiguous meaning. An additional reason why the ratings must be assumed as associative rather than analytic is the fact that the raters are encouraged to give their first impression.

A direct consequence of the use of naive raters who are urged to work themselves at a fast rate through the scales is that it becomes impossible to include scales which can only be defined with the help of technical terms. Thus one of the most striking properties of the set of scales in Table 2.1, a set specifically designed for the rating of a.o. phonation and articulation, is the absence of terms like harsh, breathy, throaty, velarized, palatalized or nasalized. Scales defined by similar terms were among the 46 scales in the original

inventory of Blom et al. They appeared to perform extremely badly, however, for reasons that are not hard to guess. Yet, it should be clear that by excluding the use of this kind of scales we sacrifice a substantial amount of detail. Another consequence is the large number of scales which can refer to both articulation and phonation. This will not only aggravate potential problems with the interpretation of the results of the ratings, but also increase the risk that different raters attach different meanings to some scales.

Even if the boundary may not always be very clear and unambiguous, the set of scales in Table 2.1 can be divided into two broad classes. The first class contains the scales which are not marked as suppliers of information on the psychological and social status of the talker, whereas the second class contains all scales which do provide such information. It may be somewhat surprising that scores which are essentially associative in nature are still considered to yield fairly objective results. But then, the general nature of the scales allows the use of large numbers of raters, so that there is hope that the 'subjective' ratings of many judges average out to a fairly objective description of relevant characteristics of the stimuli. One last comment must be made, relating to the decision to combine ratings of more 'technical' aspects of the speech samples and ratings of person characteristics in a single experiment. This procedure has the drawback of incurring the risk of contamination of levels, the more so since all para- and sociolinguistic scales can also be interpreted as articulation or phonation scales. We decided to accept this drawback, mainly since the alternative approach, viz. conducting two separate experiments, one for rating the phonation/articulation scales and a second experiment for rating the socio- and paralinguistic scales, encountered great practical difficulties. Having the same group of listeners return for a repeated rating session proved to be impossible. Using two different groups for the rating of the two sets of scales introduces the tricky problem of matching groups in a situation where it is far from being clear what subject characteristics are relevant for the task. These considerations made the scale turn in favour of an approach in which no apparent distinction was made between the two sets of scales.

2.4 TESTING THE SCALES FOR LINEARITY AND RELEVANCE

It has been said in section 2.3.1 that some aspects of the scales can be dealt with before the scales have been used in a pilot rating experiment with actual stimuli. Linearity and relevance are foremost among these features. Although we paid careful attention to these features in choosing the scales, we still deemed it worthwhile to try and test as many properties of the instrument in a paper-and-pencil experiment preceding an actual rating experiment. The format of this experiment was equal to that described in section 2.2: a large number of raters were given three complete rating forms

and were instructed to score these for the ideal male voice, the ideal female voices and the rater's own voice. The principal aim of this paper-and-pencil experiment was to minimize the risk of a failure of the very large and equally laborious and time consuming rating experiments to be described in the next chapter due to the misbehaviour of a sizable proportion of the scales selected. Since very much care had been taken in choosing the scales this risk was likely to be small but we preferred to change the assumption into more solid knowledge, the more so since a paper-and-pencil experiment was easy to perform. Thus, this preliminary experiment was not meant to test the scales for admission in the rating instrument to be used in the subsequent research. Only if a scale would appear to cut across most or all requirements it would have to be eliminated.

Scales on which the scores were not concentrated at the supposedly positive extreme were considered as possible causes of curvilinear correlations. To clarify this, take as an example the scales 'careless pronunciation $\leftrightarrow$ precise pronunciation', 'careless pronunciation $\leftrightarrow$ hypercorrect pronunciation' and 'broad $\leftrightarrow$ cultured'. The first and the last scales have their ideal point towards the right hand extreme of the scale: ideally, the pronunciation ought to be both fairly precise and fairly cultured. The middle scale, however, has its ideal point somewhere in the middle: the ideal pronunciation should neither be careless nor hypercorrect. Moreover, both hypercorrect and careless pronunciation are likely to be associated with 'broad' rather than with 'cultured'. The scores for ideal voice/pronunciation will yield additional information to adjust the instrument. Scales that appear to have their ideal point close to the middle of the range obviously are defined by means of the wrong adjectives. In any case, the allegedly positive adjective is not interpreted so by the raters.

A scale was supposed to be irrelevant if two conditions were met simultaneously, viz. ideal scores in the neighbourhood of the middle of the scale and a small variance in the scores for the own voice/pronunciation of the raters. This situation differs from the case of possible curvilinear correlations. If intermediate scores for the ideal point on the scale are combined with a large variance in the scores for the raters' own voice, the scale attribute is perhaps ill-defined but not irrelevant as testify the differences between the raters. If both the ideal is somewhere in the middle of the scale and the variance of the scores for own voice is low the attribute is likely to be irrelevant, perhaps because it is ill-defined.

For this experiment scoring forms were prepared on which the scales were printed in what might be called a balanced random order. The balancing was done with respect to two aspects: half of the scales had their (supposedly) positive pole at the right hand endpoint and the other half the other way round. This prevents the judges from being biased to one side of the scoring form. The order of the scales was balanced in such a way that similar scales were separated from each other by scales probing other aspects. Thus there were no successions of e.g. two or more articulation scales.

Many different groups, together comprising several hundreds of subjects, rated the above mentioned three concepts on the preliminary form of our semantic differential. Because we saw a possibility to extract from these scores interesting information about sex-stereotypes, we selected those groups in which both sexes were about equally represented for a detailed analysis. Those groups happened to consist of University students, viz. a group of 38 females and 31 males taking a minor in General Linguistics at the Free University in Amsterdam, a group of 29 female and 27 male first year students in Dutch at the University of Leyden, and a group of 9 females and 14 males taking courses in Dutch or English at the Catholic University in Nijmegen. All subjects were between 18 and 30 years of age. Thus the total group of raters consisted of 168 subjects, 96 of whom were females, and 72 were males. There were no relevant differences between the three subgroups.

The remaining subjects who participated in the paper-and-pencil test were students of logopedics, over 90% of which were females. A comparison of the mean scores of these 'all-female' groups with the corresponding mean scores of the females in the mixed-sex group showed, however, that there were hardly any relevant differences.

As can be seen from Table 2.2, the scores for ideal voice and articulation did indeed reveal marked differences between the sexes. Those differences corroborate the stereotypical opinion that females ought to appear more friendly, more expressive but less active, less powerful and less independent than males (Kramer, 1978). If both male and female speakers are to be judged this result might be of considerable importance, the more so if one assumes that semantic scales define a psychological continuum one anchor point of which is associated with the concept of the ideal value of the attribute. If that ideal value differs significantly for different groups or classes of objects the scale can no longer be said to have a well defined and stable meaning.

Not only the mean scores for the ideal male and female voices differed significantly on a considerable number of scales, there were also significant differences between the scores of female and male judges. Differences between ideal male and female voice ratings are much smaller in the scores of the female subjects than in those of the males. Females do not only find the distance between the ideal voice of the members of either sex smaller than the males, they also seem to make higher demands in that their scores for ideal voice are on the average closer to the extremes of the scales than the scores of the male raters. From this it can be concluded that one should be careful in combining the ratings of male and female judges, since the anchor points of the attribute continua might be dependent on the sex of the judge. In the actual rating experiment to be reported on below this sex-related bias did not pose a problem, because the overwhelming majority of the judges were females.

A more comprehensive treatment of the sex-related differences that we have come across can be found in Boves, Fagel & van Herpt (1982). For the present study it is more important to see whether the scales comply with the

Table 2.2. Results of t-test for differences between ratings of the 'Ideal male voice' and 'Ideal female voice'. Results are given for the total group of 168 raters and for the subgroup of 96 female raters. Underlined adjectives indicate the 'desirable' pole of the scales.

		SCORES OF MALES AND FEMALES (N=168)			SCORES OF FEMALES ONLY (N=96)		
		IV-♂	IV-♀	Dif.	IV-♂	IV-♀	Dif.
1	<u>pleasant</u>	1.35	1.20	0.15	1.11	1.11	0.00
2	<u>dull</u>	5.35	6.35	-1.00**** ⁴	5.33	6.40	-1.07****
3	<u>friendly</u>	1.75	1.47	-0.27*** ³	1.65	1.43	0.22
4	<u>loud</u>	3.27	4.49	-1.22****	3.26	4.33	-1.07****
5	<u>husky</u>	5.52	5.72	-0.20	5.23	5.57	-0.34*
6	<u>stereotyped</u>	6.00	6.14	-0.14	6.03	6.17	-0.14
7	<u>active</u>	2.12	2.02	0.10	2.09	1.99	0.10
8	<u>artless</u>	1.94	1.98	-0.04	1.82	1.75	0.07
9	<u>colourless</u>	6.30	6.37	-0.07	6.44	6.47	-0.03
10	<u>careless</u>	5.43	5.65	-0.22	5.39	5.68	-0.29
11	<u>low</u>	5.61	3.10	2.51****	5.74	3.38	2.36****
12	<u>firm</u>	1.88	2.57	-0.69**** ¹	1.67	2.30	-0.64****
13	<u>creaky</u>	6.20	6.46	-0.26*	6.27	6.54	-0.27
14	<u>dragging</u>	5.58	5.69	-0.11	5.77	5.84	-0.07
15	<u>melodious</u>	2.11	1.64	0.47****	1.78	1.47	0.31*
16	<u>powerful</u>	1.92	3.05	-1.13****	1.67	2.71	-1.04****
17	<u>ugly</u>	6.02	6.44	-0.42****	6.27	6.57	-0.30*
18	<u>rough</u>	4.27	5.44	-1.17****	4.23	5.40	-1.17****
19	<u>spiritless</u>	6.07	6.31	-0.24*	6.31	6.42	-0.10
20	<u>regional</u>	5.12	5.28	-0.16	5.07	5.27	-0.20
21	<u>rounded</u>	3.12	2.59	0.53****	2.96	2.41	0.55****
22	<u>deep</u>	2.04	3.63	-1.59****	1.81	3.39	-1.57****
23	<u>poor</u>	6.02	6.04	-0.02	6.22	6.24	-0.02
24	<u>tense</u>	6.13	5.92	0.21	6.23	5.97	0.26
25	<u>expressive</u>	1.70	1.66	0.04	1.47	1.50	-0.03
26	<u>steady</u>	1.71	2.10	-0.39****	1.58	1.86	-0.28*
27	<u>broad</u>	5.88	6.00	-0.12	6.12	6.27	-0.15
28	<u>selfconfident</u>	2.10	2.33	-0.23	2.06	2.16	-0.10
29	<u>whining</u>	6.08	6.36	-0.28***	6.32	6.51	-0.19
30	<u>polished</u>	1.96	1.87	0.09	1.71	1.66	0.05
31	<u>deviating</u>	5.76	5.89	-0.12	5.95	5.98	-0.03
32	<u>quick</u>	3.48	3.18	0.30*** ²	3.42	3.18	-0.24
33	<u>full</u>	1.85	2.42	-0.57****	1.62	2.13	-0.51****
34	<u>agitated</u>	5.59	5.01	0.58****	5.69	5.03	0.66****
35	<u>distinct</u>	1.51	1.56	-0.05	1.44	1.49	-0.05

1. $p < .05$
2. $p < .01$
3. $p < .005$
4. $p < .001$

requirements of linearity and relevance. From the very start of the experiment we expected the three physical reference scales ('loud $\leftrightarrow$ soft', 'high $\leftrightarrow$ low' and 'quick $\leftrightarrow$ slow') to perform badly in the ideal voice scores, as the ideal voices for members of both sexes are likely to be considered neither very loud nor very low or very quick. This expectation was confirmed with only one exception: the mean value for the ideal male voice on the scale 'high $\leftrightarrow$ low' was 5.61, which is rather close to the 'low'-extreme of the seven point scale. All other scores for ideal voice showed mean values lying in one of the three center intervals ($2.5 < \text{mean value} < 5.5$). It should be noted, by the

way, that the disambiguating addition 'for a female/male' was not added to the scale 'low $\leftrightarrow$ high' in the present experiment.

Since the majority of the raters in the actual judgement experiment were females, it is most interesting to look at the average values for the ideal male and female in the scores of the 92 female subjects. Concentrating on possible deviations from linearity of the scales, appearing from the location of the ideal somewhere in the middle of the scale, it can be seen that only four scales behave in the same way for the ideal male and the ideal female voice, viz. the scales 'loud $\leftrightarrow$ soft', 'regional $\leftrightarrow$ non-regional', 'rough $\leftrightarrow$ smooth' and 'quick $\leftrightarrow$ slow'. Not surprisingly, the remaining scales where either the ideal female or the ideal male is located towards the middle of the scale reflect stereotypes in the sense that the desirable adjective is more associated with one sex than with the other. For instance, it does not really seem to be desirable for women to sound strong and calm and to have a deep voice. On the other hand, male voices should not be too clear, but they may be somewhat husky and their articulation need not be very precise. These differences must be taken into account when interpreting relations between acoustic measures and perceptual ratings.

The scale 'rough $\leftrightarrow$ smooth' most probably has suffered from an inadequate translation of the adjective 'smooth'. The word '*glad*' in Dutch may have a strong negative connotation in the sense of 'slippery' that obviously could not be suppressed by its pairing with the negative term 'rough'. As we could not find an adjective that would do better we decided to drop this scale. 'Regional $\leftrightarrow$ non regional', a scale that might be somewhat confusing in this form, was changed into a new scale that was placed apart from the other scales; on the new scale the judges were asked to rate the degree to which a speaker has a regional accent.

Although the scale 'full $\leftrightarrow$ thin' did not appear to behave badly, there were some strong objections against it, mainly based on the doubt whether the word '*ie!*' was commonly understood. Since a considerable part of its function seems to be similar to that of the scales 'dull $\leftrightarrow$ clear', 'colorless $\leftrightarrow$ sonorous' and 'deep $\leftrightarrow$ shrill' and because we wanted to add some scales to probe the reading performance without increasing the total number of scales too much, we decided to dispose of the scale 'full $\leftrightarrow$ thin' and to replace it by two new scales, viz. one traditional bipolar scale 'jerking $\leftrightarrow$ smoothly flowing' and a special scale, presented at the very bottom of the new scoring form, on which the subjects were asked to rate the reading performance; this scale is very much like the scale on which the degree of accentedness is to be rated.

The scores that the 168 subjects gave for their own voice did not give rise to any further corrective action. All scales showed a fair degree of variance in the scores, showing that the scale attributes are relevant to voice and articulation and that in a normal population there is a considerable degree of variation with respect to the attributes.

2.5 CONCLUSION

In this chapter we have discussed a number of properties a semantic differential and the scales of which it consists must fulfil. Some of these requirements, notably the linearity and the relevance of the scales, are treated in some detail. For one part they are dealt with via a study of the literature and theoretical considerations, for the other part they have been approached by means of a paper-and-pencil test. Such a test was also carried out to ascertain that the way in which the scoring form was going to be presented to the raters would not invalidate the results of the rating experiment.

The efforts described in this chapter have resulted in the construction of an instrument consisting of 35 scales which may be considered as relevant to the stimuli to be judged and at the same time as fairly linear.

The evaluation experiment

3.0 INTRODUCTION

Having developed a semantic differential we could proceed to the actual evaluation experiment. The report on this experiment will be organized as follows: first we will describe the stimuli and the situation during the rating sessions, next we will report on the analysis of the individual scales and finally we will give the results of a factor analysis on the scores of our judges. The evaluation experiments were carried out in close cooperation with colleagues from the Institute of Phonetic Sciences of the University of Amsterdam. Because of this cooperation the experiments had to serve two purposes. The eventual aim of our research is to 'explain' perceptual ratings of speech by means of acoustic measurements. The procurement of perceptual ratings is a necessary methodological step towards that final goal. Thus for our purposes any set of ratings would do the job, provided it contains clearly identifiable subsets relating to articulation, phonation and socio- and paralinguistic aspects of the speech stimuli. The team at the University of Amsterdam aimed at the development of a maximally efficient and reliable instrument for obtaining global ratings of voice and articulation quality. Thus their major aim was a reduction of the number of scales in the semantic differential, an aim which, when reached, would facilitate our job of explaining perceptual ratings by means of acoustic measurements.

3.1 METHOD

3.1.1 *The preparation of the stimuli*

The stimuli to be judged were recordings of subjects reading a text. In order to enhance the comparability of the results we decided to share one reading text and two speakers between the projects in Amsterdam and Nijmegen. For the work done in Amsterdam both the text and the speakers were fixed, because it was decided there to replicate at least in part the work done before 1976. The most important consequence of that decision was that use had to be made of recordings made back in 1969 (Cf. Koopmans-van Beinum (1980) for a more detailed description of the speech material used in her

research; the reading texts that served as stimuli in the work done in Amsterdam had been recorded as part of her speech data base). From the five male (and five female) subjects in that data base only one, viz. DJ, was readily available for recording a new text, but he was not able to participate in the physiological measurements planned as part of our own experiment. Therefore we were more or less forced into the following construction: we have chosen five speakers, each reading two texts, to participate in our experiment, including the physiological measurements, to which we have added DJ as a sixth speaker in the evaluation experiment only; one of our five speakers was added as an eleventh speaker in the evaluation experiment carried out in Amsterdam. (Cf. chapter 2 for a motivation for the limitation of the judgment experiment to a maximum of 12 texts.) In doing so the possibility to compare the results of both experiments by constructing common scales on which speakers from either set could be assigned scale values was at least theoretically guaranteed.

The reading text used in the previous and present work in Amsterdam was adapted from a Journal published by the Dutch Tourist Club. It gives a summary account of a bicycle tour in Denmark and it is carefully edited in order to prevent any moral, ethical or political objections, even from the most sensitive of souls. This text may safely be considered as being of neutral content. It will be referred to throughout as the neutral text. One of the aims of our own work is to look after possible influences of the content of the reading text on judgements of non-content related aspects of the rendition. Therefore we decided to use a second text explicitly chosen to have a non-neutral content. This text is adapted from 'De Tribune' (The Tribune), a journal published by the Socialist Party, a small and decidedly very left wing political party that uses its journal as a medium to propagate its ideas rather than to disseminate so called objective information. The text gives an account of the situation in Iran shortly after the overthrow of the regime of the Shah and it does not eschew words like 'prison', 'porno movie', 'reactionary' and the like. The text was adapted at various places in order to make its grammar somewhat less complex and to enhance its readability.

Originally it was intended to make simultaneous registrations of the subglottal pressure, the photoglottogram and the electroglottogram together with the speech signal while our subjects read the texts (refer to chapter 4 for a detailed explanation of the physiological registrations), but due to the difficulties encountered in making the simultaneous recordings this plan proved to be impracticable. Because we thought that with a view to the acoustical processing envisaged in the other part of our research we could not tolerate the phase distortion inherent in direct recording onto analog tape, we settled for the following procedure: our subjects were seated in the terminal room next to the computer room (the same situation as during the physiological measurements (cf. Fig. 4.6)) with a Bruel & Kjaer type 41341/2" condensor microphone placed at a distance of not more than 5 cm from the lips and with

a Sennheiser type MKH 415 T condensor microphone placed at a distance of about 20 cm. The pick-up electrodes of a Fourcin and Abberton electroglottograph (Laryngograph) were attached to either side of the subjects' Adam's apple and held in place with the usual elastic band around the neck. The output signals of the Sennheiser microphone and the electroglottograph were recorded onto analog magnetic tape using a Revox A77 tape recorder. The signals of the B&K microphone and the electroglottograph were sampled with a frequency of 8 kHz each, digitized and fed into the computer that stored them onto its disc memory that provided room for 90 seconds worth of speech and electroglottograph signals. After 90 seconds of recording the contents of the disc memory had to be dumped onto digital magnetic tape, after which operation the recording could be resumed. This forced us to chop up the reading of the texts into a number of 90 seconds fragments separated by the time needed to dump the disc. All subjects were given ample time to study the texts before the beginning of the recording sessions and they were encouraged to use the time between the 90 second sections for that same purpose. The readers were told that they could freely restart sentences or parts thereof after having made a mistake or when they felt that they could do much better and almost all speakers made more or less heavy use of this possibility. The neutral text took on the average three 90 second sections to complete and the non-neutral text, that was somewhat longer, on the average four. The relevant characteristics of the speakers are given in Table 3.1.

The analog recordings were subsequently edited in order to remove as many reading errors as possible and to combine the fragments to coherent texts. Two stimulus tapes were prepared, one containing the renditions of the

Table 3.1. Speakers participating in the evaluation experiment.

Initials	Age	Regional background	Educational background
EK	40	Youth spent in the Southern part of the country. Living in Amsterdam since.	University level education.
DJ	37	Whole life in Amsterdam.	Vocational training as an instrument maker.
LB	34	Youth spent in the South. Living in Nijmegen since 1965.	University level education.
NR	37	Whole life in Nijmegen.	Vocational training as an electronic engineer.
TB	19	Youth spent in the South. Living in Nijmegen since 1979.	University student.
PD	20	Whole life in the Nijmegen region.	Student at a polytechnic school.

neutral text and the second with the non-neutral text. The reading of the neutral text by DJ was copied from the original 1969 recording; his rendition of the new, non-neutral text was recorded, together with the electroglottogram, using the facilities of the Institute of Applied Linguistics of the University of Amsterdam. There was, of course, no need to break up this recording session, but once again the speaker was encouraged to restart a sentence if he felt that the reading was not correct.

Two additional stimulus tapes were prepared for the Amsterdam experiment, each containing eleven readings of the neutral text, the difference between the tapes being the order in which the talkers were presented. Both tapes, however, always had a male speaker followed by a female.

3.1.2 *The collection of the ratings*

Ideally the scales in a semantic differential define a psychological continuum with securely anchored endpoints (defined by the opposite adjectives) and stable boundaries between the successive categories. If this conjecture applies, listeners should be able to place each speaker on the continuum regardless of whether he/she were presented in isolation or after one or more other speakers. In that case it would not be necessary to present samples of all voices to be judged before the actual ratings begin. Under less ideal circumstances there may still exist a psychological continuum, but without fixed endpoints and with more or less floating boundaries between categories. In this case the position of the endpoints and consequently those of the category boundaries are determined to a high degree by the stimuli that are maximally positive or negative with respect to the attribute alluded to by the scale. In other words, we must reckon with the possibility that the scale (rather than the psychological continuum) is adjusted to the stimuli. If this is the case, and authors like Guilford (1954) appear to doubt whether the ideal fixed endpoint situation can exist in actual practice, it is obviously necessary to precede the actual ratings by samples of all stimuli to be rated in order to enable the judges to fix the endpoints of the scales. We decided to stay on the safe side and therefore preceded all tapes by a presentation of speech samples of all speakers. This was done by splicing segments from the recordings of the speakers in the reverse order of presentation in such a way that the first part of the text was produced by a succession of speakers, each reading a sentence or a part of a paragraph. The listeners were instructed to listen carefully to this compilation and were explicitly told that it served to give them an impression of the variety of stimuli to be judged.

The perceptual ratings we are interested in are fairly global and therefore should not be dependent on slight variations in the experimental conditions during the rating experiment. For this reason it was thought to be warranted to organize the ratings in the form of classroom sessions, during which a group of judges listened to the speech samples to be judged. The recordings

were played back on a Revox A77 recorder and a pair of high quality loudspeakers.

The listeners got the usual instructions in administering a semantic differential i.e., that it is their first impression that is sought and that they are not supposed to worry or grumble over individual scales. They were allowed a maximum of four minutes to complete the scoring form for each stimulus. As the texts did not last for such a long time, the recordings were repeated to fill the time. Thus the entire rating was done while listening to the stimuli.

Prior to the presentation of the tapes the listeners were familiarized with the scales and the rating procedure by scoring the semantic differential for three concepts, viz. 'my own voice and articulation', 'the ideal voice and articulation of a male' and 'the ideal voice and articulation of a female', always in this order. Each rating session lasted about an hour; the instruction and familiarization with the procedure and the scales took approximately 10 minutes, whereas the actual rating task took almost 50 minutes. The listeners were not paid for their participation. The rating sessions replaced normally scheduled classes. This organization made it impossible to have a single group rate both sets of stimuli. The stimuli in the Amsterdam experiment were rated by groups, comprising a total of 235 judges. For the details of these rating sessions the reader is referred to Fagel, van Herpt & Boves (1983).

Four groups, comprising a total of 117 judges, participated in the evaluation experiment with the Nijmegen data. All judges were students of logopedics at three different schools in three different parts of the country. The first group consisted of 29 first year students in Amsterdam, the second and third were from a school in Nijmegen, a first year group of 24 and a second year group of 28 students, whereas the fourth group was comprised of 36 first year students of a school in Hoensbroek, in the extreme South of the country. All rating experiments were carried out during the first trimester of the school year 1980/81, so that the first year students should be considered as almost naive with respect to the formal and technical aspects and terminology of speech science.

As the present author was one of the six speakers to be rated he could, of course, not participate as an experimenter during the rating sessions. All practical work has been done by staff members of the Institute of Phonetics of the University of Amsterdam, who also took care of the coding and punching of the scores.

3.2 ANALYSIS OF THE INDIVIDUAL SCALES

This section contains a fairly detailed analysis of the formal scaling properties of the 35 bipolar rating scales making up our semantic differential. The aim of this analysis is to identify (and eliminate) those scales which flagrantly

run counter to the requirements of the factor analysis model. We realize that a detailed analysis of the scaling properties preceding a factor analysis is somewhat unusual. However, the work was motivated by the finding of Blom & van Herpt (1976-a) that even a small number of strongly 'misbehaving' scales may affect the results of a factor analysis considerably, the very robust character of the factor model notwithstanding.

3.2.1 *Introduction of the scaling model*

The analysis of the individual scales is mainly based on the 'Law of Categorical Judgement' as it was derived by Thurstone. This 'law' can be considered as a method for converting scores on ordinal rating scales to scale values for the rated objects on interval scales. The widths of the intervals on the original rating scales do not have to be equal or equal-appearing, but they must be in a rank-order relationship, i.e., if we number successive categories from 1 to N , category $n+1$ is supposed to represent a higher degree of whatever attribute is rated on the scale than category n . Comprehensive and very readable treatments of both the 'Law of Categorical Judgement' and Thurstone's General Judgement Scaling Model which underlies it are available in the literature (Torgerson, 1958). Thus it is not necessary to give a detailed explanation here. In very general terms it can be said that the assignment of a stimulus to a category on an ordinal scale is a random process. If each stimulus is rated a large number of times, the ratings assume a normal distribution. The location of the mean on the scale is called the scale value of the stimulus, the standard deviation of the distribution is called the discriminial dispersion. Every stimulus has its own scale value and discriminial dispersion, which are most likely different from the scale values and discriminial dispersions of all other stimuli. The boundaries between adjacent categories on the ordinal scale behave in much the same way as the stimuli. Since the widths of the categories do not have to be equal, and since their positions on the scale are not known beforehand, category boundaries have to be scaled at the same time as the stimuli. This model of the scaling process leads to a set of formal mathematical equations that cannot be solved because the number of unknowns is inherently larger than the number of independent equations. This dilemma is solved by assuming that both the covariance of the stimuli and the category boundaries and the discriminial dispersions of the category boundaries are constant. These assumptions lead to what is called the Condition B case of the Law of Categorical Judgement (Torgerson, 1958:209), which is formally described by the set of equations

$$t_g - s_j = x_{jg}a_j \quad (1)$$

$$j = 1, 2, \dots, J$$

$$g = 1, 2, \dots, G$$

where

J : number of stimuli rated

G : number of boundaries between adjacent categories
(i.e., we have a $G+1$ point scale)

t_g : the scale value of the g -th boundary

s_j : the scale value of the j -th stimulus

a_j : $\sigma_j + c$, i.e., the discriminial dispersion of the j -th stimulus plus an added constant which accounts for the covariance of stimuli and boundaries plus the dispersion of the boundaries

x_{jg} : unit normal deviate corresponding to the proportion of time that stimulus j is rated below the g -th category boundary

Note that the left hand side of equation (1) can be interpreted as the distance between the scale value of the j -th stimulus and the position of the g -th category boundary. The distances are only defined as long as the x_{jg} exist, which is not the case if $p_{jg} = 0$ or $p_{jg} = 1$ (where p_{jg} denotes the proportion of times that stimulus j is sorted below the g -th boundary) as the z -values corresponding to proportions of 0 and 1 are infinite. With its $2J+G$ unknowns and $J \cdot G$ observation data x_{jg} the set of equations in (1) is overdetermined. The estimation of the unknowns from the observations, therefore, entails some averaging or minimization procedure.

A number of authors have proposed methods for solving (1) for the t_g , s_j , and a_j . The methods differ from each other in the way the data are used and averaged. We had at our disposal a computer program that implements a generalized version of the least-squares minimization solution originally due to Gulliksen (Blom & van Herpt, 1976-b). In that method the error made in estimating the t_g from the observed quantities x_{jg} is minimized in a mean squares sense. The problems arising as a result of the incompleteness of the data matrix X (due to proportions p_{jg} which are either 0 or 1) are dealt with by the use of Müller-Urban weights (Boves & Cranen, 1982-a).

Thurstone's judgement model is based on a number of assumptions that need not all be fulfilled by the data. Thus the problem arises of the appropriateness of the model, or the goodness-of-fit of the model to the data. According to Torgerson a statistically sound test for the goodness-of-fit of the categorical judgement model does not exist. He mentions an index proposed by Edwards and Thurstone, based on the sum of the absolute discrepancies between the observed proportions p_{jg} and the estimated $\hat{p}_{jg}$ which can be obtained from the model by a simple rearrangement of the terms in (1):

$$\hat{p}_{jg} = \frac{1}{\hat{a}_{jg}} (t_g - s_j) \quad (2)$$

and the subsequent conversion of the estimated unit normal deviate $\hat{x}_{jg}$ to the corresponding estimated proportion $\hat{p}_{jg}$. The statistical distribution of the summed absolute discrepancies is not known, so that it cannot be tested whether it differs significantly from 0. Guilford's proposal to apply the χ^2 test, originally developed by Mosteller in the framework of Thurstone's model of comparative judgement to the discrepancies $p_{jg} - \hat{p}_{jg}$ is rejected, since neither the observed nor the estimated cumulative proportions p_{jg} and $\hat{p}_{jg}$ are mutually independent, p_{jg} always being greater than or equal to p_{jg-1} .

It is, however, not difficult to decumulate the estimated proportions $\hat{p}_{jg}$ and form the set of $G+1$ proportions $\hat{f}_{jg}$ with which stimulus j is placed in category g . The $\hat{f}_{jg}$ are mutually independent, as are the original observed proportions f_{jg} , with g degrees of freedom. Bock & Jones (1968) use the discrepancies $f_{jg} - \hat{f}_{jg}$ to form the total χ^2

$$\chi^2 = \sum_{j=1}^J \sum_{g=1}^G \frac{\{(f_{jg} - \hat{f}_{jg})N_j\}^2}{(f_{jg} - \hat{f}_{jg})N_j} \quad (3)$$

where N_j is the number of judges who have rated stimulus j . The number of degrees of freedom is $(J-1).(G-2)$, i.e., $J.G$ independently observed proportions minus $2(J-1)$ degrees of freedom consumed by the estimates s_j and a_j not depending on t_g , and G degrees of freedom consumed by the estimated t_g . We have included the computation of this measure in our program. Incidentally, Bock and Jones use the total χ^2 as the expression which is to be minimized in order to obtain the optimal solution for the t_g , s_j and a_j , starting from an approximate solution obtained by means of a fairly simple graphical method. We felt, however, that for the purpose of the present study the implementation of the minimum normit χ^2 solution could not be warranted.

The χ^2 test of Bock and Jones gives an answer to the question whether the Thurstone model does or does not fit the data. If the result of the test is negative, one does not get easy access to an explanation of the exact causes of the deficiency of the model. Some additional insight may be obtained from a test originally suggested by Blom (personal communication). This test, which is based on the computation of values of the midpoints of the categories on the scales rather than their boundaries, is very likely to fail if the scale attribute is multi-dimensional in nature. If the attribute is one-dimensional the midpoints of the categories must preserve their ordinal relation. If the scale values of the category midpoints do not come in a monotonically increasing order, a possibility which is acknowledged in Torgerson (1958:208), there is reason to assume the scale attribute to be multi-dimensional.

Computationally, the test is related to a method for solving the Law of Categorical Judgement, described in Guilford (1954:237 ff.) but which does not seem to be very popular. This method starts with an estimation of the

scale values of the category midpoints. Having computed midpoint values for the categories we can very easily arrive at scale values for the objects that have been rated. Those scale values are defined as the mean value of the distributions of the scores on the psychological continuum, so that we can estimate the s_j from the midpoint values of the categories, denoted by c_g , $g = 1, 2, \dots, G+1$, and the relative frequency f_{jg} with which object j is sorted into the g -th category, by the simple expression

$$\hat{s}_j = \sum_{g=1}^{G+1} c_g f_{jg} \quad (4)$$

In our situation we have computed values for s_j and, of course, observed f_{jg} data. We can use those data to arrive at estimates for the category midpoints. Clearly this leads to another minimization problem because we have $J+J(G+1)$ known values for the computation of $G+1$ unknowns. The approach is straightforward. Equation (4) gives an estimate of the scale values of the objects based on estimates of the values of the category midpoints. Taking the computed s_j as true values we choose the c_g in (4) such that the error in estimating s_j by the $\hat{s}_j$ in (4) is minimum in a mean squares sense. If we denote the sum of squared errors by Q , our task is to find the values of c_g that minimize

$$Q = \sum_{j=1}^J (s_j - \hat{s}_j)^2 = \sum_{j=1}^J (s_j - \sum_{g=1}^{G+1} c_g f_{jg})^2 \quad (5)$$

Differentiating (5) with respect to each c_g and equating the partial derivatives to zero gives a set of $G+1$ equations of the form

$$\frac{\partial Q}{\partial c_k} = 2 \sum_{j=1}^J (s_j - \sum_{k=1}^{G+1} c_k f_{jk}) f_{jk} \quad (6)$$

Since the f_{jg} are proportions we know that $0 \leq f_{jg} \leq 1$ and that not all $f_{jg} = 0$. Thus it follows that in order to satisfy (6) we must have

$$\sum_{j=1}^J (s_j - \sum_{k=1}^{G+1} c_k f_{jk}) = 0 \quad (7)$$

which is, in fact, a set of $G+1$ linear equations in $G+1$ unknowns that can be solved using any subroutine for the solution of a general set of simultaneous linear equations.

3.2.2 Results

In this section we will present the results of the analysis of the linearity of the scales and the analysis of the goodness-of-fit of the Categorical Judgement model.

It should be realized that the scores obtained in the rating experiments do not only reflect properties of the scales. In our scores these properties are contaminated with characteristics of the stimuli and idiosyncrasies of the raters. All analyses have been performed for all subgroups of raters separately, but there never appeared a group effect that seemed worthy of being retained. Therefore, all analyses to be described in this subsection have been carried out on the combined groups of raters.

It is, of course, not possible to combine the ratings from the Amsterdam and Nijmegen stimuli. Even if such a combination were possible, it would not be advisable. Studying both sets of ratings in parallel gives at least some means for identifying effects which can with some degree of confidence be attributed to the scales instead of to the stimuli or the raters, viz. those effects that are similar in both data sets.

3.2.2.1 *The linearity of the scales*

For the rating scores to fulfill the requirements of the factor analysis model it is most important that the correlations between the scales are linear. This requirement will be fulfilled if the categories on the scales are of equal widths. Linear correlations are also guaranteed, however, if all scales deviate from the ideal equal interval scale in about the same way.

One way to approach this problem is by cross-correlating the widths of the categories of all 35 scales, which results in $35 \times 34 / 2$ correlation coefficients, the average value of which is an indication of the overall resemblance of the scales. Another approach, which enables us to identify individual scales which behave differently from the rest, is to compute the correlation of the widths of the categories on each scale with the widths of categories on an ideal equal interval scale. Since we have computed both scale values for the category boundaries and the category midpoints, two correlation measures for the linearity of the scales can be obtained. These measures are given in Tables 3.3 and 3.4 for the Amsterdam and Nijmegen material respectively. Table 3.2 contains estimates of the category widths, obtained by averaging the distances between the boundary or midpoint values on all 35 scales. Before the averaging was done, the scales running from positive to negative were reflected. Also, the values were normalized before averaging in such a way that the midpoint of the leftmost category is assigned a scale value of 0 and the rightmost a scale value of 100.

From Table 3.2 it can be seen that both in terms of distances between category boundaries and category midpoints the scales seem to be, on the average, neatly symmetric with respect to the middle category but not especially linear: the middle categories have only half the width of the

Table 3.2. Average widths of the categories on the successive interval scales as determined from the Law of Categorical Judgement. Part A contains the distances between the six category boundaries. Part B the distances between the seven category midpoints.

A	$t_2 - t_1$	$t_3 - t_2$	$t_4 - t_3$	$t_5 - t_4$	$t_6 - t_5$
Amsterdam (N=235)	16.04	13.09	8.20	12.32	18.57
Nijmegen (N=117)	18.01	14.96	7.81	14.13	19.34

B	$c_2 - c_1$	$c_3 - c_2$	$c_4 - c_3$	$c_5 - c_4$	$c_6 - c_5$	$c_7 - c_6$
Amsterdam (N=235)	23.93	13.12	10.73	11.45	10.64	30.13
Nijmegen (N=117)	22.66	13.63	12.10	10.65	14.92	26.04

categories at the negative and positive extremes. These results are for the major part in line with the findings of Messick (1955) although the 'end effect', i.e., the increased width of the extreme categories as compared with the central ones is more symmetric in our results than in Messick's, who found the effect to be greater at the positive end of the scales. We know of no other investigations into the behaviour of semantic differential scales by means of the Law of Categorical Judgement so that we do not have a solid background to evaluate Messick's and our own results.

The average correlations of the scale values of the category boundaries with ideal equidistant values are extremely high: .991 for the Nijmegen scores and .992 in the Amsterdam data. No individual scale shows a correlation with an ideal equal interval scale that is lower than .972. This is exactly what should be expected since the boundary values are forced into an increasing order, so that the data have a rank-order correlation equal to 1. Looking at the data for the category midpoints we see that the overall picture is similar, but that now some scales appear to form exceptions most notably the scale 'distinct ↔ indistinct' in the Nijmegen material where the correlation is as low as .777. Such a low correlation suggests that there is no longer a perfect rank-order correlation with equidistant ideal values which indeed is true. Inspection of the midpoint data shows that the Nijmegen material contains four scales on which the scale values of the category mids do not form a monotonically non-decreasing sequence, viz. the scales numbered 13, 33, 34 and 35 ('creaky ↔ not creaky', 'distinct ↔ indistinct', degree of accentedness and reading performance). In the Amsterdam material no less than nine scales appear to have the category midpoints in a non-monotonic order viz. the scales numbered 1, 5, 23, 26 and 30 in addition to the same scales which appeared to be non-monotonic in the Nijmegen material. The much larger number of non-monotonic scales in the Amsterdam material explains why the pattern of average category widths determined from the scale values of the category midpoints is less symmetric in this data set than in the Nijmegen material.

The problem of establishing whether a judgement scale is unidimensional and monotonic still seems to be open. Nishisato (1980) describes a scaling technique which can be applied to categorical judgement data, the solution of which can be forced to provide monotonically non-decreasing scale values for the category boundaries as an 'optimal' solution. His model, which differs in some fundamental ways from Thurstone's general judgement model, scales the raters in addition to the category boundaries and the stimuli. It can produce several 'non-optimal' solutions, which in practice have been found to yield non-monotonic scale values for the category boundaries quite often. Nishisato suggests that his model can be extended so as to enable an analysis of the multidimensionality of the judgement space, but leaves this as 'an interesting problem for further research'.

Leaving the topic of the multidimensionality of the scales, it is time to draw a preliminary conclusion. The results up to now suggest that the rating scales can be treated as equal interval scales for all practical purposes. Indeed, Boves & Cranen (1982-a) computed correlations between the scale values of the stimuli on all scales and simple averages of the raw scores which procedure assumes equal interval scales. In the Nijmegen data the correlation coefficients exceeded .99 for all scales except 'artless $\leftrightarrow$ affected', 'creaky $\leftrightarrow$ not creaky', and the degree of accentedness scale. In the Amsterdam material only the scales 'husky $\leftrightarrow$ not husky' and 'distinct $\leftrightarrow$ indistinct' showed correlations between .97 and .99; all other scales had correlations $> .99$. These results confirm the contention that the raw scores may safely be treated as interval level data.

3.2.2.2 *The goodness-of-fit of the model*

The goodness-of-fit of the model to the data has been treated via the total χ^2 proposed by Bock and Jones. The results are summarized as part of the Tables 3.3 and 3.4. Using a significance level of .01 three scales in the Nijmegen and 13 others in the Amsterdam material appear not to fit into the categorical judgement model. There does not seem to be a systematic relation between linearity of the scales and goodness-of-fit of the model.

The fact that so many more scales resist accurate scaling in the Amsterdam material than in the Nijmegen data, cautions one to realize that the results cannot be completely due to properties of the scales themselves. Rather one must assume an interaction between properties of the scales, characteristics of the stimuli and characteristics of the rater population. The last two items differ between the data sets. In the Amsterdam data we have both male and female talkers and 235 raters, whereas the Nijmegen data pertain to 117 judges, rating speech samples from male speakers only.

It is interesting to note that the scales for which the model does not fit in the Amsterdam material can be divided into two broad categories, one comprising scales related to voice pitch and another mostly containing scales which can be interpreted as paralinguistic. It is quite conceivable that these

Table 3.3 Summary of the formal properties of the 35 rating scales in the Amsterdam experiment. The columns labelled 'Boundaries' and 'Midpoints' contain the correlations of the scale values with ideal equidistant values. 'Chi-square' and 'Probability' refer to the total chi-square measure for the goodness-of-fit of the model as defined by Bock and Jones, $df=50$. The column 'Reliability' lists Blom's R^2 .

Scale #	Boundaries	Midpoints	Chi-square	Probability	Reliability
1	.983580	.948290	51.472980	.105610	.496980
2	.985430	.985860	81.874400	.000110	.478040
3	.991040	.980640	116.623800	.000000	.315710
4	.996580	.992150	46.217860	.230920	.342510
5	.982680	.930190	65.236630	.007090	.545970
6	.986600	.977460	94.286500	.000000	.332320
7	.992860	.984360	187.773000	.000000	.341340
8	.995030	.968770	105.123100	.000000	.230710
9	.989860	.988700	84.289520	.000050	.446970
10	.991580	.985950	54.510940	.062750	.420750
11	.998480	.994850	110.594800	.000000	.353330
12	.990970	.988410	47.659970	.189260	.362010
13	.984650	.912420	43.141240	.338500	.335400
14	.996380	.993150	56.426910	.044130	.218000
15	.988000	.986920	53.695600	.072500	.437600
16	.993580	.988250	65.878750	.006130	.343700
17	.992650	.979830	148.517600	.000000	.535540
18	.992100	.969190	51.685130	.102000	.324520
19	.993870	.981660	27.728380	.928910	.439320
20	.998090	.993710	45.404920	.256920	.288760
21	.997960	.990320	61.208130	.017030	.515240
22	.995110	.989850	31.332630	.834830	.455330
23	.989770	.961090	67.086730	.004640	.356580
24	.988420	.980630	67.691470	.004030	.452770
25	.993980	.991960	47.972080	.180990	.405410
26	.990730	.974860	59.063280	.026440	.501900
27	.995540	.990460	78.689760	.000250	.405070
28	.996950	.992830	89.316560	.000010	.425370
29	.987710	.970470	57.799120	.033940	.506290
30	.993920	.926810	37.704070	.574090	.418030
31	.995530	.998200	58.995790	.026800	.267330
32	.994390	.988230	40.700010	.439460	.397040
33	.987070	.819840	46.024050	.236960	.467350
34	.985090	.814010	46.108490	.234320	.484120
35	.993070	.965240	57.314610	.037270	.515420

results are at least in part due to some uncertainty among the raters with respect to the question whether these attributes apply in the same way to males and females. A scaling procedure in which the raters are scaled along with the stimuli might have provided very illuminating results here.

3.2.3 *The reliability of the ratings*

A reliability coefficient expresses the probability that a repeated observation of some phenomenon will lead to essentially the same result as the observation under analysis. Obviously, reliable observations (measurements) are of

Table 3 4 Summary of the formal properties of the 35 rating scales in the Nijmegen experiment The columns labelled 'Boundaries' and 'Midpoints' contain the correlations of the scale values with ideal equidistant values 'Chi-square' and 'Probability' refer to the total chi-square measure for the goodness-of-fit of the model as defined by Bock and Jones, $df=55$ The column 'Reliability' lists Blom's R^2

Scale #	Boundaries	Midpoints	Chi-square	Probability	Reliability
1	.980490	.973490	33.447080	.876440	.431880
2	.985810	.979490	63.502330	.028610	.392570
3	.991550	.989510	63.277850	.029860	.234880
4	.991740	.983190	56.170120	.103190	.321940
5	.972050	.970230	49.511090	.262700	.317150
6	.994150	.995520	34.584430	.844550	.130880
7	.994000	.994820	46.754350	.359990	.196770
8	.994780	.983000	67.019410	.014230	.033310
9	.993790	.992470	37.999100	.725530	.198350
10	.993630	.987860	39.866300	.649340	.330120
11	.998890	.998210	57.737840	.080100	.382570
12	.991070	.994630	47.426940	.334710	.353080
13	.981730	.879500	44.201630	.463120	.109670
14	.995140	.991200	39.464660	.666180	.306160
15	.988410	.987510	49.010130	.279090	.231040
16	.993390	.991060	45.374330	.414510	.346740
17	.993550	.988710	51.905550	.192940	.383490
18	.990030	.979050	65.021730	.021290	.460920
19	.993870	.987090	43.062040	.511740	.203600
20	.996500	.989990	46.928530	.353360	.267270
21	.998740	.995270	44.760880	.439720	.465420
22	.998920	.995330	58.436980	.071270	.277210
23	.983700	.975790	49.246050	.271300	.328460
24	.994850	.986110	46.698580	.362130	.220650
25	.991750	.988040	111.090700	.000000	.369130
26	.990110	.991620	34.970790	.832760	.356520
27	.989670	.989360	44.053280	.469390	.393360
28	.996060	.996970	54.990430	.123850	.236290
29	.992710	.989790	60.351530	.051160	.348280
30	.993180	.938350	40.954830	.602900	.294600
31	.998380	.995030	47.721880	.323930	.263700
32	.993740	.987390	65.226700	.020440	.270770
33	.989400	.776630	70.580020	.006680	.318180
34	.975880	.902960	72.005460	.004880	.557790
35	.983060	.973190	56.025700	.105560	.451750

paramount importance. Unfortunately, however, the intuitively clear and simple concept of reliability cannot be given an unequivocal formal definition. Asendorpf and Wallbott (1979) review the literature on reliability coefficients. Their inventory of coefficients which are suitable for use with scores on successive interval scales consists of no less than nine competitors. Virtually all coefficients are based on some ratio of between stimulus variance and pooled within stimulus variance and/or residual variance. From this observation two conclusions can be drawn. Firstly, there are two contributions which may act together or independently in order to improve or

decrease the reliability: increasing the between stimulus variance will act to enhance the reliability and so will a decrease of the within stimulus variance. In other words: a rating (or any other measurement) will yield results which are easier to reproduce if the objects under measurement differ a great deal and if the measurement procedure as such is very accurate. Secondly, the use of within and between stimulus variance suggests that the measurement must at least be on the level of interval scales. Now it has been found that our rating scales, though not necessarily a priori on the level of interval measurement, can safely be treated as interval scales. Thus it is legitimate to compute the complete array of reliability measures collected by Asendorpf and Wallbott.

Blom (1980) took a somewhat different approach to the topic of the reliability of ratings on successive interval scales. He defined reliability as the proportion of variance in the raw scores that is accounted for by the Law of Categorical Judgement. The originality in Blom's proposal lies in the fact that he treats the variance in the scale values assigned to the rated objects as the explained variance, whereas the total variance is approximated as the variance in the scale values plus the sum of the squared discriminial dispersions. In formula form:

$$R^2 = \frac{\text{var}[s_j]}{\text{var}[s_j] + \sum_{j=1}^J \sigma_j^2} \quad (8)$$

This formula can be simplified by noting that the s_j are values on an interval scale, a scale on which a zero point can be chosen arbitrarily. Therefore we may shift the scale values s_j in such a way that their mean becomes zero and the variance can then be expressed as

$$\text{var}[s_j] = \sum_j s_j^2 \quad (9)$$

leading us to the extremely simple formula

$$R^2 = \frac{\sum_{j=1}^J s_j^2}{\sum_{j=1}^J s_j^2 + \sum_{j=1}^J \sigma_j^2} \quad (10)$$

Fagel (personal communication) has pointed out that R^2 , the reliability measure as derived by Blom, exactly parallels the more familiar reliability coefficient η^2 that emerges from analysis of variance schemas and that is defined as

$$\eta^2 = \frac{MS_{\text{between}} - MS_{\text{within}}}{MS_{\text{between}} + (n - 1) MS_{\text{within}}} \quad (11)$$

n the number of raters

and which is called the unadjusted intraclass correlation in Asendorpf and Wallbott. The major difference between Blom's R^2 and η^2 is that R^2 is based on model parameters whereas η^2 is based on observed scores. We have computed both R^2 and η^2 for all scales in the Amsterdam and the Nijmegen material. The results appeared to be virtually identical. This finding may be taken as a proof that the substitution of model parameters for observed scores is permissible, even if the model does not fit the data very closely.

The results of applying the R^2 formula to our scores are given in tables 3.3 and 3.4. The values look disturbingly low at first sight, but it should be realized that R^2 gives the mean reliability of a single judge. We can transform this measure into an 'effective reliability' for a greater number of judges in the same way Rosenthal (1973) transforms mean intrajudge correlations into an 'effective reliability':

$$R^2_{n,n} = \frac{nR^2_{1,1}}{1 + (n - 1) R^2_{1,1}} \quad (12)$$

where n is the number of raters and $R^2_{n,n}$ denotes the effective reliability for a group of n raters. The reason to give $R^2_{1,1}$ instead of the much more impressive $R^2_{n,n}$ (which, due to the large numbers of raters in the two experiments, always exceeds .8) is that it enables us to estimate the number of raters one should employ to reach some specified minimum level of effective reliability. It is quite obvious that for all but the worst behaving scales a satisfactory degree of reliability (somewhere near .9) can be reached with no more than about 30 judges.

A detailed examination of the results for the least reliable scales suggests that the low R^2 values are primarily due to the small between-speaker variance. The lack of reliability thus seems to be a result of the lack of true differences between the speakers that were to be rated rather than of the inability of the listeners to rate the speakers in a stable way.

3.2.4 *Conclusions from the scale analysis*

In this section an extensive analysis has been presented of a number of formal properties of the successive interval scales contained in our semantic differential. We have investigated the linearity of the scales, the degree to which the raw scores can be treated as interval data and the reliability of the ratings. From the results of the analysis it appears that the effort invested into the construction of the rating instrument has paid off: virtually all scales appear to behave well.

The single most important instrument used in the analysis of the scales is Thurstone's Law of Categorical Judgement, a formal technique for converting scores on successive category scales to values on interval scales. One of the by-products of the work (actually, of course, a prerequisite) has been the construction of a versatile Fortran program for performing the analyses. Based on the program by Blom & van Herpt (1976-b), our extended version does not only compute scale values of objects and category boundaries and discriminial dispersions of objects, but also a χ^2 test for the goodness-of-fit of the model to the data, an η^2 -like measure of reliability of the ratings and two measures of the linearity of the scales, one based on the scale values of the category boundaries, the other based on the scale values of the category midpoints. The latter measure suggests that some scales allude to an underlying feature which is multidimensional, resulting in a non-monotonic ordering of the category midpoints. It is interesting to note that the non-monotonicity was especially prominent in such popular scales like 'distinct $\leftrightarrow$ indistinct' and degree of accentedness. Owing to the lack of a sound theory of the dimensionality of rating scales it was decided to confine ourselves to signalling the potential problems here and not to remove the scales in question from the instrument.

With respect to the linearity of the scales we have found that the extreme categories of the rating scales seem to be approximately twice as wide as the central categories, a finding which confirms results from the literature. Moreover, it appeared that, on the average, the scales were neatly symmetric around their midpoint. Since all scales behave in an essentially similar way, the correlations between the scales can be assumed to be linear. This result frees the way for an unburdened application of factor analytic techniques to our data.

However paradoxically it may seem with a view to the finding that the central categories of the scales are only half as wide as the extreme ones, it was found that the averages of the raw scores correlate almost perfectly with the scale values obtained from an application of the Law of Categorical Judgement. This finding permits us to treat the raw scores as interval data.

Last but not least the reliability of the ratings was analyzed. It appeared that all scales could be scored reliably in the environment in which they were intended to be used, viz. in a semantic differential, which is never used with small groups of judges. In the few cases where the reliability seemed to be so low as to be hardly acceptable it could be proved that the problems were primarily due to the lack of variation in the stimulus materials.

3.3 THE EFFECT OF THE TEXTS ON THE RATINGS

In sociolinguistic research it has been found that the attribution of personality characteristics to talkers can be dependent upon the content of the speech, even when the judgements are based on texts read aloud (Giles & Powesland,

1975). Due to the research strategy which is normally adopted, it is not possible to decide whether this dependence results solely from the contents of the texts or rather from (context related) differences between the renditions. In order to gain some insight into this problem we decided to use two quite different texts, one of neutral and the other of clearly non-neutral content. Readings of both texts by the same subjects have been rated. Therefore, we are in the position to see if there are formal differences between the renditions, especially on the scales that are relevant with respect to para- and sociolinguistic features. We have handled this problem by carrying out a two-way analysis of variance on the raw scores of the Nijmegen material with speakers and texts as factors. Table 3.5 summarizes the results for the scales where the main factor 'text' proved to be significant. The interaction speakers $\times$ texts is significant on almost all scales. From the scales on which 'texts' is a significant main factor only number 34 (degree of accentedness) does not show a significant speaker $\times$ texts interaction.

The scales for which the texts differ significantly make the impression of a somewhat random selection. The most apparently personality-stressing scales ('friendly $\leftrightarrow$ curt', 'active $\leftrightarrow$ passive', 'artless $\leftrightarrow$ affected', 'powerful $\leftrightarrow$ weak', 'tense $\leftrightarrow$ relaxed', 'selfconfident $\leftrightarrow$ wavering') are, moreover, conspicuously absent from the list. In their place physical reference scales like 'high $\leftrightarrow$ low pitch' and 'quick $\leftrightarrow$ slow' appear.

The sign of the difference between the mean values of both texts (averaged over all talkers) is not always easy to explain. One could imagine that the non-neutral text would give rise to somewhat more positive ratings on scales

Table 3.5. Significance levels of F-values in a two-way analysis of variance to investigate the effect of the reading texts. The number of levels in the factor 'text' is 2; the factor 'speakers' has 6 levels and each cell contains 117 observations. Only those scales are given for which the factor 'text' is significant at the .1 level.

Scale		Speakers	Texts	Speakers $\times$ Texts
1. pleasant	unpleasant	.001	.001	.001
10. careless	precise	.001	.001	.013
11. high	low	.001	.008	.027
14. dragging	brisk	.001	.001	.001
17. ugly	beautiful	.001	.002	.002
18. jerky	smooth flowing	.001	.001	.001
22. poor	rich	.001	.033	.001
24. expressive	expressionless	.001	.020	.001
30. deviating	normal	.001	.035	.004
31. quick	slow	.001	.001	.001
34. accentedness (high/low)		.001	.001	.693
35. reading performance (good/bad)		.001	.003	.001

like 'pleasant $\leftrightarrow$ unpleasant' and 'ugly $\leftrightarrow$ beautiful' but it is not quite clear why it should influence tempo ratings in the direction of less quick and brisk.

There are several ways in which one could try to find explanations for the observed differences. One explanation, which cannot be easily verified, is that they are simple order effects, due to the fact that the neutral text has always been rated first. Another explanation, suggested by the fact that in all but one case there is a speakers $\times$ texts interaction is that the significant effect of the factor 'texts' is really no more than an artifact of unusually large differences in the scale positions between the two readings of one speaker. We will return to this issue in 3.4.4 below.

The analysis of variance produced global results with respect to systematic differences between the texts. However, it also showed many significant text $\times$ speaker interactions. The latter result suggests that at least some talkers have performed differently when reading one text than when reading the other, but that this difference in performance is not the same for all subjects. It seems worthwhile to investigate this topic into some depth. Blom (1980) developed an approach to assess whether the scale values of two stimuli differ significantly. His test is based on the fact that the scale values s_j of the J speakers can be interpreted as mean values of 'populations' of scores and that the associated discriminial dispersions σ_j are the standard deviations of the distributions. We can now apply the basic theorem which states that the variance of the sum or difference of samples taken from two independent distributions with variances σ_j^2 and σ_k^2 equals $\sigma_d^2 = \sigma_j^2 + \sigma_k^2$. Therefore the difference between the scale values of two speakers, j and k , has an expected value of $|s_j - s_k|$ and a variance $\sigma_{jk}^2 = \sigma_j^2 + \sigma_k^2$. The standard error of the estimate is $\sigma_{jk}/\sqrt{N}$, where N is the sample size on which the estimate is based. Under the Null-hypothesis that the scale values of speakers j and k do not differ, the probability that the observed difference exceeds that standard error by a certain factor is equal to the area under the unit normal curve up to the corresponding z -values: this leads to the test

$$z = \frac{|s_j - s_k|}{\sigma_{jk}/\sqrt{N}} \quad \text{where } \sigma_{jk}^2 = \sigma_j^2 + \sigma_k^2 \quad (13)$$

Blom cautions against a liberal application of the test, because the scale values s_j are not independent. He proposes to make no more than one comparison per scale. Although this is perhaps not fully justified, we have tested the differences between the six pairs of texts on all scales, using a significance level of .001%. An additional reason to risk this combination of tests is that it can give us some insight into the within-speaker variation and therefore into the question whether a single reading text lasting not more than two minutes is sufficient to get a stable and valid judgement of a speaker. Speaker DJ appeared to differ from himself on no less than 23

scales. This is not too surprising and disturbing a result as there were more than ten years between his readings of the neutral and non-neutral texts; furthermore, the technical quality of the recordings as played to the judges differed audibly. The other extreme is speaker NR whose readings were only significantly different on the scale 'agitated $\leftrightarrow$ calm'. The readings of speaker TB differed on 5 scales, those of LB on 8 and those of EK and PD on 9 scales.

The significant differences between the two readings of one speaker, except for the case of DJ, seem to be concentrated on a limited number of scales. This again is not very surprising as some attributes are more prone to within-speaker variation than others. Four talkers differ significantly from themselves on the scales 'jerky $\leftrightarrow$ smooth flowing', 'tense $\leftrightarrow$ relaxed' and 'selfconfident $\leftrightarrow$ wavering'. Thus at least two of the personality stressing scales are susceptible to within-speaker variation, but this variation is not contingent upon the content of the reading text. Three significantly different pairs are found on the scales 'loud $\leftrightarrow$ soft', 'careless $\leftrightarrow$ precise', 'steady $\leftrightarrow$ unsteady', 'agitated $\leftrightarrow$ calm', degree of accentedness and reading performance.

As to the scales on which 'texts' is a significant factor it appears that the direction of the difference can always be accounted for by the differences between the scale values of the texts of the speaker(s) who differ(s) significantly. Only for the scales 'ugly $\leftrightarrow$ beautiful' and degree of accentedness we do not need to invoke the significant differences to explain the direction of the texts-effect, because on those two scales the members of the pairs of texts are always in the same order.

3.3.1 *Conclusions*

Although our data do not allow very far reaching conclusions, they do strongly suggest that in a situation in which the attention is mainly called to phonation and articulation quality, the contents of the reading texts are not a matter of major concern. This conclusion is as true for scales that directly address aspects of articulation and phonation as it is for scales that probe aspects of the personality of the speakers. Note that the generalizability of this conclusion is heavily restricted by the phrase 'in a situation in which the attention is mainly called to phonation and articulation quality'. The results might have been quite different if we had used another set of rating scales or if we had not employed students of logopedics as judges in the rating experiment. Things might also have turned out completely different if we had split up the ratings into two separate sessions, one for rating phonation and articulation and another for rating para- and sociolinguistic aspects. In the latter rating session attention would certainly not have been concentrated on technical aspects of voice and articulation. In any case, the finding agrees with the results of much research that has addressed problems like the attribution of social status or personality characteristics based on a person's voice alone. If the content of a reading text really does not influence ratings on personality stressing scales it might, after all, be acceptable practice to ask

subjects to read the names of the days of the week and the month of the year or the letters of the alphabet and have these readings rated by listeners as did a number of social psychologists.

With respect to the occasions where the between-text difference was significant for one or more talkers, it should be taken into account that the cause of the difference may reside in the talker rather than in the raters. It is quite conceivable that some subjects have adapted their style of reading to the considerable differences in style and content of the texts. It should be clear that extremely intricate experiments, most probably using synthetic speech, are needed to decide this issue of primacy of apparent cause and effect.

3.4 FACTOR ANALYSIS

3.4.1 *Introductory remarks*

The purpose of carrying out factor analyses on our data was three-fold. First of all we wanted to obtain insight into the dimensionality of the judgement space. More specifically, we wanted to see whether or not the dimensions we had in mind when selecting the rating scales appear to function in reality. The next step, which was of primary interest to the team in Amsterdam, was to reduce the number of scales in the rating form by choosing the scales with the best factorial purity and the highest reliability. Lastly we wanted to obtain the coordinates (factor scores) of the talkers in the semantic space in order to see whether or not they can be explained by acoustic measurements.

More often than not factor analysis is applied to semantic differential scores without previously checking whether the correlations between items may be considered as linear. We have shown that for our data the linearity assumption is justified. It is not our intention to deal with the intricacies of factor analysis as such, but a number of short introductory remarks are necessary to make clear what we have done. The factoring method that has been used is the one known as principal factoring with iteration, i.e., the PA2 option in the SPSS program FACTOR. The mathematical model underlying the factoring approach assumes that the variance in the scores on each of a set of variables can be divided into two components, known as communality and unicity. The communality is the variance that is due to the effect of a (small) number of underlying factors that affect a number (but not necessarily all) variables simultaneously. The unicity is the variance in a variable that is not due to a common underlying factor nor in any other way related to the variance in other variables. Relations between variables are expressed in terms of their correlation coefficients. Thus, the input to a factor analysis program is a correlation matrix.

From Fig. 3.1 it is apparent that our scores form a three-dimensional structure. In order for a factor analysis to be applicable the data must be

Fig 3 / Display of the three-dimensional structure of the scores in the rating experiment

converted into a two-dimensional correlation matrix. This can be done in many ways, at least in theory. In practice, the purpose of the research precludes collapsing the scales-dimension, since it is the relation between the scales that we are interested in.

Correlations between scales can be arrived at in three ways. We could replace the scores of the individual raters by scale values of the speakers, or, for that matter, by mean scores which have been shown to be almost perfectly correlated with scale values. This removes all listener-variance from the scores whereas speaker-variance is retained. The result might, however, not be very illuminating since what we do in such a case is study the distribution of 11 or 12 points in a 35-dimensional space. This would very likely lead to the conclusion that a single dimension suffices to account for most of the between-speaker variance. An alternative way to compute correlations between scales would be to sum the scores on all speakers for each judge. This procedure would remove all speaker-related variance from the data and leave us with pure rater variance. The averaging over speakers has the disadvantage that we lose all information about the position of the speakers in the judgement space, so that it will not be possible to relate those positions to acoustic measurements. Therefore it does not suit all purposes of our undertaking. The last possible procedure, called the 'stringing out proce-

cedure' in Osgood, May & Miron (1975) is to combine speaker and listener variance by treating every score of every judge for each speaker as an independent observation. This procedure can be visualized in Fig. 3.1 by taking slices from the block parallel to the listener dimension and putting the slices on top of each other along the speaker dimension. An obvious drawback of this procedure is that it does not allow for a formal separation of speaker and listener variance. It is, however, possible to establish the positions of the speakers in the speaker-listener judgement space by averaging the factor scores of the speaker-listener combinations over listeners. Furthermore, the stringing out procedure must be expected to yield reliable results when analyzing the scores of individual groups, since even in the smallest group the product of the number of speakers and the number of listeners exceeds the number of 35 dimensions (scales) of the raw data space by far. We have decided to use the stringing out procedure since it offers the best combination of advantages, possibilities and (absence of) drawbacks for our purposes.

3.4.2 *The dimensionality of the judgement space*

Since much of the results of the Amsterdam material has been published previously (Fagel, van Herpt & Boves, 1983) and since the results obtained from both data sets are essentially similar, we will concentrate here on the Nijmegen material. First, factor analyses were performed on the scores of the four groups of raters separately. Each analysis produced five factors with an eigenvalue greater than 1, which together accounted for slightly more than 60% of the total variance in the scores. The factor structures defined in terms of the loadings of the scales were quite similar, especially as far as the first two factors, that account for the bulk of the explained variance, are concerned. The principal difference was that between the group comprised of students in Amsterdam on the one hand and the other three groups on the other in the importance given to the first two factors. The two major factors can be labelled 'Voice dynamics' and 'Articulation quality', respectively. The listener groups originating from Nijmegen and Hoensbroek attach most importance to the Voice dynamics factor, whereas in the factor solution for the Amsterdam group Articulation quality accounts for the largest proportion of the total variance. This difference in point of view is very probably related to the greater familiarity of the groups from Nijmegen and Hoensbroek with the regionally coloured speech varieties. In any case, the group in Amsterdam complained that most of the speakers sounded quite 'southern'. In all four groups the third factor was related to Osgood's potency factor and the fourth factor clearly had to do with voice colour (high loadings of the scales 'high pitch $\leftrightarrow$ low pitch' and 'deep $\leftrightarrow$ shrill'). The fifth and last factor varied between the groups and looked like a somewhat arbitrary part separated from the quite broad potency or dynamism factor (factor number 3). In no case did the last factor account for more than a few percent of the explained variance.

Table 3.6. Results of factor analysis on the ratings of 12 speakers by 117 judges on 35 scales. Only factors with eigenvalue > 1. have been extracted. Six factors account for 57.8% of the total variance in the scores. Loadings > .45 are underlined. Scales marked with an asterisk have especially low factorial purity.

			I	II	III	IV	V	VI	h^2	
	1	pleasant	unpleasant	-.362	-.543	.068	.219	.292	.206	.606
*	2	dull	clear	.340	.309	.081	-.392	.439	-.144	.635
*	3	friendly	curt	-.444	-.372	.192	-.114	.273	-.009	.460
	4	loud	soft	-.179	-.229	.031	.601	.084	.235	.509
*	5	husky	not husky	.301	.363	.020	-.400	.180	-.106	.426
	6	stereotyped	varied	.641	.103	-.151	-.178	-.015	.017	.476
	7	active	passive	-.600	-.162	.103	.344	-.027	.180	.551
*	8	artless	affected	-.116	-.131	.063	.145	.190	.072	.097
	9	colourless	sonorous	.716	.225	-.107	-.300	-.019	-.058	.668
10	careless	precise	.270	.526	-.368	-.220	-.019	-.049	.536	
11	high	low	-.056	.075	-.185	-.140	-.694	.087	.551	
12	firm	slack	-.262	-.343	.211	.593	.282	.165	.689	
*	13	creaky	not creaky	.225	.416	-.179	-.179	-.034	-.072	.294
	14	dragging	brisk	.250	.265	-.078	-.310	.044	-.585	.579
15	melodious	monotonous	-.792	-.151	.199	.047	.039	.038	.695	
16	powerful	weak	-.288	-.282	.213	.588	.318	.179	.686	
17	ugly	beautiful	.428	-.580	-.171	-.197	-.349	-.142	.730	
18	jerky	smooth flowing	.344	.374	-.588	-.056	-.246	-.124	.683	
19	spiritless	vivacious	.776	.243	-.148	-.095	.013	-.122	.707	
20	rounded	angular	-.332	-.318	.457	-.068	.358	.027	.554	
21	deep	shrill	-.038	-.150	.312	.050	.758	-.045	.700	
22	poor	rich	.542	-.377	-.302	-.107	-.266	-.035	.611	
23	tense	relaxed	.273	.138	-.654	.004	-.223	.003	.571	
24	expressive	expressionless	-.765	-.150	.249	.116	.053	.086	.693	
25	steady	unsteady	-.173	-.291	.568	.376	.220	.159	.652	
26	broad	cultured	.154	-.687	-.301	-.160	-.086	-.147	.641	
27	self confident	wavering	-.255	-.283	.649	.265	.118	-.167	.678	
28	whining	cheerful	.611	.311	-.183	-.185	-.001	-.167	.566	
29	polished	slovenly	-.252	-.669	.412	.154	.115	.023	.718	
30	deviating	normal	.239	.555	-.144	-.283	-.152	-.175	.520	
31	quick	slow	-.068	-.169	-.027	.176	-.058	.779	.675	
32	agitated	calm	.133	.226	-.557	.052	-.204	.316	.523	
33	distinct	indistinct	-.226	-.587	.372	.201	.027	.073	.580	
34	accentedness (high/low)		.108	.479	-.110	-.310	-.097	-.274	.434	
*	35	reading performance (good/bad)	-.445	-.399	.406	.165	.196	-.022	.588	
X of variance accounted for by factor			66	13.8	9.3	4.8	3.1	3.0		

Since we felt that the similarities in the factor structure of the individual groups outweighed the differences by far, we decided that it was warranted to treat the complete set of 117 listeners as a single group. The same findings and subsequent conclusion apply for the Amsterdam material. Thus the Nijmegen data were subjected to a new stringing-out analysis with 117×12 rater-talker combinations as observations. The results of this analysis are given in Table 3.6. Six factors account for 57.8% of the total variance in the scores and in terms of the amount of variance explained the first factor is by far the most important. From the scales that have high loadings on this factor it is clear that it can be unambiguously labelled as a voice dynamics or melodiousness/variegatedness factor. From the intermediate loadings of the scales 'ugly ↔ beautiful' and 'pleasant ↔ unpleasant' it must be concluded that the voice dynamics factor has some evaluative connotations. These are even stronger for the second factor that clearly relates to articulation quality and that might be called 'Articulation appreciation' to express the high loadings of 'ugly ↔ beautiful' and 'pleasant ↔ unpleasant'. The third factor,

the last one explaining a sizable amount of variance, seems to attract high loadings mostly from scales that probe paralinguistic and personality features of the speakers. The fourth factor is clearly Osgood's potency factor with its high loadings of the scales 'loud $\leftrightarrow$ soft', 'firm $\leftrightarrow$ slack' and 'powerful $\leftrightarrow$ weak'. The fifth factor is the 'voice pitch/voice colour' factor that emerged from the analysis of the separate groups and, not very surprisingly, also this factor has some evaluative connotation. The sixth factor seems to be exclusively related to tempo, since the highest loading non-tempo scale has a loading as small as .27.

3.4.2.1 *Discussion*

Perhaps the single most striking aspect of the results is that the usual evaluation and activity factors are conspicuously absent from both the Nijmegen and Amsterdam factor solutions. This is the more remarkable since we have decidedly added standard reference scales to our instrument to represent these factors. At first sight this seems to be rather puzzling, but on second thoughts our results can be given a satisfactory explanation. Before attempting such an explanation it is, however, worthwhile pointing out that in general the factor structure complies fairly well with the functions that the scales were supposed to perform. At the outset we had chosen scales to represent voice quality, articulation quality and para- and sociolinguistic features, together with reference scales for the standard semantic differential factors. The voice quality, articulation quality and paralinguistic factors emerge very clearly from the results, be it that the voice quality factor is split into a voice dynamics factor and a much less important 'voice statics' factor. Actually, the clearly separate status of a paralinguistic or personality factor might be more unexpected than the absence of the two standard factors. We had rather anticipated the paralinguistic scales to be distributed over independently motivated factors. In that case, para- and sociolinguistic judgments could have been explained by referring to the voice, articulation and/or general factors with which they had been related, much in the same way as the actual behaviour of the scales 'whining $\leftrightarrow$ cheerful' and 'friendly $\leftrightarrow$ curt', which have an understandable association with the voice dynamics factor. It is interesting, incidentally, to note that this is the situation that emerged from the research in Amsterdam, where a factor analysis on the same set of scales resulted in a five factor structure. The factors in the Amsterdam solution are identical to the factors found in the Nijmegen data, with the exception of the personality factor that is absent.

Before returning to the discussion of the status of the general factors it is worth noting that the scales 'artless $\leftrightarrow$ affected' and 'creaky $\leftrightarrow$ not creaky' have very low communalities. This is entirely in line with their low reliability. It is interesting to note that few scales seem to have a very high factorial purity, i.e., almost all scales show non-negligible loadings on more than one factor. Most striking in this respect are the scales 'dull $\leftrightarrow$ clear', 'friendly $\leftrightarrow$ curt' and 'reading performance' that have their contributions divided over

three or four factors, with the result that they do not play a role in the conceptual definition of a single factor.

This property of substantial loadings on more than one factor is shared by the three scales meant to represent the evaluation factor (scales number 1, 17 and 22); the same comment applies to the scales included to represent the potency (12, 16 and 20) factor and two of the reference scales for the activity factor (7 and 14, but not 31 which is exclusively related to the tempo factor). It has already been pointed out by Osgood, Suci and Tannenbaum that the general evaluation factor can be considered as a 'cover factor', summarizing a number of related but discriminable evaluative dimensions like moral evaluation, esthetic evaluation etc. If the objects under judgement do not call for a distinction between the various aspects of the evaluation the general cover-factor will emerge; in other situations, however, a number of aspects may be separated into independent factors.

In our results a comparable, but perhaps somewhat more complicated effect seems to be present. Not only has the evaluation factor been partitioned into aspects related to voice dynamics, voice statics and articulation appreciation, the activity factor has also been divided into a part that is related to voice dynamics and a separate tempo factor. One explanation of this result can be found in the ratio of the number of standard reference scales to the number of non-standard scales expressively addressing attributes related to speech. In our instrument the special speech scales outnumber the standard reference scales by far. As a consequence the variance in the reference scales is outrivalled by the variance in the speech scales with the result that the latter are likely to determine the factor structure. The standard reference scales fit into that structure and help to interpret it. It is clear, then, that in judging speech the general evaluation factor is partitioned and associated with what we have deliberately called Articulation appreciation and Voice pitch/Voice colour and, perhaps to a somewhat lesser extent, also with Voice dynamics. The last term is explicitly chosen to express the association of that factor with the usual activity or dynamism factor, part of which is parcelled out and shows up as the most easily discernible activity aspect of speech, viz. tempo.

3.4.3 The construction of an efficient scaling instrument

It has already been said that the development of a maximally efficient scaling instrument was the major aim of the work of our colleagues in Amsterdam. Mainly for that reason most of the work was based on the factor solution for the Amsterdam data. Since the results are readily accessible in the literature (Fagel et al., 1983), we will confine the exposition to a terse description of the method and a survey of the results in so far as they have an impact on the contents of the present book.

The work of the team in Amsterdam proceeded along the following lines. First the dimensionality of the judgement space was established. Five factors

were found, which have been explained in the previous section. Since it appeared that the Voice dynamics factor was heavily contaminated with Osgood's evaluation factor, whereas the Potency factor had an appreciable loading from static voice quality scales, it was decided to discern a total of seven 'factors', which are not necessarily completely orthogonal. Next it was decided to represent all 'factors' by two scales. Thus an optimal selection from scales with high loadings on the original factors had to be made. The choice was guided by the communality of the scales, in combination with their factorial purity, and their reliability. The resulting instrument is shown in Table 3.7. The most conspicuous feature of the set of scales in that table is the absence of socio- and paralinguistic scales. This reflects both the failure of the factor solution for the Amsterdam material to produce a sociolinguistic factor and the fact that the work in Amsterdam was mainly directed towards the development of an instrument for use in the world of logopedics. We will use the reduced rating instrument when it comes to an acoustic explanation of the perceptual ratings. By doing so redundant discussions of essentially identical scales can be prevented. We will add, however, a pair of scales representing the paralinguistic factor found in the Nijmegen material.

3.4.4 *The position of the talkers in the judgement space*

The factor scores of the readers judged in the Amsterdam experiment have

Table 3.7 The rating instrument resulting from choosing the optimal set of scales

Factor #1	monotonous	↔	melodious
Voice dynamics	expressionless	↔	expressive
Factor #2	slovenly	↔	polished
Pronunciation	broad	↔	cultured
Factor #3	weak	↔	powerful
Potency	soft	↔	loud
Factor #4	shrill	↔	deep
Pitch	high	↔	low
Factor #5	dragging	↔	brisk
Tempo	slow	↔	fast
Evaluation	ugly	↔	beautiful
'factor'	unpleasant	↔	pleasant
Voice pitch/	dull	↔	clear
Voice colour	husky	↔	not husky
'factor'			

been discussed in Fagel et al. (1983). Thus we can confine the presentation of data and their discussion to the Nijmegen data. Since we have used a stringing out approach throughout, the factor scores to be presented are averages over 117 listeners of 12×117 text-listener combinations. Every talker is present with two texts, but the texts have been treated as completely independent observations. The averaged factor scores for the twelve 'talkers' on the six factors extracted are graphically shown in Fig. 3.2. The talkers are indicated by the initial letter of their Christian name; the subscripts identify the text, *i* referring to the non-neutral and *a* to the neutral one.

The first impression conveyed by the figure is that in terms of factor scores the within-speaker consistency is considerably less than what was suggested by the number of scales on which the between-text differences are significant (refer to 3.3). Even subject NR, who differs from himself on no more than a single scale, does not have a very stable position in the judgement space. It is equally clear, however, that the within-speaker variability is greater along some dimensions than along others. The variability seems to be less pronounced on dimensions which relate to speaker characteristics that must be

Fig. 3.2 Factor scores of the 12 texts in the semantic judgement space. Roman numerals indicate the factor number. Upper case letters identify the six speakers, lower case subscripts identify the texts: *a* is neutral, *i* is non-neutral. The scales to which the plots are made are identical for all factors, with the exception of factor V, where it is halved.

supposed to be stable over time, like pronunciation quality and pitch height/voice colour. The dimensions which show the greatest amount of instability are those that are most likely to be affected by differences in reading performance, like the voice dynamics and the tempo dimensions. It is interesting to note that PD, who is very consistent on the voice dynamics dimension, made the impression of reading the neutral text much faster than the non-neutral one.

Furthermore it seems that the ratings have not suffered too much from a halo-effect, for there are no clear patterns of some texts consistently being rated either positively or negatively. The intercorrelations of the factor scores are generally very low and insignificant, with two exceptions: the scores on the pronunciation appreciation and potency factors correlate $-.69$ and those on the potency and tempo factors $.76$. The positive relation between potency and tempo is quite conceivable but the tendency for talkers to be rated less potent if their pronunciation quality decreases is less easy to explain. Apparently potency must be taken in a somewhat abstract sense ('this talker has a chance to go a long way in a professional world') than in a more concrete sense of physical strength or toughness. This relation might have turned out quite differently if our raters had not been students of logopedics, who are taught to favour the standard pronunciation of the language. Meanwhile it should be noted that in the Amsterdam data there are no significant correlations between factors. This prevents us from drawing any general conclusion from the few and weak patterns in the Nijmegen material.

It is extremely difficult to assess the meaning and importance of the apparent lack of stability of the six Nijmegen talkers in the judgement space. The variability of the factor scores may be artificially inflated, at least as compared with the scores on individual scales, by the additive effects of slight differences on a number of highly correlated scales. It is worth recalling the significant texts $\times$ speakers interaction found in the analysis of variance reported in 3.3. Apparently the texts have played a non-negligible role, but not a systematic one. Stated in overly simplistic terms: the subjects have performed better (more expressive, relaxed, etc.) in reading one text than in reading the other, but for some the neutral text was best and for others the non-neutral text. The differences in performance may have been too small to show on individual scales (again, recall that the alpha level employed in testing within-talker differences was quite high) but their additive effect on factor scores is considerable. The variability in the Nijmegen data is the more puzzling since for the Amsterdam data a very stable talker-structure was found in a situation that should be expected to give rise to much more variation. A comparison was made of the factor scores of ten speakers in the present experiment and the factor scores resulting from the rating experiment by Blom & van Herpt (1976-a). In the latter experiment the same ten speakers were rated (on a quite different version of the rating form, refer to Chapter 2 for more details) but the audio stimuli were not the readings

employed in the present study but rather recordings of a freely retold short story. Factoring of the 1976 ratings resulted in a three-factor solution. Canonical correlation of the five- and three-dimensional sets of factor scores showed that 89% of the variance in the scores for the three-factor solution could be accounted for by the five-factor solution (the former accounts for 66% of the variance in the latter). This result cannot be interpreted but as an indication that the relative positions of talkers in a judgement space seem to be fairly stable, both over time and over speaking and rating conditions. There are reasons to believe that this result is - at least in part - due to the very wide ranges in the speech material employed in Amsterdam: half of the speakers were females, the other half were males, each sex group contained two professional speakers and two talkers whose pronunciation is quite heavily coloured by a dialect, but different dialects for different subjects. In the Nijmegen material on the other hand, all subjects were males with a fairly homogeneous dialect background and none of them was a professional speaker. Since the data available at this point do not allow a further analysis of the within-talker differences, we will postpone the subject until we have more detailed information on the acoustic characteristics.

3.5 CONCLUSIONS

In this chapter we have analyzed the perceptual ratings of a number of listener groups judging two groups of talkers reading one or two texts. The first part of the chapter was devoted to an investigation of the formal scaling properties of the ratings. It was found that the ratings on carefully constructed successive category scales can safely be treated as interval level data. The reliability of the ratings proved to be quite satisfactory. If one is interested in averaged ratings, rater groups of no more than some 30 judges will yield effective reliabilities (far) in excess of .9. In most cases an effective reliability of approximately .9 can be reached with much smaller groups if the range of the scale attribute in the stimulus material is not too narrow. The judgement space proved to be five- or six-dimensional, with dimensions labelled Voice dynamics (expressiveness), Articulation appreciation, Potency, Pitch level/Voice colour, Tempo and a factor stressing the socio- and paralinguistic aspects of the rating form. The last mentioned factor did not emerge from the Amsterdam material. Although the original rating instrument did contain a sufficiently high number of standard reference scales for the customary evaluation and activity factors, these dimensions appeared not to play an independent role in the judgement of speech samples. Rather Voice dynamics, Articulation appreciation, and Pitch level/Voice colour seem to be independent aspects of a general kind of evaluation, whereas activity is most closely related to Voice dynamics. A new, condensed version of the rating form was proposed, which covers all aspects mentioned above, except the personality-stressing factor. For this part of the perceptual rating a new and independent instrument will have to be developed. The contents of the

reading texts did not have a systematic effect on the ratings, which is quite different from saying that they had no effect at all. It appeared that the six readers in the Nijmegen experiment performed somewhat variably when reading both texts. Thus even when fairly long passages are used as stimuli in a rating experiment one must envisage chance effects of differential reading performance, the more so when the stimulus material as a whole is quite homogeneous.

Separation of the contributions of phonation and articulation to the radiated speech signal

4 0 INTRODUCTION

This chapter is devoted to an attempt to analyze the acoustic speech signal in such a way as to enable us to separate the contributions of phonation from the effects of articulation. The assumption that phonation and articulation can and should be separated is one of the undisputed fundamentals of almost any phonetic theory, be it that almost equally invariably the interest of phoneticians has been focussed on the articulation part of the speech production. The study of phonation has traditionally been the domain of speech therapists and phoniatrists. The reasons for the relative neglect of phonatory aspects by phoneticians are not hard to find. The development of phonetic knowledge and theory has mainly been guided by linguistics and communication theory. From a purely linguistic point of view phonation is only of interest as far as the distinctive feature 'voiced $\leftrightarrow$ voiceless' is concerned, remaining aspects that affect things like voice quality being banished to the realm of paralinguistics. Communication engineers have often taken a comparable stance. what they want to get across is the linguistic meaning of a message and especially in military communication systems security is bound to be judged as much more important than quality or naturalness.

The belief that phonation and articulation should be distinguished and can, in fact, be separated is very clearly expressed in the pervasive linear model of acoustic speech production as developed by Fant (1960) and Flanagan (1972). This model has been the basis for a great number of acoustic phonetic studies, most of which focuss on the modelling of articulation (Ungeheuer, 1962, Vieregge, 1969; Mol, 1970, among many others). We assume the readers to be familiar with the linear model of acoustic speech production. Therefore we will confine ourselves to a short review which is mainly intended to establish our terminology and formalism.

The model has originally been developed for the continuous time (analog) signal domain. We have done all signal processing in the discrete time - quantized value domain, which is more generally known as the digital signal domain. The analog signal model can readily be transformed into the digital domain. We will present the model in its digital version. Digitized signals are

represented by sequences $s(nT)$, where n is any integer and T is the sampling interval. Because T is a constant, the expression will mostly be used in an abbreviated form as $s(n)$.

Fig. 4.1 contains a block diagram of the acoustic model of speech production in its most general form. From that figure it can be seen that the source model consists of two parts, one source for what traditionally are called voiced sounds and the other for the unvoiced sounds. The model as depicted can account for a mixed source, necessary for a precise representation of the voiced fricatives /z, v and γ/. Our attention will be focussed on the branch of the source model that is responsible for the periodic excitation.

The radiation part of the acoustic speech production model is usually considered to be a simple differentiator. Its z -transform can therefore be written as

$$L(z) = 1 - z^{-1} \quad (1)$$

[For the use of the z -transform the reader is referred to Steiglitz (1974) who gives an easy to follow introduction that should be sufficient to understand most of the mathematics to follow. More thorough treatments of the z -transform and its use in digital signal processing can be found in Jury (1973), and Oppenheim & Shafer (1975)].

The mathematical description of the transfer function of the vocal tract depends on the level of accuracy and detail one wants to attain. If one is aiming at a detailed description of vowels and consonants the vocal tract must be modelled as a pole-zero system, the transfer function of which can be written as

$$V_{pz}(z) = \frac{g \prod_{j=1}^J (1 - c_j z^{-1} + d_j z^{-2})}{\prod_{k=1}^K (1 - a_k z^{-1} + b_k z^{-2})} \quad J \leq K \quad (2)$$

If one is willing to limit the discussion to non-nasalized vowels the numerator part of the expression (2) can be equated to a constant, resulting in the so called all-pole model

$$V_{ap}(z) = \frac{g}{\prod_{k=1}^K (1 - a_k z^{-1} + b_k z^{-2})} \quad (3)$$

There exists a one-to-one relation between a_k and b_k on the one hand and the frequency and bandwidth of the k -th formant on the other.

Fig. 4.1 Block diagram of the customary linear model of acoustic speech production. The parts of the model should be taken to be 'black boxes', the functioning of which has yet to be specified.

Although (2) has a wider applicability than (3) our work is based on the assumption that the all-pole model described by (3) is a correct representation of the vocal tract. This limitation can be motivated by pointing out that manageable methods for estimating the parameters $[a_k, b_k, K; c_j, d_j, J; g]$ in (2) from the acoustic speech signal have not yet been developed. Methods for estimating the parameters $[a_k, b_k, K; g]$ of the model (3) are readily available. Finally, and perhaps most importantly, the model in Fig. 4.1 is supposed to be linear. Mathematically this means that the overall system function of the model can be expressed as the product of the z -transforms of the individual parts that are connected in series. Denoting the system function of the source model by $G(z)$ the system function of the complete model in the figure can therefore be written as

$$H(z) = G(z) \cdot V(z) \cdot L(z) \quad (4)$$

The product in (4) obeys all laws governing the products of complex numbers, including the law that products are commutative. This means that the order of the blocks in Fig. 4.1 can be changed without changing the resulting output waveform. Such a situation can only exist if the parts of the model are mutually independent.

If we look at the aerodynamic and acoustic behaviour of the separate parts of the speech production mechanism the idea that the subsystems should not influence one another can lend considerable insight into the conditions which must be fulfilled for a linear model to be applicable. To aid this discussion, Fig. 4.2 shows a highly schematized view of the speech production mechanism. During the production of vowels the glottis opening is by

Fig 4 2 Highly schematized representation of the human peripheral speech production mechanism

far the narrowest constriction that the air exhaled from the lungs has to pass. The resistance which the air flowing from the lungs to the open atmosphere has to overcome may be considered as concentrated at the glottis. In this situation the glottis can be considered as a source of volume flow which can be studied independently, i.e., regardless of the details of the shape of the supraglottal cavities. During the production of consonants, on the other hand, an additional constriction is formed in the flow bed downstream of the glottis. In this situation the air flow through the glottis is no longer completely determined by the shape of the glottis itself. Consequently the behaviour of the glottis can no longer be studied fruitfully without regard to the shape of the supraglottal articulators.

The all-pole model described by (3) is further detailed in Fig 4.3. The train of (scaled) unit sample pulses at the input of the model is a mathematical construct that has no direct physical counterpart. The pulses excite the glottal shaping filter that distorts them in such a way that the output waveform looks like a sequence of idealized glottal flow pulses. This idealized pulse form is based on both theoretical considerations and actual physiological measurements of the behaviour of the human larynx. The major aim of this chapter is to develop and test techniques for estimating the transfer function $G(z)$ of the glottal shaping filter. The idealized glottal volume flow, denoted by $U_g(n)$ is next fed into the all-pole vocal-tract filter $V(z)$ that

Fig. 4.3 Block diagram of a linear model of acoustic vowel production, in which the transfer function of the black boxes in the previous figure has been specified (at the cost of its generality).

produces an output flow wave $U_1(n)$, i.e., the volume flow that can be measured at (between) the lips. The radiation at the lips, represented by the block with system function $L(z)$, converts the flow wave at the lips into a travelling pressure wave that can be recorded at some distance in front of the lips by an ordinary microphone.

In Fig. 4.3 the number of pulses in the input pulse train is left unspecified. Now it is well-known from systems theory that the mathematical analysis of the model is simplified considerably if we take either a doubly infinite number of pulses (from $nT = -\infty$ up to $nT = +\infty$) or a single impulse at $nT = 0$. The latter possibility is more attractive, because this signal has the extremely simple z -transform:

$$\delta(nT) \Big|_{nT=0} \longleftrightarrow 1 \quad (5)$$

where $\longleftrightarrow$ denotes a z -transform pair. Since the z -transform of the output of a linear digital system equals the product of the z -transform of the input signal and the system transfer function (in z -transform notation) we have in Fig. 4.3 (supposing the input to consist of a single unit pulse at $nT = 0$)

$$S(z) = 1 \cdot G(z) \cdot V(z) \cdot L(z) = G(z) \cdot V(z) \cdot L(z) \quad (6)$$

with $S(z) \leftrightarrow s(nT)$. Thus it appears that the z -transform of the output signal is simply the system transfer function in z -transform notation, provided that the input signal is a unit sample impulse.

Let us now consider a situation where we have a speech signal $s(n)$, $n = 0, 1, \dots, N-1$ ($s(n) = 0$ for $n < 0$ and $n > N-1$). The z -transform $S(z)$ of a finite duration signal $s(n)$ can always be evaluated. Also, assume that we have estimates of the frequencies and bandwidths of the formants from which we can construct the vocal tract transfer function $V(z)$. Lastly, suppose that $s(n)$ is the output of the model from Fig. 4.3, excited by a unit impulse at $t = nT = 0$. Now we have all ingredients necessary to explain how the glottal volume velocity waveform can be recovered from the acoustic speech wave and why the

process is often called inverse filtering. According to the arguments which led to (5) the z -transform of our speech signal can be written as

$$S(z) = U_g(z) \cdot [V(z) \cdot L(z)] = U_g(z) \cdot V(z) \cdot L(z) \quad (7)$$

If we have estimates of both $V(z)$ (arrived at via formant extraction) and $L(z)$ (based on theoretical considerations) we could try and construct a new linear system, the system function of which is exactly the inverse of $V(z) \cdot L(z)$:

$$I(z) = \frac{1}{V(z) \cdot L(z)} \quad (8)$$

Taking the speech signal $s(n)$ as input to the filter with transfer function $I(z)$ we get an output signal O that can be specified by

$$O(z) = S(z) \cdot I(z) = S(z) \cdot \frac{1}{V(z) \cdot L(z)} \quad (9)$$

Substituting (6) into (8) leads immediately to the result that $O(z) = U_g(z)$ and from the z -transform theorem stating that two signals having identical z -transforms must be identical themselves it follows that the output of our inverse filter $I(z)$ exactly equals the glottal flow signal $U_g(n)$.

This somewhat lengthy and, despite of the use of some formulae, highly informal introduction is meant to provide at least some intuitive basis for explaining the intricacies of inverse filtering aiming at the reconstruction of the glottal flow wave from the acoustic speech signal. An attempt at reconstructing the glottal flow from the acoustic speech signal implies the acceptance of a model of acoustic speech (perhaps more accurately: vowel) production and the trust that the parameters of that model which are not related to the glottal component can be estimated reliably. From this it should be clear that the reconstruction of glottal flow waves can be complicated in two ways: either the actual articulation of the speech sound under analysis may not be approximated with sufficient accuracy by the underlying all-pole model, or the estimates of the model parameters may be in error. Stated differently: the reconstruction of the glottal flow from the acoustic speech signal calls for the solution of one equation with three unknowns (viz. $G(z)$, $V(z)$, and $L(z)$) and to complicate things even further it is not guaranteed that the equation really holds. Assuming that the equation $S(z) = G(z) \cdot V(z) \cdot L(z)$ holds (corresponding with the assumption that the linear model is appropriate) the accuracy of the computation of $G(z)$ depends on the accuracy with which $V(z) \cdot L(z)$ can be estimated.

From what has been said so far it should be clear that inverse filtering in order to recover the glottal flow is beset by much uncertainty. Therefore,

means for evaluating its validity and accuracy are needed. There have been some attempts to record the volume flow at the glottis directly using hot-wire anemometers (Teager & Teager, 1984). The results and interpretation of his recordings are, however, questionable. Our present knowledge of the glottal flow waves is based on a combination of indirect measurements, modelling studies and inverse filtering experiments. The indirect measurements pertain mainly to ultra-high speed movie films of the vocal folds during phonation and experiments with different kinds of glottographs. Modelling studies combine anatomic data with the outcomes of indirect measurements and the results of experimental and theoretical studies of flow phenomena to derive predictions as to the behaviour of the vocal folds and the consequent form of the flow through the glottis.

The data available today suggest that each glottal cycle can be divided into two parts, viz. an open and a closed glottis interval. During the closed interval air is prevented from flowing into the vocal tract. During the open interval the glottal area changes rather smoothly, the area as a function of time assuming a triangular-like shape. Using this knowledge, interactive inverse filtering techniques try to obtain an output which contains a flat 'zero'-interval corresponding to the closed phase of the glottal cycle, and a smooth pulse shape in the open interval. Although these criteria surely have a fairly solid empirical basis they should not be overvalued. For one thing, not all subjects seem to show complete glottal closure during an appreciable part of the glottal cycle, at least not in all modes of phonation. Moreover, Holmes (1976) has pointed out that there may be nonnegligible acoustic excitation of the vocal tract during the closed glottis interval. Thus the absence of a completely flat 'zero-flow' segment from the inverse filter output does not necessarily mean that the settings of the anti-formant parameters are incorrect. Last but not least the measurement data which serve as a reference for the inverse filter output very often stem from another subject than the person who produced the speech used in the inverse filter analysis. And even if the physiological data and the acoustic signal do stem from the same subject, they do not pertain to the same speech segment. Krishnamurthi and Childers (1981) have compared inverse filter results derived from a speech segment recorded just prior to the activation of an ultra-high speed camera with glottal area measurements derived from the same subject's utterance after the camera had reached its full speed of 4.400 frames per second, i.e., with a delay of $\pm .5$ seconds. The delay was necessary because of the extremely high noise level produced by the camera, which precludes the registration of acoustic speech signals amenable to meaningful analysis. Recently it has been possible to make useful acoustic recordings while the ultra-high speed camera is running, either by attaching a hearing aid microphone to the laryngeal mirror used to expose the vocal folds (Krishnamurthi, 1983) or by placing the camera in an acoustic enclosure (Childers, personal communication).

Ultra-high speed filming has the drawback that the laryngeal mirror prevents the subject from normal articulation. Phonation might also be affected to some extent. Therefore we have opted for a more indirect approach, in which reference information on the behaviour of the glottis can be obtained without interfering with normal articulation and phonation. The data obtained by means of indirect measurements are necessarily less detailed than the information that can be extracted from laryngeal films. Yet if it were possible to obtain reliable information on the moments of glottal opening, maximum glottal area and glottal closure in each pitch period, it would be possible to judge the accuracy of an inverse filtering technique to a considerable extent.

We have made simultaneous recordings of the electroglottogram, the photoglottogram and the subglottal pressure together with the acoustic speech signal. The procedure for making the recordings is detailed in 4.2. Section 4.3 deals with the attempts to establish the moments of glottal opening, closure, and maximum area from the three physiological signals. In 4.4 the development of an optimal inverse filtering technique suitable for use with connected speech is described. The chapter ends with the description of a large scale experiment in which the newly developed inverse filtering technique is used in order to study the glottal volume flow waveform under a number of conditions.

Section 4.1 gives an extensive and hopefully comprehensive account of the procedures for making and interpreting recordings of subglottal pressure. The material of 4.1 is included since very little has as yet been published on the subject. A comparable treatment of the electroglottogram and photoglottogram seems to be inappropriate, since much information is readily available in recent literature (Fourcin, 1974; Kitzing, 1977; Lecluse, 1977; Pedersen, 1977; Baer, Löfqvist & McGarr, 1983; Krishnamurthi, 1983).

4.1 REGISTRATIONS OF SUBGLOTTAL PRESSURE

Recordings of subglottal pressure were made in order to obtain additional information on the instants of glottal opening. The idea that subglottal pressure might yield this information was based both on a (small) number of subglottal pressure (henceforward P_{sg}) registrations in the literature and on theoretical considerations. P_{sg} registrations published by Koike & Hirano (1973) show fairly abrupt pressure drops at the moment of glottal opening as established from simultaneous ultra-high speed film data. Such a pressure drop might be accounted for by the Bernoulli effect, as will be explained shortly. Consistent relations between variations in P_{sg} and glottal area (recorded in the form of a photoglottogram) are also apparent from registrations in Flanagan, Rabiner, Christopher & Bock (1976).

The theoretical considerations for expecting a P_{sg} -drop at the moment of glottal opening derive from mathematical models of the air flow during

phonation. The aerodynamic phenomena in and below the glottis have been studied in great detail by van den Berg, Zantema & Doornenbal (1957) and later by Ishizaka & Matsudaira (1972), Scherer & Titze (1981), and Ananthapadmanabha & Gauffin (1983). If one is willing to accept a gross stylization of the cross-sectional shape of the trachea, the larynx and the pharynx and if one makes some additional simplifying assumptions (a.o. incompressibility of the air, one dimensional and lossless flow) it appears to be possible to describe the aerodynamic behaviour of the sub- en supra- glottal systems by means of three fundamental laws of physics, viz. the laws of conservation of energy (in a special formulation known as Bernoulli's law), the law of conservation of mass and the law of conservation of momentum. Broad (1979) gives an exceptionally clear and comprehensible account of the theory developed by Ishizaka and Matsudaira; moreover, he shows that it is very plausible to assume that the reasoning, originally developed to describe the aerodynamic phenomena in the glottis, is also valid for the subglottal space. Then he proceeds to derive a formula which predicts the pressure variations in the trachea as a function of the glottal area, the primary factor determining the amount of flow. ('Predicts' seems to be the better term, not 'expresses', for Broad did not have access to empirical numerical data on subglottal pressure variations). Broad's formula is as follows:

$$P(s) = P_{sg} \cdot \left(1 - \frac{A^2(g)}{A^2(s)} \cdot C\right) \quad (10)$$

with $P(s)$ the pressure at some point in the trachea, and with $A(y)$ the cross-sectional area at some point along the length of the tube made up of the trachea, larynx and pharynx through which the air flows parallel to the y -coordinate. The point $y=g$ defines the upper edge of the glottis, $y=s$ is a point somewhere in the trachea, and $y=r$ is the point of reattachment, i.e., the point in the pharynx where the jet that originates from the glottal outlet has broadened to such an extent that it reattaches to the wall of the pharynx. P_{sg} is, as usual, the mean subglottal pressure and C is a factor depending on the ratio of the cross-sectional areas of the pharynx at the point of reattachment and the glottis:

$$C = [1 - 2N - 2N^2]^{-1} \quad (11)$$

with $N = A(g)/A(r)$

In order to obtain an impression of the magnitude of the Bernoulli pressures that are to be expected in the trachea we have simulated (10) varying $A(g)$ and $A(r)$. It appeared to be difficult to find data on the absolute values of the maximum glottal area during phonation. Flanagan (1972) is among the few who give absolute data on glottal area and from them one can infer that areas exceeding .3 cm² are not very probable. Our results are summarized in Fig.

4.4 for pharyngeal areas ranging between 1.5 cm^2 and 5 cm^2 . Glottal area is varied from 0 to $.4 \text{ cm}^2$. For glottal areas less than $.15 \text{ cm}^2$ the variation in P_{sg} is virtually independent of the cross-sectional area of the pharynx; pharyngeal articulation seems to be only important for exceptionally great glottal areas. For probable, though perhaps not really small, values of the glottal area (between $.25$ and $.3 \text{ cm}^2$) the variations in P_{sg} appear to be in the order of 1% to 2% of the mean subglottal pressure. Taking mean values of approximately $10 \text{ cm H}_2\text{O}$, the high frequency variations should be in the order of $1.5 \text{ mm H}_2\text{O}$.

Assuming lossy flow, van den Berg et al. (1957) estimated that the pressure drops resulting from the flow will slightly exceed 5% of the mean subglottal pressure. Considering their way of estimating the pressure drop (by straightforward substitution of constant values for flow and flow resistance in Ohm's formula: pressure drop equals flow times flow resistance, setting peak flow to $600 \text{ cm}^3/\text{s}$ and flow resistance to $.11 \text{ cm H}_2\text{O}/100 \text{ cm}^3/\text{s}$) their result is probably only valid at a point right below the glottis and the measured pressure drop should decrease according to the distance between the glottis and the transducer.

Fig. 4.4 Subglottal pressure drop due to the Bernoulli effect as predicted by Broad (1979). Pressure drop is plotted as a function of glottal area, with pharyngeal cross-sectional area $A(r)$ as a parameter. The cross-sectional area of the trachea $A(t)$ is kept constant at 3 cm^2 .

We will now consider the question whether pressure variations in the order of 1 mm H₂O can be registered with the pressure transducers that are presently available.

4.1.1 *Techniques for recording subglottal pressure*

Over the years a number of techniques have been developed for measuring P_{sg} (Schutte, 1980). All these techniques are necessarily more or less invasive, the indirect technique of measuring the pressure in the esophagus instead of in the trachea being the least annoying for the subjects. The pressure under the glottis can be measured directly. One way of doing that is by inserting a hollow needle into the trachea through the skin just above the first tracheal ring. A somewhat less invasive direct measuring technique is to insert a catheter into the trachea via the glottal opening; it seems to be van den Berg who first described this technique in 1956. The useful frequency range of direct P_{sg} recordings will not exceed a bandwidth of some 300 Hz when classical pressure transducers are used. Edmonds, Lilly & Hardy (1971) showed that the system of needle and tube connecting the subglottal space with the pressure sensitive element in the manometer forms the limiting factor that determines the frequency response of the total set-up. In spite of the limited frequency response of the needle-tube-manometer system direct P_{sg} recordings made by means of a tracheal puncture tend to show pitch synchronous variations.

Since the introduction of miniature pressure transducers mounted at the distal end of a very thin catheter, P_{sg} registrations covering a much wider frequency range are possible. According to the literature this possibility has so far not been used very frequently. In his review of the literature Schutte (1980) mentions only two studies, viz. Perkins & Koike (1969) and Kitzing & Löfqvist (1975). Since then research in which the technique was employed has been reported by Koike (1981). In all cases mentioned the transducer was inserted into the trachea via the nasal passage and the glottis. Koike & Hirano (1973) used a miniature transducer with a female subject who was left with a fistula below the thyroid cartilage by a disease from which she had suffered a number of years before but from which she had recovered completely.

Miniature pressure transducers (such as the Millar transducers used in our study) basically consist of a pair of semiconductor strain gauges in a Wheatstone bridge. The pressure sensor is contained in a stainless steel case and its position within the case is fixed by means of hard epoxy. The steel and epoxy case opens to one side, where the sensor is covered by a silicone rubber diaphragm. The case with the sensor is mounted at the distal end of a thin plastic catheter that conducts the connecting wires (Fig. 4.5) and provides a vent tube necessary to expose the inside of the sensor to atmospheric pressure. Pressure variations cause the strain gauges to bend and thus to change their electrical resistance, thereby disturbing the balance of the bridge and

Fig. 4.5 Electrical equivalent circuit diagram of a miniature semiconductor strain gauge pressure transducer (Part A). Mechanical design of the Millar PC350 instrument used in our study. (Part B)

producing a proportional change in the voltage across the bridge terminals. The output voltage variations depend not only on the variations in the electrical resistance of the strain gauges but, of course, also on the bridge excitation voltage. After suitable amplification the output can be displayed on an oscilloscope and recorded for further analysis. We have constructed a Pressure Transducer Control unit (PTC) that meets both very stringent safety requirements and technical requirements with regard to amplitude and phase response (Boves, 1981). With a metal film resistor network replacing the transducer and in the absence of any input signal, the noise voltage at the output of the PTC amounts to 1.4 mV.

The Millar transducers show a pronounced temperature sensitivity, although variations of the output with temperature are least in the neighbourhood of 37° C, which is the normal body temperature of healthy subjects. Furthermore, the transducer seems to be slightly sensitive to variations in ambient light intensity. According to the specifications given in the older versions of the user's manual the frequency response is flat up to 20 kHz; specifications of the frequency response are absent in the most recent version of the manual. We have checked the frequency response of the transducers using a coupler similar to the one described in Edmonds et al. (1971). The results confirm that the amplitude response is completely flat in the frequency band from 0 to 4 kHz.

Apart from the frequency response, which does not seem to cause any suspicion, the sensitivity of the transducer is a factor of relevance for interpreting the recordings. According to the specifications the sensitivity of the

Millar transducers is in the range of 1.5 to 5 mV/V/100 mm H_g, the measured sensitivity of all transducers we have experimented with proved to be in the middle of that range.

Having the information given above, we are in a position to assess the smallest pressure variations that can be measured reliably. To that end we must combine the data on the noise output of the PTC and the sensitivity of the transducer. We will consider a measurement as reliable when the signal to noise ratio is greater than 3 dB. We have carried out calculations assuming a 'typical' transducer to have a sensitivity of 3 mV/V/100 mm H_g and a PTC delivering an excitation voltage of 5 VDC. It can be shown that it takes pressure variations in the order of .56 mm H₂O before a level of 3 dB above noise is reached (Boves, 1981). Doubling the bridge excitation voltage (which is possible with the Millar transducers) to 10 VDC and replacing the op-amp used in our PTC with an extremely low noise instrumentation amplifier could perhaps lower the threshold of the signals that can be measured reliably to pressure variations of about .1 mm H₂O. Pressure variations with an amplitude smaller than .1 mm H₂O can only be measured if the sensitivity of the transducer is increased. An extensive search of the market has, however, failed to yield competing instruments with a sensitivity somewhere near that of the Millar transducer.

From what is said above we can conclude that the minimal pressure variations that can be measured with our equipment are in the order of .5 mm H₂O. This compares favourably with other measurements reported in the literature. Most importantly, one can infer from the data in Edmonds et al. (1971) that the Statham type PM131TC+2.5-350 transducers, which are very popular in phonetic research, produce an output that is drowned in the noise when the pressure variations drop below 1 mm H₂O. Also the Whittacker type 1015 C transducer used by Koike and Hirano seems to have a sensitivity which is at best half that of the Millar instruments.

The most important conclusion is, however, that the Millar transducer is just sensitive enough to record the extremely small Bernoulli pressures that are predicted by Broad (1979). They are surely sufficiently sensitive to record the much greater pressure variations that follow from the considerations of van den Berg et al.

4.2 MAKING THE SIMULTANEOUS REGISTRATIONS

We decided to record four signals simultaneously, viz. the acoustic speech signal (as input to the inverse filter) and the electroglottogram (henceforward EGG), subglottal pressure and the photoglottogram (PGG) as signals from which the moments of vocal fold closure, vocal fold opening and moment of maximal glottal area should be derivable. The situation during the recor-

dings is shown in Fig. 4.6. In this section we will discuss a number of practical details of the recording procedure. First the relevant technical specifications of the apparatus will be given. Next the procedure followed to position the transducers will be described.

For a correct interpretation of the output of inverse filter methods it is extremely important that the phase response of all equipment used is linear for very low frequencies, i.e., down to at least one half of the pitch frequency (Lindqvist, 1965). Low-frequency phase distortion in the output of the inverse filter (whether caused by the inverse filter itself or already present in the recording of the original speech wave) or in the recordings of the physiological signals would make the comparison of the salient points in the output of the inverse filter with corresponding points as indicated by the physiological signals very difficult if not impossible. Therefore, much effort has been spent in assuring that all equipment has a linear phase response in the frequency range of interest or, when this proved to be impossible at least a known and smooth non-linear response which can be easily compensated for.

An important source of low frequency phase distortion is direct recording of signals on magnetic tape. There are two obvious ways around this problem, viz. analog magnetic recording using an FM-recorder or digital recording. We have chosen the latter approach, mainly because all further processing of the recorded signals will be digital (Boves, 1979).

Fig. 4.6 Schematized overview of the situation during the simultaneous recordings of the electroglottogram, photoglottogram, subglottal pressure and acoustic speech signal.

The Frøkjær-Jensen LG 900 photo-electric glottograph used in our work has some characteristics which tend to affect its suitability for our purpose. Specifically the mechanism for transmitting the light from the lamp to the subject's trachea and the mounting of the photo transistor can be improved. The light of a halogen projector lamp operated from a stabilized DC source is transmitted to the subject's neck via a rigid plexiglass rod, conically shaped to have a diameter of approximately 1 cm at its far end (where it is in contact with the skin). The subject is supposed to press the neck against the rod and hold the head in a fixed position during an experimental run. Movements of the head and variations in the force with which the rod is held against the tissue lead to considerable variations in the luminance of the tracheal cavity. Slow variations in the intensity of the light source may, however, be of only minor importance in a study focusing on fast phonatory movements of the vocal folds. The photo-sensor coming with the device is a Texas Instruments LS400 photo-transistor which is mounted in a polyethylene tube with an outside diameter of 3 mm. Attached to the photo-transducer is a transparent tube that must be swallowed into the esophagus. This set-up enables the photo-sensor to remain in a relatively securely fixed position during an experiment. This decrease in the (unwanted) mobility of the sensor is obtained at the expense of its manoeuvrability. Specifically, the sensor is constrained to positions close to the back wall of the pharynx. This might very well lead to the transducer seeing hardly more than the cartilaginous part of the glottis, whereas in our research it is the ligamental part which is of primary interest.

The electroglottograph used in our work, i.e., a device from Laryngograph Ltd., did not cause any ergonomic problems. However, both the photo- and electro-glottographs caused an electronic problem in that the vocal fold vibration component in their output signals is superimposed on a low frequency noise component, the amplitude of which exceeds that of the F_0 component by at least an order of magnitude. Thus high pass filtering is necessary in order to suppress the low frequency fluctuations. Since the LG 900 photoglottograph was developed with the measurement of low frequency glottal articulations in mind, it is not surprising that the instrument does not provide an AC coupled output. The Laryngograph, on the other hand, combines two mechanisms to suppress low frequency fluctuations which are caused by vertical movements of the larynx. The first is a feedback loop on the level of the high frequency carrier signal, whereas the second is a high pass filter in the output stage, following the demodulation circuit. Unfortunately, the cut-off frequency of the high-pass filter is rather high (approximately 40 Hz). Lecluse (1977) has convincingly shown that cut-off frequencies of that order effect a visible phase distortion of the output signal. Therefore, the high pass filter in the output stage of the Laryngograph was bypassed.

In order to obtain AC coupled signals with a minimal amount of phase distortion, compensation amplifiers were designed and built for use with both glottographs. These amplifiers, which effectively amplify the difference

of the incoming signal and its short-time integral act as high-pass filters with cut-off frequencies of approximately 5 Hz and approximately linear phase. That this circuit faithfully reproduces the slopes in the input signal can be seen from Fig. 4.7 where the input and output are compared for square waves of 10, 50 and 100 Hz. The traces are computer generated plots of signals sampled by means of the system depicted in Fig. 4.6. The excitation voltage was sampled via the DC channel normally dedicated to the P_{sg} signals. Thus it appears that the AC coupling employed with the EGG does not affect the signal characteristics to a significant extent. The same remark applies to the PGG.

As to the P_{sg} recordings, phase distortions due to the electronics were carefully avoided. This is, unfortunately, not to say that these recordings did not cause technical problems. Two of these deserve some discussion. The first one concerns the dynamic range of the recordings. If we want to obtain a faithful registration of both the slow variations in P_{sg} related to expiration - and perhaps also to intonation - and the high frequency variations due to phonation superimposed on the former, we must assign part of the 10 V range of the A/D converter to either component. Conservative settings of

Fig. 4.7 Input (lower trace) and output (upper trace) of the high-pass filters employed in the electroglottograph and photoglottograph for signals of 10 Hz (panel a), 50 Hz (panel b) and 100 Hz (panel c). The original DC component of the input signals is not shown in the figures.

amplifier gains, motivated by the knowledge that the mean subglottal pressure during normal phonation may be anywhere between 4 and 12 cm H₂O may cause a reduction of the high frequency oscillations to almost the level of the quantization noise. Emphasizing the high frequency part of the signal, on the other hand may easily lead to negative or positive clipping of the signal as the mean P_{sg} goes down or up in the normal range. The only way around this dilemma is to record two versions of the P_{sg} signal, viz. the original one and its exact derivative, both occupying the total dynamic range. This solution is, however, not practicable in our set-up because it would lead to a lowering of the effective bandwidth of all signals to about 3.2 kHz.

The second problem concerns the absolute calibration of the pressure transducer. Extensive experiments have led to the conclusion that an absolute calibration can only be carried out 'in vivo', i.e., by having the subject blow into an U-shaped tube filled with water and recording the voltage output of the PTC for several heights of the water column. We were forced to this conclusion by the fact that all attempts to arrive at an exact calibration by recording the PTC voltage output when the transducer was immersed into a water column, held at a temperature of approximately 37° C, just before and after the experiments with our subjects, invariably led to results that were inconsistent with the voltages recorded at the moments of inhalation (when P_{sg} values should be negative) and quiet expiration (when P_{sg} should be only slightly positive). For the failure to obtain valid 'in vitro' calibrations a number of reasons can be imagined, the most probable being the long-term drift of the transducer, (or perhaps rather the unpredictable response of the transducers to a temperature shock (Cranen 1983)) and - last but not least - the possibility that the sensor is pressed against the wall of the trachea, which should cause an unpredictable but considerable upward shift of the zero pressure level. Unfortunately, definite conclusions regarding the calibration problem were only reached after the last recording session with our subjects. Therefore some caution will have to be exercised in interpreting our results in terms of absolute and relative magnitude of the high frequency oscillations.

In order to minimize phase distortion in the audio recordings we have used a Bruel & Kjaer 1/2" condensor microphone type 4134, a B&K pre-amplifier type 1619 with a type 2607 measuring amplifier and a Krohn-Hite model 3341 anti-aliasing low-pass filter with its cut-off frequency set to 4 kHz. In the building where all registrations were made a high level low frequency background noise (caused by the air-conditioning) is constantly present. The superposition of the speech wave onto this low frequency signal caused intermittent heavy overloading of the output stage of the measuring amplifier. In order to prevent this overloading we have used the built-in high-pass filter of the 2607 amplifier. This filter, with its cut-off frequency of 22.5 Hz, has a non-linear phase response for very low frequencies, but this response is so smooth that it can easily be compensated for by a digital phase equalizer (Hunt, 1978).

The procedure for making the recordings has been developed in close cooperation with C.I.E. Jansen M.D., head of the department of Logopedics and Phoniatics. After a first application of Xylocaine the photo sensor is inserted and its position is adjusted by visually checking the output on an oscilloscope placed in front of the subject. Then a second dose of the anaesthesia is applied and after a few minutes the pressure transducer can be inserted into the trachea. Next it should be checked that the transducer is properly positioned below the vocal folds and that the catheter passes the glottis in the posterior commissure. This is best done by both indirect laryngoscopy and monitoring of the PTC output when the subject raises and lowers the lung pressure with his glottis closed and his mouth opened. Auditory checking of the speech produced by the subject is also very important. When positioned correctly the catheter does not affect the subject's speech in an audible way. This assertion is corroborated by the photographs reproduced in Fig. 4.8. These pictures, made by H.K. Schutte M.D. of the E.N.T. clinic of the University Hospital at Groningen, clearly show that there is no difference in the way the glottis closes during normal phonation when the catheter is present or when it is not. It is, of course, extremely difficult to control the position of the transducer in the trachea exactly. X-ray photographs made from one subject after the catheter had been inserted (courtesy of H.K. Schutte M.D.), suggest that during most recording sessions the transducer has been located in the lower third of the trachea or in the direct neighbourhood of the bifurcation (cf. Fig. 4.9).

Finally the electrodes of the electroglottograph are attached to the skin and the microphone is placed in front of the subject. In order to get the best possible signal-to-noise ratio for the acoustic signal the microphone is placed very close to the lips. It is very important for the subject to be able to monitor at least the three physiological signals on an oscilloscope.

During the recordings cooling fans from equipment in the terminal room were turned off in order to get the lowest possible background noise level. This notwithstanding, the recording conditions were, of course, far from optimal, if only because at least two experimenters were needed to monitor all equipment in order to insure an optimal use of the dynamic ranges, and to prevent heavy under- and overloading of the input amplifiers of the A/D conversion system.

Recordings have been made of five adult male speakers who produced widely varying speech material, ranging from isolated vowels to spontaneous speech.

4.3 ESTABLISHING CRITERION INSTANTS BASED ON PHYSIOLOGICAL REGISTRATIONS

In this section we will treat the processing of the physiological signals aiming at the determination of the moments of glottal closure, glottal opening and

A. Glottis during phonation with catheter pressure transducers present (medium effort phonation).

B. Glottis during medium effort phonation after removal of the catheter pressure transducer.

Fig. 4.8 Photographs of subject's LB's glottis during phonation with and without the pressure transducer inserted. Photos made by dr. H.K. Schutte, E.N.T. Clinic Groningen, using a fiber optic.

maximal glottal area. Our recordings of the EGG appear to be fairly typical for all speakers and therefore do not warrant an extensive discussion. For ease of reference a graph of a typical cycle of an EGG is given in Fig. 4.10. Suffice it to say that the moment of glottal closure can be easily and reliably established from the EGG (moment I in Fig. 4.10). For some speakers the

Fig. 4.9 X-ray photograph of subject LB, showing the position of the pressure transducer in the trachea. (Courtesy of dr. H.K. Schutte, E.N.T. Clinic Groningen.)

'knee' that according to Lecluse signals the moment of glottal opening (moment III in Fig. 4.10) is clearly present in part of the EGG traces and equally apparently absent from others. The EGG of the remaining speakers is characterized by the virtually complete absence of such a 'knee'. A careful study of traces with and without a 'knee' suggests, however, that the location

Fig 4.10 Schematic representation of a closing-opening movement of the glottis as seen in an electroglottogram. The voltage, plotted along the vertical axis, can be looked upon as representing vocal fold contact area or, alternatively, electrical conductance of the tissues including both vocal folds.

Moment I marks the onset of glottal closure, at II the closure is complete. Between II and III the folds begin to separate from below and at moment III the separation reaches the upper margin of the folds. At moment IV the folds are completely separated and further lateral movements do not affect the electrical conductance.

of the instant of glottal opening at a moment just prior to the EGG regaining its zero line will not be far off. Similar results have been reported by Krishnamurthi (1983).

The interpretation, and consequently the processing, of the PGG and P_{sg} registrations is considerably less clear and unambiguous. Therefore a more thorough discussion is called for.

4.3.1 *The photoglottogram and the moment of maximum glottal flow*

In our work the photoglottogram should serve as the signal from which the moment of maximum glottal area, and consequently the moment of maximum flow can be established. The photoglottograms made with the LG 900 glottograph in its original form were invariably of low quality. After extensive experimentation it was concluded that this must be due to the positioning of the photo transistor in the pharynx, which is tightly constrained by the mechanism for fixating its position. Recent experiments, carried out with a newly constructed photo pick-up have confirmed this conclusion. In the new pick-up we have placed the photo transistor at the very end of the catheter (Cranen, 1983; refer also to Fig. 4.12). The increased manoeuvrability enabled us to make very clear recordings of the glottal area but at the same

time it was apparent that the output is extremely sensitive to small variations in the position of the photo transistor relative to the glottis. An interesting detail is that the optimal position of the newly constructed pick-up appeared to be several centimeters higher (relative to the glottis) than the only position in which the original LG 900 pick-up produced an AC output of the same repetition frequency as the EGG.

When we were still trying to solve the problems with the quality of our PGG recordings, the first of a series of papers by Rothenberg became available reporting on his research into the relation between glottal area and glottal volume flow (Rothenberg, 1981, 1982). Rothenberg investigates an inherently dynamic situation in which a glottal source with a time varying internal resistance is loaded by a combined sub- and supraglottal impedance. He shows that the current through a time varying resistance (actually an admittance, i.e., the inverse of resistance is used) has a time function which is congruent to the time function of the resistance only as long as the load impedance is purely resistive. If, however, an inductance is inserted into the circuit, the current and the admittance functions are no longer congruent. If the time course of the glottal admittance is triangular, the current function is not only rounded, but its maximum value is delayed with respect to the top of the triangle.

If Rothenberg's approach is correct, it would defeat our intention to use the glottal area wave to determine the moment where the glottal flow should reach its maximum value. The failure is not primarily due to the fact that there is a delay between the maxima of the area and flow waves, but rather to the fact that the delay is at least in part determined by the supra-glottal articulation.

Thus it seems that the premise that the glottal area wave provides an additional criterion to assess the quality of an inverse filtering technique by determining the moment of maximal glottal flow, was wrong from the very beginning. This insight relieves the pain caused by the original failure to obtain useful PGG recordings, but it raises the new question what was wrong with Flanagan's reasoning that led him to the conclusion that area and flow functions should be essentially congruent (Flanagan, 1958).

Flanagan's analysis is based upon the assumption that the flow through the glottis can be validly approximated by a steady flow. In the case of steady flow the inductive part of the sub- and supraglottal loads can be neglected, and the essential congruence of the flow- and area functions is re-established, as is confirmed by Rothenberg (1981). For a more detailed discussion of mechanisms that may give rise to a skewing of the pulses refer to Cranen & Boves (1983).

Rothenberg's analysis, combined with work by Ananthapadmanabha & Fant (1982) as well as our own measurements force us to envisage a quite disturbing possibility, viz. that the classical aerodynamic theory of speech

production is not completely adequate. As a direct, and very unpleasant consequence of that hypothesis, not only our intention to use the PGG as a criterion for determining the moment of maximum flow would be doomed to fail, but the idea that the moment of glottal opening can be determined from the P_{sg} might also run into considerable problems. For, as explained in 4.1, that endeavor depends in a crucial way on the correctness of the aerodynamic theory put forward by a.o. Ishizaka and Matsudaira which is based on the assumption that the flow is steady.

4.3.2 *The P_{sg} recordings and the moment of glottal opening*

It was explained in section 4.1 that we did not arrive at an accurate absolute 'in vivo' calibration of the pressure transducer. Nevertheless, there is no doubt whatsoever that the high frequency oscillations caused by the periodic opening and closing of the glottis have magnitudes far exceeding the level predicted by aerodynamic theory. Variations in the order of 25% of the mean subglottal pressure are by no means exaggerated. Kitzing & Lofqvist (1975) give appreciably higher estimates.

As can be inferred from the registrations shown in Fig. 4.11 there is another reason for assuming that the pressure variations are not simply due to the Bernoulli effect. This reason is the quite complex waveshape of the pressure signals. The discrepancy between aerodynamically predicted and actually measured pressure variations can at least in part be explained by the fact that the considerations of van den Berg et al., Ishizaka and Matsudaira, and Scherer and Titze are based on a flow that is assumed to be stationary, whereas our measurements are made in a condition of periodically interrupted flow. A periodic disruption of the flow may give rise to phenomena known as the 'water hammer effect' in the field of fluid flow technology. The water hammer effect can be understood as a conversion of kinetic energy into potential energy: the sudden decrease of kinetic energy of the air particles that find their way blocked by the closing glottis should give rise to a momentary and local increase in pressure. Maxima in P_{sg} at the instant of glottal closure are obvious in all registrations made with miniature transducers (cf. Fig. 4.11).

The water hammer effect and its results cannot be described in aerodynamic models; they rather belong to the domain of acoustics. Empirical data on the acoustic properties of the subglottal system are rather scarce. The most recently published data seem to be the measurements on the input impedance of the trachea made by Ishizaka, Matsudaira & Kaneko (1976) on five laryngectomized Japanese subjects. They found clear resonances at frequencies of 640, 1400 and 2100 Hz; the variation between subjects was rather small for the lowest resonance, whereas the higher resonances showed considerable inter-subject variation. It appeared that the resonance at 1400 Hz always had the highest input impedance and consequently the smallest bandwidth.

Fig. 4.11 Sample registrations of reconstructed glottal volume flow (curves labelled 'a'), electroglottogram (b), subglottal pressure (c), photo-electric glottogram (d), speech wave e) and normalized prediction error (f), computed using a 24-point rectangular window in an 8-th order covariance analysis without pre-emphasis. The speech wave and the EGG, PEG and P_{sg} signals have been fed through a 251-point linear phase high-pass filter with cut-off frequency 40 Hz prior to analysis and display.

Formant estimates are mean values computed over the closed glottis interval only (upper figures are center frequencies, lower figures are bandwidths).

The measurements of Ishizaka et al. differ appreciably from the results of a model simulation by van den Berg, (1960), who based the values of the parameters in his model on measurements carried out on the subglottal system in human and canine cadavers. He found resonance frequencies of $(2n-1).300$ Hz, $n = 1, 2, 3, \dots$, the first resonance having the smallest bandwidth. Therefore, it is not surprising that the model of the subglottal system as designed by Ishizaka and his coworkers differs in many respects from van den Berg's model. Although the details of the model of the subglottal system are of only minor importance for the goals pursued in the present study, an extensive analysis was made of the P_{sg} recordings in order to obtain a better understanding of these signals and, more importantly, a better understanding of the mechanism of phonation. For an account of the details of the analyses the reader is referred to Boves (1983-a; 1983-b); here only a summary will be given of the results that bear directly on the mechanism of phonation.

It was found, then, that the subglottal system exhibits a number of resonances and anti-resonances, which are largely fixed for a given subject and which therefore do not depend on the setting of the supraglottal articulators. The exact locations of the frequencies and bandwidths of the resonances and anti-resonances obtained from the analysis of P_{sg} recordings made with four different subjects are given in Table 4.1. It can be seen from this table that our results differ from both those of van den Berg and of Ishizaka et al., although they are closer to the latter than to the former. The differences between Ishizaka et al. and our results were not deemed sufficiently large to motivate the development of an alternative model of the subglottal system.

The fact that the subglottal system behaves as a resonator, much like the supraglottal cavities, and the finding that the subglottal pressure is far from being constant during a complete glottal cycle, support the findings explained in the previous subsection on the PGG, viz. that an approach to phonation based on an underlying assumption of steady flow is not fully adequate.

These findings explain also why it has been possible to record pitch synchronous variations in P_{sg} with recording equipment which was not suited to deal with the extremely low amplitude phenomena predicted on the basis of the Bernoulli effect. Our original intention in making the P_{sg} recordings was, however, to use the Bernoulli effect in order to detect the moment of glottal opening. From Fig. 4.11 it can be seen that there is indeed a more or less pronounced drop in P_{sg} at the moment of glottal opening that can be inferred from the EGG. We have attempted to devise an algorithm by which the instants of glottal opening can be extracted from the P_{sg} signals. It appeared, however, that the data obtained from a fairly complex processing of P_{sg} are not more reliable and stable than comparable data that can be obtained by a very simple manipulation (differentiating and searching of local extrema) of the EGG.

Table 4.1 Frequencies and bandwidths of the resonances and anti-resonances of the subglottal system in four normal adult male speakers. The columns labelled Std. (F) and Std. (B) contain the standard deviations of the estimates of the frequencies and bandwidths, respectively. # Items gives the number of vowel tokens on which the averages are based. Unequal numbers are due to the failure to detect all higher order peaks and valleys in all vowel tokens.

A	Resonance #	# Items	Frequency	Bandwidth	Std. (F)	Std. (B)
	EK 1	26	428.	312.	42.	84.
	2	22	1065.	335.	98.	86.
	3	26	1929.	269.	85.	70.
	4	26	2645.	301.	178.	93.
	5	22	3288.	256.	185.	90.
	LB 1	9	516.	292.	60.	39.
	2	9	1289.	298.	47.	54.
	3	8	1975.	410.	162.	135.
	4	8	2697.	353.	187.	69.
	5	7	3362.	349.	183.	65.
	NR 1	17	502.	215.	50.	27.
	2	16	1159.	312.	147.	78.
	3	17	1951.	296.	202.	85.
	4	16	2659.	247.	194.	106.
	5	11	3343.	286.	239.	84.
	PD 1	12	444.	282.	40.	114.
	2	11	1185.	381.	119.	148.
	3	12	1925.	382.	111.	219.
	4	10	2582.	410.	184.	163.
	5	11	3290.	345.	186.	88.
B	Valley #	# Items	Frequency	Bandwidth	Std. (F)	Std. (B)
	EK 1	24	1094.	321.	251.	106.
	2	22	1686.	285.	191.	83.
	3	25	2483.	267.	223.	105.
	4	23	3186.	304.	218.	153.
	5	3	3688.	407.	94.	63.
	LB 1	9	1048.	338.	100.	78.
	2	8	1696.	280.	64.	66.
	3	9	2738.	525.	191.	395.
	4	8	3337.	422.	245.	109.
	5	1	3727.	310.	0.	0.
	NR 1	17	1053.	283.	171.	85.
	2	17	1777.	288.	207.	72.
	3	17	2493.	270.	130.	99.
	4	13	3237.	288.	234.	70.
	5	2	3637.	290.	35.	202.
	PD 1	11	1029.	333.	162.	131.
	2	12	1742.	347.	266.	175.
	3	12	2607.	296.	249.	82.
	4	9	3249.	386.	270.	92.
	5	3	3615.	379.	30.	90.

4.3.3 Conclusions

In this section we have reported on our attempts to identify means of obtaining additional information on the behaviour of the larynx during the

production of normal speech, with the aim of using that information for the evaluation of inverse filtering techniques. The outcomes of our work add to the growing stock of evidence that the EGG, if recorded in a correct manner, is an extremely useful and reliable signal, not only for establishing the moment of glottal closure, but also for the detection of the instant of glottal opening. The usefulness of the EGG is very much enhanced by the fact that its recording is in no way invasive and does not interfere with normal articulation and phonation.

If conditions are extremely favourable, the PGG may be an even richer source of information, since this signal can be used to detect the instant of maximum glottal area in addition to the moments of glottal opening and closure. In comparison with the EGG, the PGG is much more difficult to record, and although it does not hamper normal speech production too much, it is in a way invasive. PGG recordings are crucially dependent on the positioning of the photo pick-up in the pharynx. It is our experience that the fixation of the photocel by means of swallowing a transparent plastic tube attached to it effectively invalidates the recordings if one is aiming at glottal area variations within a glottal period.

There appears to be a growing amount of evidence supporting the conjecture that the moment of maximum glottal flow lags behind the moment of maximum glottal area. Our recordings of P_{sg} add to that evidence in a way that was not envisaged at the outset of the research. Our results support the view that the aerodynamics of speech production cannot fully adequately be described by a model that is based on the assumption of steady flow. One consequence of the finding that a model based on steady flow is not completely correct, is the possibility that characteristics of the supraglottal pressure, which are dependent on the shape of the supraglottal cavities, may enter into the driving function at the glottis. Both the time lag between the maxima of the area and flow functions and the presence of characteristics of the supraglottal cavities in the glottal driving function point into the same direction: they suggest an interdependence of the glottal source generator and the vocal tract filter, therewith questioning the validity of the linear models of Figs. 4.1 - 4.3 even for the production of vowels. If the linearity assumption does not hold the model breaks down completely. And if the linear model can be shown to be inadequate in an essential way, the whole idea of separating articulation and phonation will become unsettled.

Thus our investigation of physiological signals has led to a number of interesting conclusions. First, it has cast some doubt upon the validity of the linear model of acoustic speech production. Since virtually all presently available inverse filtering techniques are based upon that model, caution must be exercised in interpreting inverse filtering results. Secondly, and as a corollary of the first conclusion it appears that physiological signals can only serve to a limited degree as criterion measures to assess the accuracy of an inverse filtering technique. The glottal area function does not completely determine the moment of maximum glottal flow, since the latter lags behind

the former to a degree which may depend upon the position of the supraglottal articulators. Attempts to determine the moment of glottal opening from the P_{sg} are hindered by the very complicated structure of these signals, which is due to the resonances in the trachea. It appeared, however, that the EGG is not only an excellent indicator of the moments of glottal closure, but that it also allows of a fairly accurate determination of the instants of glottal opening. Thus there remain reliable means to check at least the most important property of the flow pulses, viz. the duty cycle.

4.4. INVERSE FILTERING TECHNIQUES

Since the first paper by Miller (1959) quite a number of inverse filtering techniques have been published. The single most prominent feature of almost all approaches is that they are interactive. That is, a skilled operator has to adjust the inverse filter settings in such a way that the shape of the output waveform, which is constantly monitored visually, becomes 'optimal'. The optimality criteria most often cited are the presence of a flat 'zero-flow' portion during what is supposed to be the closed glottis interval and the absence of formant-like ripples in the open glottis interval.

The two criteria, viz. presence of a flat zero interval and absence of formant ripples in the open interval, used in interactive inverse filtering are not independent. Flattening of the alleged closed glottis interval may increase the amplitude of ripple components in the open interval and vice versa. Unfortunately, there is no formal way to weigh both deviations from optimality. And to complicate things even further, it is becoming increasingly unclear what is really optimal. Referring back to the discussions in 4.0 and 4.3 it should be clear that there are essentially two problems, which are not completely independent. The first problem addresses the validity of the model in Fig. 4.3, specifically its linearity and - less importantly - the all-pole nature of the vocal tract transfer function $V(z)$. The second problem concerns the accuracy with which the parameters of $V(z)$ can be estimated from the speech signal, supposing the model to be valid.

It appears that both problems can - at least to a large extent - be circumvented by adopting a closed glottis approach to inverse filtering. During the closed glottis interval (henceforward CGI) the vocal tract can be fairly accurately modelled as a tube closed at one end and open at the other, exhibiting damped sinusoidal vibrations due to an excitation in the past. This undriven system cannot interact in any way whatsoever with a source generator which is assumed to be absent. Thus the linearity of the system is guaranteed. An exactly similar reasoning leads to the conclusion that the all-pole approximation of the vocal tract should have the greatest chance to be valid during the CGI. A more detailed motivation for adopting the CGI approach to inverse filtering is given in Appendix I. A detailed description of our inverse filtering technique is also given there. Essentially the CGI is

determined from the EGG for each successive period of the vowel to be processed. The inverse filter settings are determined from the speech signal (using only the portions within each CGI) by means of the autocovariance method of linear prediction (Markel & Gray, 1976). The inverse filter parameters are determined for each glottal period and updated pitch synchronously. This allows us to process vowels taken from running speech, where the formant pattern is hardly ever stationary. Extensive comparisons of inverse filter outputs with open and closed glottis intervals determined from the EGG have shown that the reconstructed glottal flow excellently matches the duty cycle in the EGG (cf. Figs. 11 and 12).

We have built a software package that implements a semi-automatic inverse filtering of segments (vowels) from a continuous speech signal. Input

Fig. 4.12 Sample registrations of reconstructed glottal volume flow (a), photo-electric glottogram (b) (recorded with a newly constructed phototransducer), electro-glottogram (c), speech wave (d), and normalized prediction error (e).

are sampled versions of the speech and EGG signals, LPC parameter settings and a file containing information on the signal segments to be processed.

4.5 INVERSE FILTERING OF VOWELS FROM RUNNING SPEECH PRODUCED BY SEVERAL SPEAKERS

Using the semi-automatic inverse filtering technique indicated above and detailed in Appendix I an experiment was carried out to investigate the possibility of characterising speakers on the basis of reconstructed glottal waveforms. Apart from this major goal, a number of secondary aims were pursued: an attempt was made to find out whether

- the speaker characterizing results are stable over texts
- the glottal pulses pertaining to stressed vowels can be distinguished from the pulses reconstructed from unstressed vowels
- the prediction of Rothenberg (1982) that the skewing of the glottal pulses is vowel dependent can be substantiated.

4.5.1 *Speech material*

The speech material used in this experiment consisted of the readings of the neutral text described in Chapter 3 by all five experimental speakers (EK, LB, NR, PD and TB) and of the readings of the non-neutral text by speakers EK and LB. The latter texts were used in order to establish the stability of various results. The signals used in this experiment were the original digital recordings of the reading texts mentioned in Chapter 3. The acoustic speech signal and the EGG had been sampled and digitized without prior analog recording and stored on disc and digital magnetic tape. As explained in Chapter 3 no attempt was made to record the PGG or the subglottal pressure during the sessions in which the recordings of the texts for use in the judgement experiment were obtained. Speech and EGG were sampled at 8 kHz per signal.

From each reading 60 vowel tokens were selected, approximately evenly divided over stressed and unstressed syllables. The distinction between stressed and unstressed syllables was made by the present author. His decisions were based on an auditory impression and they were not very easy to make (see below). Care was taken to insure that all possible vowels of Dutch were present in the data base. Since neither text was designed to be phonetically balanced it was not surprising to find that the relative frequency of occurrence of the vowels in the text mirrors their average frequency in normal Dutch. This made it impossible to represent the 12 monophthong vowels of Dutch with equal frequencies and there could be no question at all of equal numbers of occurrence in stressed and unstressed syllables. One of the less frequent vowels (viz. the vowel /ø/) does not appear in the neutral text at all. The widely varying frequency of occurrence of several vowels does, of

course, negatively affect the last aim of this experiment, i.e., the analysis of the effect of vowel type on the amount of skewing of the pulses. The choice of vowels to be included in the set under analysis was further restricted by the phonetic context they appear in. Only vowels surrounded by pauses and/or non-nasal consonants were made eligible, this was done in order to prevent nasal coarticulation, that may well invalidate the all-pole model of acoustic vowel production. If possible, vowels surrounded by laterals were also discarded to prevent segmentation problems.

The third aim of this study, which has to do with the influence of stress level on pulse shapes, is somewhat hampered by the fact that the extremely neutral, or perhaps even slightly boring, reading text does in no way induce a very lively performance. Consequently the number of syllables containing heavily stressed vowels is exceedingly small so that we had to be satisfied with lexical stresses in a large proportion of the tokens. At the lower end the allowable range of stress levels was restricted by the wish to constrain the analysis to vowels that were phonetically recognizable. There are, however, a fair number of /ə/ vowels included for all subjects.

For all 420 vowel tokens the inverse filter output was computed and stored on disc. The number of glottal pulses per token is, of course, dependent on its duration and the fundamental frequency. The maximum duration analyzed per token is approximately 120 ms but most tokens were shorter.

4.5.2 Processing of the inverse filtering results

The aims of this study call for a compact representation of the inverse filtering outputs which is, at the same time, amenable to statistical processing. Such representations can be arrived at either in the time or in the frequency domain. A compact time domain representation could be accomplished by a parameterization of the pulses using a parameter set like the one proposed by Fant (1979). One attractive feature of this and similar parameterization schemes is that they allow the time variation of the waveform to be captured. A single set of parameter values for each vowel token could be obtained by averaging the parameters over all pulses or by taking those values that lead to a 'best fit' for a number of consecutive pulses. If necessary, the time variation of the pulse characteristics can also be retained in a frequency domain representation, though perhaps not as naturally as in the time domain, by computing Fourier transforms of individual pulses. A spectral representation, on the other hand, lends itself more naturally to a specification of average properties of a complete vowel.

Unfortunately, the various aims of this study make somewhat contradictory demands upon the eventual representation. For characterizing the speakers some kind of average spectral representation is probably to be preferred, whereas a time domain parameterization would certainly be the best way to approach an analysis of the dependence of the falling slope of the pulses on the frequency of the first formant. The choice of a spectral representation

Fig 4.13 Mean spectra of the glottal flow pulses derived from 60 vowels taken from 5 readings of the neutral text (last identification letter is A) and 2 readings of the non-neutral one (last identification letter is I) are displayed in the upper part of each panel. The first two letters of the identification code refer to the speaker. Dotted lines indicate the $\pm 1\sigma$ interval. The lower part of the panels show the average spectral difference between flow pulses reconstructed from stressed or unstressed vowels (Display is 'Stressed-Unstressed')

reflects the relative importance attached to the aims, the analysis of speaker dependency being at the top of the hierarchy and the vowel dependency at the bottom.

The transform of the inverse filter outputs into the frequency domain is effected as follows: for each vowel token the start and end points of the credible portions of the inverse filtering results were determined by hand, using the waveform editor. A Hamming window, the duration of which was determined by the end points, was multiplied into the signal and the resulting data record padded with zeros to obtain an equal number of 1024 samples for each token. An FFT was used to produce a 512 point amplitude spectrum that was eventually converted into a 'critical band spectrum' by summing over varying numbers of consecutive samples of the amplitude spectrum.

Fig 4.13 Continued

(Maling, Morrey & Lang, 1967). The resulting critical band spectrum covers the frequency range from 80 Hz to 4 kHz (i.e., half the frequency with which the original speech signals were sampled). This spectrum analysis yields levels in 13 filter bands, three of which have a constant bandwidth of 90 Hz and center frequencies at 122, 215, and 307 Hz, whereas the remaining filters are standard 1/3 octave filters with center frequencies at 400 Hz and upwards (Plomp, 1976; Pols, 1977; van Rossum, 1981). The relation between the rank order numbers of the filters and their center frequencies is made explicit in Table 4.5. The levels at the output of the filters are specified in dB relative to the overall level as determined from the sum of the entire set of 512 Fourier coefficients. This implies an effective level normalization between vowel tokens. The conversion of FFT amplitude spectra into critical band spectra has the added advantage of effecting a considerable amount of data reduction in a transparent and presumably perceptually interpretable way.

420 critical band spectra still make for an immense amount of data, too much to process by hand (or by eye). The individual spectra can be conceived

of as points in a 13-dimensional space. Fortunately, researchers using 1/3 octave spectra of speech signals have almost invariably found that the co-ordinates of the datapoints on several dimensions are correlated (Pols 1977; cf. chapter 5 below). Consequently, a transformation can be defined that maps the 13-dimensional space into a new space of (hopefully much) lower dimensionality with negligible loss of information. One way for defining an 'optimal' transformation is provided by multiple discriminant analysis. The transformation thus derived maps the original data points into a destination space in such a way that specific groups occupy maximally different regions in the new space. The groups to be separated must, of course, be defined *a priori*. We have used a multiple discriminant program based on Overall & Klett (1972).

4.5.3 Results

4.5.3.1 *Speaker characteristics*

Fig. 4.13 contains mean spectra of the individual speakers computed by averaging the levels in the 13 frequency bands over all 60 vowel tokens. An indication of the variation within each band in terms of the $\pm 1 \sigma$ range is also given. Although there are clear differences in the detailed structure, all average spectra approach the theoretical slope of -9 dB/octave (the -12 dB/octave slope in a constant bandwidth spectrum is converted into a -9 dB/octave slope in a constant percentage bandwidth spectrum since the number of harmonics within the pass-band doubles with a doubling of the centre frequency). The mean spectra of speakers EK, TB, and LB are somewhat more regular than those of NR and PD, especially in the low frequency region. Also, EK and LB show less variation within the bands. This difference is accounted for by the fact that LB and EK have mean speaking fundamental frequencies of approximately 90 Hz whereas NR and PD are half an octave higher. With a fundamental at 135 Hz the second harmonic is in the third filterband, leaving the second band almost empty. TB seems to form a group with EK and LB. His speaking fundamental frequency is approximately 110 Hz so that on the average all filter bands will contain at least one harmonic.

The results of a more formal investigation into the speaker characterizing properties of the reconstructed glottal flow pulses are summarized in Tables 4.2 - 4.4. As indicated above, use was made of multiple discriminant analysis, followed by a classification based upon transformation to orthogonal variates. The Tables contain the confusion matrices resulting from three analysis runs. In the first one (Table 4.2) seven groups of 60 pulse spectra each were treated as originating from seven independent sources (i.e., seven different speakers). Table 4.3 contains the results of an analysis in which the 420 pulse spectra were considered to pertain to five different speakers, EK and LB being represented by twice the number of items than the remaining speakers TB, NR and PD. In the last run only the 300 pulse spectra reconstructed from

Table 4.2 Confusion matrix for 7×60 glottal pulse spectra, regardless of vowel characteristics. Rows and columns labelled EKa and EK_i refer to vowels taken from the readings of the neutral and non-neutral text by speaker EK. A similar remark applies to LBa and LB_i. The remaining data pertain to readings of the neutral text.

	EKa	EK _i	LBa	LB _i	TB	NR	PD	# items per speaker	% correct
EKa	19	23	5	6	6	1	0	60	32
EK _i	21	28	0	4	5	2	0	60	47
LBa	8	3	31	9	5	3	1	60	52
LB _i	6	2	18	20	4	6	4	60	33
TB	9	0	8	4	30	7	2	60	50
NR	0	0	1	9	4	36	10	60	60
PD	1	1	2	0	7	10	39	60	65
# items assigned to speaker	64	57	65	52	61	65	56		

vowels taken from readings of the neutral text were processed, so that all speakers were represented by the same number of tokens, taken from an optimally homogeneous material.

In the seven texts-seven speakers run 48.3% of the pulse spectra were classified correctly; in the seven texts-five speakers run the percentage correct was 63.1 and in the five texts-five speakers run 60% of all spectra were assigned to the correct speaker. Using a binomial probability test (assuming equal probabilities for all groups, an assumption which is obviously not quite valid in the second run) it appears that all classification results are much better than what should be expected by chance. The number of discriminant functions used in the first run was four, whereas only three functions were

Table 4.3. Confusion matrix for 7×60 glottal pulse spectra, regardless of vowel characteristics. No attempt has been made to distinguish between vowels taken from the neutral or non-neutral text for speakers EK and LB, who are, therefore, represented by twice the number of items for the remaining speakers.

	EK	LB	TB	NR	PD	# items per speaker	% correct
EK	92	14	8	4	2	120	57
LB	17	67	21	13	2	120	42
TB	10	10	31	5	4	60	52
NR	1	6	4	35	14	60	58
PD	0	1	8	11	40	60	67
# items assigned to 120 speaker	98	72	68	62			

Table 4 4 Confusion matrix for 5 X 60 glottal pulse spectra, regardless of vowel characteristics
Vowels taken from readings of the neutral text only

	EK	LB	TB	NR	PD	# items per speaker	% correct
EK	47	7	3	3	0	60	78
LB	11	34	11	3	1	60	57
TB	14	15	18	7	6	60	30
NR	1	1	4	42	12	60	70
PD	1	3	5	12	39	60	65
# items assigned to speaker	74	60	41	67	58		

formed in the second and third runs. Judging both from the normalized discriminant function coefficients given in Table 4.5 (which indicate the contributions of the individual variables - i.e., filter levels - to the newly formed dimensions) and the co-ordinates of the group centroids in the newly

Table 4 5 A Normalized discriminant function coefficients resulting from an analysis in which 420 average critical band spectra of glottal flow pulses pertaining to seven texts, were treated as seven groups of 60 items each **B** Same data, but treated as two groups of 120 and three groups of 60 items

	Filter number	Centre frequency (in Hz)	Normalized Function#1	Discriminant Function#2	function Function#3	coefficients Function#4
A	1	122	-.07457	.31040	.15590	-.05207
	2	215	-.23130	.38137	.43752	.55441
	3	307	.53837	-.11491	.70626	-.15445
	4	400	-.09178	.35927	-.13546	-.17400
	5	500	.53087	-.36721	-.28168	-.02414
	6	630	.19588	.03511	-.06860	-.19607
	7	800	-.65850	.48219	.50587	-.52927
	8	1000	.07674	.50431	.58042	.27035
	9	1250	.26064	-.30422	.52123	.93602
	10	1600	.06139	.27166	-1.02230	-.32156
	11	2000	.17019	-.25727	.31053	.07942
	12	2500	.06830	.53089	-.01119	-.24731
	13	3150	-.10362	-.17247	.25648	.06855
% of variance explained			51.6	24.1	12.2	8.1
B	1	122	-.07568	.25443	.13183	
	2	215	-.23174	.31379	.39932	
	3	307	.53533	-.16563	.71806	
	4	400	-.08351	.36610	-.12176	
	5	500	.53312	-.34857	-.31626	
	6	630	.19576	.01969	-.07852	
	7	800	-.66258	.46085	.58982	
	8	1000	.07169	.55577	-.53535	
	9	1250	.27373	-.38618	.38101	
	10	1600	.05042	.32692	-.98659	
	11	2000	.16103	-.22794	.36623	
	12	2500	.07524	.57126	.07479	
	13	3150	-.09049	-.23171	.16644	
% of variance explained			54.8	24.4	12.4	

formed (so called canonical) space, the three analysis runs appear to lead to a similar structure. The locations of the group centroids in the canonical space are graphically depicted in Fig. 4.14, where the results of the first analysis run (seven group classification) are given for the three lowest order discriminant functions.

The similarity of the dimensions resulting from the individual runs does not mean that they are easy to interpret. Yet there is some indication of a relation between the first discriminant function and spectral slope in the frequency region below 1 kHz (given the high negative contribution of the 800 Hz band and the fairly high positive contributions of the 500 Hz and 307 Hz bands). Following the same reasoning, the second discriminant function might be related to spectral slope in the region of medium frequencies (fairly high positive contributions from the 2400 Hz, the 1000 Hz and, again, the 800 Hz band). Speakers EK and PD differ most on the (most important) first dimension, whereas EK and LB are farthest apart along the second dimension. The interpretation of the first dimension is confirmed by an inspection of the slopes in the low frequency region of the average spectra in Fig. 4.13. Speakers EK, LB and TB show slopes of approximately -9 dB per octave in the frequency range below 800 Hz whereas the slopes in the mean spectra of NR and PD are much steeper (-13 to -14 dB per octave). A similar approach of the second discriminant function does not lead to an unconditional confirmation of a simple interpretation of spectral slope in a definite frequency region. Rather it seems as if the high frequency slope must be weighed with overall spectral balance. Considering the difficulties encountered in

Fig. 4.14 Group centroids (= speaker positions) in the three dimensional space resulting from a discriminant analysis on seven sets of 60 glottal flow pulses. Each set of pulses is comprised of vowels taken from a reading of either the neutral or the non-neutral text. Speakers BK and LB are represented by two readings, speakers NR, TB and PD are represented by a reading of the neutral text only

interpreting the first and especially the second discriminant functions it is not surprising that an attempt at an interpretation of the third one failed completely.

From the confusion matrix in Table 4.2 it is quite clear that the fairly low percentage of correctly classified pulses in the seven texts-seven speakers run is due to the difficulty encountered in separating the two readings of EK and LB. From the remaining Tables, both pertaining to runs in which no within-speaker between-texts classification was attempted, it appears that the analyses form two groups of speakers, the first one comprised of EK, LB and TB whereas the second group contains the speakers NR and PD.

4.5.3.2 *Stability of results upon repeated measurement*

Turning now to the topic of the stability of the results within speaker but across different texts it can be pointed out that the plots in Fig. 4.14 indicate that the repeated measurements for speakers EK and LB cluster together very tightly. In Fig. 4.15 the spectral difference between the average spectra computed for both texts is depicted. It appears that the level difference is less than 2 dB in all 13 filter bands. Thus it can be said that - at least for these two speakers - the average glottal pulse spectra seem to be fairly stable as long as the speech conditions do not vary too much. This result, by the way, corroborates the contention that the accuracy with which the glottal pulses are reconstructed from the speech signals is high enough to retain general speaker characteristics and between speaker differences. Further results of the examination of the within speaker stability will be presented as part of the outcomes of the remaining parts of this study.

4.5.3.3 *Glottal pulse differences between stressed and unstressed syllables*

Although some 60% correct attribution of vowels to speakers is not a

Fig. 4.15 Difference between averaged pulse spectra of vowels taken from readings of different texts by EK and LB.

Displayed is the level in the neutral text minus the level in the non-neutral one.

particularly outstanding result, it is good enough to motivate a separate treatment of the seven texts in dealing with the stressed/unstressed distinction and the influence of individual vowels. These problems were approached in the same way as the speaker separation problem, viz. via multiple discriminant analysis.

The percentages of pulse spectra classified correctly as belonging to a stressed or to an unstressed vowel are presented in Table 4.6. A binomial test of probability shows that the 63.3% correct in the neutral text of LB is just significantly different from chance at the 5% level. The 83.3% correct for PD is, on the contrary, a very significant indication of a difference between glottal pulses from stressed and unstressed syllables. The difference between average glottal pulse spectra for stressed and unstressed vowels is shown in the lower panels in Fig. 4.13. It can be seen that the differences are very small and, what is worse, far from systematic. The within-speaker differences for both EK and LB are especially striking: in the neutral text their stressed vowel pulses tend to have a slightly lower spectral level in the higher frequency bands, a situation that runs counter to the *a priori* expectation. In the non-neutral text, however, LB and EK show the expected trend, as do the neutral text data of PD and NR. Again TB seems to share the company of EK's and LB's readings of the neutral text in that in his data there is no systematic and interpretable difference between the spectra of the stressed and unstressed syllables.

4.5.3.4 *The influence of supraglottal load*

An attempt has been made to classify the glottal pulse spectra according to the vowel they pertained to, again separately for the seven texts. It was decided not to attempt to distinguish between stressed and unstressed vowels. The introduction of the stressed/unstressed distinction would have raised the number of categories to 23 and consequently lowered the average number of occurrences per class to three with some empty categories for most speakers. Ideally the classification accuracy should be at chance level, at least if the vowel type does not influence the glottal pulse shape and if the contribution of the vocal tract formants is completely eliminated by the inverse filtering. For all speakers five or six discriminant functions were

Table 4.6. Percentage of glottal flow pulses correctly classified as pertaining to a vowel taken from a stressed or from an unstressed syllable. EK_i and LB_i are the readings of the non-neutral text by speakers EK and LB. All remaining data pertain to the neutral text. The total number of vowel tokens is 60 for each text.

	EK _a	EK _i	LB _a	LB _i	TB	NR	PD
# correct	40	47	38	40	42	48	50
% correct	66.7	68.3	63.3	66.7	70.0	80.0	83.3

Fig. 4.16 Spectral differences between glottal flow pulses reconstructed from low, mid and high vowels for each of seven readings of two texts by five talkers. Upper trace in each panel is 'Low-Mid', middle trace is 'Low-High', and lower trace is 'Mid-High'

formed and the accuracy of classification appeared to vary from 60% to 75% using 12 vowel categories, i.e., all monophthong vowels of Dutch except /ə/ plus the /ə/.

Neither from the discriminant functions nor from the confusion matrices or the positions of the group means in the canonical space did a clear pattern emerge which was consistent across speakers. One way to explain the high accuracy of the classification is to consider that many categories contain no more than two or three observations and in a 13-dimensional space it is very unlikely that the procedure should be unable to find a hyperplane to separate them from the remaining groups. For that reason we decided to divide the set of 12 vowels into three larger groups by combining the high vowels /u, y, i/ the mid vowels /e, I, ε, o, ə, ʌ, ə/ and the low vowels /a, ʌ/. The regrouping

Fig. 4.16. Continued.

has the added advantage of reducing the probability that the outcome will be obscured by the need for the procedure to distinguish between individual vowel classes that must be expected to have similar effects on glottal pulse forms because of very small differences in F_1 .

For all datasets the largest possible number of two discriminant functions was formed. The number of correctly classified pulses ranged from 37 to 50, or from 61.7% to 83.3%. An examination of the confusion matrices showed a fairly consistent pattern. The number of high vowel pulses that were mistakenly classified as belonging to a low vowel never exceeded 2 and no single low vowel was mistaken for a high one. Pulses pertaining to mid vowels show a tendency to be confused with low rather than with high vowels. The interpretation of the discriminant functions is facilitated by a visual inspection of the spectral differences between the group centroids in the original

13-dimensional space. These differences are shown graphically in Fig. 4.16. The differences between the low and mid vowels is, in general, smaller than that between low and high vowels or mid and high. More interesting, however, is the fact that the pattern of distances between the high vowels and the two remaining groups is similar for all seven datasets. It is characterized by a lower level of the pulses reconstructed from high vowels in the frequency region around 600 Hz and by a tendency for the pulses of the high vowels to have a comparatively high level in the frequencies above 1.6 kHz. It can be seen from Fig. 4.16 that the spectral distances between low mid and high vowels are fairly stable across texts. Although the difference spectra computed from the readings of the neutral and non-neutral texts by EK are somewhat more similar than is the case for LB, both speakers show similar trends in both texts. In interpreting the between-text differences due attention should be paid to the fact that the vowel sets are rather inhomogeneous, both with respect to the number of specific vowels per set and to the ratio of stressed and unstressed vowels.

4.5.4 Discussion

The amount of information on personal voice quality characteristics that is retained in the time-averaged spectra of the glottal volume velocity flow pulses is not sufficient to effect a reliable recognition of speakers. Regardless of the fact that the speakers have voice qualities which are auditorily very different, the discriminant procedure has great difficulty in separating the speakers. The moderate success of the discriminant procedure is partly due to large within speaker variation. Part of the variation undoubtedly is error variance in the most straightforward meaning of that word, i.e., due to the failure of the inverse filtering procedure to recover the glottal flow pulses exactly in all vowels. We do not think, however, that the contribution of this error source is very important. Visual comparison of the duty cycle in the reconstructed glottal flow pulses with the open ratios that can be inferred from the EGG, suggests that the reconstructed pulses are on the average quite accurate. It has already been pointed out that the fairly small within-speaker differences between the average pulse spectra can also be taken as a confirmation of the accuracy of the inverse filtering technique employed in this study.

One might argue that personal voice quality is related to dynamic changes in pulse spectra rather than to stationary interpersonal differences. Dynamic aspects undoubtedly play a role in distinguishing between speakers. Yet long-term critical bandwidth spectra computed over the voiced segments in fragments of the reading texts as short as 10 seconds contain enough information to yield 100% success in a speaker verification experiment (see Chapter 5). Similar and even more convincing results can be found in Hollien & Majewski (1977). The most probable cause of the relative failure of the discriminant analysis to separate the speakers is, therefore, the fact that

glottal flow pulses pertaining to vowels of a duration not exceeding 120 ms and chosen at random from a connected text do not contain enough information to identify the speaker. Interestingly, Koopmans-van Beinum (1980), who asked 100 listeners to identify vowels excerpted from free conversation, reports an average percentage of correctly identified stimuli slightly lower than 35%. We know of no comparable experiments with speaker recognition. However things may be, coarticulation effects in running speech seem to be pervasive and strong enough to cause the standard deviation of the distribution of any individual parameter to attain extremely high values. As a direct consequence a great number of samples is needed before the sample mean is a dependable indicator of the population mean. Seen against this background the correct classification of some 60% of the glottis pulse spectra averaged over extremely short time intervals is not a particularly bad result after all.

The fact that between-speaker differences in the spectra of the glottal flow pulses are smaller than one might be inclined to anticipate on intuitive grounds is in agreement with the conclusions of Sundberg and Gauffin (Sundberg & Gauffin 1979, Gauffin & Sundberg, 1980) who found that differences between the glottal source spectra of a professional singer and an untrained speaker were very small in phonations at an intermediate effort level.

The study of the influence of stress level on the spectra of the flow pulses was mainly motivated by remarks in Miller's original paper (Miller, 1959), who suggested that in stressed vowels the level of the upper harmonics is considerably higher than in unstressed vowels. A tendency for the level of the higher harmonics to rise with increasing phonatory effort is also reported in Gauffin and Sundberg (1980). From Fig. 4.13 it is apparent that a comparable trend is not uniformly present in our results. There are considerable differences between speakers, which in some cases seem to support the hypothesis put forward by Miller; other cases however, seem to suggest the exact opposite hypothesis.

One explanation for these seemingly contradictory results can perhaps be found in the different terms used to describe the underlying phenomenon. The term 'effort' used by Gauffin and Sundberg is probably more appropriate than 'stress'. Now it is well known that in Dutch 'loudness' (which can be associated with phonatory effort) comes at best at the third place when ranking physical effects underlying the production and perception of prominence, far beyond pitch movements and syllable duration of which the former is by far the most important (van Katwijk, 1974). Thus word stresses and even sentence stresses are not likely to induce a great amount of extra phonatory effort in their production and consequently the glottal flow pulses might not unconditionally be expected to contain a considerably higher amount of energy in the upper frequency bands.

We can reconcile the seemingly contradictory results by assuming that

speakers have available a number of strategies for signalling prominence. To the extent that small changes in loudness are associated with the distinction between stressed and unstressed syllables PD, NR, and in the reading of the non-neutral text also LB, seem to increase loudness by increasing the slope of the falling edge of the glottal flow pulses, thereby raising the relative level of the higher harmonics. To the - probably negligible - extent that EK, TB and LB in reading the neutral text, use loudness as an additional cue to signal prominence they seem to make use of an increase of the amplitude of the pulses, thereby raising the overall level of the spectrum without changing its slope. Again a resemblance to results obtained by Gauffin and Sundberg should be pointed out. They found a tendency for the singer to produce a higher sound pressure level by increasing the amplitude of the glottal flow pulse without changing the form, whereas the untrained speaker made more use of the possibility to increase the slope of the falling edge, perhaps even at the expense of the amplitude of the pulse. Thus EK (who is, incidentally, a good amateur singer) would come closest to a strategy like the one employed by Sundberg's and Gauffin's singer, especially in his reading of the non-neutral text. The question should be asked whether the results obtained from the neutral text readings by EK, LB, and TB justify the assumption that pulse forms differ between unreduced unstressed and stressed vowels. The difficulties encountered in deciding which vowels from these texts should be considered as stressed suggest a negative answer to that question. Stressed/unstressed decisions were much easier in the non-neutral texts and in the readings of the neutral text by NR and PD. Where differences between pulse spectra of stressed and unstressed syllables are clearly present, LB, NR and PD seem to use an other strategy than EK. They rather join the untrained speaker from the work of Sundberg and Gauffin.

Meanwhile it can be noted that the separation of pulses derived from stressed and unstressed vowels provides one means for retaining part of the dynamic behaviour of the voices. Interestingly, EK was judged to be least tense in the evaluation experiment described in Chapter 3, and also as most pleasant to listen to. Apparently listeners prefer phonations tending in the direction of the 'flow' type over voices that tend to a 'pressed' phonation.

With regard to the interaction between voice source and vocal tract filter it is rewarding to find that the rather crude spectral representation used in this research is able to effect a clear distinction between the group of high vowels (with their low F_1) and the other vowel sounds. Due to the crudeness of our analysis the results have, however, no straightforward impact upon Rothenberg's model proposals, the more so since in his most recent work he seems to take the extremely cautious position that his model applies only to relatively open vowels (Rothenberg, personal communication).

It should be realized, incidentally, that Ananthapadmanabha & Fant (1982) offer a potentially more powerful framework for explaining systematic between-vowel differences in the spectra of glottal flow pulses in that they

incorporate the possibility that the pulse spectra contain zeros at the frequencies of the supraglottal resonances. FFT spectra of flow pulses recovered by means of inverse filtering indeed appear to contain clear zeros at the frequencies of vowel formants (Cranen & Boves, 1983).

In order to decide whether the classification results in our study are due to spectral zeros or to overall pulse shaping effects (or to both as the discriminant functions seem to suggest) a more detailed experiment has to be carried out. Ideally, such an experiment should include some independent measurement of the input impedance seen when looking from the glottis into the vocal tract.

4.6 CONCLUSIONS

The present chapter contains the report of a number of efforts aimed at the separation of the contributions of phonation and articulation to the radiated acoustic speech signal and at a characterization of a number of speakers on the basis of their phonation. To these ends a technique has been developed for reconstructing the glottal volume velocity pulses from the acoustic speech wave. It was also attempted to assess the accuracy of the inverse filtering technique by means of comparing its output with recordings of physiological activity on the level of the larynx, viz. the electroglottogram, the photoglottogram and the subglottal pressure. Since data on measuring and processing P_{sg} are as yet very scarce, our measurement techniques have been explained in detail. We have shown that the pitch synchronous variations in P_{sg} present in previously published recordings are not due to the Bernoulli effect but rather reflect the resonances of the subglottal system which are excited in much the same way as the supraglottal resonances which make up speech. With regard to the frequencies of the subglottal resonances our data seem to agree with the measurements of Ishizaka, Matsudaira and Kaneko rather than with the outcomes of a modelling study by van den Berg. Lastly, and most importantly, our study of the P_{sg} signals revealed that a model based on stationary flow had to be abandoned in favour of a model in which due attention is paid to the intermittent oscillatory nature of the flow.

The latter finding prevented the use of P_{sg} as a signal which might allow accurate estimates of the moments of glottal opening. Also the assumption that the moment of maximum glottal flow coincides with the moment of maximum glottal area is affected, as is, in fact, the complete linear model underlying the idea that phonation and articulation can be separated.

Yet it appeared possible to design a closed glottis interval inverse filtering technique which seems to have considerable theoretical validity. By using the EGG in order to determine the CGI our technique is suitable for the processing of vowel sounds taken from running speech. The inverse filtering technique was applied to 420 vowel tokens taken from seven reading texts produced by five adult male speakers, mainly in order to characterize the

speakers on the basis of their phonatory habit. It was also investigated whether speakers differ in the way in which they differentiate between stressed and unstressed vowels. Our results can be summarized in the statement that a long-term average spectrum of the glottal flow pulses captures only part of the between-speaker differences in phonation. The remaining part of the between-speaker differences must most probably be attributed to dynamic short-time variations in phonation. The latter aspect could hardly be covered by our approach.

Development of acoustic measures to explain perceptual ratings

5.0 INTRODUCTION

After the digression into the domain of physiological measurements we will return to the world of acoustics. We will revisit the topic of the extraction of glottal flow waveforms from the acoustic speech wave with the aim to investigate to what extent global information on glottal waveshapes can be obtained by means of fairly simple acoustic measurements. More specifically, we will investigate whether the long-term average spectrum, computed during the voiced intervals only, contains relevant information on speaker-dependent characteristics of the glottal wave.

The return to the acoustic domain is not difficult to motivate. Acoustic measurements are relatively easy to perform and, in any case, do not cause any discomfort to the subjects. Furthermore we will focus our attention almost exclusively on acoustic measurements that can be carried out entirely automatically. The reconstruction of glottal flow waveforms, on the other hand, still requires some operator intervention and it is not likely that the procedure can be fully automated in the near future.

Originally the idea was to develop acoustic measures to 'explain' the factor scores of the speakers obtained in the judgement experiment described in Chapter 3, i.e., to develop measures that correlate highly with one of the dimensions of the judgement space. The results of the factor analyses, however, suggested that the dimensions of the judgement space were very broad and somewhat vague. A similar result was found by Hammarberg, Fritzell, Gauffin, Sundberg & Wedin (1980). They note that, although individual semantic differential scales may have unambiguous acoustic correlates, dimensions of the judgement space need not be related to acoustic measures in a simple way, because they tend to be amalgams of fairly heterogeneous scales. This made us decide to try and develop a large number of quite different acoustic measures and investigate the correlations of those measures with both the factors from the factor analysis and a subset of the individual scales. We will limit the discussion of the relation between acoustic measures and individual scales to those retained in the condensed scaling instrument proposed in section 3.4.3.

Our search for acoustic measures has, of course, been guided to a considerable extent by the results of the analysis of the ratings. Firstly, we have attempted to obtain acoustic measures corresponding to speaker dependent properties of the glottal flow pulses. Secondly, it was clear from the very beginning that the great number of scales in our semantic differential referring to the variability in the deliveries would necessitate the development of measures of dynamic properties of the speech signals. Thirdly, our intention to separate phonation and articulation forces us to develop measures that relate exclusively to articulation.

Quite generally, the within-speaker stability of a measure and its ability to discriminate between different speakers have served as criteria in judging the usefulness of acoustic measures. It should be noticed, incidentally, that these criteria operate in much the same way as the reliability coefficients in the analysis of the perceptual ratings. For practical reasons the 'reliability analysis' has only been carried out on the Nijmegen material.

5.1 ACOUSTIC MEASURES OF TEMPO

The problem of finding global objective measures that can explain subjective ratings on scales like 'quick $\leftrightarrow$ slow' and 'dragging $\leftrightarrow$ brisk' is deceptively simple. That is to say, it is easy to measure the time each speaker spent in reading the complete text, or in reading individual paragraphs or sentences. It is also very easy to convert those time-measurements into a more homogeneous measure like the number of syllables per second, that is not dependent on the differing sizes of texts, paragraphs and sentences. Unfortunately, these simple measures appeared to be almost uncorrelated with subjective tempo judgements in our material.

This result motivated a search of the literature for comparable data and hopefully also an explanation. The number of studies on the perception of tempo in running speech appears to be surprisingly small. From the literature review in Grosjean & Lane (1981) the reader cannot but get the impression that most of the research has concentrated on studies of differences between texts read out with 'normal' speed and readings with deliberately slowed down or sped up tempo. The number of studies dealing with perceived tempo in isolated sentences is somewhat greater, but Butcher (1981) doubts whether the acoustic and articulatory parameters which have been shown to affect tempo judgements in isolated utterances play the same role in judgements of tempo in longer stretches of speech. Butcher used four parameters to describe his speech material, viz.:

- articulation rate i.e., number of syllables per second of utterance time (excluding pauses),
- speech rate i.e., number of syllables per second of time available (including pauses),
- relative break frequency i.e., number of pauses in utterance groups divided by the number of syllables,

- and percentage break time i.e., total pause time divided by total time available to the speaker.

Only the last parameter is amenable to automatic measurement. The remaining parameters involve some kind of transcription, unless one is willing to define syllables in a purely acoustic sense, in which case they may be counted automatically (Rietveld, 1983). Butcher's conclusion, however, that non-juncture pauses (hesitation pauses), juncture pauses within sentences and pauses between sentences have different effects on tempo perception seems to exclude a totally automatic signal processing.

As tempo does not play a major role in our factor solution (cf. Chapter 3) we decided to be content with a very coarse approximation of some of the measures used by Butcher. To that end we measured the duration of the first 11 sentences of both the neutral and the non-neutral texts in both stimulus sets. The total time taken from the start of the first sentence up to the end of the 11th was also measured. The measurements were made by handclocking a stopwatch when listening to the recordings played at half the original speed. Obvious false starts were measured separately and subtracted from the time per sentence and from the total time. Minor reading errors were neglected, both in measuring time and in counting syllables per sentence, which was done from paper. This means that the number of syllables was assumed to be the same for all speakers and to be equal to the count that would be obtained from a transcription of a careful and error-free reading. From the measurements speech rate was computed for individual sentences. Mean speech rate for the first part of both texts was computed as well as the relative break time for both texts.

Correlations were computed between the speech rate measures for individual sentences and the mean speech rate of the text to which the sentences belong. Table 5.1 contains a summary of the results. Correlations between individual sentences are not included in the table. Some of these correlations were quite low and sometimes even negative. This is due to the different phonetic structure of the sentences and undoubtedly also to their different length. Longer sentences are likely to contain more pauses than very short ones, a situation which is bound to give rise to an inverse relation between length of a sentence and speech rate. It must, therefore, be concluded that the conversion of speech time per sentence to speech rate does not necessarily enhance the homogeneity of the data very much. It is interesting to note in Table 5.1 that the percentage of between sentence pause time is much smaller in the Amsterdam than in the Nijmegen material. We have no explanation for this observation.

Apparently, there is no relation between the average articulation rate and the total pause time, nor with the relative pause time. In both data sets these correlations are very low and insignificant (cf. Table 5.2). With respect to the relation between total pause time and relative pause time the two data sets seem to be somewhat different, in the Amsterdam material these two mea-

Table 5.1. A: Speech rate in syllables per second (upper rows) and percentage pause time (lower rows) for the 12 stimuli from the Nijmegen and the 11 stimuli from the Amsterdam material. B: Correlation between speech rate measures obtained from 11 sentences of each text and speech rate computed from the total texts.

Nijmegen a)							Amsterdam b)						
A	EK	DJ	LB	NR	TB	PD		1	2	3	4	5	6
i	5.2	4.7	5.1	4.9	4.6	4.1	F	5.4	5.3	5.9	4.9	5.0	
a	5.9	5.4	5.5	5.3	5.1	5.3	M	5.4	6.1	4.7	5.8	5.0	5.5
i	14.61	13.54	17.87	19.64	17.12	6.97	F	8.48	7.07	6.90	4.73	5.84	
a	9.05	11.30	13.69	13.92	10.79	6.25	M	11.30	8.56	6.08	9.39	11.54	13.69
<hr/>													
B sentence number	1	2	3	4	5	6		1	2	3	4	5	6
	7	8	9	10	11			7	8	9	10	11	
<hr/>													
	.80	.20	-.18	.42	.40	.50		.86	.81	.92	.89	.63	.53
	-.48	.72	.46	-.02	-.36			.80	.78	.94	.70	.88	

- a. Upper case letters identify talkers. Lower case letters identify texts: i the non-neutral and a the neutral one.
 b. F and M refer to female and male talkers. M1 in the Amsterdam material is identical to DJa in the Nijmegen material; the same applies to M6 and LBa.

asures are virtually identical, whereas the correlation coefficient is only just significant in the Nijmegen material.

That speech rate per sentence is not a stable measure does not by necessity mean that it is useless to characterize speakers when an average is taken over a number of consecutive sentences. To investigate this issue we have carried out an analysis of variance on the data for mean speech rate per text with text and speakers as factors; the text $\times$ speaker interaction provided the error term. This analysis could, of course, only be performed on the Nijmegen data, where two complete texts of each talker were available. The results

Table 5.2. Correlations between speech rate in syllables per second, absolute pause time and relative pause time. The upper right hand of the matrix refers to the Amsterdam data; the lower left hand part pertains to the Nijmegen material.

	syl/s	absolute pause time	relative pause time
syl/s	-	.084	.297
abs. pause time	.345	-	.975
relative pause time	-.080	.588	-

indicate that the factor 'text' is highly insignificant, whereas the factor 'speakers' is significant at the 5% level. This suggests that mean speech rate might, after all, have played a systematic role in invoking perceptual judgements of our listeners.

Butcher (1981) identified percentage break time as the most important single factor affecting global tempo ratings of whole texts but at the same time found that the amount of pause time between sentences is especially important. We have measures of the percentage between-sentence pause time for both texts in the Nijmegen experiment. An analysis of variance with 'text' and 'speaker' as main effects and the text $\times$ speaker interaction as error term led to the result, however, that neither effect is significant at the 10% level.

5.2 ACOUSTIC MEASURES OF PITCH

A fair number of scales in our semantic differential refer to pitch level, pitch variation and - more generally - to voice dynamics. Moreover, these scales appear to yield the most important factor in the analysis reported in Chapter 3. When trying to deal with voice dynamics one is very strongly hampered by the consequences of the fact that phonetics has traditionally restricted its research to short, isolated utterances in which dynamic aspects do not play a role on a level above that of coarticulation between neighbouring speech sounds. To be fair: a large body of research and knowledge on intonation does exist and we feel that it should be brought to bear on the study of voice dynamics. Ideally we should have made transcriptions of the suprasegmental properties of our stimulus texts, e.g., using the intonation transcription system developed for Dutch ('t Hart & Collier, 1975) and we should have tried to relate these transcriptions to the ratings by our listeners. There are, however, two reasons for refraining from that undertaking. Firstly, transcribing the intonation in so extensive a corpus as ours is a formidable and extremely difficult task. Secondly, if a transcription could be made it is not at all certain that meaningful and manageable procedures can be developed to convert the transcriptions into variability measures amenable to statistical processing. Therefore we decided to confine ourselves to fairly global pitch measures that are both easy to obtain and to process.

All measures that have been used derive from computer processing of the output of an analog pitch period detector described in van Rossum & Boves (1978). This device performs a voiced/unvoiced decision and it measures the duration of all individual pitch periods during the voiced intervals. The output of the instrument is fed into a minicomputer, that stores the information, consisting of a sequence of durations of unvoiced intervals and pitch periods for further processing. The raw data on pitch period duration form the bases of one measure of so called pitch perturbation (see 5.2.3 below). For all remaining measures the raw data are first converted into a uniformly

sampled F_0 track, with a 10 ms sampling interval. The uniformly sampled pitch function is passed through a simplified 5-point median filter (Wichern, 1978). Visual inspection of the results for a wide variety of speech, speakers and recording qualities suggests that further correction of the pitch data is not worthwhile.

In order to investigate whether the measures that have been implemented discriminate between and are stable within speakers, the following procedure was adopted. The neutral text was divided into four non-overlapping fragments of 30 s duration each; the non-neutral text could be divided into six such fragments. A fragment begins at the same point in the text for all six speakers, but since the reading speed differs somewhat between speakers, the end points are not aligned. Thus, strictly speaking, between speakers the fragments are at best comparable but not identical. All measures were computed for each fragment from each speaker and the results subjected to an analysis of variance with texts and speakers as main effects and the texts $\times$ speakers interaction as the error term. The factor 'speakers' proved to be significant at the 1% level for all the pitch measures to be discussed, whereas the factor 'text' did never reach the 5% significance level except for the percentage of time that the signal was voiced. Therefore we may conclude that all measures to be discussed are stable within speakers and discriminate between speakers. Accordingly we will not return to this issue when dealing with individual measures.

The SPSS program ONEWAY (univariate analysis of variance with post-hoc contrasts) was used to calculate between speaker contrasts. The calculation of the contrasts enables us to see whether the significance of the speaker effect is due to a more or less regular spacing between speakers or by one strongly deviating speaker. The detailed results will be discussed when dealing with the individual measures. At this point suffice it to say that two contrasts were calculated, viz. the least (LSD) and the most conservative (Scheffé contrasts) estimates available with the program. The contrasts will mainly be used as a criterion for choosing one out of several alternative measures.

5.2.1 *Measures of central tendency*

We have computed four central tendency measures, viz. mean, mode and median value of the voiced samples expressed in Hz and the mean value of the voiced samples expressed in semitones. Given the relatively great number of samples in the distributions a high correlation between the three linear measures of central tendency was to be anticipated. The semitone means appear to be virtually indistinguishable from their linear counterparts (correlations of .997 between semitone mean and linear mean, of .892 with mode and of .994 with median in the Nijmegen material; in the Amsterdam data these coefficients are equally high). Comparable results were found in a study

on quite different (and much more varied) speech material in Cranen, Boves & van Bezooijen (1983)

Study of the post-hoc contrasts did not show sizable differences between the four measures either. Therefore, we decided - admittedly somewhat arbitrarily - to retain the linear mean, mostly since it is the measure which is easiest to compute

5.2.2 Measures of variability

As with measures of central tendency a number of different measures of variability are known from the literature. We have experimented with a large number of global variation measures, but have eventually retained only a standard deviation of the 10 ms pitch samples expressed in semitones. The reason for discarding all other measures was that they all correlated very highly. Before computing the standard deviation measure outliers were discarded by deleting all samples deviating more than 3σ from the mean.

From a perceptual and also - at this moment perhaps even more importantly - from a statistical point of view standard deviations should not be compared across samples unless they are normalized with respect to the sample mean. This suggests the somewhat less well-known coefficient of variation, defined by Pearson as the quotient of the standard deviation and mean as the most appropriate measure of the pitch variation.

Statistics such as the coefficient of variation are very useful for describing global properties of the distribution of pitch samples in the fragments. They do not, however, give insight into short-time characteristics of the pitch contours. Nevertheless, these short-time characteristics most probably influence our perception of pitch related attributes of running speech. Therefore we have implemented a number of measures that might yield some information on short-time and medium-time pitch fluctuations.

5.2.3 Pitch perturbation measures

Quite a lot of literature (refer to Askenfelt & Hammarberg, 1981, for a review) exists which discusses the relation between cycle-to-cycle fluctuations in the duration of pitch periods (which become audible when they exceed a certain threshold and are known as pitch perturbation) and perceptual qualifications like 'hoarse', 'harsh', 'unsteady' and 'unpleasant'. From the negative connotation of the adjectives it is clear that a considerable amount of pitch perturbation is not a desirable property of a voice. Minute cycle-to-cycle variations are inseparable from the action of an extremely complex vibratory system like the human vocal folds. Without disclaiming the importance of mechanical and aerodynamic contributions Bear (1979) gives compelling evidence of neuromuscular causes of at least small amounts of pitch perturbation.

A large number of physical measures of pitch perturbation have been proposed in the literature. Most of the literature and research on this topic seems to deal with pathological voices. We have implemented and compared a large number of pitch perturbation measures, but only using speech produced by healthy subjects. Also we have included a comparison of pitch perturbation measures derived from the raw pitch period duration data and measures based on the uniformly sampled (and slightly smoothed) pitch tracks. The aim of the latter comparison was to assess the importance of errors in the analog pitch detector in determining a measure of pitch perturbation. It appeared that the importance of pitch detector errors is strongly dependent on the type and especially on the duration of the speech material used. When processing 30 s fragments from connected texts read by normal subjects pitch detector errors were negligible, repeated measurements and comparable definitions led to equivalent results. In a companion study, in which short emotional utterances were analyzed, the measure based on the uniformly sampled and smoothed pitch curve appeared to differ widely from the corresponding measures derived from the raw period duration data (Cranen et al., 1983). These results suggest two implications which are not necessarily mutually exclusive. Firstly, pitch perturbation in neutral speech produced by healthy and relaxed talkers may be quite a different phenomenon than pitch perturbation in pathological voices or in emotional speech. Therefore, the different phenomena may call for different measurement procedures. Secondly, it is possible that pitch detector errors do play a crucial role in measurements of both normal/neutral and pathological/emotional speech, but that their effects are to a large extent alleviated if one is in a position to average over fairly long stretches of speech. It is conceivable that the minimum averaging time needed in order to obtain similar results from the two types of measurements will prove to be longer in pathological and emotional speech.

An analysis of the results obtained with normal speech supported the conclusion of Askenfelt and Hammarberg that the standard deviation of the distribution of between-sample differences df_n is probably the pitch perturbation measure to be preferred, if only because of its superior statistical properties. Here df_n is defined by

$$df_n = f_n - f_{n-1} \quad (1)$$

where f_n is the value of the n -th sample of a pitch track. The pitch perturbation measure DF_0W employed in the present study is simply the standard deviation of the df_n .

Pitch perturbation measures give an indication of very short-time pitch fluctuations. In addition two medium-time pitch variation measures were implemented, which are defined as

$$V_{xx} = \frac{1}{N} \sum \frac{|f_n - f_{n-xx}|}{\frac{1}{J} \sum_{j=n-xx}^n f_j} \quad (2)$$

where the summations over n and j include only pitch samples $f_n > 0$, i.e., only voiced samples. The divisors N and J are the number of non-zero samples in the summations. We have computed (2) with $xx = 5$ and $xx = 10$, i.e., the pitch changes during intervals of 50 and 100 ms. Note that these measures are very much like the Average Rate of Change of Fundamental Frequency (ARCF) measure used by Askenfelt and Sjölin (1980) in a study on the speech of depressed patients.

5.2.4 The voiced-unvoiced ratio

An additional measure which can be looked upon as related to the F_0 -distribution is the number of 10 ms samples that are voiced in each text. The proportion of the signal which is voiced might influence the perceived sonority of the speech. In any case it appeared that the number (or proportion) of voiced samples per text is characteristic of the speakers in the ANOVA terms explained at the beginning of this chapter. (Recall that for this measure there was also a significant difference between fragments of the texts.)

5.2.5 The pitch data

Table 5.3 contains the pitch data which are to be used in the acoustic explanation of perceptual ratings. The measures presented in that table are those which have proved to function more or less independently in the prediction of the judgements. Many of the pitch measures computed appeared to be highly correlated. Such measures cannot function independently when it comes to explaining some other set of observations. This remark applies to the group of measures of central tendency and to the group of variation measures.

Note that neither of the medium-time variation measures V_{xx} has been retained. These measures appeared to be virtually identical to the pitch perturbation measure DF_0W (correlations $> .93$ both in the Amsterdam and the Nijmegen material). Most probably this is due to the fact that none of our talkers had a clearly pathological voice. In the absence of gross between-cycle variations the pitch perturbation measure DF_0W is mainly determined by the slow and regular pitch fluctuations related to sentence intonation. This again confirms the suggestion that the concept of pitch perturbation when applied to essentially normal voices, must be interpreted differently than when one is dealing with pathological voices. That we have retained DF_0W instead of one of the V_{xx} measures is motivated by the fact that we have to

Table 5.3. Summary of pitch measures retained as predictors of perceptual ratings. Part A pertains to the data of the Nijmegen material. Upper case letters identify the talkers, lower case letters identify the texts (i: non-neutral, a: neutral). Part B contains the Amsterdam data Fm1-Fm5 are female, M11-M16 are males. M11 is identical to DJa; M16 is identical to LBa.

	Talker	% voiced	F ₀ (in Hz)	coefficients of variation (semi tones)	pitch perturbation
A	EKi	57.03	89.11	26.15	.205
	EKa	59.84	93.66	23.78	.207
	DJi	44.64	107.78	23.17	.181
	DJa	46.11	105.08	28.08	.296
	LBi	46.44	91.23	23.55	.194
	LBa	52.87	99.00	18.17	.164
	NRi	45.01	126.20	18.25	.211
	NRa	45.07	120.50	18.85	.206
	TBi	50.23	110.88	24.40	.189
	TBa	49.24	109.77	24.53	.180
	PDi	53.23	135.05	11.65	.162
	PDa	58.26	135.81	12.97	.173
B	Fm1	48.11	203.30	14.43	.296
	Fm2	46.19	203.78	19.97	.336
	Fm3	47.54	195.33	11.03	.256
	Fm4	51.49	236.71	9.34	.315
	Fm5	46.55	196.42	21.87	.321
	M11	46.11	105.08	28.08	.296
	M12	54.37	121.89	32.87	.385
	M13	43.69	140.65	21.40	.275
	M14	39.86	116.88	25.31	.272
	M15	44.65	122.08	26.75	.270
	M16	52.87	99.00	18.17	.164

explain ratings of huskiness. Most likely the signal characteristics which have given rise to the impression of huskiness are better retained in the DF_0W measure than in any of the V_{xx} measures.

From the figures in Table 5.3 a number of conclusions can be drawn. With respect to the Nijmegen data it can be seen that on the average the talkers are fairly stable; there are few exceptions to the rule that the measures computed for one text are virtually identical to the corresponding measure derived from the other text. This confirms the outcomes of the analyses of variance alluded to above, which invariably resulted in a significant talker-effect but not in a significant effect of the factor text. The most prominent within-person differences are the percentage of time that the signal was periodic for talker LB and the pitch perturbation for DJ. The different proportions of voicing for LB are difficult to explain. It is conceivable that the effect is related to a different reading style, since a comparison of the speech rate data in Table 5.1 shows that the neutral text was read considerably faster than the non-neutral one. Reading rate differences of comparable (and even larger) size in other talkers do not lead to corresponding differences in the proportions of time the signals are considered as voiced, however. The different percentages of between-sentence pause time do not provide a sound explanation for the greater proportion of voiced intervals in LB's neutral text either.

Contrary to what one might expect the long time between the two recording sessions with DJ is not likely to be the primary cause of the large between-text difference in pitch perturbation. Different qualities of the recordings are a more probable cause. To see this, it suffices to compare the overall levels of the pitch perturbation measure in the two data sets. Setting apart the two stimulus tapes that were shared, it appears that the pitch perturbation measures computed for the Amsterdam material are uniformly (much) higher than in the Nijmegen texts. It is extremely unlikely that this difference is due to the choice of the subjects; in any case, the Amsterdam talkers do not make the auditory impression of sounding harsher and rougher than the talkers from the Nijmegen experiment. Thus it seems that the conclusion that the perturbation measure used in this study (the standard deviation of the between-period differences df_n) is statistically stable holds only as long as the recording conditions are constant. Especially when analog equipment is used to extract the fundamental frequency, comparisons between independently recorded collections of speech material should not be made at all, or only with extreme caution. This admonition applies to normal speech where both the absolute level and the between-speaker differences in pitch perturbation are exceedingly small and where these differences might be of the same order of magnitude as differences between drop-outs and jitter of tapes and taperecorders. However, even when it comes to a comparison of pathological and healthy subjects it seems to be advisable to hold all recording and signal processing conditions as constant as possible. Further research is needed in order to enable meaningful comparison of independently published results.

5.3 LONG-TERM AVERAGE SPECTRA

Spectral representations of speech signals are, of course, very common in acoustic phonetics, where the term 'spectrum' almost invariably refers to a short-time amplitude or energy spectrum i.e., (the square of) the modulus of the Fourier transform of a short segment of a speech wave, isolated from its context by a weighting function in the time domain that has non-zero values over the interval of interest only. Irrespective of the way in which these short-time spectra are represented, they serve to identify speech segments at (or even below) the phone level. Long-term average spectra (LTAS), on the other hand, cannot be used to describe low-level linguistic units, but they might provide useful information on the individuality and/or physical and mental status of the speaker (Kosiel, 1973; Hollien & Majewski, 1977; Hammarberg et al., 1980). Long-term average spectra are mostly obtained by detecting and averaging the output of a number of band-pass filters; not only is there a choice of bandwidths and averaging techniques, the results can be represented in different forms too. A comprehensive discussion of these issues can be found in Boves & Cranen (1982-b). Here we will confine ourselves to the major points.

5.3.1 *The generation of LTAS*

The first decision to be made when designing a system to produce LTAS concerns the number and bandwidths of the filters. In all our work use has been made of the Parallel Power Density Measurement (PPDM) system described by van Rossum (1981). This system is equivalent to one designed and described earlier by Pols (1977). Essentially this is a set of three bandpass filters with a bandwidth of 90 Hz at centre-frequencies of 122, 215, and 307 Hz, 16 1/3 octave filters spanning the frequency range from 400 Hz up to 12.5 kHz and a 'linear' channel covering the band from 85 Hz up to 14 kHz. The set of filters can be said to approximate the so called critical band filters that play an important role in auditory research and theory (Plomp, 1976). All 20 channels are provided with detectors that have a logarithmic response characteristic and inherent integration time in the order of 10 ms. The outputs of the 20 detectors are scanned at a frequency of 100 Hz per channel and read into a digital computer, together with the outputs of four additional channels. All data are stored on digital magnetic tape for subsequent processing. We have used two of the extra channels to augment each short-time critical band spectrum with the voiced/unvoiced (= V/UV) and pitch period outputs of an analog pitch extractor (van Rossum & Boves, 1978).

Long-term average spectra are produced by averaging the levels in each frequency band over time. The availability of V/UV information with each spectrum frame enables us to carry out the averaging process for voiced and unvoiced segments separately. One reason to separate the voiced and unvoiced segments of the speech wave is the fact that the LTAS, when restricted to the voiced parts, might be indicative of the spectral properties of the glottal sound source. Combining voiced and unvoiced segments might obscure this relation up to the point of uselessness. Unless the contrary is explicitly stated the term LTAS in this study refers to a long-term average spectrum for which only voiced segments have been entered into the calculations.

If we denote the AC signal at the output of the j -th filter in the critical band filter set by $s_j(t)$, the long-term average of the output, measured from $t = -T$ up to $t = T$ can be written as

$$L_j = \frac{1}{2T} \int_{-T}^T s_j^2(t) dt \quad (3)$$

The L_j could be measured directly by using detectors with sufficiently great time constants. Another approach is to approximate the long-term average level by the mean of a number of successive short-time level measurements

$$L_j = \frac{1}{N} \sum_{n=-N}^{N-1} \frac{1}{\Delta T} \int_{n\Delta T}^{(n+1)\Delta T} s_j^2(t) dt \quad (4)$$

This approach is taken by Elenius (1981) and also in our set-up. Neither Elenius nor we can apply (4) directly, since both use detectors with a logarithmic response. In order to compute L_j we have to convert the logarithmic values to linear voltage values that can be averaged, after which a new log-conversion gives the eventual results in terms of dB.

5.3.2. The representation of LTAS

Each long-term average spectrum is a vector of 20 numbers and can therefore be considered as a point in a 20-dimensional space. It is, however, not likely that all dimensions are perfectly uncorrelated, so that it is only natural to look for a technique that enables us to squeeze the original 20-dimensional structure into a new space of much lower dimensionality. One obvious method to accomplish this is principal component analysis. The reader is referred to Pols (1977) for a highly readable treatment of the application of principal component analysis to critical band spectra or to the literature on multivariate statistical techniques (Cooley & Lohnes, 1971).

Notwithstanding the usefulness of principal component analysis as a mathematically well-founded technique for compressing the information in long-term average spectra, it has some drawbacks that may prevent its use for some purposes, which have to do with the structure of the eigenvectors. First of all the eigenvectors depend heavily on the set of LTAS on which the analysis is performed, therewith impeding the comparison between results of different experiments. The second drawback, closely related to the first one, is that the eigenvectors may or may not be physically or perceptually interpretable.

The drawbacks of principal component analysis cannot be remedied since they are inherent in the technique. Hammarberg et al. (1980) have, however, proposed another, computationally much simpler way to extract the most important information from LTAS. This approach is illustrated and explained in Fig. 5.1. It is not based on an attempt to squeeze as much statistical information in as few numbers as possible but on an extraction of those features of the LTAS that can be supposed to have perceptual relevance (e.g. overall spectral slope, (Plomp, 1976)) or that might be related to the physiology of voice production. Hammarberg et al. define five indices to characterize each LTAS, viz. the maximum spectral level in the band from 400-600 Hz, and the slope of the spectral envelope in four frequency regions, i.e., below 400 Hz, between 600 Hz and 1.5 kHz, between 1.5 and 5 kHz, and in the region above 5 kHz. (refer to Fig. 5.1).

Fig 5 / Definition of LTAS slope measures

However intuitively appealing the description of LTAS in terms of slopes may be, we must consider the fact that it treats the data in a rather crude and perhaps even somewhat arbitrary way. We believe that the approach is legitimate, a contention that can be fortified by a comparison of the statistical status of slope measures and formants. It is not very likely that formant values can account for the bulk of the information (= variance) in short-time speech spectra especially when only formant frequencies are retained and the bandwidths or amplitudes are discarded. Despite that, nobody will claim that the formant representation is useless. The major appeal of a representation of speech spectra in terms of formants derives from its well-understood relation to acoustic models of speech production. The formant representation contains the acoustically (and linguistically, for that matter) relevant information, which need not be identical to statistical information in the sense of variance. Moreover, a description in terms of formants is self-contained in the sense that it is not influenced by the information in other spectra in the set under analysis. Similar arguments can perhaps be put forward in favour of the slope measures, which can very probably be said to contain interpretable relations to a model of speech perception.

Although we realize that much more research is needed to clarify a number of problems attached to the use of slope measures we decided to prefer this representation over the principal component approach. The most compelling motive for this preference is the fact that we have to compare LTAS from two quite different sets of speech stimuli, one completely comprised of male voices, the other of a mixture of female and male voices.

Contrary to principal components, slope measures are not orthogonal. In fact, correlations between the measures appeared to be quite high. This suggests that it may be possible to discard some slope measures without losing too much information. Unfortunately, with one exception there are no viable reasons to decide which value(s) are to be discarded. The exception concerns the level above 5 kHz. Since we are dealing with non-pathological tape recorded speech, the maximal spectral level in the frequency region above 5 kHz is almost always equal to the level in the 5 kHz filter itself. In the very few exceptions to this rule the observed level difference is so small as to be negligible. Therefore we have discarded the slope measure defined on the maximum level above 5 kHz.

5.3.3 *The influence of fundamental frequency on LTAS*

Although most of the critical band filters are so broad as to contain more than a single harmonic of periodic speech signals, critical band spectra - and thus our LTAS - are not completely independent of the repetition frequency. This dependency is inherent in the data, but it might affect a representation in terms of spectral slopes as defined in Fig. 5.1 more than a principal components approach. It is, however, an extremely tedious undertaking to study this differential sensitivity. Since for the time being we will stick to the use of slope measures, we will restrict ourselves to an attempt to explain the problem in general and estimate the range of variation that it can bring about. The latter will be done experimentally rather than by means of theoretical computations.

From the theory of Fourier series expansion it is known that the amplitudes of the harmonics of simple periodic signals like pulses, triangles and ramps are inversely related to some power of the harmonic number. Assuming some form of overall amplitude normalisation, the level of a harmonic falling within a specific pass band, for instance the 400 Hz filter in our analyser, depends on the frequency of the fundamental.

In order to get an impression of the magnitude of the F_0 effect in a slope measure representation, we have measured the relevant filter outputs for a number of signals at fundamental frequencies of 100, 150, 200, 250 and 300 Hz. There were no great between-signal differences, so that we may as well restrict the discussion to a single signal, for instance the modified triangular wave used as the voice source in the OVE Ic vowel synthesizer. It appears that the maximum level in the 400-600 Hz range (i.e., the level that serves as the reference for the definition of the slopes in four frequency regions) varies approximately 4 dB; the same goes for the difference between the reference level and the level in the 1.6 kHz filter as well as for the level difference defined on the basis of the output of the 5 kHz filter. Thus taking the midpoint of the measured ranges as the zero point the slope measures may vary ± 2 dB as a result of variations in F_0 only. In relative terms the F_0 induced variability in

the 1.6 kHz measure is greater than in the 5 kHz measure, since the expected (and measured) level difference at 1.6 kHz is approximately 12 dB, whereas at 5 kHz it is approximately 24 dB.

5.3.4 *LTAS as measures to characterize speakers*

The fact that we need acoustic measures to 'explain' perceptual judgements made us decide to prefer the slope measures over a principal components solution. We have, however, not yet shown that their use is at all justifiable, i.e., we still have to show that the within-speaker variance is less than the between-speaker variance. Since slope measures are inherently multivariate and since the variables are not necessarily orthogonal an approach via the ONEWAY is not possible. Multiple discriminant analysis provides a viable alternative. Using the program given in Overall & Klett (1972) it was possible to assign all of the 60 LTAS to the correct speaker. This result warrants the use of slope measures as speaker characterizing features.

An informal, but quite convincing way to show that the LTAS are very stable measures if the averaging time is made sufficiently long is to visually compare the LTAS results of the two reading texts. Plots of the LTAS are shown in Fig. 5.2, from which it is quite clear that except for speaker DJ the results of the two texts are almost identical. The very great differences between the two LTAS of DJ seem to be extremely disturbing, since they might be taken as completely invalidating the claim that global spectral shapes are characteristic of a speaker. The counterevidence of the case of DJ can be explained away fairly easily by pointing out that he is the speaker who recorded the neutral text more than ten years before the other one. Moreover, the recording conditions and the recording equipment were different. All this led to two recordings the auditory impression and quality of which differ markedly. Thus it can be maintained that the judges have heard two quite different 'voices'. The remaining speakers, who recorded both texts within a very short time interval using the same equipment in the same environment, show the expected similarity in the spectra and in the auditory impression. Therefore it appears that LTAS with averaging times of 30 s can be considered as characteristic of a speaker-recording condition situation.

Our LTAS program does not only produce average spectral levels, but it also computes the variances of the level samples within each band. Informal experiments have led to the conclusion that on the whole the time variation in the (voiced and unvoiced) spectra does not yield enough information to effect a reliable speaker separation. Thus it does not seem to yield useful information about dynamic aspects of the speech. This conclusion need not to be surprising, since the variance is a measure of the overall fluctuation width of the levels in the bands, whereas the speech dynamics is most likely related to short-term changes in the spectra. One measure, however, appeared to discriminate among speakers, viz. the variance in the linear channel in the condition where voiced and unvoiced intervals were not treated separate-

Fig. 5.2 Between-text comparison of LTAS, computed from 90 s speech segments. Full lines: non-neutral, dotted lines: neutral text
The readings of DJ differ considerably, both in recording time and conditions.

ly. This variance might be interpreted as an indicator of the effective dynamic range of the speech. This intensity variation measure has been included in the set to be used in the attempts to explain perceptual ratings.

5.3.5 LTAS as representation of the glottal flow spectrum

The aim of computing LTAS of the voiced parts of the speech signals is to obtain an estimate of the overall spectrum of the glottal flow pulses. Although the position that averaging out the effects of articulation should produce a good estimate of the source spectrum can be defended on theoretical grounds, it would be nice to have more or less formal empirical evidence. We are in a position to approach the problem, since we have available both LTAS and average glottal flow spectra pertaining to seven texts read by five speakers. It was with this comparison in mind that we decided to convert the

Fig 5.3 Comparison of average critical band spectra of inverse filtered glottal flow waveforms (full lines), LTAS (dotted lines), and glottal flow corrected for radiation at the lips (dashed lines).

FFT spectra of the flow pulses to critical band spectra, as described in chapter 4.

The results of the comparison are shown in graphical form in Fig. 5.3. Each panel shows the average spectrum of the glottal pulses derived from 60 vowels (about half of which were taken from stressed syllables) in the form of a full line and the LTAS pertaining to the corresponding text in the form of a dotted line. The LTAS spectra are shifted vertically in such a way that the levels in the lowest frequency band become equal. Undeniably the two representations differ considerably. Part of the differences should have been expected, though, since they are due to the effect of the radiation impedance that is present in the LTAS but not in the glottal flow. Since the model underlying the inverse filtering technique assumes the radiation to be equivalent to a simple backward differencer, we have 'corrected' the glottal flow spectra by adding the frequency response of

Fig. 5.3. Continued.

$$L(z) = 1 - z^{-1} \quad (5)$$

which can be shown to be

$$L(2\pi f) = |\sin \pi f| \quad (6)$$

The correction was carried out by evaluating (6) at the centre-frequencies of the critical band filters and transforming the values into dB-gain relative to the level in the lowest frequency band. The results of this correction procedure are shown in the form of dashed lines.

From Fig. 5.3 it is evident that there are great between-speaker differences in the closeness with which the corrected glottal flow spectra approach the LTAS. In fact, the fit is only satisfactory for EK who is, incidentally, the speaker who is rated most positively on almost all scales. It is equally interesting to note that, except perhaps for PD (the speaker who is given the lowest rating on most scales), the asymptotic trend in the LTAS and corrected glottal flow spectra are very similar. Apparently the LTAS may be a viable predictor of the spectral slope of the glottal pulses towards high frequencies, but certainly not of the corresponding spectral level. For all but the very best speakers the LTAS lead to an underestimation of the high frequency spectral level of the glottal sound source. Thus it seems that for many speakers the high frequency boost due to the radiation at the lips is considerably less than the +6 dB per octave predicted by theory. There is even some indication that the degree to which the radiation boost falls short of the theoretical level increases with increasingly negative ratings of the voice and articulation quality of the speaker.

It is extremely difficult, if not impossible, to identify the exact causes of the differential effectiveness with which the glottal sound source is transmitted into the open air. There are at least two possibilities to choose from, but no hard data to guide the choice. The first, but probably not the most likely possibility is, that the losses in the vocal tract are greater for some speakers than for others. One way to increase overall losses is to decrease wall impedance. The measurements of Ishizaka, French & Flanagan (1976) show that the impedance of the cheeks is lower when the tissue is relaxed than when it is tensed. The impedance difference decreases, however, with increasing frequencies, at least in the frequency range (up to 150 Hz) of the measurement. The contribution of the wall impedance to between-speaker differences in losses in the frequency range of interest here must therefore be assumed to be negligible, as is also suggested by the outcomes of the calculations of Wakita & Fant (1978) who did not find an effect of vocal tract wall impedance on formant bandwidth except, of course, for low F_1 . Additional factors contributing to vocal tract losses are not very promising candidates either, since, for instance, heat conduction and viscous losses depend on the physical properties of the air which cannot, for obvious reasons, be altered by the speaker.

The alternative explanation, which cannot be substantiated either, assumes differences in the radiation impedance. Now the absolute value of the radiation impedance is inversely proportional to the area of the lip opening. Thus a speaker who opens his mouth when talking is expected to generate a higher sound level than an otherwise similar speaker, who hardly separates his lips. Furthermore, the radiation model expressed by (5) is known to be a very crude approximation to a very complex function.

Laine (1982) proposes a simple pole-zero model of lip radiation, that provides a better match to the theoretical function. His model is given by

$$L(z) = \frac{a \cdot (1 - z^{-1})}{1 + bz^{-1}} \quad (7)$$

with

$$\begin{aligned} a &= 0.0779 + 0.2373 \cdot \sqrt{A} \\ b &= 0.8430 - 0.3062 \cdot \sqrt{A} \end{aligned} \quad (8)$$

A: area of lip opening

The frequency response of this model is plotted in Fig. 5.4 for a number of realistic values of the lip opening A. It can be seen that increasing the lip opening aids in radiating the higher frequency components of the speech signal. Thus the discrepancy between LTAS and corrected glottal flow might

Fig. 5.4 Frequency response of the lip radiation model after Laine.

Full lines: pole-zero model with area of lip opening as a parameter. Dotted lines: the simple backward differencing model, plotted for reference.

be due to the use of an incorrect model $L(z)$ in the reconstruction of the flow pulses for all speakers, except EK. If this conjecture is true, the spectral slope of the average flow pulses of all speakers except EK are somewhat underestimated.

The finding that the LTAS do not exactly represent the glottal source spectra raises the question whether their intended use as predictors of ratings of phonation is legitimate. We have decided that it is, mainly for the reason that what is important auditorily is the effective sound source, i.e., the combination of $G(z)$, the glottal source, and $L(z)$ (refer to chapter 4 for a more detailed explanation of the concept of effective source). In other words we do not believe that our listeners have been able to separate true phonatory effects, related to the shapes of the glottal pulses, from radiation effects that cause exactly the same result in the slope of the spectrum of the radiated sound.

5.4 MEASURES BASED ON FORMANT ANALYSES

Despite of a vast body of knowledge of speech perception it is not yet completely understood which factors enhance or impede its intelligibility, especially with respect to normal speech, i.e., speech which has not been degraded by added noise or by some kind of electronic or acoustic processing. Yet it seems sensible to assume that there must be some relation between articulation quality and intelligibility in the first place and also between the intuitive concept of 'precision of articulation' and formant frequencies of vowels. The idea behind such a relation is that the sloppier the speech, the more the formant frequencies of all vowels will approach the target values of the neutral vowel /a/. Koopmans-van Beinum (1980) defined a so called Acoustic System Contrast

$$ASC = \frac{1}{N} \sum_{n=1}^N |\bar{V}_i - \bar{S}| \quad (9)$$

where V_i is the vector of formant frequencies defining any vowel token and S is the vector of formant values pertaining to this speaker's centroid, i.e., its average position in the formant space. ASC's were computed based on a number of stressed vowels taken from the reading texts. A speech wave editor (Butting, 1981) was used to build the data base of vowels to be analyzed.

Eight ASC measures were calculated for each text. The differences between the measures lie in the number of formants (N in formula (9)), absence or presence of a log-transform of the formant frequencies and absence or presence of a weighting function that weighs the contribution of each vowel token to the ASC according to its average frequency in normal Dutch speech.

As could be expected, the eight ASC measures were highly correlated.

Unfortunately, however, all measures failed to show a significant correlation with factor scores of the speakers on the 'Articulation appreciation' dimension resulting from the factor analysis of the judgement scores, or with the scale values of the speakers on the scales related to pronunciation (Boves & Fagel, 1982). Thus it appears that the overall distance between vowel loci in a formant space (that is treated as Euclidian) is not a useful measure of perceived articulation quality.

The idea that articulation quality might be related to the variability of the speech signals in a formant representation was not discarded completely, though. The problem was tackled again in a slightly different way. The first 60 seconds of all reading texts were sampled, digitized, stored on disc and subjected to pitch detection and formant extraction in the voiced intervals. If the speech had been completely voiced, this analysis would have resulted in 600 pitch estimates and an equal number of estimates of the lower four formants. The actual number of formant estimates per text was considerably lower since all unvoiced intervals were discarded. For each formant average frequency and average bandwidth were computed, as well as the corresponding standard deviations.

It is of some interest to know how stable the statistics obtained in this manner are. Li, Hughes & House (1969) and Zahorian & Rothenberg (1981) contend that the time-averaged covariance matrix from a spectral analysis of running speech stabilises after approximately 30 seconds. Data on long-term formant measures have not been reported in the literature. In order to obtain an estimate of the statistical stability of the long-term formant measures, all measures were computed with averaging times starting at one second and increasing in one second intervals until the final value was reached after an averaging time of 60 s. The results were plotted as a function of averaging time in order to see whether they appear to approach an asymptotic value. Part of the results for two speakers is shown in Fig. 5.5. It can be seen that fairly stable values are reached after 25-30 seconds. Thus the 60 second averages employed in the subsequent part of this study may safely be taken to be characteristic of the readings that served as stimuli in the judgement experiment.

Some comments on the formant measures have to be made. The term 'formant' must be read as 'spectral peak'. Furthermore, formants have been determined and averaged for all signal segments that were recognized as voiced. Specifically, no attempt has been made to sort the voiced segments into phoneme classes like vowels, nasals, fricatives and stops.

The results of the formant measurements are summarized in Table 5.4. A study of the data in this table leads to a number of interesting observations, the most striking one undoubtedly being that the average values of the lowest three formants in the Nijmegen material seem to differ quite clearly from the corresponding values in the Amsterdam material. The greatest difference is in F_2 , the average frequency in the Nijmegen material being 8.5% higher than the average of the males in the Amsterdam material. We have as yet no

Fig 5.5 Plot of F_1 statistics as a function of averaging time. First four panels pertain to EK's reading of the neutral text, the second four panels to PD's reading of the same text. Note the different vertical scales of the plots.

explanation for this observation. Note that the average F_2 of the males in the Nijmegen material (1490 Hz) is higher than the average F_2 of the females in Amsterdam (1395 Hz), which, in turn, is only slightly higher than the average F_2 of the Amsterdam males (1375 Hz).

Another, less important, observation is that the bandwidth estimates derived from the autocorrelation LP analysis are uniformly very high, especially in the Nijmegen material. The bandwidths of F_1 in the Amsterdam material seem to be generally somewhat smaller than in the Nijmegen material. This may be due to either idiosyncrasies of the talkers or the recording conditions (or to both). That estimates of the bandwidths of the poles are not always reliable approximations of formant bandwidths has been known for quite some time (Rabiner & Shafer, 1978.450). Nevertheless we think that, at least within the stimulus sets, the results of the analyses do represent stable characteristics of the signals that may possess some auditory relevance. Therefore we have retained these measures for use in the acoustic prediction of the perceptual ratings.

Fig 5 5 Continued

From the Amsterdam data more than from the Nijmegen results it appears that the average formant values do not represent a neutral vowel. Rather they seem to pertain to a somewhat centralized but fairly closed vowel. This reflects the fact that formants have been averaged for all voiced segments, irrespective of whether these segments are vowels or consonants. This observation has influenced the way in which the averaged formant frequencies have been converted into a modified Acoustic System Contrast, by referring measured values to the formants computed from a 'standard reference vocal tract configuration'. We have not taken the customary uniform tube of 17 cm length, but instead a tube with a length of 18 cm (to account for rounded vowels and since, according to the data in Fant (1960:115) a greater length seems to be more realistic). The cross-sectional area of this tube as a function of the distance from the lips is given in Table 5.5. The lower pharyngeal cavities are a greatly simplified copy of the shapes traced by Fant, the closure at the lips was chosen in order to obtain formant values which were close to the grand means in our measurements. Formant frequencies and bandwidths

Table 5.4. Formant frequencies and bandwidths averaged over 60s of running speech. Formant extraction was only done in voiced intervals. A: Readings of the non-neutral text (i-suffix) and the neutral text (a-suffix) by six males. B: Readings of the neutral text by five females (Fm1-Fm5) and six males (M11-M16).

		F1	F2	F3	F4	B1	B2	B3	B4
A	EKi	416	1525	2483	3422	279	282	344	322
	EKa	412	1546	2515	3520	278	258	320	293
	DJi	358	1443	2452	3427	168	340	396	246
	DJa	442	1386	2306	3535	212	254	324	258
	LBi	369	1454	2353	3300	235	250	269	387
	LBa	354	1480	2354	3267	224	241	301	361
	NRi	488	1599	2582	3528	280	239	336	279
	NRa	500	1606	2534	3486	293	253	360	345
	TBi	459	1482	2488	3459	282	344	346	340
	TBa	482	1519	2483	3420	295	344	331	363
	PDi	369	1416	2331	3497	221	298	265	368
	PDa	380	1413	2337	3500	243	308	304	363
B	Fm1	445	1394	2360	3317	201	300	360	289
	Fm2	465	1209	2310	3380	207	372	335	268
	Fm3	463	1420	2334	3555	195	281	342	332
	Fm4	517	1426	2456	3450	182	334	365	236
	Fm5	523	1522	2389	3466	151	313	394	328
	M11	442	1386	2306	3535	212	254	324	258
	M12	469	1366	2351	3530	155	287	377	274
	M13	462	1222	2323	3395	206	375	333	251
	M14	436	1391	2414	3440	181	236	290	273
	M15	435	1411	2485	3577	149	167	209	198
	M16	354	1480	2354	3267	224	241	301	361

from this tube were calculated by means of the model proposed by Wakita & Fant (1978). This model includes losses in the vocal tract due to wall vibration and heat conduction. Furthermore the model possesses a fairly realistic subglottal system and a glottal part. In our simulation we have used a 21 section subglottal system, a transglottal pressure of 10 cm water and a glottal area of .10 cm². This rather large area was chosen in order to increase the bandwidths of the resonances. The resulting reference formants and their associated bandwidths are [430, 80], [1380, 115], [2485, 105], and [3425, 120]. Due to the almost negligible differences between the averaged formants for the females and the males in the Amsterdam material it was decided to refer all data to the same reference set of formant frequencies. The modified ASC, then, was computed as follows:

$$MASC = \frac{1}{4} \sum_{i=1}^4 (f_{i,j} - f_{i,r})^2 \quad (10)$$

where $f_{i,j}$ is the i -th average formant of speaker j , and $f_{i,r}$ is the i -th formant frequency from the reference set. Experimentation with summation over different numbers of formants and with log-transformed frequencies have shown that the MASC as defined by (10) was the best (but not a particularly powerful) predictor of pronunciation appreciation. Therefore we have added MASC to the set of acoustic measures to be used in the next chapter. Apart

Table 5.5. Cross-sectional areas in cm² of the reference vocal tract used for calculating MASC. All segments have an equal length of 1 cm.

	glottis															lips	
section number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
area	2.6	1.6	1.3	2.0	3.0	4.0	5.5	5.5	5.5	5.5	5.5	5.5	5.5	4.0	3.0	2.0	1.0

from long-term average values of formant frequencies and bandwidths the standard deviations have also been computed. However, we do not know yet how to convert these measures into a predictor of articulation quality. Note that this applies not only to the standard deviations but also to the formant bandwidths themselves. Therefore we will retain these measures as a kind of raw data and investigate their use in the acoustic explanation of perceptual judgements empirically.

5.5 SUMMARY AND CONCLUSIONS

In this chapter a large number of acoustic measures have been described which can be obtained from connected speech by means of essentially automatic processing techniques. All measures pertain to overall, long-term characteristics of the signals. This is as true for the central tendency type measures as it is for the variation type measures. Our decision to retain only those measures which are stable within and discriminate between speakers emphasizes their long-term stable nature.

Most of the measures described above are well-known, as for instance average pitch, pitch perturbation and speech rate. Other measures, like the spectral slope measures, are less well-known and therefore have been explained and discussed in somewhat more detail. It has been shown that the meaning of the slope measures is fairly independent of the mean fundamental frequency of the speech, at least over quite a large F_0 range. Contrary to the naive intuition the slopes found in the LTAS of the speech signal did not correspond too closely with the slopes obtained from an analysis of the reconstructed glottal flow pulses. It has been shown that the discrepancies between the spectral slopes of the reconstructed glottal pulses on the one hand and the radiated acoustic speech wave on the other can to a large extent be explained by differences in the radiation impedance at the lips. This is an area where more research is necessary to answer all questions.

A more customary way to describe speech spectra is in the form of frequencies and bandwidths of a number of formants. Formant descriptions are, however, essentially descriptions of short-time spectra rather than of long-term speaker characteristics. Therefore an attempt was made to develop a number of global statistical descriptors based on short-time formant values. In the next chapter these, and all other, acoustic measures will be used to explain the perceptual ratings dealt with in chapter 3.

Relations between acoustic measures and perceptual ratings

6.0 INTRODUCTION

In this chapter we will present the results of correlation and regression analyses, performed in order to study the relations between acoustic measures and perceptual judgements. The perceptual judgement scores are treated as the criteria and the acoustic measures as the predictors. This reflects our belief that, at least for the time being, perceptual judgements have to be considered as more 'valid' than acoustic measures, especially when complex attributes are concerned. This view might be somewhat surprising, if only because of the extensive attention paid to more 'objective' measures that can be derived from acoustic or physiological signals. It is, however, totally in agreement with the commonly accepted idea that almost all present day psycho-acoustic knowledge pertains to the perception of essentially stationary signals (Pastore, 1981). The absence of a solid base of knowledge about the psycho-acoustics of connected speech prevents the replacement of perceptual judgements by 'objective' acoustic measures, the more so since we very often can choose from a set of alternative - and not quite similar - such measures.

This work can be considered as a contribution to our knowledge of the psycho-acoustics of non-stationary signals. Clearly, our results cannot be taken as solid knowledge, but rather as hypotheses that are worth further testing. For the generation of solid knowledge the number of stimuli is too small and the number of factors contributing to the judgements too large (and, for that matter, totally uncontrolled as far as their covariation is concerned).

In the sections that follow we will try to predict scale values and factor scores, obtained in both the Amsterdam and Nijmegen experiments, from the acoustic measures described above. The factor scores used in this part of the research are average values of the speaker-listener 'objects' in a stringing-out analysis of the raw scores on all 35 rating scales. For the Amsterdam data a five factor solution is used, with factors labelled 'Voice dynamics', 'Articulation appreciation', 'Potency', 'Voice quality' and 'Tempo' (Fagel, van Herpt & Boves, 1983). Analysis of the judgements in the Nijmegen experiment yielded a six factor solution. Factor labels were similar to those in the Amsterdam experiment with the one obvious exception of the third (extra)

sociolinguistic factor (cf. chapter 3). For the remainder of this chapter we will assume that the factor structures emerging from both experiments are essentially identical and that the sociolinguistic factor did not show up in the Amsterdam data for some unknown reason.

It was said above that we cannot reach very conclusive results, due to the small numbers of speakers. Even if we consider each rendition and the judgements it has attracted as an independent observation, we have no more than 11 or 12 measurements per scale and per acoustic variable. In a correlation analysis this small number of observations aggravates the risk that correlations are dominated or even totally determined by outliers, the more so if the outlier is not a single rendition but a speaker, present in the data with two renditions, as is the case in the Nijmegen data. To avoid unjustifiable interpretations of spurious correlations we decided not to accept a relation if a correlation between an acoustic measure and a judgement scale or factor could not be motivated and understood on independent grounds.

As an additional means of preventing unjustifiable conclusions we have compared the results obtained with the Nijmegen material (12 readings by six male speakers) with those obtained from the material employed by Fagel in his experiments in Amsterdam (six readings by males and five by females). Relations between acoustic and perceptual judgements are only considered as genuine if the results are similar in both data sets.

Unlike the previous section, where the acoustic measures formed the organizing principle, this section will be organized by 'factor'. This structure once again emphasizes the precedence of perceptual over acoustic dimensions. The discussion of individual scales will be limited to the ones that load highest on the factors. In general, all factors (except the sociolinguistic one) will be represented by the scales chosen in Fagel et al. (1983). The first part of this chapter is devoted to the study of simple correlations between acoustic and perceptual measures. In the last part of this chapter we will proceed to a multiple regression analysis, in order to investigate if a linear model, in which the individual acoustic measures are treated as additive, can lend extra insight into the ways in which signal characteristics influence perceptual ratings of speech.

6.1 SIMPLE CORRELATIONS

6.1.1 *Voice dynamics*

The voice dynamics factor, the factor which accounts for the major part of the variance in the judgement scores, has an evaluative flavour. This evaluative connotation is more apparent from the five-factor solution of the Amsterdam data than from the six-factor solution for the Nijmegen data summarized in Table 3.6.

It has already been explained that it is extremely difficult to define acoustic measures of voice dynamics that can be obtained by (semi-)automatic processing of running speech. The contamination of dynamism with general evaluation only aggravates this problem. Therefore we decided to cross-correlate all pitch measures and the intensity variation measure with the scores on the scales 'expressive $\leftrightarrow$ expressionless', 'melodious $\leftrightarrow$ monotonous', 'pleasant $\leftrightarrow$ unpleasant', 'ugly $\leftrightarrow$ beautiful' and the factor scores on the first factor in order to discern candidate measures.

The results are summarized in Table 6.1. The upper right triangle of that correlation matrix pertains to the Amsterdam experiment, the lower left contains the data on the Nijmegen experiment. From the right-most column and the bottom row of this Table it can be seen that no single acoustic measure reaches a correlation of .6 or more with the factor scores. Thus no measure can explain more than some 35% of the variance in the factor scores. Most measures fail to explain more than a very low 10% to 25% of that variance. In fact, the majority of the correlation coefficients is not significantly different from zero at the .05 level in a two-tailed test with 9 or 10 degrees of freedom. Dynamic range (or intensity variation) is negatively and significantly correlated with factor scores in one data set and positively and significantly in the other. A similar remark applies to the correlations between the proportion of voiced intervals and the scale values on the two evaluation scales, except that in the Nijmegen data the correlation coefficients are not significantly different from 0. The single most outstanding feature of the correlation matrix in Table 6.1 is the behaviour of F_0 mean in the Nijmegen data. This acoustic measure correlates highly and significantly with all four scales but not with the factor scores. In the Amsterdam data a similar behaviour of F_0 mean is absent. Although we have no means to prove this conjecture statistically, we are strongly convinced that the relation between F_0 mean and evaluative scales in the Nijmegen data is mainly an artifact, due to the fact that the two low-pitched speakers happened to show the best reading performance. None of our acoustic measures can be directly related to reading performance. Nonetheless the quality of the renditions must have influenced the scores on a number of scales to a considerable extent. Now it is quite conceivable that the impression of a high (or low) reading performance can be invoked by different combinations of acoustic variables. If this hypothesis is true, it may explain at least part of the disparate behaviour of some acoustic variables in the two data sets.

6.1.2 *Articulation quality*

Turning now to the articulation quality factor it should first of all be pointed out that this factor too has an evaluative flavour. It has already been mentioned in chapter 5 that ASC-measures as originally defined by Koopmans-van Beinum predict only a negligible part of the variance in the scores on scales related to articulation quality. Yet, ASC-like measures are attrac-

Table 6.1 Correlations between several acoustic measures and scales, resp. factor scores related to the voice dynamics factor. Numbers in the upper triangle refer to the results of the Amsterdam experiment, whereas the data for the Nijmegen experiment are given in the lower triangle. Note that the lower right 5★5 submatrix contains correlations between scale values and factor scores.

	% voiced	F ₀ mean	F ₀ variation	pitch perturbation	dynamic range	melodious	expressive	pleasant	ugly	factor #1
% voiced	-	.132	-.162	.106	-.302	.488	.456	.590 ^b	-.527 ^b	-.342
F ₀ mean	-.084	-	-.735 ^a	.369	.300	.216	.222	.344	-.287	-.245
F ₀ variation	-.218	-.738 ^a	-	.310	.290	-.429	-.431	-.237	.193	.478
pitch perturbation	-.299	-.247	.599 ^a	-	.742 ^a	-.054	-.062	.361	-.379	.129
dynamic range	.028	.513 ^b	-.575 ^b	-.548 ^b	-	-.523 ^b	-.548 ^b	-.223	.188	.563 ^b
melodious	.326	.626 ^a	-.569 ^b	-.051	.612 ^a	-	.996 ^a	.838 ^a	-.853 ^a	-.967 ^a
expressive	.285	.676 ^a	-.550 ^b	-.066	.605 ^a	.971 ^a	-	.829 ^a	-.842 ^a	-.977 ^a
pleasant	-.322	.790 ^a	-.449	-.101	.680 ^a	.559 ^b	.661 ^a	-	-.982 ^a	-.760 ^a
ugly	.230	-.841 ^a	.468	.013	-.617 ^a	-.711 ^a	-.798 ^a	-.953 ^a	-	.776 ^a
factor #1	-.507 ^b	-.488	.570 ^b	.148	-.573 ^b	-.962 ^a	-.929 ^a	-.415	.554 ^b	-

a. $p < .01$ in a two-tailed test

b. $p < .05$ in a two-tailed test

tive in that they are scalar measures, i.e., each reading text can be given a unique place on the unidimensional ASC scale. Formant statistics, on the other hand, are inherently vectorial (or multidimensional) measures, unless, of course, one is willing to limit the discussion to a single formant. As explained earlier, the sets of four average formant frequencies have been converted to a modified ASC according to (5.13). We still have four dimensional vectors of formant frequency standard deviations, of formant bandwidths and of their standard deviations. It would be nice if those three vectorial measures could be combined into a scalar measure in some meaningful way. A necessary condition for such a transformation is that the individual vector elements yield correlation coefficients with factor scores that have the same sign in the Nijmegen and Amsterdam material. An additional condition is that these signs are interpretable in themselves. It would, for instance, arouse suspicion if we were to find a strong negative relation between articulation quality and the standard deviations of F_1 and F_2 .

Comparison of the relevant correlation matrices shows that no more than four variables meet the first requirement. They are the bandwidths of the first and second formant and the standard deviations of the bandwidths of the second and third formant. This set makes the impression of a rather random choice, an impression that finds further support in the correlations themselves, none of which reaches an absolute value of .5. Perceived articulation quality increases with the bandwidth of the first formant (B_1) (a finding that gives rise to suspicion of incompatibility with the second requirement), but with decreasing B_2 (which, intuitively, makes more sense); it also increases with decreasing standard deviations of B_2 and B_3 .

A scalar physical measure of articulation quality was formed from a linear combination of the four individual measures that behave in a similar way. The three variables that correlate negatively with articulation quality were given a negative weight, whereas B_1 , the positively correlating variable, was given a positive weight. The absolute values of the weights were chosen so as to reflect the average correlations with the factor scores. This - admittedly rather unusual - procedure reflects the fact that an attempt to obtain weights via a least-squares best-fit procedure yielded weights which differed markedly both between speaker groups and within speaker groups between scales. The differences even included the signs of the regression coefficients. Therefore, the resulting 'Formant-based Articulation Quality' (FAQ)-measure, defined by

$$FAQ = .3 \times B_1 - .4 \times B_2 - .3 \times \text{s.d. } [B_2] - .4 \times \text{s.d. } [B_3] \quad (1)$$

should be handled with extreme caution. The fact that the weights do not sum to unity is of no importance, since this provisional FAQ measure is only intended for use in product-moment correlation type analyses.

The relevant results of the analysis are summarized in Table 6.2. The acoustic measures are restricted to the MASC and the composite FAQ measure. From the figures in that table we cannot but conclude that perceptual judgements of articulation quality defy any attempt to explain them from formant statistics. Obviously, the perceived clarity and correctness (and the beauty, for it has already been mentioned that this factor has non-negligible contributions of evaluative scales) of articulation bears no simple relation to average formant frequencies and bandwidths, nor to the ranges within which these frequencies and bandwidths vary. Thus it is not sufficient to make large articulatory excursions if one wants to produce clearly intelligible speech that conforms with 'Algemeen Beschaafd Nederlands' (Received Pronunciation of Dutch) and that is pleasant to listen to. Perhaps preciseness in movements and in their timing is more important than reaching extreme target values. Furthermore, it is quite improbable that voiceless consonants do in no way influence our impression of the quality of articulation, and it is not easy to think of a systematic way in which consonant articulation might have influenced the formant statistics used here to predict perceptual ratings.

The finding that our acoustic measures are not powerful predictors of perceived articulation quality does not contradict the results of Koopmans-van Beinum (1980). Comparing ASC's for two male and two female speakers, one of each sex trained and the other untrained, obtained from eight different conditions, she found that the ASC's for the trained male were generally slightly higher than the corresponding measures for the untrained

Table 6.2. Product moment correlation between a modified Acoustic System Contrast measure (ASC), a formant based articulation quality measure (FAQ) and the scale values and factor scores pertaining to perceived articulation quality. The upper right triangle of the matrix contains the results of the Amsterdam experiment, whereas the results of the Nijmegen experiment are in the lower left triangle.

	MASC	FAQ	broad cultured	polished slovenly	factor #2
MASC	-	-.203	.345	-.297	.427
FAQ	.855 ^a	-	.372	-.261	.256
broad cultured	.383	.689 ^a	-	-.964 ^a	.962 ^a
polished slovenly	-.218	-.604 ^a	-.947 ^a	-	-.913 ^a
factor #2	.321	.662 ^a	.966 ^a	-.904 ^a	-

a. $p < .01$ in a two-tailed test

b. $p < .05$ in a two-tailed test

male. The two women, however, showed exactly the opposite trend: the untrained speaker's ASC's were consistently higher. Moreover, the ASC's computed from stressed vowels taken from a retold story are approximately equal to ASC's based on unstressed vowels in text read aloud. Now it seems very unlikely that the clarity and precision of articulation in the retold story is at best (viz. in the stressed syllables) equal to that obtained in unstressed syllables in reading. Thus we are forced to conclude that there is much more to articulation quality than can be captured in the ASC measures and formant statistics employed in this study.

6.1.3 *The paralinguistic factor in the Nijmegen experiment*

The factor analysis of the ratings of the twelve texts read by male speakers only - i.e., the Nijmegen experiment - yielded a six-factor solution in which the third factor had high loadings of para- and sociolinguistic scales like 'tense $\leftrightarrow$ relaxed', 'steady $\leftrightarrow$ unsteady', 'self-confident $\leftrightarrow$ wavering' and 'agitated $\leftrightarrow$ calm' (cf. Chapter 3). A similar factor did not appear in the Amsterdam experiment.

From the contributions of scales defined by adjectives like tense, steady and agitated it seems reasonable to assume that the physical counterpart of this factor is related to tempo, loudness and pitch variations and perhaps also to pitch perturbation. Moreover, tenseness may manifest itself in the form of an increase in the level of the higher harmonics of the voiced parts of the speech signal.

Although the scales mentioned in the preceding paragraphs did not form a factor in the Amsterdam experiment, we nevertheless decided to analyse their correlations with the physical measures in the same way as was done for the Nijmegen material. There were, of course, no factor scores to go with the scale values.

The correlation coefficients are given in Table 6.3. Data with respect to pitch variation are omitted, since the coefficient of variation did not correlate significantly with any scale in either data set. The results are instructive mainly because of the differences between the two data sets. From the Nijmegen data one cannot but conclude that mean F_0 - and, be it to a lesser extent, also speaking rate, pitch perturbation and intensity variation - are useful predictors of the sociolinguistic judgements. In the Amsterdam data, however, there is not even a trace of a relation between these acoustic measures and the scale values on the scales 'tense $\leftrightarrow$ relaxed' and 'self-confident $\leftrightarrow$ wavering', except for the significant correlation between pitch perturbation and tenseness which, incidentally, is not present in the Nijmegen data. Only mean F_0 behaves roughly equal in both data sets. As mean F_0 increases the speaker makes the impression of being more tense and less self-confident.

Tempo is a factor of some importance in the Nijmegen data, but not at all in the Amsterdam data. Most probably the correlation between reading tempo and perceptual ratings of personal characteristics of the readers is

Table 6.3. Correlations between acoustic measures and scale values on scales related to the sociolinguistic factor (factor #3) from the Nijmegen experiment (data in the lower left triangle). The corresponding data from the Amsterdam experiment are given in the upper right triangle. Note that the sociolinguistic factor did not emerge in the Amsterdam material.

	Slope 1.6 kHz	Slope 5 kHz	F ₀	perturb- ation	intens. variation	syll./ sec.	tense/ relaxed	self-confident wavering
Slope 1.6 kHz	-	.062	.243	.472	.398	.274	-.507	.255
Slope 5 kHz	.300	-	.234	.089	.167	-.114	-.169	.171
F ₀	-.374	-.576 ^a	-	.369	.300	-.308	-.360	.425
perturb- ation	.219	.835 ^a	-.247	-	.742 ^a	.023	-.616 ^a	.202
intens. variation	-.147	-.661 ^a	-.513 ^b	-.548 ^b	-	-.201	-.071	-.258
syll./ sec.	-.064	.430	-.488	.358	-.662 ^a	-	-.230	.023
tense/ relaxed	.049	-.045	-.649 ^a	.330	.328	.328	-	-.777 ^a
self-confident wavering	-.111	-.295	.835 ^a	.085	-.565 ^b	-.565 ^b	-.881 ^a	-
factor #3	.008	.135	.581 ^a	.350	-.365	-.365	-.980 ^a	-.819 ^a

a. $p < .01$ in a two-tailed test

b. $p < .05$ in a two-tailed test

completely due to speaker idiosyncrasies. Yet, the signs of the correlations may be considered to confirm what is known from the literature on person perception: slowly speaking people often make the impression of being doubtful and ill at ease. It should be noted, however, that the exactly opposite 'explanation' makes sense too: if people start to talk too fast, they are considered as tense and nervous. This topic will be discussed in more general terms in 6.1.7. The Amsterdam data show much less variation in speaking rate.

Spectral slopes as defined and used in this work are virtually unrelated to perceived tenseness. The correlation between the slope in the mid range of the spectrum and tenseness scores in the Amsterdam data (which is, incidentally, significant at the .05 level in a one tailed test) is the only coefficient that agrees with our *a priori* expectations that the shallower the slope, the more tense a person sounds. A similar trend is, however, completely absent in the Nijmegen data. It is impossible to explain this finding from the data as they are. The task of reading one or two texts seemed to be more embarrassing for some subjects than for others. Yet it is possible that even for the talkers who felt most ill at ease the reading task did not induce enough stress to cause a considerable increase of muscle tension. We will return to the question of the range of variation, that has been addressed before, in the discussion of the results of the next factor, viz. potency.

6.1.4 *Potency*

The third factor in the Amsterdam material and the fourth in the Nijmegen data could unambiguously be identified as a potency factor with high loadings of scales like 'powerful $\leftrightarrow$ weak', 'firm $\leftrightarrow$ slack' and 'loud $\leftrightarrow$ soft'. It is somewhat confounded with voice quality aspects, in the Amsterdam material to a larger degree than in the Nijmegen solution. We will treat the scales 'dull $\leftrightarrow$ clear' and 'husky $\leftrightarrow$ not husky' in the next section, irrespective of the fact that they had the highest loadings on the factor under analysis in the Amsterdam solution. In the Nijmegen data set these scales had their highest loadings on the same factor, be it appreciably lower. Yet, conceptually they are better dealt with under the heading of voice quality.

It is not immediately clear which of our acoustic measures can serve as a meaningful predictor for factor scores on the potency factor and the scales that form it. The loudness scale is perhaps easiest to deal with. Since on the stimulus tapes all recordings were brought to the same VU-meter reading, the only cue to the original loudness level retained in the signals must be spectral slope. It is conceivable that potency in a narrower sense has also to do with variation measures (both amplitude and pitch) and perhaps also with tempo.

Thus we are left with an extensive set of possible predictor measures, which happens to be largely similar to the set dealt with in the previous subsection. Given the clearly paralinguistic character of 'potency' this need not be too

surprising. The results are summarized in Table 6.4. Note that for some conceptually irrelevant reason the direction of the potency factor is different in the two solutions. Thus all correlations with the factor scores should have opposite signs in the two data sets.

From the table it can be seen that the 5 kHz slope measure does what one would expect: if the spectral level increases (and slope decreases) listeners tend to perceive a greater loudness. The behaviour of the remaining measures is extremely difficult to interpret. Average F_0 , intensity variation, reading tempo and the slope at 5 kHz seem to do a fair job in explaining the scale values on the scale 'powerful $\rightarrow$ weak' in the Nijmegen material. Even the direction of some correlations could be given a convincing 'explanation'. In an all-male group of speakers those with a high average pitch will most likely be judged as less strong than the low pitched speakers. A tendency for fast talkers to make an impression of power can also be motivated, but remember the remark on the U-shaped correlation in the previous section. And indeed, if the - admittedly insignificantly low - coefficient in the Amsterdam data means anything at all, it says that in that data set the correlation is on the other leg of the curve. A similar situation is seen with respect to intensity variation. In the Nijmegen material a greater variation induces an increasing perceived weakness (which is contrary to the stereotype of weak people talking with a soft and monotonous voice); in the Amsterdam material, however, the trend is in the direction of increasing perceived power with increasing intensity variation.

In this connection there is yet another point worth mentioning. The scales 'powerful $\rightarrow$ weak' and 'loud $\rightarrow$ soft' belong to the small set where the position of the ideal female voice differs not only statistically significantly from that of the ideal male voice, but where the difference is also substantial in conceptual terms, i.e., more than one scale position on a seven point scale (Boves et al., 1982). Thus an attempt to establish relations between physical measures and perceptual ratings on these scales might well be hampered by the fact that one group of speakers comprises individuals of either sex, for it is quite possible that the functional relations are different for the two sexes.

6.1.5 *Voice quality*

In this section we will deal with the voice quality scales that load high on the potency factor and with the voice quality factor proper. The latter factor is entirely made up of two scales, viz. 'deep $\rightarrow$ shrill' and 'high pitch $\rightarrow$ low pitch'. Therefore spectral slope and average pitch level present themselves as candidate predictors. Unfortunately, spectral slope cannot be expressed as a scalar value in the three slope measure representation that is used in this study. A scalar slope measure applied to long-term average speech spectra seems, however, impossible to define in a useful way. This is -admittedly - unlike the situation in experimental psycho-acoustic research, although the stimulus ensemble used by von Bismarck (1974) cannot be described in terms

Table 6.4 Correlations between acoustic measures and scale values and factor scores related to the potency factor. The upper right hand triangle pertains to the Amsterdam material, the lower left hand triangle to the Nijmegen data.

	syl./ sec.	F ₀ mean	perturb- ation	dynamic range	Max. 400-600	left slope	slope 1.6 kHz	slope 5 kHz	powerful weak	loud soft	potency
syl./sec.	-	-.308	.023	-.201	.025	-.152	.274	-.114	.349	.546 ^b	-.764 ^a
F ₀ mean	-.488	-	.369	.300	.245	.801 ^a	.244	.234	.239	-.222	.734 ^a
perturb- ation	.358	-.247	-	.742 ^a	.538 ^b	.608 ^a	.472	.089	.159	-.011	-.180
dynamic range	-.662 ^a	.513 ^b	-.548 ^b	-	.514	.607 ^a	.398	.167	-.405	-.464	.347
Max. 400-600	.416	-.552 ^b	.795 ^a	-.758 ^a	-	.711 ^a	.495	.774	-.065	-.457	.135
left slope	.326	-.408	.924 ^a	-.523 ^b	.907 ^a	-	.591 ^b	.527 ^b	.145	-.369	.251
slope 1.6 kHz	-.064	-.374	.219	-.147	.487	.406	-	.062	.239	.053	-.148
slope 5 kHz	.430	-.576 ^a	.835 ^a	-.661 ^a	.870 ^a	.915 ^a	.300	-	-.035	-.479	.349
powerful weak	-.568 ^b	.829 ^a	-.069	.586 ^a	-.361	-.179	-.058	-.499 ^a	-	.775 ^a	-.682
loud/soft	-.513 ^b	.506 ^b	-.104	.489	-.248	-.186	-.048	-.523 ^a	.843	-	-.847
potency	-.441	.473	-.287	.467	-.342	-.358	.001	-.665 ^a	.766	.966	-

a. $p < .01$ in a two-tailed test

b. $p < .05$ in a two-tailed test

of dB per octave slope only. Nevertheless, his results quite clearly indicate an inverse relation between the spectral slope of the acoustic stimulus and perceived sharpness of the sound.

In order to overcome the technical problem we will first look at simple correlations between several slope measures and the perceptual judgements. Next we will try and see if the slope measures can be linearly combined into a single measure of the spectral balance that will do a significantly better job in predicting the judgements than a single slope measure.

The correlations between the relevant acoustic measures on the one hand and the scale values and factor scores on the other are given in Table 6.5. The condensed form in which the data are presented here can be justified by noting that the correlations between the acoustic measures themselves appear in Table 5.4. The correlations between scale values and corresponding factor scores are always very high and not very illuminating. The results fit neatly into the picture that emerged from the discussions in the foregoing sections: most measures explain a sizable amount of variance in one or more scales and/or factors, but rarely does a measure perform equally well in both data sets. Also, the explanatory power of the acoustic measures seems to be somewhat greater in the Nijmegen than in the Amsterdam material. This is especially clear in the case of the two voice quality scales which belong to the potency factor rather than to the fourth factor in the Amsterdam and the fifth one in the Nijmegen solution (viz. the scales 'dull $\leftrightarrow$ clear' and 'husky $\leftrightarrow$ not husky'). The proportion of the signal that is voiced is the only measure operative in the Amsterdam data, but it is totally insignificant in the Nijmegen material. In the latter material spectral slopes are much more powerful.

With respect to the voice colour factor, average F_0 functions according to what one would have expected. Perhaps one might have expected even higher correlations between average F_0 and perceived height of the voices. Apparently average F_0 is at least as powerful a predictor of scores on the scale 'deep $\leftrightarrow$ shrill' as it is for 'high $\leftrightarrow$ low pitch'. The relation between average F_0 and shrillness, which is not very surprising after all, was prevented from emerging in von Bismarck's experiment, since all his tonal stimuli had a constant repetition frequency of 200 Hz. The very high correlation found for the factor scores in the Amsterdam data is most probably due to the fact that the ranges are larger here because of the mixed-sex composition of the group of talkers.

Turning to the slope measures as predictors of voice colour, we see a picture that is exactly the opposite of what was observed with respect to the potency related voice quality scales: the explanatory power of the slopes is greater in the Amsterdam data. Especially the slope in the low frequency part of the spectrum seems to be very important. This is most probably due to its correlation with average F_0 , or, in other words, with the distinction between female and male talkers. It is encouraging to see that, with the exception of the slope at 1.6 kHz, the signs of the correlation coefficients are identical in

Table 6.5 Correlations between acoustic measures and voice quality ratings

	dull <--> Amst.	clear Nijm.	husky <--> Amst.	not husky Nijm.	high <--> Amst.	low Nijm.	deep <--> Amst.	shrill Nijm.	factor Amst.	scores Nijm.
% voiced	-.658 ^a	-.247	-.701 ^a	-.209	.117	.468	.053	-.520 ^b	-.018	-.558 ^b
F ₀ mean	.422	-.194	.117	-.555 ^b	-.585 ^b	-.678 ^a	.780 ^a	.662 ^a	.734 ^a	.546 ^b
perturb- ation	.150	.293	-.349	.531 ^b	-.749 ^a	.117	.698 ^a	-.096	.744 ^a	.004
Max. 400-600	.290	.620 ^a	.020	.788 ^a	-.420	-.124	.545 ^b	.130	.506	.258
left slope	.467	.526 ^a	.121	.698 ^a	-.685 ^a	-.214	.834 ^a	.284	.789 ^a	.332
slope 1.6 kHz	.039	.075	-.181	.145	-.378	.190	.410	-.148	.388	-.177
slope 5 kHz	.274	.709 ^a	.322	.899 ^a	-.103	-.097	.334	.148	.223	.241

a $p < .01$ in a two-tailed test

b $p < .05$ in a two-tailed test

both materials. This should facilitate the construction of a measure like the FAQ proposed in 6.1.2. Yet, such a measure is not likely to outperform a multiple regression analysis.

6.1.6 *Tempo*

The tempo factor essentially comprises the scales 'quick $\leftrightarrow$ slow' and 'dragging $\leftrightarrow$ brisk'. The physical measures chosen as predictor variables are the speaking rate expressed in syllables per second and the proportion between sentence pause time. As explained in 5.1.1, within and between sentence pauses were not discarded in the measurement of the speech rate. The measure is also somewhat coarse in that it has been obtained by dividing total reading time into a phonological syllable count. Thus all readings have been assumed to contain the same number of syllables, regardless of differences in pronunciation and different numbers of reading errors.

The first thing that strikes the eye when looking at Table 6.6 is that the two pause time measures account for no more than a negligible amount of the variance in the ratings and factor scores. Furthermore, it should be noted that the polarity of the tempo dimension in the six-factor solution of the Nijmegen data is reversed compared with the five-factor structure emerging from the Amsterdam solution. In both data sets speaking rate accounts for approximately 35% of the variance in the factor scores. In the Amsterdam material the correlation of tempo with the factor scores differs only significantly from 0 at the .1 level. Given the discussion in 5.1.1 this result should

Table 6.6. Correlations of tempo measures with perceptual ratings and factor scores. The upper right hand triangle pertains to the Amsterdam data, the lower left hand triangle to the Nijmegen data.

	syl./s	pause time	pause proportion	dragging brisk	quick slow	factor scores
syl./s	-	.101	.370	.355	-.686 ^a	.579 ^b
pause time	.345	-	.960 ^a	.027	.347	-.304
pause proportion	-.080	.587 ^b	-	.055	.146	-.139
dragging brisk	.737 ^a	.097	.195	-	-.786	.873
quick/slow	-.677 ^a	.144	.054	-.925	-	-.973
factor scores	-.646 ^a	.247	.159	-.892	.981	-

a. $p < .01$ in a two-tailed test

b. $p < .05$ in a two-tailed test

not come as a surprise. The fact that syl/s is a better predictor of perceived briskness than of perceived reading speed in the Nijmegen data, whereas the results for the other data set are exactly the other way round, strengthens our belief that extreme caution should be exercised in interpreting these results. Any generalisation beyond the limits of this study seems unwarranted. The only exception that might be safely made to this rule is the confirmation of Butcher's (1981) finding that speaking rate, expressed in syllables per second reading time, is not a very good indicator of perceived tempo, when the text to be read consists of several paragraphs.

6.1.7 *Discussion and conclusion*

In this subsection we will deal with a number of topics which are mainly methodological in nature, and therefore apply to all factors discussed above (and to be discussed below). As described in chapter 3 much effort has been put into ascertaining that all rating scales meet a number of requirements, among which linearity played a particularly important role. No such precautions could, however, be taken in choosing the speakers and the speech material. Yet, when it comes to evaluative judgements, many of the acoustic variables will appear to be of the beta type, i.e., form scales with two mostly negative poles. To give just one example: completely monotonous speech is very unpleasant to listen to and most probably will not induce the idea of great expressiveness. Extremely large pitch fluctuations, on the other hand, will also render the speech highly unpleasant and most likely will not give the impression of expressiveness. Thus the amount of pitch variation is only positively correlated with perceived voice dynamics and voice appreciation as long as it lies within a certain range, which need not be very large. Outside that range the amount of pitch variation correlates negatively with voice appreciation and probably also with perceived voice dynamics.

In future experiments with perceptual judgements on running speech care should be taken to independently assess the ranges within which acoustic variables used to explain the judgements correlate linearly with rating scales. This is the more important as the number of speech samples judged becomes smaller. In a study like ours, employing eleven and twelve speech samples, almost any correlation between an acoustic variable and a rating scale may be destroyed by a single speech sample obtaining a value on the acoustic variable that is on the wrong leg of an U- or W-shaped correlation with the rating scale.

A quite general finding is the large difference between the Amsterdam and the Nijmegen data, which is present in almost all factors. Although we cannot completely rule out that possibility on the basis of the statistical data alone, we think it highly unlikely that the great differences between the groups are due to the fact that the groups are on alternate legs of a highly non-linear correlation graph. We rather have the impression that the results are due to a very restricted range in most acoustic variables. This hypothesis

certainly applies to all acoustic variables that in our study fail to display a relation to a rating variable that is well established in research on speech pathology. Including pathological speakers in a group greatly expands the ranges in both the physical and the perceptual variables, thus enhancing high correlations. We have tried to make the range of differences between speakers as large as possible. Yet there are good reasons to take the 'too narrow range' explanation quite seriously if only since we purposefully excluded pathological speakers. Moreover, the numbers of speakers are far too small to ensure large ranges on all possible variables.

An interesting objection to the 'too narrow range' reasoning can be made, viz. that in normal every day life the ranges in speaker characteristics we have to deal with are not very large either, most individuals being close to the mean. Still people are able to infer - and are constantly inferring - personality characteristics from speech features. Moreover, the reliabilities of the ratings did not give rise to the suspicion that the listeners had great difficulty in reaching an overall agreement with respect to the scale positions of the speakers. In combination with the pervasive finding that the attribution of personality characteristics on the basis of speech samples is highly stereotypical (cf. Scherer, 1979, for an overview) the suggestion can be made that the ratings do not bear a simple and direct relation to single acoustic properties of the speech signals.

The single most important conclusion that must be drawn from the analysis of the product-moment correlations between acoustic measures and perceptual ratings is that only very occasionally did an acoustic variable appear to be powerful enough to account for more than half of the variance in the corresponding set of perceptual scores. If we require that a single variable explains more than 50% of the variance in the ratings in both data sets, no measure can be found that fulfills the condition. This finding is not very encouraging, but it does by no means warrant a far-reaching negative interpretation either. What we have done so far is trying to approach perceptual ratings of exceedingly complex stimuli on fairly global features by very simple one-dimensional acoustic measures. Thus it should have arisen surprise if not suspicion if we had succeeded in explaining the bulk of the information (variance) in the ratings by means of isolated acoustic measures. We will now attempt to approach the problem using the somewhat more sophisticated technique of multiple regression analysis, that allows to combine a number of measures to explain the ratings.

6.2 MULTIPLE REGRESSION ANALYSIS

Textbooks on multivariate statistical techniques often caution their readers not to use multiple regression analysis unless the number of observations is much larger than the number of predictor variables. If the number of predictors and observations is about equal, a powerful means of testing the

meaningfulness of the results of a multiple regression analysis, viz. repeating the analysis for different subsets of the observations, is no longer available. Such a means of testing the results is necessary since the regression analysis will always produce a solution of the regression equations (unless the number of observations is smaller than the number of predictors). The solution consists of regression coefficients, i.e., the weights with which the observation values must be multiplied in order for the weighed sum to be an optimal approximation of observed criterion values. More often than not these regression coefficients are very difficult to interpret. In a situation in which the regression coefficients themselves do not give much information to go by when it comes to a conceptual interpretation and evaluation of the validity of the outcomes, a repeated analysis of subsets of the original data set is a powerful means of checking the stability of the results.

In our research the number of observations equals 11 or 12; thus the predictor variables outnumber the observations. If we still dare to employ multiple regression analysis in our study, it is only because we have two comparable data sets, so that we are able to cross-check the outcomes.

Multiple regression analysis can be performed in two different ways (Cohen & Cohen, 1975). The straightforward approach solves a single mean-square minimization problem, or, in other words, all predictor variables are entered into the regression equation simultaneously. In the stepwise (or hierarchical) approach the criterion values are first approximated by the single predictor variable that does this job best (i.e., the predictor which correlates highest with the criterion). In the next step that predictor is inserted which explains the greatest part of the variance not accounted for by the predictor used in the first step, and so on until either all predictors are inserted or some statistical criterion is met. We have exclusively used the stepwise procedure as it is available in the IBM SSP program package.

As long as the predictor variables are mutually independent (have low correlations) the stepwise procedure yields two kinds of information. First it generates the weighing coefficients in the regression equation

$$b + \sum_j a_j \cdot P_j = \hat{C} \text{ for } j \in \text{subset of inserted variables} \quad (2)$$

where P_j are the vectors of predictor measures and a_j are scalars chosen in such a way that the RMS distance between the vector $\hat{C}$ of predicted values and the vector C of observed criterion measures is minimized. The second sort of information is a rank ordering of the inserted predictor variables in terms of their contribution in explaining the criterion. The reliability of the rank order information is very sensitive to non-negligible correlations between predictors. If such correlations exist, it may well be the case that a number of predictor variables account for about the same portion of the variance in the criterion. The stepwise procedure will choose the single predictor that explains most variance for insertion into the regression equation. The corresponding portion of the variance being removed from the

pool of unexplained variance, the remaining predictors which do a job that is almost as good as the one done by the candidate which had the luck of being slightly better, may not be inserted into the equation at all. This may give rise to incorrect conclusions with respect to the relative importance of the predictors in explaining the criterion.

The SSP stepwise regression program offers the possibility to force predictor variables into the regression equation. If forced variables are used, the remaining so called free variables are not eligible for insertion into the regression equation until all forced variables have been used. We have adapted the SSP stepwise multiple regression program in such a way that it can perform a completely hierarchical analysis, i.e., the user cannot only force some variables into the equation, but he or she can also establish the exact order of insertion *a priori*. A complete or a partial hierarchical analysis is an especially attractive option in a situation where one wants to know what part of the variance in the criterion variable that is not accounted for by one or more *a priori* chosen predictors can be explained by additional, perhaps not *a priori* obvious measures. We have used the expedient of the forced variables in a number of cases where those variables that were intuitively the most appropriate predictors were overruled by some other predictor that happened to have a higher correlation with the criterion measure.

From what was said in the preceding paragraphs it is clear that multiple regression analysis is a technique for forming a linear combination of predictor variables in such a way that the resulting approximation of a criterion variable is in some sense optimal. It is in this form - i.e., the formation of a linear combination of individual predictors, which are preferably mutually orthogonal - that multiple regression analysis is best known. The technique offers a very powerful, but less popular facility, viz. the introduction of interactions between independent variables and their use as predictors along with the individual variables (Cohen & Cohen, 1979). In order to be formally and conceptually correct the use of interactions requires the use of a hierarchical analysis in which the individual variables are inserted into the prediction equation before their interactions. The formation of interaction variables in this procedure is effected by the multiplication of corresponding elements of the observation vectors.

In our research the expedient of interaction variables appears extremely attractive, if only since it might enable an investigation into possible causes of the observed differences in the predictive behaviour of many acoustic variables between the Nijmegen and Amsterdam data sets. Unfortunately, the use of interactions only helps to increase the power of an analysis technique which can be considered as much too powerful for application to our data from the very beginning. Where the observations are outnumbered by the individual independent variables, the introduction of interactions - perhaps even second, third or still higher order interactions - only aggravates

the risk of producing completely spurious results. In our view the analysis of first order interactions can be justified as long as the outcomes are presented as interesting hypotheses which are worth further investigation. Furthermore, we have restricted the analysis to interactions that seem plausible, either for intuitive *a priori* reasons or were suggested by the results obtained thus far. Exercising extreme caution in the interpretation of the outcomes is the more obligatory because for one of the most interesting candidates, *viz* interactions of acoustic variables with speaker sex, we can no longer rely on a comparison of the two data sets. A similar remark applies to interactions between text and acoustics in the other data set, interactions that were suggested by the results of the analyses in chapter 3. The discussion of the latter interaction can be finished before it is begun: in no case did it contribute sizably to the explanation of the ratings. This confirms the idea that the readers in the Nijmegen experiment may have handled the texts differently, but that these differences have been far from systematic.

The variable sex of speaker might have been included in the data matrix. It appeared, however, that sex dummy code correlates .95 with mean F_0 . Therefore we have treated the sex interactions by the products of mean F_0 and other acoustic variables. A minor advantage of this approach is that a similar interaction can be formed in the Nijmegen data, be it that the meaning of the mean F_0 component is very much different there.

The multiple regression analysis generated a huge amount of data, most of which appeared to be difficult to interpret. Therefore we will not report the results in great detail, but restrict the presentation to a verbal account of findings that may have some relevance beyond the two specific data sets under study. Before proceeding to the results, an account of the procedure followed in carrying out the analyses must be given, though.

Two data matrices were formed, one for the Nijmegen, the other for the Amsterdam experiment. The Nijmegen matrix was 12 by 44, *i.e.*, 12 observations on each of 44 'variables', the Amsterdam matrix was 11 by 43. The forty odd variables were divided into two subgroups, one consisting of 12 acoustic measures and some first order interactions with mean F_0 to be used as predictors and the other comprising scale values and factor scores. The Amsterdam matrix has one column less since a five factor solution was used in contrast to the six factor solution for the Nijmegen data. The 16 scales included are the same as used in the preceding sections on simple correlations. The following acoustic measures were retained as predictors: four spectral slope measures, *viz* the maximum level in the 400-600 Hz region, and the level differences with the filter containing the F_0 mode, the filter at 1.6 kHz, and the filter at 5 kHz (*cf* Fig. 5.1), dynamic range (*i.e.*, the variance of the total signal intensity), tempo in the form of the number of syllables per second, the MASC and FAQ measures defined in by (5,13) and (6,1), the proportion of 10 ms samples where the signal is voiced, average F_0 , the

variation coefficient of the F_0 distribution in semitones, and, finally, pitch perturbation. The interactions included were F_0 mean times F_0 variation, perturbation, dynamic range, MASC, FAQ, and all four spectral slope measures. For each of the factors a fairly liberal choice of predictor variables was made and an analysis run with these predictors and the factor scores or the relevant rating scales in turn as a criterion. During a first run all predictors were left free. When it appeared that the intuitively most appropriate predictors were suppressed by less obvious ones, an additional analysis was performed using the forced variable option. The need for a second analysis run occurred quite often. With some very minor exceptions, however, forcing variables did not alter the rank order of the free predictors found in the original run, nor the total amount of variance explained after their insertion. As can be seen from the tables in the subsections on simple correlations, there are a few significant correlations between predictor variables, in the Nijmegen material slightly more than in the Amsterdam data. We have no reasons to suspect that these correlations have seriously interfered with the results, though.

6.2.1 *Voice dynamics*

If we restrict the acoustic variables used for the prediction of perceived voice dynamics and general evaluation to measures of pitch and amplitude variations, % voiced plus spectral slope measures, it appears that a few measures suffice to jointly explain between 50% and 85% of the variance in the ratings. However, in no case is the linear combination of measures which is formed similar in the two data sets. The dynamic range seems to play a major role, since it appears as the most powerful predictor for the factor scores and the scale values on the scales 'melodious $\leftrightarrow$ monotonous' and 'expressive $\leftrightarrow$ expressionless' in both data sets. Only the signs of the regression coefficients are opposite in the two data sets. Thus it must be concluded that either the seeming importance of dynamic range is completely due to chance, which is not incompatible with the common knowledge that amplitude is not the principal factor in the perception of speech (cf. van Katwijk, 1974), or that dynamic range really does matter, but that it functions differently in the two data sets. In general, the sign of the regression coefficient of dynamic range is intuitively correct in the Amsterdam material, except in the case of the factor scores; in the Nijmegen material the sign of the regression coefficient is always counterintuitive.

The coefficient of variation of F_0 comes only at the second or third place when it comes to explaining ratings on the liveliness scales and the factor scores. As a predictor it seems to be equally important as pitch perturbation. Generally, the signs of the regression coefficients are intuitively correct.

With respect to the general evaluation scales the situation is much more complicated. The proportion of time the signal is voiced is virtually the only variable that operates in an intuitively acceptable way: voices tend to be

perceived as more beautiful when the proportion of voiced intervals increases. Spectral slopes seem to play a role here too, but their operation is to a large extent dependent on the data set under analysis.

6.2.2 *Articulation appreciation*

The articulation appreciation factor is an excellent example of the power of multiple regression analysis in approximating a criterion variable. In the Nijmegen data the combination of MASC, FAQ, spectral slope at 5 kHz and spectral slope at F_0 , in that order, suffice to explain from 80% to 95% of the variance in the ratings. The two provisional formant based measures seem to account for largely different portions of the variance in the ratings. Since the MASC depends on average formant frequencies only, whereas FAQ is derived from standard deviations and bandwidths their complementary operation is not too surprising. This, however, is not to say that the joint action of formant based measures and spectral slopes is easy to explain.

In the Amsterdam material the same set of variables accounts for no more than a negligible proportion of the variance in the ratings. Adding the sex information (by adding mean F_0) and the sex $\times$ MASC interaction raises the proportion of explained variance to the same high level reached in the Nijmegen material. The problem of interpreting this result is, of course, only aggravated.

Thus the somewhat surprising conclusion is reached that extremely global measures based on formant statistics and spectral slopes can explain the bulk of the variance of perceived articulation quality, but that the finding is hard to interpret. For the time being we prefer to observe extreme caution in handling this result. More research is needed before this finding can safely be used.

6.2.3 *The sociolinguistic factor*

Since the sociolinguistic factor did not emerge from the Amsterdam factor solution, we will concentrate the discussion on the ratings on the tenseness and self-confidence scales. As to the factor scores in the Nijmegen material average F_0 , spectral slope at 5 kHz and dynamic range are the most important predictors, together accounting for 72% of the total variance in a way which is intuitively acceptable: high F_0 combined with a steeply falling spectrum gives rise to factor scores near the negative pole of the factor. Turning to the scales the pattern of the factor scores is found again in the Nijmegen material. The three variables just mentioned account for 72% of the variance in the tenseness ratings and for 80% of the variance in the ratings of self-confidence. The results can by no means be reproduced in the Amsterdam material, however. There pitch perturbation and the proportion of voiced segments take over the role of average F_0 and spectral balance, whereas the contribution of dynamic range is much larger. It is quite possible

that the differences between the data sets are due to the fact that the Amsterdam group of talkers comprised both females and males, contrary to the all-male group in Nijmegen. It is not a very likely assumption that women are judged less self-confident only because their average F_0 tends to be higher than in most males. The conversion of the average F_0 values to z-scores separately for males and females does not improve the correlations.

6.2.4 *The potency factor*

The situation is again fairly clear and simple in the Nijmegen material. Spectral balance (i.e., the slopes at F_0 and at 5 kHz) account for approximately 80% of the variance in the ratings on the loudness and power scales as well as in the factor scores. It must be noted, however, that when left free average F_0 is by far the most powerful predictor of the ratings on the scale 'powerful $\leftrightarrow$ weak'. Males are perceived as more powerful if their F_0 is lower. Some comment on the sign of the regression coefficient of the spectral slope at 5 kHz is necessary. From what is known about speech perception one would expect that perceived loudness should increase when the spectral slope at high frequencies becomes shallower. In our data the contrary is true, however; the stimuli are rated as louder when the spectral energy is concentrated in the region below 1 kHz. A similar remark applies to the ratings on the scale 'powerful $\leftrightarrow$ weak' and the factor scores. Apparently our raters preferred a lax kind of loudness over a more tense alternative, where increased loudness is the result of a greater phonatory effort.

In the Amsterdam material the situation is much less clear. Spectral balance nor average F_0 explain a sizable percentage of the variance in the ratings. Here the proportion of voiced segments is most important, but it is by no means obvious how this should be interpreted. Forcing average F_0 and spectral slopes into the regression equation leads to the result that five variables account for no more than between 50% and 60% of the total variance to be explained.

6.2.5 *Voice quality*

In the Nijmegen material spectral balance (the slopes at F_0 and 5 KHz) plus average F_0 suffice to explain between 76% and 92% of the variance in the ratings on the scales 'dull $\leftrightarrow$ clear', 'husky $\leftrightarrow$ not husky', 'high $\leftrightarrow$ low' and 'deep $\leftrightarrow$ shrill' plus the factor scores on the static voice quality factor. The most interesting finding here is that pitch perturbation is completely suppressed by average F_0 and spectral balance. The same remark is true for the Amsterdam material, with one notable exception: in this material pitch perturbation operates in the explanation of huskiness in exactly the expected way. Voices are rated as more husky when the pitch perturbation increases. The relation is not very strong, though. For the rest the explanatory power of average F_0 and spectral slopes is clearly lower in the Amsterdam material.

This is completely in line with the findings from the preceding sections.

One additional fact is worth mentioning, viz. the counterintuitive sign of the regression coefficient of the spectral slope at 5 kHz that was noted before and appears again in the voice quality factor including the scale 'deep $\leftrightarrow$ shrill'.

6.2.6 *Tempo*

In the Amsterdam material speech rate measured in syllables per second is the best predictor of ratings on the scale 'quick $\leftrightarrow$ slow' and of the factor scores. The percentage of variance explained is not very high, though. Addition of dynamic range increases the proportion of explained variance to approximately 83%. As to the scale 'dragging $\leftrightarrow$ brisk' the roles of syl/s and dynamic range are interchanged. Moreover, together they account for not more than 72% of the total variance. The contributions of absolute and relative pause time are completely negligible in all cases.

In the Nijmegen material the same pattern emerges, with one minor difference. Absolute between-sentence pause time seems to outperform dynamic range when it comes to explaining ratings of briskness and factor scores. Here again the total proportion of explained variance is lowest for the briskness, i.e., the least 'physical', scale.

The role of dynamic range in explaining tempo ratings, as well as its role in explaining all other ratings, is extremely difficult to interpret. If it were not for the fact that more often than not it operates in both data sets its emergence as a predictor would have been blamed to chance, the more so since the signs of the correlations between dynamic range and ratings are mostly opposite in the two materials. We did not succeed in tracing the cause of that effect to some conceptually interpretable interaction of dynamic range and another acoustic measure.

6.3 GENERAL DISCUSSION: METHOD AND RESULTS

It is time now to try and summarize the results of the correlation and regression analyses and to make some general comments on the method of research. The analyses aimed at finding 'hard' acoustic explanations of 'soft' perceptual judgements of a wide range of properties, running from somewhat fuzzy and broad characteristics like voice dynamics, articulation quality and potency to properties of speech signals like tempo and voice colour, which can be given a relatively precise meaning in psycho-acoustic terms. Not surprisingly, we have been less successful in finding acoustic measures to explain judgements the fuzzier the attributes rated were.

A number of quite general comments on methodology have already been made in 6.1.6. It has been noted, for instance, that our results may have suffered from the fact that the range of some variables was too small or from

the fact that they were in a range where the correlation between acoustic measures and perceptual ratings is non-linear. We will not reiterate these topics here, but rather elaborate a number of points which were implied in what was said above, but which have not yet been discussed explicitly. We have been very reticent in accepting the outcomes of the multiple regression analysis mainly because of the differences between the two data sets. In other words, we have deliberately opted for a strategy that minimizes the risk of a type I error, i.e., the acceptance of a relation where in fact the observed correlation is due to chance. However, the two data sets under comparison were different in a number of respects. One set comprised both females and males, all of whom read aloud the same text; the other consisted of males only who read out two quite different texts. It is not too difficult to think of a number of arguments why these differences should not matter very much. However, it is equally easy to argue that the differences have been disparaged and that they have in fact prevented the emergence of a simple underlying structure. Two of those arguments deserve to be mentioned.

The results of the investigation of sex-dependent standards (cf. chapter 2) suggested that the sex composition of the talker groups is not overly important, since the standards by which females and males are rated are fairly similar. This reassuring finding can, however, be refuted by the conclusions arrived at in other studies, that there is a large discrepancy between what people say is ideal and the standards which are actually operative (Gilbert Deutsch & Strahan, 1978). Thus it is quite conceivable that we would have obtained less ambiguous results if we had been able to compare two all-male or all-female groups of talkers.

At the present stage of the research the comparison of two mixed-sex groups most probably would not have led to more clear-cut results. There are reasons to assume that some (if not many) of the acoustic measures used in this study tend to turn out differently for men and women. We are still left with unsolved scaling problems related to the anatomical and physiological differences between the phonatory and articulatory organs of the sexes. Formant frequencies and bandwidths may have to be weighed and scaled differently for males and females. There are indications that the glottal pulses in females do not only have a higher repetition frequency than in the average male, but that there are additional systematic differences in spectral slopes (Monsen & Engebretson, 1977). Together with the observed tendency for females to use a more precise articulation (cf. Nuijten, 1962; Smith, 1979), these differences might complicate the pattern of acoustic measures to the point where straightforward data processing is doomed to fail to detect the system.

Another, completely independent, topic must be brought up. We have acquired judgements and made acoustic analyses of sound material that was definitely linguistic in nature. The fact that acoustic measures of the level and variability of F_0 and speech intensity do not explain perceptual ratings of melodiousness and expressiveness may be largely due to the circumstance

that the stimuli to be judged consisted of meaningful speech instead of meaningless non-stationary noises. It is quite possible that the impression of liveliness and melodiousness of a rendition is not primarily contingent on the range of pitch and amplitude fluctuations, but rather on the way in which the pitch accents are distributed through the text, the variety of pitch contours used and perhaps also the appropriateness of their choice in view of the syntactic structure and the semantic and pragmatic meaning of an utterance. In a similar vein the impression of preciseness of articulation may not primarily be a function of the ranges of formant parameters, but rather of the appropriateness with which these parameters vary as a function of time. The appropriateness of static and dynamic aspects of formant parameters seems impossible to assess without reference to the linguistic message which is conveyed by the stimuli under judgement. Global, automatically obtained acoustic measures as used in the present study are fundamentally incapable of taking the underlying linguistic structure into account. Thus it seems necessary to develop objective measures of an essentially different, much more sophisticated nature, if we eventually are to come up with satisfactory acoustic explanations of perceptual ratings of meaningful connected speech.

In chapter 3 it was found that many of the talkers in the Nijmegen experiment seem to obtain widely diverging factor scores on many factors for the two texts. In developing acoustic measures, on the other hand, the requirement that the measures be stable within speakers and differentiate between them played the principal role. One can ask whether this is not an incompatibility which renders any attempt to explain ratings acoustically futile from the very beginning. We strongly feel that there is only a seeming incompatibility here. First of all it is extremely difficult to think of sensible acoustic measures which are not essentially stable within talkers; a measure which would prove to be dependent on the part of a reading used for its computation could never be given a single value. Next, the between-text differences in the ratings seemed most prominent in those aspects of the stimuli connected to dynamic aspects of the readings or to differences in reading performance. These are exactly the aspects related to semantic/pragmatic content of the texts discussed in the previous paragraph and which our measures fail to capture. Finally it should be remembered that the between-text differences in the factor scores seemed to be inflated compared with the ratings on the individual scales, where the number of times that a talker appeared to differ significantly from himself was fairly low. Therefore we do not see any reason to doubt whether the measures employed in the correlation and regression analysis are characteristic of the stimuli.

Up to now the investigation of the causes of the lack of success of our acoustic measures to explain perceptual ratings has concentrated on the former. It is time now to scrutinize the ratings. Our work has been presented as a contribution to our knowledge of psycho-acoustics of running speech. In doing so it is implied that the ratings of our judges represent fairly objective descriptions of the stimuli. It is also implied that these descriptions are

accurate or, in other words, valid. Note that the validity of the ratings does by no means follow automatically from their observed reliability. It is quite conceivable, however, that the very rating procedure has given rise to a substantial shift of the level of description in the direction of attributions. The raters were not urged to give a careful objective description of the stimuli, but rather to give their first impressions. Thus there may very well be a more objective level of perceptual description of the stimuli, involving much less interpretation and therefore much better amenable to an explanation in terms of acoustic measurements. Such an approach, inevitable and insight promising as it may seem to be, is not without difficulties and limitations itself. Increasing the level of detail in the description naturally calls for increasing investments of time and money in training the raters, without any guarantee that the minimum level of proficiency necessary for the job will be reached. Studies on the rater reliability in the phonetic description of voice quality are virtually non-existent (van Bezooijen, 1984). Also, if reliable ratings of some seemingly basic characteristic can be obtained, e.g., ratings of pitch range that can be assumed to be more basic than ratings of melodiousness, it remains very difficult to explain these ratings by means of simple acoustic measures (Cranen, Boves & van Bezooijen, 1983). Finally, an approach based on successive moves towards greater detail in the description will inevitably invoke increasingly complex data processing techniques in order to trace the relations between acoustic characteristics of speech stimuli and the inferences made at the attribution level.

6.4 CORRELATIONAL VERSUS EXPERIMENTAL RESEARCH

Regardless of the sometimes sophisticated techniques used and the proliferation of words like 'experiment', the work reported on here is essentially correlational in nature and not experimental. We do not have manipulated certain parameters in order to study their effects in any controlled way. That is why we have stated before that the present work can at best generate hypotheses that are worth testing. In this sense our work is preparatory to psycho-acoustic research and does not constitute psycho-acoustic research itself, if only since the latter discipline is characterized by its experimental approach in the first place.

One very invidious limitation of correlational research is that any trait which happens to be absent from the sample population will not show up in the results, even if it is extremely important in actual reality. In the same vein, but quite contrary in its effect, a specific characteristic that is prominently present in the sample may be overvalued. Moreover, relations between characteristics may differ from population to population, the more so when there are no reasons why an observed relation should be causal.

An especially nasty problem in descriptive work is the possibility of natural covariation of features. It has already been pointed out in chapter 3

that differential reading performance may have interfered considerably with the ratings on many scales in the Nijmegen rating experiment. If such an all-important trait is effective, it can be there for two quite different reasons. One possible explanation is the so called 'halo-effect' (Guilford, 1954, chapter 11) that can informally be described as follows: unless specifically trained not to do so, people unconsciously make inferences on the level of attribution and when asked to rate voices on scales alluding to lower levels the actual ratings are to a large degree the result of a backwards 'inference', starting from the attribution level. The halo-effect clearly belongs to the realm of experimental artifacts. In descriptive work, however, it is extremely difficult to distinguish from the possibility that we have to do with some sort of real covariation of characteristics. In a large scale replication of the judgement experiment (using 74 speakers and 14 scales) Fagel found that the voice dynamics and the articulation quality factors merged into one very broad evaluation factor (Fagel, personal communication). One way to explain this finding, which is quite surprising given the neat separation of these factors in the two data sets on which the present study is based, is to assume that speakers who are melodious, expressive and pleasant to listen to tend to have a cultured and polished pronunciation. To the extent that this covariation does occur, purely descriptive research will fail to confirm theoretical distinctions that can be given a seemingly solid phonetic basis.

Part of the problems that have shown up in our correlational research suggest an experimental approach to their solution. It is quite possible to manipulate a limited number of features that are supposed to be most important, as has been shown by the work of Brown, Strong & Rencher (1974). A subset of the features, notably pitch range, the number and place of pitch accents, number and place of pauses etc., can be manipulated by means of an analysis-resynthesis system. An eventual test of the hypotheses with respect to the effect of combinations of certain levels of feature values should, however, preferably be carried out using a system for synthesis by rule, provided that it is capable of producing near-natural synthetic speech. If the system is built in such a way that the synthesis rules can be changed flexibly and almost automatically, the stimuli for the judgement experiment can be generated with a minimum of experimenter intervention. More importantly, however, there is an increasing body of evidence that features do not combine in a simple additive way. The auditory effect of a feature may depend on the absence or presence of other features. Moreover, a feature that plays a secondary role in the presence of some other one, may become of primary importance if the other feature is absent. For instance, removing a prominence lending pitch movement by means of an analysis-resynthesis system does not necessarily cause the corresponding syllable to lose its prominent character. Apparently, in the absence of pitch movements concomitant features like duration and preciseness of articulation can take over the normal role of pitch movements. This kind of effects cannot be easily controlled in an experiment using analyses-resynthesis techniques.

6.5 SUMMARY AND CONCLUSIONS

In this chapter we have attempted to find acoustic explanations of the perceptual ratings of the speech of two fairly small groups of speakers, using the measures developed in chapter 5.

Our preference for measures that can be obtained automatically and that are stable in the sense of representing long-term speaker characteristics clearly favours postural aspects of phonation and articulation over the more procedural, dynamic aspects. To the (very large) extent that perceptual ratings are determined by the dynamic aspects of the speech signals any approach based on long-term average measures is bound to fail when it comes to finding acoustic explanations of listener judgements. The outcomes of the work reported upon in the present chapter testify of the limitations of the postural approach and show the need to redirect the research in phonetics towards the dynamics of articulation and its acoustic and perceptual consequences.

The power of our set of measures in explaining perceptual ratings was examined by simple correlation analysis and by means of multiple regression. In order to minimize the risk of mistaking chance correlations for causal relations two data sets were analysed, one comprised of ratings of six males reading two different texts (the Nijmegen material) and one comprising five females and six males who all read the same text. The analyses were structured by the factors found in chapter 3.

Analysis of simple correlations showed that there is hardly any direct relationship between our acoustic measures and perceptual ratings. Combinations of individual measures and a very restricted subset of interactions proved to be capable of 'explaining' the bulk of the variance in the ratings but more often than not the combination of measures differed between the two data sets or was difficult to interpret. The only notable exception to this rule is voice colour or static voice quality. The exceptional behaviour of that factor can be explained by noting that it comes closest to a static, postural aspect of the speech signals and the speakers.

A number of hypotheses and methodological refinements have been proposed to arrive at a better understanding of the remaining factors. Apart from the need to develop measures that are better suited to deal with the medium-term speech dynamics than the predictors used in the present study, the need for a tighter control of the range of several parameters in the speech material was pointed out as well as of the way in which those parameters are combined. Explicit control of the ranges of the parameters is necessary since many can be suspected of non-linear correlations with perceptual ratings. Control of the ways in which parameters are combined should lend insight into their relative importance, into the effects of hierarchically structured complexes of parameters and into the manner in which parameters can change rules or can substitute each other altogether. Finally, some emphasis was laid on the possibility that the majority of our rating scales, although

explicitly meant to function on the level of objective description, have operated on the level of inferred attributions. In other words: it seems to be quite possible that our scales have suffered from the same drawbacks that were noted earlier with respect to the factors, viz. that their scores constitute the results of an intricate, and most probably highly non-linear, weighing of a large number of acoustic parameters. If this suspicion can be substantiated, it shows the need of perceptual descriptions on a much lower, phonetic level.

The status quo and perspectives

7.0 INTRODUCTION

One of the papers that marks the beginning of modern acoustic phonetics is Homer Dudley's 'The carrier nature of speech' (Dudley, 1940). In this paper the broad band speech signal is described as a high frequency carrier wave modulated by low frequency linguistic information. The analogy with radio transmission techniques is quite apparent, as is the implication: if we succeed in extracting the low frequency linguistic information before transmitting the speech a considerable savings in bandwidth will be the result.

The research efforts in fields like communication technology, acoustic phonetics and linguistics have yielded vast bases of knowledge and experience. Yet, when it comes to solving problems in every day communication between people, be it relatives, acquaintances or complete strangers, the experts are often at a complete loss. Quite naturally, this observation raises questions as to the adequacy and the comprehensiveness of the knowledge collected so far. Using some afterthought, the incapability of present day phonetics and linguistics to handle problems of every day communication is not difficult to explain. In fact, it has been explained in detail by all those scholars we have pointed out that what is known as linguistic competence is in fact part of a more comprehensive communicative competence. If the linguistic competence accounts for the knowledge an ideal speaker-hearer has of his language, a terse description of the communicative competence would be to say that it accounts for the knowledge the ideal speaker-hearer has about the non-ideal situation dependent aspects of a message. In other words, the communicative competence is an account of the real world framework that the linguistic competence needs to function in.

The notion of a linguistic competence embedded within a communicative competence raises some interesting questions, one of which addresses the trace of the boundary of the smaller one. Linguistics may be defined as the study of the conventionalized, and therewith necessarily rule based, mapping of meanings onto perceivable forms, be it sounds, pictures on paper or otherwise. If language is conceived of as a conventionalized symbolizing system, we have to answer the question what part of a perceivable linguistic form is conventional and thus carries some meaning. Thus it is necessary and worthwhile to investigate which part of what customarily are called para-

and extralinguistic phenomena are indeed conventionalized and thus essentially linguistic in nature. We have tried to approach this problem by searching for systematic relations between perceptual ratings and paralinguistic features of speech signals. The difficulties that plague research in this field are well known and have been dealt with in the literature at various occasions. Scherer (1982:87 ff) gives an extremely clear and concise summary by stating that there is no agreed upon system to classify and describe the phenomena, resulting in a perfect babel of tongues and a consequent confusion of all kinds of levels of description.

The emphasis put on the development of a methodology to deal with para- and extralinguistic phenomena in a systematic way that enables meaningful communication of the results does not mean that we believe in the eventual emergence of a description system that can be applied in a standardized manner irrespective of the goal of the research. Rather we believe that the future will see a multi-faceted system, the core of which is applicable in all circumstances but with details that must be emphasized or de-emphasized depending on the particularities of the research situation. Nor does the fact that we feel that methodological research should be given priority over research into the things that are of eventual importance mean that we think that methodological research can progress without feedback from the fields where the newly developed techniques can be applied.

The research reported in this book complies in many aspects with the course of action sketched in the preceding paragraphs. Its final aim was in the realm of linguistics, be it a brand of linguistics that conceives its object of study as wider than the (non-)behaviour of the ideal writer-reader. To recapitulate, the final aim of our work was to find out what in a speaker's voice and pronunciation makes him or her sound trustworthy, pleasant and polite, or, perhaps, exactly the contrary. Obviously two things had to be accomplished in order to reach that goal: we had to come up with reliable descriptions of a number of speakers' voice and pronunciation quality and with listener ratings relevant to the perceived personal characteristics of the speakers. However, from the very beginning of the work it became clear that for the time being we had to be satisfied with much humbler pursuits and that we would have to invest a major part of our time and effort into preparatory methodological research.

Our work, then, is comprised of three major undertakings. Firstly, we have tried to develop 'objective' acoustic (and to a lesser extent physiological) measures to separate and describe phonation and articulation. Acoustic measurements have the advantage of being easy to define and therefore of being easy to repeat. They are, however, quite often difficult to interpret with respect to their effects on speech reception. Secondly, an attempt has been made to obtain perceptual ratings of speech samples. Here the problems are an exact mirror image of those in the field of acoustic measurements. The results of the ratings can justifiably be said to have ecological validity but

they are hard to define and difficult to repeat. The third and last undertaking consists of an attempt to relate the acoustic and perceptual domains, trying to bring the problems in both domains a little bit nearer to a solution.

7.1 THE STATE-OF-THE-ART IN ACOUSTIC DESCRIPTIONS OF SPEECH SIGNALS

Coming now to a summary of the state-of-the-art in the field of acoustic measurements we can say that a large number of parameters and features can be extracted and measured quite accurately. Among the parameters that allow reliable automatic measurement are speech power (amplitude envelope or intensity) pitch, spectral envelope and formant frequencies and bandwidths. Not unexpectedly, the measurements are easier to carry out and lead to more reliable results if they entail less 'interpretation'. For example, stable and accurate estimates of the spectral envelope, which do not involve any interpretation of the data, are easy to obtain, whereas estimates of formant parameters are more difficult to obtain and at the same time more variable, since they require explicit interpretation of peaks and troughs in the spectral envelope.

From the research reported in chapter 4 it is apparent that physiological signals recorded simultaneously with the acoustic speech wave can yield useful extra information. For obvious reasons, however, that extra information can only be used as an aid or a guidance in performing measurements on the acoustic speech signal. In our view it does not make sense to try and alleviate this restriction on the use and usefulness of measures derived from physiological recordings, if only since under normal circumstances listeners do not have access to the speaker's physiology except via its impact on the acoustic speech signal. Yet, the importance of physiological recordings as suppliers of boundary conditions in the interpretation of acoustic analysis results can hardly be overestimated.

Using the electroglottogram (EGG), the photoglottogram (PGG) and the subglottal pressure (P_{sg}) alternately as guidance in the processing and as criteria against which the processing results can be evaluated, it was shown that closed glottis interval inverse filtering to recover the glottal flow pulses from vowels produced in running speech leads to accurate and reliable results. The EGG, the physiological signal that is by far the easiest to record with a negligible interference with normal speech production, appeared to be the most powerful, both in guiding the inverse filtering and when it comes to evaluating the accuracy of the results. We have developed a semi-automatic inverse filtering system that uses the EGG in order to establish the closed glottis interval. Since the inverse filtering is performed pitch synchronously, i.e., for each glottal period the closed glottis interval is determined and the optimal inverse filter established, it can be successfully applied to running speech.

The use of the EGG in order to establish the closed glottis interval is but

one of the many possibilities offered by the simultaneous recording and subsequent processing of physiological signals along with the acoustic speech wave. Even the very easily obtainable EGG offers many additional potentialities, most of which have not even begun to be explored. However, the potential power of simultaneous physiological recordings will not be brought to bear on phonetic research unless techniques for automatic processing of multiple simultaneous signals can be made available to speech scientists. Here too, most of the work remains to be done.

If it can be said that present day signal processing techniques are able to accurately extract many parameters, even when they are time varying, from the speech signal, a similar statement cannot be made with respect to the prevailing techniques for describing time varying analysis results. The availability of powerful digital signal processors which can analyse extensive data bases of running speech overnight enables one to generate huge amounts of analysis data. In order for that data to be information we need a new level of processing, the bare outlines of which are now perhaps beginning to become visible. The want of compact representations of time varying phenomena otherwise than by their mean and standard deviation has been painfully apparent all throughout the second part of this book and it has been consequential. We have been forced to be content with a static frequency domain representation of the characteristics of glottal flow pulses, only to conclude that much of the information that is most probably carried by the pulses is contained in the way in which the forms change as a function of time. Similar statements can be made with regard to intensity, fundamental frequency and formant parameters: they can be extracted from vast amounts of speech and the resulting parameter tracks must contain a wealth of information on speaker characteristics. Only we do not possess techniques to sort that information out from the masses of raw data, so that all the computation done to extract the parameters falls short in a painful way of its potential.

Although the work needed to develop techniques for an adequate representation of time varying parameters is not exactly linguistic in nature, linguistics will be among the firsts to profit from its results. For it is linguistic research and linguistic theory which must provide the criteria that determine whether a specific representation is adequate.

The last paragraph looks rather vague, but it is so by necessity. As said before, even the outline of the form and structure of the data representations we are aiming at are not known at this moment. Still we should like to make an attempt to suggest an approach to the problem. With regard to the fundamental frequency it seems worthwhile to continue research aiming at the automatic generation of a transcription in terms of the intonation grammar designed by 't Hart & Collier (1975) or in terms of competing models like the ones proposed in Scherer, Ladd & Silverman (to appear). As to postprocessing of formant data one could think of a system that first derives a segmental transcription of the analyzed speech and then computes

the values of a number of variable and tunable parameters in a speech production model in such a way that the (hypothetical) output synthesized by the model when fed with the transcription approximates the original speech signal in some optimal sense. In a much less ambitious intermediate solution the phonetic transcription could be made by a trained phonetician. In either case the actual formant tracks can be represented in terms of variable parameters of a more or less comprehensive and insight lending model.

7.2 THE STATE-OF-THE-ART IN PERCEPTUAL DESCRIPTION OF RUNNING SPEECH

The principal problems in dealing with associative perceptual ratings of running speech are their objectivity and reliability. Unfortunately, the degree of objectivity of a description is a very elusive property itself. One might contend, of course, that objectivity is identical with reliability. Even if we accept this identification, it is not of much help with descriptions that are essentially nominal in nature as is the case with almost all phonetic transcription, including transcriptions of intonation, since it is extremely difficult to define and calculate reliability measures for those descriptions. The situation seems less adverse in the case of ratings on scales that can be treated as being of interval level. Here, however, philosophical or methodological difficulties may arise with respect to the equation of objectivity, in the sense of not involving interpretations and attributions, and reliability. Statistical reliability measures per se are not able to discriminate between absence of interpretation and agreement on the way in which things are to be interpreted. To give an extreme example: ratings which are heavily influenced by stereotypical ideas will appear to be highly reliable from a statistical point of view, but few people will choose to call them objective, at least not once the stereotypical nature of the judgements has become known. An obvious way to effect this *demasqué* is by comparing the ratings with independent criterion measures.

Not only our work, but also the studies being conducted by van Bezooijen (1984) show that reliable ratings can be obtained on a surprisingly diverse set of vocal attributes, especially if the reliability of the averaged ratings is used. The claim that reliability of perceptual ratings of speech is not a great problem only holds as long as we deal with ratings on interval level scales. The segmental transcription of speech quite contrarily still poses immense problems. As long as we do not have techniques to formally establish their reliability even experiments in which more than a single phonetician describes an utterance data base are very seldomly conducted.

In the framework of the study reported here the problem of assessing the reliability of nominal measurements is only of a philosophical nature, since

we have employed equal-appearing interval scales throughout for obtaining listener judgements. Formally, *n*-point rating scales of the semantic differential type are not interval, but rather ordinal scales. Given a sufficient number of raters, however, scores on ordinal scales can be transformed to measures at interval level by means of scaling techniques like Thurstone's Law of Categorical Judgement. One of the nicest results of our work is the finding that in many cases application of a formal scaling technique is not strictly necessary. It appears that the arithmetic mean of the raw scale positions attributed to an object by a number of judges is almost identical with the scale values resulting from an application of Thurstone's Law.

It should be realized, however, that our finding that average scale positions and formal scale values are practically identical applies to scales that had all undergone an elaborate process of analysis and evaluation of their scaling properties. Specifically much effort has been put into attempts to ensure that all scales were of the so called alpha-type, i.e., that they run from a negative extreme to a positive extreme, instead of from a negative extreme via a neutral midpoint to another negative extreme. Care was also taken to ensure that the adjectives defining the extremes of the scale had an unambiguous meaning with respect to the type of stimuli to be rated. Finally, the results have been obtained with fairly homogeneous groups of raters, which were never very small; in no case did a group consist of less than 20 judges. Although the scaling technique employed in the present study appears to work well and to yield insightful results, it would be very useful to have a technique that is easier to employ but gives similar information. Specifically, our approach, which is based on Thurstone type scaling, is only applicable when the number of raters is sufficiently large to treat the ratings as approximations to normal distributions. But then, formal statistical techniques to assess whether a scale has been treated identically by all raters for all objects are perhaps only necessary when working with a fairly large group of essentially untrained and anonymous judges. When small groups of raters are used, it may well be possible to train them to handle a scale in a standardized way, irrespective of additional features of the objects to be rated which are not alluded to by the scale in question.

Turning now to the results of the factor analyses performed on the ratings obtained in our work we can say that we have arrived at a fairly comprehensive description of voice and articulation at a level that is probably closer to attributions than to 'objective' descriptions. At this level voice and articulation quality seems to be a phenomenon in which at least five mutually independent aspects can be distinguished. The most important aspect seems to be the liveliness, expressiveness and dynamism of a rendition, followed by the quality and correctness of the pronunciation. Osgood's potency factor comes next, followed by voice quality or speaking pitch level and finally, tempo. Most factors seemed to be confounded with general appreciation, in other words, with Osgood's evaluative factor.

7.3 THE RELATIONS BETWEEN ACOUSTIC AND PERCEPTUAL DESCRIPTIONS OF RUNNING SPEECH

We have seen that a large number of acoustic measures of connected speech can be obtained that can be proven to be statistically stable and capable of discriminating between speakers. Similarly, perceptual ratings on an even larger number of bipolar scales can be obtained that are highly reliable if the groups of raters are not overly small. Yet acoustic measures and perceptual judgements seem to belong to two separate worlds. In any case, our attempts to relate those two domains to each other via the way of product-moment correlation and multiple regression has not been too successful.

We have already pointed out a number of facts and circumstances that might in part explain the difficulties encountered in trying to predict perceptual ratings by means of acoustic measures. At this point there is no need to reopen these discussions, since no new arguments have appeared in this chapter. Apparently, there is a long and difficult way ahead that has to be travelled before we really can begin to understand how the acoustic properties of a speaker's voice and articulation determine listener's judgements. We might find eventually that at the level of attribution there is hardly any systematic relation with the primary acoustics since listener preferences and, perhaps, prejudices outweigh the information that the rater takes in via his or her ears. Even if this pessimistic (from the point of view of one-dimensional causal scientific thinking; optimistic from a number of alternative points of view) can be proved to be correct, there remains a lower and more 'objective' level of perceptual description where acoustic information is stronger than the personal evaluation. It is this level that future efforts in linking speech acoustics to speech evaluation should concentrate upon. And in order for these efforts not to be futile to begin with, much work has to be done to develop compact descriptions of the dynamic aspects of voice and articulation. It is hoped that the present book contains some information that will prove valuable for the scientists who are going to tackle these problems in their future research.

Development of an accurate closed glottis interval technique for inverse filtering

A1 0 INTRODUCTION

This appendix gives a detailed description of the closed glottis interval inverse filtering technique in which the electroglottogram (EGG) is used to determine the interval of each glottal cycle during which the glottis is closed

A1 1 The closed glottis interval approach to inverse filtering

The contention that the vocal tract during the production of a vowel can be considered as an all-pole filter is based on a model that treats the supra glottal cavities as a single tube of non-uniform cross-sectional area which is open at one end (at the lips) and closed at the other end (*viz* at the glottis), cf Figs. 4 2 and 4 3 Usually, the infinite termination impedance at the glottis is incorporated into the voice source This results in a model consisting of a constant current source which feeds into a vocal tract with a very low input impedance Allocating the glottal termination impedance to the source still leaves us with a vocal tract the transfer function of which is well defined Also the separability of source and system is not affected by this slight shift of the boundary between the two parts Even the replacement of the infinite glottal impedance by a very high, but finite constant impedance does not fundamentally change the picture

The situation changes drastically, however, if one realizes that during the open glottis interval the termination impedance at the glottis will obtain finite values which need not be much larger than the input impedance of the vocal tract Thus it appears that the voice production mechanism may pendulate between two states during each glottal cycle, one (*viz* the closed glottis interval, henceforward CGI) during which the linear all-pole model of Fig 4 3 holds and another (*viz* the open glottis interval) during which the values of the internal impedance of the source and the input impedance of the vocal tract are no longer very different and, even worse, constantly changing with time If the termination impedance is no longer constant, concepts like transfer function and formants loose much of their original meaning

Of course the glottal area is small during most part of the open interval and a high subglottal pressure will aid to provide a high impedance termination that need not be too different from the situation during the closed interval

Yet a detailed study of vowel waveforms leads to the conclusion that the formant frequencies differ slightly between the open and closed intervals. Differences in formant bandwidth (or damping) are even more apparent, especially for vowels with a high frequency F_1 (Fant 1979; see also the speech signals displayed in Fig. 4.12, especially the /a/ and /o/).

The variations of the formant parameters within each glottal cycle has a number of consequences. The first, and most apparent, is that the 'optimal' parameters of the anti-formant circuits are no longer defined in an unambiguous way. Taking the CGI values will lead to an inverse filter which is mistuned during the open glottis interval and vice versa. Setting the anti-formant circuits to an average of the optimal values for open and closed intervals leads to an inverse filter which is slightly mistuned during the entire period. Another consequence of the alternation between an infinite (or at least extremely high) and a finite (perhaps even fairly low) termination impedance of the vocal tract within one glottal period is that it becomes impossible to unambiguously define the source or, what is essentially the same, to separate source and filter. During the open interval the flow from the trachea into the vocal tract interacts with the formant oscillations. Thus the effective flow through the glottis may depend on the frequency of the formants, at least to some extent. This is a rather unsatisfactory situation that cannot, unfortunately, easily be remedied. Fant (1979) proposed to solve the problem by dividing the glottal periods into two parts and define source and filter for the open and closed parts separately. This is an attractive proposal if one is dealing with (deterministic) production models. We are, however, looking for an approach suitable for a situation in which non-deterministic signals are to be analysed. In such a situation it is most attractive to define the vocal tract transfer function $V(z)$ as the function during the CGI i.e., during the interval where the all-pole model is most appropriate. The consequence of this decision is that the glottal flow is defined as the actual flow through the glottis, including possible formant induced ripples superimposed on the transglottal flow during the open interval. The latter approach is also taken by Fant in a recent paper (Ananthapadmanabha & Fant, 1982). For a more detailed discussion of this topic from our own point of view the reader is referred to Cranen & Boves (1983).

A1.2 Determination of the CGI from the speech signal

From the discussion in the preceding paragraph it follows that, from a theoretical point of view, an attempt to separate phonation and articulation by means of inverse filtering will only yield meaningful results if the anti-formant settings are tuned to the natural frequencies of the vocal tract during the CGI. Wong, Markel & Gray (1979) have published an inverse filtering technique which is based on the detection of the CGI from the acoustic speech wave. It is interesting that their original motivation for determining the CGI was not in the first place the desire to comply with the theoretical considerations explained above. What Wong et al. were aiming at was the

development of an automatic inverse filtering technique in which the expertise of a trained operator was replaced by a formal (least-squares) optimization of the results. The reasoning underlying their least-squares approach runs, however, essentially parallel to the theory which led us to the CGI approach to inverse filtering. Since the formal derivation of the technique presented by Wong et al. is extensively explained in their paper, which is readily available, there is no need to repeat it here. It is sufficient to note that their approach assumes that the glottal volume flow function consists of an alternation of intervals of zero flow during the CGI and smooth pulses during the open glottis interval. It is also assumed that the first derivative of the flow function contains exactly one discontinuity, viz. at the moment of glottal closure. The differentiated glottal flow can be considered as the 'effective input' to the vocal tract model in Fig. 4.3 by interchanging the position of the radiation and the vocal tract boxes, a procedure that is perfectly legitimate as long as the complete system is linear. If an all-pole system is driven by an input which is an alternation of a spike, a zero interval and an interval of minor and smooth variation, the output signal consists of a superposition of damped sinusoids during the zero interval at the input and a combination of the input and damping sinusoids during the presence of the smoothly varying input. The spike in the input serves to excite the damped sinusoids. Each damped sinusoid corresponds with a formant. Wong et al. noted that a covariance linear prediction analysis (Markel & Gray, 1976) using an analysis window short enough to fit completely into the closed glottis interval should lead to a zero prediction error. Shifting the window one sample at a time and performing a covariance linear prediction analysis (LPA) for each window position should yield a zero prediction error as long as the window is completely within the CGI; as soon as the window contains a point of the signal where a non-zero contribution of the input is present, the prediction error can no longer be zero, because the signal is no longer an exact superposition of damped sinusoids. The prediction error is very large for all window positions which contain the spike exciting the system at its input. During the intervals of zero (theoretically, near zero in practice) prediction error the results of the LPA are an accurate estimate of the transfer function $V(z)$ of the vocal tract. Therefore, an inverse filter based on this estimate must be optimal and should be capable of reconstructing the glottal flow exactly, without any need of expert tuning.

In applying the technique developed by Wong et al. we have encountered two serious problems. The first, but the less important one, is that it is not at all easy to detect the CGI from the prediction error function. According to Wong et al. a unique local minimum marking the beginning of each CGI should be present in each period. They even suggest the error function as an attractive means for pitch detection. From the considerations underlying their approach it is clear that a unique local minimum is only present if the effective glottal flow function does not contain more than one discontinuity, which should be at the point of glottal closure. Rosenberg (1971) has propo-

sed a number of glottal flow functions, some of which do indeed have derivatives containing only a discontinuity at the point of closure, but others contain two or even three discontinuities, viz. at the moments of opening, closing and maximal glottal area. The glottal flow pulse proposed by Fant (1979) also has three discontinuities in its first derivative. The possibility, put forward by Holmes (1976) of vocal tract excitation during the closed glottis interval would, of course, completely invalidate the arguments of Wong et al. Even if Holmes is right there would, however, be no direct need to discard the assumption that the vocal tract transfer function is all-pole during the CGI.

The second problem is the strong vowel dependence of the overall level of the error function, especially the fact that for some vowels the error is uniformly very small. The effect is most apparent for rounded back vowels /u, o/. The cause of this vowel dependence is not yet clear. One explanation could be that the prediction error increases with increasing coupling between supra- and subglottal systems during the open glottis interval. This coupling is assumed to be great for open vowels, which have a fairly narrow pharynx, and small for closed vowels, which have a wide pharynx and consequently a bad impedance match at the glottis. This argument would predict a small prediction error for /i/-like vowels too, which we have not found in our work. The contention that the variance in the predictor error depends on the differences in formant frequencies and bandwidths between the open and closed glottis intervals does not seem to hold either. Therefore we tend to think that the most likely explanation is to be sought in the linear prediction analysis procedure itself. Makhoul and Wolf (1972) have proved that, given a spectral dynamic range D and total energy E , a two level spectrum, i.e., a spectrum with a rectangular envelope, will yield an absolute minimal prediction error. From all vowels /u/ and /o/ have spectra that approximate a rectangular envelope most closely.

Whatever the causes of the complexity of the error signal and its dependency on vowel type may be, the error signal is by no means an attractive vehicle for establishing the CGI. The seemingly obvious consequence, viz. that the closed glottis interval approach to inverse filtering is not meaningful, is not warranted, though. However complex the behaviour of the normalized error might be, in almost any case it shows a minimum in the neighbourhood of the point of glottal closure. More importantly, results of an extensive and detailed analysis of the spectra of speech signals during the CGI do not affect the assumption that the transfer function is all-pole. It is only when dealing with rounded back vowels, that yield a uniformly very low normalized prediction error, that one could doubt the advantages of estimating the formant parameters from the closed glottis interval. Here too, however, the transfer function as determined during the open glottis interval differs dramatically from the one determined from the CGI (Cranen & Boves, 1983). Consequently, the location of the analysis window from which the formant parameters are extracted is at least as important for rounded back vowels as it is for other vowel classes.

There is, however, yet another explanation for the problems encountered in dealing with /u, o/-like vowels, suggested by Fujimura (Fujimura & Lindqvist, 1971; Fujimura, personal communication), viz. that for the rounded back vowels the impedance of the yielding walls of the vocal tract makes a non-negligible contribution to the vocal tract transfer function. Moreover, this contribution is in the form of spectral zeros rather than added or altered resonances. This suggestion is worth of being pursued further. If it can be substantiated it would follow that the all-pole model is not accurate for rounded back vowels.

A1.3 Estimating the CGI from the EGG

A number of authors (Strube, 1974; El Malawany, 1977) have proposed slightly different approaches to the problem of automatically determining the CGI from the acoustic speech wave, but all seem to run into the same kind of problems, viz. multiple candidate locations in the analysis of many open vowels and virtual absence of candidate locations in /u, o/-like vowels. If the problem of reliable automatic determination of the CGI can be solved at all, it will probably be in the form of extremely time consuming iterative procedures. For our aims this increase in computational complexity cannot be justified. Moreover, we have access to a simple and extremely reliable indicator of the moments of glottal closing and also, though somewhat less accurate, the moments of glottal opening in the form of the EGG.

It should be noted that this decision is not fully compatible with the intention to use the EGG in order to evaluate the quality of an inverse filtering technique. It is not a valid procedure to use the same information twice, firstly to guide an analysis and next to formally evaluate the results of that analysis. We feel, however, that the decision to develop an inverse filtering technique which uses the EGG for determining the CGI is completely warranted. First of all the eventual aim of our effort is to be able to separate phonation and articulation in order to gain a better understanding of their contributions to perceptual judgement. Thus accurate inverse filtering is a tool rather than a goal in itself. Secondly, it remains possible to assess the accuracy of our two-signal approach to inverse filtering by comparing the reconstructed glottal wave with the pertinent physiological signals. Only it does no longer seem appropriate to do this comparison in some formal way, since the test signal and the most important criterion signal are no longer completely independent. An extensive visual comparison of the inverse filtering results with the physiological signals has shown that our approach does yield very accurate results for all types of vowels and that it can be successfully applied to vowels taken from running speech. Also it appeared to be capable of dealing with vowels from both stressed and unstressed syllables. Last but not least, the study of Krishnamurthi (1983), who compared inverse filtering results with glottal area data and simultaneous EGG recordings, has resulted in an inverse filtering technique which is similar in all respects to ours that is explained below. Krishnamurthi proces-

sed an extensive data base of glottal area data (derived from ultra-high speed films), EGG and inverse filtering results. He concluded that a closed glottis inverse filtering technique that determines the CGI from the EGG is far superior than any competing approach.

A1.4 The implementation of a closed glottis interval inverse filtering technique using the EGG

In this section we will give a detailed account of the inverse filtering technique developed to produce a large data base of vowels excerpted from running speech. Point of departure is a series of two disc files, one containing the sampled speech wave, the other containing the sampled EGG. The signals are, of course, supposed to be time aligned. Using an interactive waveform edit program (Buitting, 1981) the starting and end points of the vowels to be processed are determined and stored in a third disc file in the form of labels. The information in the label files is used to drive a linear phase filtering program that takes as input a labelled segment of a speech file and the corresponding segment of the EGG file and that produces as its output two smaller files containing the filtered versions of the segments. The FIR filter used is a 251-point high-pass filter designed by means of the program published by McClellan, Rabiner and Parks (1973); the cut-off frequency is at 40 Hz. High-pass filtering of the speech signals is necessary to remove any low frequency trend which otherwise would cause overload problems in the integrator following the inverse filter. High-pass filtering of the EGG is done to remove low frequency trends present in almost all EGG recordings, despite of the AC coupling applied during the recording process (refer to Fig. 4.6). Next the filtered speech wave is time reversed, passed through a phase correcting network and restored to its normal time scale (Hunt, 1978). The phase correction is necessary to undo the combined effects of the microphone pre-amplifier and the high-pass filter in the B&K measuring amplifier.

A fairly simple and straightforward algorithm is used to determine the sample numbers where a glottal closure occurs from the filtered EGG. These sample numbers are stored on disc. The inverse filter program proper begins by reading the information on the closure instants and then proceeds to reading the header of the filtered speech file. The header is supposed to contain information on the window length, on the number of predictor coefficients to be used and on the pre-emphasis factor. Then a portion of the speech wave is retrieved from disc, the rectangular analysis window is positioned in the closed glottis interval of the first pitch period to process and a covariance LP analysis is carried out. This analysis results in a set of filter coefficients, specifying the recursive production system with system function $V(z)$ in equation (4.9). The construction of the non-recursive inverse system is then straightforward and the complete pitch period is passed through the inverse filter, integrated (in order to undo the effects of the lip radiation component $L(z)$ in (4.9)), and eventually written to an output file. The analysis window is then positioned in the closed glottis interval of the next

pitch period, and the optimal inverse filter for that period is established and applied. Thus we have implemented a 'pitch synchronous' inverse filtering system, in which the coefficients of the inverse filter are updated at the beginning of each pitch period. The pitch synchronous approach is necessary for the processing of vowels taken from normal running speech, where the formant parameters are hardly ever stationary for a number of consecutive periods.

The window lengths that can be employed depend, of course, on the fundamental frequency of the speech and on the open ratio. Useful window lengths appeared to range from a low 20 samples for relatively high pitched voices combined with large open ratios to a high 45 samples for low pitched voices. In general, window length can be given a constant value for each speaker. Adapting the window length for each vowel in order to optimize the output of the inverse filter does not seem to be worth the time and inconvenience. For most speakers and most vowels predictor orders in the range between eight and twelve appeared to yield the best results. Here too it appeared that for each speaker a fixed value performed well in almost all analyses.

Initially, experiments were carried out with predictor orders that were slightly higher than the number needed to represent all formants. The idea behind these experiments was to try to use enough coefficients to model both the glottal volume flow shaping filter and the vocal tract filter in Fig. 4.3 and isolate the vocal tract part of the combined model by means of conventional formant extraction techniques (Flanagan, 1972: chapter 5; Markel & Gray, 1976: chapter 7). Formant extraction was attempted in several ways, but eventually we settled for a formal solution of the predictor polynomial for its zeros using a Newton-Raphson iteration (Boves & Cranen, 1979) followed by some logic and a bandwidth criterion. It appeared, however, that this procedure tended to result in occasional failures to detect all formants in some periods. Missing a formant causes the absence of the corresponding anti-formant network in the inverse filter, with the consequential undesirable effects on the reconstructed glottal flow. Therefore, it was decided to abandon this theoretically superior approach and to prefer, for purely practical reasons, an approach which - at least in theory - is suboptimal but which - in practice - performs quite well. In this approach the predictor order is set according to the number of formants that are expected to show up and the LP filter is inverted without making any attempt to establish whether all pairs of complex conjugate poles can indeed be considered as formants. The problems with occasionally missing formants in the theoretically optimal approach could, of course, have been handled by quite common techniques for smoothing raw formant tracks and for interpolating at points where formants seem to be missing. The differences between the output from the direct inversion of the LP filter and the output of the optimal method where it performed correctly was so insignificant, however, that the extra step of formant smoothing and interpolation was deemed to be not worth the trouble.

Pre-emphasis factors used were typically in the range between 90 and 98. The optimality of this choice was confirmed in a number of informal experiments in which the pre-emphasis factor was adapted from period to period on the basis of the ratio between the first and second term in the autocovariance function of the windowed signal. Optimal pre-emphasis values obtained in that way appeared to range from 80 to 99 but choosing 'non-optimal' values in the range from 90 to 98 invariably gave results that did not look worse than those obtained with 'optimal' values. Equally invariably, omitting pre-emphasis yielded results of a squarely unacceptable quality, most probably due to problems in obtaining a correct spectral estimate of the non pre-emphasized speech signal.

A1.5 Accuracy of formant estimates obtained from the CGI

At this point some question might arise as to the accuracy of the analysis results obtained with windows of extremely short duration. The routine advise to prospective users of LPA is to ensure that the analysis window encompasses at least three complete glottal periods. However, this advise applies to non-pitch synchronous analysis, i.e., a fully automatic LPA in which the window is shifted by a fixed number of samples regardless of the number of samples in a pitch period (Markel & Gray, 1976:156). In pitch synchronous analysis, on the other hand, the window duration is constantly adapted to the length of the pitch periods and so is the number of samples the window is shifted. Usually the window length is set equal to the number of samples in the pitch period and the leftmost point of the window is aligned with the leftmost (first) point of the period under analysis. That analysis windows with a duration of one pitch period or even appreciably less still can lead to highly accurate estimates of the (formant-)parameters of the vocal tract follows from the fact that LP analysis is not a 'blind' approach, on the contrary, it makes optimal use of knowledge about limitations of the speech production mechanism for making clever guesses with respect to the form of the signal outside of the window (Schroeder, 1982). Formally, LPA is based on the autocovariance function of the signal to be analysed. Now it can be shown that the autocovariance function of the output of an undriven all-pole system is completely determined by a number of samples equal to the number of poles, i.e., equal to twice the number of formants. All remaining autocovariance samples can be obtained by means of a simple recursive relation. From this it follows that a window length of $2m$ samples suffices to compute the autocovariance samples from which $m/2$ formants (frequencies and bandwidths) can be determined. What we get in practice is, of course, only an estimate of the autocovariance function.

From the definition of the (discrete) autocovariance function of the signal $s(m)$

$$r(n) = \sum_{m=0}^M s(m) s(m-n) \quad (\text{A1.1})$$

it is clear that it is an average-type measure. Therefore it is apparent that we should obtain more stable estimates if we take the average over a larger number of samples. The last statement, however, is only true if additional samples are drawn from the same population. In the case of the autocovariance function of a signal this means that the accuracy with which samples of the function can be estimated will only be increased by increasing M in (A1,1) as long as additional samples $s(m)$ are taken from a segment of the signal where its characteristics do not change.

On a microscopic level the signals making up voiced speech segments can be considered as highly non-stationary, as is evidenced by the fact that the goodness-of-fit of an LP model is much greater during the closed than during the open glottis intervals. Taking this into account it appears that it might be profitable to constrain the analysis window to very short lengths so that it can fit into a closed glottis interval, since the stationarity of the signal is only warranted within the boundaries of a CGI. Extension of the window beyond the length that can fit within a stationary interval might deteriorate rather than improve the accuracy of the analysis results.

The latter conclusion is not only rather surprising, it also seems to violate the uncertainty principle which states that the accuracy of spectral estimates increases with increasing time windows. However, the uncertainty principle is only seemingly violated, since it assumes stationary signal conditions. Furthermore, it is assumed in the customary formulation of the uncertainty principle that we have no a priori knowledge at all about the system which has produced the signal under analysis. Thus the explicit assumption that the system, the parameters of which we try to estimate is all-pole enables us to base the estimates on extremely short signal segments without having to sacrifice frequency resolution. It should be clear, of course, that these estimates are completely in error if it appears that the production system is fundamentally different from an all-pole system. But in that case our approach to the problem of inverse filtering is bound to break down completely anyhow.

References

- Ananthapadmanabha, T.V. & Fant, G. (1982) Calculation of true glottal flow and its components. *Speech Communication* 1: 167-184.
- Ananthapadmanabha, T.V. & Gauffin, J. (1982) Some results on the acoustic and aerodynamic factors in phonation. *STL-QPSR* 1/1982: 1-17.
- Asendorpf, J. & Wallbott, H.G. (1979) Masse der Beobachterubereinstimmung: Ein systematischer Vergleich. *Zeitschrift fur Sozialpsychologie* 10: 243-252.
- Askenfelt, A. & Sjolín, A. (1980) Voice analysis in depressed patients: Rate of change of fundamental frequency related to mental state. *STL-QPSR* 2-3/1980: 71-84.
- Askenfelt, A. & Hammarberg, B. (1981) Speech waveform perturbation analysis revisited. *STL-QPSR* 4/1981: 49-68.
- Baer, Th. (1979) Vocal jitter: A neuromuscular explanation. In Lawrence, V., Weinberg, B. (Eds.): *Transcripts of the eighth symposium 'Care of the professional voice' Part I: Physical factors in voice, vibrato, registers*. New York. The Voice Foundation.
- Bear, Th., Lofqvist A. & McGarr, N.S. (1983) Laryngeal vibrations. A comparison between high speed filming and glottographic techniques. *J. Acoust. Soc. Am.* 73: 1304-1308
- Berg, J.W. van den (1956) Direct and indirect determination of the mean subglottic pressure. *Folia Phoniatrica* 8: 1-24
- Berg, J.W. van den (1960) An electrical analogue of the trachea lungs and tissues. *Acta Physiol. Pharmacol. Neerlandica* 9: 361-385.
- Berg J.W. van den, Zantema, J.T. & Doornenbal, P. (1957) On the air resistance and the Bernoulli effect of the human larynx. *J. Acoust. Soc. Am.* 29: 626-631
- Bezoijen R. van (1984) *Characteristics and recognizability of vocal expressions of emotion*. Dissertation Nijmegen.
- Bismarck G. von (1974) Timbre of steady sounds: A factorial investigation of its verbal attributes. *Acustica* 30: 146-159.
- Blom, J.G. (1980) Some notes on the reliability of scale values. In: Fagel, W.P.F. *Literatuuronderzoek uitspraakbeoordeling*. Report Nr. 61. Inst. Phon. Sc. Univ. of Amsterdam, Amsterdam.
- Blom, J.G. & Koopmans-van Beinum, F.J. (1973) An investigation concerning the judgement criteria for the pronunciation of Dutch. *Proceedings Inst. Phon. Sc. Univ. of Amsterdam* 3: 1-24.
- Blom, J.G. & van Herpt, L.W.A. (1976-a) The evaluation of jury judgments on pronunciation quality. *Proceedings Inst. Phon. Sc. University of Amsterdam* 4: 31-47.
- Blom, J.G. & van Herpt, L.W.A. (1976-b) A number of computer programs for scaling. *Proceedings Inst. Phonetic Sciences Univ. Amsterdam* 4: 81-112.
- Bock, R.D. & Jones, L.V. (1968) *The measurement and prediction of judgment and choice*. Holden Day, San Francisco.
- Boves, L. (1979) A set-up for the comparison of glottal inverse filtering with physiological registrations. *Proc. Inst. Phon. Nijmegen* 3: 103-123.

- Boves, L. (1981) On measuring and modelling subglottal pressure signals. *Proc. Inst. Phon. Nijmegen* 5: 41-51.
- Boves, L. (1983-a) Subglottal pressure as a means to control fundamental frequency and SPL. *Proc. Inst. Phon. Nijmegen* 7: 9-18.
- Boves, L. (1983-b) Frequency analysis of broad-band subglottal pressure signals. *Proc. Inst. Phon. Nijmegen* 7: 19-30.
- Boves, L. & Cranen, B. (1979) Die Zuverlässigkeit der Rekonstruktion der Schallschnellefunktion an der Glottis in Abhängigkeit der Formantextraktion. *Proc. 4 DFG Kolloquium Sprachverarbeitung*, Darmstadt: 105-108.
- Boves, L. & Cranen, B. (1982-a) The use of the Law of Categorical Judgement in phonetic research. *Proc. Inst. Phon. Nijmegen* 6: 9-19.
- Boves, L. & Cranen, B. (1982-b) On long-term average spectra and their representation. *Proc. Inst. Phon. Nijmegen* 6: 20-30.
- Boves, L. & Cranen, B. (1982-c) Comparison of reconstructed flow waveforms with physiological registrations. *Fortschritte der Akustik, FASE/DAGA '82*, Göttingen: 979-982.
- Boves, L. & Cranen, B. (1982-d) Evaluation of glottal inverse filtering by means of physiological registrations. *Conf. Records ICASSP '82*: 1988-1991.
- Boves, L. & Fagel, W. (1982) Relations between semantic differential ratings of running speech and acoustic characteristics of the speech signals. *Fortschritte der Akustik, FASE/DAGA '82* Göttingen: 1095-1098.
- Boves, L., Fagel, W. & Herpt, L. van (1982) Opvattingen van vrouwen en mannen over de spraak van mannen en vrouwen. *De Nieuwe Taalgids* 75: 1-23.
- Broad, D. (1979) The new theories of vocal cord vibration. In Lass, N.J. (Ed.): *Speech and language. Advances in basic research and practice*, Vol. 2, Academic Press, New York. 203-257.
- Brown, B.L. & Lambert W.E. (1976) A cross-cultural study of social status markers in speech. *Canadian J. Behav. Scs.* 8: 39-55.
- Brown, B.L., Strong, W.J. & Rencher, A.C. (1974) Fifty-four voices from two: The effect of simultaneous manipulation of rate, mean fundamental frequency and variance of fundamental frequency on ratings of personality from speech. *J. Acoust. Soc. Am.* 55: 313-318.
- Brown, B.L., Strong, W.J. & Rencher, A.C. (1975) Acoustic determinants of perception of personality from speech. *Int. Journal Sociology of Language* 6: 11-22.
- Buiting, H.J.A.G. (1981) MOSES: a simple speech editing facility. *Proc. Inst. Phon. Nijmegen* 5: 58-63.
- Butcher, A. (1981) Phonetic correlates of perceived tempo in reading and spontaneous speech. *Work in Progress, Phon. Lab. Univ. Reading* 3: 105-117.
- Cohen, J. & Cohen, P. (1975) *Applied multiple regression/correlation analysis for the behavioral sciences*. Lawrence Erlbaum Ass., Hillsdale, N.J.
- Cooley, W.W. & Lohnes, P.R. (1971) *Multivariate data analysis*. New York, Wiley & Sons.
- Cranen, B. (1983) Preliminary results on the simultaneous registration of seven laryngeal parameters. *Proc. Inst. Phon. Nijmegen* 7: 31-45.
- Cranen, B. & Boves, L. (1983) The effects of the glottal termination impedance on the formants of speech signals. In Schussler, H.W. (Ed.) *Signal Processing II: Theories and Applications*. Elsevier Science Publishers, Eurasp. 343-346.
- Cranen, B., Boves, L. & Bezooijen R. van (1983) Acoustic correlates of auditory ratings of some voice quality settings. In: Winkler, P. (Ed.) *Investigations of the speech process*. Bochum, Studienverlag Dr. N. Brockmeyer. 264-288 (= *Quantitative Linguistics* 19).
- Crystal, D. (1975) *The English tone of voice. Essays in intonation, prosody and paralanguage*. Edward Arnold, London.
- Dudley, H. (1940) The carrier nature of speech. *Bell Syst. Tech. J.* 19: 495-515.

- (Reprinted in Flanagan, J L & Rabiner, L R (1973) *Speech Synthesis* Dowden Hutchinson & Ross, Stroudsbury Pa)
- Edmonds, T D , Lilly, D J & Hardy, J C (1971) Dynamic characteristics of air-pressure measuring systems used in speech research *J Acoust Soc Am* **50** 1051-1057
- El Malawany, I I (1977) Area function extraction over the closed glottis interval In Carré, R , Descout, R and Wajskop, M (Eds) *Articulatory modeling and phonetics* G A L F 65-76
- Elenius, K (1980) Long time average spectrum using a 1/3 octave filterbank *STL-QPSR* **4**/1980 14-22
- Fagel, W Herpt, L van & Boves, L (1983) Analysis of the perceptual qualities of Dutch speakers' voice and pronunciation *Speech Communication* **2** 315-326
- Fant, G (1960) *Acoustic theory of speech production* Mouton, The Hague
- Fant G (1979) Glottal source and excitation analysis *STL-QPSR* **1**/1979 85-107
- Finkelstein, S (1983) Customized LPC vocabulary lets cars talk *Speech Technology* **2** 65-69
- Flanagan, J L (1958) Some properties of the glottal sound source *J Speech Hearing Res* **1** 99-116
- Flanagan, J L (1972) *Speech analysis, synthesis and perception* Springer, Berlin
- Flanagan, J L , Rabiner, L R , Christopher, D , Bock, D F & Shipp, T (1976) Digital analysis of laryngeal control in speech production *J Acoust Soc Am* **60** 446-456
- Fourcin A J (1974) Laryngographic examination of vocal fold vibration In Wyke B (Ed), *Ventilatory and phonatory control systems* Oxford Univ Press, London 315-326
- Fritzell, B , Hammarberg, B & Wedin, L (1977) Clinical applications of acoustic voice analysis Part I Background and perceptual factors *STL-QPSR* **2-3**/1977 39-43
- Fujimura, O & Lindqvist, J (1971) Sweep-tone measurements of vocal-tract characteristics *J Acoust Soc Am* **49** 541-558
- Gauffin, J , Sundberg, J (1980) Data on the glottal voice source behavior in vowel production *STL-QPSR* **2-3**/1980 61-70
- Gilbert, L A , Deutsch, C J & Strahan, R F (1978) Feminine and masculine dimensions of the typical, desirable and ideal woman and male *Sex Roles* **4** 767-779
- Giles, H & Powesland, P F (1975) *Speech style and social evaluation* Academic Press, London
- Grosjean, F , Lane, H (1981) Temporal variables in the perception and production of spoken and sign languages In Eimas, P D , Miller, J L (Eds) *Perspectives on the study of speech* Lawrence Erlbaum Ass , Hillsdale, N J 207-237
- Guilford, J P (1954) *Psychometric methods* Tatra - McGraw Hill, Bombay-New Delhi
- Hammarberg, B , Fritzell B , Gauffin J , Sundberg, J & Wedin, L (1980) Perceptual and acoustic correlates of abnormal voice quality *Acta Otolaryng* **90** 441-452
- Hart, J 't & Collier, R (1975) Integrating different levels of intonation analysis *J of Phonetics* **3** 235-295
- Hollien, H & Majewski, W (1977) Speaker identification by long-term spectra under normal and distorted speech conditions *J Acoust Soc Am* **62** 975-980
- Holmes, J (1976) Formant excitation before and after glottal closure *Conf Records ICASSP-'76* 39-42
- Hunt, M J (1978) Automatic correction of low-frequency phase distortion in analogue magnetic recordings *Acoustic Letters* **32** 6-10
- Ishizaka, K , French, J L & Flanagan, J L (1976) Direct determination of vocal tract wall impedance *IEEE Trans Acoust , Speech Signal Processing*, ASSP-**23** 370-373
- Ishizaka, K & Matsudaira, M (1972) Fluid mechanical considerations of vocal cord vibration *SCRL Monograph No 8*, SCRL, Santa Barbara

- Ishizaka, K., Matsudaira, M. & Kaneko, R. (1976) Input acoustic impedance measurement of the subglottal system. *J Acoust Soc. Am.* **60**: 190-197.
- Jury, E.I. (1973) *Theory and application of the z-transform method*. Robert Krieger Publ. Co., Huntington, N.Y.
- Katwijk, A. van (1974) *Accentuation in Dutch*. Van Gorcum, Assen.
- Kitzing, P. (1977) Methode zur kombinierten photo- und elektroglossographischen Registrierung von Stimmlippenschwingungen *Folia Phoniat.* **29**: 249-260
- Kitzing, P. & Lofqvist A. (1975) Subglottal and oral air pressure during phonation - Preliminary investigation using a miniature transducer system. *Medical and Biological Engineering* **13**: 644-648.
- Koike, Y. (1981) Sub- and supraglottal pressure variation during phonation. In Stevens, K.N. & Hirano, M. (Eds.), *Vocal Fold Physiology*, Univ. of Tokyo Press, Tokyo: 181-192.
- Koike, Y. & Hirano, M. (1973) Glottal area time function and subglottal pressure variation. *J Acoust. Soc. Am.* **54**: 1618-1627.
- Koopmans-van Beinum, F.J. (1980) *Vowel contrast reduction An acoustic and perceptual study of Dutch vowels in various speech conditions*, Academische Pers B.V., Amsterdam
- Kosiel, U. (1973) Statistical analysis of speaker-dependent differences in the long-term average spectrum of Polish speech. in: W. Jassem (Ed.), *Speech analysis and synthesis, Vol. 3*. Polish Scientific Publ., Warsaw: 181-208
- Kramer, Chr. (1977) Perceptions of female and male speech. *Language and Speech* **20**: 151-161.
- Krishnamurthi, A.K. (1983) *Study of vocal fold vibration and the glottal sound source using synchronized speech, electroglottography and ultra high speed laryngeal films*. Doct Diss., Univ. of Florida, Gainesville.
- Krishnamurthi, A.K. & Childers, D.G. (1981) Vocal fold vibratory patterns: Comparison of film and inverse filtering *Conf. Records ICASSP-81*: 133-136.
- Labov, W. (Ed.) (1972) *Sociolinguistic patterns*. Univ. of Pennsylvania Press, Philadelphia PA.
- Ladefoged, P. (1973) The features of the larynx. *J Phonetics* **1**: 73-83.
- Ladefoged, P. & Maddieson, I. (1983) Techniques for the analysis of phonation types in natural languages. *J. Acoust Soc. Am* **73**: S88 (Abstr.).
- Laine, U K. (1982) Modelling of lip radiation impedance in the z-domain. *Conf. Records ICASSP-82*: 1992-1995.
- Laver, J. (1980) *The phonetic description of voice quality* Cambridge Univ. Press, London
- Lecluse, F.L F. (1977) *Elektroglossografie*. Diss. Rotterdam.
- Li, K P., Hughes, G W & House, A S. (1969) Correlation characteristics and dimensionality of speech spectra. *J Acoust. Soc Am* **46**: 1019-1025.
- Lindqvist, J. (1965) Studies of the voice source by means of inverse filtering. *STL-QPSR* **2/1965**: 8-13
- Makhoul, J.I. & Wolf, J.J. (1972) *Linear prediction and the spectral analysis of speech*. Technical Report No. 2304. Bolt, Beranek and Newman Inc., Cambridge, Mass.
- Maling, G., Morrey, W., Lang W. (1967) Digital determination of third-octave and full-octave spectra of acoustic noise. *IEEE Trans. Audio Electroacoust.* **AU-15**: 98-104.
- Markel, J.D. & Gray Jr, A.H. (1976) *Linear prediction of speech*. Springer, Berlin.
- McClellan, J.H., Parks, T.W. & Rabiner, L.R. (1973) A computer program for designing optimum FIR linear phase digital filters. *IEEE Trans Audio Electroacoust.* **AU-21** 506-526.
- Messick, S.J (1955) Metric properties of the semantic differential. *Educ. and Psych. Measurement* **17**: 200-206. Also in Snyder, J G and Osgood, Ch. E. (Eds): *Semantic differential technique. A sourcebook*. Aldine Publ. Co., Chicago: 161-167

- Messick, S.J. (1960) Dimensions of social desirability. *Journ. of Consulting Psychol.* 24: 279-287.
- Millar, H.D. & Baker, L.E. (1973) A stable ultraminiature catheter-tip pressure transducer. *Medical & Biological Engineering* 11: 86-89.
- Miller, R.L. (1959) Nature of the vocal cord wave. *J. Acoust. Soc. Am.* 31: 667-677.
- Mol, H. (1970) *Fundamentals of phonetics. Part II - Acoustical models generating the formants of the vowel phonemes*. Mouton, The Hague
- Monsen, R.B. & Engebretson, A.M. (1977) Study of variation in the male and female glottal wave. *J. Acoust. Soc. Am.* 62: 981-993.
- Nishisato, S. (1980) *Analysis of categorical data: Dual scaling and its applications*. Univ. of Toronto Press, Toronto.
- Nuijten, E.Th.G. (1962) *De tweetalige mens. Een taalsociologisch onderzoek naar het gebruik van dialect en cultuurtaal in Borne*. Van Gorcum, Assen.
- Oppenheim, A.V. & Schaffer, R.W. (1973) *Digital signal processing*. Prentice Hall, Englewood Cliffs, N.J.
- Osgood, C.E., May, W.H. & Miron, M.S. (1975) *Cross-cultural universals of affective meaning*. Univ. of Illinois Press, Urbana.
- Osgood, C.E., Suci, G.J. & Tannenbaum, P.H. (1957) *The measurement of meaning*. University of Illinois Press, Urbana.
- Overall, J.E. & Klett, C.J. (1972) *Applied multivariate analysis*. McGraw Hill, New York.
- Pastore, R.E. (1981) Possible psychoacoustic factors in speech. In: Eimas, P.D. & Miller J.L. (Eds.) *Perspectives on the study of speech*. Lawrence Erlbaum Ass., Hillsdale, N.J.: 165-205.
- Pedersen, M.F. (1977) Electroglottography compared with stroboscopy in normal persons. *Folia phoniat.* 29: 191-199.
- Perkins, W.H. & Koike Y. (1969) Patterns of subglottal pressure variations during phonation. *Folia Phoniat.* 21: 1-8.
- Plomp, R. (1976) *Aspects of tone sensation A psychophysical study*. Academic Press, London.
- Pols, L.C.W. (1977) *Spectral analysis and identification of Dutch vowels in monosyllabic words*. Academische Pers b.v., Amsterdam.
- Rabiner, L.R. & Schaffer, R.W. (1978) *Digital processing of speech signals*. Prentice-Hall, Englewood Cliffs N.J.
- Rietveld, A.C.M. (1983) *Syllaben, klemtonen en de automatische detectie van beklemtoonde syllaben in het Nederlands*. Dissertation Nijmegen.
- Rosenberg, A. (1971) Effect of glottal pulse shape on the quality of natural vowels. *J. Acoust. Soc. Am.* 49: 583-590.
- Rosenthal R. (1973) Estimating effective reliabilities in studies that employ judges' ratings. *J. Clinical Psychology* 29: 342-345.
- Rossum, N. van (1981) Parallel power density measurement system. *Proc. Inst. Phon. Nijmegen* 5: 142-159
- Rossum, N. van & Boves L. (1978) An analog pitch-period extractor. *Proc. Inst. Phon. Nijmegen* 2: 1-17.
- Rothenberg, M. (1981) Acoustic interaction between the glottal source and the vocal tract. In Stevens, K., Hirano, M. *Vocal Fold Physiology*, Univ. of Tokyo Press, Tokyo: 305-328.
- Rothenberg, M. (1982) An interactive model for the voice source. *STL-QPSR* 1/1982: 1-17.
- Scherer, K.R. (1978) Personality inference from voice quality: The loud voice of extraversion. *European J. Social Psych.* 8: 467-487.
- Scherer, K.R. (1979) Personality markers in speech. In: Scherer, K.R. & Giles, H., *Social markers in speech*, Cambridge Univ. Press, Cambridge: 147-209.
- Scherer, K.R. (Hrsg.) (1982) *Vokale Kommunikation*. Beltz Verlag, Weinheim und Basel

- Scherer, K.R., Ladd, D.R. & Silverman, K.E.A., Vocal cues to speaker affect: testing two models. Submitted for publication in *J. Acoust. Soc. Am.*
- Scherer, R.C. & Titze, I.R. (1981) Pressure-flow relationships in a model of the laryngeal airway with a diverging glottis. Paper presented at the *2nd Int. Conf. Physiology and Biophysics of Voice*, Madison, Wisconsin.
- Schroeder, M.R. (1982) Linear prediction, extremal entropy and prior information in speech signal analysis and synthesis. *Speech Communication* 1: 1-20.
- Schutte, H.K. (1980) *The efficiency of voice production*. Dissertation, Groningen.
- Smith, P.M. (1979) Sex markers in speech. In: Scherer, K.R. & Giles, H. (Eds.), *Social markers in speech*, Cambridge Univ. Press, London: 109-138
- Steiglitz, K. (1974) *An introduction to discrete systems*. J. Wiley, New York.
- Strube, H.W. (1974) Determination of the instant of glottal closure from the speech wave. *J. Acoust. Soc. Am.* 56: 1625-1629
- Sundberg, J., Gauffin, J. (1979) Waveform and spectrum of the glottal voice source. In: Lindblom, B., Öhman, S. (Eds.) *Frontiers of speech communication research*, Academic Press, London.
- Takahashi, H. & Koike, Y. (1975) Some perceptual dimensions and acoustical correlates of pathological voices. *Acta Oto-Laryngologica Suppl.* 338: 1-24.
- Teager, H.M. & Teager, S.M. (1984) Active fluid dynamic voice production models. To appear in: Scherer, R. & Titze, I. (Eds.) *Proceedings of the Conf. on Physiology and Biophysics of Voice*, Univ. of Iowa City Press, Iowa City.
- Titze, I.R. & Talkin, D.T. (1979) A theoretical study of the effects of various laryngeal configurations on the acoustics of vibration. *J. Acoust. Soc. Am.* 66: 60-75.
- Torgerson W.S. (1958) *Theory and methods of scaling*. J. Wiley, New York
- Ungeheuer, G. (1962) *Elemente einer akustischen Theorie der Vokalartikulation*. Springer, Berlin.
- Uldall, E. (1960) Attitudinal meanings conveyed by intonational contours. *Language and Speech* 3: 223-234.
- Vieregge, W.H. (1969) *Untersuchungen zur akustischen Struktur der Plosivlaute*. Dissertation Bonn.
- Voiers, W. (1964) Perceptual bases of speaker identity. *J. Acoust. Soc. Am.* 36: 1065-1073.
- Wakita, H. & Fant, G. (1978) Towards a better vocal tract model. *STL-QPSR* 1/1978: 9-29
- Wichern, P. (1978) On the design of efficient median smoothers. *Proc. Inst. Phon. Nijmegen* 2: 18-28.
- Wong, D.Y., Markel, J.D. & Gray Jr. A.M. (1979) Least squares glottal inverse filtering from the acoustic speech wave form. *IEEE Trans. Acoust., Speech, Signal Processing* ASSP-27: 350-355.
- Zahorian, S.A. & Rothenberg M. (1981) Principle-components analysis and low-redundancy encoding of speech spectra. *J. Acoust. Soc. Am.* 69: 832-845

Samenvatting

Dit proefschrift handelt over de samenhang tussen fysisch (akoestisch) beschrijfbare kenmerken van spraaksignalen en subjectieve oordelen over de spraak en over de spreker die gegeven worden door ongeoefende luisteraars. Door het hele werk ligt de nadruk eerder op het ontwikkelen en op kleine schaal beproeven van methoden en technieken van onderzoek dan op een poging om een omvattend inhoudelijk antwoord op de opgeworpen vraag te geven.

Het eerste hoofdstuk van het boek geeft een korte inleiding tot het probleem en probeert er een kader voor te scheppen. Dat kader moet op de eerste plaats gezocht worden in onderzoek van het effect van mondelinge communicatie, meer specifiek het onderzoek naar de invloed van vormkenmerken van een boodschap op de wijze waarop die boodschap bij de ontvangers overkomt. Betoogd wordt dat de voortgang van dat onderzoek tot dusver weinig bevredigend geweest is en dat dit voor een belangrijk deel te wijten is aan methodologische tekortkomingen. Daarmee is meteen het motief voor de sterke methodologische oriëntatie van het onderhavige werk gegeven.

Het tweede hoofdstuk behandelt de constructie van een schaleninstrument waarmee statistisch verwerkbare subjectieve (perceptieve) oordelen over spraaksamples verkregen kunnen worden. Na een korte beschrijving van de voornaamste eisen waaraan beoordelingsschalen moeten voldoen, wordt uitvoeriger ingegaan op de wijze waarop de schalen gekozen zijn en op de manier waarop nagegaan is of de gekozen schalen aan een aantal onafhankelijk toetsbare criteria voldoen. Het ontwikkelde beoordelingsinstrument heeft de vorm van een semantische differentiaal, d.w.z. een verzameling van 7-vaksschalen, waarvan de polen gemarkeerd zijn door aan elkaar tegengestelde adjectieven. Het voorlopige instrument waarvan getoetst is of de schalen voldeden aan de eisen van lineariteit en relevantie ten aanzien van de te beoordelen stimuli bestond uit 35 schalen. Een aantal daarvan was bewust gekozen voor het representeren van de evaluatie, potency en activiteitsdimensies, de drie 'standaard'-dimensies die in zeer veel onderzoek met semantische differentialen naar voren blijken te komen. Daarnaast waren grotere aantallen schalen opgenomen die betrekking hadden op verschillende aspecten van uitspraak, stemgebruik en para- en sociolinguïstisch relevante spraak- en sprekerkenmerken.

De schalen, die uiterst zorgvuldig gekozen waren op basis van een uitvoerig literatuuronderzoek, zijn op hun bruikbaarheid getoetst in een papier-en-potlood experiment waarin grote aantallen proefpersonen het instrument gescoord hebben voor drie concepten, nl. 'de ideale stem en uitspraak van een man', 'de ideale stem en uitspraak van een vrouw' en 'de eigen stem en uitspraak'. Het idee achter dit experiment was dat schalen alleen voldoen aan de eis dat één pool negatief en de andere positief is als de beide scores voor ideale stem en uitspraak in de buurt van een van beide schaaluiteinden liggen. Een grote spreiding in de scores voor ideale stem zou er bovendien op kunnen wijzen dat er tussen de proefpersonen onenigheid bestaat over de interpretatie van de betreffende schaal, een uitermate ongewenste situatie. Een zeer kleine spreiding van de scores voor eigen stem daarentegen geeft aanleiding om te veronderstellen dat het schaalattribuut niet op spraak van toepassing is dan wel dat normale sprekers met betrekking tot dat attribuut nauwelijks variatie vertonen.

Het is gebleken dat vrijwel alle schalen in voldoende mate aan de gestelde eisen tegemoet kwamen. Desondanks is om onafhankelijke redenen besloten alsnog een klein aantal wijzigingen in het instrument aan te brengen, overigens zonder het totale aantal van 35 schalen te veranderen.

Een nadere analyse van de scores bracht aan het licht dat stereotiepe denkbelden over sexeverschillen nauwkeurig gereproduceerd worden: de ideale stem en uitspraak van vrouwen is vriendelijker en expressiever dan die van mannen maar minder krachtig en actief. Verder bleken de verschillen in de scores van de mannelijke proefpersonen groter dan in de scores van de vrouwelijke deelnemers aan het experiment (allen universiteitsstudenten).

Hoofdstuk drie beschrijft de experimenten ter verkrijging van perceptieve oordelen over twee verzamelingen van spraaksamples. In alle gevallen ging het om voorgelezen teksten. De eerste verzameling omvatte elf sprekers, zes mannen en vijf vrouwen, die allen dezelfde neutrale tekst voorlezen. Deze verzameling wordt aangeduid als 'het Amsterdamse materiaal' omdat de erop gebaseerde beoordelingsexperimenten deel uitmaakten van een nauw verwant onderzoek aan het Instituut voor Fonetische Wetenschappen van de Universiteit van Amsterdam. De andere verzameling, aangeduid als 'het Nijmeegse materiaal', bestond uit zes mannelijke sprekers die allen twee teksten voorlezen, nl. de neutrale tekst die in het Amsterdamse materiaal gebruikt was en een tekst waarvan de inhoud duidelijk politiek gekleurd was. Het Amsterdamse materiaal is beoordeeld door 235 luisteraars, het Nijmeegse door 117.

De scores zijn allereerst uitvoerig geanalyseerd om na te gaan of alle schalen zich zodanig gedragen hadden dat een verwerking van de uitkomsten met behulp van factoranalyses verantwoord was. Voor de analyse van de metrische eigenschappen van de schalen is voornamelijk gebruik gemaakt van de 'Wet van het Categorisch Oordeel' zoals oorspronkelijk geformuleerd door Thurstone. Er wordt tamelijk uitvoerig ingegaan op twee op die wet

gebaseerde technieken, een voor de beoordeling van de dimensionaliteit van de schalen, de ander voor de bepaling van de betrouwbaarheid van de scores, die in een eerder stadium voorgesteld waren maar nog niet eerder formeel beschreven. Gebleken is dat strikt genomen geen van de schalen beschouwd mag worden als een schaal met zeven even brede intervallen: de intervalbreedte neemt toe naarmate men dichterbij een van de schaaluiteinden komt. Omdat echter alle schalen zich in dit opzicht wezenlijk identiek lijken te gedragen, mogen we ervan uitgaan dat de correlaties tussen de schalen lineair zijn. Alle schalen blijken betrouwbaar gescoord te kunnen worden: bij een aantal van 30 beoordelaars ligt de effectieve betrouwbaarheid boven de .90; de meeste schalen halen dit niveau al met een aanzienlijk kleiner aantal beoordelaars. Verder blijkt de betrouwbaarheid meer een functie te zijn van de variatie in het stimulusmateriaal dan van de formele eigenschappen van de schaal. Een soortgelijke conclusie kan getrokken worden voor wat de dimensionaliteit van de schalen betreft.

Hoewel er op een klein aantal schalen, voornamelijk verband houdend met stem en uitspraak, statistisch significante verschillen bleken te bestaan tussen de scores voor de twee teksten in het Nijmeegse materiaal, was het niet mogelijk om een systematisch of inzichtelijk verband te vinden tussen de inhoud van de teksten en de oordelen van de luisteraars. De indruk ontstaat veeleer dat de scoreverschillen voornamelijk te wijten zijn aan toevallige verschillen in voorleesprestatie van de sprekers.

Factoranalyses van de scores leidden tot beoordelingsruimten die voor de Amsterdamse en Nijmeegse stimulusverzamelingen wezenlijk overeenkomstig waren. De in termen van verklaarde variantie verreweg belangrijkste factor kon benoemd worden als 'Stemdynamiek/Expressiviteit'. Op de tweede plaats komt de factor 'Uitspraakkwiteit'. De volgende factoren verklaren nog maar een klein percentage verklaarde variantie, hun benoemingen zijn 'Sociolinguïstische oordelen' (alleen in de Nijmeegse data), 'Potency', 'Stemkleur' en 'Tempo'. Het ontbreken van de evaluatiefactor in alle factoroplossingen kan verklaard worden door erop te wijzen dat een diffuse algemene evaluatie geconcretiseerd en daardoor opgesplitst is in termen van meer specifieke evaluatieve oordelen over stemgebruik, uitspraak en voordracht. De factoranalyses hebben geleid tot een gereduceerd beoordelingsinstrument, bestaande uit 14 schalen die alle relevante aspecten bestrijken, dat in de rest van dit onderzoek en in vervolgonderzoek gehanteerd zal worden.

Hoofdstuk vier behandelt een poging om een akoestische beschrijving van het aspect 'Stem' te verkrijgen, onafhankelijk van het aspect 'Uitspraak'. Deze poging is gebaseerd op invers filtering van het spraaksignaal, een techniek waarmee het mogelijk zou moeten zijn om het oorspronkelijke stemgeluid zoals dat bij de stembanden gegenereerd wordt uit het afgestraalde spraakgeluid te reconstrueren. Het hoofdstuk begint met een grotere intuïtieve beschrijving van de basis onder die techniek, nl. het

lineaire bron-filter model van akoestische spraakproductie. Nochtans blijkt uit die beschrijving dat er aan de invers filter techniek een aantal problemen kleven die een onafhankelijke toetsing van haar adequaatheid noodzakelijk maken.

Teneinde een dergelijke toetsing te kunnen uitvoeren zijn er simultane registraties gemaakt van het electroglottogram (EGG), fotoglottogram (PGG), subglottale druk (P_{sg}) en uiteraard het akoestisch spraaksignaal. Het was daarbij de bedoeling om uit de P_{sg} het moment van openen van de stembanden af te leiden, uit het PGG het moment van maximale glottisopening en uit het EGG het moment van sluiten van de glottis, in elke stembandperiode. Allereerst wordt uitvoerig ingegaan op de wijze van registreren van P_{sg} , met behulp van een miniatruurdrupknemer, die via de neusholte en de achterste commissuur van de glottis in de luchtpijp gebracht wordt, en op de verwerking en interpretatie van de registraties. Aangetoond wordt dat de synchroon met de stembandbeweging lopende variaties in P_{sg} niet primair te maken hebben met het Bernoulli effect, maar veeleer een gevolg zijn van excitatie van de eigenmodes van het subglottale systeem. Deze bevinding tast de basis aan waarop de poging om uit P_{sg} het moment van openen van de glottis te bepalen gebaseerd was.

De PGG registraties bleken technisch grote problemen met zich mee te brengen, die voornamelijk verband hielden met de positionering van de foto-opnemer boven de stembanden. Recent gepubliceerde theoretische overwegingen tonen echter aan dat er geen eenvoudig verband bestaat tussen het moment van maximale glottisoppervlakte (het criterium dat uit het PGG afgeleid zou moeten worden) en het moment waarop de luchtstroom door de glottis zijn maximum bereikt (en dat rechtstreeks af te leiden is uit het te toetsen invers filter resultaat). Daarmee blijft het EGG over als het enig betrouwbare signaal dat bruikbaar is voor een toetsing van de resultaten van een invers filter techniek. Daarbij moet aangetekend worden dat er een groeiende hoeveelheid gegevens ontstaat waaruit blijkt dat naast het moment van sluiten van de glottis ook het moment van openen nauwkeurig uit het EGG afgeleid kan worden.

In de onderhavige studie is afgezien van een poging om de resultaten van een puur akoestische invers filter techniek formeel te toetsen tegen de openen sluitmomenten afgeleid uit het EGG. Daarentegen is de voorkeur gegeven aan de ontwikkeling van een invers filter techniek waarin het EGG gebruikt wordt om het gesloten interval in elke stembandperiode te bepalen, waarna uit het akoestische signaal gedurende het gesloten interval het optimale inverse filter bepaald wordt. De techniek en haar merites worden meer uitvoerig beschreven in Appendix 1.

De 'closed glottis interval inverse filtering technique' is toegepast om een beschrijving te verkrijgen van het stemgebruik van vijf sprekers in het Nijmeegse materiaal. Daartoe zijn van vijf sprekers elk 60 klinkers uit de neutrale leestekst en van twee sprekers eveneens 60 klinkers uit de niet-neutrale tekst geanalyseerd teneinde het bronsignaal te reconstrueren. Uit alle

teksten zijn ongeveer evenveel beklemtoonde als onbeklemtoonde klinkers gekozen. Voor zover de teksten dit toelieten zijn de klinkers zo gekozen dat alle klinkerfonemen vertegenwoordigd waren. Voor de verdere verwerking van de uitkomsten is van alle stembronsignalen het kritische bandspectrum gemiddeld over de hele klinker bepaald. Het is gebleken dat er voor elke spreker een niet onaanzienlijke spreiding in de spectra van het brongeluid bestaat, zodanig dat een eenvoudige techniek voor 'sprekerherkenning' slechts tot matige resultaten leidt. Tussen twee teksten is de variatie echter zeker niet groter dan binnen een tekst. Wel lijken de sprekers te verschillen voor wat betreft de manier waarop zij in hun stemgebruik verschil maken tussen beklemtoonde en onbeklemtoonde lettergrepen: met name in de slechtere stemmen lijkt er een trend te bestaan in de richting van een rijker harmonisch spectrum in de beklemtoonde klinkers. Naarmate de spreker beter getraind is lijkt deze 'effort'-component een kleinere rol te spelen. Tenslotte is uit de verwerking van de analyseresultaten gebleken dat er inderdaad sterke aanwijzingen bestaan in de richting van een klinkerafhankelijke mate van interactie tussen de stembron en de mond-keelholte.

Nadat in hoofdstuk vier een poging ondernomen was om een beschrijving te verkrijgen van specifiek met de stembron gerelateerde kenmerken van het spraakgeluid, beschrijft hoofdstuk vijf een aantal technieken voor het bepalen van globale akoestische kenmerken van spraaksignalen. Bij het ontwikkelen en toetsen van alle te gebruiken maten is speciale aandacht besteed aan de statistische stabiliteit ervan. Waar mogelijk is nagegaan of de variatie binnen sprekers significant kleiner was dan de variatie tussen de sprekers. Achtereenvolgens komen aan de orde tempo, uitgedrukt in lettergrepen per seconde en percentage pauzetijd tussen zinnen, toonhoogte en toonhoogtevariëatie, long-term average spectra en representaties gebaseerd op de resultaten van formantanalyses. Bij de beschrijving van de toonhoogtegegevens wordt vooral ingegaan op pogingen om maten te vinden voor de korte, midellange en lange termijn fluctuaties van de toonhoogte. Met betrekking tot de long-term average spectra (LTAS) wordt allereerst kort ingegaan op de ter beschikking staande technieken om LTAS te bepalen. Uitvoeriger wordt ingegaan op methoden voor een zodanige verwerking van LTAS dat deze zich lenen voor verdere statistische verwerking. Gekozen is voor kritische bandspectra, waarbij alleen gemiddeld is over de stemhebbende delen van het spraaksignaal, gerepresenteerd in de vorm van spectrale afstanden tussen de energieniveaus in een referentieband (400-600 Hz) en de banden bij 1.6 kHz en 5 kHz en de band die de F_0 -modus bevat.

Om na te gaan of LTAS berekend over de stemhebbende intervallen overeenkomt met het gemiddelde spectrum van het stembronsignaal zijn de resultaten voor zeven teksten grafisch vergeleken. De overeenkomsten bleken minder goed te zijn dan verwacht. Voor een belangrijk deel laten de discrepanties zich echter verklaren met een verwijzing naar mogelijke verschillen tussen de sprekers voor wat betreft de frequentiekaracteristiek van de afstraling aan de lippen.

De eerste 60 s van alle leesteksten zijn onderworpen aan een formantanalyse die voor alle als stemhebbend geklassificeerde intervallen voor elke 10 ms een schatting van de frequentie en bandbreedte van de laagste vier formanten opleverde. Nagegaan is of de gemiddelden van alle parameters, bij middeling over toenemende tijdsintervallen, naar een stabiele waarde convergeerden. Gebleken is dat middeling over meer dan 30 s tot wezenlijk stabiele, voor de spreker karakteristieke, waarden leidt. De gemiddelden voor de formantfrequenties zijn verwerkt in een maat voor het contrast in het akoestisch klinkersysteem van elke spreker door ze te vergelijken met de formantfrequenties van een hypothetische referentiespreker.

In het zesde hoofdstuk wordt geprobeerd om akoestische maten en perceptieve oordelen aan elkaar te relateren. De perceptieve oordelen zijn hierbij de criteriumwaarden die met behulp van de akoestische maten benaderd, voorspeld of, als men wil, verklaard moeten worden. Het leggen van relaties gebeurt in statistische termen: door middel van correlatie en multiple regressie. De relatief kleine aantallen stimuli maakten het noodzakelijk om een strategie te kiezen die de kans minimaliseert dat toevallige verbanden aangezien zouden worden voor wezenlijke relaties. Gekozen is voor een strategie die verbanden alleen dan als zinvol accepteert als ze op overeenkomstige wijze aanwezig zijn in zowel het Amsterdamse als het Nijmeegse materiaal.

Voor alle factoren is geprobeerd de factorscores voor de sprekers en de schaalwaarden van de sprekers op de voor de factor in kwestie meest relevante schalen met behulp van akoestische maten te verklaren. Correlaties tussen akoestische maten en perceptieve oordelen (in de vorm van factorscores en schaalwaarden) waren in het algemeen tamelijk tot zeer laag, met de 'Stemkwaliteit/Stemkleur'-factor als belangrijkste uitzondering. Voor dit resultaat wordt een aantal, elkaar meestal aanvullende, verklaringen voorgesteld. Het beeld zou in ieder geval positiever geweest zijn als niet de eis gesteld was dat een op zich plausibele relatie in beide materiaalverzamelingen moest terugkeren. Met name het Nijmeegse materiaal bevat een aantal relaties tussen akoestische maten en perceptieve oordelen die op zich zeer plausibel zijn, maar die in het Amsterdamse materiaal niet gereproduceerd worden. Verder wordt gewezen op de mogelijkheid dat de correlaties tussen veel akoestische maten en ermee gerelateerde perceptieve oordelen niet lineair hoeven te zijn. Een zeer kleine toonhoogtevariatie bijv. wordt waarschijnlijk evenzo negatief gewaardeerd als een zeer grote; daartussen in ligt een - mogelijk klein - bereik waarin mate van toonhoogtevariatie gelijk op gaat met waardering. Verder zijn de gebruikte globale akoestische maten niet bij machte om te verdisconteren voor de linguïstisch bepaalde oordelen: met name de gebruikte maten voor de variatie in allerlei parameters kunnen geen onderscheid maken tussen wat linguïstisch noodzakelijk is, wat wel of niet in overeenstemming is met de betekenis van het gezegde en wat bijdraagt aan de expressiviteit van de voordracht. Tenslotte moet ook rekening gehouden worden met de mogelijkheid dat voor veel akoestische parameters en/of

perceptieve oordelen de spreiding in de materiaalverzamelingen te klein was om duidelijke resultaten te behalen.

Het beeld van de resultaten bij de correlatie analyse wordt door de uitkomsten van de multiple regressie analyse bevestigd. Het voornaamste resultaat van deze exercitie is het inzicht dat enkelvoudige interacties tussen akoestische maten nauwelijks een bijdrage leveren aan de verklaring van de perceptieve oordelen.

In een algemene discussie wordt nog een keer ingegaan op de mogelijke gevolgen van de gekozen strategie. Met name wordt overwogen dat de vergelijking van een groep sprekers bestaande uit vrouwen en mannen met een groep alleen uit mannen bestaande voor problemen gezorgd kan hebben. Het opereren van verschillende normen voor mannen en vrouwen en de mogelijkheid dat om fysiologische redenen sommige akoestische maten voor vrouwen anders geïnterpreteerd moeten worden dan voor mannen worden aangevoerd ter verklaring van het feit dat in vrijwel alle gevallen de relatie perceptie-akoestiek in het Nijmeegse materiaal sterker was dan in het Amsterdamse. Verder wordt erop gewezen dat het onderzoek opgezet was als een correlationele studie. Dit heeft tot gevolg dat alleen die relaties gevonden worden die min of meer geprononceerd in de materiaalverzamelingen aanwezig zijn. Gegeven de omvang van de verzamelingen is het niet onwaarschijnlijk dat sommige reële relaties toevallig ontbreken. Ook wordt nog een keer het belang onderstreept van het feit dat de beoordelaars te maken hadden met talige, betekenisvolle stimuli. Daarom moet rekening worden gehouden met de mogelijkheid dat de oordelen niet rechtstreeks als reactie op de akoestische eigenschappen van de stimuli gegeven zijn, maar veeleer indirect, terugredenerend vanuit een linguïstisch gestuurde waarneming van de stimuli.

Het zevende en laatste hoofdstuk bestaat uit een poging om de stand van zaken op de in het voorafgaande bestreken deelgebieden nog eens samen te vatten en om de belangrijkste te verwachten of noodzakelijke ontwikkelingen voor de naaste toekomst te beschrijven.

Met betrekking tot de akoestische beschrijving van lopende spraak wordt gesteld dat er twee ontwikkelingen noodzakelijk zijn: op de eerste plaats moeten datarepresentatietechnieken ontwikkeld worden die het mogelijk maken om de manier waarop parameters in de tijd variëren compact weer te geven, zodanig dat verdere statistische verwerking kan plaats vinden. Daarbij moet het tevens mogelijk zijn om een aantal parameters simultaan te verwerken, rekening houdend met het feit dat veel parameters onderlinge afhankelijkheden vertonen. Op de tweede plaats moeten de mogelijkheden verruimd worden om bij de verwerking van tijdsvariante akoestische meetresultaten rekening te houden met de linguïsticiteit van het geanalyseerde geluid.

Ten aanzien van de perceptieve beschrijving van lopende spraak wordt gesteld dat de semantische differentiaalscores van ongetrainde luisteraars

waarschijnlijk te ver afstaan van de akoestische kenmerken van de spraaksignalen om een eenduidige relatie tussen akoestiek en perceptieve oordelen te kunnen opleveren. Er wordt derhalve geconstateerd dat het noodzakelijk is om verder onderzoek te doen dat moet leiden tot technieken die meer 'objectieve' perceptieve oordelen over lopende spraak opleveren.

Curriculum vitae

L. Boves werd geboren te Maastricht op 21 april 1946. Hoewel hij zijn jeugd niet in Maastricht doorbracht en dus het dialect van die plaats niet beheerste, slaagde hij er toch in om in 1964 het gymnasium β diploma te behalen aan het Henric van Veldeke College (te Maastricht dus). Na een afgebroken studie Wis- en Natuurkunde te Utrecht behaalde hij in 1969 het kandidaats- en in 1973 het doctoraaldiploma Nederlandse taal- en letterkunde aan de Katholieke Universiteit Nijmegen. Bij het uitvoeren van zijn afstudeeronderzoek verkreeg hij onontbeerlijke steun van het Instituut voor Fonetische Wetenschappen van de Universiteit van Amsterdam.

Sinds juli 1973 is L. Boves als wetenschappelijk medewerker verbonden aan het Instituut voor Fonetiek van de K.U.N. Zijn wetenschappelijke belangstelling is vooral gericht op onderzoek van stemvoering en op mogelijkheden om digitale signaalverwerkingstechnieken toe te passen voor linguïstisch relevant spraakonderzoek.

- 1 De veel gehoorde bewering dat het onmogelijk is om het onderzoek van vakgenoten op zijn inhoudelijke kwaliteit te beoordelen lijkt niet zelden te berusten op angst voor de veroordeling van het eigen werk
- 2 Nu het vertrouwen in de kwaliteit en de productiviteit van de onderzoekers het doorslaggevende criterium lijkt te gaan worden bij de beoordeling van programma's voor de voorwaardelijke financiering van onderzoek, is er geen enkele reden om te vrezen dat werkelijk hoogwaardig en vruchtbaar kleinschalig onderzoek bij voorbaat kansloos is. Hetzelfde criterium maakt wel beginnende (groepen van) wetenschappers bij voorbaat vrijwel kansloos, waardoor het gemakkelijk aanleiding kan zijn tot een verstarring van het wetenschapsbedrijf
- 3 Nu vrijwel elk drie- of vierjarig project, dat vaak door de begeleider(s) tot in detail geformuleerd en gepland is, moet leiden tot de promotie van de uitvoerder, doet zich de vraag voor of er in het Nederlandse academische systeem niet de behoefte ontstaat aan een pendant van het Duitse Habilitationsschrift als bewijs van de bekwaamheid tot het werkelijk zelfstandig opzetten en uitvoeren van wetenschappelijk onderzoek en van geschiktheid voor het bekleden van vooraanstaande academische posities
- 4 De bewering dat een verbreding en veralgemenisering van de academische eerste fase opleiding (die onvermijdelijk ten koste gaat van de diepgang) de kansen van de afgestudeerden op de arbeidsmarkt vergroot wordt afdoende weerlegd door de klachten van werkgevers uit alle sectoren van de maatschappij over de beperkte inzetbaarheid van abiturienten van beroepsopleidingen met een sterk algemeen vormend karakter.
- 5 Indien de beroepsopleidingen op het terrein van de taal- en spraakpathologie in hun curricula geen aandacht gaan besteden aan het omgaan met signaalverwerkende hulpmiddelen voor diagnose en therapie, zal binnen enkele jaren de situatie ontstaan dat ofwel de kwaliteit van de hulpverlening onnodig laag is, ofwel de huidige beroepsbeoefenaren in toenemende mate verdrongen zullen worden door klinisch werkende fysici of electronici

6. Het feit dat de taalwetenschap voornamelijk beoefend en gedoceerd wordt in faculteiten waarin een gedegen mathematische en methodologische scholing niet strikt noodzakelijk geacht wordt, betekent een ernstige belemmering voor haar ontwikkeling.
7. In een land waarin zelfs massacommunicatie-organisaties van een uitgesproken rechtse signatuur het beginsel van 'hoor en wederhoor' belijden, is het onbegrijpelijk dat bij de meeste openbare plechtigheden de persoon die aanleiding tot de viering vormt niet in de gelegenheid gesteld wordt om te reageren op datgene wat door feestredenaars over haar/zijn persoon en werk gedebiteerd wordt.