Speech processing strategies for the profoundly hearing impaired

Nic J.D.M.M. van Son

Speech processing strategies for the profoundly hearing impaired

Nic J.D.M.M. van Son

The work presented in this thesis was performed at the Laboratory of Experimental Audiology, Dept. of Otorhinolaryngology of the Academic Hospital of the Universiteit Utrecht, as part of a research programme on the "Rehabilitation of the Deaf and Hearing Impaired".

This research was financially supported by SMA, an Universiteit Utrecht Incentives Fund for Research in Social Areas for Special Attention.

Omslagfoto: Trudy Wiendels
Druk: OMI *Offset*, Utrecht

CIP-GEGEVENS KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Son, Nicolaas Jacobus Derk-Jan Martinus Maria van

Speech processing strategies for the profoundly hearing impaired / Nicolaas Jacobus Derk-Jan Martinus Maria van Son. - Utrecht : Universiteit Utrecht, Faculteit Geneeskunde
Proefschrift Universiteit Utrecht. - Met lit. opg. - Met samenvatting in het Nederlands.
ISBN 90-393-0138-7
Trefw.: audiologie.

Speech processing strategies for the profoundly hearing impaired

Strategieën voor de bewerking van spraaksignalen
ten behoeve van ernstig slechthorenden
(met een samenvatting in het Nederlands)

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit Utrecht
op gezag van de Rector Magnificus Prof. Dr. J.A. van Ginkel
ingevolge het besluit van het College van Dekanen
in het openbaar te verdedigen
op dinsdag 5 oktober 1993 des namiddags te 16.15 uur

door

Nicolaas Jacobus Derk-Jan Martinus Maria van Son

geboren op 30 november 1959 te Helmond

Promotor: Prof.dr. G.F. Smoorenburg
Verbonden aan de Faculteit der Geneeskunde van de Universiteit Utrecht

Het in dit proefschrift beschreven onderzoek werd uitgevoerd aan het Laboratorium voor Experimentele Audiologie van de afdeling Keel-, Neus- en Oorheelkunde van het Academisch Ziekenhuis te Utrecht, als onderdeel van het onderzoekprogramma "Revalidatie van Doven en Slechthorenden".

Het onderzoek werd gefinancierd door het Stimuleringsfonds Maatschappelijke Aandachtsgebieden (SMA) in het kader van het aandachtsgebied "Emancipatie van Gehandicapten".

Dankwoord

Gedurende de afgelopen vier jaren heb ik van verschillende mensen in mijn omgeving grote steun gehad, al dan niet van wetenschappelijke aard. *Bedáánk állemoal* zou daarvoor voldoende dankbetuiging moeten zijn, ware het niet dat ik (a) normaal wat breedsprakiger ben, zodat ik me bij alleen die twee woorden wat ongemakkelijk voel, en (b) de Utrechtse tongval niet beheers, waar ik niet rouwig om ben maar wat me extra duidelijk maakt dat die zelfde twee woorden niet echt van mij zijn. Derhalve een wat uitgebreider dankwoord.

Ten eerste gaat mijn dank uit naar mijn collega Arjan Bosman voor zijn steun op vele terreinen. Zijn technische en wetenschappelijke kwaliteiten hebben mij voor al te grote ontsporingen behoed, een stuk of wat opgeblazen hoofdtelefoontjes daargelaten. Zonder zijn eenzaam gezwoeg om een goedwerkende omgeving voor spraakanalyse en -bewerking te creëren zouden vele ideeën voor het experimentele werk binnen het onderzoekprojekt "Revalidatie van doven en slechthorenden" nooit verwezenlijkt zijn. Mijn password weet ik nog, maar een fles wijn lijkt me niettemin op zijn plaats.

Voor een gedegen wetenschappelijke ondersteuning ben ik veel dank verschuldigd aan Guido Smoorenburg en Pieter Lamoré. Hun rijkdom aan ideeën en kritische beoordeling van teksten hebben mij veel tijd gekost maar minstens zoveel genoegen opgeleverd. Verder ben ik Bert Schouten van het Utrechtse Onderzoekinstituut voor Taal en Spraak (OTS) zeer erkentelijk voor veel tijd en energie die hij heeft gestoken in het meedenken over experimentele opzetten van de eerste proeven en het bediscussiëren van hun resultaten.

Wetenschappelijk onderzoek gedijt alleen in een goede werkomgeving waarin het verantwoorde denkwerk af en toe eens flink verstoord wordt door onverantwoorde prietpraat en koffieleut. Voor een dergelijke sfeer hebben mijn collega's en diverse stagiaires door de jaren heen garant gestaan. Hun geldt voornamelijk de hierboven genoemde uit Frits zijn hart gegrepen dankbetuiging: Frits Meeuwsen, John de Groot, Arjan Bosman, Jurgen Lentz, Paul Kolston, Guido Smoorenburg, Wilma Ruizendaal, Ferry Hendriksen, Hans Mülder, Hans van Dijk, Tirtsa Huiskamp, Marius Rodenburg, Jacques Berk, Jorien Schreuder, Ippei Takagi, Takehiro Hanada, Isabella Terpstra, Stan Turkawski, Piet Lamoré, Yvonne van Holsteijn, Carlien Theewis, Maarten van Emst, Sjaak Klis, Winanda Popken, Marieke van Hal, Matthijs Killian, Margreet Langereis en Lydeke Fransen. Met name het immer aanwezige enthousiasme van Tirtsa en de warme belangstelling en steun van Margreet in de laatste fase van het vele schrijfwerk hebben mij veel goed gedaan; hun vriendschap is het belangrijkste resultaat van de afgelopen vier jaar.

Tenslotte gaat mijn grote dank uit naar de vaste groep proefpersonen die door de jaren heen steeds bereid bleken om allerlei vervelende en langdurige experimenten te doorstaan. Voor hen was het onderzoek bedoeld en slechts door hun welwillende medewerking hebben we de in dit proefschrift genoemde resultaten kunnen bereiken. Aan hen draag ik dan ook dit proefschrift op, daarbij de wens uitsprekend dat de ontwikkelingen rondom de toepassing van spraakbewerking voor ernstig slechthorenden (al dan niet in hoortoestellen) in de toekomst vruchten zal afwerpen.

Als allerlaatste wil ik Trudy bedanken. Jouw enige zichtbare bijdrage aan dit proefschrift is de omslagfoto. Een voor velen onzichtbare en onschatbare bijdrage is geweest dat jij er gewoon altijd was. Ik hoop dat je er nog een lange tijd gewoon zult zijn.

Nic van Son

ring the bell that still can ring
forget your perfect offering
there is a crack in everything
that's how the light gets in

Leonard Cohen, *Anthem (1992)*

Contents

List of phonetic symbols

phonetic symbol (IPA[1])	example words	symbol used in this thesis
Dutch consonants		
p	paal	p
t	tak	t
k	kat	k
b	bok	b
d	dak	d
f	fout	f
s	soep	s
ʃ	sjaal	ʃ
x	goed	x
v	voet	v
z	zeep	z
m	maan	m
n	nat	n
ŋ	lang	ŋ
l	loop	l
r, [R][2]	ruit, waar	r, [R]
ʋ, [w]	water, ruw	ʋ, [w]
j	jas	j
h	huis	h
Additional English consonants		
g	good	g
θ	three	θ
ð	this	δ
ʃ	show	ʃ
ʒ	garage	*ʒ*
tʃ	cheap	tʃ
d	July	d*ʒ*

[1] The phonetic symbols in this list were taken from the *international phonetic alphabet*, published by the "International Phonetic Association" (IPA).

[2] Symbols in square brackets indicate allophonic variants normally used in syllable-final position.

List of phonetic symbols (continued)

phonetic symbol (IPA)	example words	symbol used in this thesis
Dutch vowels		
i	bier	i
I	vis	I
e	geel	e
ɛ	bes	ɛ
a	maan	a
ɑ	kan	ɑ
ɔ	pot	ɔ
o	rood	o
u	soep	u
y	vuur	y
ø	leuk	ø
œ	mug	œ
ɛi	wijn	ɛi
œy	huis	œy
au	goud	au
ə	lopen	ə
Additional English vowels		
æ	bad	æ
ʌ	but	ʌ

List of abbreviations and acronyms

A	auditory
ALSCAL	a specific MDS technique
AM	amplitude modulation/modulated
ANOVA	analysis of variance
AV	auditory-visual
AXB, BXA, *etc.*	discrimination paradigms
C	consonant
CDT	connected discourse tracking
CL95, CL99	95% and 99% confidence limits
CV, CVC, VCV, *etc.*	syllable types, composed of consonants and vowels in the order indicated
Δf	frequency difference between two pure tones
DL_F	difference limen for frequency
EGG	electroglottograph
f_0	fundamental frequency
F_1, F_2	first formant, second formant
FM	frequency modulation/modulated
HC	hierarchical clustering
HI	hearing impaired
HL	hearing level
INDSCAL	a specific MDS technique (individual scaling)
ITA	information transmission analysis
LDL	loudness discomfort level
LFPTA	low-frequency PTA [.25, .5, 1 kHz]
LPC	linear predictive coding
LPF	low-pass filtering/filtered
MANOVA	multivariate ANOVA
MCL	most comfortable loudness
MDS	multidimensional scaling
NH	normally hearing
PTA	pure-tone average [.5, 1, 2 kHz]
RMS	root mean square
sd	standard deviation
SiVo	"Sinusoidal Voicing", a signal processing hearing aid presenting f_0 and amplitude information
S/N ratio	signal-to-noise ratio
SPL	sound pressure level
THR	pure-tone threshold
UCL	uncomfortable loudness level
V	visual (referring to lipreading experiments); vowel (referring to phonemes)

Chapter 1.

General introduction

1.1 Origin and objectives of this study

1.1.1 Speech processing for the profoundly hearing impaired

Many profoundly hearing impaired people have such limited auditory capabilities that their perception of speech is severely restricted to the extent that it is not possible without lipreading. In general, they wear hearing aids in order to hear warning sounds and make acoustic contact with their environment, but most aids are of little use when it comes to speech perception. Even the more advanced types of hearing aids, utilizing various digital processing techniques, have been of little avail to the (speech) perception of the profoundly hearing impaired because they do not improve discrimination among speech sounds (Boothroyd, 1990).

Profound hearing impairment may be defined by an average hearing loss of 90 dB HL or more at the audiometric frequencies of .5, 1 and 2 kHz (pure-tone average or PTA). It is normally associated with higher thresholds at high frequencies or even a reduction of the audible frequency span to about 2 kHz, a strongly reduced dynamic range, often covering not more than 10 dB, and limited frequency resolution; temporal resolution is generally less deteriorated (Lamoré *et al.*, 1985; Faulkner *et al.*, 1990a, 1990b; Rosen *et al.*, 1990; Bosman and Smoorenburg, 1993b).

Because of the limited residual auditory capacities, the profoundly hearing impaired listener is hardly able to extract useful information from the complex speech signal. Several processing techniques have been endeavoured in order to simplify the speech signal and thus emphasize the main characteristics. Techniques like frequency lowering (frequency band compression), high-frequency emphasis, peak clipping, and fast amplitude compression (syllabic compression) generally provide little success for the (profoundly) hearing impaired (Braida *et al.*, 1979; Boothroyd, 1990). The frequency processing techniques do not have an answer to the problem of reduced frequency resolution, while high-frequency emphasis has no effect when the frequency span is strongly reduced. As to compression of the speech envelope amplitude, fast compression rates may have a devastating effect on the natural

amplitude variation over time, which is often the only speech cue that the impaired person can perceive reasonably well (*e.g.* Plomp, 1988).

A better solution to the loss of discrimination of speech sounds may be found in advanced processing schemes making use of our knowledge of speech, such as speech cue enhancement (Montgomery and Edge, 1988; Simpson *et al.*, 1990) or speech feature extraction (Breeuwer, 1985; Breeuwer and Plomp, 1986; Boothroyd, 1990). Some of these techniques have been used for a considerable number of years now in rehabilitation programmes for the deaf who can be helped with electrical stimulation of the cochlear nerve through cochlear implants or inner ear prostheses (Tyler *et al.*, 1988), or with vibro-tactile stimulation (Pickett and McFarland, 1985). In a number of implant systems, the acoustic waveform is processed on the basis of our knowledge of speech. Relevant speech information is extracted from the acoustic signal and recoded in a form that can be transmitted to the nerve. With the relative success of speech processing algorithms and information transmission in cochlear implants, the interest in rehabilitation of profoundly hearing impaired persons has been renewed. Successful rehabilitation may be achieved on a similar basis as rehabilitation of the deaf with a cochlear implant: by extracting and coding essential speech information. In this case, the speech information will be transmitted by an acoustic device rather than an inner ear prosthesis.

In recent years the development of these so-called signal processing or speech processing hearing aids has received a fair amount of interest among researchers (Rosen *et al.*, 1987; Grant, 1987; Fourcin, 1990; Boothroyd, 1990; Smoorenburg, 1990a; Smoorenburg and Bosman, 1993). The main idea is to present to the profoundly hearing impaired a highly simplified speech signal, only containing the most necessary information, in such a way that it is properly matched with the limited auditory capacities. This implies that it should only contain information that can actually be perceived, removing all other, possibly confusing or masking, information. In the development and application of signal processing hearing aids, four areas of interest may be distinguished:

(1) selection of the most relevant speech features to be presented to the profoundly hearing impaired;
(2) the use of robust feature extraction algorithms;
(3) encoding of a new acoustic signal on the basis of the extracted information, optimally matched with the auditory capacities of the profoundly hearing impaired;
(4) structured rehabilitative training, in which the patient has to learn and interpret the new speech code (and other non-speech signals).

As the profoundly hearing impaired have no perception of sound whatever without using a hearing aid, there is some freedom to create new optimal relations between physical and perceptual dimensions in profoundly impaired hearing (Smoorenburg, 1990b). To give an example: if high-frequency noise is the type of information that is deemed important (for identification of plosive and fricative consonants), it may be transposed to a region within the audible frequency span. In that way, it becomes an audible cue which the listener only has to learn to interpret as a meaningful speech sound. In such a way, a new speech code may be defined which is entirely different from the existing one, but which is more adequately matched to the auditory system.

1.1.2 Objectives of this study

The main objective of the study described in this thesis was to investigate the information processing capacities of the profoundly hearing impaired for speech or speech-like patterns, both auditory and visual (lipreading). Closely related to that is the second objective, the definition of a well-matched speech signal aimed at improved speech perception when compared to an unprocessed speech signal. The study was thus mainly concerned with questions within areas (1) and (3): what are the relevant features in speech for profoundly hearing impaired people, and in what way should these features be presented to them, regarding the limited auditory capacities that they have available ?

The questions have been studied at a basic phonetic level, that of the phoneme, so as not to involve higher levels of linguistic processing. In this way, perception by the profoundly hearing impaired could be evaluated without interference from confounding levels of ability that do not tell how basic acoustic and articulatory patterns are actually processed. Most of the study has been concerned with vowels or vowel-like acoustic signals; vowels carry an important part of the information for the profoundly hearing impaired, especially when auditory-visual speech perception is considered.

The characteristic approach in all of the experiments described was the study of psychophysical relationships, both auditory and visual. This involves the study of how acoustic or articulatory distinctions between (speech) sound patterns are reflected in perceptual distinctions, and the interpretation of the underlying perceptual dimensions that may account for the results. Furthermore, it involves the investigation of basic auditory capacities in order to assess the relative contributions of hearing and lipreading in speech perception.

1.2 Topics investigated in this study

1.2.1 Visual perception: lipreading of consonants and vowels

For most hearing impaired people, lipreading is an indispensable source of understanding speech. In the case of profound hearing impairment, people may need to rely on it to a great extent. In our opinion, therefore, the presentation of relevant speech features should be studied in light of the phonetic information that can be accessed through lipreading. When considering processed speech sounds as a supplement to lipreading for the profoundly hearing impaired, it is a valid approach to offer only or mainly that information acoustically that cannot be extracted from the visible speech signal.

Considering the earlier mentioned interest in psychophysical relationships, the following issues were addressed in the experiment described in chapter 2:

(1) To what extent are different consonants and vowels, characterized by distinct articulatory movements, actually perceived differently ? It may be assumed that certain phonemes cannot be distinguished visually because they have identical or similar visible articulations. Groups of visually indistinguishable phonemes are normally referred to as *visemes*, a term coined by Fisher (1968) from *visual phonemes*.

(2) What perceptual dimensions are underlying consonant and vowel viseme classifications ? Not just the actual viseme groups are of interest, but also the perceptual information that defines their similarities or distinctions. Subsequently, we are interested in the acoustic information related to these perceptual dimensions; in other words, what acoustic speech information emanates from articulatory movements that are plainly visible ?

(3) Do viseme classifications differ with better lipreading abilities ? This is an interesting point with regard to individual strategies of signal processing: when a good lipreader uses other perceptual dimensions than the average lipreader, he may profit from different speech features presented auditorily.

As will be discussed in chapter 2, the classifications and underlying perceptual dimensions give indications as to what information is and is not available through lipreading, and may consequently be presented auditorily in a recoded speech signal.

1.2.2 Auditory perception: similarity of harmonic complexes and identification of synthetic vowels

We would like to define the information processing capacity of the profoundly impaired ear in terms of suprathreshold perception of speech or speech-like signals, rather than in terms of basic auditory functions. In such a way we may be able to describe the ability to extract feature information from the multidimensional speech signal. In general, profoundly hearing impaired listeners are reported to make use of mainly temporal and fairly low-frequency information in speech (Erber, 1972a; Boothroyd, 1984; Faulkner *et al.*, 1990b; Shannon *et al.*, 1992). These cues may be used in perceiving suprasegmental speech structures, and certain aspects of consonants, such as voicing and manner of articulation. Until recently, most realizations of processed speech signals for the profoundly hearing impaired have only included these types of information (see chapter 4). However, a number of studies have shown that also relatively high-frequency spectral information may be processed, if there is some residual frequency selectivity (Lamoré *et al.*, 1985; Rosen *et al.*, 1987, 1990; Faulkner *et al.*, 1990b). This ability may therefore be exploited in those profoundly hearing impaired with residual frequency selectivity.

With respect to speech perception, we are especially interested in the ear's capacity to simultaneously process information from different frequency channels. This is related to such questions as whether profoundly hearing impaired listeners are able to cope with two-formant representations of vowels, or whether they are able to distinguish different (low-frequency) noise spectra. In chapter 3 a study is described in which profoundly hearing impaired subjects were asked to rate the similarity of a number of related harmonic complexes, differing as to the distribution of spectral components over the spectrum. The perceptual dimensions underlying the subjects' responses provide information about the acoustic cues that were used in judging pattern similarity. Following this abstract task, an identification experiment was run with synthetic two-formant vowels. Different vowel sets were constructed on the basis of the results from the harmonic complexes study.

1.2.3 Perceptual dimensions and the potentials of processed speech

The results from the studies in chapters 2 and 3 may be described in terms of perceptual dimensions used in lipreading and hearing, and their relation to the acoustic features of speech signals. On the basis of these

results, coded speech signals may be defined that contain the selected information to be presented to the lipreader in a form that is well matched to the profoundly impaired auditory system. In order to place our experimental work within the framework of other research carried out in these areas, chapter 4 provides a review of lipreading research and research with processed acoustic supplements to the profoundly hearing impaired.

Speech perception naturally involves processing of information by eye and by ear. In case one of the modalities cannot be used or is impaired to a certain degree, information from the speech signal is processed altogether or mainly through the other modality. In consonant perception, the information processed through lipreading is mainly based on visible places of articulation, whereas auditory perception mainly gives information on voicing and manner of articulation, certainly in cases of profound hearing impairment. This complementary relationship in consonant perception has often been stressed (Summerfield, 1983). Generally speaking, lipreading provides cues based on high-frequency acoustic information, while audition mainly serves for processing low-frequency acoustic information. The complementary relationship does not show in vowel perception. Auditory cues for vowel recognition are based on acoustic formant information, and so are visual cues: the lipreader bases identification on visible degrees of lip rounding and vertical lip opening, related to F_2 and F_1, respectively. Temporal information in the acoustic signal is also (partly) accessible both through vision and audition. In running speech, finally, prosodic information, which is important for the linguistic organization of speech, is hardly visible but can well be processed auditorily as it is cued mainly in temporal and low-frequency information.

Research into the potentials of encoded speech signals is rather limited. Most of the signals only carry fundamental frequency (f_0) and amplitude information, thus making use of the complementary relationship: these two cues are not accessible through lipreading. By and large, the results fall short of the expectation that extracted speech features would be more beneficial to the profoundly hearing impaired lipreader than an unprocessed amplified speech signal. In the discussion section of chapter 4 it is argued that f_0 and amplitude cues may provide too little information for most patients, and that this minimal amount may well be extended to include some spectral speech cues.

In chapter 5 a vowel discrimination experiment is described, in which we have made use of the results from earlier studies (chapters 2 and 3) in order to define a well-matched supplementary speech signal for the profoundly hearing impaired lipreader. Discrimination is therefore mainly

(but not exclusively) studied with respect to auditory-visual perception.

The group of profoundly hearing impaired subjects is a very heterogeneous one, both as to their auditory capacities and with regard to their lipreading abilities. In the interpretation of our results we will have to take this phenomenon into account; the potentials of feature extraction and coded-speech presentation cannot exclusively be discussed for the group as a whole. It was clear from the experiment in chapter 5 that *the group* of profoundly hearing impaired does not benefit from the speech signal we designed, but that *some individuals* actually did. This was also found in a study running parallel to our vowel discrimination experiment, in which supplemented sentence lipreading ability was measured using a set of everyday sentences in various unprocessed and processed speech conditions (Bosman and Smoorenburg, 1993a, 1993b). This individual performance may call for the individual selection of speech features and design of the acoustic supplement instead of making one signal do for the whole group of profoundly hearing impaired. The topic of individual performance on the part of our subjects is touched upon in various sections of the chapters 2, 3 and especially 5, where discrimination performance for individual subjects is related to their audiometric data (threshold and dynamic range).

Finally, the conclusions and implications of this study of speech processing for the profoundly hearing impaired are given in chapter 6. Topics of interest for the near future are discussed.

Chapter 2.

Viseme classifications of Dutch consonants and vowels

This chapter is a slightly adapted version of a paper by N. van Son, T.M.I. Huiskamp, A.J. Bosman and G.F. Smoorenburg, provisionally accepted by the Journal of the acoustical society of America (1993c)

Abstract

Videotaped lists of meaningless Dutch syllables were presented in quiet to four subject groups, differing with respect to their knowledge of and experience with lipreading (lipreading expertise). Syllables consisted of all Dutch consonants within three vowel contexts, and of all Dutch vowels within four consonant contexts. Three speakers pronounced all syllable lists. The aim of the research was (1) to establish viseme classifications of Dutch consonants and vowels; (2) to interpret the visual-perceptual dimensions underlying these classifications and relate them to acoustic-phonetic parameters; (3) to establish the effect of lipreading expertise on the classification of visually similar phonemes (visemes). In general, viseme classification proved consistent with different subject groups: lipreading expertise is not related to viseme recognition. Important visual features in consonant lipreading are *lip articulation*, *degree of oral cavity opening* and *place of articulation*, leading to the following viseme classification: /p,b,m/, /f,v,ʋ/, /s,z,ʃ/, and /t,d,n,j,l,k,x,r,ŋ,h/. In the acoustic domain, these features may be related to spectral differences. Important vowel features in lipreading are *lip rounding*, *vertical degree of lip opening* and *vowel duration*, yielding the following visemes: /i,I,e,ɛ,ɛi,a,ɑ/, /u,y,œ,ɔ/, /ø,o/, and /au,œy/. In the acoustic domain, lip rounding may roughly be related to the second formant, lip opening to the first formant.

2.1 Introduction

2.1.1 Visemes

For most hearing impaired people, lipreading is an important means of understanding speech. However, various phonemes have identical or highly similar visible articulation movements and as such are visually non-contrastive or indiscriminable. This indicates that the baseline for purely visual speech perception is quite at a disadvantage as compared to the situation in auditory perception. A set of visually indiscriminable

phonemes is often referred to as a *viseme*[1]. Normally, the term is reserved for visual consonant groupings (Fisher, 1968; Walden *et al.*, 1977, 1981; Owens and Blazek, 1985), but it can be applied analogously to vowel groupings. Briefly, a viseme may be defined as a set of phonemes which are visually confused to a high degree. The degree of confusion should exceed a criterion set to the researcher's own discretion. It may have a statistical basis. Changing the variables in a visual perception task may lead to different numbers and compositions of visemes. Empirical viseme classifications depend on factors like speakers, stimuli, subjects, and statistical criteria. Nevertheless, some visemes consistently emerge in almost all studies concerned, and as such viseme categorization may offer a general description of lipreading performance at the phoneme level.

In general, visemes are taken as the basic information that a lipreader should have at his disposal in order to be able to communicate through lipreading. The more visemes can be recognized, the better the baseline for lipreading ability is. Most lipreading training courses will give a short account of the various articulatory movements and positions associated with (groups of) speech sounds. In rehabilitation of the deaf and profoundly hearing impaired by means of speech processing techniques in cochlear implants or signal processing hearing aids, the knowledge of what speech features are visually accessible may be used for determining which acoustic information is to be presented to the lipreader as a supplement to the optical signal (Erber, 1972b; Risberg and Agelfors, 1978; Summerfield, 1983; Fourcin, 1990).

In fact, the simplest way to establish a viseme classification is to look into the confusion matrices and find the cells that yield the highest numbers of confusions. In addition, visemes may be the result of various analysing techniques that reduce a set of individual phoneme confusions to a smaller number of sets or clusters of phonemes sharing certain properties. Each of these techniques has its advantages and drawbacks, but their great value is that they all provide ideas about the way in which properties of the stimuli are reflected in the observer's perception

---

[1] The name viseme suggests a strong analogy with the phonological concept of the phoneme. In terms of perception, however, there is a clear difference between them. A phoneme is a linguistically defined speech unit. Factors such as acoustic environment or hearing impairment may influence a listener's perception so that certain phonemes will be confused, but nevertheless the same units are dealt with. In lipreading tasks phonemes may also be confused, in which case the confused units may be referred to as visemes. The auditory counterpart of the viseme, then, is not the phoneme, but ought to be called an 'audeme', rather. It is important to realize that whereas a phoneme is a strictly defined unit, the viseme is not.

of these stimuli. The most widely used methods are those of Information Transmission Analysis (ITA) as described by Miller and Nicely (1955), various Hierarchical Clustering (HC) techniques (Johnson, 1967), and various multidimensional scaling (MDS) techniques (Kruskal, 1964; Carroll and Chang, 1970; Schiffman *et al.*, 1981). In ITA, one studies how well information from *a priori* defined features is transmitted from stimulus to observer. In HC and MDS, phonemes are grouped into clusters on the basis of perceptual similarities. Depending on the different aspects involved in perception, clusters will be arranged in one or more dimensions, each of which may be related to certain (combinations of) physical properties of the phonemes. This method of *a posteriori* labelling of perceptual dimensions may provide new insights.

2.1.2 Objectives of the viseme study

The central issue of the study reported in this chapter was the establishment of a viseme classification for Dutch. The study is set in the context of a research programme for rehabilitation of the profoundly hearing impaired by means of signal processing hearing aids. In this approach the most relevant speech features are extracted from the acoustic signal and recoded into a less complex signal (Boothroyd, 1990; Fourcin, 1990), which is then presented to the hearing impaired in support of lipreading. Relevant speech features may be defined as features that provide speech information that cannot or can hardly be lipread. Integration of visual and auditory information is a prerequisite for successful use of the signal processing aid, so lipreading training will be part of the rehabilitation. Definition of a viseme classification of Dutch phonemes is therefore important for (a) the design of speech processing strategies, and (b) the set-up of an integrated auditory-visual speech perception training programme.

Two factors that may influence viseme categorization had our special attention: the role of the *spoken language* involved (in this case Dutch), and a factor which we would like to refer to as *expertise* with regard to lipreading. The *language factor* is important because not all languages include the same phonemes, nor even the same places and manners of articulation of consonants and vowels. As a consequence, it would be unwise or even impossible to copy results from English studies for specific applications in other languages. Structured lipreading instruction and testing programmes in Dutch should therefore be based upon Dutch categorizations of visually similar speech sounds.

Table 2.1 Data from three Dutch studies on lipreading phonemes. The classification of the data from Eggermont (1964), and from Breeuwer (1985) and Breeuwer and Plomp (1986) were based on our own analysis.

method	visemes[1]	
Eggermont (1964)	*consonants*	
stimuli: 120 CVC syllables, presented twice	1. /p,b,m/	
subjects: 156 deaf children participated, but the results of only 78 children were analysed completely	2. /f,v/	
	3. /w[2]/	
	4. /t,d,s,z/, /l/, /n,j,k,x,ŋ,R[3],h/	
speaker: one speaker was recorded on film, material was read aloud		
procedure: film was replayed under two conditions, auditory-visual and visual-only	*vowels*	
	1. /i,I,e,ɛ,ɛi/, /a,ɑ/	
data: consonant and vowel confusions were gathered in matrices and analysed	2. /ɔ,o,u,y,ø,œ,au,œy/	
Corthals (1984)	*non-sonorant*	*sonorant*
no experimental data are available, viseme classification is based on a theoretical discussion of auditory-visual lipreading of phonemes	1. /p,b/	1. /m/
	2. /f,v/	2. /w/, /u,y,ø/
	3. /ʃ,ʒ/	3. /ɔ,o/
	4. /s,z/, /t,d,k,x/	4. /i,I,e/
		5. /ɛ/
		6. /a,ɑ/
		7. /l,r,n,j/, /ŋ,R,h/, /œ/
Breeuwer (1985), Breeuwer and Plomp (1986)	*consonants*	
stimuli: one list of 36 /aCa/ syllables (18 consonants presented twice), and one list of 24 /hVt/ syllables (12 vowels presented twice)	1. /p,b,m/	
	2. /f,v,ʋ/	
	3. /s,z/, /l,r/, /t,d,n,j,k,ŋ/, /x,h/	
subjects: 24 normally hearing persons without any experience in speechreading		
speaker: female speech therapist, with clear pronunciation, recorded on colour videotape	*vowels*	
	1. /i,I,e/, /ɛ/, /a,ɑ/	
procedure: videotape was replayed under nine conditions: four auditory-visual, four auditory-only, and one visual-only	2. /ɔ,o,u,y,ø,œ/	
data: consonant and vowel confusions were gathered in confusion matrices		

To be continued on next page

Table 2.1. Continued.

Notes:

[1] Main viseme groups are indicated by numbers 1., 2., etc. If subsets can be identified within a viseme, these are separated by commas (so, /i,I,e,ɛ,ɛi/ and /a,ɑ/ under Eggermont (1964) are two subsets forming the unrounded vowel viseme).

[2] [w] is the rounded bilabial, [ʋ] the labiodental allophonic variant of the approximant /ʋ/. [ʋ] is the standard Dutch form, while [w] will mostly be found in southern Dutch and Flemish dialects, and in standard Dutch when in syllable-final position.

[3] [r] is the alveolar, [R] the uvular allophonic variant of the roll /r/. [R] is the more regular form in standard Dutch, although it may be freely interchanged with alveolar [r] in prevocalic and intervocalic position. In final position, [R] is the regular form, seldom replaced by [r].

For Dutch, there have been no such extended and systematic studies into empirical viseme categorization as for British, American or Australian English (Summerfield, 1983; Busby *et al.*, 1984, 1988; Owens and Blazek, 1985). Table 2.1 shows a summary of the results from three Dutch studies involving lipreading (Eggermont, 1964; Corthals, 1984; Breeuwer, 1985). Eggermont (1964) presented meaningless syllables to a large group of severely hearing impaired and deaf children. The classification of visemes from his study is the result of our own analysis of confusion matrices that Eggermont provides. Corthals' (1984) classification is not based on experimental data, but is the result of a mixture of data from reported studies (mainly English and American) and classical phonetic and phonological classifications. Unlike Eggermont (1964) and Breeuwer (1985), Corthals (1984) bases his classification on a condition where lipreading is supplemented with crude auditory information. Therefore, he distinguishes sonorant and non-sonorant phonemes rather than consonants and vowels. Finally, interesting data were provided by Breeuwer and Plomp (Breeuwer, 1985; Breeuwer and Plomp, 1986). In a number of experiments they studied which auditory information is feasible for supplementing lipreading. The confusion matrices provided in their work were analysed by us, resulting in the classification given in table 2.1.

The *expertise factor* relates to such lipreader qualities as explicit knowledge of the principles of lipreading or a certain degree of proficiency due to the necessity to use lipreading in everyday communication. Owens and Blazek (1985) concluded that there was no

difference in viseme recognition between normally hearing and hearing impaired subjects. Detailed inspection of their subjects' lipreading performance led them to two more conclusions: (a) for hearing impaired lipreaders, "skill in visual recognition of consonants is not necessarily related to the ability to lipread sentences", and (b) for normally hearing people, "knowledge and experience related to hearing impairment do not necessarily constitute an advantage in [visual phoneme recognition]" (p. 392). The authors did not base these conclusions on other specific investigations. Yet, the issue of expertise is an interesting one. The study reported here addressed the effect of expertise on general viseme classifications of consonants and vowels. Lipreading expertise, in this case, is defined as relating to (a) explicit knowledge of the principles of lipreading in normally hearing subjects, and (b) lipreading ability of hearing impaired subjects in everyday communication.

Summarizing, the purpose of this study was threefold:

(1) to establish general viseme classifications for Dutch vowels and consonants;
(2) to interpret the visual-perceptual dimensions underlying these classifications, and to relate these perceptual dimensions to acoustic speech cues;
(3) to establish the effect of lipreading expertise on viseme categorization.

2.2 Method

2.2.1 Stimuli

Four lists of syllables were used, each with a different target phoneme in various phonemic contexts. Table 2.2 shows the syllable types (with underlined target phoneme), the number of syllables per list, and the phonemes and phonemic contexts used for construction of the syllables.

Syllable types CV, VCV, and VC were constructed by combining consonants with the three vowels /i,u,a/. These were chosen because they hold three extreme articulatory positions (close unrounded or spread, close rounded, and open unrounded, respectively). In initial and medial syllable position 18 consonants were combined with these vowels, rendering 54 syllables of each type CV and VCV. In final syllable position, only 36 stimulus items were selected. Voiced plosives and fricatives do not appear in final position in Dutch, and, due to phonotactic constraints, the consonant /w/ can be preceded by /i/ only, while /j/ was preceded only by /a,u/. So, 11 final consonants were

combined with all three vowels, giving 33 syllables; the remaining three syllables were /iw/, /aj/, and /uj/. Finally, CVC syllables were constructed by combining 15 vowels with consonants /p,f,t,k/ (symmetric contexts: /pVp/ *etc.*). They were chosen because they cover a large range of articulatory consonant positions: bilabial, labiodental, alveolar and velar place of articulation.

The syllables within a list were randomly ordered. Each list was preceded and followed by three dummy syllables. Moreover, a 20-syllable training list was designed in which all four syllable types were equally represented. All phoneme combinations were phonotactically acceptable in Dutch, with the possible exception of /iŋ/, /aŋ/, and /uŋ/.

The lists were read out by one male and two female speakers. Both female speakers were speech therapists. The speakers were recorded head-and-shoulders on colour videotape using a Sony Umatic VO-7630 recorder and camera. The studio was normally lit, no special illumination was used in order to illuminate the oral cavity, the background was neutrally coloured, and horizontal and vertical viewing angles were zero (see Erber, 1974a). Before uttering a syllable, the speaker held his or her mouth in the 'neutral' position of the vowel /ə/. Each 7 seconds a syllable was pronounced. In between two syllables, the speaker looked down. Just before reading aloud, the speaker looked up, straight into the camera, so that the head movement indicated to the subject that a new syllable would soon follow.

Table 2.2 Target phonemes (C=consonant, V=vowel), types of stimulus syllable, number of different syllables used per target phoneme, phoneme sets used, and phonemic contexts per syllable type.

target phoneme	syllable type	N	phoneme set	phonemic context
$C_{INITIAL}$	C<u>V</u>	54	/p,t,k,b,d,f,s,ʃ,x,v,z,m,n,l,r,ʊ,j,h/	V = /i,u,a/
C_{MEDIAL}	V<u>C</u>V	54	/p,t,k,b,d,f,s,ʃ,x,v,z,m,n,l,r,ʊ,j,h/	V = /i,u,a/
C_{FINAL}	V<u>C</u>	36	/p,t,k,f,s,x,m,n,ŋ,l,R,w,j/	V = /i,u,a/
Vowel	C<u>V</u>C	60	/i,I,e,ɛ,a,ɑ,ɔ,o,u,y,ø,œ,ɛi,au,œy/	C = /p,f,t,k/

2.2.2 *Subjects*

Four groups of 12 subjects participated in the main part of this study:

1. normally hearing medical students, lacking explicit knowledge about lipreading (further called group *MSt*);
2. normally hearing speech therapy students, having specific knowledge about the principles of lipreading (group *STh*);
3. hearing impaired persons with normal everyday lipreading experience (group *HIn*);
4. hearing impaired persons who were considered to be experienced and skilled lipreaders (group *HIx*).

Table 2.3 gives relevant data about the subject groups. All subjects had normal or corrected-to-normal vision. They were paid for their services.

At first, 12 hearing impaired subjects had been selected from a regular pool in our department to form group 3. Later on, group 4 was formed by selecting new subjects on the basis of informal reports by themselves and by their close relatives that they were good lipreaders in everyday communication situations. At the end of the experiment, subjects from groups 3 and 4 performed a sentence lipreading task, the results of which largely confirmed their assignment to the groups[1], but which were not used as *post-hoc* group assignment criteria.

In order to check the reliability of the results, the two hearing impaired groups *HIn* and *HIx* were measured a second time, after about one year. In between the two measuring sessions the subjects did not receive any special training, apart from what any person had always been doing with his own speech therapist. Also, a second group of medical students was measured after a year. The second session was run by a second experimenter.

2.2.3 *Experimental design and procedure*

Within each subject group the presentation of speakers was varied in such a way that four subjects watched speaker 1, four subjects watched speaker 2 and four subjects watched speaker 3. Those subjects watching the same speaker received different list orders. This incomplete design implies that the interaction effect of speaker with subject group cannot

---

[1] The mean sentence-lipreading performance, expressed as correctly reported words-in-sentences, was 13.0% (sd=11.7%) for group *HIn* and 25.2% (sd=8.0%) for group *HIx*. One subject was clearly in the 'wrong' group: this was a subject in group *HIn* who proved to be by far the best sentence-lipreader, having a purely visual recognition score of 74% (which explains the large standard deviation in group *HIn*).

very well be tested statistically. More importantly, however, in this way all stimulus material was presented only once to each subject, so as to avoid the risk of learning effects.

Each subject was tested individually in a quiet, normally lit room. The distance to the video monitor was about 2 metres (Erber, 1974a). No audio signal was presented. At the beginning of the session, subjects were informed what to do by means of a printed instruction form. First, the training list was shown, followed by the four stimulus lists. Before each list the syllable type was made known, and attention was drawn to the target phoneme. The subject reported the syllable (or just the target phoneme) that he thought had been uttered. No feedback was given. When necessary, the experimenter emphasized that a response be given to each stimulus, even if the subject had no idea of what utterance had been presented. After each list there was a short break. On the average, the entire procedure took 45 minutes.

Table 2.3 Information on sex, age, pure-tone average (PTA; .5, 1, 2 kHz) and hearing impairment of subjects in four experimental groups: *MSt* are normally hearing medical students, *STh* are normally hearing speech therapy students, *HIn* are hearing impaired *normal* lipreaders, and *HIx* are hearing impaired *expert* lipreaders.

group	sex	age (years)		PTA [dB HL]		hearing impairment
	M/F	mean	range	mean	range	
MSt	2/10	23	20-26			
STh	0/12	22	19-23			
HIn	9/3	51	19-80	64	20-103	congenital, early and later acquired
HIx	5/7	24	11-74	95	55-115	congenital or early acquired

2.3 Results

2.3.1 Identification scores

The results from the replication study for the *HIn* and *HIx* groups indicated that, apart from an overall decrease in identification score, the subjects' response behaviour was comparable to that in the main study. The same holds for the second group of students when compared to

group *MSt*. The confusion patterns of both sessions were quite similar so that we may conclude that the data from the main part seem representative for actual lipreading performance. Consequently, they were used for evaluation of phoneme lipreading performance.

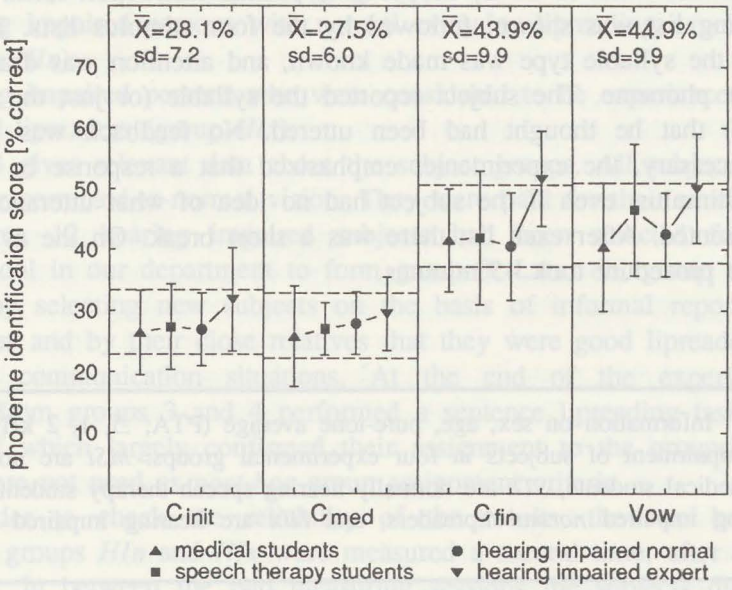

Figure 2.1 Overall identification scores of four subject groups in four phoneme positions. The dashed lines indicate 95% limits of confidence (CL95) that group means do not deviate significantly from the overall mean at a given phoneme position.

Figure 2.1 gives the overall identification scores per phoneme position of four subject groups. The significance of differences among group means was tested in a multivariate analysis of variance (MANOVA; design: four variables, four groups of 12 subjects each). Although the expert lipreaders in group *HIx* generally show the best performance, no statistically significant multivariate effect was found: Wilks' Lambda=.69, corresponding to F(12,108.8)=1.38 (p=.19). The dashed lines in figure 2.1 indicate the 95% confidence limits (CL95) for mean scores, based on the normal approximation to the binomial distribution for probabilities of $1/N_{PHONEMES}$ (0.056 for 18 initial and medial consonants, 0.077 for 13 final consonants, and 0.067 for 15 vowels). The limits thus indicate, for each phoneme position (*Cinit*, *Cmed*, *Cfin*,

Vow), the mean score that a subject group should attain before deviating significantly from the mean over all subject groups (given for each phoneme at the top of the figure). The CL95 limit is exceeded by the *HIx* group in the final consonants, and univariate F tests applied on individual consonant and vowel scores after MANOVA indicate that, indeed, the *HIx* lipreaders are significantly better than the other lipreaders in identifying the final consonants: $F(3,44)=3.92$, $p=.015$. However, no significant difference in group means was found for any of the other phonemes.

2.3.2 Patterns of phoneme confusions

For a good appreciation of lipreading performance also phoneme confusions were studied, since different confusion patterns might arise from the responses by different subject groups. In order to give the reader a general idea of the phoneme confusions made by the subjects, four confusion matrices are displayed in tables 2.4 to 2.7. In all matrices, stimulus/response pairs from 48 subjects were aggregated. The utmost left column indicates the stimuli, the top line indicates the responses. Cells contain percentages confusions; the bold figures in the main diagonal indicate percentages correct identifications. The utmost right column ('?') gives the percentages missing responses. The bottom line gives the number of times that that phoneme was returned by the subjects (N_{RESP}).

Although the patterns of confusions among lipread phonemes (or *viseme classification*) discussed below will not be based on these overall matrices but rather on the confusion matrices from individual subjects, the confusion patterns may be recognized in global lines from the matrices in tables 2.4 to 2.7. From these confusion patterns, as well as from the identification scores, it appeared that the subjects responded equally to the initial and medial consonants in our study. Consequently, we decided to use only pooled responses to initial and medial consonants for further analysis.

Consonant and vowel confusion matrices of 48 subjects were transformed into symmetric similarity matrices, using an algorithm proposed by Houtgast (Klein *et al.*, 1970). Subsequently, the 48 similarity matrices per phoneme position were analysed by means of an INDSCAL technique according to Carroll and Chang (1970). Each analysis yielded a solution in two dimensions explaining a good part of the total variance, between 88.3% and 94.6%.

Table 2.4 Initial consonant confusion matrix (18 consonants, 48 subjects). All scores and confusions are based on 144 presentations; the total number of presentations was 2,592. Overall identification score was 28.1%.

stim \ resp	p	b	m	f	v	ʋ	s	z	ʃ	t	d	j	n	l	k	x	r	h	?
p	**65.3**	17.3	11.8	.	.	.	.	.	.	.	.	0.7	.	.	0.7	.	.	0.7	3.5
b	46.5	**27.8**	25.0	.	.	.	.	.	.	.	.	.	.	.	0.7	.	.	.	.
m	31.9	18.1	**44.5**	.	0.7	2.8	0.7	.	.	.	.	.	.	.	.	.	0.7	0.7	.
f	.	1.4	.	**62.5**	23.6	8.4	0.7	.	.	1.4	2.1	.	.	.	.	.	.	.	.
v	0.7	.	.	49.3	**29.9**	9.1	.	0.7	.	2.1	3.5	.	0.7	.	0.7	.	.	2.8	0.7
ʋ	2.1	0.7	0.7	48.6	28.5	**13.9**	1.4	.	.	.	4.2	.	.	.	.	.	.	.	.
s	.	.	.	.	.	0.7	**37.5**	12.5	0.7	17.3	8.3	6.3	7.0	3.5	0.7	2.1	0.7	2.1	0.7
z	.	.	0.7	0.7	1.4	1.4	31.9	**14.6**	.	20.1	9.7	2.1	7.7	3.5	1.4	.	1.4	2.8	0.7
ʃ	.	0.7	.	.	.	0.7	40.3	15.3	**1.4**	17.3	5.6	3.5	0.7	4.2	2.8	1.4	3.5	1.4	1.4
t	.	.	0.7	0.7	.	.	9.0	4.9	.	**36.8**	20.1	4.9	9.0	2.8	4.2	0.7	1.4	2.1	2.8
d	0.7	0.7	0.7	.	0.7	0.7	14.6	2.1	.	18.8	**25.7**	3.5	11.8	7.0	6.3	2.1	3.5	1.4	.
j	.	.	0.7	.	0.7	.	10.4	4.9	0.7	18.1	10.4	**11.1**	8.4	8.4	12.5	3.5	5.6	4.9	.
n	.	.	.	.	.	0.7	7.0	2.1	.	6.3	11.8	7.0	**12.5**	22.9	7.7	3.5	7.0	9.7	2.1
l	.	1.4	1.4	.	0.7	.	0.7	0.7	.	3.5	4.2	4.9	8.3	**41.0**	5.6	3.5	7.0	15.3	2.1
k	.	0.7	.	.	0.7	1.4	3.5	0.7	.	5.6	6.3	7.7	9.7	8.4	**20.8**	9.7	7.7	16.0	1.4
x	.	.	0.7	.	0.7	.	2.1	.	.	7.0	6.3	4.9	4.2	11.8	16.7	**9.0**	15.3	20.2	1.4
r	.	.	0.7	0.7	.	.	2.1	2.1	.	4.9	4.2	2.8	3.5	20.8	25.7	7.0	**9.7**	15.3	0.7
h	1.4	.	.	.	1.4	0.7	2.8	0.7	.	3.5	2.1	4.9	4.9	8.3	9.7	9.0	5.6	**41.7**	3.5
N_{RESP}	214	99	126	234	128	58	237	88	4	234	179	92	127	205	167	74	99	197	30

Table 2.5 Medial consonant confusion matrix (18 consonants, 48 subjects). All scores and confusions are based on 144 presentations; the total number of presentations was 2,592. Overall identification score was 27.5%.

stim \ resp	p	b	m	f	v	ʋ	s	z	ʃ	t	d	j	n	l	k	x	r	h	?
p	**79.2**	13.2	4.9	0.7	.	.	0.7	.	.	0.7	.	.	.	0.7	.	.	.	.	.
b	67.4	**14.6**	14.6	.	0.7	0.7	.	0.7	.	.	0.7	.	0.7	.	.	.	.	.	.
m	57.6	12.5	**27.1**	.	.	1.4	0.7	.	.	.	.	.	.	0.7	.	.	.	.	.
f	.	1.4	.	**77.1**	13.2	4.2	.	0.7	.	1.4	.	.	1.4	.	.	.	0.7	.	.
v	.	0.7	0.7	63.2	**23.6**	6.3	.	.	.	2.1	.	.	0.7	0.7	.	.	.	.	2.1
ʋ	.	0.7	0.7	62.5	20.2	**7.0**	.	.	.	3.5	1.4	.	0.7	.	.	.	2.1	1.4	.
s	.	0.7	0.7	.	0.7	2.1	**45.1**	5.6	0.7	22.9	2.1	4.2	2.8	4.9	2.1	0.7	2.8	0.7	1.4
z	.	.	0.7	.	.	0.7	34.1	**9.8**	0.7	24.3	4.9	2.8	8.4	3.5	2.1	1.4	4.9	0.7	1.4
ʃ	.	.	0.7	0.7	.	0.7	44.5	9.7	**5.6**	18.1	2.1	4.2	5.6	4.2	2.1	.	0.7	0.7	0.7
t	.	.	.	2.1	1.4	1.4	9.7	2.1	.	**54.9**	8.3	1.4	2.8	4.2	4.9	0.7	.	2.1	4.2
d	0.7	.	.	0.7	.	1.4	4.2	0.7	.	38.9	**10.4**	2.8	5.6	11.8	9.0	1.4	5.6	4.2	2.8
j	.	0.7	0.7	.	0.7	1.4	7.0	.	.	19.5	8.4	**10.4**	11.8	14.6	9.1	1.4	9.7	2.1	2.8
n	.	.	.	.	.	0.7	4.9	4.9	.	21.5	6.3	1.4	**11.8**	20.2	6.3	4.9	11.1	5.6	0.7
l	.	.	.	0.7	.	0.7	0.7	0.7	.	2.1	1.4	2.8	4.2	**47.9**	3.5	2.8	20.2	10.4	2.1
k	0.7	.	.	.	.	0.7	5.6	0.7	.	11.8	6.3	5.6	11.1	11.8	**16.7**	7.7	14.6	4.9	2.1
x	0.7	.	.	0.7	0.7	0.7	.	0.7	.	6.3	4.2	5.6	4.2	6.3	20.8	**12.5**	27.1	7.7	2.1
r	0.7	2.1	0.7	.	0.7	1.4	5.6	0.7	.	13.2	4.9	3.5	7.0	13.9	11.8	6.3	**19.4**	6.9	1.4
h	1.4	0.7	.	.	1.4	1.4	2.1	0.7	.	1.4	3.5	2.1	6.3	8.4	12.5	9.0	21.5	**22.9**	4.9
N_{RESP}	300	68	74	300	91	47	237	54	10	349	93	67	122	221	145	70	202	101	41

Table 2.6 Final consonant confusion matrix (13 consonants, 48 subjects). All scores and confusions are based on 144 presentations, except those for /w/ (48 presentations) and for /j/ (96 presentations); the total number of presentations was 1,728. Overall identification score was 43.9%.

stim	resp p	m	f	w	s	t	j	n	l	k	x	ŋ	R	?
p	**97.9**	1.4	.	.	.	.	.	.	.	0.7	.	.	.	.
m	36.1	**63.9**	.	.	.	.	.	.	.	.	.	.	.	.
f	1.4	.	**93.8**	.	1.4	3.5	.	.	.	.	.	.	.	.
w	2.1	.	60.4	**33.3**	2.1	.	.	.	.	.	.	.	.	2.1
s	0.7	.	0.7	.	**48.6**	35.5	2.1	4.9	2.1	0.7	0.7	.	4.2	.
t	1.4	.	2.1	0.7	9.1	**63.2**	0.7	8.4	4.2	5.6	.	.	3.5	1.4
j	.	.	1.1	.	4.2	2.1	**40.6**	7.3	14.6	8.4	3.1	.	15.6	3.1
n	1.4	.	0.7	.	4.9	27.8	6.3	**16.0**	13.2	7.7	4.9	1.4	15.3	0.7
l	.	.	0.7	0.7	1.4	16.7	3.5	7.7	**35.4**	7.0	3.5	.	22.9	0.7
k	0.7	.	.	0.7	5.6	17.4	6.3	12.5	10.4	**21.5**	3.5	0.7	18.8	2.1
x	.	.	0.7	0.7	7.7	14.6	3.5	7.6	6.3	22.2	**9.1**	0.7	22.2	3.5
ŋ	.	.	2.1	.	6.3	8.4	4.2	3.5	13.2	19.4	7.7	**0.0**	27.1	8.3
R	.	.	0.7	1.4	2.1	9.8	6.3	6.3	7.0	16.7	9.0	.	**38.9**	2.1
N_{RESP}	202	94	176	22	130	285	86	103	146	154	58	4	235	31

Table 2.7 Vowel confusion matrix (15 vowels, 48 subjects). All scores and confusions are based on 192 presentations; the total number of presentations was 2,880. Overall identification score was 44.9%.

	resp															
stim	i	I	e	ɛ	ɛi	a	ɑ	u	y	ɔ	œ	o	ø	œy	au	?
i	**52.6**	18.2	2.1	7.8	3.7	1.1	6.8	.	.	1.6	2.6	.	.	0.5	.	3.1
I	26.6	**31.3**	5.8	14.6	4.7	1.1	13.0	0.5	.	0.5	1.1	.	.	0.5	.	0.5
e	21.4	4.7	**24.0**	8.4	29.7	8.3	0.5	.	.	0.5	1.1	.	.	0.5	0.5	0.5
ɛ	14.6	21.9	7.3	**17.2**	16.2	4.2	15.1	1.1	.	.	0.5	.	0.5	0.5	0.5	0.5
ɛi	4.2	0.5	12.0	3.7	**54.2**	20.9	1.1	.	.	0.5	.	.	.	1.6	.	1.6
a	.	.	5.2	2.6	20.9	**62.5**	4.7	.	.	1.1	.	.	.	2.1	1.1	.
ɑ	4.7	4.7	.	4.7	.	4.7	**70.9**	0.5	.	2.1	1.1	0.5	.	2.6	3.2	0.5
u	0.5	.	0.5	.	.	.	0.5	**73.0**	2.1	12.0	4.2	5.8	0.5	.	.	1.1
y	.	1.1	.	.	.	.	.	65.6	**5.7**	18.8	3.2	5.3	.	.	.	0.5
ɔ	.	.	.	.	.	.	1.1	15.6	1.6	**61.5**	5.7	11.5	2.1	.	0.5	0.5
œ	1.1	.	.	.	0.5	0.5	5.8	12.5	1.1	44.8	**21.9**	7.3	3.2	.	.	1.6
o	.	.	.	.	.	.	0.5	3.7	.	8.4	.	**81.3**	2.6	2.1	0.5	1.1
ø	.	.	0.5	0.5	.	1.1	.	4.7	0.5	20.8	0.5	59.4	**5.2**	3.1	3.2	0.5
œy	0.5	.	.	0.5	.	3.2	3.2	.	0.5	0.5	0.5	6.8	1.6	**39.1**	43.3	0.5
au	0.5	.	.	.	0.5	1.6	2.1	0.5	.	2.6	.	6.3	.	12.5	**72.9**	0.5
N_{RESP}	243	158	110	115	250	209	240	341	22	337	81	353	30	125	241	25

Consonant confusions

Figure 2.2 displays the object, or stimulus, spaces (left) and the subject, or condition, spaces (right) for 18 initial and medial consonants (top) and for 13 final consonants (bottom). The condition spaces for all consonants will be discussed in section 2.3.3 below. The stimulus configurations of figure 2.2 make clear that confusions are almost exclusively made within three major sets:

[1] /p,b,m/ bilabial consonants
[2] /f,v,ʋ/ labiodental consonants
[3] /t,d,s,z,n,l,ʃ,j,k,r,x,ŋ,h/ nonlabial consonants.

As mentioned before, the velar nasal /ŋ/ only appears in final position, while the voiced plosive and fricative consonants /b,v,d,z/ and glottal /h/ only appear in syllable-initial or syllable-medial position in Dutch (see table 2.2).

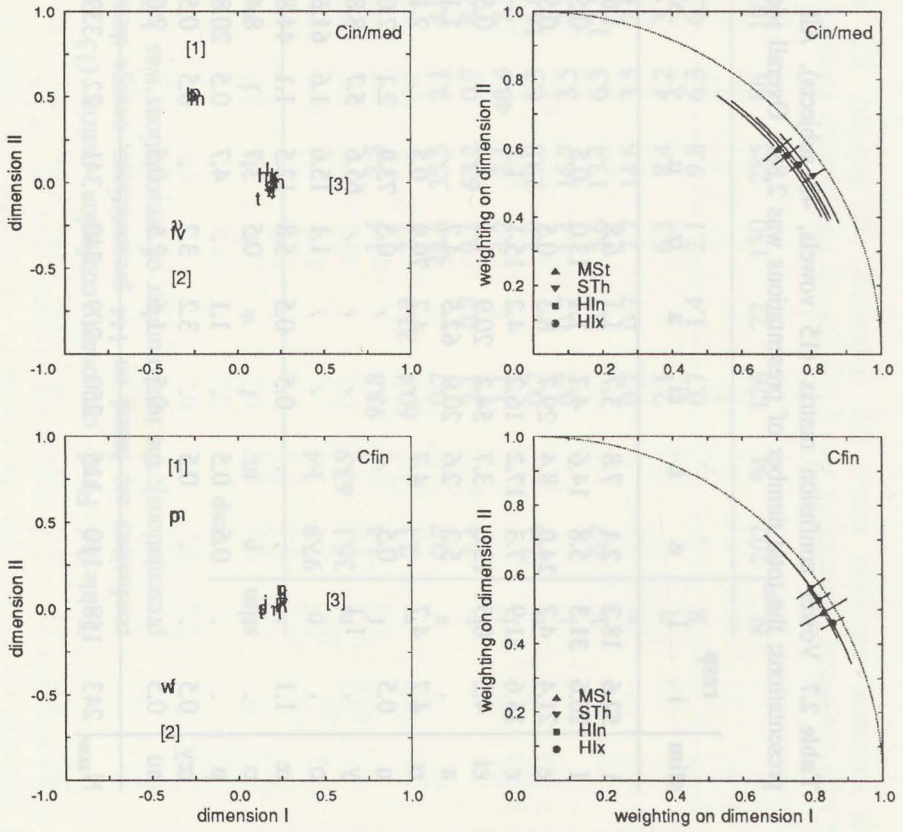

Figure 2.2 Stimulus and condition spaces of 18 initial/medial consonants (top panels) and 13 final consonants (bottom panels). [1] *bilabials*, [2] *labiodentals*, [3] *nonlabials*. Explained variance for $C_{IN/MED}$ is 88.3% (dim.I 56.4%, dim.II 31.9%), that for C_{FIN} is 94.6% (dim.I 68.4%, dim.II 26.2%).

The three sets form a degenerate solution: due to the large dissimilarity among the sets and the strong coherency of the phonemes within each set, the sets are displayed as three distinct spots. For this reason it is difficult to interpret the two dimensions along which the sets are ordered. Clearly, visible *lip articulation* is the most important cue for distinguishing the sets. Sets [1] and [2], bilabial and labiodental consonants respectively, are clear-cut: their members are visually equal, because they share the same visible articulatory movements. This does not hold, however, for set [3], which includes phonemes that have varying visible articulatory shapes. For an appreciation of the similarities among these consonants, we analysed them apart from the consonants in sets [1] and [2]. The resulting stimulus and condition spaces, explaining 64.5% (initial/medial) and 74.1% (final) of the total variance, are displayed in the plots of figure 2.3.

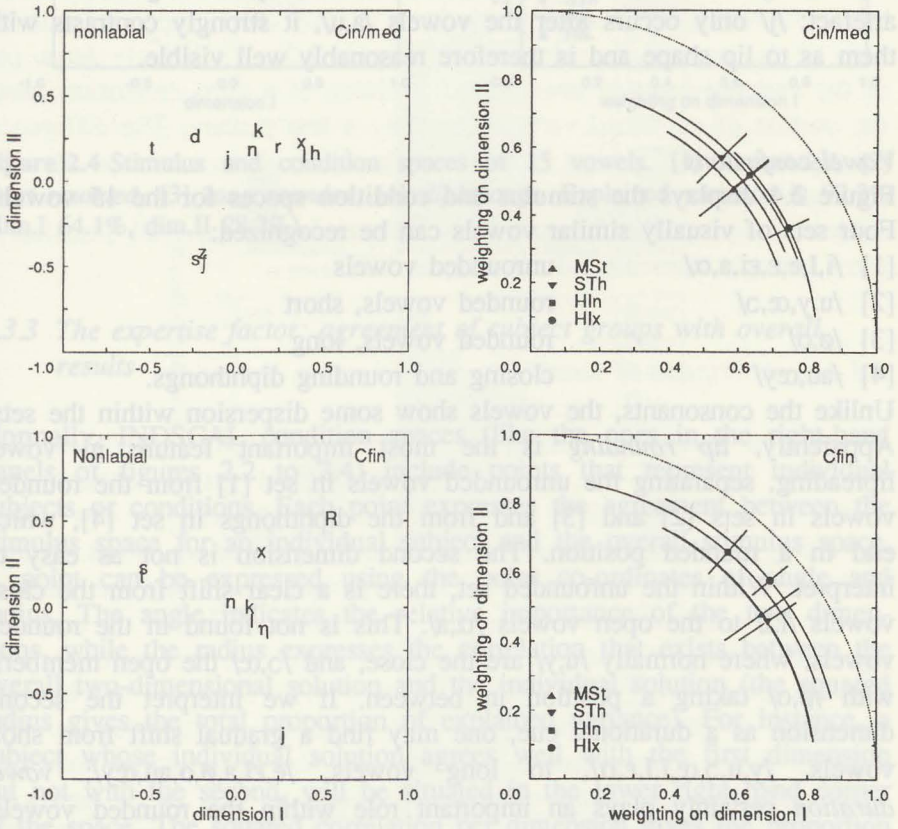

Figure 2.3 Stimulus and condition spaces of 12 *nonlabial* initial/medial consonants (top panels) and 9 *nonlabial* final consonants (bottom panels). Explained variance for $C_{IN/MED}$ is 64.5% (dim.I 39.7%, dim.II 24.8%), that for C_{FIN} is 74.1% (dim.I 41.9%, dim.II 32.2%).

Along the first dimension of the initial/medial consonant plot in the top panel of figure 2.3 one may observe a gradual shift from front-articulated consonants (such as alveolar /t,d/) to back-articulated consonants (such as uvular /R/ or glottal /h/), although the order of consonants does not correlate strongly with the actual *place of articulation.* The second dimension displays a difference between the fricatives /s,z,ʃ/ and the other nonlabial consonants. Identification of these fricatives may be based on the *(limited) degree of opening of the oral cavity,* due to the necessary constriction near the front of the oral cavity. The first dimension of the final consonants in the bottom panel of figure 2.3 is less clear. It may be compared with the second dimension of the initial/medial consonants. In final position, /s,t/ may be identified on the basis of the relatively closed oral cavity (remember that /z,ʃ/ do not occur). The second dimension mainly displays a difference between /j/ and the other consonants. This may be regarded as an artefact: /j/ only occurs after the vowels /a,u/, it strongly contrasts with them as to lip shape and is therefore reasonably well visible.

Vowel confusions

Figure 2.4 displays the stimulus and condition spaces for the 15 vowels. Four sets of visually similar vowels can be recognized:

[1] /i,I,e,ɛ,ɛi,a,ɑ/	unrounded vowels
[2] /u,y,œ,ɔ/	rounded vowels, short
[3] /ø,o/	rounded vowels, long
[4] /au,œy/	closing and rounding diphthongs.

Unlike the consonants, the vowels show some dispersion within the sets. Apparently, *lip rounding* is the most important feature in vowel lipreading, separating the unrounded vowels in set [1] from the rounded vowels in sets [2] and [3] and from the diphthongs in set [4], which end in a rounded position. The second dimension is not as easy to interpret. Within the unrounded set, there is a clear shift from the close vowels /i,I/ to the open vowels /ɑ,a/. This is not found in the rounded vowels, where normally /u,y/ are the close, and /ɔ,œ/ the open members, with /ø,o/ taking a position in between. If we interpret the second dimension as a durational cue, one may find a gradual shift from short vowels, /y,u,ɔ,œ,i,I,ɛ,ɑ/, to long vowels, /e,ɛi,a,ø,o,au,œy/; *vowel duration* certainly plays an important role within the rounded vowels, separating long /o,ø/ from short /u,y,œ,ɔ/. In the diphthongs /au,œy/ both *lip rounding* and *lip opening* play a role. The movement in the diphthongs is perceived when the change from less to more rounding as well as the change from openness to closeness is perceived. In the third

Dutch diphthong, /ɛi/ in set [1], there is no lip rounding at all, nor is there an appreciable amount of change along the open/close dimension. Therefore it must be regarded as an unrounded monophthong in the visual sense.

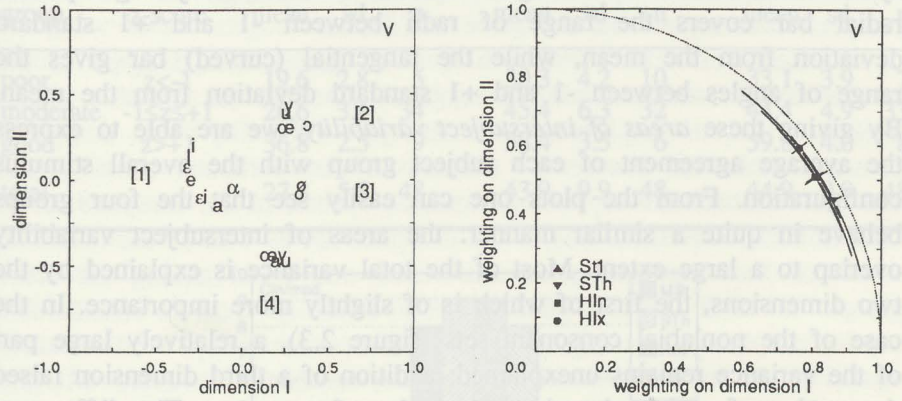

Figure 2.4 Stimulus and condition spaces of 15 vowels. [1] *unrounded*, [2] *short rounded*, [3] *long rounded*, [4] *diphthongs*. Explained variance is 92.3% (dim.I 64.1%, dim.II 28.2%).

2.3.3 The expertise factor: agreement of subject groups with overall results

Normally, INDSCAL condition spaces (like the ones in the right-hand panels of figures 2.2 to 2.4) include points that represent individual subjects or conditions. Each point expresses the agreement between the stimulus space for an individual subject and the overall stimulus space. A point can be expressed using the polar co-ordinates of angle and radius. The angle indicates the relative importance of the two dimensions, while the radius expresses the correlation that exists between the overall two-dimensional solution and the individual solution (the squared radius gives the total proportion of explained variance). For instance, a subject whose individual solution agrees well with the first dimension but not with the second, will be situated in the lower right hand corner of the space. The squared correlation per dimension gives the proportion of variance that is explained by the use of that dimension. The maximum of 100% explained variance by the two dimensions is indicated by the quarter circle drawn in the plots. The further a subject

is removed from that arc, the less variance will be explained by the given two-dimensional stimulus configuration - in other words, the less will his individual solution correspond to the overall solution.

In the condition spaces of figures 2.2 to 2.4 we have not indicated 48 individual subjects, but rather the average position taken by each of the four subject groups (averaged over 12 subjects). In the figures, the symbols indicate the mean angle and radius for each subject group. The radial bar covers the range of radii between -1 and +1 standard deviation from the mean, while the tangential (curved) bar gives the range of angles between -1 and +1 standard deviation from the mean. By giving these *areas of intersubject variability*, we are able to express the average agreement of each subject group with the overall stimulus configuration. From the plots one can easily see that the four groups behave in quite a similar manner: the areas of intersubject variability overlap to a large extent. Most of the total variance is explained by the two dimensions, the first of which is of slightly more importance. In the case of the nonlabial consonant sets (figure 2.3), a relatively large part of the variance remains unexplained; addition of a third dimension raised the portion of explained variance by only a few percent. The differences among the subject groups show most clearly in these nonlabial stimulus spaces, implying that viseme categorization is less consistent across subject groups within this subset as compared to the overall situation of three strongly concentrated sets (figure 2.2).

2.3.4 A description of visemes in good phoneme lipreaders

In order to gain some insight into differences between good and poor phoneme lipreaders we abandoned our original group categorization described in section 2.2.2, and assigned the 48 subjects to new groups following a *post-hoc* analysis of phoneme identification scores. Subject assignment to any one of three score groups was based on standard scores: *poor* lipreaders are those with standard score below -1, *moderate* lipreaders have a standard score between -1 and +1, and *good* performers are those with standard scores over +1. Standard scores were computed for initial/medial and final consonants and for vowels. With each of these three phoneme positions different score groups emerged. Table 2.8 gives the scores of the three groups as well as the scores for the entire group (compare figure 2.1). According to the normal distribution, about 16% of the scores should be higher than z=+1 (which equals 7.7 subjects), and 16% should be lower than z=-1. Table 2.8 gives the actual numbers of subjects per score group per phoneme position.

Table 2.8 Phoneme identification scores (%) on initial/medial and final consonants, and vowels (mean percentages and standard deviations). Scores of three groups composed according to z-scores: *poor*, *moderate* and *good* lipreaders. *n* indicates the number of subjects per score group.

score group	z-score	$C_{IN/MED}$ mean	sd	n	C_{FIN} mean	sd	n	Vowel mean	sd	n
poor	z<-1	19.6	2.8	5	30.3	4.2	10	33.1	3.9	7
moderate	-1≤z≤+1	26.6	3.4	34	45.4	6.3	32	44.4	4.9	33
good	z>+1	36.8	2.5	9	58.4	3.5	6	59.8	4.0	8
total		27.8	5.8	48	43.9	9.9	48	44.9	9.9	48

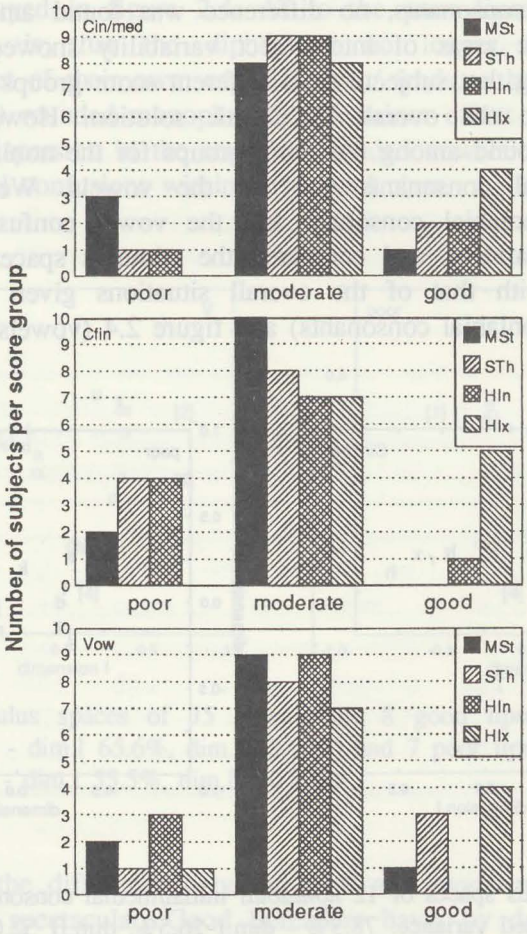

Figure 2.5 Numbers of subjects from original groups in new score groups of *poor* (z<-1), *moderate* (-1≤z≤+1) and *good* (z>+1) phoneme lipreaders.

Figure 2.5 shows how members from the original subject groups are distributed over the newly formed score groups. It is clear that the *HIx* lipreaders are overrepresented in the *good* class and underrepresented in the *poor* class; in general, they do not belong to the *poor* phoneme lipreaders. This emphasizes the earlier mentioned (section 2.3.1) tendency for expert lipreaders to have slightly higher phoneme identification scores than other lipreaders, although in an overall perspective no significant difference was found.

In order to check how the three score groups would agree to the joint stimulus configurations (48 subjects) given in the stimulus spaces of figures 2.2 to 2.4, areas of intersubject variability were computed (averaged over the number of subjects in that particular score group) and plotted in the corresponding condition spaces. In the complete set of initial/medial consonants and in the complete set as well as the nonlabial subset of final consonants, no difference was found among the score groups: all three areas of intersubject variability showed considerable overlap, implying that subjects from different score groups tend to attach equal weight to the overall INDSCAL solution. However, different behaviour was found among the score groups for the nonlabial subset of the initial/medial consonants and for the vowels. We subsequently analysed the nonlabial consonant and the vowel confusions for each score group separately, and compared the stimulus spaces of the three score groups with that of the overall situations given in figure 2.3 (initial/medial nonlabial consonants) and figure 2.4 (vowels).

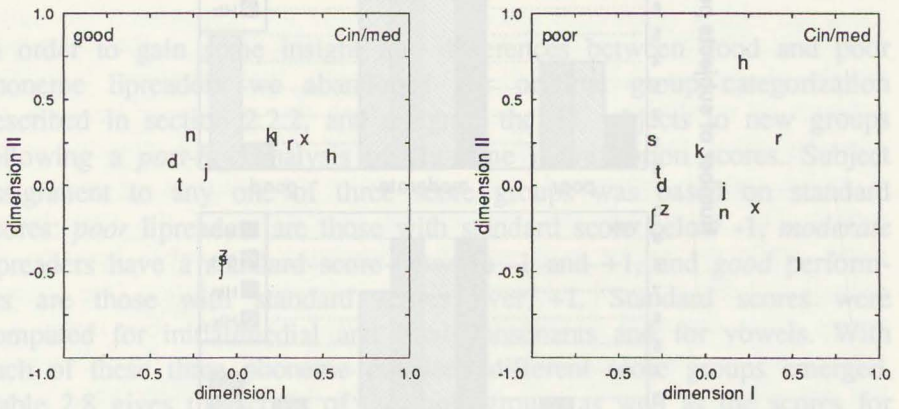

Figure 2.6 Stimulus spaces of 12 *nonlabial* initial/medial consonants for 9 good lipreaders (explained variance: 78.5% - dim.I 26.5%, dim.II 52.0%) and 5 poor lipreaders (explained variance: 65.6% - dim.I 38.6%, dim.II 27.0%).

Figure 2.6 shows the nonlabial consonant spaces of the good and poor lipreaders (as expected, the stimulus space of the moderate lipreaders is almost identical to the average situation, for which reason we do not show it here). Good phoneme lipreaders tend to discriminate among three nonlabial sets: front fricatives /s,z,ʃ/, front consonants /t,d,n,j/, and back consonants /k,x,r,h/ plus alveolar /l/, the only consonant that seems to be in the 'wrong' position. Poor phoneme lipreaders appear not to recognize these categories: /s,z,ʃ/ are integrated in the set, and no more than a slight shift from front to back consonants can be found.

The resulting stimulus spaces of good and poor vowel lipreaders are displayed in figure 2.7 (again, moderate lipreaders are perfect representatives of the overall situation). The good lipreaders seem to make fewer vowel confusions within the class of rounded vowels, but on the whole their vowel configuration agrees well with the overall situation displayed in figure 2.4. As to the poor lipreaders, the most striking thing is that the diphthongs no longer form a tightly concentrated set of their own; especially /au/ tends to be confused more with the (long) rounded monophthongs. A minor effect seems to be the reduction of dispersion in the set of unrounded vowels: apparently there are more vowel confusions within this vowel class.

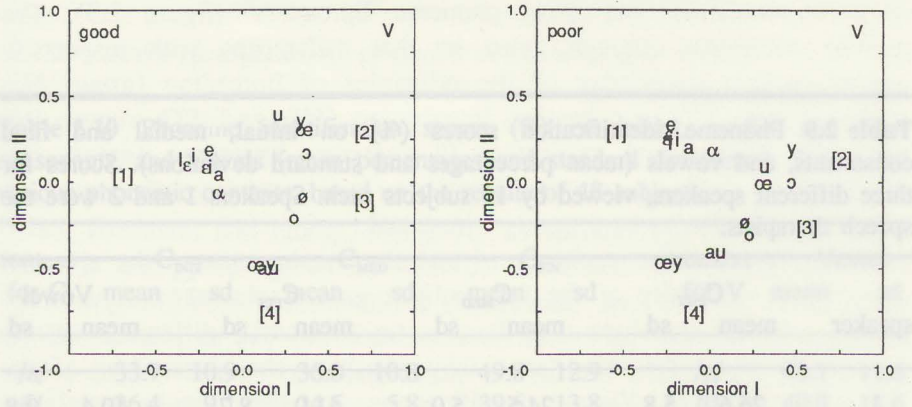

Figure 2.7 Stimulus spaces of 15 vowels for 8 good lipreaders (explained variance: 95.3% - dim.I 65.6%, dim.II 29.7%) and 7 poor lipreaders (explained variance: 86.3% - dim.I 55.5%, dim.II 30.8%).

Summarizing, the difference between *good* and *moderate* phoneme lipreaders is not spectacular. Good lipreaders have, by definition, higher identification scores, but they make use of the same visual features, and

phonemes tend to cluster into the same sets or visemes. There is considerable difference, however, between *good* and *poor* lipreaders. This difference actually becomes manifest in different viseme classifications: poor lipreaders hardly make discriminations within the set of nonlabial consonants and they make considerably more confusions between round vowels and the diphthongs /au,œy/, which end in a rounded lip shape.

2.3.5 Some remarks on effects from speakers and phonemic contexts

Although studying speaker and context effects was not our main goal, the design of the experiment gives opportunities for further exploration of scores on these variables. In this section we will therefore present the more illuminating overall results.

It is known from other studies that the speaker may have an effect on the performance of lipreaders (Walden *et al.*, 1981; Kricos and Lesner, 1982). In our experiment we used three different speakers in order to level out possible speaker-specific effects. Any one subject received the syllable lists pronounced by only one speaker, so that eventually each speaker had been viewed by 16 subjects, four from each subject group. In table 2.9 the overall scores are displayed for the three speakers.

Table 2.9 Phoneme identification scores (%) on initial, medial and final consonants, and vowels (mean percentages and standard deviations). Scores for three different speakers, viewed by 16 subjects each. Speakers 1 and 2 were the speech therapists.

	C_{INIT}		C_{MED}		C_{FIN}		Vowel	
speaker	mean	sd	mean	sd	mean	sd	mean	sd
1	26.6	5.8	24.5	5.0	41.0	8.7	40.4	9.8
2	25.9	7.3	29.3	5.9	44.6	12.3	48.2	10.4
3	31.7	7.4	28.8	6.3	46.0	8.1	45.9	8.1

The differences among the scores are marginal. As to phoneme identification scores, the only significant effects were found in the initial consonants, $F(2,45)=3.36$, $p=.04$, and medial consonants, $F(2,45)=3.28$,

p=.05. The differences were not large enough for *post-hoc* comparisons (Scheffé's F-test) to be significant. Furthermore, inspection of INDSCAL plots showed that the areas of intersubject variability around the speakers' mean points (averaged over 16 subjects) overlapped considerably. On the whole, then, we may conclude that the use of three different speakers has not led to clearly different viseme classifications.

As to phonemic context effects, it has been shown occasionally (Benguerel and Pichora-Fuller, 1982; Owens and Blazek, 1985; Montgomery *et al.*, 1987) that the use of certain vowels and consonants as a context to the stimulus phoneme may facilitate lipreading, as is the case *e.g.* with the English open vowel /æ/. In our study we found some clear effects. Table 2.10 displays mean phoneme identification scores in relation to context; the scores were based on the pooled results from all 48 subjects.

In all three consonant positions the unrounded open-front vowel /a/ led to significantly higher scores: $F(1,44)=111$, $p=.00$. There was no statistically significant difference between the unrounded close-front vowel /i/ and the rounded close-back vowel /u/: $F(1,44)=.46$, $p=.50$. As to the vowels, the nonlabial consonants /t/ and /k/ led to significantly higher scores than the labial consonants /p/ and /f/: $F(1,44)=15.4$, $p=.00$. No significant effect was found for the difference between /p/ and /f/, $F(1,44)=1.6$, $p=.22$, nor for that between /t/ and /k/, $F(1,44)=.09$, $p=.76$.

Table 2.10 Phoneme identification scores (%) on initial, medial and final consonants, and vowels (mean percentages and standard deviations). Scores for various phonemic contexts, based on the results of 48 subjects.

context for C	C_{INIT} mean	sd	C_{MED} mean	sd	C_{FIN} mean	sd	context for V	Vowel mean	sd
/a/	33.7	10.9	36.0	10.8	49.8	12.9	/p/	43.1	11.6
/i/	26.4	9.1	24.5	5.8	39.6	13.8	/f/	40.8	11.6
/u/	24.2	9.1	22.1	9.1	42.2	12.6	/t/	48.2	16.6
							/k/	47.4	14.1

Obviously, facilitating contexts are those that provide some view into the oral cavity; this bears on the feature of *lip opening*, rather than viseme membership. The vowels /i,u/ and the consonants /p,f/ do not

permit a lipreader to see articulatory movements inside the oral cavity; the open vowel /a/ gives ample view, while the consonants /t,k/, if not giving a clear view, at least do not obstruct this view when half-open or open vowels are pronounced. After checking the identification scores of individual phonemes in each of the contexts, we found that the facilitating contexts /a/ and /t,k/ resulted merely in overall higher scores and that no dramatic changes in confusion patterns were found, but for one clear exception: in the /a/ context, the lateral alveolar /l/ is much less often confused with other nonlabial consonants when compared to the overall situation.

2.4 Discussion and conclusions

2.4.1 Lipreading expertise

From the statistics given in section 2.3.1 we may conclude that, in terms of identification scores, there is little reason for treating the four subject groups as separate entities, although the lipreaders of group *HIx* tend to score slightly higher. The fact that 'skilled everyday lipreaders' are also good phoneme lipreaders is again brought forward by the distribution of lipreaders from the four groups over the *post-hoc* defined score groups of *poor*, *moderate* and *good* phoneme lipreaders[1] (figure 2.5). The general conclusion, though, must be that differences with respect to neither explicit knowledge of the principles of lipreading (group *MSt* versus group *STh*), nor hearing impairment (*MSt* and *STh* versus *HIn* and *HIx*), nor special experience in everyday lipreading (group *HIn* versus *HIx*) lead to clear differences in identifying phonemes in meaningless syllables in a purely visual task.

With regard to the positions of and the overlap among the areas of intersubject variability of the four subject groups, shown in figures 2.2 to 2.4, we may conclude that there is no substantial difference among the groups as to the importance they attach to the dimensions of the INDSCAL solutions. Viseme classifications may therefore be concluded not to be influenced by lipreading expertise; it seems that neither knowledge of the principles of lipreading nor a certain degree of proficiency in everyday lipreading is an advantage for recognizing

---

[1] Interestingly, a counterexample to this observation is the excellent sentence-lipreader mentioned in note 1 on page 30, whose purely visual phoneme identification scores are very moderate: with z-scores ranging from -.50 to +.60 this subject truly belongs to the group of *moderate* phoneme lipreaders.

visemes, which conclusion is in full accordance with the findings of Owens and Blazek (1985). It is obvious that untrained and unaware lipreaders are equally well able to recognize a number of visemes as well-trained and experienced persons. This conclusion implies that viseme classifications can be established without having to pay too much attention to the composition of subject groups. From various studies we have so far learned that variables like presence and degree of hearing impairment, skill and experience are not strongly related to lipreading phonemes (Conrad, 1977; Risberg and Agelfors, 1978; Caplan Hack and Erber, 1982; Owens and Blazek, 1985). The sometimes considerable differences among individual subjects should be explained by other factors, *e.g.* linguistic capacities (Stoker and French-St. George, 1984).

Visemes only tend to change with extremely poor phoneme lipreaders; besides having low identification scores, which defines them as being poor performers, they actually make more phoneme confusions across the overall viseme boundaries. This naturally leads to different viseme classifications of both consonants and vowels. The difference between good and moderate phoneme lipreaders, however, does not lead to different viseme classifications. This observation implies, once more, that lipreading expertise does not influence viseme groupings, even though it may have some effect on correct phoneme identification.

2.4.2 Viseme classifications of Dutch consonants and vowels

In terms of consonant and vowel visemes, there is good agreement between the results of our experiment and those of the Dutch studies by Eggermont (1964), Corthals (1984) and Breeuwer and Plomp (Breeuwer, 1985; Breeuwer and Plomp, 1986). As far as language differences allow, there is also good agreement with the multitude of English studies, at least in the global classifications (Woodward and Barber, 1960; Erber, 1972b; Binnie *et al.*, 1974; Jackson *et al.*, 1976; Walden *et al.*, 1977, 1981; Caplan Hack and Erber, 1982; Owens and Blazek, 1985; Busby *et al.*, 1984, 1988). The most striking difference between English and Dutch is that in English one normally finds more consonantal visemes: at least six, but in post-training situations up to nine visemes are occasionally reported (see the excellent survey by Owens and Blazek, 1985). The difference may simply be explained by articulatory features. In Dutch, (inter)dental fricatives like English /θ,ʒ/ do not appear. Moreover, English comprises a number of consonants which have *lip rounding or protrusion* as a highly visible secondary articulation cue besides the actual place of articulation. This makes *e.g.* such palatal

sounds like /ʃ,tʃ,ʒ,dʒ/, and retroflex /r/, distinguishable from other intra-oral phonemes. In other words, the number of clearly visible nonlabial consonants is much larger in English. In terms of visemes, Dutch may be concluded to distinguish up to four consonantal and four vocalic visemes.

Consonantal visemes

The primary cue for visual discrimination of consonants is *lip articulation.* On the basis of plainly visible lip movements or positions, /p,b,m/ and /f,v,ʋ/ are readily identified by Dutch lipreaders. Within the class of nonlabial phonemes /s,z,ʃ/ is discriminated as a group probably on the basis of *restricted oral cavity opening*, due to the necessary constriction near the front of the oral cavity. It is interesting to note that both Eggermont (1964) and Corthals (1984) mention these front fricatives as a possibly distinct visual set. *Place of articulation*, finally, can only be used to make a crude discrimination between front and back consonants. In general, only the better phoneme lipreaders are able to distinguish the minor differences between *e.g.* alveolar and velar consonants.

It is worthwhile mentioning the two allophonic variants of /ʋ/ that are used in Dutch. Both Eggermont (1964) and Corthals (1984) use the bilabial, or rather rounded, variant [w], as it is current in many languages, including English and French. This rounded feature is typical of the southern Dutch (Eggermont) and Flemish (Corthals) pronunciation, and as such it might constitute a viseme of its own in Dutch. However, the labiodental variant [ʋ] is normally regarded as the standard Dutch form. It was also the variant used by our three speakers. In this sense, phoneme /ʋ/ was logically included in the labiodental viseme.

Dutch /r/ also has (at least) two allophonic variants: alveolar [r] and uvular [R]. It is often argued that alveolar [r] is a highly visible phoneme, certainly if the pronunciation is slightly exaggerated. For that reason it is often used by speech therapists and teachers of the deaf. However, the more usual variant in Dutch is the uvular [R]. Our speakers had not been instructed to use any of the two forms, so they used whatever was their own habit. The two speech therapists used the alveolar form in a number of cases in prevocalic or intervocalic position. In final position [R] was always used. Speaker 3 only used the uvular variant [R]. Consequently, phoneme /r/ was often confused with the back phonemes and not with /l/, which is sometimes reported. Indeed, in the work by Breeuwer (1985) we found /l,r/ as a subset within the set of nonlabial phonemes (see table 2.1); this is undoubtedly due to the

pronunciation of the speech therapist participating in that study.

The alveolar lateral /l/ takes an ambiguous position if one considers the results from the different Dutch studies on visual phoneme recognition. From Eggermont's (1964) study it appears as a subset of its own, Corthals (1984) incorporates it within a group of alveolar and palatal sounds, and according to Breeuwer's (1985) data it is generally confused with alveolar [r]. Finally, our own results classify /l/ with the back rather than the alveolar consonants; moreover, they indicate that its visibility is heavily influenced by its vocalic context. All information taken together, /l/ might perhaps best be classified under the alveolar subset of consonants, with a tendency to be distinguished from them in cases where open vowels permit some view into the oral cavity.

Considering all information from the various Dutch studies, we may conclude that Dutch lipreaders distinguish among four consonantal visemes; the subsets [4a] and [4b] will only be recognized as such by the keener observers:

[1]	/p,b,m/	bilabial consonants
[2]	/f,v,ʋ/	labiodental consonants
[3]	/s,z,ʃ/	nonlabial front fricatives
[4a]	/t,d,n,j,l/	other nonlabial front consonants
[4b]	/k,R,x,ŋ,h/	other nonlabial back consonants

Vocalic visemes

With respect to visually perceived vowels, the primary divide is that between the classes of rounded and unrounded vowels. This is also reported in various English vowel studies (Wozniak and Jackson, 1979; Caplan Hack and Erber, 1982; Busby *et al.*, 1984). From our study we may conclude that there is also an apparent difference between monophthongs and diphthongs. Some reserve may be expressed as to the diphthong /ɛi/, which behaves like a simple front unrounded monophthong in CVC syllables. It is well possible that its diphthong character (especially its transition from open to close vowel) is more obvious in final position, as is actually taught in most Dutch lipreading instruction manuals (*e.g.* van der Zee-Zetstra, 1985). Besides *lip rounding*, also the *vertical degree of lip opening* seems to play a role in visual perception of vowels. This conclusion is confirmed by the results from other studies. The Breeuwer and Plomp data (Breeuwer, 1985; Breeuwer and Plomp, 1986) show that this cue is especially important for distinctions among the unrounded vowels. As mentioned in section 2.1.2, Corthals (1984) classified Dutch vowels into six separate classes on the basis of *lip rounding* and *lip opening* cues. As was shown in our

experiment, this classification is a rather optimistic view of the general situation, but the keener lipreaders indeed tend to see slightly more distinctions among the vowels, especially as rounded vowels are concerned. From our data we concluded that *degree of lip opening* was confounded with a third cue for visual vowel perception, *vowel duration.* This cue plays a role in distinguishing between short and long rounded vowels: /u,y,œ,ɔ/ versus /ø,o/. The use of that cue may be the reason why Corthals' (1984) classification does not agree completely with our data. Keeping in mind that the following is an optimal situation, Dutch vowels may be grouped into four visemes:

[1a] /i,I,e,ɛ/ close and half-close front vowels (unrounded)
[1b] /ɛi,a,ɑ/ half-open and open vowels (unrounded)
[2] /u,y,œ,ɔ/ short back vowels (rounded)
[3] /ø,o/ long back vowels (rounded)
[4] /au,œy/ closing and rounding diphthongs

The distinction between subsets [1a] and [1b] was not found in our group of good phoneme lipreaders, but is a result of taking all Dutch phoneme lipreading data together: /a,ɑ/ forms a subset in the studies by Eggermont (1964), Corthals (1984) and Breeuwer (1985), while our own confusion data indicate that /ɛi/ and /a/ are often confused (/ɛi/ was not used in the studies by Corthals and Breeuwer).

2.4.3 Integration of visual and auditory information

Normally, classifications of lipread consonants and vowels are strictly separated; most experiments are designed in such a way that the viewer knows whether he has to report a consonant or a vowel. However, the number of consonantal and vocalic visemes cannot simply be added: our four consonantal and four vocalic visemes will probably not add up to eight Dutch visemes. It is by no means a clear case that lipreaders are consistently able to tell the difference between syllables of the types VCV, CVC, VCVC, etc. The syllabic organization of spoken syllables, words and sentences is not readily available through purely visual perception (Risberg, 1974; Summerfield, 1983). It is probably one of the reasons why there is no straightforward relation between lipreading of phonemes and (auditory-visual) lipreading of longer speech segments. Observations like these indicate the limited role that pure lipreading plays in speech processing by many hearing impaired. In order to achieve good understanding of spoken messages in everyday conversation, the (profoundly) hearing impaired should rely on auditory as well as on visual perception. The complementary relationship between

the modalities of audition and vision has often been stressed (Binnie *et al.*, 1974; Erber, 1974b; Summerfield, 1983). Much important information in speech is not easily or not at all accessible through the visual sense, *e.g.* consonant voicing and prosodic information, both present in the f_0-signal, and consonant manner of articulation, which is related to various acoustic cues, both spectral and temporal. On the other hand, the cues that are available purely from lipreading have also been established, in our experiment as well as those of many others: consonant place of articulation (although this appears to be a much stronger cue in English than in Dutch), which is acoustically related to spectral differences, oral fricative information, related to high-frequency spectral energy and relatively long duration, and lip rounding and degree of lip opening in vowels, which cues in Dutch are roughly related to the second and the first formant, respectively.

In view of the aforementioned rehabilitation of the profoundly hearing impaired by means of speech processing hearing aids (see section 2.1.2), it is noted that speech processing and coding strategies may be based on the auditory and visual classifications of Dutch phonemes as have been established by now. If one chooses to supplement lipreading by means of suitable auditory information, feature extraction and signal recoding may be restricted to those cues that cannot be perceived very well through the visual sense alone. The way in which this is done largely depends on the auditory information that can be handled by the profoundly hearing impaired (van Son *et al.*, 1993a; next chapter).

2.4.4 General conclusions

1. Dutch lipreaders are able to recognize four consonantal and four vocalic visemes.
2. Consonants are classified on the basis of *lip articulation, degree of oral cavity opening* and *place of articulation.* In the acoustic domain these features are mainly related to relatively high-frequency spectral information.
3. Vowels are classified on the basis of *lip rounding, vertical degree of lip opening* and *vowel duration.* Lip rounding is globally related to the second formant, lip opening to the first formant. Vowel duration is a temporal cue.
4. Dutch viseme classifications are not influenced by lipreading expertise, relating to explicit knowledge about lipreading and a certain degree of lipreading proficiency in everyday spoken communication.

5. Even though experienced and skilled lipreaders do not recognize more visemes than other lipreaders, they tend to show slightly better phoneme identification.
6. There are generally no speaker effects on identification nor on confusion patterns.
7. A positive context effect on phoneme identification is observed when the view into the oral cavity is not obstructed. This is the case with open vowel /a/, and nonlabial consonants /t,k/. No dramatic effects were found on confusion patterns.
8. Dutch and English viseme classifications mainly differ in the number of consonantal visemes that can be observed (4 for Dutch and 6 to 7 for English).

Acknowledgement

We thank Carlien Theewis and Marieke van Hal for their contribution to this study. As former students at our laboratory they collected most of the data, computed part of the results and brought their own ideas into the discussion. They described their work in two internal papers available from our department.

Chapter 3.

The perception of complex harmonic patterns by profoundly hearing impaired listeners

This chapter is a slightly adapted version of a paper by N. van Son, A.J. Bosman, P.J.J. Lamoré and G.F. Smoorenburg, Audiology 32, 1993a, 308-327.

Abstract

When providing profoundly hearing impaired persons with processed speech through a signal processing hearing aid, it is important that the new speech code matches their auditory capacities. This processing capacity for auditory information was investigated in this study. In part I, the subjects' ability to judge similarities among eight different but related harmonic complexes was studied. The patterns contained different numbers of harmonics to a 125 Hz fundamental frequency; the harmonics had been spread over the spectrum in various ways. The perceptual judgments appeared to be based on a temporal cue, beat strength, and a spectral cue, related to the balance of high and low frequency components. In part II, three sets of synthetic vowels were presented to the subjects. Each vowel was realized by summing harmonically related in-phase sinusoids at two formant frequencies. The sets differed in the number of sinusoids per formant: one, two or three. It was found that the subjects used spectral cues and vowel duration for differentiating among the vowels. The overall results show the limited but perhaps usable ability of the profoundly impaired ear to handle spectral information. Implications of these results for the development of signal processing hearing aids for the profoundly hearing impaired are discussed.

3.1 Introduction

In this chapter an experimental study is described in which we investigated the extent to which profoundly hearing impaired persons can discriminate and identify different harmonic complexes. Part I (sections 3.2 up to and including 3.4) deals with their ability to discriminate among eight different but related harmonic complexes. More specifically, we were interested in the perceptual dimensions that profoundly hearing impaired listeners use in assessing similarities among these patterns. In the discussion of the results we will dwell upon some aspects of the signal processing capacity of the profoundly impaired hearing, as well as the relevance of the perceptual dimensions found in this laboratory study

to speech coding schemes for signal processing hearing aids. In the second part (sections 3.5 up to and including 3.7), which is strongly tied up with part I, we presented the subjects with three sets of synthetic vowels, generated by summing harmonically related sinusoidal components at the frequencies of the first two vowel formants. In defining the different vowel sets, we applied some of the findings of part I. In this experiment, we investigated the subject's ability to identify these highly simplified vowels. As with the non-vowel harmonic patterns, we were mainly interested in the underlying perceptual dimensions on which the listeners based their identification.

3.2 Similarities among different harmonic complexes

Sections 3.3 and 3.4 are based on the results from three experimental sessions, run within a period of ten months. The sessions were similar to each other as far as experimental method is concerned. In the second and third sessions some improvements were made to the procedure of loudness equalization of the stimuli. These alterations will be described in the sections below, together with the methods that were applied in all three sessions.

3.3 Methods

3.3.1 Stimuli

Figure 3.1 shows schematic spectra of the eight auditory patterns used in this study. Pattern 1 is a summation of 12 sine waves, the frequencies of which are multiples of the fundamental frequency of 125 Hz. All sine waves are in sine phase. The other signals were derived by removing certain components from this pattern. All signals were designed to elicit a pitch equal to that of a 125 Hz sine wave.

The patterns were constructed by varying three factors:

(1) *Number of sine wave components*, ranging from 2 to 12.
(2) *Spectral distribution.* Components were spread regularly over the spectrum (signals 1 and 2), or in a grouped fashion, either using two 'peaks' (signals 3, 4, 5, and 6) or only one (signals 7 and 8).
(3) *Presence of adjacent harmonics.* Spectral 'peaks' were represented either by adjacent harmonics (signals 3, 5, 7 and 8) or by single harmonics (4 and 6).

The eight patterns were matched for loudness (for the procedure see section 3.3.3, below). The steady-state part of each stimulus was 300 msec, with a 100 msec cosine-shaped onset and offset.

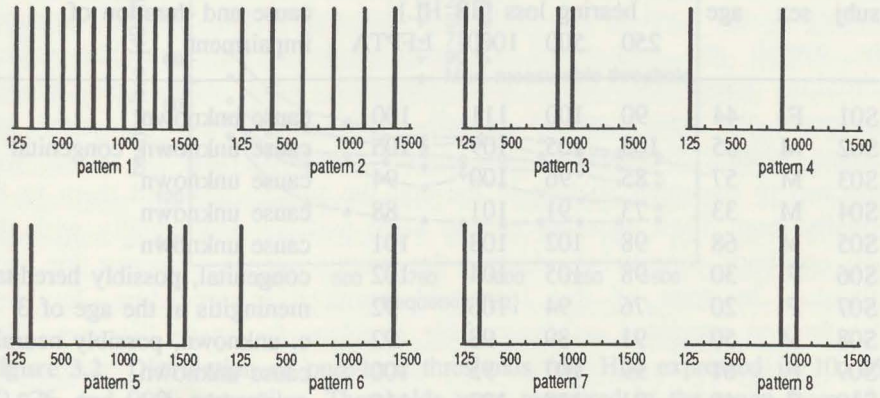

Figure 3.1 Schematic spectra of 8 harmonic patterns. The frequencies of all harmonics are multiples of a fundamental of 125 Hz.

3.3.2 Subjects and their dynamic ranges

Ten sensorineurally hearing impaired subjects participated in the first two sessions. One of them did not participate in session 3, where another subject took his place. Subjects were paid for their services. Table 3.1 gives some relevant data about the subjects. Tympanometry indicated that middle ear functions were normal, with the exception of subject S04, who had middle ear problems as a result of a cold during the time of session 3. All subjects are regular hearing aid wearers, but in this study the aid was not used.

In session 1, pure-tone thresholds were measured in both ears for tones between 125 and 1500 Hz, at multiples of 125 Hz. A simple *up-down* procedure was applied (staircase method; Levitt, 1971), and the threshold was defined as the arithmetic mean of the last six out of nine reversals. After this measurement, the best ear was selected for measurement of the loudness discomfort level (LDL) and for further experimentation. In sessions 2 and 3, thresholds and LDLs were measured in this ear only. In this way, pure-tone thresholds were measured for ten subjects in three sessions. In figure 3.2 the distribution of these thresholds is expressed in 10, 25, 50, 75 and 90% percentiles (dB HL).

Table 3.1 Relevant data of 11 subjects. Hearing loss is expressed in the losses (dB HL) at three frequencies and the pure-tone average for these three frequencies (low-frequency PTA or LFPTA).

subj	sex	age	hearing loss [dB HL] 250	500	1000	LFPTA	cause and duration of impairment
S01	F	44	90	100	111	100	cause unknown
S02	M	35	103	105	109	105	cause unknown, congenital
S03	M	57	85	96	100	94	cause unknown
S04	M	33	73	91	101	88	cause unknown
S05	M	68	98	102	103	101	cause unknown
S06	F	30	98	105	104	102	congenital, possibly hereditary
S07	F	20	76	94	106	92	meningitis at the age of 3
S08	M	50	91	89	98	92	c. unknown, possibly neural
S09	F	61	99	107	95	100	cause unknown
S10	F	40	91	118	122	110	meningitis at the age of 2
S11	M	23	98	107	116	107	cause unknown, congenital

Subject S08 participated in sessions 1 and 2, and subject S11 participated in session 3 only; all other subjects participated in all three sessions.

Loudness discomfort was measured by presenting a series of tones with increasing intensity, using 1 dB steps. The series of tones was stopped when the subject indicated that loudness became uncomfortable. That point was taken as the estimate of the LDL. In most cases LDL could not be established as the headphone (TDH 39P) output was not sufficient (maximum level about 140 dB SPL; calibrated with B&K artificial ear, type 4152, and B&K measuring amplifier, type 2608). In figure 3.2, this maximum level is indicated by the bottom graph. For one subject, S02 in table 3.1, the dynamic range at the two lowest frequencies was extremely narrow; in session 3 it was zero, because the threshold could not be established.

After measuring a subject's dynamic range, the stimuli were defined for that individual according to three criteria:

(1) all components of a pattern had to fall into the dynamic range;
(2) the total pattern should not exceed the loudness discomfort level;
(3) all patterns had to be equally loud.

Especially the third was of great concern to us. We were only interested in the subjects' capacity to process signal waveforms that differed in spectral content and the associated different temporal shapes. We therefore wanted them to base their responses solely on spectral cues.

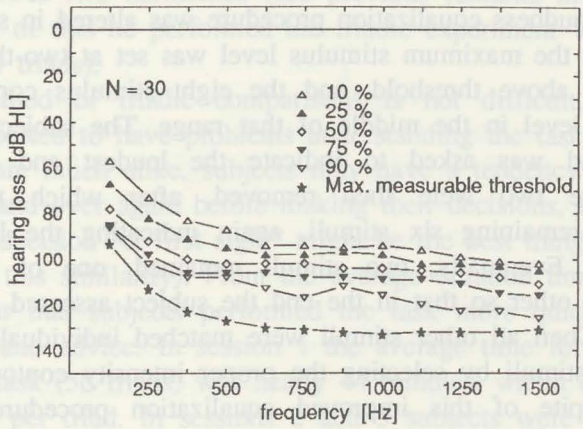

Figure 3.2 Distribution of pure-tone thresholds (dB HL) expressed in 10, 25, 50, 75, and 90% percentiles. Thresholds were measured in the range from 125 to 1500 Hz (30 measurements in three sessions).

3.3.3 Equally loud stimulus patterns

In session 1, loudness equalization was done according to the following procedure. First, the stimulus intensity contour, as described by the levels of the individual components, was set in the middle of the dynamic range. If the loudness of the stimulus was uncomfortable for the subject, the contour was situated at a lower level, reducing each component proportionally to the dynamic range at that frequency, until the loudness was acceptable. Finally, the subject compared the selected patterns and was asked to indicate if any one pattern was clearly louder or softer than the other patterns. In that case, the outlying signal was computed anew by determining a new contour.

The same procedure was applied in session 2, but this time subjects also had to rate the loudness of each stimulus on a 10-point scale: the louder the stimulus, the higher the mark. This loudness rating procedure served as a check on the equalization procedure. If equalization was fully successful, the loudnesses of all stimuli would be rated equally. However, in some subjects the entire stimulus shaping procedure appeared problematic.

Two problems occurred: either a particular stimulus remained louder than other ones, even if the intensity contour was at a minimum level (just above threshold); or, on the other hand, a stimulus remained softer

than other ones, even if the intensity contour was at the maximum level allowed in session 2 (in the middle of the dynamic range). As a consequence, the loudness equalization procedure was altered in session 3.

In session 3 the maximum stimulus level was set at two-thirds of the dynamic range above threshold, and the eight stimulus contours were defined at the level in the middle of that range. The subject compared the stimuli and was asked to indicate the loudest and the softest stimulus. These two were then removed, after which the subject compared the remaining six stimuli, again indicating the loudest and softest stimuli. Eventually, two stimuli remained, one of which was matched to the other so that in the end the subject assessed them to be equally loud. Then all other stimuli were matched individually with the two reference stimuli by selecting the proper intensity contour for each stimulus. In spite of this improved equalization procedure, loudness equalization remained an insuperable problem for some listeners. This was mainly due to their own restricted dynamic range and consequently to the restricted range of stimulus intensities.

3.3.4 Design and procedure

The task of the subject, explained on a printed instruction form, was to give assessments of similarity/dissimilarity among the eight stimuli. These ratings were acquired by means of triadic comparisons (Levelt *et al.*, 1966). Altogether 56 triads were presented, which is the number of combinations of 3 out of 8 elements. In this set each stimulus occurred 21 times and each pair of stimuli occurred 6 times. In each triad the subject had to indicate, in this order, which were the two *most similar* and which were the two *least similar* stimuli of the three. The triads were presented in a random order to each of the subjects. They knew that the stimuli were different, but they were not aware of the nature of the differences.

Stimuli were generated by means of a digital audio signal processor (van Raaij, 1986) and replayed through a Philips F4220 power amplifier and TDH 39P headphones in Peltor circumaural cushions. The subject was seated in a sound-insulated booth. The stimuli were presented monaurally to the best ear. In sessions 1 and 2 the subject called up the stimuli and gave responses using a specially designed triangular response box. In session 3 the procedure was carried out on the keyboard of the PC controlling the experiment. Subjects were allowed to listen to each stimulus as long as they deemed necessary. The ratings of each subject were gathered in a dissimilarity matrix.

In session 3, subject S02 did not perceive pattern 7, which contains only the lower two harmonics (see previous remarks in section 3.3.2). As a result of this he performed the triadic experiment with only seven patterns (35 triads).

The method of triadic comparisons is not difficult. None of the subjects reported to have problems understanding the task. When stimuli in a triad are much alike, subjects may have a tendency to listening to them over and over again before making their decisions, while in such a situation a decision "at first sight" might be the best thing to do (exactly because of this similarity). From the average duration time of a session, it was clear that subjects performed the task more quickly after some experience and advice. In session 1 the average time for completion of the triadic task (56 triads) was nearly 44 minutes, which comes down to 47 seconds per triad. In sessions 2 and 3 subjects were advised not to linger too long over one triad, because it makes the complete task tiring; in these sessions completion times were, respectively, 31 minutes (33 sec/triad) and 26 minutes (28 sec/triad).

3.3.5 Processing (dis)similarity assessments

The assessments in the triadic comparisons were computed as dissimilarity ratings, *i.e.* a 'distance' of 2 units was allotted to the pair of least similar stimuli, a distance of 0 units was given to the most similar stimuli and 1 unit distance was given to the remaining pair. On this basis a dissimilarity matrix was computed for each subject in each of the three sessions. Twenty-nine matrices were analysed by means of an INDSCAL technique according to Carroll and Chang (1970), so as to represent the dissimilarities in an *n*-dimensional space. The third dissimilarity matrix of subject S02 was excluded from this analysis, as it had been based on similarity assessments of seven patterns only (see previous section).

Using INDSCAL, one normally seeks a minimal number of (interpretable) dimensions accounting for a maximal proportion of variance. As will be explained in the next section, a two-dimensional representation gave satisfactory results; a three-dimensional representation raised the proportion of accounted variance only marginally, while the extra dimension did not lead to new insights into the psychophysical relationship between stimuli and subjects.

3.4 Results and discussion

3.4.1 Reliability of (dis)similarity assessments in three sessions

The INDSCAL analysis of twenty-nine dissimilarity matrices, ten from sessions 1 and 2, and nine from session 3, yielded the two-dimensional object space and condition space plotted in figure 3.3. The object space gives the overall perceptual distances among the eight patterns on the basis of two independent dimensions. The condition space shows the relative weight that subjects in three sessions attach to each of the two dimensions; because this space is based on both subjects and sessions, we prefer calling it 'condition' space instead of merely 'subject' space. The two dimensions explain 62.7% of the total variance in the data (dimension I 37.4% and dimension II 25.3%).

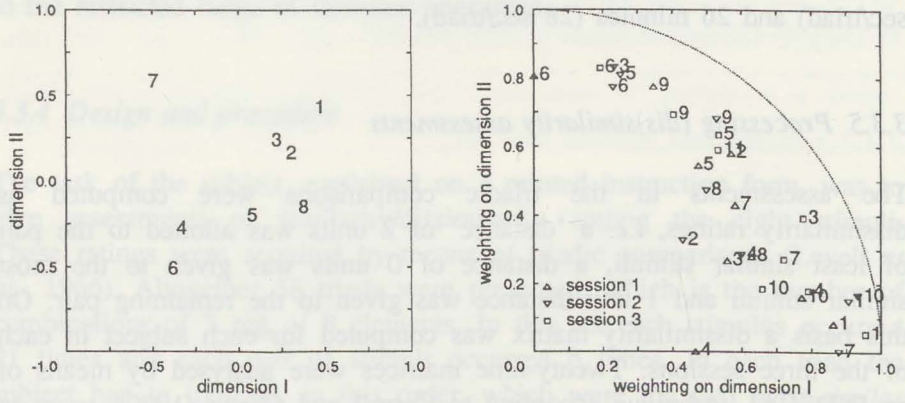

Figure 3.3 Two-dimensional INDSCAL solution, based on pattern dissimilarities of 29 subject measures in three experimental sessions. The two dimensions explain 62.7% of the total variance (dim.I 37.4%, dim.II 25.3%). Left panel: object space of eight complex harmonic patterns. Right panel: condition space of 11 subjects accounting for 29 measures in three sessions (ten subjects in sessions 1 and 2, nine in session 3).

Figure 3.3 shows the overall situation, based on three sessions, but the solutions from individual sessions were not identical. Therefore, in figure 3.4 the object and subject spaces pertaining to the individual sessions are shown. In order to give a quantitative measure of the equality of both dimensions in the three sessions and in the overall solution, we used the following procedure. For each dimension independently, the positions of patterns in the object spaces of the three sessions (expressed

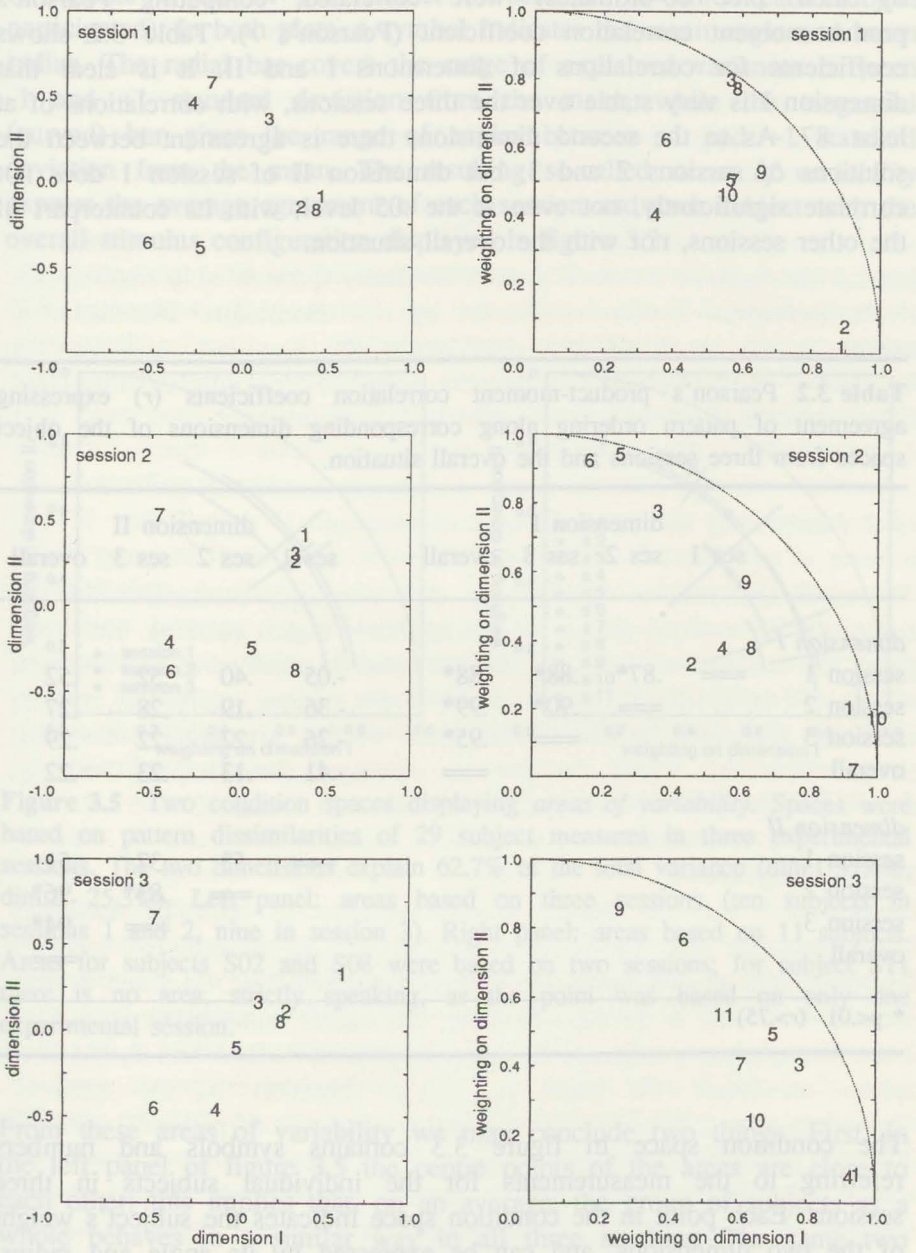

Figure 3.4 Two-dimensional object and subject spaces for three sessions. Session 1: configuration based on measures of ten subjects; explained variance is 63.5% (dim.I 37.2%, dim.II 26.3%). Session 2: configuration based on measures of ten subjects; explained variance is 69.6% (dim.I 40.2%, dim.II 29.4%). Session 3: configuration based on measures of nine subjects; explained variance is 67.8% (dim.I 44.5%, dim.II 23.3%).

by their plot co-ordinates) were correlated, computing Pearson's product-moment correlation coefficient (Pearson's *r*). Table 3.2 shows coefficients for correlations of dimensions I and II. It is clear that dimension I is very stable over the three sessions, with correlations of at least .87. As to the second dimension, there is agreement between the solutions of sessions 2 and 3, but dimension II of session 1 does not correlate significantly, not even at the .05 level, with its counterpart of the other sessions, nor with the overall situation.

Table 3.2 Pearson's product-moment correlation coefficients (*r*) expressing agreement of pattern ordering along corresponding dimensions of the object spaces from three sessions and the overall situation.

	dimension I				dimension II			
	ses 1	ses 2	ses 3	overall	ses 1	ses 2	ses 3	overall
dimension I								
session 1	===	.87*	.88*	.88*	-.05	.40	.52	.52
session 2		===	.90*	.99*	-.36	.19	.28	.27
session 3			===	.95*	-.26	.22	.22	.29
overall				===	-.41	.13	.23	.22
dimension II								
session 1					===	.53	.37	.51
session 2						===	.83*	.96*
session 3							===	.94*
overall								===

* p<.01 (*r*>.75)

The condition space in figure 3.3 contains symbols and numbers referring to the measurements for the individual subjects in three sessions. Each point in the condition space indicates the subject's weight of the two dimensions, and can be expressed by its angle and radius. The angle indicates the position relative to the two dimensions, while the squared radius gives the proportion of variance explained by both dimensions. In figure 3.5 the same condition space is displayed, but instead of 29 session/subject points, we have indicated in two condition spaces the average positions of three sessions (arithmetic means of ten or nine subjects per session) and the average positions of eleven subjects

(arithmetic means of the number of sessions in which each subject participated). In both plots, a symbol indicates the mean angle and mean radius. The radial bar covers the range of explained variances between -1 and +1 standard deviation from the mean, while the tangential (curved) bar gives the range of angles between -1 and +1 standard deviation from the mean. The resulting so-called *areas of variability* express the average agreement of each session and each subject with the overall stimulus configuration displayed in figure 3.3.

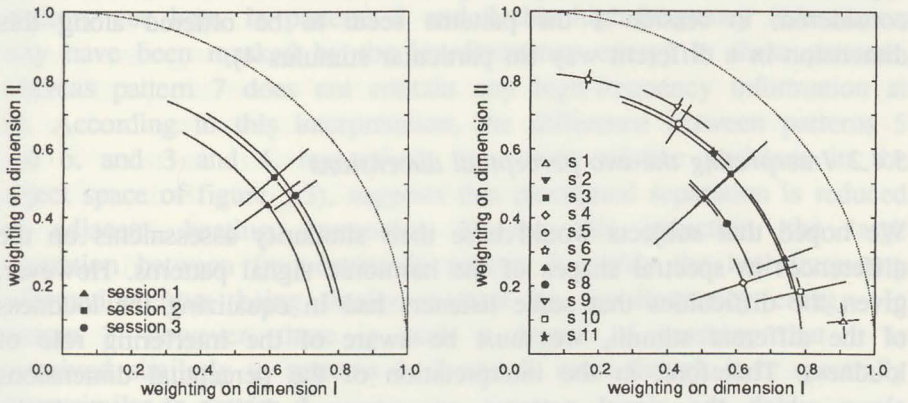

Figure 3.5 Two condition spaces displaying *areas of variability*. Spaces were based on pattern dissimilarities of 29 subject measures in three experimental sessions. The two dimensions explain 62.7% of the total variance (dim.I 37.4%, dim.II 25.3%). Left panel: areas based on three sessions (ten subjects in sessions 1 and 2, nine in session 3). Right panel: areas based on 11 subjects. Areas for subjects S02 and S08 were based on two sessions; for subject S11 there is no area, strictly speaking, as the point was based on only one experimental session.

From these areas of variability we may conclude two things. First, in the left panel of figure 3.5 the centre points of the areas are close to each other: this implies that, on an average, the group of subjects as a whole behaves in a similar way in all three sessions. The same two dimensions emerge from all three sessions, and with roughly the same proportions of explained variance. The fact that the mean points of sessions 2 and 3 are slightly closer to the quarter circle than that of session 1 indicates a trend of a better fit in the two measures with better loudness equalization. Secondly, the right panels of figures 3 and 5 show that most of the subjects behave in a consistent way over the

three sessions. However, subjects S03, S05 and S07 tend to change the importance that they attach to each dimension over the three sessions. Especially in session 2 they seem to change their opinions, giving more weight to dimension II (subjects S03 and S05) or dimension I (subject S07). These changes could not be attributed to specific differences in the spectral contents of some individual stimuli in session 2, such as, for example, more high-frequency energy for subjects S03 and S05.

All in all, the results indicate that dimension I is quite stable. In general, the subjects assessed pattern similarity quite reliably making use of the perceptual cue(s) represented by this dimension. Dimension II appears to be used reliably only when the last two sessions are considered; in session 1 the patterns seem to be ordered along this dimension in a different way (in particular stimulus 4).

3.4.2 Interpreting the two perceptual dimensions

We hoped that subjects would base their similarity assessments on the differences in spectral shapes of the harmonic signal patterns. However, given the difficulties that some listeners had in equalizing the loudness of the different stimuli, we must be aware of the interfering role of loudness. Therefore, in the interpretation of the perceptual dimensions along which the signal patterns are arranged three aspects will be discussed: temporal aspects, spectral aspects, and loudness.

Temporal aspects

Dimension I can be related to the presence of adjacent frequency components in the higher frequency region as a cue for auditory discrimination, as this dimension separates patterns 1, 2, 3, 5 and 8 from patterns 4, 6 and 7. Strictly speaking, the higher components in pattern 2 are not as closely adjacent as those in patterns 1, 3, 5 and 8, but they appear to create the same perceptual sensation in the subjects. Conversely, the low-frequency adjacent components of pattern 7 evoke a different percept in the subjects, one that is highly similar to patterns 4 and 6. The exact nature of the percept of adjacent high-frequency components is a matter of conjecture; to normally hearing listeners (colleagues at our laboratory) these patterns sounded rather rough or 'rattle-like', while patterns 4, 6 and 7 sounded clearly smooth and 'flute-like', like the sound from organ pipes. Adjacent harmonics lying within a critical band may create *beat-like* sensations to the ear that are due to temporal cues, audible temporal fluctuations, rather than spectral cues, as we suppose that a profoundly impaired ear cannot resolve

high-frequency components lying only 125 Hz apart. So, the presence of beats, resulting from two adjacent frequency components in the high-frequency region of the spectrum, seems an important cue in auditory discrimination by the profoundly hearing impaired.

Spectral aspects

The perceptual cue represented by dimension II appears to involve the processing of spectral information. From the left panel of figure 3.3 one may recognize a divide between patterns 1, 2, 3 and 7 on the one hand, and patterns 4, 5, 6 and 8 on the other hand. It appears that the perceptual presence of high-frequency components is the important auditory cue here. In patterns 1 and 2 this high-frequency information may have been masked by the low-frequency energy in these patterns, whereas pattern 7 does not contain any high-frequency information at all. According to this interpretation, the difference between patterns 5 and 6, and 3 and 4, respectively (see their relative positions in the object space of figure 3.3), suggests that perceptual separation is reduced by adjacent, beating, harmonics. Despite this reduction, the large separation between frequencies in pattern 5 avoids the high-frequency information from being totally masked by low-frequency energy. In pattern 3, however, there is such a degree of masking that it is perceived similarly to patterns 1, 2 and 7, even though it is spectrally more similar to pattern 4.

It is interesting to note here that a three-dimensional INDSCAL solution of the three sessions yields a second dimension that can be related to spectral information. Dimension I is similar to the one described above (explained variance 37.1%) and relates to the presence of adjacent high-frequency components producing beats, while perceived separation of low and high components is represented by dimension III (explained variance 15.5%). The extra dimension II (19.4% explained variance) is related to the presence of low-frequency energy: either it is relatively strong, as in patterns 1, 2, 3, 5 and 7, or it is relatively weak (patterns 4 and 6) or fully absent (pattern 8). We have not been concerned too much with this three-dimensional solution for two reasons. First, the proportion of variance explained by the three dimensions did not increase substantially, from 62.7% with two dimensions to 72.0% with three. Secondly, and most importantly, both dimensions seem to be related, at least partly, to the same spectral cue: clear presence of energy in a certain frequency region. In dimension II the focus is on the presence of low-frequency energy, irrespective of higher spectral components, while in dimension III the importance of separation of low and high-frequency components is stressed.

Loudness

The comments of a few subjects during the loudness equalization procedure suggested that they confounded loudness and timbre: they perceived certain signal patterns as louder than others because these "sounded sharper" or because they "sounded rough". It has been observed before that the more different two sounds are, the more difficult it is to match their loudness (see *e.g.* Zwicker *et al.*, 1957).

The loudness ratings of session 2 had been collected to give some insight into the importance of the loudness percept. With very large critical bandwidths, to be expected in our profoundly hearing impaired subjects, loudness is independent of the spectral distribution of components, but is rather related to the amount of energy in the signal. This, in its turn, is related to both the number of components in the signal and the subject's dynamic range. Pearson's correlation coefficients were computed, expressing the agreement between loudness ratings of individual subjects and pattern positions along each dimension in the solution of session 2 (centre panel of figure 3.4). Only three out of the ten subjects in session 2 show a significant correlation ($p<.01$) between their loudness ratings of the eight patterns and the patterns' positions along dimension I (subjects S01 and S10, $r=.93$ and $r=.97$, respectively) or dimension II (subject S09, $r=-.85$).

From these results in session 2 we may conclude that in only three subjects it is not totally clear whether spectral shape or loudness was the cue on which they based their judgment. Even though the stimuli had been matched for loudness, it may have played an interfering role in their similarity judgments.

3.5 Identification of synthetic two-formant vowels

It was not immediately clear to us in how far the important 'beat cue' might be applied in speech pattern coding. To our own (normally hearing) ears synthetic vowel spectra with two-component formant representations lend a more speech-like quality to the vowel than spectra with simple one-component formants. We also had the impression (but so far we have not checked this) that this led to better identification. Although we had no inkling whether our profoundly hearing impaired subjects would experience a clear difference in sound quality as well, the interesting question was whether the difference between one-component and two-component formant representation may indeed lead to better vowel identification scores. In order to answer this

question we presented the subjects with synthetic vowel spectra, in which we used different representations of vowel formants F_1 and F_2.

3.6 Methods

3.6.1 Subjects

The vowel identification task was run at the end of the third session of the pattern similarity assessment experiment. Most of the subjects involved in that session participated in the vowel identification task, with the exception of subject S10 (table 3.1); so, altogether nine subjects were presented with the vowel stimuli, *viz.* subjects S01 to S07, S09, and S11.

3.6.2 Stimuli

Twelve two-formant Dutch vowels were generated by summing harmonically related in-phase sinusoids of the proper formant frequencies. These frequencies had been adopted from Pols *et al.* (1973) and were slightly changed in order to provide harmonic relationship to the spectrally absent fundamental of 100 Hz. All formant values remained within the standard deviation boundaries that Pols *et al.* give (sd_{F1}=57 Hz; sd_{F2}=134 Hz); table 3.3 gives the two formant values for the 12 vowels. Three sets of 12 vowels were created, differing as to the number of sinusoidal components used to represent the formant. In set 1 each formant was simply represented by one component. In set 2 a second component was added to the higher frequency side, and in set 3 two components were added, one on either side of the original component. The vowel spectra were not shaped to the individual subjects' dynamic ranges; all component levels were equal in terms of dB SPL. By and large, this choice matched the average dynamic range which is, to a first approximation, flat in dB SPL (not in dB HL, as was shown in figure 3.2). Each vowel started and ended with a 50 ms cosine-shaped onset and offset, and the steady states lasted 100 ms for short vowels /I,ɛ,ɑ,ɔ,œ/ and 200 ms for long vowels /i,e,a,o,u,y,ø/ (see table 3.3). As an example, figure 3.6 shows the spectra of the Dutch vowels /i/, /u/ and /a/ in all three sets.

Table 3.3 Formant frequencies [Hz] of 12 monophthong Dutch vowels, and their phonological classification with respect to the features of *vowel place*, *lip opening*, *lip rounding*, and *vowel duration*.

vowel	formant		place	opening	rounding	duration
	F_1	F_2	front/back	open/close	round/spread	long/short
/i/	300	2300	+	+	+	+
/ɪ/	400	2000	+	+	+	+
/e/	400	2100	+	+	+	+
/ɛ/	600	1700	+	+	+	+
/a/	800	1300	+	+	+	+
/ɑ/	700	1100	+	+	+	+
/ɔ/	600	900	+	+	+	+
/o/	500	900	+	+	+	+
/u/	300	800	+	+	+	+
/y/	300	1700	+	+	+	+
/ø/	400	1500	+	+	+	+
/œ/	500	1400	+	+	+	+

Formant frequencies were adapted from Pols *et al.* (1973) so that each spectral component is harmonically related to the (spectrally absent) fundamental of 100 Hz. Duration applies to the steady-state portion of the vowel.

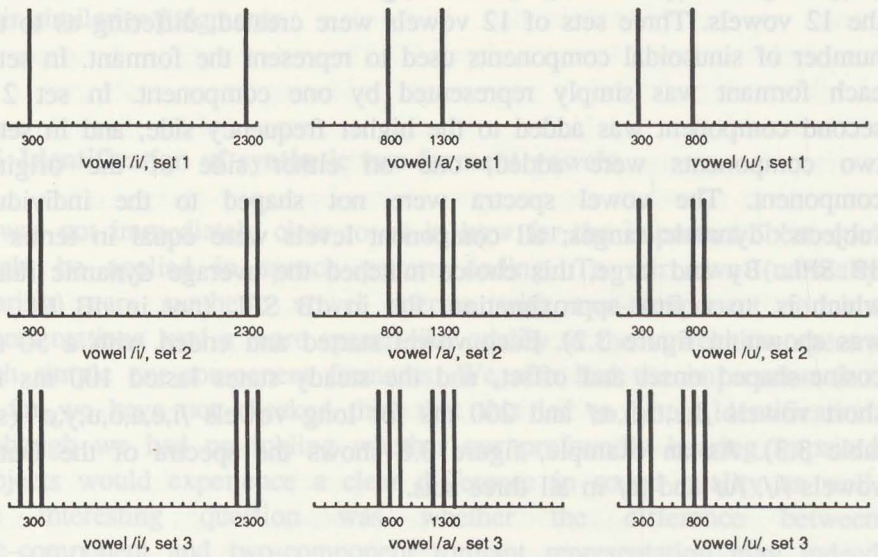

Figure 3.6 Vowel spectra of Dutch /i/, /a/ and /u/ in three vowel sets.

3.6.3 Procedure

Vowel stimuli were generated by means of a digital audio signal processor and replayed through a Philips F4220 power amplifier and TDH 39P headphones in Peltor circumaural cushions. The subject was seated in a sound-insulated booth. The complete set of 36 vowels was presented twice to the subject (monaurally to the best ear), in two different random orders. The experiment was run in a 12-alternative forced-choice format. The subject listened to each vowel and then had to indicate on a list of printed vowels the one he had just perceived. Altogether 648 stimuli were presented (12 vowels by 3 sets by 2 runs by 9 subjects), only six of which were not responded to. Stimuli and responses were compared and confusion matrices were filled for each subject for each vowel set. The nine individual matrices of each set were analysed by means of the above mentioned INDSCAL technique (Carroll and Chang, 1970).

3.7 Results and discussion

3.7.1 Identification scores

Figure 3.7 presents mean overall identification scores with standard deviations and score range for each vowel set (left-hand panel). A two-factor analysis of variance showed statistically significant differences for subjects, $F(8,16)=3.7$, $p<.01$, and vowel sets, $F(2,16)=5.8$, $p<.01$, showing that sets 2 and 3 provide clearly better vowel intelligibility than set 1.

We also analysed scores for individual subjects on the phonetic vowel features of *vowel place*, *lip opening*, *lip rounding*, and *vowel duration*. Table 3.3 gives the classification used for these features. For each feature individually, scores were analysed by means of a three-factor analysis of variance (vowel set by feature by subject). The right-hand panel of figure 3.7 gives all feature scores, averaged over nine subjects, for the three vowel sets.

Three significant differences were found:

(1) Both short and long vowels in sets 2 and 3 are better identified than those in set 1: $F(2,16)=7.3$, $p<.01$.

(2) With sets 2 and 3, the degree of lip opening is better identified than with set 1: $F(2,16)=11.4$, $p<.01$.

(3) In the lip opening feature, there is a significant interaction, $F(2,16)=16.0$, $p<.01$, between vowel set and degree of lip opening:

close vowels have higher scores than open vowels in set 1, whereas in sets 2 and 3 the situation is contrary.

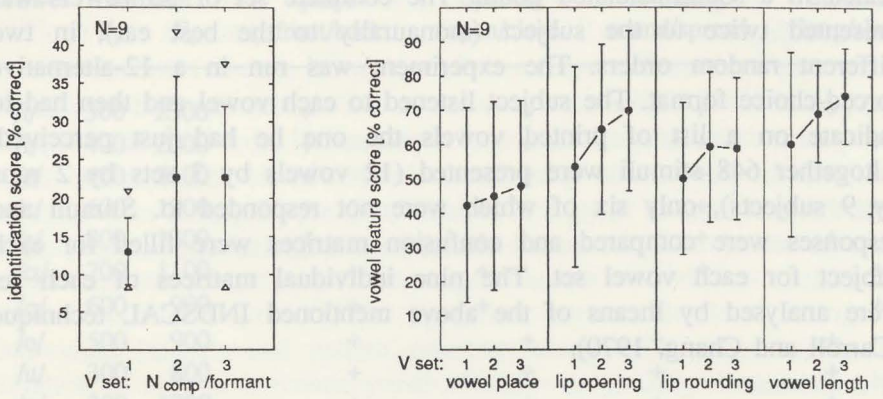

Figure 3.7 Left panel: vowel identification scores in three sets. The plot gives mean scores (filled circles), standard deviations (vertical lines), and lower and upper points of score range (triangles). Right panel: vowel feature scores for three sets and four features. Mean scores (filled circles) and standard deviations (vertical bars) are given. All scores are averaged over nine subjects.

We may conclude from these statistics that vowels with two-component or three-component formants are better identified than those with only one component per formant. The results suggest that more components in the vowel spectrum lend the signal a more vowel-like quality and thus facilitate vowel identification. Analysis with respect to vowel features shows that increased identification may be related to a better identification on the lip opening feature and vowel duration. Lip opening is partly associated with lower frequency information available from the first formant, whereas vowel duration is clearly a temporal cue. Both features have been reported to be available to persons with even severe hearing impairments (Erber, 1972a; Boothroyd, 1984; Shannon *et al.*, 1992).

3.7.2 Perceptual dimensions

Twenty-seven individual vowel confusion matrices were filled for the nine subjects and the three vowel sets. On this basis similarities among

vowels were analysed using INDSCAL (Carroll and Chang, 1970). The resulting one-dimensional solutions explained about 80% of all variance in the data, but inspection of the plots showed that this dimension could represent all three physical parameters, especially F_1 and F_2. In order to unravel this intricate perceptual-physical relationship, we opted for two-dimensional solutions, explaining up to 90% of the variance (figure 3.8). In these solutions one dimension related to spectral, and the other to temporal information. The left-hand panels of figure 3.8 show the object spaces of the three vowel sets, while the right-hand panels give the corresponding condition spaces. Pearson's *r* was computed between the plot co-ordinates of the vowels on each dimension and the parameter values of F_1, F_2 and vowel duration. Table 3.4 gives the *r* values for all three vowel sets. Dimension I still relates to both formants, the relation to F_2 being somewhat stronger. Dimension II is strongly related to vowel duration in sets 2 and 3, whereas this relation is rather weak in set 1. For comparison, table 3.4 also shows Pearson's *r* for correlations between plot co-ordinates of the one-dimensional solution and the three parameter values.

How can we interpret the (cor)relations ? It is clear that dimension I represents a spectral cue, namely vowel formants. The relation to F_2, however, should not be interpreted by assuming that our profoundly hearing impaired subjects are able to perceive differences in this relatively high frequency region better than those in the lower first formant region. It is more likely to be a case of a difference between vowel spectra with two formants (F_1 and relatively low F_2) and vowel spectra with only or mainly one formant (F_1; F_2 is too high to be perceived). The perceptual basis for this difference is therefore probably the absence or presence of F_2, not the actual place of the second formant.

Dimension II is clear-cut, although the interpretation of Pearson's *r* suffers somewhat from the fact that vowel duration is a two-valued variable, which places some restrictions on the possible values that the coefficient may assume. Nevertheless, it is clear that short vowels are generally placed to one end of the object spaces of vowel sets 2 and 3, while the long vowels are generally placed to the other end. In set 1 the situation is not clear: the dimension holds a weak relation to both formants ($p<.05$), whereas especially the positions of long vowels /a,u,ø/ make the relation to vowel duration insignificant.

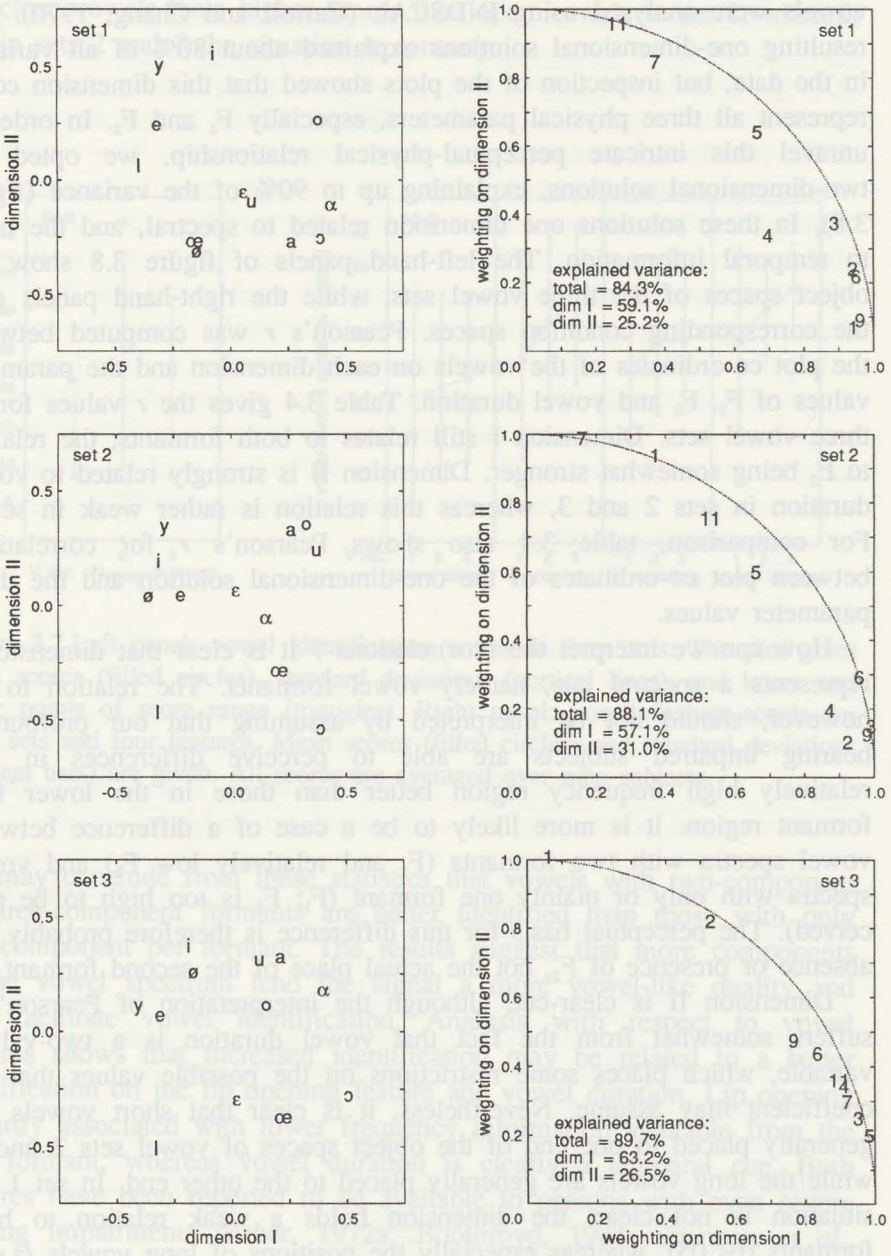

Figure 3.8 Two-dimensional INDSCAL plots for a 12-vowel similarity matrix. In the upper left and right corners of the plots the vowel set is identified, the number indicating the number of components per formant. Left panels: object spaces of 12 Dutch vowels. Right panels: condition spaces of nine subjects, giving percentages of variance explained by the use of these INDSCAL dimensions.

Table 3.4 Pearson's product-moment correlation coefficients (r), expressing agreement between the physical parameters F_1, F_2 and vowel duration, and the positions of 12 vowels, expressed in plot co-ordinates, along the dimensions of two-dimensional and one-dimensional INDSCAL solutions.

	dimension I				dimension II		
	F_1	F_2	duration		F_1	F_2	duration
two-dimensional solution							
set 1	.66*	-.75*	-.15	set 1	-.58	.53	.42
set 2	.53	-.86*	-.22	set 2	-.12	-.04	.80*
set 3	.69*	-.77*	-.43	set 3	-.10	-.12	.80*
one-dimensional solution							
set 1	.76*	-.78*	-.32				
set 2	.61	-.84*	-.32				
set 3	.66*	-.78*	-.54				

* $p<.01$ ($r>.63$)

3.8 General conclusions and implications

The ultimate goal of our research programme is to define or describe speech coding procedures which enable us to develop signal processing hearing aids that are beneficial to profoundly hearing impaired people who hardly profit from current techniques. Choices must be made as to what speech patterns will be conveyed to the profoundly hearing impaired listener as well as to how this information will be presented. These two issues are independent; ideally speaking, the speech information that is most important for the listener should be incorporated in an acoustic signal that is optimally processed by the listener.

The results of the study described in this chapter hint at the sort of acoustic signals that can be applied in conveying relevant speech information. In judging similarities or dissimilarities among various harmonic complexes, we concluded that profoundly hearing impaired subjects based their assessments mainly on two perceptual cues:

(1) primarily, a *temporal cue*: the presence of adjacent high-frequency components, resulting in audible beats; for some subjects, this cue was clearly confounded with loudness perception;

(2) secondarily, a *spectral cue*: the presence of high-frequency components, provided that the separation from low-frequency components is substantial.

The question arises whether the physical properties that are related to these perceptual dimensions can be used to distinguish between certain phonetic features in speech that are otherwise indistinguishable for the profoundly hearing impaired. The results give some evidence, at least, to assume that profoundly hearing impaired listeners are able to process more than only temporal information: it seems possible to create audible differences between two signal patterns by using higher-frequency energy that is not masked by strong low-frequency energy. Generalizing this to speech perception, it means that two independent frequency channels can be utilized for presentation of the speech signal, when coded in a particular way.

Profoundly hearing impaired listeners rely greatly on speechreading for their understanding of spoken messages. But as much speech information is hardly available from visible articulation, presentation of a well-chosen acoustic signal can be a great support to the lipreader. In many studies it has been shown that speechreading may be successfully supplemented by acoustic signals that provide mainly temporal, voicing and amplitude information (Erber, 1972a; Walliker *et al.*, 1983; Breeuwer and Plomp, 1986; Grant, 1987; Van Tasell *et al.*, 1987; Boothroyd, 1988; Faulkner *et al.*, 1992; see chapter 4). Additional presentation of spectral cues may give extra support, which may even provide more than strictly complementary information. In that case overlapping auditory and visual information may yield a small amount of redundancy, which would be welcome to the lipreader.

As an example, let us see how Dutch vowels could be represented by presenting the spectral cue that was used by our profoundly hearing impaired subjects. They are able, to some extent, to simultaneously process high-frequency and low-frequency information. However, appreciating the fact that substantial separation between high-frequency and low-frequency information is required, simultaneous presentation of two vowel formants is not likely to be profitable in all cases. This suggests that only one of the vowel formants should be represented in the frequency region up to about 1500 Hz. Dutch lipreaders are well able to distinguish rounded from unrounded vowels (van Son *et al.*, 1993c; previous chapter), which is roughly the same as distinguishing front vowels, with a relatively high second formant, from back vowels, with a relatively low second formant (compare table 3.3). So, F_2 information is accessible through the visual sense. It seems rather obvious, then, to present F_1 in a supplementary acoustic signal. On the

other hand, F_1 is related to the degree of lip opening, which is also an important visual cue in lipreading, especially for the better lipreaders. This would rather call for an acoustic presentation of F_2. In terms of numbers of discriminable vowels, there is no difference between representation of either F_1 or F_2 (van Son, 1993; see also section 4.2.4 of this thesis). Nevertheless, we would be in favour of presenting F_1 for two reasons. First, because it seems most profitable for the average lipreader; and second, F_1 might be perceived in a virtually normal way by the profoundly hearing impaired, whereas F_2 would have to be mapped onto an unusual frequency region. Besides F_1, it is possible to present simultaneous high-frequency information. We may opt for presentation of relatively high-frequency noise (*e.g.* between 1200 and 2000 Hz) in consonants, or we may consider presenting F_2 for those vowels that normally have a low F_1 and a high F_2: /i,I,e,y/. In that case a fixed F_2 value of maximally 1800 Hz (within the hearing span of most profoundly hearing impaired) may be chosen.

In part II of this study we have investigated how the 'beat cue' can be applied in vowel perception. We have shown that two-component (or even three-component) vowel formants produced higher maximal identification scores than single-component formants. The main conclusion from this vowel task must therefore be that if we want to speech code vowels by giving one or more formants, each formant should be realized by two closely adjacent components. Moreover, two-component representation of voiced speech parts may be used to encode pitch movements. This obviates the need for presenting the f_0 signal as a spectral component, so that usable frequency channels can be reserved for higher-spectral information, such as a formant together with a noise band.

We must realize that speech is essentially characterized as sound energy fluctuating over time: it is a dynamic signal. Explicit knowledge as to how coded speech information is processed by the profoundly hearing impaired must eventually be studied in relation to dynamic acoustic patterns. In chapter 5 we will describe an experiment that deals with the perception of naturally produced, but recoded, speech stimuli, and specifically with the integration of visual and auditory information.

Acknowledgement

We wish to thank Dr. M.E.H. Schouten from the Research Institute for Language and Speech (OTS), Universiteit Utrecht, for his fruitful contribution to this study, both in the preparatory stage and in the discussion of its results and implications.

Chapter 4.

Perceptual dimensions in lipreading and supplementary acoustic cues for the profoundly hearing impaired: a critical look at the potentials of signal processing hearing aids

This chapter is an abridged version of a review paper by N. van Son, submitted to Audiology (1993)

Abstract

In the application of signal processing hearing aids for the profoundly hearing impaired two issues should be addressed: (1) the kind of speech information that needs be presented, and (2) the way in which this information should be presented so as to match optimally with the residual hearing capacities. For many profoundly hearing impaired lipreading is an indispensable source of information. Research has shown that visible (labial) articulatory movements provide the lipreader with vowel lip rounding and lip opening information and with consonant place information. Lipreading does not provide useful prosodic cues. In order for lipreaders to profit from supplementary acoustic signals, these should provide information about vowel duration, consonant voicing and manner of articulation, as well as prosodic cues. It has been argued that profoundly hearing impaired persons are not able to extract this information from the complex speech signal. Therefore, feature extraction techniques have been developed so as to present only selected speech features. Most strategies have been restricted to the presentation of f_0 and amplitude information. So far, the approach has failed to yield better results than presentation of unprocessed speech conveyed by well-fitted hearing aids. It may be argued that the information presented is too minimal, especially if some degree of residual frequency selectivity is present. For such persons, extension of the encoded signal with spectral cues might be profitable.

4.1 Introduction

This chapter provides a review of experimental studies into phoneme lipreading in visual and auditory-visual conditions. Its aim is to give an account of the perceptual dimensions that play a role in visual and auditory-visual speech perception and to present a critical evaluation of the supplementary acoustic cues that have so far been presented to the profoundly hearing impaired lipreader.

As argued in chapter 1, section 1.2.1, the starting point for establishing the informational content of a supplementary acoustic signal may be the type and the amount of information that can be derived from lipreading. An analysis of perceptual errors in lipreading is required for explaining why certain speech features are perceived while others are not. A similar analysis in auditory-visual speech perception may subsequently reveal whether addition of acoustic signals brings any new dimensions of information about the speech signal. In this way an inventory can be made of speech information that can be derived from optical and acoustic patterns.

Most of the implications of speech recognition research discussed in this chapter are based on phoneme identification results only. Phoneme perception studies provide the least confounded insight into the basic visual speech elements and consequent definitions of supplementary acoustic information, because no higher levels of linguistic information are involved. For this reason, this review is mainly concerned with phonemes, and the survey on relevant research is accordingly restricted.

Sections 4.2.1 and 4.2.2 give a survey of experimental studies into vowel and consonant lipreading, both in visual and auditory-visual conditions. The main results are summarized in tables 4.1 and 4.2. In section 4.2.3 some attention is paid to the function of prosodic cues in speech, and their importance for the (profoundly hearing impaired) lipreader. Section 4.3 presents a critical evaluation of the supplementary acoustic cues that have so far been presented to normally hearing and profoundly hearing impaired lipreaders. In section 4.4, finally, conclusions will be drawn about the actual benefit that profoundly hearing impaired subjects have experienced from the presentation of selected speech features. The question is put forward whether the speech signals so far used fully employ the auditory capacities, or whether further extension is possible and required.

4.2 Dimensions in visual and auditory-visual speech perception

In chapter 2, section 2.1.1, the concept of the *viseme* was introduced as an often used measure of lipreading ability. Viseme classifications result from the analysis of visual phoneme confusions and as such shed light on the perceptual dimensions used by the lipreader. Techniques used to arrive at the classifications were briefly described, and they will be mentioned time and again in the surveys of sections 4.2.1 and 4.2.2. of this chapter. In those sections an overview of research into visual and auditory-visual vowel and consonant perception is provided. Its purpose

is to state in global terms, without paying too much attention to the widely varying methods of experiment and data analysis, what the relevant perceptual dimensions are in (auditory-)visual phoneme perception. In section 4.2.3 prosodic information is briefly dealt with. In section 4.2.4, finally, the most important findings are summarized, and an example of a speech coding strategy using those results is given.

4.2.1 Visual and auditory-visual vowel perception

In table 4.1 a summary of relevant perceptual dimensions in visual (V) and auditory-visual (AV) vowel perception is given, as well as recognition scores (V only) and score improvements (AV-V) when available (some studies only investigated purely visual conditions). Improvement is only indicated for hearing impaired subjects or for normally hearing subjects receiving a non-optimal signal; auditory-visual scores for normally hearing subjects under optimal conditions invariably added up to almost 100%. Some of the papers discussed below provided confusion tables, which we analysed ourselves in order to arrive at our own interpretation of the dimensions. Vowel confusions were analysed by converting the confusion matrices to symmetric similarity matrices using an algorithm by Houtgast (Klein *et al.*, 1970), which were subsequently analysed by means of a Kruskal-based MDS (1964). In most cases, of course, our own findings agree with the conclusions stated in the paper, but sometimes it was possible to obtain additional results that, to our opinion, give interesting information.

Eggermont (1964) investigated language acquisition in a group of 156 hearing impaired and deaf Dutch children with varying hearing losses and usable residual hearing[1]. The perception of consonants and vowels, embedded in CVC syllables, was measured in two conditions: lipreading alone and lipreading plus low-frequency auditory information. The results in half of the group, 78 children, were analysed completely, *i.e.* recognition scores were computed and phoneme confusions were evaluated. The results showed that the acoustic signal did not improve vowel recognition. In both presentation modes (V and AV), a clear

---

[1] The children had been assigned to five different groups on the basis of their hearing loss, which was merely specified as "high-tone and low-tone losses" of a certain degree. The author was primarily interested in the results of groups (4) and (5), with high-tone losses of more than 90 dB HL and low-tone losses between 60 and 83 dB HL (4) and beyond 83 dB HL (5). These two groups included 62 of the 78 subjects whose results were thoroughly analysed.

perceptual difference was found between rounded and unrounded vowels. Our own MDS analysis of the vowel confusion data confirmed that lip rounding was the important perceptual dimension in visual perception as well as in aided lipreading; as a secondary dimension also vertical lip opening played a role, but this only showed in the unrounded vowels, where the close vowels /i,I/ were clearly distinguished from open /a,ɑ/.

Dutch lipreading data can be found in two other investigations. Breeuwer and Plomp (1984, 1985, 1986) studied vowel and consonant perception in a group of normally hearing subjects. They provided confusion tables for unaided visual consonant and vowel perception as well as ITA results for the consonant confusions (see also Breeuwer, 1985). Their results agree very well with Eggermont's findings of two vowel dimensions: lip rounding and, especially in the unrounded vowels, vertical lip opening. Results of AV phoneme perception are described in section 4.3.2. Van Son *et al.* (1993c; see chapter 2) presented meaningless syllables (V only) to four groups of subjects, two normally hearing (medical students and speech therapists) and two hearing impaired ('normal' and 'expert' lipreaders). They found, again, that vowels were primarily distinguished on the basis of lip rounding. The second important dimension was interpreted differently for the classes of unrounded and rounded vowels. Within the unrounded vowels vertical lip opening played a role in further discrimination of especially the open vowels /a,ɛi,ɑ/ from the close vowels /i,I,e/. Distinction among the rounded vowels was more likely based on different vowel duration: long /o,ø/ versus short /ɔ,œ,u,y/. The diphthongs /au,œy/, changing from unrounded to rounded and from open to close, were clearly separated from all other vowels.

The relationship between perceptual dimensions in vowel lipreading and physical measures of lip movements during articulation was studied by Jackson *et al.* (1976) and Montgomery and Jackson (1983). In the first study, ten normally hearing subjects rated dissimilarities among 15 American English vowels, presented visually only, on a 7-point scale. The ratings were analysed by means of an INDSCAL procedure (Carroll and Chang, 1970), and the resulting five perceptual dimensions were identified to represent various cues of lip rounding, lip opening and diphthong perception. Montgomery and Jackson (1983) presented the same material to the same subjects, but within a closed-format vowel recognition paradigm. The recognition score was 54%. Stimulus/response pairs were gathered in a pooled confusion matrix and subjected to MDS. The dimensions were found to represent lip rounding and tongue height (vertical lip opening). Multiple regression indicated, however, that the perceptual/physical relations were moderate (50% accounted variance)

Table 4.1. Summary of studies on vowel lipreading, giving the underlying dimensions in visual perception and the extra dimensions in auditory-visual perception. The last two columns give recognition scores in the V-only condition (V score) and improvement scores of auditory-visual recognition over visual recognition (AV-V score).

perceptual dimensions		scores	
visual	auditory-visual	V	AV-V
Eggermont (1964)[1]			
lip rounding	no extra dimension	29%	0%
vertical lip opening (unrounded vowels only)			
Jackson et al. (1976)			
lip rounding	does not apply		
vertical lip opening			
general lip opening			
diphthong perception			
Caplan Hack & Erber (1982)[1]			
lip rounding	no extra dimension	52%	4-14%
vertical lip opening			
Montgomery & Jackson (1983)[1]			
lip rounding	does not apply	54%	
vertical lip opening			
Busby et al. (1984)			
lip rounding	first formant vowel duration	64%	8%
Breeuwer (1985)[1]			
lip rounding	see table 4.3	55%	27%[2]
vertical lip opening (unrounded vowels only)			
Van Son et al. (1993c)			
lip rounding	does not apply	45%	
vertical lip opening (unrounded vowels only)			
vowel duration (rounded vowels only)			

[1] Conclusions are also based on our own MDS analysis of provided confusion tables.

[2] Improvement score for normally hearing subjects provided with best supplementary signal: F_1 plus F_2 (see table 4.3).

and that considerable intertalker differences occurred.

Caplan Hack and Erber (1982) studied auditory, visual and auditory-visual perception of a set of ten American English vowels by 18 severely hearing impaired children (losses of more than 80 dB HL). The children had been assigned to three performance groups on the basis of word recognition scores: good, moderate and poor. Auditory recognition scores for these groups were 56%, 30% and 12%, respectively. In the visual condition there was no difference between the three classes; the average identification score was 52%, and vowels were concluded to be distinguished on the basis of lip rounding. Our own Kruskal (1964) analysis of their confusion data shows that, besides lip rounding, also vertical lip opening is an important dimension. In the auditory-visual condition the good and moderate groups improved their scores by maximally 14%, while the poor group showed only a marginal improvement (+4%) in auditory-visual perception over lipreading alone.

Busby *et al.* (1984) endeavoured to identify the underlying dimensions for auditory, visual and auditory-visual vowel perception in a group of four severely hearing impaired children with a mean PTA[.5,1,2 kHz] of 100 dB HL. Responses to all Australian English monophthong vowels, within a /hVd/ context, were analysed by means of an MDS technique (ALSCAL). Multiple regression analysis was used to verify the relations between perceptual dimensions and physical parameters. In auditory vowel perception three dimensions were found to be of significance: first and second formant frequency, and vowel duration. In pure lipreading only the width of internal lip opening (a measure for lip rounding) was used as a cue. The situation in the combined presentation mode is less clear. The order of vowels along ALSCAL dimensions I and III correlated significantly with both internal lip opening width and the first formant frequency; these two parameters were not independent, however. Dimension II was concluded to represent vowel duration, although the correlation failed the .05 level of significance. Nevertheless, the authors argued that in auditory-visual vowel perception cues from both auditory and visual perception were effectively combined.

4.2.2 Visual and auditory-visual consonant perception

Table 4.2 gives a summary of perceptual dimensions underlying visual and auditory-visual consonant perception. In most studies consonant confusions were analysed by means of ITA; in some cases we analysed the confusions ourselves. The last columns in the table provide visual

scores and score improvements in auditory-visual recognition (only for hearing impaired subjects and for normally hearing subjects provided with non-optimal speech signals).

In the study by Eggermont (1964), referred to in the previous section, aided lipreading proved somewhat more favourable for consonants than for vowels. The auditory-visual condition yielded a raise in identification score of 4%, averaged over initial and final consonants. Low-frequency signal presentation had no effect, however, on the pattern of confusions: the very same perceptual dimensions played a role in unaided and aided lipreading. The main perceptual difference is that between labial and nonlabial consonants. Our own MDS and ITA analyses of Eggermont's consonant confusions showed that four groups (visemes) could be distinguished: bilabial /p,b,m/, labiodental /f,v/, bilabial /w/, and nonlabial /t,d,s,z,n,l,j,k,x,ŋ,r,h/. The dimension underlying this four-group distinction is clearly labialization; the type of labialization is also an indicator for the actual place of articulation, for which reason the underlying dimension in consonant lipreading is referred to as *labial* place of articulation in table 4.2.

The other two Dutch studies already mentioned (Breeuwer, 1985; van Son *et al.*, 1993c) are fully in line with Eggermont's data. Dutch consonants are mainly distinguished visually on the basis of labial articulation; actual place of articulation plays a minor role within the set of nonlabial consonants. Interestingly enough, it appeared that the fricative consonants /s,z,ʃ/ tended to form a subset within the nonlabial group especially in the van Son *et al.* data, probably because of the narrow lip opening resulting from the necessary constriction in the palato-alveolar region of the oral tract.

A consonant lipreading study was performed by Woodward and Barber (1960). Twenty-four consonants were used in /Ca/ syllables. A selection of 229 consonant pairs was presented on black-and-white film to three groups of normally hearing adults, each of which received one of three presentation modes: auditory, visual, auditory-visual. The subjects' task was to note whether the members of a consonant pair were 'different' or 'alike'. Visual difference ratings were computed and a hierarchy of visual contrasts was established. In the visual condition 43% of all consonant pairs was found to contain different members; in the other two conditions this percentage was considerably higher: auditory 77%, auditory-visual 83%. In terms of visually contrastive units, the following four sets of consonants could be classified: bilabial, labiodental, rounded labial and nonlabial. Labialization is again indicative of the actual place of articulation and is therefore referred to as *labial* place of articulation (table 4.2). Because of the high scores in

the auditory and auditory-visual conditions, the respective perceptual dimensions underlying lipreading performance could not be established by us using MDS.

Erber (1972b) performed a closed-format identification study on a limited set of consonants: /p,b,t,d,k,g,m,n/. The consonants were presented visually, auditorily and auditory-visually in /aCa/ syllables to three groups of five children, normally hearing, severely hearing impaired (hearing losses between 80 and 100 dB HL) and profoundly deaf (hearing losses over 100 dB HL). The auditory phoneme scores for the three groups were 100%, 50% and 21%, respectively. In the lipreading condition, all three groups were able to distinguish the consonants as to place of articulation. The average recognition score was 42%. In auditory-visual conditions the normally hearing and hearing impaired groups achieved almost perfect recognition (100% and 88%, respectively), but the profoundly deaf group profited much less from the additional acoustic information (a wideband speech signal): recognition scores improved by only 15%. Our own ITA, performed on consonant confusions pooled across the three subject groups, showed that manner and voicing information was transmitted considerably better in the auditory-visual condition than in the visual condition: increases amounted from 13% to 71% (manner) and from 1% to 47% (voicing), so performance was generally well above the chance level of 0% transmitted information.

Binnie *et al.* (1974) investigated the roles of audition and vision in the perception of consonants by normally hearing subjects. Sixteen English consonants embedded in /Ca/ syllables were presented in eight experimental conditions: auditory in quiet, visual (in quiet), auditory in noise (S/N ratios of -18, -12 and -6 dB) and auditory-visual in noise (same S/N ratios). The authors conducted an ITA with some of the Miller and Nicely (1955) articulatory features and showed that low-frequency information (voicing, nasality) was the strongest auditory cue, that place of articulation information was the first to disappear under noisy conditions, and that lipreading substantially reduced the number of confusions among phonemes with different places of articulation. The visual condition yielded scores of up to 43% correct identifications versus 95% in the auditory condition in quiet. In the S/N ratio conditions of -6 and -12 dB, auditory-visual recognition scores were about 86%, so the gain over pure lipreading was 43%.

Walden *et al.* (1977) undertook a study in order to evaluate the effect of structured lipreading training (purely visual) on visual consonant recognition. An intensive period of training was preceded and followed by an evaluative test, and improvement was measured in terms of

Table 4.2. Summary of studies on consonant lipreading, giving the underlying dimensions in visual perception and the extra dimensions in auditory-visual perception. The last two columns give recognition scores in the V-only condition (V score) and improvement scores of auditory-visual recognition over visual recognition (AV-V score).

perceptual dimensions		scores	
visual	auditory-visual	V	AV-V
Woodward & Barber (1960)			
labial place of articulation	no extra dimension	43%	40%
Eggermont (1964)[1]			
labial place of articulation	no extra dimension	40%	4%
Erber (1972b)[1]			
place of articulation	voicing manner of articulation (nasality, plosion)	42%	15-46%
Binnie et al. (1974)			
place of articulation	nasality voicing	43%	43%[2]
Walden et al. (1977)[1]			
place of articulation	does not apply	44%[3]	
Owens & Blazek (1985)[1]			
place of articulation	does not apply	34%	
Breeuwer (1985)[1]			
labial place of articulation	see table 4.3	38%	31%[4]
Busby et al. (1988)			
place of articulation	manner of articulation	48%	13%
Van Son et al. (1993c)			
labial place of articulation vertical lip opening	does not apply	36%	

[1] Conclusions are also based on our own ITA of provided confusion tables.
[2] Improvement score for normally hearing subjects with S/N-ratios of -6 and -12 dB.
[3] Post-training scores.
[4] Improvement score for normally hearing subjects provided with best supplementary signal: f_0 plus amplitude (see table 4.3).

increased recognition scores, numbers of visemes and within-viseme responses. After training, the number of visemes increased from five to nine. The fact that in the post-training situation /s,z/ could be distinguished from /ʃ,ʒ/, where lip rounding is involved, and labialized /r/ was split off from /t,d,n,j,k,g/, indicates that lipreaders actually learn to use labialization as a feature. Because /s,z/ and /ʃ,ʒ/ have different places of articulation, lip movement or rounding may serve as an indicator for this (as is the case in /p,b,m/ and /f,v/).

Owens and Blazek (1985) investigated viseme identification by normally hearing and hearing impaired subjects, and also studied the effect on viseme classification of various vocalic contexts. Twenty-three consonants were presented in VCV syllables, with contexts /a,i,u,ʌ/, in a purely visual condition. No differences between the two groups were found, but vocalic context appeared to have a certain influence on viseme recognition; especially /u/ limited recognition to /p,b,m/ and /f,v/ only. The overall recognition score was 34%. With regard to other viseme research, Owens and Blazek (1985) concluded that a general six-viseme classification, based on place of articulation, could be established for English consonants.

Busby *et al.* (1988), finally, studied the perception of 14 consonants in /aCa/ syllables, presented to four hearing impaired children (the same as in their 1984 study: mean PTA of 100 dB HL) in auditory, visual and three auditory-visual conditions. Identification scores were 28% (auditory), 48% (visual), and 61% (auditory-visual). An ITA of consonant confusions showed that voicing and manner of articulation were strong cues for auditory perception, while in visual perception place of articulation appeared most important. In the combined condition both manner and place of articulation played a role, but voicing did not.

4.2.3 Prosodic information

We will digress a moment to lipreading of longer speech segments, and find that not only segmental information is of importance, but also suprasegmental or prosodic information. Prosodic structures involve intonation contours, rhythmic patterns, word and sentence stress, and pauses in the speech signal. Whereas prosody in itself does not make speech intelligible, it plays an important role in organizing speech parts syntactically, and may be of help to the listener in understanding speech under adverse perceptual conditions. It has been shown that prosodic structures offer restrictions to the listener in the sense of possible response patterns (Kozhevnikov and Chistovich, 1966; Wingfield, 1975).

The division of running speech into syntactically coherent speech parts depends on the perceptual interaction between the prosodic structure and the order of recognized words. In case of difficult circumstances, where words will not be recognized quite easily, it seems natural that more weight is placed on the prosodic structure to perceive syntactic parts in the speech signal.

Compared to phoneme lipreading studies, not much research has concentrated on the question as to what prosodic information is at all available through lipreading and what acoustic information can or should consequently be added for the lipreader. A valuable exception is the work reported by Risberg and Agelfors (1978). Prosodic information is normally regarded to be (almost) unavailable through visual speech perception. Only gross temporal information and pitch movements are conveyed through the sequence of opening and closing lip movements (Pickett, 1963; Risberg, 1974; Erber, 1979) and by means of non-verbal gestures (facial expressions, nodding, arm movements, *etc.*). Subtle prosodic information is hardly conveyed visually: lipreaders are not likely to see which words in a sentence are stressed, or where there is a break between main and subordinate clauses. Generally, prosodic cues are automatically treated as the first and often only kind of speech information to be presented auditorily (*e.g.*, Rosen *et al.*, 1987; Boothroyd, 1988; Boothroyd *et al.*, 1988). This approach is even more understandable as prosodic information, in contrast to segmental information, is cued in low-frequency information and relatively slow temporal modulations of the speech envelope, which may be perceived fairly well by the profoundly hearing impaired listener.

4.2.4 Supplementary information for the profoundly hearing impaired

The results from the studies given in the previous sections indicate the importance of the following dimensions in visual phoneme perception: *lip rounding* (spread/round contrast) and *vertical lip opening* (open/close contrast) define the visual contrasts among vowels, while the primary cue in consonant lipreading is *(labial) place of articulation* (bilabial/ labiodental/labialized/nonlabial contrasts). Prosodic information is hardly available. These findings firmly support Woodward and Barber's (1960) observation that "the fact that only labial articulations are discriminated consistently seems to justify the term 'lipreading' for this level of activity in visual speech reception" (p. 220). Applying these results, one might expect observers to be able to distinguish, in unaided lipreading, among four sets of monophthong vowels (spread/close, spread/open,

round/close, round/open) and possibly a distinct set of diphthongs, and four to six sets of consonants (bilabial, labiodental, labialized, interdental, front and back nonlabial)[1].

Most of the studies also indicate that, when a lipreader is presented with auditory as well as visual information, speech perception is enhanced substantially. Basically, this may be achieved in two ways. On the one hand, new information is added: speech features may become available that cannot be perceived by unaided lipreading, *e.g.* consonant voicing and prosodic cues. On the other hand, one may present acoustic signals carrying information that was already available through the visual sense, such as vowel formants F_1 and F_2. When listeners are largely deprived of conclusive evidence about the contents of the spoken message, such overlapping information may provide just that little bit of help that is needed to "confirm or reject perceptual hypotheses about oral patterns that they receive through lipreading" (Erber, 1979, p. 267).

As argued in the introductory section, acoustic supplements for the profoundly hearing impaired lipreader may involve newly formed speech signals, based on feature extraction techniques, rather than presentation of the unprocessed speech signal. This approach includes both the selection of speech cues that are most helpful to the lipreader and appropriate matching of the encoded signal to the residual auditory capacities. The primary aim is to enhance the discriminability of speech sounds, and this may be achieved by creating any new speech code carrying the selected information. As the profoundly hearing impaired listener does not perceive any sound whatever when not using a hearing aid, there is some freedom, in principle, to create new perceptual sensations in the profoundly hearing impaired listener by using any newly formed acoustic signal (Smoorenburg, 1990b).

The following example will illustrate how such an entirely new speech code may be defined. Figure 4.1 gives an F_1/F_2 space for Dutch monophthong vowels, in which the three Dutch vocalic visemes that the average lipreader may recognize (van Son *et al.*, 1993c) are indicated (short and long rounded vowels combined). Formant frequencies were adopted from Pols *et al.* (1973)[2]. Definition of a supplementary acoustic

---

[1] The exact assignment of vowels and consonants to contrastive sets depends on the phonetic classification system one adopts, and will differ across languages (see *e.g.* chapter 7 in O'Connor, 1973).

[2] In figure 4.1, the formant frequencies of /o,ɔ,ø,œ/ were slightly changed in order to be able to draw the outlines for indication of the viseme structure. Normally, /o/ and /ɔ/ have almost identical formants, and so do /ø/ and /œ/. The members of the pairs are mainly distinguished on the basis of vowel duration.

signal should lead to an optimization of the number of within-viseme vowel groups (one might call them *audiovisemes*). If there is some residual frequency selectivity, vowels may be represented unidimensionally by F_1, F_2, or, alternatively, a weighted combination of the two formants, whatever is the most useful supplement for a given profoundly hearing impaired lipreader. The effect of three potential coding schemes is given in figure 4.2. After extraction of the two vowel formants from the full speech waveform, arithmetic rules have been applied such that all vowels are represented as one formant (frequencies are logarithmically scaled) in the frequency range between about 300 and 1050 Hz, which most profoundly hearing impaired still have at their disposal (Faulkner *et al.*, 1992; Bosman and Smoorenburg, 1993a, 1993b; van Son *et al.*, 1993a). In the figure, the visible difference between the three Dutch visemes is expressed by separating them by means of dashed lines. Auditory separation between the vowels within a viseme may be possible with Δf of 10-15%, which is the average difference limen for frequencies (DL_F) in profoundly hearing impaired listeners (Bosman and Smoorenburg, 1993b).

Coding scheme 1 applies the following rule, based on F_1: $1.4*(F_1-70)$. The new vowels thus created can be grouped into six or seven auditory-visual sets: /i/, /I,e/, /ɛ,ɑ,a/, /o,ø/, /y,u/ and /œ,ɔ/ (within /ɛ,ɑ,a/ a difference may be heard (and seen) between /ɛ/ and /a/).

Coding scheme 2 is based on F_2 in the following rule: $0.53*(F_2-250)$. Seven sets may be discerned: /ɑ/, /a/, /ɛ,I,e,i/, /o/, /ø/, /u,ɔ/ and /œ,y/. Vowel /ɛ/ may be distinguished from /i/. The clearest effect of this scheme is that, in the higher F_2-region (close front vowels), all formants lie close to each other.

Finally, *coding scheme 3* is based on a weighted ratio of the two vowel formants: $115*((F_2/F_1)+1.1)$. This scheme renders the largest number of discriminable sets, *viz.* ten: /ɑ,a/, /ɛ/, /e,I/, /i/, /o/, /ø/, /ɔ/, /u/, /œ/ and /y/.

In this example, the rules have been defined arbitrarily, only in order to show their effects. The new speech codes thus presented are not like natural speech, certainly not those of schemes 2 (where F_2 is shifted to a different frequency region) and 3 (which is a weighted combination of the two formants). Scheme 1 leads to vowels in an extended F_1 range. It is apparent that, for successful use, listeners must be able to learn and interpret the code, which may prove to be more easy with scheme 1 than with schemes 2 or 3. On the other hand, if learning is no problem, scheme 3 offers the greatest potential in auditory-visual distinction of vowels.

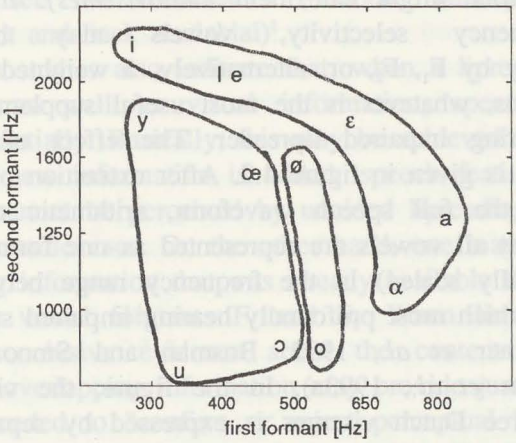

Figure 4.1. F_1/F_2 space of 12 Dutch vowels (formant frequencies are logarithmically scaled), with outlines indicating visemes. Formant frequencies after Pols *et al.* (1973).

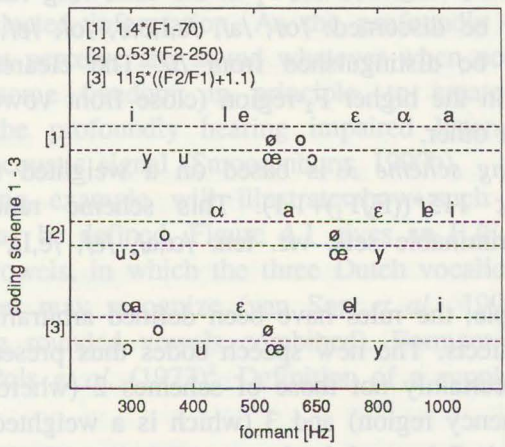

Figure 4.2. The effect of three coding schemes applying simple arithmetic rules to the frequency values of formants F_1 and F_2. Formant frequencies are those given by Pols *et al.* (1973), see also figure 4.1. The arithmetic rules were defined to the purpose that all vowels are represented as a single formant within a range from about 300 to 1050 Hz. The dashed lines separate the three Dutch vocalic visemes. The effects of the arithmetic rules are explained in more detail in the text.

4.3 Lipreading supplemented with processed acoustic information

4.3.1 Research with acoustic supplements

The efficacy of various sources of information supplementing speechreading has been investigated mostly in normally hearing subjects (section 4.3.2). The aim of this research, surveyed in table 4.3, is usually to find out which acoustic signals carry relevant speech information that best serves the lipreader. Whether this information is actually of any help to the profoundly hearing impaired is the subject under study in section 4.3.3. Only little research has been carried out in this area so far; the results, summarized in table 4.4, come from experiments with laboratory-simulated or prototype signal processing hearing aids for the profoundly hearing impaired.

Appreciating the importance of pitch contours for lending stress, marking of questions and cueing of syntactic boundaries, many investigators have provided lipreaders with fundamental frequency information (pitch), in order to give them a framework that may structure the speech signal into smaller, linguistically relevant, word groups. Speech envelope amplitude information has also been considered to provide useful speech cues in aid of lipreading. In only very few studies has the effect of spectral cues been the specific topic of evaluation. Most of the studies summarized below have mainly employed f_0 and the speech envelope amplitude.

4.3.2 F_0 and amplitude cues for normally hearing lipreaders

Van Tasell *et al.* (1987) concluded from studies by Horii *et al.* (1971) and Erber (1972a) that the envelope may convey phonemic information on the basis of amplitude and duration cues. Subsequently, they designed a study for investigating the nature of the time/intensity envelope cues for consonant recognition. They derived from a set of 19 /aCa/ stimuli three different waveform envelope amplitude-modulated low-pass noises by low-pass filtering the envelope at three cutoff frequencies: 20 Hz, 200 Hz, and 2000 Hz. The stimuli were presented in one unprocessed and three processed conditions to a group of 12 normally hearing students. Confusion data were collected and analysed by means of MDS and HC. Four distinct envelope groups (called *envemes*) were established on the basis of three perceptual dimensions, labelled "voicing envelope", "amplitude envelope", and "burst envelope": /p,t,k/, /f,θ,s,ʃ/, /b,d,g,v,ð,z,ʒ/, and /m,n,l,r,j/. There was no difference

between the 200 Hz and 2000 Hz conditions, but both of them yielded significantly higher recognition scores than the 20 Hz low-pass filtering. Van Tasell *et al.* (1987) asserted that a subject who is able to extract all possible information from lipreading as well as from the waveform envelope will receive 95% of the consonant information in the 19-stimulus set. This seems an overestimation, however, of a realistic performance by the subjects; the rather low percentages of information actually transmitted (29% and 35% in the best two processing conditions) indicate that information extraction from the waveform envelope is no *sinecure*.

In a comparison study Breeuwer and Plomp (1986) investigated the effect of four different acoustic speech signals as a supplement to speechreading CVC syllables and sentences by normally hearing lipreaders: two-formant information (*F1+F2*), fundamental frequency (*F0*), fundamental frequency plus amplitude information (*F0+A*), and sound-pressure information from octave bands around 500 Hz and 3160 Hz (*SPI*). As a supplement to vowel lipreading, *F1+F2* appeared to be beneficial. As to consonants, the other three signals were most favourable in auditory-visual consonant lipreading, especially because of added voicing and manner information. For general speechreading, the authors concluded that three of the supplementary signals proved appropriate, the differences among them not being significant: *F1+F2*, *F0+A*, and *SPI*.

The contribution of voicing and amplitude cues to speechreading has also been investigated by Grant *et al.* (1985). They studied connected discourse tracking (CDT) performance by three normally hearing subjects in two presentation modes: speechreading alone and speechreading plus various acoustic signals. An unprocessed-signal condition yielded the maximal performance score, and all other scores were expressed as a percentage of this score. The results showed that all auditory-visual conditions led to significantly higher performance than the speech-reading-only condition. Low-pass filtered speech and amplitude-modulated fundamental frequency proved significantly better than either an amplitude-modulated fixed-frequency sinusoid or a constant-level fundamental frequency. Altogether, the authors concluded that normally hearing lipreaders may improve their performance when they are aided by acoustical signals carrying durational, amplitude and fundamental frequency information. These signals are especially valuable in that they provide prosodic information otherwise unavailable to the lipreader. However, regarding the poorer auditory capacities, the authors are quite doubtful as to the degree in which profoundly hearing impaired lipreaders may profit from the prosodic cues provided in the signal types

investigated.

Boothroyd and co-workers have mainly been concerned with the f_0 signal as a useful supplement to lipreading. In most of their experiments, whether they studied perception of speech contrasts, words or sentences, they were able to demonstrate speechreading enhancement when presenting fundamental-frequency information to normally hearing lipreaders (Boothroyd, 1988; Boothroyd *et al.*, 1988; Hnath-Chisolm and Boothroyd, 1992). F_0 contours also have a great potential for (postlingually) hearing impaired lipreaders.

Boothroyd (1988) measured the effect of f_0 presentation on the perception of a number of phonologically significant speech contrasts, both suprasegmental and segmental. The f_0 signal had been extracted from the full waveform, and was represented as a constant-amplitude square wave. In an auditory condition it was found that all suprasegmental contrasts were perceived almost perfectly, while final consonant voicing was the only segmental score that was significantly above chance level. The contrasts accessible from f_0 contours were concluded to be cued by presence and duration of voicing (stress, final consonant voicing), average f_0 (talker sex), and the amount and direction of change of f_0 (stress, pitch rise/fall, pitch range). A visual condition provided enough information for above-chance perception of the segmental contrasts of vowel height and place (respectively corresponding to *lip opening* and, indirectly, to *lip rounding*), initial consonant place and, partially, final consonant place and continuance (plosives and affricates versus fricatives). Auditory-visually, contrasts become accessible through both modalities; the performance is close to that of the better one.

Boothroyd *et al.* (1988) employed two electroglottograph (EGG) signals representing f_0, one unfiltered and one low-pass filtered at 350 Hz. The results of an everyday-sentence test showed that low-pass filtering led to a significant decrease of the enhancement effect: the unfiltered EGG signal gives 46% score enhancement over lipreading alone, whereas this effect is 39% when a filtered EGG version is presented to the lipreader. The difference between the two EGG signals can be attributed to the filtering out of low-frequency spectral information. The authors argued that, even though further investigation is required, f_0 provides sufficient supplementary information in speechreading for all subjects, whether they are competent in speechreading or not, or whether they are normally hearing or (postlingually) hearing impaired.

Table 4.3. Summary of studies on the efficacy of supplementary speech cues auditorily presented to normally hearing lipreaders, giving the types of signals used and the speech material presented to the subjects. The last two columns give recognition scores in the V-only condition (V score) and improvement scores of auditory-visual recognition over visual recognition (AV-V score).

acoustic signal	speech material	scores	
		V	AV-V
Grant et al. (1985)			
FM sinusoid (f_0)	connected discourse	38%[1]	30%
AM sinusoid			30%
AM+FM sinusoid (f_0)			39%
LPF speech (300 Hz)			40%
Breeuwer & Plomp (1986)[2]			
F_1 and F_2	vowels (F_1 and F_2)	55%	27%
f_0	consonants (f_0 and ampl)	38%	31%
f_0 and amplitude	syllables in sentences (SPI)	28%	59%
sound pressure in two octave bands (SPI)			
Van Tasell et al. (1987)			
envelope-amplitude modulated noise bands (three low-pass filters)	intervocalic consonants	does not apply[3]	
Boothroyd (1988)			
f_0	prosodic contrasts[4]	13%	82%
	vowel contrasts[4]	93%	1%
	consonant contrasts[4]	43%	21%
Boothroyd et al. (1988)			
EGG[5] (unfiltered)	words in sentences	29%	48%
EGG (LPF at 350 Hz)			41%
Hnath-Chisolm & Boothroyd (1992)			
quantized f_0 (4 conditions)	words in sentences	32%	22%[6]
EGG (LPF at 350 Hz)			42%

To be continued on next page

Table 4.3. Continued.

Notes:

1. CDT rates (words per minute) are expressed relative to the rate in a normal-speech condition.
2. All speech materials were presented in all signal conditions. Score improvements are only given for the best conditions (in parentheses).
3. Presentation was only auditory. Identification scores were 19% (2 kHz), 15% (200 Hz), and 13% (20 Hz).
4. Presentation was only auditory. In these cases, score improvement should be read as AV-A score.
5. EGG stands for "electroglottograph"; this signal is a registration of the vocal fold vibration, picked up from skin electrodes placed at both sides of the larynx.
6. Score improvement of the best quantized-f_0 condition (12 steps).

Finally, Hnath-Chisolm and Boothroyd (1992) investigated the sensitivity of the f_0 contour reproduction: is exact reproduction of f_0 required or will an approximation be sufficient ? They addressed this question by means of a sentence recognition task run in six conditions, including unaided speechreading, speechreading supplemented with a low-pass filtered EGG signal, and speechreading supplemented with processed and quantized f_0 contours. Quantization was applied in 1 step (an on/off signal representing the voiced/voiceless distinction), 4, 8, and 12 steps. Again, the efficacy of auditorily presented f_0 as an aid to the lipreader was demonstrated by improvement rates of maximally 22% in case of the 12-step quantized signal and 42% in case of the filtered EGG signal. Quantization into less than 12 steps proved significantly worse. The discrepancy between the 12-step quantization and the filtered EGG signal was found to be largely due to distortion effects of processing errors (*e.g.* pitch extraction) rather than to the quantization itself.

4.3.3 F_0 and amplitude cues for profoundly hearing impaired lipreaders

Following the strategy of an earlier designed extracochlear device for the totally deaf (Fourcin *et al.*, 1979, 1983), Fourcin and co-workers developed a feature extraction hearing aid designed for profoundly hearing impaired lipreaders, the SiVo aid (*Sinusoidal Voice*). The basic aid provides voice fundamental frequency information during voiced speech parts and no output during voiceless speech or silence. The

signal was a frequency-modulated sinusoid at the most comfortable level (Rosen *et al.*, 1987). In a comparison between the SiVo aid and a high-quality conventional reference aid (Faulkner *et al.*, 1990a), consonant recognition proved the same for both aids, 68%, while only one subject profited significantly more from the SiVo aid. Feature transmission scores (voicing and manner) indicated that in two out of eight patients the SiVo aid offered significantly better transmission of voicing than the reference aid. A follow-up study (Faulkner *et al.*, 1992) also showed that speech reception performance (intervocalic consonant recognition, connected discourse tracking, and question/statement) was the same whether the SiVo aid or the reference aid was used. The overall conclusion, then, must be that the SiVo aid does not prove better or worse than a well-fitted conventional hearing aid. As a general measure of the practical utility of the SiVo aid the authors reported that half of the subjects preferred its daily use over that of a conventional hearing aid. For them, the SiVo aid provided better or at least equal performance to normal amplification.

It appeared that some patients were able to extract consonant manner information and even some low-frequency vowel formant cues from the amplified unprocessed speech signal. Faulkner *et al.* (1992) suggested that the residual abilities that were apparently present in these patients could be better utilized by means of an extended version of the SiVo. In this version voiceless speech was represented by means of a low-frequency noise band (between 50 Hz and 400-1000 Hz), while amplitude information was provided by amplitude modulation of the noise band and the voice-based sinusoid. Intervocalic consonant recognition and CDT was tested in five of the original eleven patients, comparing amplified speech and compound speech pattern presentation. In general, the listeners profited from the speech pattern information added to fundamental frequency. In consonant identification the combination of f_0, speech noise and amplitude information was especially advantageous for the perception of fricative and plosive information, thus raising overall scores (88%) significantly as compared to unprocessed speech (81%). In CDT, only f_0 plus amplitude information showed some advantage of processed speech over amplification, while there was no difference between amplified speech and the full compound pattern signal.

Table 4.4. Summary of studies on the efficacy of supplementary speech cues auditorily presented to profoundly hearing impaired lipreaders, giving the types of processed signals used and the speech material presented to the subjects. In all experiments, speech materials were presented in auditory-visual conditions, except where noted, and comparisons were made with unprocessed speech, either low-pass filtered or conveyed through a hearing aid. The last two columns give recognition scores with the best unprocessed signal (unproc) and the best processed signal (proc).

acoustic signal	speech material	scores	
		unproc	proc
Grant (1987)			
f_0 (various conditions)	connected discourse	68 wpm[1]	68 wpm
	intonation judgment[2,3]	94%	99%
	stress judgment[2,3]	95%	96%
Faulkner et al. (1990a)			
basic SiVo (f_0)	intervocalic consonants	68%	68%
	question/statement[3]	70%	81%
Faulkner et al. (1992)			
basic SiVo (f_0)	intervocalic consonants	73%	67%
	question/statement[3]	74%	83%
	connected discourse	62 wpm	61 wpm
Faulkner et al. (1992)			
extended SiVo (f_0,	intervocalic consonants	81%	88%
amplitude, noise)	connected discourse	59 wpm[4]	63 wpm

[1] wpm stands for "words per minute".

[2] Mean closed-set scores in an amplitude-modulated f_0 condition, averaged across two different tests; the mean scores were based on the best three subjects only, as the other two subjects generally scored at or below chance level.

[3] These materials were presented in an auditory-only condition.

[4] Mean tracking score in an f_0 plus amplitude condition (without noise).

Finally, Grant (1987) studied CDT performance and stress and intonation judgment in five prelingually profoundly hearing impaired subjects provided with normal speech, low-pass filtered speech and various processed f_0 signals. Basically, two patterns of results were observed.

The three better subjects were able to improve visual CDT rates when presented with additional auditory information; there was no difference between the normal-speech and processed conditions. Their stress and intonation judgments were roughly the same with speech and with f_0 information. The two other subjects showed no CDT rate improvement when auditory information was added. There was substantial improvement in stress and intonation judgment with the best processed signal over normal speech presentation. These latter two subjects had rejected the use of a hearing aid early in their development and had little auditory experience, for which reason Grant supposed that they had difficulty integrating sound and visual information. Because it was for subjects like these that improvement by means of f_0 presentation was sought, the CDT performance was rather discouraging.

4.4 Conclusions and implications

From the overviews of lipreading abilities in and signal presentation to normally hearing and profoundly hearing impaired lipreaders some firm conclusions can be drawn:

1. In lipreading, visible (labial) articulation movements provide the lipreader with vowel lip rounding and lip opening information and with consonant place information. This information is related to vowel formants and relatively high-frequency consonant information in the acoustic domain.
2. Lipreading does not provide useful prosodic cues, which are important for syntactically structuring running speech (by dividing it into coherent word groups).
3. Normally hearing lipreaders profit from supplementary auditory information, providing overlapping cues (F_1 and F_2 information, related to vowel lip rounding and lip opening) or extra, complementary, cues (vowel duration, consonant voicing and manner of articulation, and prosodic cues such as intonation and temporal aspects).
4. Additional acoustic signals providing these cues have also been shown to successfully support profoundly hearing impaired lipreaders with residual hearing for low frequencies and slow temporal modulations.
5. The feature extraction approach of providing only selected speech information, however, has so far failed to yield better results than is possible with presentation of unprocessed speech conveyed by well-fitted hearing aids.

Especially the fifth conclusion is an interesting and perhaps somewhat dismaying one, considering the idea that underlies the development of speech processing hearing aids. According to this idea, the complete speech signal contains too much and, as a consequence, confusing information for a profoundly impaired ear, and might therefore be better replaced by a signal that only proffers minimal auditory information. Observing the results from the experiments discussed, one may wonder if the minimal information that has so far been provided, and which is primarily based on prosodic cues, is enough for increasing speech understanding. The question is: is not this minimal information too minimal ? And is not an extension of the signals to include also spectral cues possible or even required ?

Until now there has been little experience with applications of spectral cues in speech-pattern presentation for the profoundly hearing impaired. Yet, there is evidence to suggest that some profoundly hearing impaired listeners do have residual frequency selectivity and consequently may be able to process more than just prosodic and consonant voicing information. Lamoré *et al.* (1985) found that residual hearing capacities, including phoneme discrimination, may be present up to an average hearing loss (Fletcher index) of 105 dB. Rosen *et al.* (1987, 1990) reported that two listeners with a certain degree of frequency selectivity were able to make use of the first formant information in vowel identification. Faulkner *et al.* (1990b) reported that one of their listeners also showed this ability. The frequency selectivity was argued to make a gross distinction between temporal information carried in the regions of the fundamental and the F_1 frequency. Interestingly, a similar result was found in our own study on similarity ratings of harmonic complexes (chapter 3; van Son *et al.*, 1993a): profoundly hearing impaired subjects were concluded to have based their ratings on temporal and spectral information available from two frequency channels, roughly below and above 800 Hz. The synthetic-vowel identification task showed that the same subjects were able to use F_2 information; perception may be attributed to the difference between vowels with a relatively low F_2 and those with a high F_2 (probably lying beyond the hearing span), rather than to the exact F_2 position. Finally, in section 4.3.3 it was already mentioned that Faulkner *et al.* (1992) extended, with some success, the original SiVo design so as to include some spectral information cueing consonant manner of articulation.

All of these results suggest that at least some profoundly hearing impaired, retaining a certain degree of residual frequency selectivity, might profit more from presentation of (low-frequency) spectral cues in addition to prosodic information than from prosodic information alone.

On the basis of our findings discussed in chapters 2 and 3, we defined two supplementary signals for lipreaders that contain possibly useful spectral cues, offering segmental information. The potentials of these two signals have been evaluated in two lipreading experiments: a sentence recognition task (Bosman and Smoorenburg, 1993a, 1993b) and a vowel discrimination task (van Son *et al.*, 1993b). The latter study will be described in chapter 5.

Chapter 5.

Discrimination of natural and processed Dutch vowels by profoundly hearing impaired lipreaders

This chapter is an extensive version of a paper in preparation by N. van Son, A.J. Bosman and G.F. Smoorenburg (1993b)

Abstract

A vowel discrimination task was performed by 12 profoundly hearing impaired persons. Three speech conditions were presented auditory-visually: amplified speech (AMPL), coded speech containing first formant information (F1), coded speech containing first and second formant information (F1F2). In order to check the relative contributions of auditory and visual perception, two control measures were run with some of the subjects: visual-only (n=11) and auditory-only (n=6). Another auditory-only control measure was run with a group of six normally hearing subjects, only presented with the F1 and F1F2 speech codes, in order to measure optimal discrimination scores for processed vowels. Discrimination among vowels was measured within and across three visemes (rounded and unrounded vowels, and diphthongs). All scores were corrected for chance (p=.5). Control measures in the normally hearing subjects showed that coded speech permitted them to reach score levels of 97% (F1F2) and 83% (F1). For the profoundly hearing impaired subjects, the visual-only discrimination score for vowels belonging to different visemes was 94%. When sound was added, this across-viseme performance increased by a few percent; no difference was found among the speech conditions. As to discrimination within visemes, all scores were well beyond chance level. The visual-only score was 58%, while auditory-visual scores were significantly higher: 79% (AMPL), 73% (F1F2), and 75% (F1). The AMPL scores were significantly higher than those for the two speech codes, which were not significantly different from one another. Response analysis (INDSCAL) indicated that, in all conditions, discrimination was based on two perceptual dimensions, explaining 80-90% of the total variance. Dimension I related to F_1 or vertical degree of lip opening, dimension II to vowel duration.

5.1 Introduction

5.1.1 Formant-based synthesized speech signals

In chapter 4 the rationale of the presentation of visual and well-matched auditory speech cues for the profoundly hearing impaired has been discussed. Experimental research has shown that prosodic speech cues

allow lipreaders to structure an otherwise hardly interpretable stream of visible articulatory movements, but that it does not help intelligibility much in terms of phonemic identification. Moreover, the (limited) available evidence suggests that recoded signals containing only f_0 and amplitude information hardly yield better results than whole-speech presentation through a conventional hearing aid. The information thus offered to the profoundly hearing impaired lipreader is not sufficient, and might well be extended by spectral information cueing segmental speech features.

So far, few applications of spectral cue presentation have been designed. This is mainly a result of the assumption that the speech signal easily becomes too complex for the profoundly impaired hearing to cope with all spectral and temporal information contained in it. Especially frequency selectivity is often grossly reduced so that *e.g.* variations in F_1 may not be distinguished in all cases. Several studies, however, have shown that profoundly hearing impaired listeners may be able to use temporal as well as spectral speech cues, if they have residual frequency selectivity (Lamoré *et al.*, 1985; Rosen *et al.*, 1987, 1990; Faulkner *et al.*, 1990b, 1992; van Son *et al.*, 1993a, chapter 3). In order to use those limited capacities, we synthesized two versions of a speech signal containing formant information, so as to provide segmental speech information. The efficacy of these speech codes were investigated in a vowel discrimination experiment described in this chapter.

5.1.2 Objectives of this study

Auditory-visual vowel discrimination was studied in a group of profoundly hearing impaired persons, most of whom had also participated in the study described in chapter 3. From the lipreading study in chapter 2 (van Son *et al.*, 1993c) we concluded that, on the basis of lip rounding and vertical lip opening, at least three Dutch vocalic viseme groups can be distinguished: unrounded and rounded monophthongs, and diphthongs (correct perception of vowel length in the rounded vowels extends the number to four visemes). The three basic visemes formed the baseline of our discrimination experiment:

[1] /i,I,e,ɛ,ɛi,a,ɑ/ unrounded vowels
[2] /u,y,œ,ɔ/, /ø,o/ rounded vowels (*short* and *long* subsets)
[3] /au,œy/ closing and rounding diphthongs.

Further discrimination within each viseme may result from acoustic information that is well-matched to the profoundly hearing impaired subjects' auditory capacities.

The basic objective of this study was to evaluate the efficacy of the two formant-based synthesized speech signals in comparison with an unprocessed amplified speech signal and with unaided lipreading. All stimuli (CVC syllables; see section 5.2.1) were processed as follows: voiced speech parts (the vowels) were represented by either F_1 or both F_1 and F_2, while the voiceless speech (consonantal contexts) were simply represented as silent periods. According to the rationale expressed in section 3.8, each formant was represented by two adjacent harmonics to the f_0 estimate, so that pitch is cued by the inter-harmonic distance (the *residue* effect). The speech conditions will further be referred to as AMPL (unprocessed), F1 and F1F2. The visual-only presentation mode will be regarded as a speech condition as well, referred to as VIS.

The second objective is to measure vowel discrimination both *across visemes* and *within visemes*. Near-perfect discrimination between vowels from different visemes should be possible purely on the basis of visible information. We expect that the addition of auditory information will have no influence on this. On the other hand, within-viseme discrimination would not be possible without supporting acoustic cues, because a viseme is conceptually characterized as a group of phonemes (in this case vowels) that have such visual similarities that they are perceived as one item. If listeners have residual capacities for processing temporal and spectral information, supplementary auditory information is expected to enhance the lipreaders' within-viseme vowel discrimination.

For a small number of within-viseme vowel pairs, the best discrimination may be expected when different acoustic codes are used. On the basis of F_1, discrimination between /u/ and /y/, /ɔ/ and /œ/, and /o/ and /ø/ is hardly possible, because the two vowels of a pair have similar first formants (see *e.g.* figure 4.1). When the second formant is added, discrimination is more likely, but the highest degree of discrimination is expected in the case that only the vowel with the highest second formant is indeed represented by two formants (F_1 and F_2 for vowels /y,œ,ø/), while its counterpart is represented by the first formant only (vowels /u,ɔ,o/). This expectation was based on our observation, in the harmonic complexes study described in chapter 3, that profoundly hearing impaired subjects were able to discriminate between patterns that differ as to the presence of high-frequency components, *e.g.* the difference between patterns 3 and 5 (figure 3.1). The same reasoning may hold for the vowel pair /ɑ,a/, in which both first and second formants are in each other's proximity (but they differ in length as well). The best discrimination will be expected by coding /ɑ/ by the F1 algorithm and /a/ by the F1F2 algorithm. The third objective of this study thus refers to what will be called *within-viseme*

different-codes discrimination.

Finally, performance will not only be evaluated in terms of discrimination scores. Physical vowel parameters will be measured and we will attempt to establish, by interpreting multidimensional scaling results (INDSCAL), which physical cues are actually used by the observer in order to make correct discriminations. The fourth objective is therefore to investigate the relationship between perceptual dimensions in vowel discrimination and the underlying physical cues.

5.2 Method

5.2.1 Preparation of the stimuli

Vowel discrimination was measured in a set of 15 Dutch vowels, twelve monophthongs and three diphthongs. Each vowel was paired once with all other vowels in a set (AB pairs). The comparison pairs were derived from three basic sets:

[1] 14 pairs for *across-viseme* comparisons:
unrounded /i/ *vs.* rounded /y,u/, 2 pairs
unrounded /e/ *vs.* rounded /ø,o/, 2 pairs
unrounded /ɛ/ *vs.* rounded /œ,ɔ/, 2 pairs
unrounded /ɛi/ *vs.* diphthongs /œy,au/, 2 pairs
unrounded, rounded, and diphthong /a,o,au/, 3 pairs
unrounded, rounded, and diphthong /ɑ,œ,œy/, 3 pairs

[2] 37 pairs for *within-viseme* comparisons:
unrounded vowels /i,I,e,ɛ,ɛi,a,ɑ/, 21 pairs
rounded vowels /ɔ,o,u,y,ø,œ/, 15 pairs
diphthongs /au,œy/, 1 pair

[3] 4 pairs for *different-codes* comparisons (code F1 is indicated by subscript 1, code F1F2 by subscript 2):
/u_1/ *vs.* /y_2/, 1 pair rounded
/o_1/ *vs.* /$ø_2$/, 1 pair rounded
/$ɔ_1$/ *vs.* /$œ_2$/, 1 pair rounded
/$ɑ_1$/ *vs.* /a_2/, 1 pair unrounded

This set was repeated three times, producing 12 vowel pairs.

The complete stimulus set was constructed as follows. First, the basic set of 63 vowel pairs was combined with four different consonantal contexts, symmetric /p_p/ and /k_k/, and asymmetric /p_k/ and /k_p/. Within the symmetric set and within the asymmetric set the original AB pairs were rearranged so as to yield 50% AB pairs and 50% BA pairs. Next, AB and BA pairs were extended to AXB and BXA triplets by

making X equal to the first or second vowel using a quasi-random balanced design. Finally, the complete list of 252 AXB-triplets was turned into a randomly ordered list and then divided into 10 smaller lists of 25 triplets (27 triplets in list 10).

Four extra lists of 25 AXB-triplets were created, comprising various within-viseme and across-viseme comparisons. During the measuring sessions, these were used as preamble lists for the four conditions AMPL, F1, F1F2, and VIS. In this way, the subject was familiarized to the stimulus type that was to follow.

5.2.2 Recording and processing of the speech material

The AXB-triplet lists were read aloud by a female speaker with standard Dutch pronunciation (ABN). A head-and-shoulders front view was colour videotaped by means of a Panasonic MS50 camera and a Panasonic NV-FS100 Super-VHS video recorder. The speaker's face was illuminated shadow-free; no special illumination of the oral cavity was used. The speaker pronounced all syllables of a triplet clearly, and with generally the same intonation and loudness. She kept at a distance of 50 cm of a B&K 4144 condenser microphone. The signal was amplified by means of a B&K 2608 measuring amplifier, after which it was passed to one of the hifi tracks on the videotape. The interval between the start of one triplet and the start of the next was 7 sec, as marked by a 1 kHz tone burst recorded onto the other track of the tape.

Next, each triplet was low-pass filtered at 4.4 kHz, digitized at a sample frequency of 10 kHz and stored into a file. Speech signals were analysed off-line using Entropics software running on a SUN workstation. The analysis produced RMS measures and estimates of fundamental frequency, voiced/unvoiced classification, and filter coefficients (LPC-12, providing formant frequencies). Analysis results were used to synthesize stimulus material for the F1 and F1F2 conditions: each formant was created by generating the two harmonics of f_0 that were closest to the estimated formant value, while the amplitude was based on the time envelope of the original speech waveform. The synthesis programmes were written by the second author (AJB). A manual correction of altogether 1.6% of the synthesized CVC syllables was applied. Finally, the files were processed in two ways, resulting in the two speech codes F1 and F1F2. For the *different-codes* comparisons of set [3] (section 5.3.1) special files were prepared with the proper combinations of F1 and F1F2 vowels.

Three videotapes were prepared, one for each auditory-visual

condition (AMPL, F1, F1F2), by copying the video signal and the tone burst onto new tapes. Subsequently, while copying the audio signals, the tone burst was used to ensure synchrony of sound and image. The maximum delay between the visible and audible signal, due to delay in tone burst detection, was less than 10 msec. This is amply within the critical delay time for asynchrony detection and subsequent performance decrease on the part of the viewer (about 40-80 msec; McGrath and Summerfield, 1985; Pandey *et al.*, 1986).

Not all comparisons (set [1]: *within-viseme*, set [2]: *across-viseme* and set [3]: *different-codes*) were eventually evaluated in all speech conditions; to be precise, an evaluation of the following comparisons was performed:

* Tape 1 (condition AMPL): only comparisons from sets [1] and [2] were evaluated.
* Tape 2 (condition F1): comparisons from all three sets were evaluated. Moreover, set [3] comparisons were compared with the corresponding vowel comparisons of set [2], which we will henceforth refer to as *equal-codes* comparisons. For tape 2 this means comparing F1 versus F1/F1F2 discriminations.
* Tape 2 (condition F1F2): comparisons from all three sets were evaluated. Here also, *different-codes* discrimination was compared with *equal-codes* discrimination for corresponding vowel pairs: F1F2 versus F1/F1F2 discriminations.

5.2.3 Vowel parameters

Vowel duration, signal level and the formant frequencies F_1 and F_2 were measured semi-automatically in all 756 unprocessed vowels for comparison with the perceptual data. Vowel duration was defined to be the length of the signal in between the silent gaps of the plosive consonants /p,k/, and the RMS level was computed for this interval. All measurements were performed by the first author (NvS).

Semi-automatic measuring involves subjective decisions on the part of the data analyst, *e.g.* as to the exact begin and end points of the vowel, and the decision whether a maximum in the spectral envelope indeed indicates a correct formant position. In order to check the reliability of the analysis, 120 vowels (16%) were measured independently by a second analyst. Correlated *t* tests revealed that there was no difference between both measures of level and formant frequencies. The second analyst's vowel durations, however, were found to be significantly longer: $t(14)=3.6$, $p<.01$. The amount is negligible, though: 6 msec.

Table 5.1 Mean values (and standard deviations in parentheses) for duration, level, and formant frequencies F_1 and F_2 of 12 Dutch monophthong vowels and 3 diphthongs, pronounced by a female speaker. For duration and level, vowels are given in ascending orders. Formant transitions in the diphthongs are expressed by giving formant frequencies for the first and second part.

V	duration [msec]		V	level [dB]		V	F_1 [Hz]		F_2 [Hz]	
/ɪ/	95.8	(10.3)	/ɛi/	71.2	(1.4)	/i/	321	(31)	2417	(103)
/i/	96.3	(14.2)	/y/	71.4	(1.1)	/ɪ/	499	(23)	2151	(142)
/u/	103.5	(13.3)	/i/	71.5	(1.4)	/e/	515	(17)	2223	(173)
/y/	104.7	(13.5)	/ɪ/	71.6	(1.1)	/ɛ/	629	(95)	1874	(70)
/ɔ/	107.1	(13.7)	/u/	71.6	(1.8)	/a/	995	(73)	1617	(56)
/ɑ/	108.7	(12.3)	/ɛ/	72.1	(1.2)	/ɑ/	855	(90)	1400	(94)
/œ/	108.8	(12.8)	/au/	72.1	(1.2)	/ɔ/	594	(64)	1026	(70)
/ɛ/	112.5	(11.8)	/œy/	72.1	(1.3)	/o/	529	(54)	1015	(43)
/e/	178.3	(19.1)	/a/	72.2	(1.1)	/u/	348	(43)	845	(48)
/o/	193.1	(17.8)	/ɑ/	72.4	(1.4)	/y/	351	(48)	1786	(94)
/ɛi/	195.2	(18.5)	/e/	73.2	(1.5)	/ø/	523	(32)	1659	(53)
/a/	196.5	(17.0)	/o/	73.3	(1.1)	/œ/	520	(27)	1688	(90)
/ø/	198.4	(17.9)	/œ/	73.3	(2.2)	/ɛi/	670	(108)	1887	(53)
/au/	203.1	(21.6)	/ø/	73.4	(1.2)		505	(75)	2171	(217)
/œy/	207.1	(19.2)	/ɔ/	74.1	(1.2)	/au/	841	(109)	1410	(71)
							486	(136)	956	(112)
						/œy/	826	(135)	1631	(53)
							502	(44)	1648	(76)

Table 5.1 presents mean values and standard deviations for vowel duration, level, F_1 and F_2. For both duration and level, the vowels are ranked in an ascending order. As for duration, the normal division into short and long vowels shows. Two minor comments can be made: first, vowels /i,ɪ/ may be labelled "very short", as *t* statistics for uncorrelated samples show that they are significantly shorter than /u,y,ɔ,ɑ,œ,ɛ/; second, /e/ is significantly shorter than the other long vowels - as the difference with the next vowel /o/ is about 15 msec, while the whole range from /o/ to /œy/ is only 14 msec, /e/ may perhaps be called "half-long". With respect to the diphthongs, for /ɛi,au,œy/ formant values are given for the first and second part. In figure 5.1 all individual vowel realizations by our female speaker are represented in

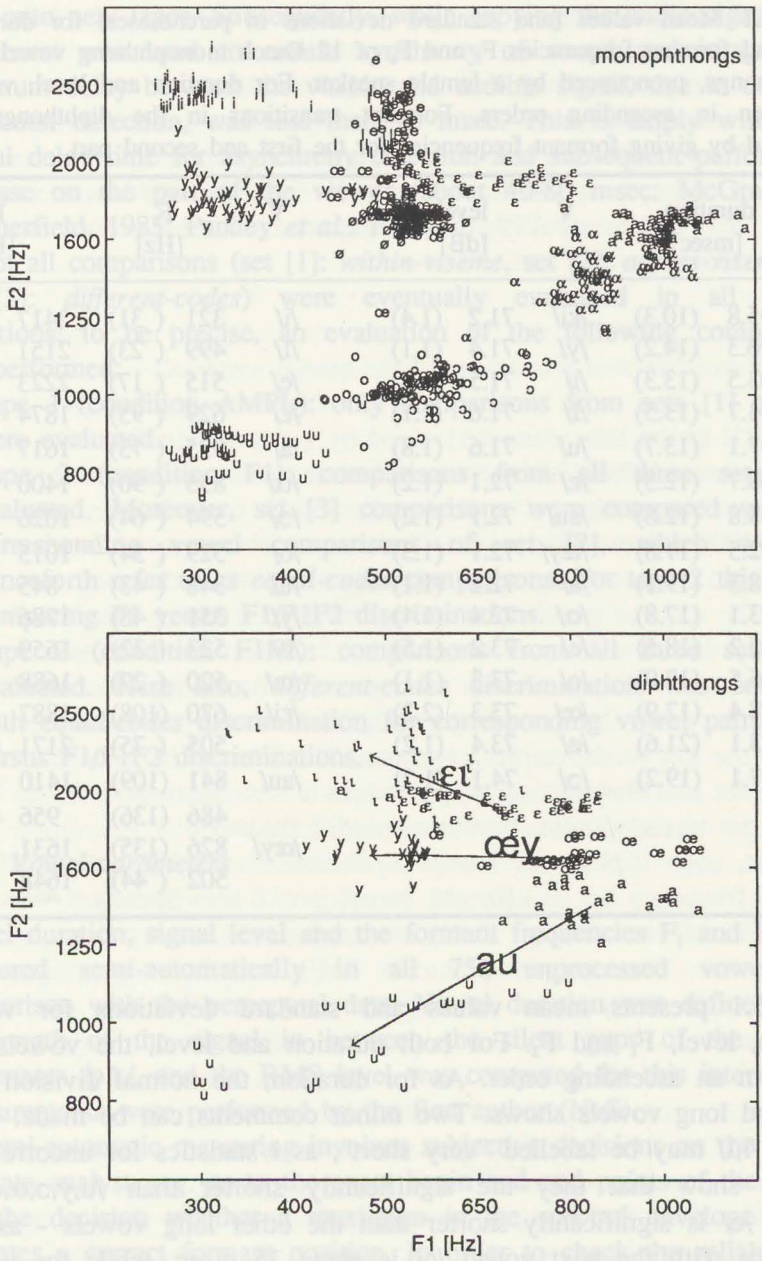

Figure 5.1 F_1/F_2 spaces of 12 Dutch monophthongs (top panel) and 3 diphthongs (bottom panel). All individual vowel realizations by our female speaker are placed in the space. The transitions in the diphthongs are represented by the three vectors. In both panels logarithmically divided frequency axes are used.

F_1/F_2 spaces. The monophthong vowel positions of our female speaker agree well with those of the average female speaker as given by van Nierop *et al.* (1973). The only exceptions are the positions of /i,I/, shifted to a slightly higher F_1 range, and those of /a,ɑ/, which have shifted to higher F_1 and F_2 ranges, thus taking different positions from the van Nierop *et al.* (1973) results.

5.2.4 Subjects

A group of 12 profoundly hearing impaired subjects participated in the experiment. Figure 5.2 shows their hearing thresholds (THR), most comfortable levels (MCL) and levels of discomfort (UCL) at octave frequencies from 125 Hz to 4 kHz, measured in this study according to the following method. Stimulus tones were generated by a digital audio signal processor (van Raaij, 1986) and presented through Beyer DT 48 headphones in supra-aural cushions. The threshold was measured by means of a simple *up-down* procedure (staircase method; Levitt, 1971). MCL was obtained from loudness ratings as indicated by the subject on a modified version of the Würzburger loudness triangle (Moser, 1987). UCL was measured by presenting a series of tones, with increasing intensity, using 1 dB steps. The tone series was terminated when the subject indicated that loudness became uncomfortable.

Table 5.2 contains a number of relevant data like age, sex, average hearing loss, mean dynamic range, years and origin of impairment. As to the audiometric data, PTA and LFPTA are correlated significantly at the 1% level (r=.82), and so are the corresponding mean dynamic ranges (r=.84). Furthermore, there is a moderate negative correlation between the average hearing thresholds and the dynamic ranges, both mean and at various frequencies; r values are about -.56 ($p<.05$). Within the dynamic range measures, the highest correlations were found between average dynamic range at PTA frequencies and the dynamic range at 1 kHz (r=.94, 89% explained variance), and between average dynamic range at LFPTA frequencies and the dynamic range at 500 Hz: r=.97, 94% explained variance.

Nine of the subjects had also taken part in the experiment on harmonic complexes (van Son *et al.*, 1993a, chapter 3). For this reason, the original numbering of subjects is used: S01 to S11 are the same as those in table 3.1, S04 and S08 did not participate this time, while S12 to S14 are 'new' subjects.

A control group of six normally hearing subjects performed the discrimination task in a limited number of conditions. Their hearing

levels were normal. All subjects, normally hearing and profoundly hearing impaired, had normal or corrected-to-normal vision, and were paid for their services. Most of the profoundly hearing impaired and three of the normally hearing subjects had ample experience in experimental settings involving auditory tasks; about half of them had also participated in previous lipreading studies.

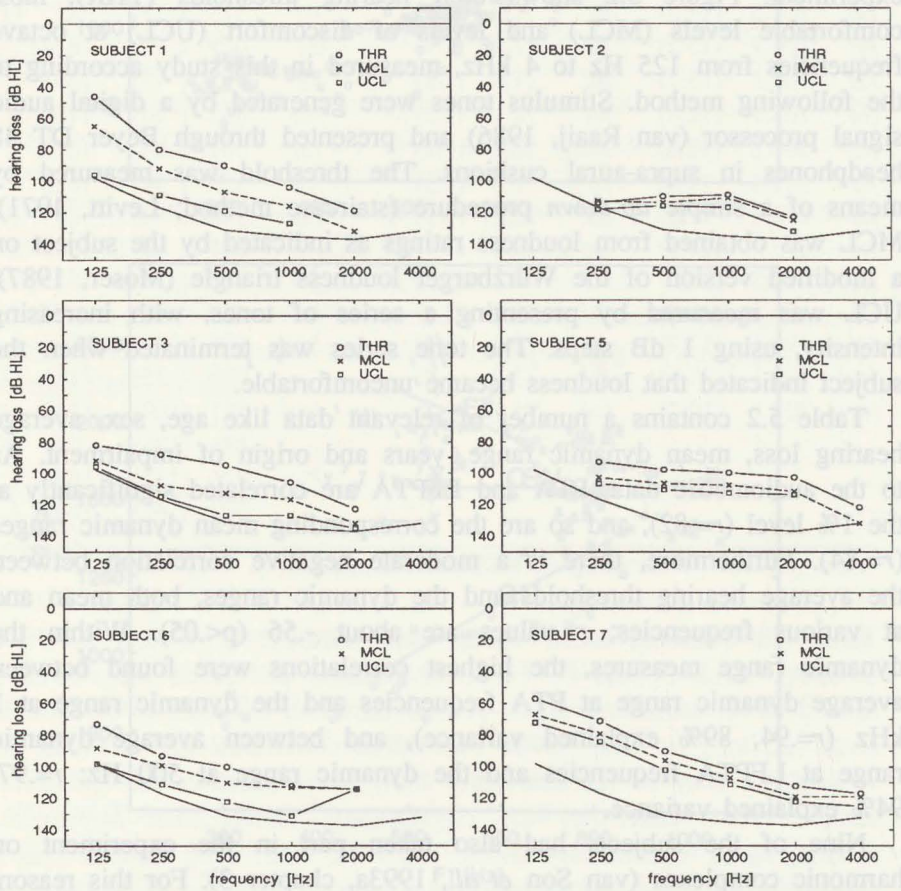

Figure 5.2a Hearing thresholds (THR), most comfortable levels (MCL) and levels of discomfort (UCL) for six profoundly hearing impaired subjects (S01 to S07). The bottom graph of each audiogram indicates the maximally measurable threshold imposed by the Beyer DT 48 headphone.

5.2.5 Design

In the auditory-visual condition, stimulus tapes (AMPL, F1, F1F2) were presented individually to each of the 12 profoundly hearing impaired subjects. In order to minimize order and learning effects, tapes were presented in different orders: each of six possible orders (3!) were presented to two subjects. With each tape a different order of AXB lists was replayed. Three or four lists were kept together, so that there was no need for forward and backward searches every three minutes.

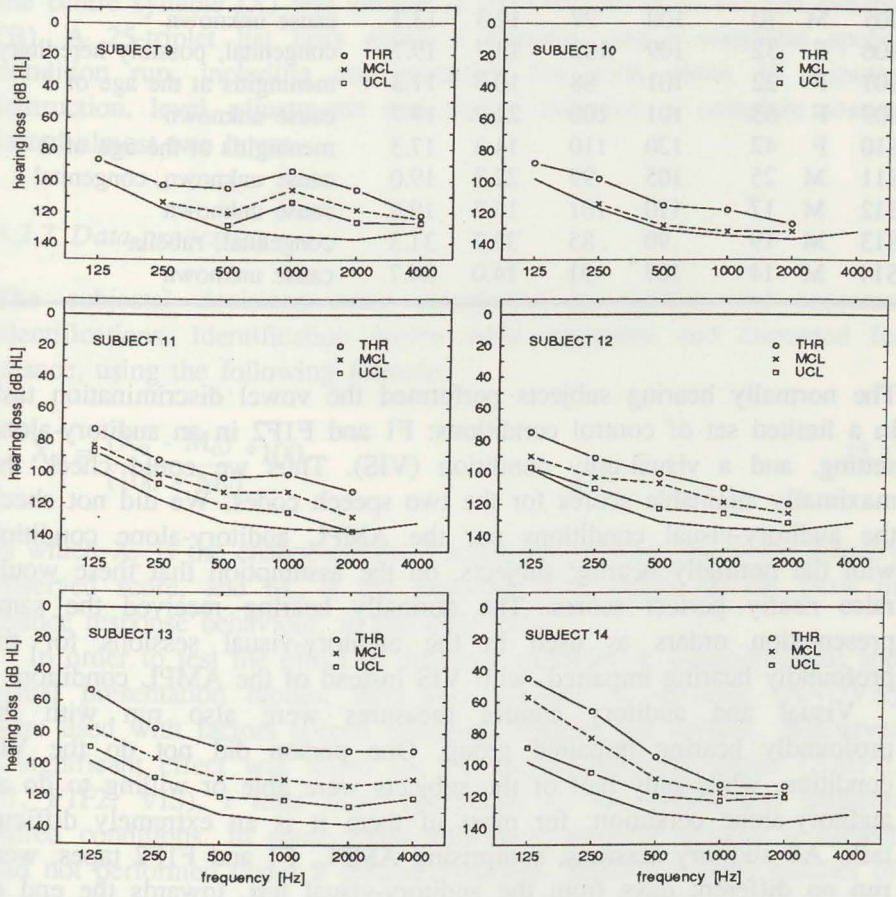

Figure 5.2b Hearing thresholds (THR), most comfortable levels (MCL) and levels of discomfort (UCL) for six profoundly hearing impaired subjects (S09 to S14). The bottom graph of each audiogram indicates the maximally measurable threshold imposed by the Beyer DT 48 headphone.

Table 5.2 Relevant data of 12 profoundly hearing impaired subjects. Hearing loss is expressed in dB HL as the pure-tone average (PTA) of .5, 1 and 2 kHz, and as the low-frequency PTA (LFPTA) of .25, .5 and 1 kHz. Mean dynamic ranges in dB are given at the PTA and LFPTA frequencies.

subj	sex	age	hearing loss		dynamic range		cause and duration of impairment
			PTA	LFPTA	PTA	LFPTA	
S01	F	46	104	91	23.3	27.0	cause unknown
S02	M	37	114	111	7.7	6.0	cause unknown, congenital
S03	M	59	108	96	21.7	26.7	cause unknown
S05	M	70	100	97	12.3	13.3	cause unknown
S06	F	32	109	102	13.7	19.7	congenital, possibly hereditary
S07	F	22	101	88	10.3	11.3	meningitis at the age of 3
S09	F	63	101	100	21.7	19.7	cause unknown
S10	F	42	120	110	14.3	17.3	meningitis at the age of 2
S11	M	25	105	99	22.3	19.0	cause unknown, congenital
S12	M	17	110	101	16.7	19.0	cause unknown
S13	M	19	90	85	32.7	31.3	congenital, rubella
S14	M	14	107	91	14.0	24.7	cause unknown

The normally hearing subjects performed the vowel discrimination task in a limited set of control conditions: F1 and F1F2 in an auditory-alone setting, and a visual-only condition (VIS). Thus we could check the maximally attainable scores for the two speech codes. We did not check the auditory-visual conditions nor the AMPL auditory-alone condition with the normally hearing subjects, on the assumption that these would raise nearly perfect scores. The normally hearing received the same presentation orders as used in the auditory-visual sessions for the profoundly hearing impaired, with VIS instead of the AMPL condition.

Visual and auditory control measures were also run with the profoundly hearing impaired group. One person did not do the VIS condition, while only half of the subjects were able or willing to do an auditory-alone condition; for most of them it is an extremely difficult task. All auditory sessions, comprising AMPL, F1 and F1F2 tapes, were run on different days from the auditory-visual test, towards the end of the study (after the auditory-visual conditions had already been worked through).

5.2.6 *Experimental procedure*

At the start of the vowel discrimination task, the subject was given a printed instruction form with information about the complete procedure. The subject watched the video on a Philips RGB monitor (Matchline). In order to ensure that acoustic information was presented at supra-threshold level at all frequencies, speech spectra were shaped according to the MCL curves using a Boss GE-131 1/3-octave equalizer. The output signal was presented monaurally, at a comfortable listening level, to the subject's best ear through Beyer DT 48 supra-aural headphones.

The subject watched and listened to the triplet and decided whether the centre syllable (X) was similar to the first (A) or to the last syllable (B). A 25-triplet list lasts about 3 minutes, and a complete speech condition run, including one preamble list, took about 35 minutes. Instruction, level adjustments and pauses included, a complete session lasted almost two hours.

5.2.7 *Data processing*

The subjects' decisions were interpreted as correct and incorrect identifications. Identification scores were computed and corrected for chance, using the following formula:

$$X_c = \frac{(X_o - M_G)}{(100 - M_G)} * 100 \qquad [5.1]$$

in which X_c is the chance-corrected score (in percent), X_o is the actually observed score, and M_G is the expected mean score on the basis of chance response behaviour, in this case 50%.

In order to test the effect of different variables in auditory-visual and visual presentation modes, three-way analyses of variance (ANOVA) were used with factors speech condition, subject and context. Whenever a significant effect was found for the speech condition factor (AMPL, F1, F1F2, VIS), *t* tests were performed to test differences between paired conditions. In some cases data were missing because a subject had not performed under a certain condition. When the same number of persons had been subjected to the conditions, correlated *t* tests were performed; with unequal numbers, uncorrelated *t* tests had been used. All *t* tests are based on 4 context scores from 11 or 12 subjects, except where noted.

Apart from discrimination scores, error patterns were analysed. Vowel confusions were collected into matrices, one for each subject in each

condition. The confusions were regarded as perceptual distances (similarities), and matrices were converted into symmetric similarity matrices using an algorithm proposed by Houtgast (Klein *et al.*, 1970). Subsequently, the similarity matrices were subjected to INDSCAL analysis (Carroll and Chang, 1970), which provides an *n*-dimensional perceptual space in which the stimuli are arranged according to certain properties or characteristics they possess. For the interpretation of these INDSCAL dimensions, the parameter values (table 5.1) of vowel length and signal level and the logarithmically transformed formant frequencies were correlated with the positions of vowels along the perceptual dimensions found (Pearson's product-moment correlation coefficient *r*).

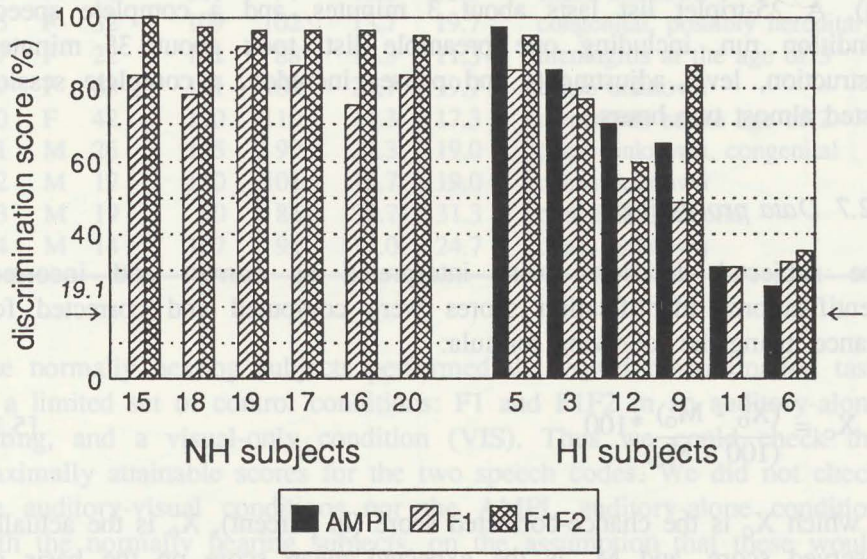

Figure 5.3 Mean within-viseme discrimination scores of 6 normally hearing and 6 profoundly hearing impaired subjects in an auditory-only condition. The arrows indicate the 99% chance levels (CL99) for chance-corrected scores, based on the normal approximation to the binomial distribution.

5.3 Control results: auditory discrimination

5.3.1 Discrimination scores

The mean subject scores in figure 5.3 are based on within-viseme discriminations (set [2]) pooled across contexts (no context effects had been found). The arrows indicate the 99% chance levels (CL99) for chance-corrected scores, based on the normal approximation to the

binomial distribution. This score assumes that all subjects guessed on all 148 within-viseme AXB's. It was computed by means of the following rule (Boothroyd, 1988):

$$CL99 = z*\sqrt{[p(1-p)/n]}*100*g \qquad [5.2]$$

in which z indicates the z-score below which 99% of the scores in the normal distribution will be found (z=2.325), p is the chance probability of a correct discrimination (p=.5), n is the total number of items on which a subject's mean score is based (148), and g is a multiplying factor, due to chance correction (g=2).

In the figure, the normally hearing subjects (NH) are rank ordered as to their F1F2 scores. It is clear that perfect discrimination is possible with the F1F2 code (97%), and that the F1 coding scheme allows reasonably high scores, around 83% on an average.

The profoundly hearing impaired subjects (HI) are rank ordered as to their performance on the AMPL condition. Subject S11 did not take the F1F2 auditory-alone condition. Even with this difficult presentation mode, most mean scores are well beyond the CL99 level. The HI subjects show much more variability than the NH subjects, an observation which will be discussed in section 5.6.1.

5.3.2 Perceptual dimensions

INDSCAL analysis of auditory within-viseme vowel confusions yielded one-dimensional solutions for the three speech conditions. Stimulus and subject spaces for unrounded and rounded vowels are plotted in figure 5.4. Because the group is quite small (n=6 in AMPL and F1, n=5 in F1F2), the total percentage of explained variance is easily influenced by an outlying individual, as can be seen in the subject plots for the unrounded AMPL and rounded F1F2 vowels. In those cases S05 has a zero weight on the dimension, probably because of the lack of confusions (see this subject's scores in figure 5.3). With this subject's data removed, the percentages of explained variance changed to 61% (AMPL, unrounded vowels) and 78% (F1F2, rounded vowels), respectively.

For interpretation of the perceptual dimension, plot co-ordinates of the vowels were correlated with the four parameters; the outcome was rather ambiguous in that both F_1 and vowel duration explained a significant part of the variance of vowel positions along dimension I (ranging from 53% to 92% explained variance). INDSCAL was therefore run again for a two-dimensional solution, but this did not reveal a

clearer picture: dimension I still correlated significantly with the same parameters, while dimension II was only correlated weakly with duration in the F1 condition for unrounded vowels and the F1F2 condition for the rounded vowels. In general, then, auditory vowel discrimination is mainly based on first formant and vowel duration information.

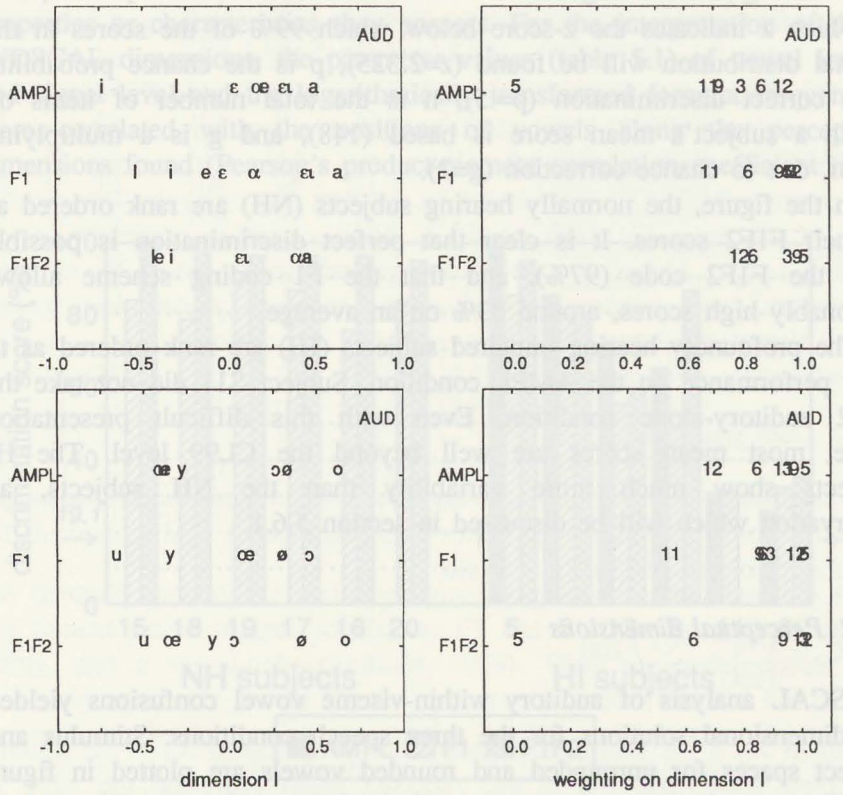

Figure 5.4 One-dimensional INDSCAL solutions of auditory within-viseme vowel confusions, based on the results of 6 profoundly hearing impaired subjects (for F1F2, n=5). Top panels: stimulus and subject spaces for 7 unrounded vowels. Explained variance: 50.8% (AMPL), 76.1% (F1) and 80.8% (F1F2). Bottom panels: stimulus and subject spaces for 6 rounded vowels. Explained variance: 78.4% (AMPL), 71.9% (F1) and 62.3% (F1F2).

5.4 Results: visual and auditory-visual discrimination scores

5.4.1 Across-viseme discrimination

Figure 5.5 shows across-viseme discrimination scores for four contexts in four speech conditions. Mean points for AMPL, F1 and F1F2 are

based on the scores of 12 subjects, that of VIS on those of 11 subjects (S02 did not participate in the visual-only condition). The means and standard deviations in the top of the figure are those computed across subjects and contexts. In the figure the 99% chance levels (CL99) for chance-corrected scores are indicated, assuming that all subjects guessed on all 14 pairs within an across-viseme context list.' It was computed by application of formula [5.2], with n=168 for auditory-visual conditions and n=154 for the VIS condition.

All scores are extremely high and thus well above chance level. This indicates that across-viseme discrimination, as expected, is nearly perfect. Only one significant effect was found, that for the subject factor: F(11,33)=5.3, p=.00.

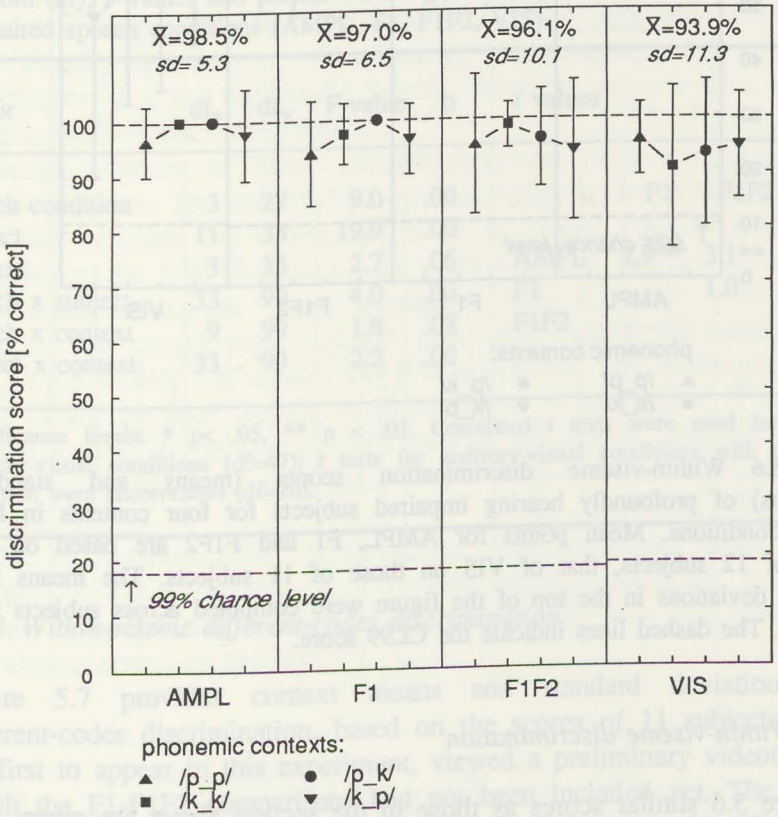

Figure 5.5 Across-viseme discrimination scores (means and standard deviations) of profoundly hearing impaired subjects for four contexts in four speech conditions. Mean points for AMPL, F1 and F1F2 are based on the scores of 12 subjects, that of VIS on those of 11 subjects. The means and standard deviations in the top of the figure were computed across subjects and contexts. The dashed lines indicate the CL99 score level.

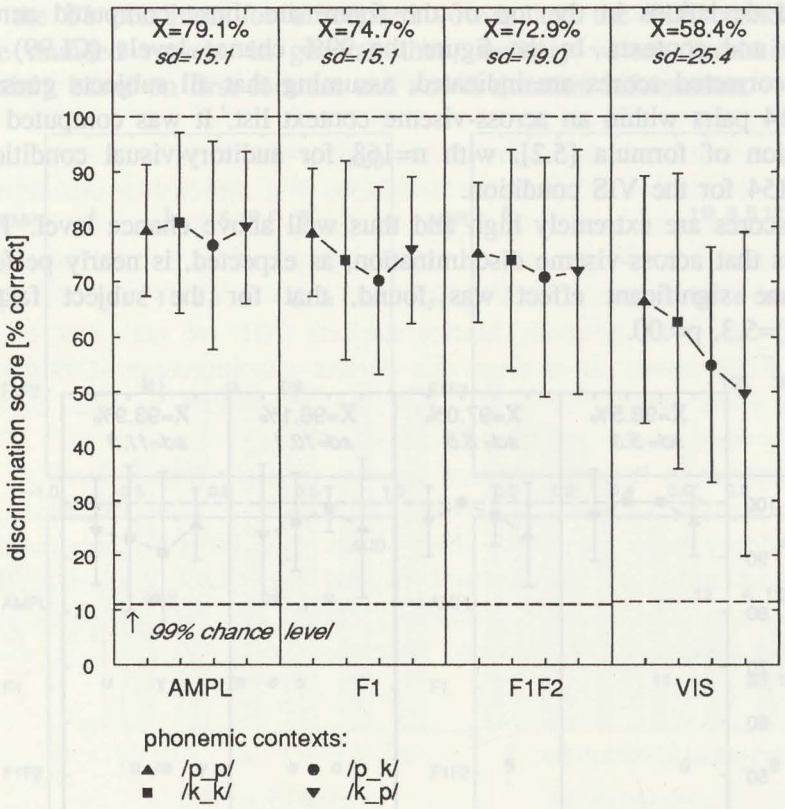

Figure 5.6 Within-viseme discrimination scores (means and standard deviations) of profoundly hearing impaired subjects for four contexts in four speech conditions. Mean points for AMPL, F1 and F1F2 are based on the scores of 12 subjects, that of VIS on those of 11 subjects. The means and standard deviations in the top of the figure were computed across subjects and contexts. The dashed lines indicate the CL99 score.

5.4.2 Within-viseme discrimination

In figure 5.6 similar scores as those in the section above are given, but this time for vowel comparisons within visemes. CL99 was computed with n=444 for the auditory-visual conditions and with n=407 for the visual condition (there are 37 items in each context list). Again, mean scores are well above chance level in all four speech conditions. The results of a three-way ANOVA and subsequent *t* tests are summarized in table 5.3. Two main effects are significant, those for speech conditions

and for subjects. All auditory-visual conditions are significantly better than the visual condition; within the auditory-visual conditions, AMPL is better than the two speech codes, which do not differ from each other beyond chance level. Furthermore, two interaction effects were found to be significant, both involving the subject factor. The speech-by-subject interaction is not significant when only the three auditory-visual conditions are analysed together, so it is mainly due to the large subject variability in the VIS condition.

Table 5.3 Results from a three-way ANOVA, giving factors, degrees of freedom (df), F-values and critical values (p). Subsequent *t* test results are given for paired speech conditions (AMPL, F1, F1F2, VIS).

Factor	df_B	df_W	F-value	p	*t* values			
						F1	F1F2	VIS
speech condition	3	27	9.0	.00				
subject	11	33	19.9	.00				
context	3	33	2.7	.06	AMPL	2.9**	3.1**	4.8**
speech x subject	33	99	4.0	.00	F1		1.0	3.8**
speech x context	9	99	1.8	.08	F1F2			3.1*
subject x context	33	99	2.2	.00				

Significance levels: * p< .05, ** p < .01. Correlated *t* tests were used for paired auditory-visual conditions (df=47); *t* tests for auditory-visual conditions with the VIS condition were uncorrelated (df=90).

5.4.3 Within-viseme different-codes discrimination

Figure 5.7 provides context means and standard deviations for different-codes discrimination, based on the scores of 11 subjects: S11, the first to appear in this experiment, viewed a preliminary videotape in which the F1-F1F2 comparisons had not been included yet. The CL99 score was consequently computed with n=132 (11 subjects by 12 items). For the sake of comparison, figure 5.7 also gives mean scores for the different-codes comparisons' counterparts, *equal-codes* comparisons: the same vowel pairs but without the F1-F1F2 opposition (so only F1 or only F1F2 AXB's). As these pairs were unequally distributed over context lists, only overall means across subjects and contexts can be

given. Furthermore, context means are provided for the same vowel pairs presented in the VIS condition.

As the pairs of differently coded vowels in the two coding conditions are identical, no statistical effect was expected for this speech code factor. However, a *t* test for correlated samples indicated that there was a difference at the .01 significance level: $t(43)=2.7$. Furthermore, differently coded vowels in the F1 condition were significantly better discriminable than their VIS counterparts: uncorrelated $t(86)=2.3$, $p<.05$. Otherwise, there were no differences in discrimination among the various conditions, whether vowels in a pair were differently or equally coded. As the performance of individuals differs so much in this within-viseme different-codes condition, further comments are left to section 5.6.2.

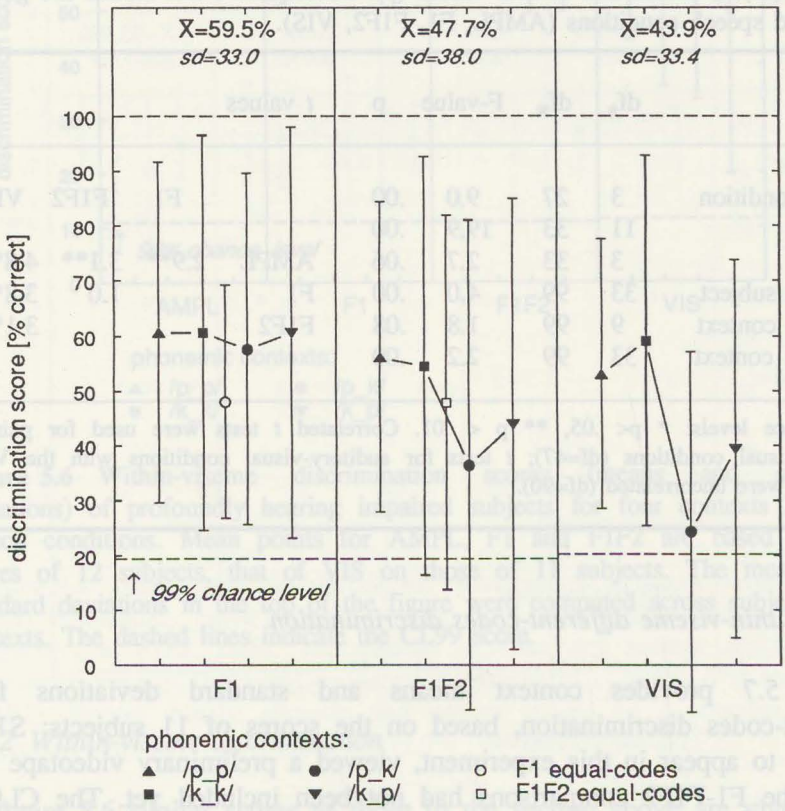

Figure 5.7 Within-viseme different-codes discrimination scores (means and standard deviations) of 11 profoundly hearing impaired subjects in F1 and F1F2 speech conditions. As a reference, mean and standard deviations are given for their equal-codes counterparts in F1 and F1F2 conditions, and for the context lists in the visual-only (VIS) condition. The means and standard deviations in the top of the figure were computed across subjects and contexts. The dashed lines indicate the CL99 score.

5.5 Results: visual and auditory-visual vowel confusions

5.5.1 Confusion matrices and INDSCAL dimensions

Due to the high number of correct discriminations, the vowel confusion matrices presented in the following sections contain relatively few off-diagonal entries. This holds especially for the across-viseme comparisons and, of course, the within-viseme different-codes comparisons, because they could only be confused pair-wise. As a consequence, for the across-viseme and within-viseme different-codes comparisons we will only present matrices, whereas for the within-viseme comparisons both matrices and INDSCAL solutions will be presented. Separate analyses were run on the unrounded and rounded vowel visemes, based on individual subject matrices. The relation between perceptual dimensions and physical parameters will give an indication of the source of information that our profoundly hearing impaired subjects have actually used in their vowel discrimination.

In all matrices to follow, stimulus/response pairs were aggregated across contexts. The utmost left column indicates the stimuli (X), the top line indicates the responses (A or B). Cells contain percentages confusions; the bold figures in the main diagonal indicate percentages correct discriminations. These percentages have not been corrected for chance. The utmost right column gives the number of presentations (N_{PRES}) on which the scores and confusions for that particular vowel were based; the total number of presentations of all vowels is indicated by the number in the lower right corner of the matrix. The bottom line, finally, gives the number of times that a phoneme was returned by the subjects (N_{RESP}).

For each vowel, 99% chance scores were computed, based on the assumption that all subjects were guessing on all AXB items. Again, the normal approximation to the binomial distribution was used, computing the chance scores as follows:

$$CL99 = M_G + z*\sqrt{[n*p(1-p)]} \qquad [5.3]$$

CL99 is the 99% score expected from chance, M_G is the mean percentage expected from chance, z indicates the z-score below which 99% (z=2.325) of the scores in the normal distribution will be found, p is the chance probability of a correct discrimination, and n is the total number of items on which the scores are based. All these statistics are summarized in table 5.4, so that it is quite easy to observe which scores are above chance level.

Table 5.4. Percentages vowel scores attainable by chance. The following statistics are shown for three comparison modes: chance probabilities of correct discriminations (p), the total number of stimulus presentations (N), expected mean scores and standard deviations (M and sd), and chance scores for 99% confidence levels (CL99).

Vowel set	p	N	M (%)	sd (%)	CL99 (%)
		auditory-visual (12 subjects)			
across-visemes					
/i,e,ɛ,ɛi,a,ɑ/	1/3	48	33.3	6.8	49.2
/ɔ/	1/3	24	33.3	9.6	55.7
/u,y,ø/	1/2	24	50.0	10.2	73.7
/o,œy,au/	1/5	72	20.0	4.6	30.3
/œ/	1/2	72	50.0	5.9	63.7
within-visemes					
/i,I,e,ɛ,ɛi,a,ɑ/	1/7	144	14.3	2.9	21.1
/ɔ,o,u,y,ø,œ/	1/6	120	16.7	3.4	24.6
/œy,au/	1/2	24	50.0	10.2	73.7
within-visemes different-codes/equal-codes					
/a,ɑ,ɔ,œ,o,ø,u,y/	1/2	66[1]	50.0	6.2	64.3
		visual-only (11 subjects)			
across-visemes					
/i,e,ɛ,ɛi,a,ɑ/	1/3	44	33.3	7.1	49.9
/ɔ/	1/3	22	33.3	10.1	56.7
/u,y,ø/	1/2	22	50.0	10.7	74.8
/o,œy,au/	1/5	66	20.0	4.9	31.4
/œ/	1/2	66	50.0	6.2	64.3
within-visemes					
/i,I,e,ɛ,ɛi,a,ɑ/	1/7	132	14.3	3.0	21.4
/ɔ,o,u,y,ø,œ/	1/6	110	16.7	3.6	24.9
/œy,au/	1/2	22	50.0	10.7	74.8
within-visemes different-codes/equal-codes					
/a,ɑ,ɔ,œ,o,ø,u,y/	1/2	66	50.0	6.2	64.3

[1] Scores are based on 11 subjects: S11 did not participate in this condition.

5.5.2 Across-viseme vowel confusions

Whatever speech condition was presented, discrimination of vowels belonging to different visemes was always nearly perfect. As an example of (the lack of) confusions we will only provide the matrix of the 'worst' condition, VIS (table 5.5). This matrix clearly indicates that lipreaders have no difficulty in recognizing the three different visemes.

Table 5.5 Vowel confusion matrix of across-viseme discrimination in the VIS speech condition. The overall mean score (11 subjects) is 96.9%.

stim \ resp	i	I	e	ɛ	ɛi	a	ɑ	u	y	ɔ	œ	o	ø	œy	au	N_{PRES}
i	**100.0**	.	.	.	.	.	.	.	.	.	.	.	.	.	.	44
I	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0
e	.	**.100.0**	.	.	.	.	.	.	.	.	.	.	.	.	.	44
ɛ	.	.	**97.7**	.	.	.	2.3	.	.	.	.	.	.	.	.	44
ɛi	.	.	.	**.100.0**	.	.	.	.	.	.	.	.	.	.	.	44
a	.	.	.	.	**.100.0**	.	.	.	.	.	.	.	.	.	.	44
ɑ	.	.	.	.	.	**93.2**	.	.	.	.	.	.	4.6	2.3	.	44
u	.	.	.	.	.	.	**.100.0**	.	.	.	.	.	.	.	.	22
y	.	.	.	.	.	1.5	.	**96.9**	.	.	.	.	.	.	1.5	66
ɔ	.	.	.	.	.	.	.	.	**.100.0**	.	.	.	.	.	.	22
œ	9.1	.	.	.	.	.	.	.	.	**90.9**	.	.	.	.	.	22
o	.	.	9.1	.	.	.	.	.	.	.	**90.9**	.	.	.	.	22
ø	.	.	.	1.5	.	.	1.5	.	.	.	.	**97.0**	.	.	.	66
œy	.	.	.	.	.	.	1.5	.	.	.	.	3.1	**95.5**	.	.	66
au	.	.	.	.	1.5	.	.	.	4.6	.	.	.	.	.	**93.9**	66
N_{RESP}	46	0	46	44	45	45	43	23	67	22	20	20	68	64	63	616

Table 5.6 Vowel confusion matrix of within-viseme discrimination in the AMPL speech condition. The overall mean score (12 subjects) is 89.5%.

stim \ resp	i	I	e	ɛ	ɛi	a	ɑ	u	y	ɔ	œ	o	ø	œy	au	N_{PRES}
i	**97.3**	1.4	.	.	1.4	.	.	.	.	.	.	.	.	.	.	144
I	4.2	**90.3**	3.5	0.7	.	.	1.4	.	.	.	.	.	.	.	.	144
e	2.8	3.5	**82.0**	2.8	3.5	3.5	2.1	.	.	.	.	.	.	.	.	144
ɛ	1.4	2.8	3.5	**88.9**	2.8	.	0.7	.	.	.	.	.	.	.	.	144
ɛi	0.7	1.4	3.5	2.8	**90.3**	1.4	.	.	.	.	.	.	.	.	.	144
a	0.7	.	1.4	3.5	4.2	**87.5**	2.8	.	.	.	.	.	.	.	.	144
ɑ	.	1.4	1.4	1.4	.	.	**95.8**	.	.	.	.	.	.	.	.	144
u	.	.	.	.	.	.	.	**93.3**	.	0.8	.	3.3	2.5	.	.	120
y	.	.	.	.	.	.	.	1.7	**91.7**	0.8	0.8	3.3	1.7	.	.	120
ɔ	.	.	.	.	.	.	.	3.3	.	**91.7**	1.7	.	3.3	.	.	120
œ	.	.	.	.	.	.	.	0.8	0.8	0.8	**95.0**	1.7	.8	.	.	120
o	.	.	.	.	.	.	.	0.8	6.7	1.7	1.7	**83.3**	5.8	.	.	120
ø	.	.	.	.	.	.	.	3.3	.	0.8	5.8	.	**90.0**	.	.	120
œy	.	.	.	.	.	.	.	.	.	.	.	.	.	**45.8**	54.2	24
au	.	.	.	.	.	.	.	.	.	.	.	.	.	37.5	**62.5**	24
N_{RESP}	154	145	137	146	145	133	148	124	119	116	126	110	125	20	28	1776

Table 5.7 Vowel confusion matrix of within-viseme discrimination in the F1 speech code. The overall mean score (12 subjects) is 87.3%.

stim \ resp	i	I	e	ɛ	ɛi	a	ɑ	u	y	ɔ	œ	o	ø	œy	au	N_{PRES}
i	**95.2**	2.1	1.4	1.4	.	.	.	.	.	.	.	.	.	.	.	144
I	8.3	**82.7**	4.9	0.7	2.8	.	0.7	.	.	.	.	.	.	.	.	144
e	2.8	2.8	**88.9**	2.1	2.1	1.4	.	.	.	.	.	.	.	.	.	144
ɛ	2.8	4.9	1.4	**83.3**	2.8	.	4.9	.	.	.	.	.	.	.	.	144
ɛi	.	1.4	3.5	4.2	**86.1**	3.5	1.4	.	.	.	.	.	.	.	.	144
a	2.1	.	0.7	2.1	5.6	**85.4**	4.2	.	.	.	.	.	.	.	.	144
ɑ	.	2.1	0.7	1.4	1.4	.	**94.5**	.	.	.	.	.	.	.	.	144
u	.	.	.	.	.	.	.	**90.8**	1.7	.	.	4.2	3.3	.	.	120
y	.	.	.	.	.	.	.	3.3	**90.8**	0.8	.	2.5	2.5	.	.	120
ɔ	.	.	.	.	.	.	.	.	.	**86.7**	9.2	0.8	3.3	.	.	120
œ	.	.	.	.	.	.	.	.	0.8	6.7	**92.5**	.	.	.	.	120
o	.	.	.	.	.	.	.	.	9.2	.	2.5	**85.0**	3.3	.	.	120
ø	.	.	.	.	.	.	.	5.8	0.8	0.8	5.0	0.8	**86.7**	.	.	120
œy	.	.	.	.	.	.	.	.	.	.	.	.	.	**54.2**	45.8	24
au	.	.	.	.	.	.	.	.	.	.	.	.	.	50.0	**50.0**	24
N_{RESP}	160	138	146	137	145	130	152	120	124	114	131	112	119	25	23	1776

Table 5.8 Vowel confusion matrix of within-viseme discrimination in the F1F2 speech code. The overall mean score (12 subjects) was 86.4%.

	resp														
stim	i	I	e	ɛ	ɛi	a	ɑ	u	y	ɔ	œ	o	ø	œy	au	N_{PRES}
i	**92.4**	1.4	2.8	1.4	0.7	0.7	0.7	.	.	.	.	.	.	.	.	144
I	4.9	**86.1**	4.9	3.5	.	0.7	.	.	.	.	.	.	.	.	.	144
e	4.2	2.8	**82.6**	4.2	3.5	2.8	.	.	.	.	.	.	.	.	.	144
ɛ	2.8	5.6	2.1	**82.7**	3.5	.	3.5	.	.	.	.	.	.	.	.	144
ɛi	.	2.8	2.8	3.5	**87.5**	3.5	.	.	.	.	.	.	.	.	.	144
a	.	.	0.7	2.8	2.8	**91.7**	2.1	.	.	.	.	.	.	.	.	144
ɑ	.	0.7	1.4	1.4	0.7	.	**95.8**	.	.	.	.	.	.	.	.	144
u	.	.	.	.	.	.	.	**87.5**	1.7	3.3	1.7	3.3	2.5	.	.	120
y	.	.	.	.	.	.	.	2.5	**89.2**	0.8	.	5.8	1.7	.	.	120
ɔ	.	.	.	.	.	.	.	4.2	3.3	**83.3**	3.3	1.7	4.2	.	.	120
œ	.	.	.	.	.	.	.	1.7	0.8	5.0	**90.8**	.	1.7	.	.	120
o	.	.	.	.	.	.	.	0.8	11.7	1.7	0.8	**80.8**	4.2	.	.	120
ø	.	.	.	.	.	.	.	10.0	0.8	0.8	6.7	1.7	**80.0**	.	.	120
œy	.	.	.	.	.	.	.	.	.	.	.	.	.	**62.5**	37.5	24
au	.	.	.	.	.	.	.	.	.	.	.	.	.	37.5	**62.5**	24
N_{RESP}	150	143	140	143	142	143	147	128	129	114	124	112	113	24	24	1776

Table 5.9 Vowel confusion matrix of within-viseme discrimination in the VIS speech condition. The overall mean score (11 subjects) is 79.2%.

stim \ resp	i	I	e	ɛ	ɛi	a	ɑ	u	y	ɔ	œ	o	ø	œy	au	N_{PRES}
i	**85.6**	2.3	1.5	5.3	0.8	0.8	3.8	.	.	.	.	.	.	.	.	132
I	6.8	**80.3**	5.3	0.8	1.5	1.5	3.8	.	.	.	.	.	.	.	.	132
e	4.6	2.3	**76.5**	3.0	6.8	5.3	1.5	.	.	.	.	.	.	.	.	132
ɛ	3.8	6.8	6.8	**75.8**	3.8	0.8	2.3	.	.	.	.	.	.	.	.	132
ɛi	0.8	3.0	9.1	3.8	**80.3**	2.3	0.8	.	.	.	.	.	.	.	.	132
a	2.3	0.8	3.8	9.8	3.8	**75.0**	4.6	.	.	.	.	.	.	.	.	132
ɑ	.	1.5	3.1	3.8	0.8	.	**90.9**	.	.	.	.	.	.	.	.	132
u	.	.	.	.	.	.	.	**80.0**	0.9	5.5	1.8	7.3	4.5	.	.	110
y	.	.	.	.	.	.	.	7.3	**77.3**	4.5	1.8	5.4	3.6	.	.	110
ɔ	.	.	.	.	.	.	.	7.3	4.5	**70.9**	7.3	2.7	7.3	.	.	110
œ	.	.	.	.	.	.	.	2.7	3.6	8.2	**78.2**	2.7	4.5	.	.	110
o	.	.	.	.	.	.	.	2.7	5.4	1.8	1.8	**80.0**	8.2	.	.	110
ø	.	.	.	.	.	.	.	8.2	2.7	.	8.2	0.9	**80.0**	.	.	110
œy	.	.	.	.	.	.	.	.	.	.	.	.	.	**81.8**	18.2	22
au	.	.	.	.	.	.	.	.	.	.	.	.	.	40.9	**59.1**	22
N_{RESP}	137	128	140	135	129	113	142	119	104	100	109	109	119	27	17	1628

5.5.3 Within-viseme vowel confusions and perceptual dimensions

Tables 5.6 up to and including 5.9 provide vowel confusion matrices for the speech conditions AMPL, F1, F1F2, and VIS. The confusions are based on stimulus/response pairs of 12 subjects (11 in VIS). Although the INDSCAL analyses presented below are based on confusion matrices from individual subjects, the overall matrices give a general idea of the perceptual errors that were made.

In general, there are relatively few confusions, and all scores within the unrounded and rounded visemes are well beyond the CL99 score. Inspection of tables 5.6 to 5.9 indicates that individual vowel scores are higher in the auditory-visual conditions than in the visual, except in the case of the diphthongs. Apparently, acoustic support does more harm than good to correct diphthong discrimination. The positive bias, in the purely visual condition, towards correct discrimination of diphthong /œy/ is remarkable; in fact we have noticed the opposite response behaviour in our identification data of the viseme study in chapter 2.

The INDSCAL solutions are plotted in the two-dimensional stimulus and subject spaces of figures 5.8 (unrounded vowels) and 5.9 (rounded vowels). The high proportions of explained variance indicate that these dimensions suffice for an adequate representation of the perceptual confusions. Moreover, it is clear that the individual solutions of almost all subjects agree very well with the overall solution, considering the fact that most of them have nearly all of their variance explained by the two dimensions (positions close to the quarter circle). Obviously, this does not hold for S13 in the AMPL condition, for S03 in the F1 condition, and to a lesser degree for a few subjects in the VIS condition. In section 5.6.3 the subject spaces will be discussed in more detail.

Correlations of the vowel positions along the two dimensions and physical parameters show a clear picture of the information that is actually used in order to make decisions in the discrimination task. Table 5.10 presents accounted variances, correlation coefficients and the parameters that make for significant correlations. In the auditory-visual condition, dimension I clearly represents formant F_1, while dimension II represents vowel duration. In the VIS condition F_1 is also an important dimension; obviously, it is not an acoustic cue but it rather reflects vertical degree of lip opening as an important lipreading cue (besides lip rounding; Jackson *et al.*, 1976; Caplan Hack and Erber, 1982; Busby *et al.*, 1984; van Son *et al.*, 1993c).

Figures 5.8 and 5.9 (pp. 132-133). Two-dimensional INDSCAL solutions of auditory-visual and visual within-viseme vowel confusions, based on the results of 12 profoundly hearing impaired subjects (11 in VIS). **Figure 5.8:** From top to bottom stimulus and subject spaces are shown for unrounded vowels in four speech conditions: AMPL (83.2% explained variance), F1 (81.5%), F1F2 (89.4%) and VIS (77.4%). **Figure 5.9:** From top to bottom stimulus and subject spaces are shown for rounded vowels in four speech conditions: AMPL (93.6% explained variance), F1 (91.5%), F1F2 (90.1%) and VIS (88.0%). The proportions of explained variance per dimension are given in table 5.10.

Table 5.10 Accounted variances (acv, %) of two perceptual INDSCAL dimensions in within-viseme vowel discriminations, the vowel parameters (param) that have a significant relation to them, and correlation coefficients (Pearson's *r*) plus significance levels expressing this relationship.

speech condition	dimension I acv	param	r	dimension II acv	param	r	total acv
unrounded vowels (figure 5.8)							
AMPL	35.1	F1	.78*	48.1	duration	.88**	83.2
					F1	-.66*	
F1	54.8	F1	.83*	26.7	duration	.95**	81.5
F1F2	37.1	F1	.85**	52.3	duration	.83**	89.4
VIS	40.1	F1	.78*	37.3	duration	.90**	77.4
rounded vowels (figure 5.9)							
AMPL	44.4	F1	.71*	49.2	duration	.86**	93.6
F1	45.4	F1	.95**	46.1	duration	.94**	91.5
F1F2	26.3	F1	.80*	63.8	duration	.95**	90.1
VIS	62.7	duration	.84**	25.3	F2	.81*	88.0
		F1	.72*				

Notes:
1. Significance levels: * $p < .05$, ** $p < .01$
2. In the unrounded vowels, F_1 and F_2 are strongly related: $r = -.91$, $p<.01$
3. In the rounded vowels, F_1 is strongly correlated with level: $r = .99$, $p<.01$

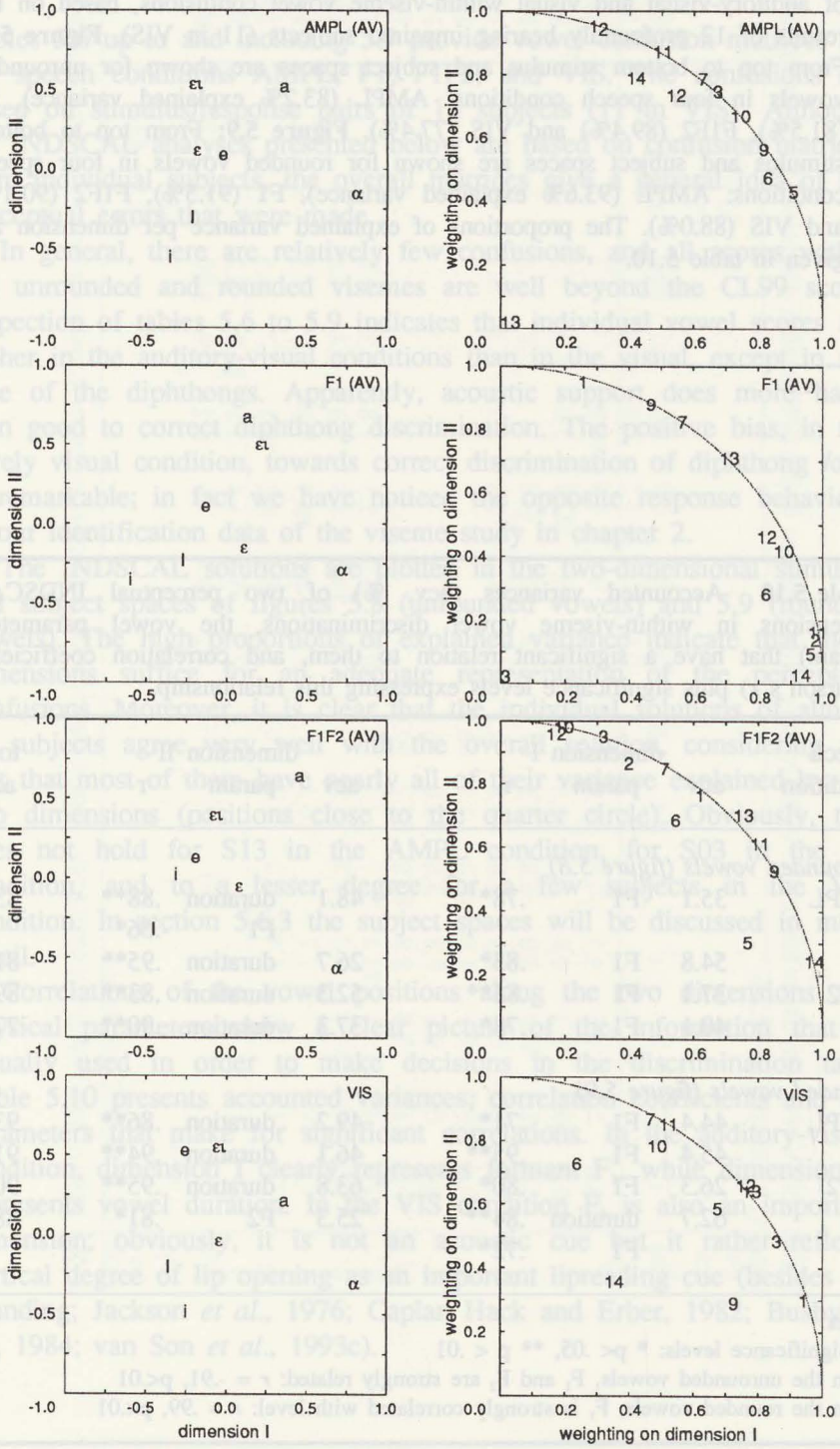

Figure 5.8 Legend see page 131.

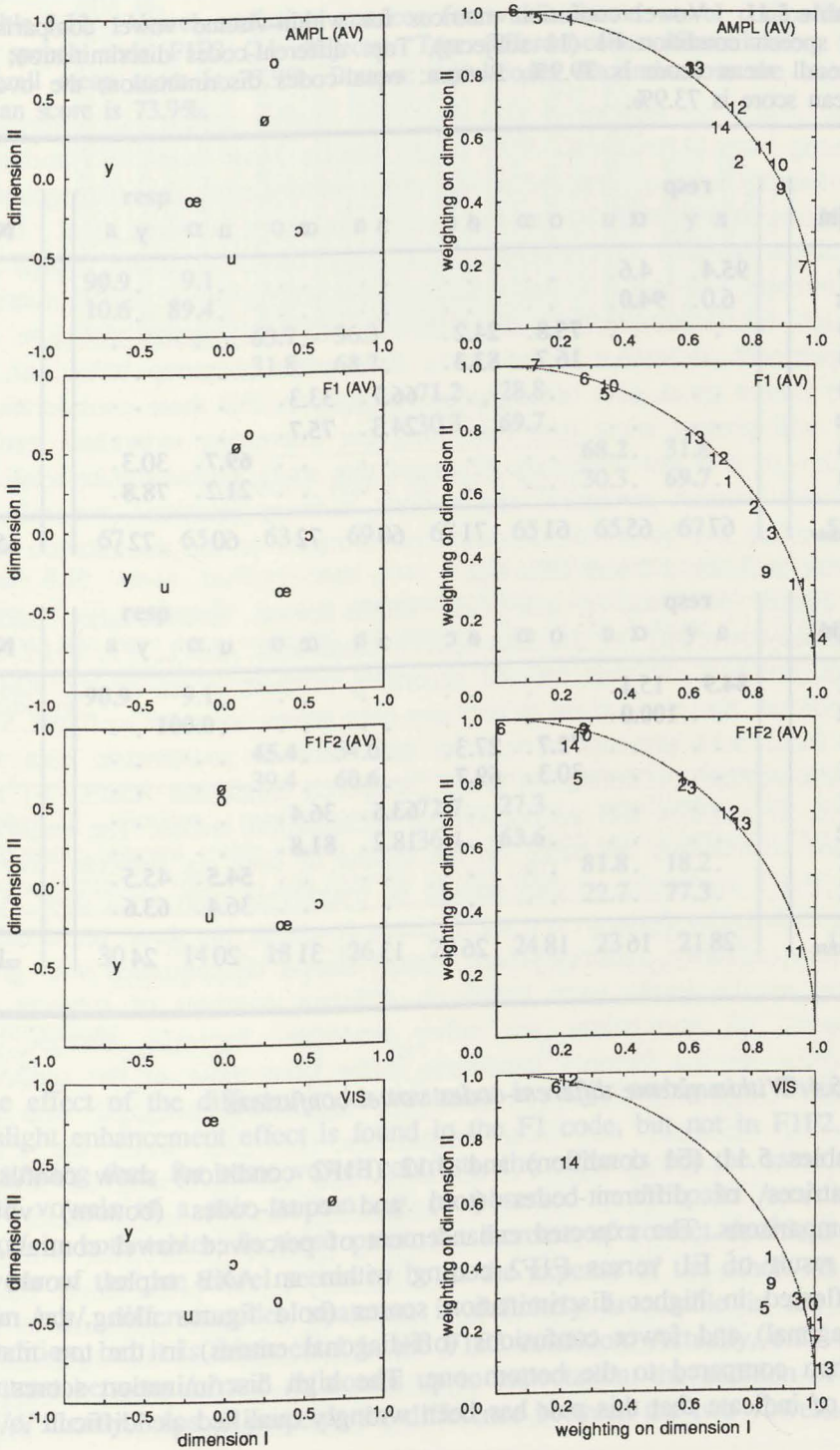

Figure 5.9 Legend see page 131.

Table 5.11 Vowel confusion matrices for within-viseme vowel comparisons of speech condition F1 (11 subjects). Top: different-codes discrimination; the overall mean score is 79.9%. Bottom: equal-codes discrimination; the overall mean score is 73.9%.

	resp								
stim	a	ɑ	o	ø	ɔ	œ	u	y	N_{PRES}
a	**95.4**	4.6	.	.	.	.	.	.	66
ɑ	6.0	**94.0**	.	.	.	.	.	.	66
o	.	.	**75.8**	24.2	.	.	.	.	66
ø	.	.	16.7	**83.3**	.	.	.	.	66
ɔ	.	.	.	.	**66.7**	33.3	.	.	66
œ	.	.	.	.	24.3	**75.7**	.	.	66
u	.	.	.	.	.	.	**69.7**	30.3	66
y	.	.	.	.	.	.	21.2	**78.8**	66
N_{RESP}	67	65	61	71	60	72	60	72	528

	resp								
stim	a	ɑ	o	ø	ɔ	œ	u	y	N_{PRES}
a	**84.9**	15.1	.	.	.	.	.	.	33
ɑ	.	**100.0**	.	.	.	.	.	.	33
o	.	.	**72.7**	27.3	.	.	.	.	11
ø	.	.	30.3	**69.7**	.	.	.	.	33
ɔ	.	.	.	.	**63.6**	36.4	.	.	11
œ	.	.	.	.	18.2	**81.8**	.	.	33
u	.	.	.	.	.	.	**54.5**	45.5	22
y	.	.	.	.	.	.	36.4	**63.6**	22
N_{RESP}	28	16	18	26	13	31	20	24	176

5.5.4 Within-viseme different-codes vowel confusions

Tables 5.11 (F1 condition) and 5.12 (F1F2 condition) show confusion matrices of different-codes (top) and equal-codes (bottom) vowel comparisons. The expected enhancement of perceived vowel contrast, as a result of F1 versus F1F2 coding within an AXB triplet, would be reflected in higher discrimination scores (bold figures along the main diagonal) and fewer confusions (off-diagonal entries) in the top matrix when compared to the bottom one. The high discrimination scores for /a,ɑ/ indicate that this pair has been wrongly qualified as 'difficult'.

Table 5.12 Vowel confusion matrices for within-viseme vowel comparisons of speech code F1F2 (11 subjects). Top: different-codes discrimination; the overall mean score is 73.9%. Bottom: equal-codes discrimination; the overall mean score is 73.9%.

stim \ resp	a	ɑ	o	ø	ɔ	œ	u	y	N_{PRES}
a	**90.9**	9.1	.	.	.	.	.	.	66
ɑ	10.6	**89.4**	.	.	.	.	.	.	66
o	.	.	**63.7**	36.3	.	.	.	.	66
ø	.	.	31.8	**68.2**	.	.	.	.	66
ɔ	.	.	.	.	**71.2**	28.8	.	.	66
œ	.	.	.	.	30.3	**69.7**	.	.	66
u	.	.	.	.	.	.	**68.2**	31.8	66
y	.	.	.	.	.	.	30.3	**69.7**	66
N_{RESP}	67	65	63	69	67	65	65	67	528

stim \ resp	a	ɑ	o	ø	ɔ	œ	u	y	N_{PRES}
a	**90.9**	9.1	.	.	.	.	.	.	33
ɑ	.	**100.0**	.	.	.	.	.	.	33
o	.	.	**45.4**	54.6	.	.	.	.	11
ø	.	.	39.4	**60.6**	.	.	.	.	33
ɔ	.	.	.	.	**72.7**	27.3	.	.	11
œ	.	.	.	.	36.4	**63.6**	.	.	33
u	.	.	.	.	.	.	**81.8**	18.2	22
y	.	.	.	.	.	.	22.7	**77.3**	22
N_{RESP}	30	14	18	26	20	24	23	21	176

The effect of the different-codes treatment is rather ambiguous. Overall, a slight enhancement effect is found in the F1 code, but not in F1F2. It is striking that, for some vowel contrasts, the effect is not the same for both vowels of a pair, as can *e.g.* be observed in the pairs /a,ɑ/ and /ɔ,œ/ in both tables. In these pairs, the increase of correct decisions in favour of the one vowel seems to be at the expense of the other. As to pair /u,y/, different-codes treatment is definitely favourable in the F1 condition, but it is detrimental in the F1F2 condition. Actually, only the contrast between /o/ and /ø/ comes up to expectation: the addition of F_2 to /ø/ clearly helps to perceive the difference between the two vowels.

5.6 Results: individual performance and audiometric data

5.6.1 Within-viseme discrimination

The results in sections 5.3 to 5.5 have shown substantial differences among the 12 subjects. Even with similar basic auditory functions, profoundly hearing-impaired people differ widely as to their capacities of handling speech signals, not to mention lipreading (Lamoré *et al.*, 1985; Faulkner *et al.*, 1990b, 1992). The only condition under which no difference was found is the auditory-visual across-viseme discrimination, which is almost perfect for all subjects. It would therefore be impertinent to restrict oneself to a description of group behaviour, as one cannot speak of a homogeneous population. For these considerations we will present some data that bring out interesting individual results. Although it is not within the scope of this study to investigate relations between speech perception and various auditory functions, we will try and relate the observed individual performance to the audiometric data given in figure 5.2 and table 5.2.

Figure 5.10 shows individual within-viseme discrimination scores (chance-corrected) for the four speech conditions AMPL, F1, F1F2 and VIS. Subjects are rank ordered according to their scores on the AMPL condition. For comparison to auditory-only scores we refer to figure 5.3.

Table 5.13 presents correlation data between audiometric data and within-viseme discrimination scores in speech conditions AMPL, F1 and F1F2 in auditory and auditory-visual presentation modes. The auditory-alone correlations are based on the results of only six subjects (five in the F1F2 condition), so they should be interpreted with great care. Also the differences between the significance levels in the three presentation modes (auditory, auditory-visual, visual) should be compared with great care, as the levels were based on different numbers of subjects. In general, all correlations are rather moderate. Typically enough, the patterns for the auditory conditions differ from those of the auditory-visual conditions. Auditory scores are negatively correlated with the hearing threshold for lower frequencies (LFPTA), while the F1F2 condition, involving higher frequencies, also correlates with the regular PTA. So, with more favourable thresholds, discrimination scores are higher. Moreover, F1F2 discrimination is positively correlated with the dynamic range at 2 kHz. In the auditory-visual presentation mode, discrimination performance is not related to threshold measures at all, except for the moderate correlation between AMPL and the threshold at 1 kHz. On the other hand, the dynamic range at 2 kHz seems to explain some of the variance in auditory-visual discrimination scores: the larger

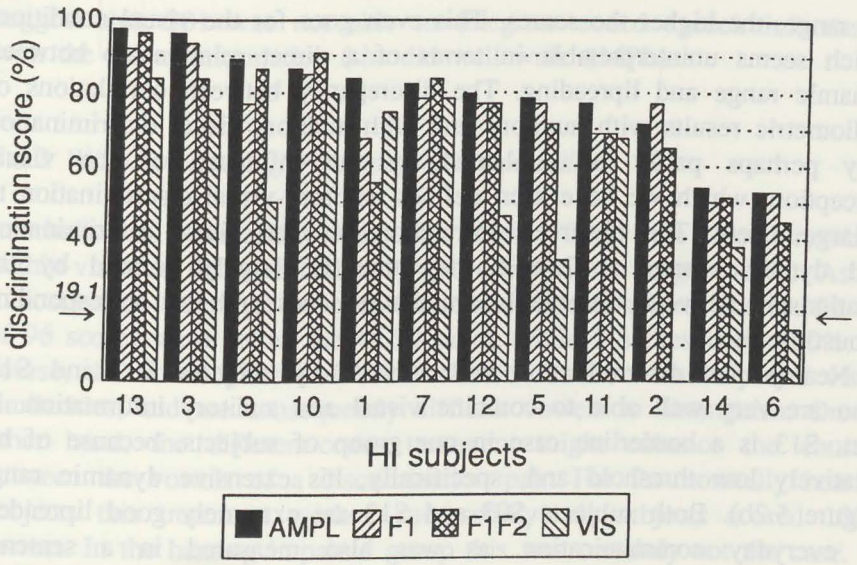

Figure 5.10 Mean auditory-visual and visual within-viseme discrimination scores of 12 profoundly hearing impaired subjects. Scores are averaged over contexts and have been corrected for chance. The arrows indicate CL99 points for chance-corrected scores, based on the normal approximation to the binomial distribution.

Table 5.13 Pearson product-moment correlation coefficients for the relationship between within-viseme vowel discrimination and audiometric data of hearing thresholds (dB HL) and dynamic ranges (dB).

	Hearing threshold				Dynamic range			
	1 kHz	2 kHz	PTA	LFPTA	1 kHz	2 kHz	PTA	LFPTA
auditory-alone								
AMPL	-.27	-.08	-.34	-.68*	-.61	.02	-.14	-.08
F1	-.13	.00	-.27	-.77*	-.58	-.13	-.22	-.01
F1F2	-.83*	-.50	-.78*	-.76*	-.27	.77*	.27	-.22
auditory-visual								
AMPL	-.55*	-.19	-.33	-.20	.50*	.66**	.58*	.27
F1	-.38	-.09	-.33	.01	.31	.48*	.34	.05
F1F2	-.44	-.25	-.25	-.09	.30	.55*	.37	.07
VIS	-.26	.06	-.12	-.27	.39	.59*	.48	.27

1. PTA: pure-tone average (.5, 1, 2 kHz); LFPTA: low-frequency PTA (.25, .5, 1 kHz)
2. Significance levels: * $p<.05$, ** $p<.01$

the range, the higher the scores. This even goes for the visual condition, which seems uninterpretable in terms of a direct relationship between dynamic range and lipreading. The discrepancy between correlations of audiometric results with auditory and with auditory-visual discrimination may perhaps partly be explained by the influence of the visual perception, which we have seen defines auditory-visual discrimination to a large extent. The observed correlation between visual discrimination and dynamic range at 2 kHz may thus be largely defined by the relationship between observable articulation cues and the corresponding acoustic effect.

Near-perfect discrimination was attained by subjects S03 and S13, who are very well able to combine visual and auditory information. In fact, S13 is a borderline case in our group of subjects, because of his relatively low threshold and, specifically, his extensive dynamic range (figure 5.2b). Both subjects S03 and S13 are extremely good lipreaders in everyday communication, as was also measured in a sentence lipreading study running parallel to this vowel discrimination experiment (Bosman and Smoorenburg, 1993a, 1993b). Quite opposite to that is the position of subject S05, who shows better discrimination in auditory-alone conditions than in auditory-visual conditions, for all speech conditions (*cf.* figure 5.3). In the auditory-visual session, he had difficulty keeping concentrated on the video screen; in the auditory session, however, he was very well concentrated; apparently, the video presentation actually distracted him from carrying out the task.

In two cases in figure 5.10, a subject has a slightly higher score on one of the coded-speech conditions (S02 and S10). From the data in table 5.2 it can been seen that these two subjects have by far the two highest thresholds (PTA and especially LFPTA), and consequently narrow dynamic ranges. Four more subjects (S05, S06, S07, S13) show little difference between the AMPL condition and at least one of the speech codes: less than 1.5% difference. Of these, S05, S06 and S07 also have relatively narrow dynamic ranges.

Subjects S05, S09, S12 and especially S06 have rather low visual scores and benefit well by the additional acoustic signals, whatever the information offered. S02 did not take the visual-only condition. Finally, it is interesting to note that subject S01 has higher scores when only lipreading than when simultaneously listening to a coded speech signal. Actually, she is one of the persons with better auditory capacities, as reflected in the threshold curve (figure 2a) and other psychoacoustic data which are not presented in this thesis (Bosman and Smoorenburg, 1993b). The coded speech signals probably offer too little, and even distracting information for her to improve lipreading scores. This may

suggest that S01 was not able to integrate coded speech information with the visual information (she did well with AMPL).

5.6.2 *Within-viseme different-codes discrimination*

Interindividual variability appears most from the auditory-visual within-visemes different-codes vowel comparisons (figure 5.7). As an indication of the range of chance-corrected mean subject scores, 22 out of 96 scores were below the CL99 point, while 9 scores were 100%. Of course, the vowel pairs had been selected for their auditory and visual similarities, and are consequently difficult vowels to distinguish. On the other hand, the different-codes treatment might enhance the contrast between the vowels of a discrimination pair. For the normally hearing subjects this treatment worked successfully: all subjects scored 100% correct. In the hearing impaired group this was certainly not the case.

The interesting questions, of course, are which subjects benefit from the extra cue in order to increase vowel contrasts, and whether this can be related to audiometric data. Benefit is a slightly different concept in the F1 condition than in F1F2: in the F1 condition benefit is received by adding an extra F_2 cue, in F1F2 by leaving out the F_2 information. The effect of both is acoustic contrast improvement. Benefit may be expressed by a raised different-codes score in comparison to the equal-codes counterparts.

We first computed score differences for F1 and F1F2 by subtracting the equal-codes scores from the different-codes scores (the *benefit score*). The average benefit score (11 subjects) in the F1 condition is 12.1%, in F1F2 it is nought. Standard deviations are considerable: 22% in both cases. Different coding within a vowel pair has a detrimental effect of more than 10% for subjects S03 and S06 in the F1 condition, and subjects S03, S05 and S06 in the F1F2 condition. Only six subjects reacted in the same way to increased contrasts in both speech codes: S03 and S06 had negative benefit scores (so, equal-code discrimination is better) of between 21% and 33%, while S01, S07, S10 and S12 really profited by raised scores of between 8% and 50%.

Next, the between-codes scores and the benefit scores were correlated with audiometric data. Few significant correlations were found, and all of them moderate (between 25% and 31% explained variance). F1 and F1F2 between-codes scores appear to be related to the dynamic range at 2 kHz (the F1F2 correlation just touches the .05 significance level). The benefit scores in F1 are not correlated with any audiometric measure, but in those of the F1F2 condition mild but significant negative

correlations were found with the dynamic range at 500 Hz (r=-.53) and the average dynamic range for the low-frequency PTA (r=-.50).

5.6.3 Within-viseme perceptual dimensions

The INDSCAL solutions presented in section 5.5.3 clearly indicated two important physical cues underlying perceptual dimensions of vowel discrimination: F_1 frequency and vowel duration. In the subject spaces of figures 5.8 and 5.9 the agreement of nearly all individuals with the general solution is clearly expressed by the high proportions of variance; whatever the speech condition, most of the subjects are situated on or close to the quarter circle describing 100% variance explained by use of the two dimensions. Only in the visual-only condition some subjects have a deviating own perceptual configuration.

The fact that the group of subjects is placed along the entire quarter circle indicates, again, that there are clear interindividual differences. A different position along the quarter circle implies different weightings of the two dimensions. For example, if a subject attaches all weight to dimension I, his position will be in the right-hand lower corner of the subject space (100% variance explained by dimension I, nought by dimension II). When both dimensions are equally important (50% explained variance each), his position will be halfway the quarter circle.

If a subject behaves consistently over the various speech codes, his or her position on the quarter circle would agree in the corresponding subject spaces. For instance, S01 weighs primarily on dimension II in most of the auditory-visual subject spaces of figures 5.8 and 5.9, implying that for her vowel duration is the most important cue. Different examples are S12 and S13, who in most conditions hold positions near the middle region of the quarter circle, indicating that they attach equal importance to both dimensions. In general, however, most of the subjects seem to change their positions from condition to condition, which may indicate that there are no clear optical-acoustic differences in vowel shapes that lead a person to consistent behaviour. Obviously, F_1 and durational cues are used all the time, although preferences for one or the other may change easily. There is one exception to this, shown by the positions of the subjects in the VIS condition of rounded vowels: either the lipreader makes use of differences in vowel duration when discriminating /o,ø/ from the other rounded vowels, or he is keen on the subtle lip opening differences (related to F_1). There seems to be no middle way, as is the case in all other conditions of figures 5.8 and 5.9

5.7 Discussion

Compared to unprocessed speech, the two speech codes F1 and F1F2 offer less potentials to the profoundly hearing impaired. The information in the AMPL speech condition yields significantly higher within-viseme discrimination scores. Apparently, the subjects are able to process more information than the limited amount contained in the F1 and F1F2 signals; within the group of subjects there are only few persons who profit more from a coded signal than from the unprocessed speech. Again, it is shown that speech signals constructed on the basis of extracted speech features do not provide the results that is often expected from this approach. So far, the optimum of information is still derived from the entire speech waveform, even though this amount is limited. Our attempts to simplify this signal in order to bring out more clearly the relevant features has had little or even adverse effect. At least two factors may play a role in this.

First of all, the synthesized signals that the profoundly hearing impaired had to listen to were unnatural, even though we made use of straightforward processing techniques, not involving drastic frequency mappings of formants or noise bands or encoding schemes such as proposed in chapter 4 (figure 4.2, schemes 2 and 3). Nevertheless, some training will be required in order to learn and interpret the new speech code. The patient who receives a speech-processing hearing aid should be able to integrate visual and auditory information and has to build up new phonological patterns of speech sounds. As discrimination is purely based on the perception of physical differences in speech sounds, it seems very unlikely that the lack of training with our speech codes has had much effect on auditory(-visual) vowel discrimination performance.

A second and more important factor, but also less transparent, is the degree of auditory functioning on the part of the listener. We have seen that some of the subjects benefit more from the F1 or F1F2 speech code than from the unprocessed signal. In general, performance seems to be related, in some sense, to the audiometric threshold at the lower frequencies (LFPTA: .25, .5, 1 kHz) and to the dynamic range at 1 and especially 2 kHz, the region of the higher second formants. Of course, the threshold is compensated for by adequate amplification of the speech signals, so other auditory functions, related to the threshold, are likely to play an important role here (*e.g.* difference limens for frequencies and the amount of upward spread of masking). In our opinion, one of the most important questions to be addressed in the near future is why some subjects profit from processed speech.

Vowel discrimination in all four speech conditions (amplified speech,

F1 code, F1F2 code and unaided lipreading) may be described by two independent perceptual dimensions, which explain equal parts of the variance and are sufficient for an adequate description (the total proportion of explained variance is between 80% and 90%). The first dimension is related to first formant information, the second to vowel duration. In purely auditory discrimination, the same cues seem to play a role, but they are both related to one perceptual dimension. An added second dimension does not make for clearer interpretation of the psychophysical relationship.

It seems that vowel discrimination in auditory-visual conditions depends mainly on purely visual cues. As most phoneme lipreading studies have shown, the vertical degree of lip opening in lipreading is clearly related to the first formant frequency: open vowels have relatively high first formants. Moreover, vowel duration has been shown to be a useful cue in vowel lipreading: in Dutch, this cue leads to the formation of the viseme /o,ø/ (van Son *et al.*, 1993c).

Yet, auditory-visual discrimination is not completely explained by lipreading performance. This is shown in two ways. First, the same cues of F_1 and vowel duration are related to perception in auditory conditions. Second, the statistically significant difference in discrimination scores between all auditory-visual conditions and the visual condition shows that acoustic information is really presented to the benefit of the lipreader. In summary, presentation of an acoustic signal (processed or not) increases visual vowel discrimination, but it does not add new information: the same underlying dimensions play a role in auditory and visual vowel perception by the profoundly hearing impaired. In this way, the auditory-visual speech signal contains redundant information (see also chapter 4).

The effect of increasing vowel contrasts by using different codings for vowels in a comparison pair, the different-codes condition, was not unambiguously successful. First, the presentation of identical pairs in the F1 and F1F2 conditions did not give rise to equal scores. In general, benefit from extra contrast was greater in the F1 than in the F1F2 condition. Secondly, the correlations between audiometric data and discrimination, if present, are rather weak. It is therefore not quite clear why some subjects receive benefit by this coding effect and others do not.

In retrospect, the different effects in F1 and F1F2 conditions may not be too surprising. In the F1 condition, adding a second formant to one of the vowels indeed enhances the contrast if the extra information is perceived. This may not be the case when excessive upward spread of masking by the first formant occurs; this we have seen in the pattern

perception study of chapter 3, where a substantial separation between low and high frequencies was needed for perceived differences. So, if the spectral distance between the two formants is not large enough, the different-codes treatment may have no effect at all. It is not clear why a detrimental effect should occur. In the F1F2 condition, leaving out a second formant will only help, also, when there is no excessive upward spread of masking - so, if it was actually perceived at all. Most of the evidence is that discrimination is mainly related to F_1. So even when the contrast is actually enhanced, *i.e.* acoustically, the perceptual effect may be negligible (on an average this is the case). But a detrimental effect may also occur. If an F_2 is removed, it is always the one of the vowel with the lowest second formant: /u,ɔ,o,ɑ/. This is a relatively low frequency and may thus be more strongly present, perception-wise, than F_2 of the other vowel: /y,œ,ø,a/. This stronger presence of a relatively low-frequency F_2 might be a cue to base discrimination on, and when it is removed, perceptual contrast is actually reduced. If this explanation is a valid one, a different-codes strategy might be used when applying an F1 speech coding scheme, because it may or may not help, but is unlikely to have a negative impact. In an F1F2 coding scheme it is not advisable to use it, because it may have a countereffect from the one intended.

We found it a disquieting observation, at first sight, that visual within-viseme vowel discrimination scores, of both hearing impaired and normally hearing subjects, are well beyond chance level. With regard to the working definition of a viseme, a set of phonemes that are hardly or not at all discriminable, we had expected within-viseme discrimination to be around chance level. But clearly, discrimination is quite different from identification (most viseme structures are based on identification tasks; see Owens and Blazek, 1985) in that relatively subtle cues can be used in order to make similarity decisions, although they provide too little information for correct phoneme labelling. However, the fact that better discrimination is possible than suggested by identification tasks implies that it may be possible to train lipreaders to observe these subtle differences among vowels in a consistent way. Whether such a training automatically leads to better identification (where no direct comparisons are possible) cannot be substantiated as yet.

Finally, one comment should be made on the discrimination task itself. With such high scores as we found in this study, our interpretation of auditory-visual vowel discrimination performance may suffer from ceiling effects, obscuring differences among the acoustic supplements. However, this risk is disproved by statistical analysis showing even now a significant advantage of unprocessed over

processed speech. Nevertheless, also in light of the reasonably high average auditory-only scores, we must regard this vowel discrimination task as a questionably easy one, prone to the negative effects of ceiling scores.

5.8 General conclusions

1. Profoundly hearing-impaired lipreaders profit significantly from the addition of acoustic cues in order to discriminate among vowels from the same viseme.
2. As an acoustic supplement, formant-based synthesized speech signals are mostly less beneficial than unprocessed amplified speech that is well-matched to the subject's dynamic range.
3. Visual and auditory-visual vowel discrimination are related to the same physical cues: vowel formant F_1 and vowel duration. Both cues may be accessible through the auditory as as well as the visual sense (F_1 is correlated with the visible vertical degree of lip opening).
4. Purely visual discrimination of vowels from the same viseme is done well beyond chance level for rounded and unrounded vowels. In this light, the definition of the viseme as a set of mutually indiscriminable phonemes is incorrect; viseme classifications based on identification studies do not reflect the more subtle vowel distinctions very well.
5. The high within-viseme discrimination scores imply that training may be feasible for learning to observe subtle distinctions among vowels in a consistent way.
6. Performance in auditory vowel discrimination by the profoundly hearing impaired is moderately related to the hearing threshold for low frequencies (LFPTA); in auditory-visual discrimination a correlation is found with the dynamic range at 2 kHz.

Acknowledgement

We are grateful to Margreet Langereis, who analysed part of the material in order to provide an independent check on the correctness of the semi-automatic formant analysis procedure.

Chapter 6.

Miscellaneous topics for future research

6.1 Processed speech and auditory functions

6.1.1 Coding schemes for individuals

From the summary in chapter 4 (table 4.4) and from our own results in chapter 5 one may conclude that the feature extraction approach of providing only selected speech information has so far failed to yield better results than is possible with presentation of unprocessed speech conveyed by well-fitted hearing aids. Even those processing schemes applying spectral speech features in addition to F_0 and amplitude cues do not lead to better performance than that obtainable by the use of amplified speech.

Nevertheless, these findings do not imply, just like that, that the approach of feature extraction and acoustic recoded-signal transmission has definitely failed. The studies described in section 4.3.3 and in chapter 5 have demonstrated that subjects with comparable hearing losses (with respect to various auditory functions) may react quite differently to various speech codes. The group of profoundly hearing impaired persons appears to be a very heterogeneous one, if various abilities of visual and auditory speech perception are taken into account. For clinical settings this implies that statistical descriptions of general group behaviour is of very limited use when evaluating the efficacy of certain speech patterns. It thus seems to suggest that speech codes should be defined on an individual basis rather than trying to make one strategy do for a whole group of profoundly hearing impaired persons.

One of the important implications of this need for individual fitting is that one is required to have some sense of the factors that predetermine the choice of certain speech features. It is of eminent importance to establish predictive descriptors of the potentials of various strategies, and firm selection criteria as to what speech features might be profitable for which impaired persons should therefore be developed. Ideally, a prescription scheme should indicate what kind of information might be profitable for a patient given his or her auditory as well as lipreading abilities. It would be a desirable situation if lipreaders who are able to handle more than only very basic acoustic information could be

provided with more speech cues. Someone who has no residual frequency selectivity and a very narrow dynamic range may only profit from f_0 information presented at MCL. Another person's dynamic range may allow the extra use of amplitude information. When some degree of frequency selectivity is present, (unidimensional) formant information may be utilized; the choice of which formant may even depend on the individual's lipreading skill.

6.1.2 Predictive tests for patient selection

The call for useful predictors for selection of patients and speech processing schemes may be one of the most important issues in the near future. Faulkner *et al.* (1992) mention a few useful clinical indicators for the appropriateness of a SiVo-like aid, such as extreme audiometric loss, often associated with lack of frequency selectivity in the region of the fundamental frequency, and a restricted dynamic range. Bosman and Smoorenburg (1993b) and van Son *et al.* (1993b) are less conclusive as to significant correlations that exist between various auditory functions, including absolute threshold and usable dynamic range, and the performance on different speech perception tasks for those subjects that profit from processed-speech presentation. In the absence of clear correlations, a useful strategy in patient selection may be to measure intelligibility of sentences presented auditory-visually in unprocessed-speech as well as various processed-speech conditions. Although this is a rather pragmatic solution, the method is straightforward and time-saving when compared to psycho-acoustic measurements, which may place a considerable load on the patient and only give satisfying results after quite some practice.

In some (profoundly) hearing impaired outstanding lipreading skills can be witnessed, so that ceiling effects may occur. Such performance was found by Bosman and Smoorenburg (1993b); four out of their twenty subjects correctly reported up to 80% of the syllables in a set of semantically unrelated everyday sentences (Plomp and Mimpen, 1979), presented in a vision-only condition (Bosman and Smoorenburg, 1993b). In such cases the lipreaders appear to profit well from only basic speech information, *e.g.* voicing and amplitude information. In fact, any added speech sound would do to raise the scores up to a near-perfect level, whether processed or unprocessed speech is used. For these highly skilled lipreaders a signal processing hearing aid might be evaluated as if it were a 'normal' hearing aid rather than a device primarily serving as a lipreading aid. This implies that testing should also be considered

in auditory-alone conditions, or that, at least, test materials of a higher degree of difficulty should be available (*e.g.* the so-called semantically anomalous sentences, having correct syntactic structures but little redundancy; Nakatani and Dukes, 1973).

6.2 Choice of speech features in individual languages

In section 6.1.1 we mentioned the desirability for the establishment of an ordered scheme of feature selection for individuals in relation to gradations of profoundness of hearing impairment or proficiency level at any given speech perception task. Speech feature selection can also be viewed from the point of view of the phonological features of individual languages.

It has been pointed out by Fourcin (1990) that native lipreaders in different languages may profit from the presentation of different features, depending on the phonological system. Lipreading complexity may thus be a factor in selecting the proper speech features as supplementary cues, and hence a rank order of important speech features to be presented to different lipreaders should be established for each language individually. *E.g.*, in tone languages, such as Chinese, different pitch levels are used for lexical contrasts, so the absence of this cue in purely visual speech perception adds to the 'normal' difficulties for the lipreader. Presentation of pitch information, then, will be of great help to the (hearing impaired) lipreader. A different example is the Arabic language, containing especially many fricative contrasts which are also invisible to the lipreader. In this case, good support may be lent by presenting information providing some of the fricative contrasts.

In order to elucidate the influence of phonological systems of native languages on speech feature selection, let us elaborate a bit on the similarities and differences between English and Dutch consonants. Table 6.1 displays the consonantal systems for English and Dutch, assigning each consonant to a class based on the features of visibility (visemes), voicing and waveform amplitude envelope. First of all, we may recognize 24 English versus 19 Dutch consonants. Most of the English consonant lipreading studies discussed in chapter 4 (section 4.2.2) mention relatively large numbers of visemes as compared to the Dutch studies, which is due to the presence of such highly visible consonants as (inter)dental /θ,ð/, and to the fact that a number of consonants have lip rounding as a secondary articulatory cue: /ʃ,ʒ,tʃ,dʒ,w,r/. In the table, the English consonants are classified into seven visemes (Owens and Blazek, 1985; combined visemes in contexts /a,i,ʌ/) and the Dutch

consonants into four visemes (van Son *et al.*, 1993c; chapter 2). It is clear that the higher number of visual contrasts in English sets a better baseline for the lipreader: for English the phonemic information available from lipreading is 29% (24 phonemes assigned to 7 visemes), while in Dutch 21% is available. If the lipreader is provided with voicing information, these percentages are raised to 54% and 42%, respectively. Further extension of the number of contrastive elements may be provided by speech waveform amplitude information, mainly separating non-continuants (plosives and affricates) from continuants (fricatives and sonorants), based on perceived gaps. The combination of lipreading, voicing and amplitude envelope features then leads to a number of 19 distinctive elements for English, or 79% of the total phonemic information passed, and 11 distinctive phonemic elements for Dutch, 58%. Further distinctions, *e.g.* those among nasals, liquids and semivowels, or between Dutch /v/ and /ʋ/, ought to involve presentation of spectral cues. Because of the substantial difference in transmitted phonemic information between English and Dutch, 79% versus 58%, the use of extra spectral cues may be argued to be more urgent for the Dutch than for the English lipreader.

Of course, this simple model does not take into account the relative ease with which each cue may be perceived. The perception of voicing contrasts is probably more easy than perception of amplitude contrasts (see *e.g.* the relatively low identification scores in the consonant study by Van Tasell *et al.*, 1987). Spectral cues may pose even more difficulties for many profoundly hearing impaired, especially if their residual capacities for frequency selectivity are severely limited.

Linguistically relevant factors such as frequency of use of individual consonants or actual occurrence of minimally contrastive word pairs may also play a role in the decision for or against the use of a certain speech cue. For example, the need for further distinctions within the Dutch viseme /s,z,ʃ/ does not seem really urgent. On the one hand, palatal /ʃ/ does not frequently occur, and there are only few word pairs differentiated by the use of /s/ or /ʃ/. Also, few contrastive word pairs are based on the /s,z/ distinction. On the other hand, even when minimal word pairs exist, they will not be easily confused because of strong contextual constraints. *E.g.*, in a number of Dutch dialects speakers do not make a voicing distinction in their production of such minimal pairs as /s,z/ and /f,v/, but this seldom leads to misunderstandings (although the author must admit that the mention of his surname almost invariably evokes the question "Is that an *s* or a *z* ?").

Table 6.1 Consonantal systems for English and Dutch. The assignment of consonants to classes is based on visemes, voicing and amplitude information.

viseme	voicing	English continuance -	English continuance +	Dutch continuance -	Dutch continuance +
[1] bilabials	-	p		p	
	+	b	m	b	m
[2] labiodentals	-		f		f
	+		v		v,ʋ
[3] (inter)dentals	-		θ		
	+		ð		
[4] rounded approximants	-				
	+		w,r		
[5] front fricatives and affricates[1]	-	tʃ	ʃ		s,ʃ
	+	dʒ	ʒ		z
[6] nonlabials[2]	-	t	s		
	+	d	z		
[7] nonlabials	-	k		t,k	x
	+	g	n,l,j,ŋ,h	d	n,l,j,R,ŋ,h

[1] In English, viseme [5] is recognized on the basis of lip rounding (and slight lip protrusion), in Dutch on the basis of restricted lip opening.

[2] In most English consonant studies /t,d,s,z/ are classified as a separate viseme. However, the difference between visemes [6] and [7] is not always very clear.

The consonant system in table 6.1 clearly demonstrates that further distinction among the consonants would especially be needed for the voiced continuants in the nonlabial viseme group [7]. Until now we have been mainly concerned with the question as to how Dutch vowel information might best be presented acoustically to the lipreader. Apart from the lipreading experiment in chapter 2, consonants have not been the topic of the study described in this thesis. The following strategy may be adopted for consonant representation. Voiced consonants can be treated in the same way as vowels, *i.e.* they may be represented by

formant and amplitude information. This implies that some crude perceptual differences may occur within the voiced continuants /n,l,j,r,ŋ,h/ in nonlabial viseme group [7]. Voiceless consonants may simply be represented by white noise or, if more differentiation is feasible, by different noise bands. Another possibility of consonant treatment may be the accentuation of certain temporal cues such as duration of the gap of voiceless plosives.

As to fricative noise representation, table 6.1 shows that if the visemes are readily recognized, only fricatives /s/ and /ʃ/ need be distinguished. This, however, is not an urgent case for reasons explained above. On the other hand, Dutch lipreaders who do not make a clear difference between visemes [5] and [7] may profit from the use of different fricative noises. In that case /s,ʃ/ may be represented by relatively high-frequency noise (1200-2000 Hz) so as to distinguish them from /x/ with low-frequency (50-1000 Hz) noise; a similar strategy may be utilized for the /t/-/k/ distinction (different burst noises).

In Dutch, the difference between continuants and non-continuants is cued by the temporal structure of the plosive sounds (and the odd affricate, occurring in a number of foreign loan words, like "chip" and "gel"). A plosive may roughly be characterized as a sequence of a silent or nearly silent period and a noise burst. In running speech, the plosive may be indicated by an interruption of the amplitude envelope. In order to emphasize the difference between plosives and fricatives, the plosive gap might be artificially lengthened. In view of the temporal resolution of the average profoundly hearing impaired this may be a useful procedure: the average minimal length of the silent gap in speech is found in the voiced plosives /b,d/ and amounts to 39 msec (sd=12 msec; Kuijpers, 1993), while the average gap width that profoundly hearing impaired listeners may detect in a 200 msec period of noise is 30 msec, with a standard deviation of about 17 msec (Bosman and Smoorenburg, 1993b). Moreover, gap lengthening in voiceless plosives may be a strategy to accentuate the difference between voiced and voiceless plosives, which partly depends on the duration of the oral tract closure during plosive consonant production (Lisker, 1957; Kuijpers, 1993).

6.3 Technology and training

6.3.1 Coded speech and structured training

One of the aspects that is crucial to the relative success of speech processing hearing aids is the amount of training required for inter-

pretation of the new speech code. A person who switches from conventional hearing aids, conveying the entire speech signal, to a feature extraction aid like *e.g.* SiVo[1], cannot be expected to be able to interpret the new speech signal as if it were a slightly different normal speech signal. It is not only a matter of accommodating to the different sound, but it actually involves learning to cope with a new speech system.

The 'naturalness' of the coded signal may have consequences for the ease of learning to interpret the new speech code. 'Naturalness' in this sense simply indicates the degree of resemblance to the formerly perceived speech signal (through a conventional hearing aid). In principle, a new speech code should be designed such that maximal discriminability among speech sounds is achieved. The only condition is that the coded information matches the limited auditory capacity. If such a new, maximally efficient, speech system cannot be incorporated by the profoundly hearing impaired lipreader, however, it may be more profitable to define a less unnatural speech code that is less efficient as to discriminative potential but which approximates unprocessed speech more closely and is thus more easily learned by the lipreader. In chapter 4, figure 4.2, we have presented three coding schemes for unidimensional vowel representation, the third of which had great potential for vowel differentiation, but was highly unnatural. If this speech code is very problematic for listeners to learn and interpret, application of the less potential but much more natural speech code [1], based on F_1 in the correct, low frequency region, may be the better strategy to follow.

Discussing rehabilitational aspects of speech processing hearing aid use lies outside the scope of this article. Nevertheless, in view of the results of the studies reported in section 4.3.3, where speech processing techniques were compared with unprocessed-speech presentation, one should be aware of the fact that the reported results may not be the maximally attainable scores. Experience in this line of research has been rather limited so far, which makes it difficult to assess whether a few training sessions applied within the context of the experiment make for maximum performance on the part of the profoundly hearing impaired.

---

[1] The SiVo (*Sinusoidal Voice*) is a speech processing device which extracts f_0 information from the speech waveform and represents it as a sine wave. In the first version of the SiVo the sine wave was presented at a constant level (MCL); in a later version the amplitude was modified according to the amplitude envelope of the full speech waveform (Rosen *et al.*, 1987; Faulkner *et al.*, 1992).

6.3.2 Natural speech and structured training

It may be argued that unprocessed-speech perception by the profoundly hearing impaired can also profit from structured auditory training. In rehabilitation following conventional hearing aid fitting this aspect is generally neglected. When a hearing impaired person receives a new hearing aid, it is common practice that he needs some time to get used to the sound quality of the new apparatus. It is unusual, however, for a hearing aid wearer to attend auditory training classes (or individual sessions) where he can practice, in a highly structured way, hearing abilities at various levels (especially discrimination and identification). Auditory training is commonly given to profoundly hearing impaired or deaf children attending special schools and to persons who have received a cochlear implant or SiVo type of aid and are following a rehabilitational training course.

It is an interesting question in how far profoundly hearing impaired persons using a conventional type of hearing aid would profit from an auditory training structured in the way that the current post-implantation trainings are. It is not only a theoretically interesting question (does implantation itself or post-operative training have the best effect in effective cochlear implant use ?), but a positive effect from auditory training on conventional hearing aid use might make conventional hearing aid fitting plus auditory training a good alternative to signal processing hearing aid fitting plus auditory training, or may even be a strong argument for auditory training for the profoundly hearing impaired, whether a signal processing aid is considered or not.

However, as hardly any systematic auditory training is given to conventional hearing aid users, it is impossible to give a direct answer to the question whether it would be helpful. In an indirect way some evidence is available. At the department of Otorhinolaryngology of the Academic Hospital Utrecht, a 22-channel cochlear prosthesis has recently been implanted in two patients who previously wore a hearing aid in the other ear. In these cases, the use of the hearing aid before implantation was rather limited: auditory phoneme recognition scores were less than 15% (which is one of the criteria for implantation). Prior to implantation pre-operative tests were taken, in which among other things auditory CVC word perception with the hearing aid was measured. After implantation the patients attended a training programme in which much attention was paid to discrimination and identification of speech sounds. The training was carried out with both hearing aid and implant, and included regular sessions at the hospital, guided by a speech therapist,

and regular practice at home, with the help of a partner. After six months the effects of this training on hearing aid use and implant use were evaluated. The CVC word perception results of both patients showed substantial increases: from 20% to 45%, and from 8% to 20%. There was little difference between the scores obtained with the hearing aid only and scores with the hearing aid and implant.

The effect of training and that of the new hearing sensations introduced by the use of the cochlear implant cannot be separated; both may have contributed to the score increases. Nevertheless, it seems that training may have a favourable effect on speech perception through a conventional hearing aid if a well-structured training programme is carried out. On the evidence from these two patient results, we may therefore carefully conclude that it is worth considering giving systematic auditory training to profoundly hearing impaired persons using conventional hearing aids. For some of them it may be a good alternative to using a signal processing hearing aid, and as long as no efficient signal processing aids are available, auditory training may be regarded as a good form of rehabilitation in its own right. In other words, structured auditory training should not be automatically considered to be feasible only in combination with technologically advanced hearing prostheses, but ought to be available to (profoundly) hearing impaired persons as a normal facility. In this light it is interesting to mention that the Audiological Centre of the Amsterdam Medical Centre started an experimental auditory training some five years ago, using exercises and evaluative tests of their own. Recently, they have included parts of the cochlear implant training programme used at the Academic Hospital Utrecht. Their results appear to be quite good (personal communication), although the training effect has not been evaluated in a systematic way, nor have any results been reported yet.

6.3.3 The signal processing hearing aid as a training device

The observation that both training and the multichannel implant system may have contributed to the post-operative results with only the hearing aid can be interpreted in a way that may have consequences for the practical use of signal processing aids. The implant wearer receives selected speech information (f_0 and formant information) and learns to interpret this as significant linguistic information. The various auditory exercises with constant feedback will make him aware of speech cues that he was formerly not able to extract from the complex speech waveform. The implant system may thus be regarded as a perfect tool

for training perception of selected speech features. An implant system basically functions exactly in the way that a signal processing hearing aid is intended to do: the speech processor reduces a highly complex signal to an efficient speech code that contains nothing but useful information. This implies that a well-designed signal processing hearing aid, even if it does not replace the conventional aid in daily use, may be very useful as a training device.

In fact, Fourcin (1990) mentions this practical use of the SiVo device in relation to the acquisition of speech elements to young profoundly hearing impaired children. He claims that a signal processing aid for the profoundly hearing impaired child might be one of the most important future applications. Fourcin asserts that, at a certain age, only those features should be presented which are normally acquired in that stage of development. So, in the first few postnatal stages only intensity, periodicity and temporal organization of voice factors are being acquired, and it would thus be senseless to present more than this information to the developing child. Only in later stages it will be necessary to consider fundamental frequency and, still later, vowel and consonant aspects. The slow adaptation to ever new sounds may be feasible for the profoundly hearing impaired child, and this may be easily achieved with the help of speech feature presentation. The main value of such a process of gradual familiarization may be the fact that speech features are presented in isolation so that recognition may be triggered more effectively than when the same features should be extracted from the entire speech waveform containing multiple other information cues that have no meaning yet for the child. Even if the child will not go on using the signal processing hearing aid as a daily aid, it may have potential value during various stages of the process of speech and language acquisition.

6.4 Speech in noise and non-speech signals

Two topics that have not received attention in this thesis, but which are of eminent importance in the development of a practical signal processing hearing aid, are the robustness of the processing algorithms especially in noise, and the way in which non-speech signals are treated. In this section these topics are briefly touched upon.

The development and application of robust algorithms for extraction of selected speech features is an important but - in most laboratory research settings - neglected point. Most of the perception tasks are being carried out in quiet listening conditions, with optimal S/N ratios.

Algorithm performance is seldom tested with low S/N ratios. Nevertheless, the robustness of the extraction algorithm to adverse acoustic environments (noise, reverberation) may be of especially great help to the profoundly hearing impaired, as he himself faces immense problems understanding speech in noise. Research activities in this area are going on at the University College of London, where the potentials of speech technology solutions to the noise problem are being investigated. Simulated neural net methods are applied in order to extract voice fundamental period information and other speech features robustly in low S/N ratios (Walliker and Howard, 1990). In the framework of a European Research Programme, robustness measures of several speech processing algorithms, including our own (see chapter 5) and that used in the SiVo device, are being carried out at the Laboratory of Experimental Audiology of the Academic Hospital Utrecht. Unpublished results indicate that the SiVo device is able to make a reasonable amount of correct voiced/unvoiced decisions with S/N ratios down to -5 dB, which performance may be better than the profoundly impaired auditory system is capable of when processing the entire sound waveform entering from the hearing aid.

The perception of environmental non-speech signals is of eminent importance in daily circumstances, especially when the sounds have an alarming function such as car horns or fire alarms. Most processing algorithms are especially useful for speech signals, but their performance with non-speech signals may be problematic. In the case of our LPC-based signal processing and generating algorithm, many different non-speech sounds may be returned as white noise because no voicing was detected. Differential spectral characteristics are lost, so that discrimination of non-speech sounds will mainly be based on temporal patterns. However, with more advanced processing schemes non-speech signals may be treated in a way that their special characteristics will not be lost, however different the generated sound may be from the original. Evidence from cochlear implant patients shows that training will help them discriminate among different environmental sounds, such as footsteps, barking dogs, ambulance sirens, whistles, alarm clocks, *etc.* So, as long as the speech processing algorithm is able to generate different outputs with different inputs, interpretation of the resulting output sounds may be learnt in a similar fashion as interpretation of speech sounds. A different solution to the problem would be the implementation of a multi-programme signal processing hearing aid with at least one programme for speech processing and one for non-speech processing. Other programmes may be designed for noisy and quiet backgrounds. The programmes may be selected by the hearing aid

wearer or, in a (semi-)automatic way, by the hearing aid itself. Until now, no such applications have been implemented, nor has much research been carried out as to the effects of processing schemes on environmental sounds.

6.5 Suggestions for future research

In this and previous chapters we have made some suggestions for research that might be carried out in the near future. Recapitulating, the following topics are of interest:

1. Establishing relationships between processed-speech perception and basic auditory functions.
2. Investigation of factors that predetermine successful use of the presentation of certain speech features.
3. Development of reliable and valid tests for patient selection.
4. Development of a structured training programme, or redesign and application of existing programmes, such as are used in the rehabilitation of cochlear implant wearers.
5. Evaluation of the feasibility of these training programmes for profoundly hearing impaired persons using conventional hearing aids.
6. Application of the speech processing hearing aid as a useful training device, in order to make profoundly hearing impaired persons aware of linguistically relevant speech information that they could not derive from the speech waveform passed through a conventional hearing aid.
7. Development and application of robust speech processing algorithms, especially with regard to acoustically adverse situations.
8. Attention for processing of non-speech signals.

Acknowledgement

I wish to thank the cochlear implant team of the Academic Hospital Utrecht, chaired by ENT specialist Dr. A.F. van Olphen, for their permission to mention preliminary post-operative evaluation results from two of their patients. Special credit is due to Margreet Langereis, speech and hearing therapist of the team, for providing me with detailed information on the training programme and testing procedures.

Chapter 7.

References

Benguerel, A.-P. and Pichora-Fuller, M.K. (1982). Coarticulation effects in lipreading. *Journal of speech and hearing research 25*, 600-607.

Binnie, C.A., Montgomery, A.A. and Jackson, P.L. (1974). Auditory and visual contributions to the perception of consonants. *Journal of speech and hearing research 17*, 619-630.

Boothroyd, A. (1984). Auditory perception of speech contrasts by subjects with sensorineural hearing loss. *Journal of speech and hearing research 27*, 134-144.

Boothroyd, A. (1988). Perception of speech pattern contrasts from auditory presentation of voice fundamental frequency. *Ear and hearing 9*, 313-321.

Boothroyd, A. (1990). Signal processing for the profoundly deaf. *Acta oto-laryngologica suppl. 469* (Stockholm), 166-171.

Boothroyd, A., Hnath-Chisolm, T., Hanin, L. and Kishon-Rabin, L. (1988). Voice fundamental frequency as an auditory supplement to the speechreading of sentences. *Ear and hearing 9*, 306-312.

Bosman, A.J. and Smoorenburg, G.F. (1993a). Speech processing for the profoundly hearing impaired. In: E. Ballabio, I. Placencia-Porrero and R. Puig de la Bellacasa (Eds.), *Rehabilitation technology. Studies in Health Technology and Informatics, Vol. 9.*, pp. 113-117. Amsterdam: IOS Press.

Bosman, A.J. and Smoorenburg, G.F. (1993b). Speech processing for the profoundly hearing impaired: aided lipreading of sentences and auditory functions. In preparation.

Braida, L.D., Durlach, N.I., Lippmann, R.P., Hicks, B.I., Rabinowitz, W.M. and Reed, C.M. (1979). Hearing aids - a review of past research on linear amplification, amplitude compression, and frequency lowering. *ASHA Monographs no. 19.*

Breeuwer, M. (1985). *Speechreading supplemented with auditory information.* Unpublished doctoral thesis. Free University, Amsterdam, The Netherlands.

Breeuwer, M. and Plomp, R. (1984). Speechreading supplemented with frequency-selective sound-pressure information. *Journal of the acoustical society of America 76*, 686-691.

Breeuwer, M. and Plomp, R. (1985). Speechreading supplemented with formant-frequency information from voiced speech. *Journal of the acoustical society of America 77*, 314-317.

Breeuwer, M. and Plomp, R. (1986). Speechreading supplemented with auditorily presented speech parameters. *Journal of the acoustical society of America 79*, 481-499.

Busby, P.A., Tong, Y.C. and Clark, G.M. (1984). Underlying dimensions and individual differences in auditory, visual and auditory-visual vowel perception by hearing-impaired children. *Journal of the acoustical society of America 75*, 1858-1865.

Busby, P.A., Tong, Y.C. and Clark, G.M. (1988). Underlying structure of auditory-visual consonant perception by hearing-impaired children and the influences of syllabic compression. *Journal of speech and hearing research 31*, 156-165.

Caplan Hack, Z. and Erber, N.P. (1982). Auditory, visual and auditory-visual perception of vowels by hearing-impaired children. *Journal of speech and hearing research 25*, 100-107.

Carroll, J.D. and Chang, J.-J. (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition. *Psychometrika 35*, 283-319.

Conrad, R. (1977). Lip-reading by deaf and hearing children. *British journal of educational psychology 47*, 60-65.

Corthals, P. (1984). Een eenvoudige visementaxonomie voor spraakafzien [A simple viseme taxonomy for speechreading]. *Tijdschrift voor logopedie en audiologie 14*, 126-134.

Eggermont, J.P.M. (1964). *Taalverwerving bij een groep dove kinderen [Language acquisition in a group of deaf children]*. Groningen: J.B. Wolters.

Erber, N.P. (1972a). Speech-envelope cues as an acoustic aid to lipreading for profoundly deaf children. *Journal of the acoustical society of America 51*, 1224-1227.

Erber, N.P. (1972b). Auditory, visual, and auditory-visual recognition of consonants by children with normal and impaired hearing. *Journal of speech and hearing research 15*, 413-422.

Erber, N.P. (1974a). Effects of angle, distance and illumination on visual reception of speech by profoundly deaf children. *Journal of speech and hearing research 17*, 99-112.

Erber, N.P. (1974b). Auditory-visual perception of speech: a survey. *Scandinavian audiology suppl. 4*, 12-30.

Erber, N.P. (1979). Speech perception by profoundly hearing-impaired children. *Journal of speech and hearing disorders 44*, 255-270.

Faulkner, A., Ball, V., Rosen, S., Moore, B.C.J. and Fourcin, A.J.

(1992). Speech pattern hearing aids for the profoundly hearing impaired: speech perception and auditory abilities. *Journal of the acoustical society of America 91*, 2136-2155.

Faulkner, A., Fourcin, A.J. and Moore, B.C.J. (1990a). Psychoacoustic aspects of speech pattern coding for the deaf. *Acta oto-laryngologica suppl. 469* (Stockholm), 172-180.

Faulkner, A., Rosen, S. and Moore, B.C.J. (1990b). Residual frequency selectivity in the profoundly hearing-impaired listener. *British journal of audiology 24*, 381-392.

Fisher, C.G. (1968). Confusions among visually perceived consonants. *Journal of speech and hearing research 11*, 796-804.

Fourcin, A.J. (1990). Prospects for speech pattern element aids. *Acta oto-laryngologica suppl. 469* (Stockholm), 257-267.

Fourcin, A.J., Douek, E.E., Moore, B.C.J., Rosen, S., Walliker, J.R., Howard, D.M., Abberton, E. and Frampton, S. (1983). Speech perception with promontory stimulation. *Annals of the New York academy of sciences 405*, 281-294.

Fourcin, A.J., Rosen, S.M., Moore, B.C.J., Douek, E.E., Clarke, G.P., Dodson, H. and Bannister, L.H. (1979). External electrical stimulation of the cochlea: clinical, psycho-physical, speech-perceptual and histological findings. *British journal of audiology 13*, 85-107.

Grant, K.W. (1987). Encoding voice pitch for profoundly hearing-impaired listeners. *Journal of the acoustical society of America 82*, 423-432.

Grant, K.W., Ardell, L.H., Kuhl, P.K. and Sparks, D.W. (1985). The contribution of fundamental frequency, amplitude envelope, and voicing duration cues to speechreading in normal-hearing subjects. *Journal of the acoustical society of America 77*, 671-677.

Hnath-Chisolm, T. and Boothroyd, A. (1992). Speechreading enhancement by voice fundamental frequency: the effects of F_0-contour distortions. *Journal of speech and hearing research 35*, 1160-1168.

Horii, Y., House, A.S. and Hughes, G.W. (1971). A masking noise with speech-envelope characteristics for studying intelligibility. *Journal of the acoustical society of America 49*, 1849-1856.

Jackson, P.L., Montgomery, A.A. and Binnie, C. A. (1976). Perceptual dimensions underlying vowel lipreading performance. *Journal of speech and hearing research 19*, 796-812.

Johnson, S.C. (1967). Hierarchical cluster schemes. *Psychometrika 32*, 241-254.

Klein, W., Pols, L.C.W. and Plomp, R. (1970). Vowel spectra, vowel spaces, and vowel identification. *Journal of the acoustical society of*

America 48, 999-1009.

Kozhevnikov, V.A. and Chistovich, L.A. (1966). *Speech: articulation and perception [translated from Russian]*. Washington, D.C.: Joint Publications Research Service, U.S. Dept. of Commerce.

Kricos, P.B. and Lesner, S.A. (1982). Differences in visual intelligibility across talkers. *Volta review 84*, 219-225.

Kruskal, J.B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika 29*, 1-27.

Kuijpers, C.T.L. (1993). *Temporal coordination in speech development.* Unpublished doctoral thesis. University of Amsterdam, The Netherlands.

Lamoré, P.J.J., Verweij, C. and Brocaar, M.P. (1985). Investigations of the residual hearing capacity of severely hearing-impaired and profoundly deaf subjects. *Audiology 24*, 343-361.

Levelt, W.J.M., Geer, J.P. van de and Plomp, R. (1966). Triadic comparisons of musical intervals. *British journal of mathematical and statistical psychology 19*, 163-179.

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *Journal of the acoustical society of America 49*, 467-477.

Lisker, L. (1957). Closure duration and the intervocalic voiced voiceless distinction in English. *Language and speech 33*, 42-49.

McGrath, M. and Summerfield, Q. (1985). Intermodal timing relations and audio-visual speech recognition by normal-hearing adults. *Journal of the acoustical society of America 77*, 678-685.

Miller, G.A. and Nicely, P.E. (1955). An analysis of perceptual confusions among some English consonants. *Journal of the acoustical society of America 27*, 338-346.

Montgomery, A.A. and Jackson, P.L. (1983). Physical characteristics of the lips underlying vowel lipreading performance. *Journal of the acoustical society of America 73*, 2134-2144.

Montgomery, A.A. and Edge, R.A. (1988). Evaluation of two speech enhancement techniques to improve intelligibility for hearing-impaired adults. *Journal of speech and hearing research 31*, 386-393.

Montgomery, A.A., Walden, B.E. and Prosek, R.A. (1987). Effects of consonantal context on vowel lipreading. *Journal of speech and hearing research 30*, 50-59.

Moser, L.M. (1987). Das Würzburger Hörfeld, ein Test für prothetische Audiometrie. *Hals, Nase und Ohr 35*, 318-321.

Nakatani, L.H. and Dukes, K.D. (1973). A sensitive test of speech communication quality. *Journal of the acoustical society of America 53*, 1083-1092.

Nierop, D.J.P.J. van, Pols, L.C.W. and Plomp, R. (1973). Frequency

analysis of Dutch vowels from 25 female speakers. *Acustica 29*, 110-118.

O'Connor, J.D. (1973). *Phonetics*. Harmondsworth, Middlesex, England: Penguin Books Ltd.

Owens, E. and Blazek, B. (1985). Visemes observed by hearing-impaired and normal-hearing adult viewers. *Journal of speech and hearing research 28*, 381-393.

Pandey, P.C., Kunov, H. and Abel, S.M. (1986). Disruptive effects of auditory signal delay on speech perception with lipreading. *Journal of auditory research 26*, 27-41.

Pickett, J.M. (1963). Tactual communication of speech sounds to the deaf: comparison with lipreading. *Journal of speech and hearing disorders 28*, 315-330.

Pickett, J.M. and McFarland, W. (1985). Auditory implants and tactile aids for the profoundly deaf. *Journal of speech and hearing research 28*, 134-150.

Plomp, R. (1988). The negative effect of amplitude compression in multichannel hearing aids in the light of the modulation-transfer function. *Journal of the acoustical society of America 83*, 2322-2327.

Plomp, R. and Mimpen, A.M. (1979). Improving the reliability of testing the speech reception threshold for sentences. *Audiology 18*, 43-52.

Pols, L.C.W., Tromp, H.R.C. and Plomp, R. (1973). Frequency analysis of Dutch vowels from 50 male speakers. *Journal of the acoustical society of America 53*, 1093-1101.

Raaij, J.L. van (1986). De audiosignaalprocessor: eigenschappen en technische omschrijving [The audio signal processer: characteristics and technical description]. *TNO Institute for perception report no. I-4*.

Risberg, A. (1974). The importance of prosodic speech elements for the lipreader. *Scandinavian audiology suppl. 4*, 153-164.

Risberg, A. and Agelfors, E. (1978). Information extraction and information processing in speech-reading. *STL-QPSR 2-3*, 62-82.

Rosen, S., Faulkner, A. and Smith, D.A.J. (1990). The psychoacoustics of profound hearing impairment. *Acta oto-laryngologica suppl. 469* (Stockholm), 16-22.

Rosen, S., Walliker, J.R., Fourcin, A.J. and Ball, V. (1987). A microprocessor-based acoustic hearing aid for the profoundly impaired listener. *Journal of rehabilitation research and development 24*, 239-260.

Schiffman, S.S., Reynolds, M.L. and Young, F.W. (1981). *Introduction to multidimensional scaling: theory, methods and applications*. New York: Academic Press, Inc.

Shannon, R.V., Zeng, F.-G. and Wygonski, J. (1992). Speech recognition using only temporal cues. In: M.E.H. Schouten (Ed.), *The auditory processing of speech: from sounds to words*, pp. 263-274. Berlin/New York: Mouton de Gruyter.

Simpson, A.M., Moore, B.C.J. and Glasberg, B.R. (1990). Spectral enhancement to improve the intelligibility of speech in noise for hearing-impaired listeners. *Acta oto-laryngologica suppl. 469* (Stockholm), 101-107.

Smoorenburg, G.F. (Ed.) (1990a). *Hearing impairment and signal-processing hearing aids. Acta oto-laryngologica suppl. 469 (Stockholm).*

Smoorenburg, G.F. (1990b). Physical versus perceptual dimensions in cochlear implants. In: J.M. Miller and F.A. Spelman (Eds.), *Cochlear implants: models of the electrically stimulated ear*, pp. 105-113. New York: Springer-Verlag, Inc.

Smoorenburg, G.F. and Bosman, A.J. (Eds.) (1993). Hearing impairment and signal-processing hearing aids II. *In preparation as Scandinavian audiology suppl. 38.*

Son, N. van (1993). Perceptual dimensions in lipreading and supplementary acoustic cues for the profoundly hearing impaired: a critical look at the potentials of signal processing hearing aids. *Submitted to Audiology.*

Son, N. van, Bosman, A.J., Lamoré, P.J.J. and Smoorenburg, G.F. (1993a). The perception of complex harmonic patterns by profoundly hearing-impaired listeners. *Audiology 32*, 308-327.

Son, N. van, Bosman, A.J. and Smoorenburg, G.F. (1993b). Discrimination of natural and processed Dutch vowels by profoundly hearing impaired lipreaders. In preparation.

Son, N. van, Huiskamp, T.M.I., Bosman, A.J. and Smoorenburg, G.F. (1993c). Viseme classifications of Dutch consonants and vowels. *Accepted for publication in Journal of the acoustical society of America.*

Stoker, R.G. and French-St. George, M. (1984). Factors influencing speechreading performance: research findings. In: W.H. Northcott (Ed.), *Oral interpreting: principles and practices*, pp. 95-122. Baltimore, Ma.: University Park Press.

Summerfield, A.Q. (1983). Audiovisual speech perception, lipreading and artificial stimulation. In: M.E. Lutman and M.P. Haggard (Eds.), *Hearing science and hearing disorders*, pp. 131-182. London: Academic Press.

Tyler, R.S., Tye-Murray, N. and Lansing, C.R. (1988). Electrical stimulation as an aid to speechreading. *Volta review 90*, 119-148.

Van Tasell, D.J., Soli, S.D., Kirby, V.M. and Widin, G.P. (1987). Speech waveform envelope cues for consonant recognition. *Journal of the acoustical society of America 82*, 1152-1161.

Walden, B.E., Erdman, S.A., Montgomery, A.A., Schwartz, D.M. and Prosek, R.A. (1981). Some effects of training on speech recognition by hearing-impaired adults. *Journal of speech and hearing research 24*, 207-216.

Walden, B.E., Prosek, R.A., Montgomery, A.A., Scherr, C.K. and Jones, C.J. (1977). Effects of training on the visual recognition of consonants. *Journal of speech and hearing research 20*, 130-145.

Walliker, J.R. and Howard, I. (1990). Real-time portable multi-layer perceptron voice fundamental-period extractor for hearing aids and cochlear implants. *Speech communication 9*, 63-71.

Walliker, J.R., Rosen, S., Douek, E.E., Fourcin, A.J. and Moore, B.C.J. (1983). Speech signal presentation to the totally deaf. *Journal of biomedical engineering 5*, 316-320.

Wingfield, A. (1975). The intonation-syntax interaction: prosodic features in perceptual processing of sentences. In: A. Cohen and S.G. Nooteboom (Eds.), *Structure and process in speech perception*, pp. 146-160. Berlin, Heidelberg, London: Springer-Verlag.

Woodward, M.F. and Barber, C.G. (1960). Phoneme perception in lipreading. *Journal of speech and hearing research 3*, 212-222.

Wozniak, P. and Jackson, P.L. (1979). Visual vowel and diphthong perception from two horizontal viewing angles. *Journal of speech and hearing research 22*, 354-365.

Zee-Zetstra, J.A. van der (1985). *Een goed verstaander...: geprogrammeerde instructie in verbale communicatietechniek voor groepen volwassenen met een verworven auditieve handicap [A word is enough...: Programmed instruction in verbal communication techniques for adult groups with postlingually acquired hearing impairment]*. The Hague, The Netherlands: Nijgh & Van Ditmar Educatief.

Zwicker, E., Flottorp, G. and Stevens, S. (1957). Critical band width in loudness summation. *Journal of the acoustical society of America 29*, 548-557.

Chapter 8.

Summary: speech processing strategies for the profoundly hearing impaired

This thesis discusses a number of aspects that are relevant to the development and application of speech processing techniques for the profoundly hearing impaired. In chapter 1 an introduction is given into the problem of speech perception by the profoundly hearing impaired, and possible solutions by application of advanced digital techniques making use of our knowledge of speech.

Many profoundly hearing impaired people have such limited auditory capacities that their perception of speech is severely restricted to the extent that it is not possible without lipreading. In general, profound hearing impairment may be characterized by an average hearing loss of more than 90 dB HL, frequency span reduction to about 2 kHz, dynamic range reduction to as little as 10-20 dB, and limited frequency resolution. These residual auditory capacities hardly permit the profoundly hearing impaired listener to extract useful information from the complex speech signal. As a consequence, little discrimination among speech sounds is possible. Hearing aids do not compensate for this discrimination loss and are thus of little use in speech perception.

Several processing techniques have been endeavoured in order to simplify the speech signal and thus emphasize the main characteristics. Techniques that compensate for dynamic range and frequency span reduction, however, do not offer a solution for reduced frequency selectivity; sometimes they have a countereffect, rather. One solution to the loss of speech sound discrimination may be found in speech feature extraction, an approach which has been applied for some years now in a number of inner ear prostheses for the deaf. In this technique, essential speech information is extracted from the speech signal (while other, possibly disturbing, information is removed) and recoded into a new signal, such that it can be optimally transmitted to the perceiver. With an inner ear prosthesis the speech signal is transformed to an electric signal which is then presented to the cochlear nerve. In the case of profound impairment, the speech information will be transmitted by an acoustic device rather than an inner ear prosthesis. In the development and application of this so-called *speech or signal processing hearing aid*, four areas of interest may be distinguished:

(1) selection of the most relevant speech features to be presented to the profoundly hearing impaired;
(2) the use of robust feature extraction algorithms;
(3) encoding of a new acoustic signal on the basis of the extracted information, optimally matched with the auditory capacities of the profoundly hearing impaired;
(4) structured rehabilitative training, in which the patient has to learn and interpret the new speech code (and other non-speech signals).

The main objective of the study described in this thesis was to investigate the information processing capacities of the profoundly hearing impaired for speech or speech-like patterns, both auditory and visual (lipreading). Closely related to that is the second objective, the definition of a well-matched speech signal aimed at improved speech perception when compared to an unprocessed speech signal. The study was thus mainly concerned with questions within areas (1) and (3): what are the relevant features in speech for profoundly hearing impaired people, and in what way should these features be presented to them, regarding the limited auditory capacities that they have available ?

In all experiments major attention was paid to psychophysical relationships, both auditory and visual. This involves the study of how acoustic or articulatory distinctions between (speech) sound patterns are reflected in perceptual distinctions, and the interpretation of the underlying perceptual dimensions that may account for the observed arrangements. Furthermore, it involves the investigation of basic auditory capacities in order to assess the relative contributions of hearing and lipreading in speech perception.

The study was structured as follows. Chapters 2 and 3 present data on the accessibility of speech or speech-like information through lipreading and through the profoundly impaired hearing sense. The underlying dimensions in visual and auditory perception are distinguished. Chapter 4 then gives an overview of experimental work in visual and auditory-visual phoneme perception, and provides conclusions about the type of speech information that may be presented as an acoustic support to the (profoundly) hearing impaired lipreader. The second part of chapter 4 describes a number of studies tackling the problem of supplementary speech cue presentation and discusses the general outcomes. On the basis of the conclusions from the studies in chapters 2 and 3, two supplementary speech codes were designed; their efficacy for the profoundly hearing impaired lipreader was evaluated in a study described in chapter 5. Chapter 6, finally, discusses a number of important or interesting topics of research for the near future.

The study described in chapter 2 investigated the amount and type of information that a lipreader can extract from visually presented phonemes. The aim of the research was (1) to establish sets of visually similar phonemes (visemes) for Dutch consonants and vowels; (2) to interpret the visual-perceptual dimensions underlying this classification and relate them to acoustic-phonetic parameters; (3) to establish the effect of lipreading expertise on these viseme classifications.

Four subject groups participated in the study. They differed with respect to their knowledge of and experience with lipreading (lipreading expertise): normally hearing medical students (*MSt*) and speech therapy students (*STh*), and hearing impaired persons with normal lipreading ability (*HIn*) and with expert lipreading ability (*HIx*). All subjects were presented, individually and in quiet, with videotaped lists of meaningless Dutch syllables, consisting of all Dutch consonants within three vowel contexts, and of all Dutch vowels within four consonant contexts. Three speakers had pronounced all syllable lists.

Identification scores amounted to 28% for initial and medial consonants, and 44% for final consonants and vowels. No significant difference was found among the subject groups, although the expert lipreaders of group *HIx* tended to have slightly higher scores. Important visual features in consonant lipreading are *lip articulation*, *degree of oral cavity opening* and *place of articulation*, leading to the following viseme classification: /p,b,m/, /f,v,ʋ/, /s,z,ʃ/, and /t,d,n,j,l,k,x,r,ŋ,h/. In the acoustic domain, these features may be related to spectral composition of the speech sounds. The most important vowel features in lipreading are *lip rounding* and *degree of vertical lip opening*. Secondarily, *vowel duration* plays a role in further differentiation. The following visemes may be recognized: /i,I,e,ɛ,ɛi,a,ɑ/, /u,y,œ,ɔ/, /ø,o/, and /au,œy/. In the acoustic domain, lip rounding may roughly be related to the second formant, lip opening to the first formant. As with the identification scores, no difference was found among the subject groups with respect to viseme recognition. It was concluded that different degrees of lipreading expertise, as we defined it, are not reflected in different viseme classifications.

The processing capacity for auditory information was the topic of the study reported in chapter 3. In this study the profoundly hearing impaired listener's perception of complex harmonic patterns was investigated.

In part I of chapter 3, the subjects' ability to judge similarities among eight different but related harmonic complexes was studied. The patterns contained different numbers of harmonics to a 125 Hz

fundamental frequency (maximally to the 12th harmonic). The harmonics had been spread over the spectrum in various ways: evenly or in a 'peaked' fashion, with peaks consisting of either one harmonic or two adjacent harmonics. In a triadic task, a subject had to compare three complexes at a time and indicate the pair of most similar and the pair of least similar complexes. These similarity ratings were analysed by a method of multidimensional scaling (INDSCAL); they could be represented in a two-dimensional perceptual space. The perceptual dimensions could be related to a temporal and a spectral cue. The temporal cue was based on the presence of adjacent harmonics in the high-frequency region, creating a *beat-like* sensation in the listener. The spectral cue related to the balance of high and low frequency components: high-frequency components could only be perceived when there was no low-frequency energy present or when there was a substantial separation between low and high frequency components.

In part II of chapter 3, the effect of adjacent harmonics on the identification of synthetic vowels was investigated. Three sets of synthetic vowels were presented to the subjects. Each vowel was realized by summing harmonically related in-phase sinusoids at two formant frequencies. The sets differed in the number of sinusoids per formant: one, two or three. Set 1 produced significantly lower identification scores than sets 2 and 3 (13% versus 22%), indicating that two-component or three-component formants lead to better vowel recognition than when only one component per formant is used. Vowel confusions were analysed by means of INDSCAL, and two-dimensional solutions were found to give an adequate representation of similarities among the vowels. The two dimensions could be related to spectral cues, especially F_2, and to vowel duration.

The overall results showed the limited but perhaps usable ability of the profoundly impaired ear to handle spectral information from two frequency channels. They suggest that vowels may be represented by a single formant or even by two formants if there is ample separation between the two (in such vowels as /i,I,e,y/). The use of two harmonic components per formant obviates the use of a spectral f_0 component cueing pitch; pitch may rather be coded on the basis of the inter-harmonic distance (the *residue* effect).

Chapter 4 gives a review of research activities that have provided relevant data for the development and application of signal processing hearing aids for the profoundly hearing impaired. The review especially addresses the following two issues: (1) the kind of speech information that needs be presented in order to support the profoundly hearing

impaired lipreader, and (2) the way in which this information should be presented so as to match optimally to the residual hearing capacities.

As to the first issue, research has shown that visible (labial) articulatory movements provide the lipreader with vowel lip rounding and lip opening information and with consonant place information. Lipreading hardly provides useful prosodic cues. In order for lipreaders to profit from supplementary acoustic signals, these should provide at least information about vowel duration, consonant voicing and manner of articulation, as well as prosodic cues. This calls for the presentation of voicing and amplitude information.

With respect to the second issue, it has been argued that profoundly hearing impaired persons are not well able to extract useful information from the complex speech signal. Therefore, feature extraction techniques have been developed in order to simplify the speech signal and emphasize its main characteristics. Various processed speech patterns carrying f_0 and amplitude information have been evaluated as supplements to both normally hearing and profoundly hearing impaired lipreaders. It has been convincingly shown that lipreaders profit from these highly simplified speech patterns; so far, however, the feature extraction approach has failed to yield better results than presentation of unprocessed speech conveyed by well-fitted hearing aids. Chapter 4 concludes with a discussion of these unexpectedly meagre results. It is argued that the minimal information provided (f_0 and amplitude) is too minimal for most of the profoundly hearing impaired who still retain some degree of frequency selectivity. For them, addition of spectral cues may be considered.

On the basis of the experimental results from the studies described in chapters 2 and 3, and following the discussion of chapter 4, two versions of a resynthesized speech signal containing spectral information were defined and evaluated with respect to their functionality as a supplement to lipreading. This evaluation was investigated in a vowel discrimination experiment presented in chapter 5. The experiment included the following four objectives:

(1) comparing the coded signals to the original, unprocessed, amplified speech signal;

(2) to measure vowel discrimination within and across three visemes (rounded and unrounded vowels, and diphthongs). As visemes are clearly distinct groups of visually similar phonemes, across-viseme discrimination was expected to be nearly perfect purely on the basis of lipreading, while within-viseme discrimination would only be possible in the auditory-visual presentation mode;

(3) the comparison between the two speech codes in a small number of vowel pairs;
(4) investigating the underlying perceptual dimensions and their relationship to physical vowel parameters.

The vowel discrimination task (AXB paradigm) was performed by a group of 12 profoundly hearing impaired persons. AXB triplets were pronounced by a female speaker and recorded onto colour video tapes. Three speech signals were created: linearly amplified speech (AMPL), and two coded speech signals, containing first formant (F1) or first and second formant (F1F2) information. The tapes were replayed in an auditory-visual condition. In order to check the relative contributions of auditory and visual perception, two control measures were run with some of the subjects: visual-only (n=11) and auditory-only (n=6). Another auditory-only control measure was run with a group of six normally hearing subjects, only presented with the F1 and F1F2 speech codes, in order to measure optimal discrimination scores for processed vowels. As said, discrimination among vowels was measured within and across three visemes (rounded and unrounded vowels, and diphthongs). All scores were corrected for chance (p=.5). Vowel confusions were analysed by means of INDSCAL. In order to be able to relate the perceptual dimensions to physical vowel properties, duration, level and first and second formant frequencies were measured from the original vowel realizations.

The control measures showed that coded speech permits normally hearing subjects to reach score levels of 97% (F1F2) and 83% (F1). For the profoundly hearing impaired subjects, the visual-only discrimination score for vowels belonging to different visemes was 94%. When sound was added, this across-viseme performance increased by a few percent; no difference was found among the speech conditions. As to discrimination within visemes, all scores were well beyond chance level. The visual-only score was 58%, while auditory-visual scores were significantly higher: 79% (AMPL), 73% (F1F2), and 75% (F1). The AMPL scores were significantly higher than those for the two speech codes, which were not significantly different from one another. The INDSCAL analyses indicated that, in all conditions, discrimination was based on two perceptual dimensions, explaining 80-90% of the total variance. Dimension I related to F_1 or vertical degree of lip opening, dimension II to vowel duration.

Chapter 6, finally, discusses the implications of the general conclusions drawn from the various parts of the study. One of the main topics for further research in the near future is the study of the relationship

between auditory functions and speech perception performance using processed speech, in order to answer the question: what is the profile of the profoundly hearing impaired person who profits more from processed than from unprocessed speech ? In future, individual fitting procedures, including the selection of speech features to be presented, may be established. This would call for a (battery of) predictive test(s) for patient selection. For the moment, it seems clear that defining one speech processing scheme for an entire group as heterogeneous as the group of profoundly hearing impaired is not a promising road to follow.

A second important topic is the use of a well-structured training programme, not only for future signal processing hearing aid wearers, but also for those profoundly hearing impaired who use a conventional type of hearing aid with limited success. There is evidence from cochlear implant wearers who also wear a hearing aid in the other ear that their hearing aid use improves as a result of intensive auditory training with the implant system. This improvement is partly a result of the training itself. Alternatively, the implant system may have transmitted speech feature information that the patient formerly had not been able to extract from the entire speech waveform passed through the hearing aid. Speech processing may thus be considered as a useful tool in training, for which reason the signal processing hearing aid may be used in future as a practical training device.

Chapter 9.

Samenvatting: strategieën voor de bewerking van spraaksignalen ten behoeve van ernstig slechthorenden

Dit proefschrift behandelt een aantal aspecten die relevant zijn voor de ontwikkeling en toepassing van spraakbewerkingstechnieken ten behoeve van de ernstig slechthorenden. Hoofdstuk 1 geeft een inleiding tot de problemen van ernstig slechthorenden met betrekking tot het verstaan van spraak; mogelijke oplossingen daarvan in de vorm van geavanceerde digitale technieken die gebaseerd zijn op onze kennis van het verschijnsel spraak worden besproken.

Veel ernstig slechthorenden hebben een dusdanig geringe hoorcapaciteit dat het waarnemen van spraak niet of nauwelijks mogelijk is zonder liplezen. In het algemeen kan ernstige slechthorendheid als volgt worden gekenschetst: een gemiddeld gehoorverlies van 90 dB HL, een beperking van het frekwentiebereik tot ongeveer 2 kHz, een verminderd dynamisch bereik tot 10 á 20 dB en een verminderd frekwentieoplossend vermogen. Met dergelijke geringe hoorresten is de ernstig slechthorende nauwelijks in staat om enige bruikbare informatie uit het spraaksignaal te halen. Het gevolg hiervan is dat weinig onderscheid tussen verschillende spraakklanken waarneembaar is. Hoortoestellen bieden voor dit discriminatieverlies geen oplossing en hebben voor het spraakverstaan derhalve weinig waarde.

In de loop der tijd zijn verschillende technieken ontwikkeld om het complexe spraaksignaal te vereenvoudigen en aldus de belangrijke informatie te benadrukken. Helaas sorteren veel oplossingen voor het probleem van verminderd dynamisch en frekwentiebereik weinig effect waar het erom gaat het verlies aan frekwentieresolutie te compenseren; soms wordt dat probleem zelfs verergerd. Een mogelijk bruikbare oplossing is de techniek van de *kenmerkextractie*. Dit is een aanpak die al vele jaren met succes gebruikt wordt in een aantal systemen voor cochleaire implantatie bij doven. De techniek houdt in dat een of meer zinvol geachte spraakkenmerken uit het signaal worden gehaald (geëxtraheerd), terwijl mogelijk storende componenten worden verwijderd. De zinvolle informatie wordt als een nieuwe spraakcode aangeboden aan de luisteraar op een manier die optimaal aansluit bij de beperkte auditieve vermogens. Met een cochleair implantaat (binnen-

oorprothese) wordt het spraaksignaal omgezet in een elektrisch signaal en rechtstreeks via de gehoorzenuw aangeboden. In het geval van ernstige slechthorendheid wordt het nieuw gecodeerde signaal akoestisch aangeboden via een hoortoestel. In de ontwikkeling en toepassing van deze zgn. *spraak- of signaalbewerkende hoortoestellen* kunnen vier aandachtsgebieden onderscheiden worden:

(1) de keuze ten aanzien van de aan ernstig slechthorenden aan te bieden spraakkenmerken;
(2) het gebruik van robuuste algoritmen voor kenmerkextractie;
(3) het definiëren van een nieuw (spraak)signaal op basis van de geëxtraheerde informatie;
(4) het ontwerpen en gebruiken van gestructureerde revalidatie-programma's waarin de hoortoesteldrager de nieuwe code (spraak en omgevingsgeluiden) als zinvolle informatie leert te interpreteren.

De belangrijkste doelstelling van de in dit proefschrift beschreven studie was het onderzoek naar de informatieverwerkende capaciteit van ernstig slechthorenden in relatie tot hun waarneming van spraakgeluid, waaronder zowel de auditieve als de visuele waarneming wordt verstaan (horen en liplezen). Het tweede doel van het onderzoek volgt hier rechtstreeks uit: de definitie van een goed aan de hoorresten aangepast bewerkt spraaksignaal waarmee door de slechthorende liplezer een beter spraakverstaan mogelijk is dan met het oorspronkelijke, onbewerkte, spraaksignaal. Deze twee doelstellingen geven aan dat het onderzoek met name binnen de aandachtsgebieden (1) en (3) valt. De vragen zijn aldus: wat is de voor ernstig slechthorende meest zinvolle informatie uit het spraaksignaal en op welke manier kan die informatie zo gunstig mogelijk worden overgedragen, gezien de beperkte auditieve vermogens ?

In alle beschreven experimenten is steeds veel aandacht uitgegaan naar psychofysische relaties, zowel in auditief als visueel opzicht. Dit heeft betrekking op de bestudering van de wijze waarop akoestische of articulatorische verschillen tussen spraaksignalen weerspiegeld worden in perceptieve verschillen, en de interpretatie van de perceptieve dimensies die aan de gevonden verbanden ten grondslag liggen. Verder worden in het onderzoek ook de auditieve basisfuncties bestudeerd, als grondslag voor de beoordeling van de relatieve bijdragen uit de auditieve en de visuele waarneming van de spraakinformatie.

Het onderzoek is als volgt opgebouwd. In de hoofdstukken 2 en 3 worden gegevens gepresenteerd over de toegankelijkheid van informatie uit het spraaksignaal via liplezen en via het ernstig verminderde gehoorvermogen. De dimensies die ten grondslag liggen aan de visuele en auditieve waarneming van die informatie worden gekenschetst. In hoofdstuk 4 wordt ten eerste een overzicht gegeven van experimenteel

werk op het terrein van visuele en auditief-visuele foneemperceptie. Hieruit worden conclusies getrokken ten aanzien van de spraakinformatie die als aanvulling op het lipbeeld zou kunnen worden aangeboden aan de (ernstig) slechthorende liplezer. Het tweede deel van hoofdstuk 4 beschrijft een aantal onderzoeken naar de bruikbaarheid van specifieke lipbeeld-ondersteunende spraaksignalen; de algemene resultaten hiervan worden besproken. Op basis van de conclusies uit het experimentele werk dat beschreven staat in hoofdstukken 2 en 3 hebben we zelf twee spraakcodes gedefinieerd. De waarde ervan hebben we geëvalueerd in een experiment naar klinkerdiscriminatie door ernstig slechthorende liplezers. Dit onderzoek staat beschreven in hoofdstuk 5. In hoofdstuk 6 wordt tenslotte een aantal zaken op een rijtje gezet. De belangrijkste vragen en onopgeloste problemen worden in het kort besproken en suggesties voor nader onderzoek worden gegeven.

In hoofdstuk 2 wordt een lipleesexperiment beschreven waarin onderzocht werd welke informatie uit spraak (in syllaben ingebedde fonemen) in welke mate toegankelijk is door puur liplezen, zonder ondersteuning van geluid. Het doel van het experiment was drieledig. Ten eerste werd gezocht naar een classificatie van klinkers en consonanten op basis van hun visueel waarneembare eigenschappen. Wanneer spraakklanken sterk gelijkende eigenschappen hebben zullen ze door de liplezer regelmatig met elkaar verwisseld worden. Een dergelijke groep visueel moeilijk te onderscheiden fonemen wordt ook wel aangeduid met de term *viseem*. Het tweede doel van het onderzoek was het vaststellen van de (articulatorische) kenmerken die leiden tot clustering in visemen en het relateren ervan aan akoestisch-fonetische kenmerken. Ten derde werd het effect van lipleesexpertise op de viseemclassificatie onderzocht.

Aan het onderzoek deden vier groepen proefpersonen mee. Ze werden onderscheiden met betrekking tot hun expliciete kennis over en persoonlijke ervaring met liplezen (*lipleesexpertise*): normaalhorende geneeskundestudenten (*MSt*) en logopediestudenten (*STh*), en slechthorenden met een gewone lipleesvaardigheid (*HIn*) en een bijzonder goede lipleesvaardigheid in de dagelijkse omgang (*HIx*). Alle proefpersonen kregen videobanden te zien met door drie sprekers voorgelezen lijsten consonanten en klinkers, ingebed in specifieke konteksten (drie klinkers voor de consonanten, vier consonanten voor de klinkers).

Foneemidentificatiescores liepen uiteen van 28% voor initiële en mediale consonanten tot 44% voor finale consonanten en klinkers. Hoewel de expert-liplezers gemiddeld steeds iets hoger scoorden dan de andere groepen, waren de verschillen tussen de vier groepen statistisch

niet significant. De voor het liplezen van consonanten belangrijke kenmerken zijn *liparticulatie, verticale graad van lipopening* en *plaats van articulatie.* De waarneming van deze kenmerken leidt tot de volgende viseemindeling: /p,b,m/, /f,v,ʋ/, /s,z,ʃ/ en /t,d,n,j,l,k,x,r,ŋ,h/. De kenmerken kunnen gerelateerd worden aan de spectrale samenstelling van de spraakklanken. De belangrijkste klinkerkenmerken in het liplezen zijn *lipronding* en *mate van lipopening.* Daarnaast speelt ook *klinkerduur* een rol in met name de ronde klinkers. Vier visemen kunnen worden onderscheiden: /i,I,e,ɛ,a,ɑ/, /u,y,œ,ɔ/, /ø,o/ en /au,œy/. Lipronding is globaal gerelateerd aan de tweede-formantfrekwentie, verticale graad van lipopening aan de eerste-formantfrekwentie. De viseemclassificatie was voor alle proefpersoongroepen hetzelfde. Hieruit werd geconcludeerd dat verschillende niveaus in lipleesexpertise niet tot uiting komen in verschillen in viseemindeling.

De verwerkingscapaciteit voor auditieve informatie door ernstig slechthorenden was het onderwerp van het onderzoek dat in hoofdstuk 3 staat beschreven. We hebben daarin specifiek de waarneming van complexe harmonische patronen onderzocht.

In deel I van hoofdstuk 3 hebben we bestudeerd in hoeverre ernstig slechthorende proefpersonen in staat waren om verschillen tussen harmonische complexen waar te nemen en te beoordelen. Elk van acht patronen bestond uit een aantal spectrale componenten die in harmonische relatie stonden tot de grondtoon van 125 Hz. De verdeling van de componenten over het spectrale bereik (125-1500 Hz) was volgens drie criteria bewerkstelligd: het aantal componenten, uiteenlopend van minimaal twee tot maximaal twaalf; een gelijkmatige of een gegroepeerde (gepiekte) verdeling; de realisatie van een piek: hetzij één enkele hetzij twee aangrenzende componenten. In een triadisch experiment moest een proefpersoon voor elk drietal harmonische complexen aangeven welke twee patronen het meest, en welke twee het minst op elkaar leken. Deze gelijkheidsoordelen werden geanalyseerd met behulp van een multidimensionele schalingsmethode (INDSCAL). De oordelen bleken langs twee perceptieve dimensies gerepresenteerd te kunnen worden, die geïnterpreteerd werden als een temporele en een spectrale dimensie. De temporele dimensie was gebaseerd op de waarneembare verschillen tussen patronen met enkele en patronen met aangrenzende componenten in het hoogfrekwente deel van het spectrum. De aangrenzende componenten verleenden aan het complex een ruwheidssensatie die het gevolg was van zwevingen. De spectrale dimensie was gerelateerd aan verschillen in de hoog/laagbalans van componenten in verschillende patronen. Patronen met hoogfrekwente

componenten werden alleen perceptief onderscheiden van de andere patronen wanneer er geen maskering optrad door laagfrekwente componenten. Dit was het geval indien er geen laagfrekwente energie aanwezig was of indien de afstand tussen laag- en hoogfrekwente componenten groot genoeg was.

In deel II van het derde hoofdstuk werd het effect gemeten van de aanwezigheid van aangrenzende componenten op de identificatie van synthetische klinkers. Drie sets klinkers werden gemaakt door het optellen van harmonisch gerelateerde componenten op de posities van formanten F_1 en F_2. De interharmonische afstand was 100 Hz. De twee formanten van iedere klinker werden op drie manieren gerealiseerd: als een enkele component (set 1), als twee aangrenzende (set 2) of als drie aangrenzende (set 3) componenten. Set 1 leverde een identificatiescore op van 13%, significant lager dan de scores op sets 2 en 3: 22%. Het gebruik van meerdere componenten om een formant te realiseren levert dus een betere identificatie op. De opgetreden klinkerverwisselingen werden met INDSCAL geanalyseerd en gerepresenteerd langs twee perceptieve dimensies, wederom een spectrale en een temporele dimensie. De spectrale dimensie bleek met name gerelateerd aan F_2; de temporele dimensie had betrekking op de klinkerduur.

De gezamenlijke resultaten van de twee delen tonen aan dat de ernstig slechthorende in staat is om, zij het in geringe mate, spectrale informatie te verwerken uit twee afzonderlijke frekwentiekanalen. Dit suggereert dat het mogelijk is om klinkers te presenteren als één- of zelfs tweeformantige klinkerpatronen (mits de afstand tussen de formanten groot genoeg is, zoals in /i,I,e,y/). Het gebruik van aangrenzende componenten per formant maakt bovendien het gebruik van de grondtoon als spectrale component overbodig: de toonhoogte van het spraaksignaal kan afgeleid worden uit de interharmonische afstand (het *residu*-effect).

Hoofdstuk 4 bevat een overzicht van onderzoek dat belangrijke resultaten heeft opgeleverd voor de ontwikkeling van spraakbewerking in akoestische hoortoestellen. Het overzicht is bedoeld om twee onderwerpen te belichten: (1) de informatie die nodig is als ondersteuning voor de liplezer, en (2) de manier waarop deze informatie het best kan worden overgedragen, gezien de beperkte gehoorvermogens.

Met betrekking tot de benodigde informatie ter aanvulling van het lipbeeld is duidelijk aangetoond dat het liplezen informatie verschaft over lipronding en lipopening van klinkers en over de (labiale) plaats van articulatie van consonanten. Prosodische informatie wordt door middel van liplezen nauwelijks verkregen. Indien liplezers willen

profiteren van ondersteunend spraakgeluid, dan zal dit informatie moeten leveren over klinkerduur, stemhebbendheid en manieren van articulatie van consonanten en prosodische eigenschappen. Dit suggereert sterk de aanbieding van stemhebbendheid en amplitude.

Voor wat betreft de manier van overdracht speelt de techniek van de kenmerkextractie een grote rol. Ernstig slechthorenden hebben grote problemen met het waarnemen van de belangrijke spraakkenmerken uit het complexe spraaksignaal, en een kenmerkextractietechniek zou gebruikt kunnen worden om het spraaksignaal als het ware voor te bewerken en te hercoderen tot een goed te percipiëren signaal. Tot nu toe heeft men vooral signalen die slechts f_0- en amplitude-informatie bevatten gebruikt voor ondersteuning van de (normaalhorende en ernstig slechthorende) liplezer. Uit dergelijk onderzoek is in ieder geval naar voren gekomen dat die geringe spraakinformatie voldoende is om de liplezer aanzienlijk te ondersteunen in zijn visuele spraakperceptie. Een verbetering ten opzichte van het gebruik van onbewerkte spraaksignalen, zoals een gewoon hoortoestel die aanbiedt, is echter nog niet opgetreden. Hoofdstuk 4 besluit met een bespreking van deze, gezien de verwachtingen toch geringe, resultaten: er wordt op gewezen dat signalen met alleen f_0 en amplitude wellicht te weinig informatie bevatten voor een aantal ernstig slechthorenden, zeker als er nog enige frekwentie-selectiviteit aanwezig is. Voor die personen zou extra ondersteuning gevonden kunnen worden in de aanbieding van spectrale kenmerken.

Naar aanleiding van de resultaten uit de onderzoeken naar visemen en harmonische complexen (hoofdstukken 2 en 3) hebben we twee geresynthetiseerde spraakcodes gedefinieerd op basis van de eerste twee formanten. Hun bruikbaarheid als aanvulling op het lipbeeld is onderzocht in een klinkerdiscriminatie-experiment, beschreven in hoofd-stuk 5. Het onderzoek had de volgende vier doelstellingen:

(1) het vergelijken van de twee spraakcodes met het oorspronkelijke, onbewerkte spraaksignaal;
(2) het meten van klinkerdiscriminatie *binnen visemen* en *tussen visemen* (ronde en ongeronde klinkers, tweeklanken). De veronderstelling was dat discriminatie tussen verschillende visemen al mogelijk zou zijn op basis van alleen het lipbeeld, terwijl daarentegen voor discriminatie binnen visemen een ondersteunend akoestisch signaal noodzakelijk zou zijn;
(3) een vergelijking maken tussen de twee spraakcodes binnen een geselecteerde klinkerset;
(4) het onderzoeken van de perceptieve dimensies die aan klinkerdiscriminaties en -verwisselingen ten grondslag liggen en het

leggen van verbanden tussen die dimensies en een aantal fysische parameters waarmee de klinkers akoestisch worden beschreven (F_1, F_2, klinkerduur en klinkerniveau).

De klinkerdiscriminatie (een AXB paradigma) werd uitgevoerd door een groep van 12 ernstig slechthorende proefpersonen. De AXB-drietallen werden door een vrouwelijke spreker uitgesproken en vastgelegd op videoband. Naast het onbewerkte spraaksignaal (AMPL) zijn twee spraakcodes gebruikt, die ofwel alleen F_1 ofwel F_1 plus F_2 bevatten (codes F1 en F1F2). De drie stimulusbanden werden gepresenteerd in auditief-visuele condities. Om de relatieve bijdrage van de afzonderlijke modaliteiten aan de auditief-visuele perceptieve te controleren, deden een aantal van de 12 proefpersonen ook mee aan puur visuele (n=11) en puur auditieve (n=6) metingen. Daarnaast werd een extra controle uitgevoerd op de optimaal haalbare scores met bewerkte spraak, waarvoor een groep van zes normaalhorende het experiment uitvoerde in puur auditieve condities (alleen F1 en F1F2). Klinkerdiscriminatie werd gemeten. Zoals gezegd werd discriminatie binnen visemen en tussen visemen gemeten. Alle discriminatiescores werden gecorrigeerd voor kansniveau. De klinkerverwisselingen werden door middel van INDSCAL geanalyseerd. Om de perceptieve dimensies te kunnen interpreteren werden van iedere onbewerkte klinkerrealisatie de klinkerduur, het niveau en de formantfrekwenties F_1 en F_2 bepaald.

De controlemetingen lieten zien dat normaalhorenden met bewerkte spraak discriminatiescores behalen van 97% (F1F2) en 83% (F1). De ernstig slechthorenden behaalden met puur liplezen een gemiddelde discriminatiescore voor klinkers uit verschillende visemen van 94%. Bij toevoeging van geluid werd deze tussen-visemenscore nog enkele procenten hoger; er was echter geen enkel verschil tussen de drie spraakcondities. Discriminatie binnen visemen bleek in alle condities ruim boven het kansniveau te liggen. De puur visuele score was (58%), en in de auditief-visuele condities werden significant hogere scores gevonden: 79% (AMPL), 73% (F1F2) en 75% (F1). AMPL bleek significant beter te zijn dan de twee coderingen, die onderling niet significant van elkaar verschilden. De INDSCAL analyses gaven aan dat in alle condities twee dimensies 80-90% van alle variantie verklaarden. Dimensie I kon daarbij gerelateerd worden aan F_1 of de verticale graad van lipopening, dimensie II aan klinkerduur.

De implicaties van de verschillende experimentele studies worden in hoofdstuk 6 verder besproken. Eén van de belangrijke onderwerpen voor toekomstig onderzoek is de relatie tussen de perceptie van bewerkte spraak en de aanwezige auditieve capaciteiten. Dit onderzoek zal een

antwoord moeten geven op de vraag wat het profiel is van de ernstig slechthorende die daadwerkelijk profijt ondervindt van bewerkte spraak in vergelijking met het oorspronkelijke spraaksignaal. In de toekomst zullen wellicht individuele aanpassingsstrategieën, inclusief keuze van aan te bieden spraakkenmerken, vastgesteld worden, waarvoor een aantal predictieve testen voor patiëntselectie benodigd zou zijn. Voorlopig lijkt het in ieder geval duidelijk dat het weinig zin heeft om één bepaalde codering te gaan gebruiken voor een groep die zo heterogeen is als deze groep van ernstig slechthorende liplezers.

Een tweede belangrijk onderwerp is het gebruik van een goed gestructureerde revalidatietraining, niet alleen ten behoeve van ernstig slechthorenden die in aanmerking zouden komen voor een signaalbewerkend hoortoestel, maar ook voor hen die een conventioneel hoortoestel gebruiken waar ze relatief weinig profijt van hebben. Uit onderzoek met implantaatdragers die in het andere oor een hoortoestel dragen krijgen we aanwijzingen dat hun hoortoestelgebruik (sterk) verbetert als gevolg van de gestructureerde training die ze ontvangen. Daarnaast lijkt het erop dat een impuls voor het hoortoestelgebruik juist ook komt uit het implantaat, dat geselecteerde spraakkenmerken overdraagt die de dove voorheen niet kon waarnemen met zijn hoortoestel. In die zin kan de techniek van de spraakbewerking dus als een goed trainingsmiddel opgevat worden; het signaalbewerkend hoortoestel zou derhalve in de toekomst kunnen gaan fungeren als een trainingshulpmiddel in de revalidatie van ernstig slechthorenden.

Toelichting op het proefschrift

Op weg naar een 'slim' hoortoestel voor ernstig slechthorenden

Waar dit proefschrift over gaat

In dit proefschrift wordt een onderzoek beschreven waarin getracht is een oplossing te vinden voor een probleem waar veel ernstig slechthorenden mee te kampen hebben: een sterk verminderd vermogen om spraak te verstaan. De huidige hoortoestellen hebben als voornaamste funktie het versterken van (spraak)geluid, zodat men weer gaat *horen*, maar bieden in het geval van ernstige slechthorendheid geen oplossing voor het verbeteren van het spraak*verstaan*. Ernstig slechthorenden zijn hiervoor sterk afhankelijk van liplezen; het hoortoestel dient vaak slechts als ondersteuning van het liplezen.

In een onderzoek van de afdeling KNO/Experimentele Audiologie van het Academisch Ziekenhuis Utrecht is onderzocht hoe deze ondersteunende funktie van het hoortoestel geoptimaliseerd kan worden. Uitgangspunten hierbij waren: (1) de informatie die in het spraakgeluid zit, en (2) de informatie die een zeer slecht gehoor kan verwerken. Door nu het spraakgeluid zodanig aan te passen dat alleen de informatie die verwerkt kan worden ook daadwerkelijk gebruikt wordt (al het andere wordt weggegooid) kan een efficiëntere ondersteuning van het liplezen bereikt worden. Uiteindelijk zou deze aanpak moeten resulteren in een hoortoestel dat spraakgeluid niet alleen versterkt, maar op een 'slimme' manier ontleedt en bewerkt. Een dergelijk *spraakbewerkend hoortoestel* neemt dan als het ware een stukje verstaan van de slechthorende over.

Het in dit proefschrift beschreven onderzoek is bedoeld als een eerste fase: niet de ontwikkeling van een toestel stond centraal, maar de omschrijving van een bewerkingsmethode van spraak die mogelijk in de vorm van een hoortoestel toepasbaar is.

Deze toelichting is speciaal bedoeld voor lezers die geïnteresseerd zijn in het onderwerp, maar die weinig of in het geheel niet vertrouwd zijn met het soort onderzoek dat wordt uitgevoerd. Ze belicht voornamelijk de achtergronden van het onderzoek, zonder al te veel in te gaan op alle technische details. Het doel is om vooral de grote lijnen zichtbaar te maken: waar was het onderzoek voor bedoeld, wat heeft het opgeleverd, en wie heeft er uiteindelijk wat aan ?

Slechthorendheid: verzwakking en vervorming

Slechthorendheid uit zich in de eerste plaats in het feit dat alle geluiden zachter doordringen. Dit probleem wordt aangeduid met de term *verzwakking* (alle geluiden komen verzwakt door). Hoe erger de slechthorendheid, hoe groter het probleem om alle omringende geluiden (verkeer, radio, gesproken woord, wekker, deurbel) waar te nemen. Voor het oplossen van dit probleem kan de slechthorende gebruik maken van hulpmiddelen, van zeer eenvoudige (hand achter het oor, tv harder zetten, aan de spreker vragen om harder te praten) tot zeer gecompliceerde (signaleringen door middel van lichtflitsen in plaats van geluid, allerlei soorten moderne hoortoestellen). Wanneer verzwakking het enige probleem is dan helpt een hoortoestel meestal goed, terwijl vaak ook relatief eenvoudige operaties mogelijk zijn.

Naast verzwakking kan een bijkomend fenomeen optreden, aan te duiden met de term *vervorming*. De waargenomen geluiden kunnen vervormd klinken, "doffer" of "onduidelijk zoals wanneer de radio net verkeerd staat afgestemd". Het is een vorm van slechthorendheid die, bij grotere gehoorverliezen, meestal veel vervelender effecten dan verzwakking heeft. In die gevallen helpt versterking door harder praten of door gebruik van een hoortoestel alleen maar om het geluid weer hoorbaar te maken; het karakter van het geluid blijft echter vervormd. Dit houdt in dat de slechthorende op een gegeven moment de spreker wel hoort, maar niet verstaat - hoe hard er ook gesproken wordt. Deze vervormingscomponent kan met geavanceerde hoortoestellen slechts ten dele worden bestreden. Operatief ingrijpen is niet mogelijk.

Wanneer de vervormingsslechthorendheid - we spreken dan over *perceptieve slechthorendheid* of *perceptief gehoorverlies* - zeer ernstige vormen aan gaat nemen, dan is het voor de persoon in kwestie vaak heel moeilijk om het gesproken woord nog te kunnen verstaan. Zoals gezegd, het versterken van het geluid via een hoortoestel maakt het geluid vooral beter hoorbaar, maar niet veel beter verstaanbaar. Om het verminderde *spraakverstaan* te compenseren moet de slechthorende het gesproken woord van de lippen van de spreker aflezen. Dit *liplezen* is bepaald geen gemakkelijke klus; uit het zichtbare spraaksignaal zijn nu eenmaal veel minder 'klanken' af te leiden dan uit het hoorbare spraaksignaal (de geheim agent die in de andere hoek van de donkere en rokerige nachtclub twee verdachte lieden ziet praten over een uit te voeren staatsgreep bestaat echt alleen maar in Hollywood-films). Wanneer het gehoor ernstig is aangetast, is liplezen echter vaak de enige manier om nog iets van de boodschap mee te krijgen. In vergelijking met de normale gesproken communicatie lijkt de situatie bij grote

gehoorverliezen net andersom te zijn: normaal is horen het belangrijkste en liplezen kan soms een ondersteuning zijn ("Nee meneer, ik heet van Son, niet van Som"), maar bij een ernstig gehoorverlies is het liplezen vaak het belangrijkste en levert het geluid een ondersteuning. Dit houdt tevens in dat, wanneer liplezen onmogelijk is, in feite de gesproken communicatie ook onmogelijk wordt: telefoneren lukt dus niet, net zomin als lekker kletsen in een gezellig donker hoekje van de kroeg.

Het verstaan van spraak - hoe doen we dat ?

Blijkbaar is ons gehoorsysteem niet alleen bedoeld of geschikt om te *horen*, maar ook om te *verstaan* en uiteindelijk te *begrijpen* wat er gezegd wordt. Hoe gaat dit spraakverstaan in zijn werk ?

Globaal gesproken kan het gehoororgaan in drie delen opgesplitst worden. Het eerste deel, bestaande uit oorschelp, gehoorgang, trommelvlies en drie kleine gehoorbeentjes, vangt het geluid op en versterkt het. Het tweede deel rafelt elk geluid uiteen in een aantal afzonderlijke signalen en stuurt die door naar de hersenen. Dit deel bestaat uit het binnenoor (wegens zijn vorm ook wel slakkehuis genoemd) en de gehoorzenuw. Het derde deel is het gehoorcentrum in de hersenen zelf: hier worden de afzonderlijke signalen weer gecombineerd tot zinvolle informatie, zodat de geluiden worden geïnterpreteerd. De hersenen zorgen er dus voor dat de boodschap begrepen wordt, dat de autotoeter als een waarschuwing wordt opgevat of dat muziek als muziek in de oren klinkt (wanneer het tenminste geen *house* is).

De mensen waar het onderzoek in dit proefschrift voor bedoeld was hebben een sterk verminderde functie van het binnenoor: ze hebben daardoor problemen met het uiteenrafelen (*analyseren*) van alle binnenkomende geluiden. De meeste geluiden om ons heen zijn samenstellingen van allerlei verschillende tonen. De samenstelling is zodanig dat we de geluiden als een geheel waarnemen, en niet als afzonderlijke toontjes. Dit geldt ook voor spraakgeluiden, ofwel voor de verschillende klanken (klinkers en medeklinkers) die elke taal kent. De klanken klinken verschillend omdat ze zijn samengesteld uit verschillende tonen. Iedereen kan dit voor zichzelf makkelijk controleren door bv. afwisselend een langgerekte *ssssss* en een lang gerekte *shshshsh* uit te spreken: de *ssssss* klinkt veel hoger omdat er meer hoge tonen in de samenstelling zitten.

Een goed functionerend gehoor neemt de kleine verschillen tussen klanken waar. Hierdoor kan een goedhorende alle spraakklanken van het Nederlands onderscheiden. Wanneer iemand echter ernstig slechthorend

is, en het probleem ligt in het binnenoor, dan zijn de kleine verschillen tussen spraakklanken vaak niet meer hoorbaar. Het gevolg is dat veel spraakklanken op elkaar gaan lijken, of dat ze als het ware door elkaar gaan lopen: ze worden verweven tot een grote brij. Een goedhorende kan zich er moeilijk een voorstelling van maken hoe dat klinkt. Een medische ingreep is uitgesloten: een eenmaal aangetast binnenoor is niet te herstellen door een operatieve ingreep, noch met behulp van medicijnen. En zoals gezegd, met de bestaande hoortoestellen wordt dit probleem nauwelijks opgelost.

Een oplossing: liplezen met *minder* maar wel *betere* informatie

Spraakverstaan wordt dus bemoeilijkt omdat het gehoor de verschillen tussen veel spraakklanken niet meer aankan. Ofwel: het spraaksignaal bevat veel informatie waar de slechthorende weinig of niets mee kan doen. Het idee achter de oplossing voor het probleem is eigenlijk simpel: als het binnenoor niet zoveel informatie tegelijk aan kan, moet die hoeveelheid verminderd worden. En als het binnenoor de kleine verschillen tussen spraakklanken niet verwerkt, dan moeten die verschillen zo mogelijk maar duidelijker gemaakt worden. Dit betekent: *minder* informatie aanbieden, maar wat er wordt aangeboden moet *beter* waarneembaar zijn. Het nieuwe hoortoestel biedt de ernstig slechthorende liplezer als het ware een nieuw klanksysteem voor het Nederlands aan.

Natuurlijk is zo'n toestel geen futuristische wondermachine die het spraakverstaan ineens weer mogelijk maakt. Het dient vooral om de slechthorende beter te helpen bij het liplezen of spraakafzien, beter dan nu mogelijk is met een gewoon hoortoestel. De slechthorende zal nog steeds moeten liplezen om een gesprek te kunnen voeren, maar een aantal dingen die echt niet te onderscheiden zijn bij liplezen zou het toestel hoorbaar kunnen maken. Simpel gezegd: wat je niet kunt horen moet je liplezen, wat je niet kunt liplezen moet je kunnen horen. Twee voorbeelden zullen dat illustreren.

In woorden als 'fuiven', 'suizen' of 'gesjacher' komen de Nederlandse *ruis-* of *sisklanken* voor: *f*, *v*, *s*, *z*, *g*, *sj* of *ch*. De verschillen tussen de meeste sisklanken zijn klein, en veel slechthorenden horen dan ook geen verschillen tussen bv. 'sjaal' en 'zaal', tussen 'zet' en 'set' of tussen 'zuil' en 'vuil'. In sommige gevallen zijn de klankverschillen wel te liplezen: kijk in de spiegel voor het verschil tussen 'zuil' en 'vuil' of 'refter' en 'rechter'. De onhoorbare verschillen tussen deze klanken zijn voor de gemiddelde liplezer geen onoverkomelijk probleem. Voor de niet-zichtbare verschillen, echter, zou een bewerking van het geluid een

mogelijke oplossing zijn. Het toekomstige hoortoestel zou bv. de sis van de *sj*, *ch* of *g* om kunnen zetten in een soort zoemtoon, zodat de slechthorende dan beter het verschil hoort tussen een scherpe sisklank *s* en de zwaardere zoemklanken *sj*, *ch* en *g*. Op die manier kunnen 'shop' en 'sop', 'recht' en 'rest' of 'lachen' en 'lassen' makkelijker uit elkaar gehouden worden.

Een tweede voorbeeld. De klinkers *ie* en *uu* klinken voor veel slechthorenden gelijk; hetzelfde geldt voor *oe* en *oo*. Toch worden de woorden 'vier' en 'vuur' bijna nooit met elkaar verward, omdat ze met liplezen makkelijk te onderscheiden zijn. Anders ligt dat met 'roet' en 'rood' - het mondbeeld lijkt sterk op elkaar. Om deze woorden uit elkaar te houden hoeft de luisteraar bij 'vier' en 'vuur' alleen maar goed te kijken, maar bij 'roet' en 'rood' zou het spraakgeluid uit het toestel kunnen helpen. Wederom zou het toestel het verschil hoorbaar kunnen maken door bv. de *oe* als een zwaardere zoemtoon te laten klinken.

Gecodeerde spraak herken je niet zonder oefenen

Hoortoestellen kunnen we tegenwoordig onder de *hi-tech* apparatuur scharen. In het spraakbewerkende hoortoestel komt in feite een miniatuurcomputertje te zitten. Dat apparaatje herkent het binnenkomende spraakgeluid, gooit de storende informatie weg en zet de belangrijke informatie om in goed te onderscheiden spraakklanken. Omdat een computer het werk doet, kan de spraak in de meest vreemde klanken worden omgezet. Technisch is er geen enkele beperking. Maar de vraag is natuurlijk: kan iemand die vreemde klanken ooit begrijpen ?

In andere talen dan het Nederlands komen andere klanken voor. Het Vlaams klinkt anders, maar heeft zeer herkenbare klanken. Duits en Engels hebben al weer wat meer andere klanken, maar veel klanken lijken ook op het Nederlands. Luisteren we echter naar Russisch of Arabisch, dan horen we allerlei voor ons heel vreemde klanken. Toch leren veel Nederlanders vreemde talen, ook Russisch, Arabisch, Japans of Indonesisch. Als een spraakbewerkend hoortoestel dus een gedeeltelijk nieuw klanksysteem voor het Nederlands maakt, dan moet dat ook te leren zijn. Dit betekent dat het aanmeten van een spraakbewerkend hoortoestel wat anders is dan het aanmeten van een gewoon hoortoestel. Het is geen kwestie van een keer naar het *Audiologisch Centrum* stappen en vervolgens bij de hoortoestellenverkoper (de audicien) het hoortoestel laten aanpassen. De slechthorende die in aanmerking komt voor een spraakbewerkend hoortoestel zal moeten leren omgaan met het toestel en met het nieuwe geluid; hij of zij zal als het ware het nieuwe

klanksysteem moeten leren interpreteren. Bij het aanpassen van het hoortoestel hoort dus een oefenfase, waarin men in een aantal sessies met een hoortherapeut leert hoe het bewerkte geluid past bij het mondbeeld dat men ziet van degene waarmee men spreekt.

Liplees- en luisterproeven

Tot zover de achtergronden - even terug nu van de toekomst naar het huidige onderzoek. Tot nu toe is gesproken over het veranderen van klanken, opbouwen van een nieuw klanksysteem enzovoorts alsof het de gewoonste zaak van de wereld is. Maar het grote probleem is: hoe weten we als onderzoekers nu op welke manier we zo'n nieuw klanksysteem moeten creëren ? Het is natuurlijk onzin om zomaar wat te proberen en als het niet lukt zeggen: jammer, volgende keer beter. Op die manier kunnen we lang bezig zijn. Juist aan deze vraag - hoe vinden we uit wat we precies moeten doen ? - is het grootste deel van het onderzoek besteed.

Voor het antwoord op die vraag hebben we in het eerste deel van de vier jaar verschillende liplees- en luisterproeven gedaan, vooral bij proefpersonen met een groot gehoorverlies. We zijn daarin nagegaan welke klanken de proefpersonen wél en welke ze juist níet konden waarnemen of van elkaar konden onderscheiden. Op die manier zijn we bijvoorbeeld te weten gekomen dat wanneer mensen liplezen ze zeer goed in staat zijn om het verschil tussen getuite lippen zoals in het woord *vuur* en gespreide lippen zoals in *vier* te zien. Of het verschil tussen lippen op elkaar en lippen van elkaar: *paard* of *haard*. Sommige dingen zijn niet of nauwelijks te zien, bv. klemtoon ("Jan is ziek, niet Piet" of "Jan is ziek, niet weg"). Maar: klemtoon is wel goed te horen, zelfs al heeft iemand een groot gehoorverlies. Uit de experimenten hebben we dus kunnen afleiden wat mensen kunnen liplezen en wat niet, en wat ze nog kunnen horen en wat juist niet. In de dagelijkse communicatie is het meestal geen kwestie van óf liplezen óf luisteren; vaak is het zo dat juist uit de combinatie van liplezen en luisteren kan worden opgemaakt wat er gezegd wordt. Het combineren van klank en mondbeeld is dus van groot belang.

In het tweede deel van het onderzoek hebben we de proef op de som genomen. Met de resultaten van de eerdere proeven in de hand hebben we eerst een paar verschillende methoden van spraakbewerking ontwikkeld. Vervolgens hebben we een video-opname gemaakt van iemand die een grote lijst losse klinkers en eenvoudige zinnen voorlas. Het geluid van de videoband hebben we daarna met behulp van een

computer bewerkt. Tenslotte hebben we de videobanden met het nieuwe geluid voorgelegd aan een groep slechthorende proefpersonen, die niets anders hoefden te doen dan na te zeggen wat ze meenden te horen en zien. Als vergelijking hebben we hun ook de oorspronkelijke videoband laten zien en horen. Na afloop van dit experiment hebben we gekeken hoeveel van de klinkers en zinnen goed zijn waargenomen, zodat we een score konden berekenen. Door nu een vergelijking te trekken tussen de scores met het bewerkte geluid en scores met het gewone geluid konden we tenslotte zeggen of onze aanpak al dan niet succesvol is geweest.

En wat heeft het onderzoek nu opgeleverd ...?

En dus was de grote vraag: kunnen mensen wat met die veranderde spraak ? Is het beter of slechter dan het gewone spraakgeluid ? Zijn we erin geslaagd tot een oplossing te komen ? En zo ja, voor wie kan de ontwikkeling van belang zijn ?

Onderzoekers hebben de neiging voorzichtig te zijn; wij zeggen daarom ook: voorlopig hebben we nog niet heel goede resultaten bereikt, maar we hopen verder te kunnen gaan met het onderzoek. We hebben onze methode van geluidaanpassing vergeleken met een geluid dat via een gewoon hoortoestel wordt doorgegeven, en voor de meeste (ernstig slechthorende) proefpersonen bleef het gewone geluid toch het beste, hoewel het verschil met de nieuwe geluiden gering was. Voor een paar proefpersonen bleek het nieuwe geluid echter beter te voldoen dan het gewone geluid, maar ook hier was het verschil niet groot. Het lijkt er dus op dat we voorlopig nog niet echt een verbetering hebben gevonden.

Hoewel het leuk is om met een nieuw procédé een echte verbetering te vinden (*'te scoren'*), zijn we niet helemaal ongelukkig met de resultaten. Ten eerste hebben de proefpersonen helemaal geen training gehad met de nieuwe spraakcode, ondanks dat dat, zoals we hierboven beschreven hebben, best belangrijk kan zijn. Helaas ontbrak ons de tijd om een gedegen training uit te voeren, maar het is dus mogelijk dat met wat gerichte oefening er nog meer proefpersonen zijn die beter uit de voeten kunnen met het nieuwe spraakgeluid. Ten tweede is het voor de proefpersonen die er echt wat aan hadden wel hoopvol dat er misschien een nieuw hulpmiddel voor hun aan zit te komen. Voor ons, onderzoekers, is de vraag vooral: hoe komt het dat de ene er wel wat aan heeft en de andere niet. Ofwel: wat is nou de typische eigenschap van de succesvolle nieuwe-spraakcode-luisteraar ? Heeft het alleen met het soort gehoorverlies te maken, of heeft de manier van luisteren er ook mee van doen (een goed verstaander...) ? En zijn het vooral de goede

liplezers die er wat aan hebben, of kunnen ook de wat mindere liplezers leren om het vertrouwde mondbeeld en het nieuwe geluid te combineren ?

Er ligt nog een flink aantal vragen om te beantwoorden. De resultaten zijn goed genoeg om met het onderzoek door te gaan, maar het is nog te vroeg om te zeggen dat een spraakbewerkend hoortoestel echt helpt, ongeacht degene die het gaat gebruiken. Bovendien hebben we het nog niet eens gehad over de produktie van zo'n toekomstig hoortoestel. Een fabrikant kijkt tenslotte niet alleen naar het nut van een produkt, maar vooral ook of het verkocht kan worden: is er een markt voor en levert het nog wat op ? Of het spraakbewerkende hoortoestel er daadwerkelijk ooit van komt is voorlopig dan ook nog een heel groot vraagteken.

Curriculum vitae

De schrijver van dit proefschrift is geboren te Helmond en getogen te Tilburg. In juni 1978 behaalde hij zijn diploma Gymnasium B aan het Mill Hill College te Goirle, waarna hij zijn studie Engelse Taal- en Letterkunde aanving aan de KU te Nijmegen. Na de kandidaatsfase deed hij een hoofdvakstudie Fonetiek binnen de Vrije Studierichting Letteren. Hij behaalde in november 1985 met lof het doctoraaldiploma Fonetiek met als bijvakken Moderne Engelse Taalkunde en Toegepaste Taalkunde. Daarnaast haalde hij ook een eerstegraads lesbevoegdheid Engels. Zijn wetenschappelijke carrière begon hij in 1986 met een kort uitvoerdersschap aan het Instituut voor Fonetiek te Nijmegen waar hij werkte aan het daar in ontwikkeling zijnde regelsynthesesysteem. Van januari 1987 tot en met juni 1989 was hij aan het Instituut voor Fonetische Wetenschappen van de Universiteit van Amsterdam aangesteld als uitvoerder van een ESPRIT-projekt. Dit projekt betrof een Europese samenwerking op het gebied van toepassingsgerichte ontwikkelingen in de spraaktechnologie. De werkzaamheden werden uitgevoerd op het TNO Instituut voor Zintuigfysiologie te Soesterberg. Na een verfrissende adempauze in het zonnebadende Spanje werd de schrijver in oktober 1989 Assistent in Opleiding (AiO) aan het Laboratorium voor Experimentele Audiologie, afdeling KNO, van het Academisch Ziekenhuis te Utrecht. Hij heeft in het kader van het onderzoekprogramma "Revalidatie van doven en slechthorenden" onderzoek uitgevoerd naar de auditieve en visuele waarneming van spraakkenmerken door ernstig slechthorenden. Het onderzoek is beschreven in dit proefschrift, het verschijnen waarvan min of meer samenvalt met het einde van de AiO-aanstelling. Naast het wetenschappelijke audiologisch onderzoek heeft de schrijver vooral belangstelling voor de beleving door slechthorenden van hun auditieve handicap en het schrijvenderwijs voorlichting geven over allerlei aspecten van slechthorendheid, zowel aan slecht- als aan goedhorenden. Deze belangstelling heeft in 1991 geleid tot zijn toetreding tot de redaktie van het tijdschrift HOREN, een tweemaandelijkse uitgave van de Nederlandse vereniging voor slechthorenden (NVVS).

In de loop van de verschillende onderzoeken heeft de schrijver bijgedragen aan de volgende rapporten, proceedings en artikelen:

Boves, L., Kerkhoff, J., Loman, H. & Son, N. van (1987). Automatic text-to-speech synthesis by allophones. *IFN-proceedings 11*, 24-26.

Loman, H., Son, N. van, Boves, L. & Kerkhoff, J. (1987). Some phonetic rules of the text-to-speech conversion system and naive listeners' judgments about the realisation of synthesized CVC utterances. *IFN-proceedings 11*, 57-74.

Son, N. van & Pols, L.C.W. (1987). Evaluation of a French diphone-based synthesis system. *TNO/IZF-report C-33, IFA-report 93.*

Son, N. van, Pols, L.C.W., Sandri, S. & Salza, P.L. (1988). First quality evaluation of a diphone-based speech synthesis system for Italian. *Procedures of SPEECH '88, 7th FASE Symposion, Edinburgh 1988, Book 2*, 429-436.

Son, N. van & Pols, L.C.W. (1989). First quality evaluation of a diphone-based synthesis system for Italian. *TNO/IZF-report 15, IFA-report 105.*

Son, N. van & Pols, L.C.W. (1989). Second quality evaluation of a French diphone-based synthesis system: identification and quality ratings of consonant clusters. *TNO/IZF-report 16, IFA-report 104.*

Son, N. van & Pols, L.C.W. (1989). Final evaluation of three multipulse LPC coders: CVC intelligibility, quality assessments and speaker identification. *TNO/IZF-report 17, IFA-report 103.*

Huiskamp, T.M.I., Smoorenburg, G.F., Son, N. van & Frowein, H. (1990). Normal and videotelephony in a picture sorting task. *Rapport van het Lab. voor Exp. Audiologie, Universiteit Utrecht, en het PTT onderzoekinstituut Dr. Neherlab.*

Son, N. van, Bosman, A.J., Lamoré, P.J.J. & Smoorenburg, G.F. (1992). Auditory pattern perception in the profoundly hearing impaired. In: M.E.H. Schouten (Ed.), *The auditory processing of speech: from sounds to words*, pp. 275-282. Berlin/New York: Mouton de Gruyter.

Son, N. van, Bosman, A.J., Lamoré, P.J.J. & Smoorenburg, G.F. (1993). The perception of complex harmonic patterns by profoundly hearing-impaired listeners. *Audiology 32*, 308-327.

Son, N. van, Bosman, A.J., Lamoré, P.J.J. & Smoorenburg, G.F. (1993). Auditory pattern perception in the profoundly hearing impaired and lipreading of Dutch phonemes. *Geaccepteerd voor publicatie in Scandinavian audiology supplement 38.*

Son, N. van, Huiskamp, T.M.I., Bosman, A.J. & Smoorenburg, G.F. (1993). Viseme classifications of Dutch consonants and vowels. *Geaccepteerd voor publicatie in Journal of the acoustical society of America.*

Son, N. van & Smoorenburg, G.F. (1993). Future hearing aids for the profoundly hearing impaired using advanced digital processing techniques. *Te verschijnen in de proceedings van het 15e Danavox Symposium, gehouden te Kolding, Denemarken, van 30 maart t/m 2 april 1993.*

Son, N. van (1993). Perceptual dimensions in lipreading and supplementary acoustic cues for the profoundly hearing impaired: a critical look at the potentials of signal processing hearing aids. *Aangeboden aan Audiology.*

Son, N. van, Bosman, A.J. & Smoorenburg, G.F. (1993). Discrimination of natural and processed Dutch vowels by profoundly hearing impaired lipreaders. *In voorbereiding.*

Son, N. van & Pols, L.C.W. (1989). First quality evaluation of a diphone-based synthesis system for Italian. *TNO/IZF-report 15, IFA-report 105.*

Son, N. van & Pols, L.C.W. (1989). Second quality evaluation of a French diphone-based synthesis system: identification and quality ratings of consonant clusters. *TNO/IZF-report 16, IFA-report 106.*

Son, N. van & Pols, L.C.W. (1989). Final evaluation of three multipulse LPC codecs: CVC intelligibility, quality assessment and speaker identification. *TNO/IZF-report 17, IFA-report 103.*

Huiskamp, T.M.I., Smoorenburg, G.F., Son, N. van & Prowein, H. (1990). Normal and videotelephony in a picture sorting task. *Rapport van het Lab voor Exp. Audiologie, Universiteit Utrecht, en het PTT onderzoekinstituut Dr. Neherlab.*

Son, N. van, Bosman, A.J., Lamoré, P.J.J. & Smoorenburg, G.F. (1992). Auditory pattern perception in the profoundly hearing impaired. In: M.E.H. Schouten (Ed.), *The auditory processing of speech: from sounds to words*, pp. 275-282. Berlin/New York: Mouton de Gruyter.

Son, N. van, Bosman, A.J., Lamoré, P.J.J. & Smoorenburg, G.F. (1993). The perception of complex harmonic patterns by profoundly hearing-impaired listeners. *Audiology 32*, 308-327.

Son, N. van, Bosman, A.J., Lamoré, P.J.J. & Smoorenburg, G.F. (1993). Auditory pattern perception in the profoundly hearing impaired and lipreading of Dutch phonemes. *Geaccepteerd voor publicatie in Scandinavian audiology supplement 38.*

Son, N. van, Huiskamp, T.M.I., Bosman, A.J. & Smoorenburg, G.F. (1993). Viseme classifications of Dutch consonants and vowels. *Geaccepteerd voor publicatie in Journal of the acoustical society of America.*

Son, N. van & Smoorenburg, G.F. (1993). Future hearing aids for the profoundly hearing impaired using advanced digital processing techniques. *Te verschijnen in de proceedings van het 15e Danavox Symposium, gehouden te Kolding, Denemarken, van 30 maart t/m 2 april 1993.*

Son, N. van (1993). Perceptual dimensions in lipreading and supplementary acoustic cues for the profoundly hearing impaired: a critical look at the potentials of signal processing hearing aids. *Aangeboden aan Audiology.*

Son, N. van, Bosman, A.J. & Smoorenburg, G.F. (1993). Discrimination of natural and processed Dutch vowels by profoundly hearing impaired lipreaders. *In voorbereiding.*