PRODUCTION AND PERCEPTION OF FAST SPEECH

ESTHER JANSE

Published by
LOT
Trans 10
3512 JK Utrecht
The Netherlands

phone: +31 30 253 6006
fax: +31 30 253 6000
e-mail: lot@let.uu.nl
http://wwwlot.let.uu.nl/

Cover illustration by Inge Zwitserlood: Hockett's Easter egg analogy

Throughout the history of the study of speech, researchers have been trying to model the relation between phonetics and phonology. One famous description of both domains is the Easter egg analogy by Hockett (1955). Hockett's analogy of speech production involves a row of Easter eggs on a conveyer belt being smashed between the two rollers of a wringer. By this, Hockett implies that the units of speech are distinct and serially ordered (perhaps also invariant) before they are all smeared together in the process of speech articulation: "The flow of eggs before the wringer represents the impulses from the phoneme source; the mess that emerges from the wringer represents the output from the speech transmitter" (Hockett 1955, p.210). It is important to note that Hockett does not imply that it is impossible to recover the original eggs. Hockett remarks that the hearer examining the passing mess can "decide, on the basis of the broken and unbroken yolks, the variously spread-out albumen, and the variously coloured bits of shell, the nature of the flow of eggs which previously arrived at the wringer". In relation to the present study on fast speech, one can imagine that the faster the conveyer belt moves, the bigger the mess becomes. Hence, it becomes all the more difficult for the listener to recover the original eggs.

ISBN 90-76864-30-6
NUR 632

PRODUCTION AND PERCEPTION OF FAST SPEECH

Productie en perceptie van snelle spraak

(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. W.H. Gispen
ingevolge het besluit van het College voor Promoties
in het openbaar te verdedigen
op vrijdag 31 januari 2003
des middags te 4.15 uur

door

ESTHER JANSE

Geboren op 23 november 1973 te Grijpskerke

Promotor: Prof. Dr. S.G. Nooteboom
Copromotor: Dr. H. Quené

ACKNOWLEDGEMENTS

Writing a dissertation is never a one-(wo)man thing, and I would like to take the opportunity to thank several people here.

First, I would like to thank my promotor Sieb Nooteboom for his supervision throughout the years and for keeping an eye on the major line of research. I thank my copromotor Hugo Quené for his theoretical and practical input. Discussing my ongoing research with the two of you has been a great help.

I thank my thesis committee, Anne Cutler, Vincent van Heuven, Jan Odijk, Louis Pols, Jacques Terken, and Frank Wijnen for their constructive comments on the dissertation manuscript.

I owe much to Theo Veenker who has always been available for practical help and technical assistence. I thank my phonetics colleagues, Bert Schouten, Gerrit Bloothooft, Guus de Krom, and the phonetics PhDs, for a good atmosphere.

A number of phonetics students were involved in experiments reported in this thesis. I thank Anke Sennema, Anneke Slis, Eva Sittig, Fiona Sely, Eelco de Vries, Ritske Hermelink and Agnes Doorduin for their contribution to my research.

Frank Wijnen, Arthur Dirksen and Talitha Francke are thanked for lending their voices for recordings.

I thank Johan Wouters (OGI School of Science and Engineering, now at SVOX) for his suggestions on my study and for his attempt to produce more natural-sounding artificially time-compressed speech (cf. Chapter 5).

Tina Cambier-Langeveld, Johanneke Caspers, and Olga van Herwijnen are thanked for being great company during conferences. I thank the Utrecht institute of Linguistics OTS for paying conference trips to San Francisco, Beijing, Leipzig, Aalborg and Denver, and for their generosity towards my numerous subjects. An extra thanks to those colleagues who willingly volunteered as subjects.

Lastly, I thank my office mates and colleagues with whom I could share the ups and downs of doing linguistic research, or just talk about something else for a change: Saskia te Riele, Sergio Baauw, Brigit van der Pas, Maya van Rossum, Ellen Gerrits, Inge Zwitserlood (whose artistic skills decorate the front cover), Elma Blom, Silke Hamann, Rianneke Crielaard, Willemijn Heeren, and the other members of the phon&phon group.

Sincere thanks to all of you.

Esther Janse

CONTENTS

AUTHOR'S NOTE **11**

CHAPTER 1
INTRODUCTION **13**
1.1 Motive: the visually handicapped 14
1.2 Literature on perception of time-compressed speech 16
1.2.1 Time-compression techniques and perception studies from the 1950s to the 1970s 16
1.2.2 Time scaling with LPC and PSOLA 19
1.2.3 More studies on the perception of fast speech 20
1.3 The present study on fast speech 22
1.3.1 Robustness and ease of processing 23
1.3.2 Adaptation to fast speech rates 24
1.3.3 Higher speech rates in perception than in human speech production 25
1.3.4 Naturally produced fast speech easier to process than artificially time-compressed speech? 27
1.4 Outline of present study 31

CHAPTER 2
INTELLIGIBILITY OF TIME-COMPRESSED WORDS AND NONWORDS **35**
2.1 Introduction 36
2.2 Method 40
2.2.1 Material 40
2.2.2 Time compression 42
2.2.3 Order of presentation 44
2.2.4 Experimental design and procedure 45
2.2.5 Subjects 46
2.3 Results 46
2.4 Discussion 57
2.5 Conclusion 60

CHAPTER 3
PERCEPTION OF NATURAL AND SYNTHETIC SPEECH AFTER TIME COMPRESSION **63**
3.1 Introduction 64
3.2 Experiment 1: Intelligibility of time-compressed natural and synthetic speech 67
3.3 Experiment 2: Processing speed 72
3.4 Discussion 82
3.5 Conclusion 86

CHAPTER 4
TIMING OF NATURAL FAST SPEECH AND WORD-LEVEL INTELLIGIBILITY OF TIME-COMPRESSED SPEECH **87**
4.1 Introduction 88
4.2 Experiment 1: Fast speech timing 91
4.3 Experiment 2: Intelligibility of words and nonwords after linear or nonlinear time compression 99
4.4 Experiment 3: Three ways of time-compression 105
4.5 Experiment 4: The speech-interference technique 110
4.6 Experiment 5: A phoneme detection study 113
4.7 General discussion 118
4.8 Conclusion 122

CHAPTER 5
WORD PERCEPTION IN FAST SPEECH: COMPARING TIME-COMPRESSED SPEECH AND NATURAL FAST SPEECH **123**
5.1 Introduction 124
5.2 Two pilot tests 128
5.2.1 Experiment 1: Intelligibility sentence material Chapter 4 128
5.2.2 Experiment 2: Processing speed of natural fast vs. time-compressed speech 130
5.3 Experiment 3: Word-perception in natural fast speech and time-compressed speech 134
5.4 Experiment 4: Subjective preference test 142
5.5 Intermediate discussion 145
5.6 Experiment 5: Compressing pauses more than speech 146
5.7 General discussion 151
5.8 Conclusion 153

CHAPTER 6
GENERAL DISCUSSION AND CONCLUSION **155**
6.1 Introduction 156
6.2 Summary of main results 156
6.3 Implications for theories of speech perception and production 159
6.4 Practical conclusions 163
6.5 Suggestions for future research 165
6.6 Conclusion 170

NOTE **173**

ON MEASURING MULTIPLE ACTIVATION USING THE CROSS-MODAL SEMANTIC PRIMING PARADIGM **175**
1 Introduction 176
2 Previous cross-modal priming studies 178
3 Replication experiment I 181
4 Replication experiment II 189
5 Discussion 194
6 Conclusion 196

REFERENCES **199**

APPENDICES **213**

Appendix A Overview table of all experiments on fast speech 213
Appendix B Non-word and real word material of Chapter 2 214
Appendix C Confusion matrix of onset consonants (nonwords) Chapter 2 217
Appendix D Confusion matrix of coda consonants (nonwords) Chapter 2 218
Appendix E List of nouns used in Chapter 4 (experiment 1) 219
Appendix F Word and sentence material used in semantic priming study 220

SAMENVATTING (SUMMARY IN DUTCH) **225**

CURRICULUM VITAE **231**

AUTHOR'S NOTE

This dissertation mainly concerns the production and perception of fast speech. However, the initial PhD project proposal concerned something different, namely whether and how sentence context affects activation of multiple lexical representations during the initial stages of auditory word recognition. The experimental paradigm for this on-line word recognition study was to be cross-modal semantic priming with partial primes. After we had found that the experimental task was not suited for the type of research we had in mind, we completely changed the topic: from an on-line study into how sentence context affects speech perception to a more general study into the production and perception of fast speech. Since we have not yet been able to publish our negative findings on the experimental task of cross-modal semantic priming elsewhere, one separate methodological section on this topic is included in this dissertation as a rather lengthy appendix (pp.175-197).

1

Introduction

Abstract

This chapter provides the theoretical background for the study into fast speech. An overview of the literature on production and perception of fast speech is provided. Furthermore, the aims of the present study are motivated and an outline of the experimental questions is given.

1.1 Motive: the visually handicapped

There are many applications in which artificial time compression of speech can be useful. It is used in telephone voicemail systems to enable fast playback of voicemail messages. Time compression might also be used for fast playback of lectures or talks, or as a way of browsing long recordings. The possibility to adjust the playback rate is also a feature of many text-to-speech systems. Text-to-speech systems can be used in combination with software programs that read the content of the computer screen (screen readers) for people who either cannot read or have difficulty with reading, such as blind, visually impaired, or dyslectic people. In the text-to-speech system Fluent Dutch (version 1.6), a speech rate between 0 and 10 can be chosen, where '5' stands for the default 'normal' rate. Speed '10' is 2.5 times as fast as normal rate; and speed '0' is 1.6 times slower than normal rate. In this way, users of the text-to-speech system can flexibly adjust the speech rate to whatever speed they prefer for the content of the speech fragment. This is analogous to reading: different people read documents at different rates: people read complex technical documents at a slower rate than glossy magazines.

In 1991 a project was started in the Netherlands, set up for people with a visual handicap. This project was called the ELK project: *Electronisch Lezen van een Krant* ('Electronic Reading of a Newspaper'). The aim of the project was to provide easier access to the news for the visually handicapped. Apart from listening to radio and television, the visually handicapped already had access to spoken versions of magazines and newspapers on audiocassette. Obvious disadvantages of these spoken versions are that not all newspaper articles are covered, that the readers have no influence on this selection, and that the news has become outdated before these spoken texts have been recorded and distributed. The ELK project provided the visually handicapped with a full digital daily newspaper (the Dutch newspaper *Trouw)*, which could be made audible by means of a text-to-speech synthesis system. Furthermore, a search system was developed to enable the listeners to browse through the newspaper.

One of the speech synthesisers which were used and evaluated by the visually impaired subjects was the Apollo system. This is a hardware speech synthesiser, consisting of a small box containing an amplifier and a loudspeaker. This box is attached to the serial port of a computer. The system is used mainly in combination with screen readers for the visually handicapped. The building blocks of the Apollo system are allophones: speech segments the size of individual sounds. These allophones have not been cut out of recordings of human speech, but are machine-made. For each allophone, a number of acoustic parameters are determined via table look-up.

Depending on the phonetic environment, these parameter values are modified via adaptation rules. At the higher sentence-level, an intonation contour and an accent pattern are computed and added. This enriched representation of parameter values is then fed to a formant synthesis system and converted to speech. Apollo is a 'flexible' system in the sense that the listener can choose between different pitch ranges, voice qualities and playback rates.

The speech quality and segmental intelligibility of the Apollo system were, however, actually quite poor. Phoneme identification percentages were between 50 and 88% correct in CVC words and nonwords, depending on the amount of training subjects had had with the system. (Jongenburger & van Bezooijen 1992). The overall quality of the system was rated 6.9 out of 10 by the users. Despite this relatively poor intelligibility, some subjects indicated that they preferred a playback rate that was faster than normal for the newspaper texts. This is not discussed in the evaluation report, but is documented on the following website (in Dutch): http://www.let.uu.nl/~Hugo.Quene/personal/demos/apollo.html#TNS1. This website has some downloadable Apollo audio demonstrations. Although the actual fast rates reported on the website are not extremely fast (rate 6 equals 8.6 syllables/second, or 263 words/minute), the audio demonstrations show how robust speech perception is: intelligibility is poor because of the primitive speech synthesis, and yet listeners prefer a faster playback rate. According to the information on this website, the visually impaired Mr Horsten remarked that speech presented at this fastest rate requires all his attention, but is still intelligible to him.

Zemlin, Daniloff & Shriner (1968) showed that comprehension of time-compressed speech is relatively unaffected by compression up to twice the normal rate, but that the difficulty of listening increases markedly. Zemlin et al. (1968) quote a study by Foulke (1966) that for visually impaired listeners 35-40% time compression (i.e., playback is 1.35-1.4 as fast as normal) is the preferred rate for listening to speech. Zemlin et al. (1968) argue that visually impaired subjects will trade increased difficulty or effort in listening to speech for increased information rate and time savings.

It makes sense that visually handicapped people are more willing to put extra effort into listening. They are highly motivated to have fast access to new information. Secondly, it is interesting to see that they adapt so quickly to difficult listening situations. In the ELK evaluation report, intelligibility of CVC words was measured of lists of 50 CVC words: each list had both real words and nonsense words. This was done at three points in time: right at the beginning of the project (t1), after about one month (t2), and after about two months of experience with the electronic newspaper (t3). From t1 to t2, there is a significant improvement in intelligibility, but there is no further improvement between t2 and t3. This learning process arguably consists of

learning the systematicity of the Apollo rule-based synthesis. Once listeners have captured the rules of the synthesis system, their performance does not increase any further. It seems reasonable to hypothesise that this adaptation is a general property of speech perception, rather than being a property of visually impaired listeners. Evidence for this hypothesis will be provided in the next section.

The findings of the ELK project have led to the present study on the perception of fast speech. First, there is the interesting finding that intelligibility of time-compressed speech is relatively well preserved up to a very fast playback rate. Secondly, the robustness of speech perception is emphasised by the fact that listeners prefer higher playback rates, even if the speech they listen to is synthetic and of a relatively poor quality. Thirdly, the process of adaptation, both to the synthetic speech as such, and to the faster speech rates, is remarkably quick.

In the next section a literature survey is provided to give a general overview of studies on the perception of fast speech. In section 1.3 the research questions for the present study are laid out.

1.2 Literature on perception of time-compressed speech

1.2.1 Time-compression techniques and perception studies from the 1950s to the 1970s

The simplest method of producing time-compressed speech is to increase the speed of the playback device above the speed at which the speech was originally recorded. This is called the *speed changing method* in an overview article on speech rate studies by Foulke (1971). The speed changing procedure not only increases the rate of speech, but increases all frequencies as well. This is the effect of playing an LP record at 45 revolutions per minute (the speed required for 'singles'). If playback speed is doubled, all frequencies are doubled as well. The American cartoon series *The Chipmunks* made use of this speed changing method. Since all frequencies are made higher, listeners get the impression that the high-pitched fast speech comes from small creatures with small vocal tracts. Changing speech playback rate in this way is technically very easy.

The technique of time-compressing speech while retaining the original pitch derives from investigations into the intelligibility of interrupted speech (Miller & Licklider 1950). Because speech in which segments had been deleted at regular intervals still remained intelligible, they argued that the speech signal was redundant. Miller & Licklider (1950) found that listeners had no problem comprehending the speech as long as the interruptions occurred at a frequency of 10 times per second or more. The

intelligibility of monosyllabic words did not drop below 90% until 50% of the signal had been discarded. Inspired by these results, Garvey (1953) argued that if almost half of the signal could be discarded, and if the remaining samples could be connected, the result would be time-compressed intelligible speech with the original pitch. Garvey (1953) proved his point by way of tape-splicing. This *sampling* method was further developed by Fairbanks, Everitt & Jaeger (1954) who made electronic speech compressors. Scott (1965) used a computer to compress speech by the sampling method. The computer first segments the speech and then samples it according to a rule for which it has been programmed. The durations of the discarded and retained segments can be varied.

Fairbanks & Kodman (1957) varied the size of the discard interval while keeping the time-compression ratio constant. The ultimate time-compression ratio (R_c) is a result of the size of the discard interval (I_d) and the sampling frequency (f_s): $R_c = I_d f_s$. Fairbanks & Kodman (1957) found that, at equal time-compression ratios, enlarging the discard interval makes intelligibility worse. This makes sense because the discarded portions are then distributed less evenly across the signal.

Several studies have shown that there is a nonlinear relation between speech rate or word rate and comprehension. Zemlin, Daniloff & Shriner (1968) showed that comprehension of time-compressed speech is relatively unaffected by compression up to twice the normal rate. Fairbanks, Guttman & Miron (1957) found that doubling the speech rate only reduced the comprehension score to 90% of that of subjects who had heard the original uncompressed version. This led them to the idea of repeated exposure. If a fragment (in their study, technical instructions to Air Force trainees), time-compressed to 50% of the original duration, can be presented twice in the same time as the original fragment's duration, comprehension may be higher than after single presentation at the original rate. Their results indeed showed a small comprehension advantage of double presentation at double speed over single presentation at the original speed.

The two methods of time compression, the speed-changing and the time-compression method, turned out to differ in intelligibility. Garvey (1953) found that time-compressed words (using the sampling method via tape splicing) were more intelligible than speeded words. The pitch and spectral distortion involved in the speed-changing method apparently interferes with intelligibility. The same was found by De Haan (1977; 1982) who used the electromechanical sampling method: time-compressed speech scored higher than speeded speech in terms of intelligibility and comprehension. In Foulke (1966), however, no difference was found between the sampling method and speed-changing method. In this study, listeners were visually impaired school children,

who were accustomed to listening to speeded speech. Foulke concluded that the superiority of the sampling method is slight, and may be erased by experience.

Garvey (1953) also compared his sampling method of time compression with the periodic interruption results of Miller & Licklider (1950). The retained samples in Garvey's method were connected, whereas Miller & Licklider's samples were not: the retained samples were interrupted but not connected. Garvey (1953) found that there was no difference in intelligibility between time-compressed and interrupted words when 50% of each word was discarded. When more of the word was discarded, the interrupted words were more intelligible than time-compressed words. Since this cannot be due to differences in speech information, it must be related to the speeded rate of occurrence of speech sounds (Foulke 1971).

In the conclusion of Foulke's overview article on the perception of time-compressed speech (1971), he remarks that until then only the effects of unselectively compressed speech had been examined. He mentions the need for experiments in which listeners are tested for the comprehension of selections in which words, phrases, sentences, or other syntactical units have been highly compressed, while the intervals between these units have been left intact in order to make processing time available. This expectation is based on the assumption that at very fast rates, the incoming items are added to short term memory at a faster rate than they can be processed. As a result, there will not only be confusion about the order in which the words are processed, but some words will actually fall out of the crowded memory. As long as listeners are given enough processing time between stretches of highly time-compressed speech, complete processing of speech is possible. Foulke relates this to the concept of buffer capacity (Miller 1956) in which the listener has a finite capacity for handling information. This capacity is defined by the rate at which the short-term representations of stimulus events can be processed for long-term storage. In fact, according to this view, the problem is not so much that there is a limit on the identifiability of the time-compressed representation, but problems arise mainly because of the lack of processing time. This is an interesting thought, but it seems quite improbable that the representation of the stimulus itself should not have anything to do with ultimate intelligibility or comprehension.

Beasley & Maki (1976) present studies in which elderly listeners with normal hearing were presented with time-compressed speech. As age increased, the effect of time compression upon intelligibility increased. This is again attributed to the idea that channel capacity for information transfer was exceeded for the elderly people, but not for the younger people. Later studies have related this age effect to temporal acuity: with increasing age, temporal acuity, or resolution, decreases (Konkle, Beasley & Bess 1977; Versfeld & Dreschler 2002). Both at the hearing, or perception level, and at a

higher cognitive processing level, elderly people will have more problems with time-compressed speech than younger people.

Another suggestion for further research made in Foulke (1971) concerns the extent to which the perception of spoken language is influenced by its temporal organisation. He raises the hypothesis that the temporal organisation of spoken language is relatively unimportant at a normal rate, but that the temporal organisation becomes more critical to comprehension the more the speech rate is increased. By varying the temporal organisation of time-compressed speech, experimental results might suggest what temporal organisation is actually most efficient for comprehension.

1.2.2 Time scaling with LPC and PSOLA

LPC (Linear Predictive Coding) can be used for the analysis, coding and re-synthesis of speech (Markel & Gray 1976). Storing fragments of speech requires a certain amount of memory. This memory load can be reduced by coding the speech signal: each speech fragment is analysed in a number of coefficients. Storing these coefficients is more economic than storing the original signal. The coefficients of each fragment can be predicted on the basis of previous fragments. LPC analysis computes the values of coefficients, such that the sum of the prediction errors, which is the difference between the predicted and the actual value, is minimal over a certain time interval. The analysis window is generally set to a duration of between 10 and 25 ms. The analysis results reflect values during short time intervals (or frames) during which the signal is assumed to be more or less stationary. Each 5 or 10 ms ('step size'), a new analysis is done for the 10-25 ms analysis window. If LPC is used with 10 coefficients, these coefficients describe the first five formants plus their bandwidths. These 10 coefficient values, plus the voiced/unvoiced value and the fundamental frequency, can be used to resynthesise the speech signal. A pulse signal is fed through a filter defined by the 10 LPC coefficients. The result is then a resynthesised version of the original speech, which is of lower quality than the original because much of the signal's detail is lost in reducing the signal to 10 values. The temporal characteristics of speech can be manipulated by updating the parameters of the synthesiser at a rate different from the rate of extraction at the analyzer.

Nowadays, PSOLA (Pitch-Synchronous OverLap Add) is the most widely used technique that can be used for time-scale modification of speech. Unlike LPC, the signal does not have to be parametrised for PSOLA manipulation. The high output quality and low system complexity make PSOLA an attractive technique. As a rule of thumb, PSOLA modified speech is high intelligible and of high quality if it involves modifications of up to a factor of two (Kortekaas 1997). The pitch or duration

manipulations that can be obtained with PSOLA are applied directly to the signal itself (Charpentier & Stella 1986; Moulines & Charpentier 1990). First, a pitch detection algorithm places markers at each consecutive pitch period. Unvoiced portions of speech are simply labelled into chunks equal to the size of the mean pitch period. At the pitch-mark locations, the signal is decomposed into separate but overlapping windows by means of Hanning windowing (Moulines & Charpentier 1990). The length of the window usually is twice the fundamental frequency with vanishing smoothing functions on either side. The maxima of the windows coincide with the pitch markers. When a higher pitch is required, the windows are made to overlap more. When a lower pitch is required, the windows are moved apart. For time-domain manipulations, the simplest option would be to cancel or duplicate some of the pitch periods. With PSOLA, pitch periods are deleted from the signal, as many as necessary to realise the shorter duration. In constructing the new waveform, the speech signal of the descending ramp is added to that of the next ascending ramp. The result is then a signal with fewer pitch periods than the original. Since the information is averaged across now overlapping pitch period windows and not simply deleted, the signal retains many of the brief acoustic events like release bursts that are important to phonetic perception. However, when speech is time-compressed to less than half of its original duration, inevitably, neighbouring pitch periods are removed. This means that very short events, such as these release bursts, may be removed entirely from the signal. Thus, time compression deteriorates the segmental quality of the speech signal. More details about PSOLA are provided in section 2.2.2.

1.2.3 More studies on the perception of fast speech

Section 1.2.1 showed that the perception of heavily time-compressed speech received much attention from the 1950s to the 1970s, but then there seems to be a gap. It was not until the 1990s that the subject is on the agenda again. A line of research which did receive attention from the 1970s up to the present is the effect of speech on phoneme perception. A number of studies have shown that speech rate affects the acoustic information specifying phonetic segments (Gay 1978; Gottfried, Miller & Payton 1990; Lindblom 1963; Miller, Green & Reeves 1986; Nooteboom 1981; Summerfield 1975). In turn, these rate-dependent modifications have perceptual consequences. Speaking rate affects the perception of long vs. short vowels (Ainsworth 1974; Gottfried et al. 1990), and the perception of Voice Onset Time (Summerfield 1975; Wayland, Miller & Volaitis 1994). Miller, O'Rourke and Volaitis (1997) show that the duration of the initial transition (distinguishing /b/ from /w/) is judged depending on the rate of the utterance. Thus, with varying speech rate, listeners change the precise mapping between

the acoustic signal and phonetic categories. In most of these rate studies, perceptual rate normalisation operates in a forward fashion: the rate of a precursor phrase affects the categorisation of a following segment. In a study which was set up to investigate the relative contribution of preceding and following context material to the perceptual normalisation of speech rate, Nooteboom (1979) found that there was no significant effect of speech rate from the preceding context material. There was, however, an effect of speech rate from the following context material (i.e., backward normalisation). Nooteboom argues that perceptual normalisation mainly operates where it is needed, i.e., in cases with ambiguous stimuli. If the stimulus is ambiguous, the perceptual decision on such a segment is delayed. This delay allows the decision process time to use all the information available to resolve the ambiguity. Acoustic information from the speech fragment following the ambiguity is still coming in, and the speech rate of this fragment may then influence the categorisation of the earlier ambiguous segment. Kidd (1989), on the other hand, not only confirmed that forward rate normalisation does indeed occur, but also showed that the effect of speech rate on phoneme identification is not restricted to a phoneme's immediate articulatory context. Kidd argues that patterns of rate changes lead listeners to build up certain expectancies. His results demonstrate that the effect of a change in articulatory rate is not simply a function of the amount of rate-altered speech or its distance from a target syllable. When the pattern of rate changes in a precursor phrase is manipulated, the rate of speech that precedes a target syllable by more than three syllables can have a greater effect on phoneme perception than the articulatory rate in the immediate context. Such global-rate effects support theories of speech perception in which timing is an independent parameter for the control of production and perception (extrinsic timing models), rather than intrinsic timing models in which timing is an integral part of the specification of an articulatory gesture. In an intrinsic timing model, rate should not affect perception beyond the local context of a particular articulatory gesture (Fowler 1980).

In the 1990s, research into the perception of heavily time-compressed speech emerged again. Studies by Pallier, Sebastian-Gallés, Dupoux, Christophe & Mehler (1998) and Dupoux & Green (1997) illustrate how listeners can adapt to speech which is presented at twice or three times the original rate. These two studies are both concerned with the adaptation process, and, in particular, the level of processing at which adaptation occurs. Improvement in performance can be assumed to occur at different levels, ranging from rather low-level adaptation in the processing of acoustic properties to ad-hoc higher-level strategies in the integration of information. Adjustment to highly time-compressed speech was found to occur over a number of sentences, where the time that was needed to adjust depended on the compression rate

(Dupoux & Green 1997). The adjustment process was not influenced by changes in either talker or compression rate. When intervening uncompressed sentences were presented, adjustment to time-compressed speech did not return to baseline performance. Dupoux & Green (1997) speculate that adjustment to time-compressed speech may be the result of two processes operating simultaneously: short-term adjustment to local speech rate parameters, and longer-term adjustment which reflects a more permanent perceptual learning process. The first, short-term adjustment, is related to local rate normalisation in phonetic processing. This normalisation is investigated in the rate effect studies by, amongst others, Miller and colleagues (Miller et al. 1997). The long-term adjustment process is said to operate on a level of representation abstract enough that the acoustic differences between talkers no longer matter.

A more refined answer with respect to the locus of the adjustment is given by Pallier et al. (1998). They found that adaptation to time-compressed speech in one language carries over to another language, but only when the languages are phonologically, i.e., rhythmically, related. Their experiments also show that understanding of the time-compressed material is not necessary for adaptation to occur: monolingual Spanish subjects, adapted with Catalan sentences which were totally incomprehensible to them, performed better on Spanish time-compressed speech than control subjects who had not been adapted. A later study by Sebastián-Gallés, Dupoux, Costa & Mehler (2000) showed that lexical information is not a determining factor in adaptation in cross-linguistic speech processing: Spanish monolingual subjects who were adapted with Greek, which is rhythmically, but not lexically related to Spanish ('lexically related' meaning with respect to lexicon, morphological system and syntax) also showed transfer of adaptation when they were then presented with Spanish time-compressed sentences. Altmann & Young (1993) did not observe a transfer from French to English, or the reverse, for monolinguals of either language. This led Pallier et al. (1998) to the conclusion that adaptation does not rely on acoustic properties, but rather on linguistic mechanisms that map the acoustic information onto lexical representations.

1.3 The present study on fast speech

In the previous sections an overview was given of studies concerning the perception of fast speech. Some of the questions addressed by those studies are relevant to the present research. In the present study the overall question will be addressed of how the

perception of artificially time-compressed speech compares to the perception of naturally produced fast speech.

This overall question is translated into a number of sub-issues. These are introduced below.

1.3.1 Robustness and ease of processing

Section 1.2 on the intelligibility of strongly time-compressed speech illustrated the robustness of speech perception: much of the speech signal is actually redundant and can be missed. Time-compressed speech remains intelligible up to almost three times the original rate, in particular when subjects have been given some time to adapt to the extremely fast rate. However, the faster rate is at the expense of the ease of processing: even though speech time-compressed to twice the original rate may be perfectly intelligible, listeners have to put more effort into the perception process. This indicates that the redundancy of speech is helpful for the listener: it makes speech processing easier and it makes it more robust against distortions from, e.g., interfering noise.

In the present study, the robustness of speech against time compression will be investigated in Chapter 2. We have seen that Foulke (1971) relates listeners' inability to cope with extremely strongly time-compressed speech to the concept of storage capacity (Miller 1956): the listener has a finite capacity for handling information. This capacity is defined by the rate at which the short-term representations of stimulus events can be processed for long-term storage. As long as there is processing time available in between stretches of highly time-compressed speech, complete processing of speech is possible. It seems illogical that the segmental intelligibility of the speech itself should remain untouched. If single strongly time-compressed monosyllabic words are presented at a rate of one word per 5 seconds, short-term memory will not be overcrowded. Yet, this does not necessarily mean that listeners will be able to identify the strongly time-compressed words. Segments may become so short that they exceed the limits imposed by the temporal resolution of the hearing system. So, even though an information handling limit may certainly play a role in the processing of longer stretches of strongly time-compressed speech, we predict that the robustness against time-scale distortions of the speech signal also depends on the segmental make-up. Some segments will resist time compression better than others, depending on the length of their steady-state parts. It is reasonable to assume that segments with a longer steady-state part (such as vowels and fricatives) will resist time compression better than those with a shorter or no steady-state part (such as plosives). When the identification of a speech segment relies on a rather rapid spectral change, time compression will hinder identification: the change then becomes so rapid that the limits of temporal

resolution may be exceeded. Normal-rate transitions have lower rates of frequency change than time-compressed transitions. Discrimination studies have shown that just noticeable differences in endpoint frequency decrease (i.e., auditory sensitivity increases) with increasing transition duration. Thus, sensitivity to changes in the size a frequency transition is higher for longer stimuli than for short stimuli, due to an increase in processing time (van Wieringen 1995; van Wieringen & Pols 1995). This means that the shorter the transition, the more difficult it is to detect a change in frequency. Apart from this, discrimination can also be based on bandwidth cues. The spectrum of a signal is a function of duration. As the duration of the transition becomes smaller, the signal bandwidth increases. This can impose a limit on frequency discrimination in short or time-compressed stimuli (van Wieringen 1995).

Apart from investigating differences between phonemes in segmental intelligibility, we will also study the role of lexicality. In Chapter 2, intelligibility of words and nonwords will be investigated at normal speech rate, and in two time-compressed conditions. In this way, it can be established how segmental intelligibility and lexical redundancy both contribute to the intelligibility of words. By disentangling these two factors, we hope to shed more light on the mechanisms underlying the robustness of the speech perception mechanism. Lexical redundancy in real words can be helpful in filling in the difficult segments. One can expect that lexical redundancy becomes more helpful when the speech signal is more degraded. Thus, with higher rates of time compression, the difference in intelligibility between real words and nonwords will increase.

Thirdly, we have seen that even though time-compressed speech may be perfectly intelligible, it is more difficult to process than normal-rate speech. Listening to fast speech requires more attention, and thus the faster rate is at the expense of the ease of processing. This issue will be addressed in Chapter 3. In Chapter 3, processing speed of normal-rate speech is compared with that of fast-rate speech. Phoneme detection time is used as a measure of processing speed. The hypothesis is that processing of normal-rate speech is easier than that of time-compressed speech, and thus, phoneme detection times are expected to be faster when listeners are presented with normal-rate speech than when they are presented with time-compressed speech.

1.3.2 Adaptation to fast speech rates

Several studies were mentioned that emphasised the speed of the adaptation process: in Dupoux & Green (1997) adjustment to highly time-compressed speech was found to occur over only a small number of sentences. It is certainly not the case that listeners

have to be subjected to intensive training, or have to be as eager and motivated as the visually impaired mentioned in section 1.1.

The Apollo findings reported in section 1.1 show how listeners can actually get used to the ins and outs of a primitive synthesis system. Schwab, Nusbaum & Pisoni (1985) suggest that several sentences are required to adjust to synthetically generated speech. The amount of training necessary depends on the quality of the synthesis system. It is assumed here that adaptation is a quick process of tuning in.

In the present study, we will not be concerned with the question at which level of processing adaptation takes place. This question is addressed by Dupoux & Green (1997) and Pallier et al.(1998). We adopt their claim that adaptation takes place at some pre-lexical level. It is irrelevant for our purposes, however, whether the exact locus of adaptation is the phonological level or a lower, acoustic/phonetic, level of processing. If adaptation takes place at a pre-lexical level, it should also occur when subjects are presented with phonotactically legal nonwords. This is investigated in Chapter 2. The expectation that will be addressed in Chapter 2 is that adaptation to time-compressed speech is relatively fast: within the duration of the test a significant improvement in intelligibility is expected, both for real words and for nonwords.

1.3.3 Higher speech rates in perception than in human speech production

An important observation is that the very fast speech rates that listeners can adapt to are much higher than the fastest speech rate they can produce themselves. Listeners can understand speech which is artificially time-compressed to two to three times the original rate, but the maximum speech rate that speakers can attain is lower than that.

Goldman-Eisler (1968) investigated the influence of the actual speed of articulation on overall speech rate. She found that what seemed to be a variation in the speed of talking turned out to be variation in the amount of pausing. Whereas speech rate was variable across speakers and across different speech production tasks, articulation rates were relatively stable. Thus, the first thing that speakers do when they speak faster than normally, is to reduce the amount and duration of pauses. Of course, speakers can speed up their articulation rate. Greisbach (1992) argues that there is a maximal speed in articulation: if one goes on trying to read faster and faster, there is a point at which articulation breaks down; one has to stop and start again. This means that there is a maximal speed of reading aloud, both absolute and speaker dependent. The mean maximum speed reached by the fastest speakers in the Greisbach (1992) study is 9 to 11 syllables per second. Speakers were able to reduce the durations of the fragments they had to read to nearly 50% of the normal-rate durations (this includes pause durations).

Note that speech rate may thus be doubled, but even though speakers try very hard, they will probably not be able to double their articulation rate (without pauses).

This asymmetry is caused by restrictions on speech production. These restrictions may be at the lowest physiological level, the motor command level, or the higher speech planning level, or indeed at all three levels (and more intermediate ones) at the same time. There is a maximum rate of raising and lowering the jaw, e.g., for the production of nonsense strings such as 'mamamama'. The heavier articulators, such as the jaw, are relatively slow: slower than, e.g., the tongue tip. Early work by Miller (1951), who asked subjects to repeat simple syllables (such as 'tat tat tat') as fast as they could, also found that articulatory movements involving the tip of the tongue could be produced faster than those involving the back of the tongue. This illustrates how there may be restrictions at the lowest physical level (Kiritani 1977; McClean 2000; Perkell 1997). Perception is limited to the finite capacities of temporal resolution and, at a higher cognitive level, rate of information procesing.

The Motor theory of speech perception (Liberman, Cooper, Shankweiler & Studdert-Kennedy 1967), as revised in Liberman & Mattingly (1985), claims that "the objects of speech perception are the intended phonetic gestures of the speaker". Perception of these gestures is claimed to occur in a specialised mode, which is different from the auditory mode. The auditory mode is available for use in e.g., discrimination. Now, is this asymmetry between maximum rate of production and perception in contrast with the claims of the Motor theory of speech perception? Furthermore, does the claim that "to perceive an utterance, then, is to perceive a specific pattern of intended gestures" lead to the expectation that naturally produced patterns of gestures will be easier to perceive than a time-compressed pattern of gestures? This time-compressed pattern cannot possibly be a pattern of intended gestures produced by a human speaker since human speakers are not capable of such rates. Nevertheless, in the Motor theory framework, it is acknowledged that unnatural types of speech do not have to be problematic per se. For the perception of synthetic speech, Liberman & Mattingly (1985) claim that synthetic speech will be treated as speech if it contains sufficiently coherent phonetic information. In their view, "it makes no difference that the listener knows, or can determine on auditory grounds, that the stimulus was not humanly produced; because linguistic perception is informationally encapsulated and mandatory, he will hear synthetic speech as speech" (p.28). Consequently, the fact that people can listen to speech which is time-compressed to much faster rates than can be produced by human speakers does not provide a strong argument against the Motor theory. Even though listeners may need an intermediate transformation or internal time-scaling step, time-compressed speech is still sufficiently phonetically coherent to be perceived as speech.

Ohala (1996) argues against the claims of the Motor theory by stating that "speech perception is hearing sounds, not tongues". Ohala backs up his argument with phonological data, ranging from obstruent production to vowel inventories and sound change. He claims that in any signaling system, the underlying units should be physically as different as possible for the purpose of maximum contrast. If the underlying units are articulatory or gestural events, one would expect speech sound inventories to have this differentiability between articulations. Yet, it seems that sounds, rather than gestures or articulations are the domain where this maximum differentiation is found. Furthermore, Ohala argues that nonhuman species (such as chinchillas and macaques) have been shown to be capable of differentiating speech sounds. It is unlikely that these animals recover the underlying vocal tract gestures. A further argument against the Motor theory's assumption is that in first and second language acquisition, the ability to differentiate sounds auditorily usually precedes the ability to produce these contrast.

A problematic aspect of the Motor theory is that it is difficult to infer testable predictions from it. It is unclear whether one might infer from the Motor theory that naturally produced fast speech would be easier to perceive than artificially time-compressed speech. This question is worked out further in sub-issue 4.

1.3.4 Naturally produced fast speech easier to process than artificially time-compressed speech?

One of the main arguments in favour of the Motor theory is that it can naturally cope with assimilation and coarticulation processes, or in other words, with the fact that different sounds lead to the same phonetic percept. Liberman & Mattingly (1985) argue that variation in the acoustic pattern results from overlapping of invariant gestures, which indicates that the gestures, rather than the acoustic pattern itself, are the object of perception. As said, it is not clear whether this theory would predict that naturally produced fast speech should be easier for listeners than time-compressed speech. Both types of speech have been produced naturally, and are thus phonetically coherent enough, and listeners need only a rather simple time-scaling step in order to decode artificially time-compressed speech.

Perception studies have shown that coarticulation and assimilation can play a facilitating role in speech perception: when sounds influence each other, segments provide acoustic cues to upcoming and preceding segments. Evidence for the facilitating effect of coarticulation was found by Whalen (1991) and Martin & Bunnell (1981), who both found that listeners can use coarticulatory information in one segment to speed processing of the next. Still, even though assimilation and coarticulation are natural phenomena in connected speech, it is conceivable that

increased articulatory overlap in very fast speech hinders the perception process. It may be true that listeners expect a certain amount of slurring when they are presented with fast speech, but the fact that they can understand speech at faster rates than anyone can attain suggests that perception is not seriously impeded by fast speech not meeting this expectation. If the speaker is pressed for time, not all acoustic or articulatory targets can be reached. Because some articulatory structures are relatively slow, articulatory gestures will be smaller than usual. Acoustically, this means that formant tracks are more smooth than for normal rate speech. Auditorily, very rapidly articulated speech gives the impression of mumbled and less intelligible speech.

The model of target undershoot, as formulated in Lindblom (1963), Gay (1981), and Lindblom (1983), and revised by Moon & Lindblom (1994), concerns the 'reduced' articulation of segments, due to faster speech rates. Lindblom (1963) showed a systematic reduction in the differences among vowels, with respect to the frequencies of the first two formants, when these vowels were spoken at rapid rates. In this early version of the target undershoot model, shorter duration always implied levelling off of the formant tracks. Several findings have challenged the target undershoot model. Van Son & Pols (1990; 1992) have shown that an increase in speech tempo is not necessarily accompanied by target undershoot. The vowel formant values measured in normal rate speech did not differ significantly from those measured in fast rate speech (van Son & Pols 1990). Furthermore, there was no significant levelling off of the formant tracks (van Son & Pols 1992), except for the F1 tracks of the open vowels. The F2 'targets' are reached in fast speech: tongue movements can be executed relatively fast. The failure to reach the F1 targets, F1 reflecting degree of mouth opening, may be due to the relative slowness of jaw movements. It is important to note that the reduction in duration of the vowels in the van Son and Pols studies (1990; 1992) was 15%. Speakers can be pushed to speak faster than that. Still, it means that some people are able to speed up their speech to some extent without slurring. Further evidence that rate does not exert a systematic influence on the formant frequencies of vowel nuclei comes from a study by Miller (1981). In a revised version of the target undershoot model, Moon & Lindblom (1994) argue that formant patterns depend on three variables: the 'locus-target' difference, vowel duration, and rate of formant frequency change. The latter is defined as 'an indirect index of articulatory effort' (p.53). In this revised model, shorter durations are not necessarily accompanied by 'undershoot'. Studies which did not find a systematic 'undershoot' at faster rates may have involved a change in articulatory effort (i.e., a change in speaking style) to compensate for the shorter segment durations.

At very fast articulation rates, however, it seems that even the neatly articulating speakers have to give up on care of articulation. Greisbach (1992) investigated the intelligibility of speech read aloud at maximal speed. For his speakers, maximal speed

seems to have no influence on speaking style; speakers who generally pronounce precisely were better understood than those with generally lax pronunciation. Conversely, slow or normal rate speech can be articulated in a reduced way as well. However, when speakers are asked to speed up beyond the moderate speed-up factor of about 1.2 times the original rate, i.e., clearly beyond the 15% increase in van Son & Pols (1990), reduced articulation, and hence, reduced intelligibility, are almost inevitable. Assuming that fast articulation must result in reduced articulation, one may hypothesise that naturally produced fast speech has a lower intelligibility than artificially time-compressed speech. The very neat articulation may even be the reason why time-compressed speech remains intelligible at extremely fast rates.

The third theoretical framework relevant to the issue of whether natural fast speech is easier to perceive than artificially time-compressed speech is the Hyper- and Hypo-speech theory (henceforth H&H theory) by Lindblom (1990). The H&H theory states that much of the variability of speech stems from the ways speakers adapt their speech to what is needed by the listener to comprehend the message (Lindblom 1990). On the one hand, the speaker wants to be understood, and this output-oriented goal forces him to use hyperspeech. On the other hand, he does not want to spend too much energy on redundant parts of speech. This system-oriented, low-cost, goal allows the speaker to use hypospeech. Thus, the speaker continuously estimates how much care of articulation is minimally needed or permitted by the audience. Nooteboom & Eefting (1994) find support for the H&H model in a series of production and perception experiments. They challenge the idea of Crystal & House (1990) who argue that variation in articulation rate observed between successive phrases is solely a function of the phonological characteristics of the phrases concerned (with respect to number of phones, number of stressed syllables, etc.). Nooteboom & Eefting's results (1994) demonstrate that rate of articulation in interpausal units is dependent on context. Furthermore, deviations from the intended rate are noticed by listeners and have a negative effect on perceived naturalness. The authors conclude that contextual factors are included by the speaker in his control of articulation rate, for the sake of the model listener in his mind.

If speakers tailor their speech to the needs of the listener, they may also be expected to do this when they are asked to speak faster than normally. It is conceivable, then, that speakers will speed up more during parts of speech which they consider to be less informative, and will speed up less during the most informative parts. Remember that one of Foulke's (1971) suggestions for further research was to look into the importance of temporal organisation. Foulke (1971) raised the hypothesis that the temporal organisation of spoken language is relatively unimportant at a normal rate, but that the temporal organisation may become more critical to comprehension, the more the

speech rate is increased. Duration studies of normal and fast rate speech have indeed shown that speakers do not speed up in a linear way: some parts are reduced more than others. It has been found that consonant durations are reduced less, relatively, than vowel durations (Gay 1978; Lehiste 1970; Max & Caruso 1997). Furthermore, durations of sentence-stressed syllables are reduced less, relatively speaking, than durations of unstressed syllables (Peterson & Lehiste 1960; Port 1981). As a result, the relative difference in duration between stressed and unstressed syllables increases in faster speech, thereby making the prosodic pattern more prominent. Stressed syllables normally carry more information than unstressed syllables, so it seems that the nonlinear way of speeding up reflects a strategic and communicative principle, namely that speakers tend to preserve the parts of information in the speech stream that are most informative. Thus, on the basis of this particular interpretation of the H&H theory, one can expect that making the temporal organisation of artificially time-compressed speech more like that of natural fast speech would improve its intelligibility and ease of processing.

A number of studies have shown that prosodic patterns are a very important source of information in adverse listening conditions. When the speech signal is degraded, prosodic information is preserved better than segmental information because it is spread over larger chunks of the speech signal. External noise, damping by thick walls or degradation over a telephone line all have an effect on the spectral content of speech, but not on the timing, pitch and loudness information. A degraded speech signal may therefore cause listeners to rely more on prosodic cues than when speech quality is high. The intelligibility of deaf speech has been found to improve significantly when sentence intonation was corrected and when a more natural temporal pattern was implemented on the original utterances (Maassen & Povel 1984). Secondly, correct sentence-level phrasing is helpful in the understanding of time-compressed speech (Wingfield, Lombardi & Sokol 1984). An earlier study by Wingfield (1975) showed that intelligibility of sentences with anomalous intonation declined steeply as time compression increased, whereas the decline was much more gradual for sentences with normal intonation. Although Wingfield's method of cross-splicing leaves open the possibility that the temporal pattern, rather than the intonation pattern (or in fact both), was responsible for this effect, it is clear that the correct prosodic pattern adds extra information to the speech signal which can be exploited in difficult listening situations.

Thus, applying the temporal structure of natural fast speech to time-compressed speech is expected to yield an intelligibility or ease-of-processing advantage over linearly time-compressed speech in Dutch. This issue is addressed in Chapters 4 and 5. Conversely, the reduced segmental articulation of naturally produced fast speech is assumed to hamper intelligibility and ease of processing. We will investigate whether

perception is in fact inhibited by the increased assimilation, coarticulation and slurring that accompany natural fast speech.

The question of how important segmental intelligibility is for overall intelligibility and comprehension is also addressed in Chapter 3, in which perception of natural speech will be compared with synthetic speech: both in normal-rate and in time-compressed conditions. Although the main part of this thesis is concerned with perception of time-compressed natural speech, time compression in speech applications may mainly be used in combination with synthetic speech. It is therefore important to investigate how robust both natural and synthetic speech are against time compression. If synthetic speech is more difficult to process than natural speech, does this also imply that perception of synthetic speech is less robust against artificial time compression than natural speech? Or, conversely, is it the case that the unnatural hyperarticulation of synthetic, particularly diphone, speech may even become advantageous under difficult listening conditions?

1.4 Outline of present study

This thesis contains the description of a number of experiments to compare the perception of artificially time-compressed and naturally produced fast speech.

In Chapter 2 the intelligibility of strongly time-compressed speech is studied. This study investigates whether whether real words resist time compression better than nonwords (because of lexical redundancy), and whether some segments resist time compression better than others (because of differences concerning the presence or length of the steady-state interval),. Furthermore, the adaptation process is investigated. The following research questions are addressed:

- How much sentence material do listeners need to adapt to highly time-compressed speech? A significant improvement in intelligibility is expected within the duration of the test.
- Do segments with a long steady-state part resist time compression better than segments with a shorter or no steady-state part?
- Does lexical redundancy become more helpful the more difficult the listening situation? In other words, does the difference in intelligibility between real words and nonwords increase with higher rates of time compression?

To answer these questions, two perception experiments were run: one with non-word stimuli, and one with real word stimuli. These experiments were set up to provide a first general impression of the importance of extra-segmental factors (such as lexical redundancy), and of how time compression affects segmental intelligibility of artificially time-compressed speech.

The issue of segmental intelligibility is worked out further in Chapter 3, which focusses on the perception of natural speech vs. synthetic speech. The research questions addressed in Chapter 3 are listed below:

- Does increased playback speed (to a rate at which the speech is still perfectly intelligible) make speech perception more difficult, in terms of increased processing load?
- Does the processing advantage of natural over synthetic (i.e., diphone) speech decrease when both types of speech are time-compressed? Or, in other words, is processing of time-compressed synthetic speech helped by the greater hyperarticulation in diphone speech?

To answer these questions, a number of experiments were set up in order to compare intelligibility and ease of processing of normal-rate and time-compressed natural and synthetic speech.

In Chapter 4 prosody, or more specifically, temporal patterns, play a central role. We investigated how speakers speed up when they are asked to speak fast, and whether their way of speeding up is helpful for the listener. In naturally produced fast speech, the temporal pattern turned out to be more pronounced than at a normal speech rate. On the basis of our interpretation of the Hyper- and Hypospeech theory, the following hypotheses were tested:

- Speakers will reduce lexically unstressed syllables more, relatively, than stressed syllables.
- The durational correlate of pitch accent will become more prominent at faster speech rates because unaccented words (referring to 'given' information) are reduced more, relatively, than accented words (containing 'new' information).
- Word-level intelligibility of artificially time-compressed speech can be improved by taking into account the changes in temporal organisation going from normal to fast speech.

The first two questions were addressed in a production study. Four perception experiments were devoted to the third question. However, the perception results

showed that intelligibility, or ease of processing, of artificially linearly time-compressed speech cannot be improved by making its temporal pattern more like that of natural fast speech.

In Chapter 5 we investigated whether these results may have been due to the fact that the timing pattern of the fast speech of Chapter 4 was typical of very fast and slurred speech in which the speaker does not care about intelligibility. Therefore, a new experiment was set up to address the last question of Chapter 4 again, this way by studying a fast, but not very slurred, articulation rate. However, the hypothesis was no longer that making speech more natural with respect to the prosodic pattern would help the listener. Rather, the expectation was that the more pronounced prosodic pattern is not meant to make perception easier, but is due to restrictions on articulation. Furthermore, it seemed that the only nonlinear aspect of natural fast speech timing that should be imitated in order to improve the intelligibility of artificial time compression of speech is the stronger reduction of pauses than of speech. The following two hypotheses were tested in Chapter 5:

- Processing of fast speech is hampered by a more natural speech signal: removing either the temporal or both the temporal and the segmental characteristics of natural fast speech (as in artificially time-compressed speech) will make processing easier.
- Pause removal, combined with linear time compression, can improve intelligibility of heavily time-compressed speech over strictly linear time compression (but only when the other prosodic phrase boundary markers, such as pitch and preboundary lenghtening, are left intact).

The study is concluded in Chapter 6 with a summary of the main findings and conclusions, and a discussion of the implications of our findings in a wider perspective. Finally, a number of suggestions for future research will be discussed.

An overview table of all experiments on the production or perception of fast speech is provided as Appendix A.

2

Intelligibility of Time-Compressed Words and Nonwords

Abstract

This chapter investigates word intelligibility in artificially time-compressed speech. Listeners are presented with words and nonwords, in order to analyse the separate contributions of segmental intelligibility and lexical redundancy.

As expected, segments with a short or no steady-state part turn out to be less resistant against time compression than segments with a longer steady-state part. This shows that the segmental make-up of the words themselves determines their robustness against time compression. Secondly, the advantage of real words over nonwords increases when the speech signal is time-compressed further: whereas non-word identification is low, real words are still identified quite well. Within the duration of the experiment, plateau performance is reached, which indicates how fast listeners adapt to strongly time-compressed speech. After a few months, listeners have lost this initial adaptation. Thus, perception flexibly adjusts to the current listening conditions. However, the robustness of speech perception in strongly time-compressed conditions should mainly be attributed to non-segmental sources of information.

2.1 Introduction

Listeners can adapt to very fast rates of speech. They can quite easily learn to understand speech which is compressed to rates that are much faster than can ever be attained in natural fast speech. In the Introduction Chapter, the question was raised whether this fact provides a challenge to the Motor theory of speech perception. The central claim of the Motor theory is that "to perceive an utterance, then, is to perceive a specific pattern of intended gestures". But what then, if what listeners perceive cannot possibly be a pattern of intended gestures produced by a human speaker? For the perception of synthetic speech, Liberman & Mattingly (1985) claim that synthetic speech will be treated as speech if it contains sufficiently coherent phonetic information. In their view, "it makes no difference that the listener knows, or can determine on auditory grounds, that the stimulus was not humanly produced; because linguistic perception is informationally encapsulated and mandatory, he will hear synthetic speech as speech" (p.28). Consequently, the fact that people can listen to speech which is time-compressed to much faster rates than can be produced by human speakers is not a strong argument against the Motor theory. Time-compressed speech is still sufficiently phonetically coherent to be perceived as speech. Listeners will only have to perform a time-scaling step in order to derive the original gestures.

In normal everyday speech, speaker and listener tune in to each other. Listeners need to adapt to the speaker's voice characteristics and dialect or regional accent. On the speaker's side, speakers adapt their speech to the requirements of the communicative situation (Lindblom 1990; Nooteboom & Eefting 1994). An example of this type of co-operative behaviour is accentuation and deaccentuation. Accentuation is used by the speaker to guide the listener's attention to new and informative words in the speech stream, whereas given or more redundant information is usually deaccented. Likewise, speech rate can also be varied according to contextual redundancy. Speakers may have to speak relatively slowly and carefully when they are conveying new information, but they can use a relatively fast speech rate when they are, e.g., recapitulating what they have just said. However, this pact between speaker and listener does not hold for time-compressed speech. Now the listener is presented with a global speech rate which is much faster than the speaker intended. In this chapter we hope to give some insight into how listeners deal with these unco-operative situations.

In order to adapt to strongly time-compressed speech (two to three times the original rate), listeners need only a small amount of training (Pallier et al. 1998). When adapting to time-compressed speech, listeners are assumed to learn to make acoustic transformations on the signal in order to derive the correct speech segments and words.

Dupoux & Green (1997) speculate that the adjustment to time-compressed speech may be the result of two processes operating simultaneously: a short-term adjustment to local speech rate parameters, and a longer-term, more permanent, perceptual learning process. The studies by Pallier et al. (1998), Altmann & Young (1993), and Sebastián-Gallés, Dupoux, Costa & Mehler (2000) were set up to investigate the mechanisms that are responsible for these adaptation effects. They argued that adaptation to time-compressed speech may also involve certain phonological processes. Subjects who were trained with time-compressed sentences in a foreign language which was phonologically similar to their own native language (e.g., Spanish-speaking subjects adapted to Italian or Greek) showed an adaptation effect when they were subsequently presented with time-compressed sentences of their own language. Even though they did not understand the language they had been presented with first, they still performed better when presented with sentences of their own language in comparison to subjects who had not had any training at all. More importantly, these listeners also performed better than listeners who had been trained with a language that was phonologically distant from their own language (e.g., Spanish-speaking subjects adapted to English or Japanese). The authors argue that certain pairs of languages show transfer of adaptation, while others do not. This suggests that adaptation does not rely on raw acoustic properties, but on phonological or rhythmic properties. This is in line with the distinction between different broad language classes, as laid out by, amongst others, Abercrombie (1967). In his study, stress-timed languages (such as English, Dutch and German), which are said to exhibit nearly equal intervals between stresses or rhythmic feet, are distinguished from syllable-timed languages (such as Italian and Spanish), which display near isochrony between successive syllables (a third category are mora-timed languages, such as Japanese). More recent work on rhythmic differences between languages has refined this dichotomy between stress-timed and syllable-timed languages. Dauer (1983) observed that the rhythmic classes differ with respect to, amongst others, syllable type inventory and spectral vowel reduction. Ramus, Nespor & Mehler (1999) argue that the proportion of vocalic intervals in an utterance (%V) is the best acoustic correlate of rhythm class: stress-timed languages having, on average, a lower %V than syllable-timed languages. Low, Grabe & Nolan (2000) propose a pairwise variability index (the mean absolute difference between successive pairs of vowels, combined with a normalisation procedure for speaking rate) to capture rhythmic differences between languages or between language varieties. Importantly, a number of studies have shown that languages within such a class show particular language processing mechanisms (Cutler & Mehler 1993; Cutler, Mehler, Norris & Segui 1986; Ramus et al. 1999). Processing a language that belongs to the same class as one's native language should then be easier than processing a more distant language.

Altmann and Young (1993) showed that adaptation also occurs when listeners are trained with time-compressed (phonotactically legal) nonwords. Subjects who had been trained with time-compressed nonsense sentences (sentences in which all content words had been replaced by nonsense words) performed equally well on (meaningful) test sentences as subjects who had been trained with time-compressed meaningful sentences.

In this thesis, we will not be concerned with the exact level at which adaptation takes place. Adaptation is assumed to take place at some pre-lexical level, whether phonological or not. The fact that adaptation has been shown to be fast (Dupoux & Green 1997; Pallier et al. 1998) suggests that it is not an explicit learning procedure, but rather a quick process of tuning in.

It seems reasonable to assume that lexical redundancy plays an important role in the perception of time-compressed speech. The more degraded the segmental information is, the more one has to rely on extra non-segmental information. Most of the studies employing time-compressed speech have used meaningful sentences as test material. Thus, listeners could make use of both the segmental information and the non-segmental sources of information. The present study was set up to examine segmental intelligibility and the effect of lexical redundancy separately. By disentangling these two factors, we hope to shed more light on the mechanisms underlying the robustness of the speech perception mechanism.

Listeners' inability to cope with extremely strongly time-compressed speech has been ascribed to a limit on storage capacity by Foulke (1971). According to Foulke, complete processing of speech is possible as long as there is processing time available in between stretches of highly time-compressed speech. This would mean that the identifiability of the time-compressed representations is not so much at stake, but that problems mainly arise because of the lack of processing time and because some words actually fall out of the crowded memory. However, we assume that the segmental intelligibility of speech is also affected by time compression, regardless of whether there is enough processing time. Segments may become so short that they exceed the limits imposed by the temporal resolution of the hearing system. So, even though an information-handling limit may certainly play a role in the processing of longer stretches of strongly time-compressed speech, the robustness against time-scale distortions of the speech signal is predicted to depend on the segmental make-up. Some segments will resist time compression better than others, depending on the length of their steady-state part. Segments with a longer steady-state part (such as vowels and fricatives) are expected to resist time compression better than e.g., plosives which have no steady-state part at all (apart from the silent interval or voice bar, obviously). When the identification of a speech segment relies on a rather rapid spectral change, time

compression will hinder identification: the change then becomes so rapid that the limits of temporal resolution may be exceeded. In order to study segmental intelligibility after time compression, the intelligibility of different phoneme classes in different positions in the word was studied. Nonwords were used to avoid the effect of lexical redundancy. Studies with speech presented in noise showed that segments are mainly confused with segments from the same phoneme class: plosives are confused with plosives and nasals with nasals (Miller & Nicely 1955; Pols 1983). However, we did not expect these broad classes to surface after time compression. If the identification of a segment relies on rapid transitions, listeners may not be able to recover the broad phoneme class at all after time compression, and they might confuse the segment with any other segment that is similar in place of articulation. The expectation that segments with longer steady-state parts resist time compression better than those with shorter or no steady state leads to three sub-hypotheses. First, vowels, fricatives and sonorants were expected to resist time compression better than plosives, which have no steady-state part at all.

Secondly, if segment length plays a key role in how well segments are identified after time compression, consonants are expected to be better identified when they occur as singleton consonants than when they occur in consonant clusters. Waals (1999) showed that consonants in clusters are significantly shorter in clusters than when they occur on their own. On the other hand, there are phonotactic restrictions on which consonants can appear next to each other in clusters. If one of the members of the cluster is recognised, it is relatively easy to guess the other member. This should make it easier to identify consonants in clusters. Therefore, it is an open question which of the two effects is stronger: the durational factor or the phonotactic restrictions within clusters.

Thirdly, if duration of the segment is important in how well segments resist time compression, lexical stress might also play a role in how well segments are identified. Lexical stress not only makes the segment longer, but the greater care of articulation accompanying lexical stress also makes the transitions into or out of the vowel better audible. Segments in stressed syllables were therefore expected to be more robust against time compression than segments in unstressed syllables.

Non-segmental sources of information can be expected to become all the more important for the listener, the more degraded the speech signal is. Pisoni (1987) carried out a word identification study with natural speech and several types of synthetic speech. Word identification was studied in two contexts: first, in syntactically correct and meaningful sentences, and secondly, in syntactically correct but semantically anomalous sentences. The results showed that semantic constraints are relied on much more by listeners as the speech becomes progressively less intelligible. Likewise, the lexicality effect is expected to become more important, the more the signal is time-

compressed. This study was set up to investigate the interaction between increasing speech rate and the lexicality effect. The difference in intelligibility between words and nonwords is expected to become greater with increasing speech rate. Furthermore, the difference in intelligibility will be greater for disyllabic than for monosyllabic words. More identification errors can be made in the disyllabic nonwords, and at the same time, lexical neighbourhood density is smaller for disyllabic real words than for monosyllabic real words. In other words, identification of nonwords becomes more difficult when more segments are involved, while recognition of real words is easier the longer the word is.

The present exploratory study was set up to test the following expectations:

1. Adaptation to time-compressed speech is relatively fast: within the duration of the test (i.e., within 30 minutes) a significant improvement in subjects' performance is expected.
2. Segments with a long steady-state part resist time compression better than segments with a shorter or no steady-state part. This hypothesis is translated into three sub-hypotheses spelled out above.
3. Lexicality becomes more helpful the more difficult the listening situation. In other words, the difference in intelligibility between real words and nonwords increases with higher rates of time compression.

2.2 Method

One large perception experiment was set up, which was presented as two experiments to the listeners. The main reason for presenting it as two parts was that the amount of material was too large and subjects should not get too bored.

In the first experiment, subjects were asked to identify nonwords in carrier sentences; in the second experiment the same subjects were asked to identify real words in carrier sentences. In the following sections, the test material, the rate of time compression, and the procedure will be discussed for the two identification experiments.

2.2.1 Material

On the basis of 240 real words, 120 nonwords were formed: each non-word was based on the combination of parts of two real words. There were 60 monosyllabic nonwords

and 60 disyllabic nonwords (30 with initial stress and 30 with final stress). For the monosyllabic nonwords, the onset and vowel of a real monosyllabic word were combined with the vowel and coda of another real monosyllabic word, such as exemplified in (1). All combinations resulted in phonotactically legal nonwords.

(1) den (/dɛn/: 'pine') + speld (spɛlt/: 'pin') → delt (/dɛlt/)

For the disyllabic nonwords, the onset and vowel of a real disyllabic word were combined with the second vowel and final coda of another disyllabic real word (as exemplified in (2)), or with the first vowel, the medial consonant, and the second vowel of another disyllabic real word (as exemplified in (3)).

(2) radar (/ˈradɑr/: 'radar') + lepel (/ˈlepəl/: 'spoon') → rakel (/ˈrakəl/)
(3) sandaal (/sɑnˈdal/: 'sandal') + antiek (/ɑnˈtik/: 'antique') → santiel (/sɑnˈtil/)

In order to avoid too obvious similarity to the original real words, the nonwords were not combinations of the onset and vowel of one word with the medial consonant, second vowel and last consonant (cluster) of another word (so *sandaal* and *antiek* were not combined into *santiek).* Thus, the identification of onsets, medial consonants and codas could be related to the identification of those consonants in real words.

The 60 monosyllabic items formed 5 sets of 12 words: each set was defined by the type of onset (e.g., sonorant onset) and the type of coda (e.g., obstruent-obstruent cluster coda). The 60 disyllabic items were also grouped as 5 sets of 12 items; stress position was balanced in each set. Furthermore, the onset, medial and coda consonants within each group appeared in stressed and in unstressed position. This means that within each set, the identification of e.g., onset /p/ could be studied in stressed and in unstressed word-initial position (in nonwords *'pammek* vs. *pa'meek*, as in the real words *'passie* ('passion') vs. *pa'niek* ('panic')). All words and nonwords are listed in Appendix B.

One speaker read all the material (words and nonwords) embedded in short carrier phrases. The default carrier phrase was *Je moet … typen* ('You must….type'). The word or non-word had sentence accent. If the (non-)word ended in /t/, the carrier phrase *Je moet …. schrijven* ('You must …. write') was used. This was done to avoid degemination because all the words were to be cut out of their original phrase. If the (non-) word started with /t/, the carrier sentence *Dit als ….. typen* ('This like …. type') was spoken. If the (non)word started and ended with /t/, the carrier sentence *Dit als … schrijven* ('This like … write') was used. One version of the carrier phrase *Je moet… typen* and one version of *Je moet… schrijven* were chosen as standard carrier phrases. All words and nonwords were then cut out of their original carrier sentences and were pasted into

either of these two standard carrier sentences (only those ending in /t/ were pasted into the *Je moet… schrijven* phrase).

The standard carrier phrases had an unreleased /t/ in *moet* ('must'): cf. upper graph in Figure 2.1. However, if the word or non-word started with a voiceless fricative, a released /t/ was used because the original carrier sentence always had a released /t/ in *moet:* cf. lower graph in Figure 2.1. The waveforms of two target words embedded in carrier phrases (after the cut-and-paste operation) are displayed in Figure 2.1.

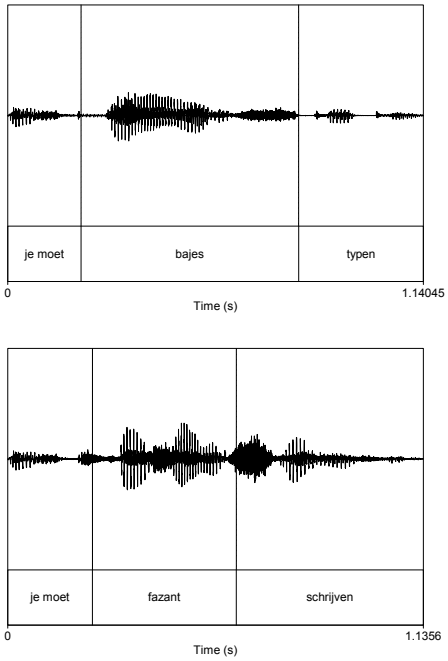

Figure 2.1. Waveform displays of *Je moet bajes typen* ('You must jail type') with unreleased /t/ in *moet* (upper graph); and *Je moet fazant schrijven* ('You must pheasant write') with released /t/ in *moet* (lower graph).

2.2.2 Time compression

All time-scale modifications were applied with the PSOLA technique (Charpentier & Stella 1986), as implemented in the speech editing program Gipos v2.3. These manipulations retained the original pitch level. The intelligibility and quality of PSOLA-modified speech is generally very high, but PSOLA can lead to audible artefacts such as

roughness, hoarseness and 'tube effects'. Even if PSOLA modification does not lead to such audible distortions, it does have an effect on the signal's spectral content (Kortekaas 1997). The reason why PSOLA is so successful is that these spectral changes are either not perceived at all by human listeners, or are phonetically irrelevant to speech perception (Klatt 1982; Kortekaas 1997). Furthermore, time-domain modifications are less likely to result in annoying artefacts than pitch modifications (Kortekaas 1997).

In a pilot experiment a degree of time compression was established in which performance for the real words would not show ceiling effects and performance for the nonwords would not show floor effects. The percentages of correct identification for real words and nonwords turned out to differ enormously and therefore two rates of compression were chosen. Correct identification for real words started to collapse at compression to 35% of the original duration. Because the correct identification of the nonwords at compression to 35% was rather low, two degrees of compression were chosen, namely compression to 40% and to 35% of the original duration. The original uncompressed items (100% of the original duration) were also tested to provide baseline identification. In this way, the decrease in intelligibility as a result of time compression could be established.

In Figure 2.2 below, the waveform displays of the original uncompressed target word *kachel* ('stove'; /kɑχəl/; upper graph) and its time-compressed version (to 35%; middle and lower graph) are shown. Because the upper and the lower waveforms are x-axis-aligned in the figure, one can see that time-compression is linear: the silent interval is affected to the same degree as the vowels (remember that, as a first step of the PSOLA process, a pitch detection algorithm places labels at each consecutive pitch period, whereby unvoiced portions of speech are simply labelled into chunks equal to the size of the mean pitch period; pitch periods are then deleted from the signal in a linear fashion, as many as necessary to realise the shorter duration). Even though in constructing the new waveform, the speech signal of the descending ramp is added to that of the next ascending ramp such that the information is averaged across now overlapping pitch period windows, a heavy compression rate inevitably means that neighbouring pitch periods are deleted from the signal. Thus, with these drastic amounts of time compression, short events may be deleted entirely from the signal.

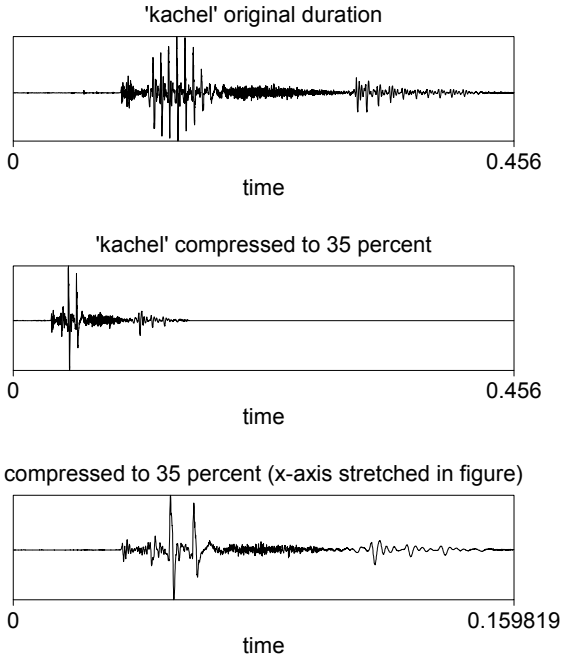

Figure 2.2. Waveform displays of original target word 'kachel' (upper graph) and PSOLA time-compressed version (compressed to 35% of original duration; middle and lower graph). In the lower graph the x-axis is stretched.

2.2.3 Order of presentation

Two orders of presentation were chosen to balance for a possible adaptation effect within the experiment. In the first experiment with nonwords, the 120 nonwords were split into three groups: 40 per degree of time compression. The items were put on a list, blocked by speech rate. Half of the subjects started with the most heavily time-compressed block, and proceeded with the less heavily time-compressed block and the uncompressed block. The other half of the subjects started with the uncompressed block, then proceeded with the compressed block and the most heavily time-compressed block. After the three blocks had been presented, the items were presented again, but now the blocks were in the reverse order.

Because two real words were used to form one non-word, experiment 2 (real words) consisted of twice as much material as experiment 1 (nonwords). The real words were also presented twice, but in order to make the amounts of material comparable, only the results of the first presentation were analysed. These real word

results are compared with the non-word results, which are collapsed over both presentations. Again, half of the subjects started with the most heavily time-compressed block, whereas the other half started with the uncompressed block and proceeded with the faster rates after that.

2.2.4 Experimental design and procedure

There were three degrees of time compression (including the uncompressed condition). There was an additional factor concerning the manipulation of the vowel durations, but the results of this manipulation will not be discussed here.[1] This latter factor, combined with the three degrees of compression yielded a 6-condition design. Because each subject can only be presented with the same item once, and because all items have to be presented in all 6 conditions, 6 experimental lists were made. The experimental conditions were balanced across the items according to a Latin-square design. This means that on each list, the 120 items were distributed over the 6 conditions. The 120 items were split in 6 groups of 20 items: the 6 experimental conditions were divided among these 6 groups. On each of the 6 lists, a particular group of items was presented in a different condition.

The items were blocked for degree of time compression. The order of blocks was either from normal to very fast or from very fast to normal. This introduced an extra experimental factor, namely Order of Presentation. The six experimental conditions, combined with the two orders of presentation yielded 12 experimental lists. Each block was preceded by a practice block of 9 items so that subjects could get used to the speech rate. In this way, a certain amount of adaptation already occurred during the practice session.

Subjects were seated in sound-treated booths and used closed earphones. They were instructed to identify the non-word (in experiment 1) or the real word (in experiment 2) embedded in the carrier phrase and to type in their response. There was no time pressure and subjects could modify their typing errors before entering their response. Subjects were told which two carrier phrases they were going to hear. During the first practice block, subjects' performance was monitored by the experimenter and after the presentation of the first block, subjects could ask questions if anything was unclear. The entire experiment was self-paced: after subjects had finished their response, they hit the Enter key to proceed to the presentation of the next item. In between the blocks, a

1 This additional factor was used to study whether manipulating the duration of the unstressed vowel could save the intelligibility of words and nonwords at very fast rates. This question was more neatly addressed in the study described in Chapter 4, and the results of this manipulation are therefore left out of this chapter.

Continue button appeared on which subjects had to click in order to proceed with the next block. In this way, subjects could choose to rest a while in between the blocks. On average, subjects took 30 minutes to finish experiment 1 (nonwords) and 50 minutes to finish experiment 2 (real words).

2.2.5 Subjects

In both experiments, 96 subjects participated: 8 subjects were randomly assigned to each of the 12 lists. For experiment 2, which took place approximately five months after experiment 1, the subjects of experiment 1 were asked to participate again. However, 14 of these 96 subjects were unable to participate in the second experiment and these were replaced by other subjects.[2] The subjects were all students of Utrecht University and received a small payment for their participation.

2.3 Results

Some subjects reported that they noticed typos only at the moment they hit the Enter key. All responses were therefore checked for obvious typing errors or different spellings (e.g., *nu;l* for *nul* or *stiekum* for *stiekem*).

It was predicted that the adaptation to time-compressed speech in experiment 1 would not carry over to performance in experiment 2. This was investigated by comparing the results of the 'new' subjects with those of the subjects who had participated in both experiments. The mean recognition score for the time-compressed real words was 70% correct for the subjects who had also participated in the first experiment; and it was 68% correct for the 'new' subjects. The difference between the recognition scores of the two groups was not significant (Mann-Whitney, $z=-0.83$, $p=0.41$). A second analysis on arcsine transformed data yielded the same insignificant result. This suggests that there may be no long-term memory component involved in adaptation to fast speech: the adaptation obtained in these experiments is rather short-lived.

The results will be discussed in three sections: each section will be headed by the hypothesis in question. For the real words, only the recognition results for the first presentation will be shown, unless clearly indicated otherwise.

2 Per list, not more than two subjects had to be replaced.

2.3.1 Hypothesis 1: Learning effect during experiment

In Dupoux & Green (1997) the learning curve reaches a plateau after subjects have been presented with 20 time-compressed sentences. If this also holds for the present experiment, then an important part of the adaptation process took place during the practice session that preceded the experimental blocks. Subjects were warned beforehand that the fastest speech would probably sound totally unintelligible to them at first, but that it would become easier after some time. After the first practice block, during which subjects' performance was monitored by the experimenter, all subjects agreed that this was indeed the case.

The present experiment was not set up in such a way that a learning curve could be established. However, half of the subjects started with the most heavily time-compressed condition, whereas for the other half the speech rate was increased throughout the first three blocks. This enables us to look at the identification scores of these two groups of subjects on the first most heavily time-compressed block of sentences. In Table 2.1 the identification scores for the nonwords in all three compression conditions are shown, broken down by the order of presentation in which the blocks were presented to the listeners, and by the first and second presentation (all nonwords were presented twice). The group of listeners who were presented with the normal rate first is called group 1; the group of listeners with the reverse order (fastest rate first) is called group 2. Each mean identification score is collapsed over three experimental lists and thus over all 120 items. Each column represents a later block in the experiment. In the bottom row, the mean identification for that particular block is shown, collapsed over the two listener groups.

Table 2.1. Identification scores (% correct) for the nonwords in all three compression conditions, broken down by two listener groups (group 1: those presented with normal rate first vs. group 2: those with fastest rate first) and by first and second presentation of the items.

	First presentation			Second presentation		
	block 1	block 2	block 3	block 4	block 5	block 6
Group 1	normal – 80	fast – 25	fastest – 19	fastest – 21	fast – 30	normal – 83
Group 2	fastest – 12	fast – 26	normal – 81	normal – 83	fast – 32	fastest – 22
Mean	46	25	50	52	31	53

Interestingly, the data show that the adaptation effect is largest for the most difficult condition: for group 2, the identification score for the fastest condition almost doubles from the first (12%) to the second presentation (22%). In line with Altmann and Young 's results (1993), the present data show that adaptation occurs when listeners are presented with time-compressed (phonotactically legal) nonwords.

The means represented in the bottom row are an indication of the adaptation during the experiment. At blocks 1, 3, 4, and 6 the mean percentages of correct identification can be computed, collapsed over the two listener groups. At each of these blocks, one of the groups is presented with the normal rate, and the other group is presented with the fastest rate. By comparing the mean identification scores at these subsequent blocks, the increase in identification scores throughout the experiment can be studied. The increase in the mean, collapsed over the two groups, is largest within the first presentation from block 1 (46%) to block 3 (50%). The subsequent increases to block 4 and block 6 are smaller. This suggests that the learning curve is indeed approaching a plateau.

In Table 2.2 the identification scores for the real words in all three compression conditions are shown, broken down by the two listener groups, and by the first and second presentation.

Table 2.2. Identification scores (% correct) for the real words in all three compression conditions, broken down by order of presentation of the blocks and by first and second presentation of the target items.

	First presentation			Second presentation		
	block 1	block 2	block 3	block 4	block 5	block 6
Group 1	normal – 99	fast – 80	fastest – 70	fastest – 75	fast – 84	normal – 99
Group 2	fastest – 64	fast – 79	normal – 99	normal – 99	fast – 82	fastest – 73
Mean	81	79	85	87	83	86

Table 2.2 shows that correct identification of the real words also increases throughout the presented blocks, but one should note that the set of real words consisted of twice as much material as the nonwords. Although the identification score of group 2 almost doubles from the first to the second presentation for the nonwords, this improvement is relatively smaller for the identification of the real words. The mean increases from block 1 to block 3 and from 3 to 4, but decreases from blocks 4 to 6. So, for the real words, the 'learning curve' approaches a plateau as well.

The correct identification percentages of real words and nonwords were arcsine transformed to avoid the risk of a data distribution that is too much skewed. These transformed data were then entered into analyses of variance to test the effects of Lexical Status, Compression Rate and Repetition (first vs. second presentation). In order to keep the amounts of stimulus material comparable, only half of the real words were compared with the nonwords (i.e., for each non-word only one of the real words that it corresponded with). With respect to all ANOVA results reported in this thesis, effects are only counted as significant when both the subject analysis and the item

analysis have $p<0.05$. The main effect of Lexical status was highly significant ($F_1(1,95)=4702$, $p<.001$; $F_2(1,119)=402$, $p<.001$), and so were the effects of Compression Rate ($F_1(2,94)=2291$, $p<.001$; $F_2(2,118)=357$, $p<.001$) and of Repetition ($F_1(1,95)=290$, $p<.001$; $F_2(1,119)=165$, $p<.001$). There was a significant interaction between the effect of Repetition and Compression Rate ($F_1(2,94)=47$, $p<.001$; $F_2(2,118)=34$, $p<.001$). This statistical analysis and Tables 1 and 2 thus show that performance increases most in the most difficult (i.e., the most heavily compressed) condition. Models of word recognition, such as Morton's logogen model (1969), argue that once a word has been recognised, its activation is reset to a somewhat higher level than competing logogens which have not fired and which are reset to their original 'rest' level. This means that once words have been recognised, they are easier to recognise a second time. The interaction between Repetition and Compression Rate indicates that the better performance in blocks 4-6 is not only due to a lower recognition threshold for already recognised items because many of the time-compressed nonwords had not been recognised the first time. Therefore, the adaptation effect, and not a lowering of the recognition threshold, must be responsible for the interaction between Compression Rate and Repetition. The three-way interaction between Lexical Status, Compression Rate and Repetition was not significant ($F_1(2,94)<1$, n.s.; $F_2(2,118)=1$, n.s.). Although the increase in performance from the first to the second presentation seems to be greater for the time-compressed nonwords than for the time-compressed real words, this difference is apparently not significant.

A separate analysis of variance on subjects was carried out for the first presentation of the nonwords to establish the effect of Order of Presentation and Compression Rate. In this analysis, subjects were nested under Order of Presentation. There was no main effect of Order of Presentation: $F_1(1, 94)<1$, n.s. However, the interaction between Order of Presentation and Compression Rate was highly significant ($F_1(2,93)=7.3$, $p=.001$). Since items together form the Order of Presentation, an item analysis could not be carried out.

The results of the present experiment show that the relative increase in performance throughout the experiment becomes smaller over time. This suggests that listeners do not need intensive training in order to adapt to heavily time-compressed speech.

2.3.2 Hypothesis 2: Segments with longer steady-state parts resist time compression better than segments with a short or no steady-state part

The hypothesis that segments with longer steady-state parts resist time compression better than segments without or with a short steady-state part is translated into three sub-hypotheses which will be explained below. In this section only data from the nonwords are discussed because the hypothesis solely concerns segmental intelligibility.

If the duration of the steady-state part is crucial to how well segments resist time compression, the first sub-hypothesis is that vowels, particularly long vowels, should be least affected by time compression. Furthermore, fricatives and sonorants will resist time compression better than plosives, which have no steady-state part at all.

In Table 2.3 the percentages of correct identification of vowels and consonants in monosyllabic nonwords are given at all three compression rates. For the consonants, the results are broken down by onset and coda position. For the onset fricatives, responses with errors in voicing value were regarded as correct.

Table 2.3. Percentages of correct identification of vowels and consonants in the monosyllabic nonwords, broken down by compression rate, and onset vs. coda position (for the consonants only).

	Onset			Coda		
	original duration	compressed to 40%	compressed to 35%	original duration	compressed to 40%	compressed to 35%
vowels	97	89	81			
sonorants	93	80	73	92	60	59
fricatives	98	92	81	97	92	87
plosives	99	49	43	99	77	67

The percentage of correct identification of vowels is relatively high. Most of the vowels in the monosyllabic words were short vowels, and the results in Table 2.3 above were therefore not broken down by vowel length. The identification percentages of long and short vowels were compared for the stressed vowel in the disyllabic nonwords, and there appeared to be a slight tendency for long vowels to be recognised better than short vowels (97% correct vs. 93% correct). However, the materials are not balanced well enough to warrant any conclusive remarks on the effect of vowel length.

Table 2.3 shows that the identification of vowels and fricatives is indeed least affected by time compression. Whereas the identification of plosives clearly suffers from time compression, the identification of the fricatives only drops to 84% correct[3].

3 Strictly speaking, the identification results of the fricatives in the heavy compression condition (to 35% of the original duration) drop to 44% correct when voicing errors are not allowed. In the

The consonant results were entered into analyses of variance to study the effects of Set (consonant class; singleton consonants only), Position in the Syllable (onset vs. coda) and Compression Rate on percentages of correct identification. The main effect of Set was significant only by subjects ($F_1(2,94)=38.7$, $p<0.001$; $F_2(2,21)=2.5$, n.s.). The effect of Position in the syllable was also only significant by subjects ($F_1(1,95)=7.2$, $p=0.009$; $F_2(1,21)<1$, n.s.), but the main effect of Compression Rate was significant in both analyses ($F_1(2,94)=208.5$, $p<0.001$; $F_2(2,20)=42.6$, $p<0.001$). The interaction between Set and Position was significant only by subjects ($F_1(2,94)=46.5$, $p<0.001$; $F_2(2,21)=2.2$, n.s.). The interactions between Set and Compression Rate ($F_1(2,94)<1$, n.s.; $F_2(4,42)=1.7$, n.s.) and between Position and Compression Rate were not significant in either analysis ($F_1(2,94)=1.4$, n.s.; $F_2(2,20)<1$, n.s.). Lastly, the three-way interaction between Set, Position and Compression Rate was significant in the subject analysis ($F_1(2,94)=15.3$, $p<0.001$; $F_2(4,42)=1.7$, n.s.).

Obviously, the small number of items makes significant results in the item analysis fairly problematic (only 6 plosives in onset position that can be compared with 6 fricatives and 12 sonorants in onset position). These statistical results only provide some weak evidence for the three-way interaction: the identification of particularly plosives in onset position is affected most by time compression. The results can therefore only be taken as providing weak support for the idea that fricatives and sonorants resist time compression better than plosives in onset position.

Confusion matrices were made to investigate the error responses to each of the singleton consonants in onset and coda position and to check whether most of the confusions are within or across category. They are presented as Appendices C (for onset position) and D (for coda position). Because of the absence of a voicing distinction for obstruents in coda position in Dutch, fricatives and plosives were expected to be easier to identify in coda than in onset position. The confusion matrix shows that these are indeed easier to recognise, but this is not due to voicing misperceptions in onset position: onset /t/ and /p/ were hardly confused with their voiced counterparts. Onset plosives were often 'overheard' (subjects reporting no onset consonant at all). Secondly, they are often confused with nasals of the same place of articulation (/p/ or /b/ becomes /m/; /t/ or /d/ becomes /n/). Obviously, the silent interval becomes too short to be noticed as such, and listeners are only left with formant transition cues related to place of articulation. A third tendency is that plosives are confused among themselves: plosives are confused with plosives having the wrong place of articulation or having a wrong voicing value. Fricatives, on the other hand, are only rarely left out, and furthermore, they are not often confused with fricatives having

---

uncompressed condition with the same restriction, only 71% of the onset fricatives are correctly identified.

the wrong place of articulation. The trend that coda-plosives are identified better after time-compression than onset-plosives agrees with the results of van Wieringen & Pols (1995) who found a higher auditory sensitivity for transition size differences in final (i.e., VC) than in initial (i.e., CV) position. This may be a recency effect: later incoming information is remembered best.

It should be noted that most of the consonant confusions (e.g., place confusion in stops, confusion between nasals) agree with those that many researchers have found since the original Miller & Nicely experiment (1955); cf. e.g., Soli & Arabie (1979) for references. Pols (1983), who investigated consonant identification and confusion in several noise and reverberation conditions, notes that there is much more confusion between voiced and voiceless fricatives in Dutch than in (American) English data. Thus, with respect to consonant confusion, it seems that only the observation that plosives are often not perceived at all, or are confused with nasals, is a typical consequence of time compression.

The second sub-hypothesis is that if segment length plays a key role in how well segments are identified after time compression, consonants should be better identified when they occur as singleton consonants than when they occur in consonant clusters. However, this effect may be outweighed by the effect of phonotactic restrictions: if one of the members of the cluster is identified, there is only a limited choice with respect to the identity of the other member. It is an open question which of the two effects is strongest: the durational effect, or the phonotactic effect.

Because plosives suffered most from time compression, the identification of plosives will be looked at when they occur as singleton segments, when they occur in a cluster with /s/ and when they occur in an /s/-plosive-liquid cluster. Plosive duration measurements (including the silent interval) were carried out on the experimental material. The percentages of correct plosive identification, and the mean plosive duration are shown in Table 2.4 below. No statistics can be provided on these durations because of the low number of observations these are based on. The identification scores are collapsed over the two time-compressed conditions (identification scores were 100% correct in the original normal-rate condition).

Table 2.4. Correct identification of plosive (collapsed over the two time-compressed conditions), broken down by C-cluster complexity.

	singleton plosive	plosive in /s/-plosive cluster	plosive in /s/-plosive-liquid cluster
/p/	48%	52%	74% (/spr,spl/)
	(/p/: 140 ms)	(/p/: 88 ms)	(/p/: 90 ms)
/t/	0%	68%	91% (/str/)
	(/t/: 100 ms)	(/t/: 64 ms)	(/t/: 84 ms)

Note, first of all, that the duration of the plosive decreases when /s/ is added to the onset. The plosive's duration is not reduced further when the cluster consists of three members. This is also in agreement with Waals (1999), who notes that "the overall duration and internal temporal structure of /s/ + obstruent is the same regardless of whether a liquid follows. That is, the addition of a liquid to a cluster of /s/ + obstruent does not result in further compression of the segments in that cluster." (p.29). Even though plosives are shorter in clusters than when they occur as singletons, they seem to be easier to identify in clusters. This means that the effect of duration is outweighed by that of coarticulation effects within the onset cluster and/or phonotactic knowledge. As far as coarticulation is concerned, one can say that the information concerning the plosive is spread over a longer time-interval when the plosive occurs in a cluster than when it occurs on its own. Still, the clusters /sp/ and /st/ also elicited /sm/ and /sn/ responses, respectively. The low numbers of items in singleton vs. cluster occurrence render statistical analyses of the results impossible.

Obviously, the data suggest that segments are easier to identify when they occur in a cluster than when they occur as singleton consonsants. This means that the duration of the segment itself is not the most important factor in whether segments are correctly identified. Phonotactic knowledge about possible consonant clusters in Dutch and/or coarticulatory traces in the other members of the cluster are helpful in recovering the plosive from the signal.

The third sub-hypothesis concerned the role of lexical stress. Lexical stress not only makes the segment longer, but the greater care of articulation also makes the transitions into or out of the vowel better audible. Consonants or consonant clusters in stressed syllables were therefore expected to be more robust against time compression than segments in unstressed syllables. In Table 2.5 below the correct identification scores are listed for word-initial and word-final consonants or consonant clusters in disyllabic nonwords. For the word-final consonant, a subset of the material was analysed (36 out of 60 disyllabic items) because only in this subset was the word-final consonant the same in stressed vs. unstressed condition (cf. Appendix B, items 73-84 and items 97 to 120).

The results are broken down by whether the syllable bears lexical stress or not (Table 2.5). The results for the two time-compressed conditions (compressed to 40% and to 35% of the original duration) were collapsed.

Table 2.5. Identification results (% correct) for first (word-initial) and last (word-final) consonants of disyllabic nonwords, broken down by stress position. In the compressed condition, identification results are collapsed over the two time-compressed conditions.

	word-initial C		word-final C	
	stressed	unstressed	stressed	unstressed
normal rate	97	91	95	88
compressed	56	44	68	50

If the greater segmental length and care of pronunciation that accompany lexical stress increase robustness against time compression, one expects the interaction between Stress Position and Compression Rate to be significant: stressed consonants or consonant clusters should resist time compression better than unstressed ones. The results were analysed statistically for both the word-initial consonant and the word-final consonant position (compressed conditions collapsed). The correct identification percentages (after arcsine transformation) were analysed for the effect of Compression Rate and Stress in repeated measures ANOVAs. For the word-initial consonant (or consonant cluster), the main effect of Stress was significant ($F_1(1,95)=65.4$, $p<0.001$; $F_2(1,58)=4.17$, $p=0.046$). This means that, overall, segments are recognised better when they are part of a stressed syllable. The effect of Compression was significant as well ($F_1(1,95)=880$, $p<0.01$; $F_2(1,58)=120$, $p<0.001$). The interaction between Stress Position and Compression Rate was not significant, however ($F_1(1,95)<1$, n.s.; $F_2(1,58)<1$, n.s.).

For the word-final consonant, the effect of Stress was not significant ($F_1(1,95)<1$, n.s.; $F_2(1,34)=2.65$, n.s.). The effect of Compression Rate was highly significant ($F_1(1,95)=607$, $p<0.001$; $F_2(1,34)=75.8$, $p<0.001$). The interaction between Stress and Compression was significant only in the subject analysis ($F_1(1,95)=6.53$, $p=0.012$; $F_2(1,34)=1.25$, n.s.). The statistics thus provide only weak support, if any, for the hypothesis that lexical stress increases robustness against time compression.

Some additional support for the effect of Stress may be found in the percentages of truncated responses. Disyllabic nonwords were expected to elicit truncated monosyllabic responses because unstressed syllables may become too short to be noticed. The target disyllabic words were expected to be affected by final lenghtening because target words embedded in short carrier phrases are often articulated almost as if in isolation. That is why particularly the finally stressed target words (weak-strong words) were expected to elicit monosyllabic responses (consisting of only the last strong syllable). Table 2.6 shows the percentages of truncated responses elicited by the disyllabic nonwords.

Table 2.6. Percentages of truncated responses to disyllabic nonwords.

	original duration	compressed to 40%	compressed to 35%
Initial stress (SW)	1	16	29
Final stress (WS)	0	5	7

Contrary to our expectations, there are relatively few truncated responses. The most striking finding, however, is that most truncations are found for the initially stressed (strong-weak) target words, contrary to the expectation. A possible explanation for this can be found in the choice of material. Target words with initial stress often had a typical unstressed syllable containing schwa, such as /–əl/ or /–ər/. Unstressed vowels in finally stressed, or weak-strong, words always contained 'full' unstressed vowels, which make the unstressed syllable more salient, and longer, than when it contains schwa. These data then do provide some support for the hypothesis that the identification of unstressed segments suffers more from time compression than stressed segments.

Summing up, we have found some evidence that segments with longer steady-state parts resist time compression better than segments with shorter or no steady-state parts. Vowels and fricatives are least affected by time compression, whereas plosives suffer most. There is no firm evidence that segments in stressed syllables resist time compression better than segments in unstressed syllables. A sub-hypothesis that was proven to be wrong was the hypothesis that consonants in clusters suffer more from time compression than singleton consonants because of their shorter duration in clusters. The present results suggest that plosives are actually easier to recover in clusters than when they occur on their own.

2.3.3 Hypothesis 3: Lexical redundancy will become more helpful the more difficult the listening situation

If the segmental intelligibility of a speech signal is poor, non-segmental top-down sources of information become all the more important. Thus, lexical redundancy is expected to become more important, the more the signal is time-compressed. This means that the difference in intelligibility between words and nonwords becomes greater with increasing speech rate.[4] Furthermore, identification of nonwords becomes

[4] Note that this hypothesis requires a comparison between two groups of subjects (14 subjects needed to be replaced for the second experiment). However, because this is a minority of the total number of subjects, and because the lexicality effects are large enough, this comparison seems to be warranted.

more difficult when more segments are involved, while recognition of real words is easier for longer words because of the relatively small lexical neighbourhood density.

The results presented in Table 2.7 (and again in Figure 2.3) show that the difference in identification scores between real words and nonwords becomes greater with increasing speech rate. The identification scores of the disyllabic real words remain highest in the most extreme condition, whereas those of the disyllabic nonwords are lowest. This clearly shows how lexicality is most helpful for disyllabic real words.

Table 2.7. Correct identification scores for monosyllabic and disyllabic items, broken down by lexical status (real vs. nonwords) and compression condition.

	Monosyllabic		Disyllabic	
	nonwords	real words	nonwords	real words
Original duration	85	99	78	99
Compressed to 40%	37	76	19	83
Compressed to 35%	26	66	10	69

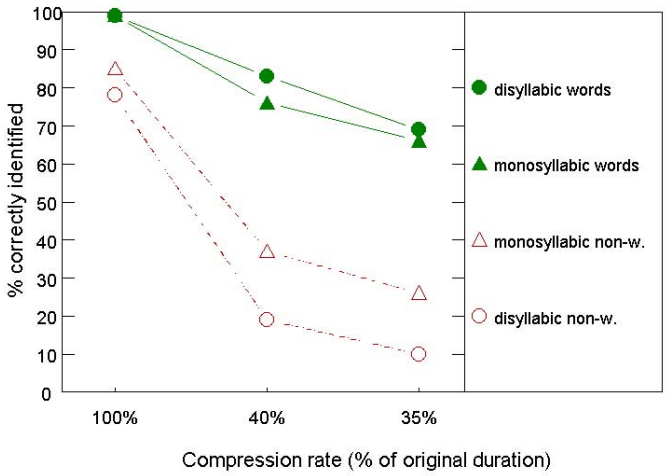

Figure 2.3. Identification results for real words and nonwords (disyllabic and monosyllabic), broken down by compression rate (as % of the original duration).

The results presented above were entered into analyses of variance (after arcsine transformation) to test the effects of Lexical Status, Compression Rate and Syllable Number. The overall effect of Lexical Status was highly significant ($F_1(1,95)=2816$,

$p<0.001$; $F_2(1,118)=412$, $p<0.001$), and so was the overall effect of Compression Rate ($F_1(2,94)=1752$, $p<0.001$; $F_2(2,117)=424$, $p<0.001$). The predicted interaction between Lexical Status and Compression Rate was also significant ($F_1(2,94)=143$, $p<0.001$; $F_2(2,117)=45.2$, $p<0.001$). The two-way interaction between Compression Rate and Syllable Number was not significant ($F_1(2,94)=4.17$, $p=0.018$; $F_2(2,117)<1$, n.s.). The effect of Syllable Number was significant ($F_1(1,95)=93.1$, $p<0.001$; $F_2(1,118)=5.49$, $p=0.021$), and so was the interaction between Lexical Status and Syllable Number ($F_1(1,95)=246$, $p<0.001$; $F_2(1,118)=22.1$, $p<0.001$). This means that the difference in identification score between disyllabic and monosyllabic words is affected by Lexical status: for real words, the difference between disyllabic and monosyllabic words is relatively small, whereas this difference is considerable for the nonwords. Compression Rate further affects this interaction, as shown by the significant three-way interaction between Lexical Status, Compression Rate, and Syllable Number ($F_1(2,94)=22.0$, $p<0.001$;$F_2(2,117)=3.71$, $p=0.027$).

Thus, the more the signal is time-compressed, the more important the role of lexical redundancy becomes. This is most obvious for the disyllabic words. The low recognition scores for the disyllabic nonwords show how poor segmental intelligibility actually is, but lexical redundancy can be used to fill in unintelligible segments for disyllabic words having only a limited number of close neighbours in the mental lexicon.

2.4 Discussion

The results of this experiment have shown that speech perception is quite robust against distortions in the time-scale of the speech signal. Listeners were capable of recognising about 60-70% of the real words in speech time-compressed to almost three times the normal rate.

The first expectation concerned the adaptation process. Dupoux & Green (1997) showed that the duration of the adjustment process depends on compression rate: the heavier the speech is time-compressed, the longer it takes to approach plateau performance. Still, Dupoux & Green (1997) also note that improvement in recall of compressed speech can occur with exposure to only 5 or 10 compressed sentences. In the present experiment each experimental block was preceded by a practice block of 9 items. This means that an important part of the adaptation process already took place during the practice blocks. Subjects were warned beforehand that the fastest speech would probably sound totally unintelligible to them at first, but that they would get

used to it soon. Within the duration of the experiment, the mean performance per block increases by ever smaller steps. Adaptation shows an exponential trend: the relative increase in performance becomes smaller over time. This suggests that subjects do not need intensive training: towards the end of the experiment, performance approached a plateau.

If plateau performance is approached during the experiment, performance will hardly get better with more training. In other words, there seems to be a limit to what listeners can adapt to. Secondly, the fact that there was no significant difference between the performance of subjects who had been exposed to highly time-compressed speech before and new subjects illustrated that the adaptation may be quite short-lived. Dupoux & Green (1997) showed that the adjustment process was not influenced by changes in compression rate. When intervening uncompressed sentences were presented, adjustment to time-compressed speech did not return to baseline performance. On the basis of this, they speculate that adjustment to time-compressed speech involves both short-term adjustment to local speech rate parameters, and longer-term adjustment which reflects a more permanent perceptual learning process. However, even though adaptation to time-compressed speech may not disappear immediately as soon as the rate of speech is slowed down, the present results show that it cannot be a transformation which is stored in long-term memory either. If adapting to highly time-compressed speech involves learning a 'trick', this would mean that once people have adapted to time-compressed speech, this ability is immediately available again even when they have not been presented with very fast speech for some time. An analogous example of learning such a trick or strategy is the ability to read geometrically transformed (e.g., mirrored or rotated) letters and text. Once people have learnt how to do this, the ability does not disappear, not even a year after the the initial experiment (Kolers 1975). Conversely, adapting to speech rate seems to be a more gradual process of 'tuning in': if one is not presented with time-compressed speech any more, the adaptation advantage disappears again. Still, these results cannot tell us whether repeated exposure to time-compressed speech would lead to a more lasting effect.

The present results have shown that performance for the nonwords apparently cannot be improved over 20-30% correct identification, whereas identification of the real words is about 75% correct. This disproves the idea of Foulke (1971) that listeners can cope with time-compressed speech as long as the storage capacity allows it. In the present study, subjects had plenty of time to process strongly time-compressed monosyllabic words (presented in short carrier phrases), which means that short term memory will not have been overcrowded. Still, segments may become so short that they exceed the limits imposed by the temporal resolution of the hearing system. The results of the listening experiment showed that some segments resist time compression better

than others. The difference in robustness against time-compression between consonantal classes can, to some extent, be predicted from the presence and length of the steady-state part. Plosive consonants, having no steady-state portion, are affected most by increasing time compression. This agrees with the finding that the shorter the transition, the more difficult it is to detect a change in frequency (van Wieringen 1995).

Furthermore, lexical stress seems to have an effect on how well segments are identified after time compression: because of the longer duration and the greater care of articulation, segments in stressed syllables resist time compression better than segments in unstressed syllables. This effect was weaker than we expected, however.

Although plosives were generally shorter in clusters, the identification scores were higher than when they occurred as singletons. Phonotactic restrictions on legal consonant sequences are responsible for the higher identification scores in clusters, together with coarticulatory information in the surrounding consonants. Summerfield, Bailey, Seton & Dorman (1981) show what this coarticulatory information may consist of. By manipulating the duration of the silent interval and the intensity fall-time of the fricative, they found out that plosive-perception depends heavily on abrupt amplitude changes in the preceding fricative, in distinguishing *slit* from *split*. This explains why the identification of plosives was higher when they occurred in onset clusters with /s/ than when they occurred as singleton consonants: the amplitude envelope of /s/ cued the presence of a following plosive. Singleton (voiceless) plosives have to be identified as plosives by the presence of the silent interval. In many cases, the duration of the silent interval was apparently too short: plosives were confused with nasals or were just not perceived at all. In combination with the amplitude envelope cue to the presence of a plosive, the duration of the silent interval does not have to be that long for successful identification of the plosive.

In plosive-liquid clusters coarticulatory (i.e., formant) cues to the presence of the plosive are dispersed throughout the liquid. The more distributed the information concerning the plosive is, the better it resists time compression. Although the duration of the plosive itself may be shorter when it is part of a cluster, the temporal window over which the cues to the presence of a plosive are distributed is wider for clusters than for singletons. Listeners then use their phonotactic knowledge to attribute this coarticulation information to phonotactically legal clusters.

The last hypothesis was that lexical redundancy becomes more helpful when the signal is more degraded. The results showed that the difference in intelligibility between real words and nonwords increases with higher rates of time compression. This was most obvious for the disyllabic words. Subjects make use of this lexical redundancy right from the start: the identification scores for the real words are much higher to start

with than those of the nonwords. The effect of adaptation is therefore more prominent for the nonwords than for the real words.

2.5 Conclusion

Subjects need only a limited amount of speech material to show significant improvement, or even plateau performance, in the identification of highly time-compressed speech. At the same time, the adaptation effect does not seem to be lasting. This proves the flexibility of the speech perception mechanism: listeners tune in to a fast speech rate quickly, but once they are no longer presented with time-compressed speech, they gradually lose the initial adaptation to it.

Secondly, the results disprove Foulke's (1971) point that successful processing of heavily time-compressed speech is possible as long as the listener has enough processing time. The duration of a segment's steady-state portion, and also the care of articulation of the segment itself, determine whether the segment can be identified after strong time compression. Obviously, there is a limit to what listeners can adapt to: the segmental intelligibility of speech, as measured in the identification of the nonwords, remains rather low at these heavy time-compression rates.

Lexical redundancy proved to play an important role. Lexical redundancy can be said to interact with 'robustness against time compression': for those consonant classes which were most affected by heavy time compression, lexical redundancy was more important for filling in the missing consonant. The robustness of the speech perception mechanism should thus mainly be attributed to non-segmental sources of information. It must be the combination of lexical redundancy, context information, knowledge of the world and adaptation to the segmental quality of time-compressed speech that makes visually impaired people able to 'read' the newspaper after fairly heavy time compression.

Listeners can make do with relatively poor segmental quality as long as the context suffices to fill in the missing information. Even though time compression of speech involves interfering with the pact between speaker and listener, listeners appear to be able to recover speech sounds and words. Now, if listeners can cope with these fast rates of speech for contextually redundant speech signals (such as newspaper text), why then do speakers not speak faster when this is allowed by the communicative situation? There are at least two possible answers to this question. One must note that we have only looked at intelligibility in this chapter, and not at processing difficulty. We do not know how tiring it is to listen to speech which is almost three times faster than normal.

Although our subjects performed quite well in the experiment, perhaps only visually impaired listeners will eventually be willing to trade increased difficulty for increased information rate and time savings. Another important aspect is the inability of the speaker to speak very fast. Some articulators are relatively slow and many speakers stumble when they try to speak very fast because of limited capabilities in motor programming and at higher, speech planning, levels. In Chapters 4 and 5 it will be made clear that time-compressed fast speech is easier to decode than naturally produced fast speech because of the inevitable slurring and reduced quality that accompany fast articulation rates.

The importance of segmental intelligibility to the overall intelligibility of speech is analysed further in the next chapter which focusses on the perception of natural speech vs. synthetic speech. These two types of speech differ in their segmental intelligibility, or at least in the ease of (phonetic) processing. It will be investigated whether time compression has the same effect on the two speech types, or whether one of them is more robust against time compression than the other.

3

Perception of Natural and Synthetic Speech After Time Compression

Abstract

Even though synthetic speech is generally perfectly intelligible when presented at a normal speech rate, the intelligibility of synthetic speech suffers more than that of natural speech when both types of speech are time-compressed. However, since some studies suggest that perception of fast speech is helped by segmental redundancy, the hyperarticulation often found in synthetic speech might turn into an advantage at a fast rate. If the segmental redundancy of hyperarticulated diphone speech, consisting only of hyperarticulated stressed building blocks, is helpful, then the processing advantage of natural over synthetic speech might decrease after artificial time compression. A second expectation was that processing time-compressed speech was expected to put a higher processing load on listeners than normal-rate speech. A phoneme detection experiment was set up to test processing speed of normal-rate and time-compressed natural and synthetic speech. The results showed that the processing advantage of natural over synthetic speech did not decrease, but rather tended to increase. Although the fact that all syllables in diphone speech are stressed and neatly articulated might help phonetic processing, the alternation of weak and strong syllables helps listeners to group syllables together and to start lexical access. Synthetic speech which, for the most part, lacks such speaking effort fluctuation becomes rather blurred at faster playback rates, which in turn hinders processing of, particularly, polysyllabic words. Phoneme detection times, which were assumed to be an indication of processing load, were faster in the time-compressed conditions, contrary to our expectation. This suggests that listeners adjust their response deadline to the input rate.

Part of this chapter also appeared in Janse (2002). 'Time-compressing natural and synthetic speech', Proceedings 7th International Conference on Spoken Language Processing, Denver, September 2002, pp. 1645-1648.

3.1 Introduction

In the previous chapter, segmental intelligibility, as measured in nonwords, was shown to be severely affected by time compression. The recognition scores of real words, on the other hand, are relatively high at these very fast rates. Lexical redundancy in real words helps listeners to fill in the 'difficult' segments. Thus, listeners have two sources of information when listening to real words: segmental intelligibility and lexical redundancy. In this chapter, segmental intelligibility is looked at from a different angle, namely by comparing the intelligibility of natural versus synthetic speech. Natural and synthetic speech differ in ease of processing at normal rates, and the central question in this chapter is how and whether the difference in processing speed between the two speech types changes under the influence of increased playback rate. The study described in this chapter focusses on low-level ease of processing of both types of speech, rather than on differences in perceived quality or naturalness between natural and synthetic speech.

Segmental intelligibility seems to be important for the perception of fast speech. In the study described in Chapter 4, speakers were asked to produce sentences at normal and very fast rates (cf. section 4.2). In a pilot study (presented in section 5.2.2), the intelligibility of this naturally produced very fast speech was compared with speech which was originally spoken at a normal rate, and then time-compressed to that same fast rate. The intelligibility of the time-compressed speech appeared to be much higher (90% correct word identification) than that of the fast articulated speech (64% correct identification). The slurring, coarticulation and assimilation processes that inevitably accompany a very fast speech rate obviously do not contribute to the intelligibility of speech, even though listeners might expect these processes to occur at such a fast rate.

Quené & Krull (1999) used a word detection task to investigate whether word recognition is speeded up by assimilation or is hampered by it. The type of assimilation was deletion of /t/ between consonants, as in Dutch *pos/t/ brengen* 'mail deliver'. Various studies (Gaskell & Marslen-Wilson 1996, 1998; Marslen-Wilson, Nix & Gaskell 1995) had suggested that, at normal speech rates, there is no perceptual advantage for assimilated over unassimilated articulations of a word form, given the appropriate phonological context. Quené & Krull (1999) thought that this could have been due to the rate and style of the experimental material in those three studies. If rate and style were intermediate, it could have been truly optional whether assimilation occurred or not. In that situation, listeners may not be biased towards either the assimilated, or the unassimilated form. Given a faster speech rate, however, listeners should expect certain assimilation processes to occur. Therefore, Quené & Krull (1999) argued, if listeners

expect assimilation, as in fast speech, then assimilated forms should be recognised faster than unassimilated forms. The results of the word detection study were rather surprising: whereas people detected the assimilated form of the word *post* faster than the unassimilated form at normal speech rate, the reverse was found for fast speech rate. Even though the unassimilated form was rather unnatural given the fast speech rate, listeners were faster in recognising it.

Thus, assimilation and coarticulation seem to deteriorate the intelligibility of fast speech. Word recognition and intelligibility in fast or time-compressed speech are helped by segmental redundancy, even if that segmental redundancy is artificially high. How does this relate to the idea that speakers take into consideration the needs of the listeners? Normally, speakers would only speed up if they assume that the listener can handle this loss of information. In our laboratory situations, speakers are asked to speed up their global speech rate, regardless of the communicative situation. Hence, they may be forced to neglect the needs of the listener, for articulatory reasons.

In this chapter, the question is raised whether the segmental redundancy, or hyperarticulation, present in synthetic diphone speech could be turned into an advantage, when the perception of time-compressed natural speech is compared with that of time-compressed synthetic speech. One of the most widely used speech synthesis systems for Dutch is Fluent Dutch (a commercial product of Fluency), which is based on diphone concatenation. The diphones of the Fluent Dutch diphone database are all cut from neatly articulated, stressed nonsense syllables. Concatenated strings of diphones are therefore, in a sense, segmentally maximally redundant: all syllables are originally stressed and carefully articulated. Since unstressed syllables in synthetic speech are given a short duration, as in natural speech, the hyperarticulation is all the more overdone. This hyperarticulation may sound rather unnatural for speech presented at a normal rate in good listening conditions. Still, although unnatural, segmental redundancy may be helpful for perception in difficult listening situations. At faster playback rates, the unnaturalness of hyperarticulated speech may be outweighed by its increased segmental redundancy (Quené & Krull 1999). Under difficult listening situations, perception might be helped by this type of overspecification. Even though natural speech is expected to have a processing advantage over synthetic speech, this advantage is expected to decrease when the playback rate is increased. Thus, we are not so much interested in the absolute difference in processing time between natural and synthetic speech, but rather in how the difference between the two changes when playback rate is increased. This means that the difference in perception between the two speech conditions at a normal rate must be evaluated against the difference between the two after time compression. As a consequence, intelligibility is not suitable for measuring differences between the two conditions, as both speech conditions are

perfectly intelligible at a normal rate. Phoneme detection speed has been shown to differ for natural and synthetic speech conditions when both conditions are perfectly intelligible. Phoneme detection thus provides a useful and sensitive measure of low-level acoustic/phonetic processing difficulty (Nix, Mehta, Dye & Cutler 1993; Pisoni 1997). Research on the intelligibility of time-compressed speech has also shown that speech time-compressed to about 1.5 times normal rate is still almost perfectly intelligible (cf. Chapter 1). Hence, phoneme detection can be used to evaluate the difference in processing speed between natural and synthetic speech both at a normal rate, and after moderate time compression.

The prediction is that after time compression, the processing advantage of natural over synthetic diphone speech is smaller than at a normal rate, because processing of time-compressed synthetic speech is helped by its greater segmental redundancy. Furthermore, the fact that all diphones have been cut from stressed syllables might mean that the advantage over natural speech comes out most clearly for polysyllabic words. Polysyllabic synthetic words are more redundant than polysyllabic natural words, as each syllable stems from an originally stressed syllable. Attempts have been made to improve the perceived naturalness of synthetic speech by recording both stressed and unstressed diphones (Drullman & Collier 1991), or by controlling articulation effort (Wouters & Macon 2002a, 2002b). Wouters & Macon's acoustic analyses of natural speech showed that spectral rate of change of vowel transitions increases with linguistic prominence (Wouters & Macon 2002a). The spectral rate of change can be predicted on the basis of the prosodic structure of the utterance. They describe an approach to integrate this knowledge into a concatenative speech synthesis system. Their results show that controlling the degree of articulation improves the perceived naturalness of speech (Wouters & Macon 2002b). Conversely, the use of unstressed diphones for unstressed syllables did not systematically result in more natural-sounding speech than when - temporally reduced - stressed diphones were used (Drullman & Collier 1991). The present study investigates whether the unnatural aspect of having equal stress (spectrally, but not temporally) on all segments might be turned into an advantage.

A second prediction is that increased playback rate makes perception more difficult. Even though speech is still perfectly intelligible when speech is accelerated about 1.5 times, time compression is expected to put a higher processing load on the listeners. This higher processing load may either be the result of the increased rate of information, or of the reduced segmental intelligibility of time-compressed speech (witnessed by the lower identification percentages in nonwords than in real words, cf. Chapter 1), or indeed of both. This higher processing load is expected to translate into slower detection times in the fast condition than in the normal rate condition.

The two hypotheses are repeated below:

- The difference in processing speed between natural and synthetic speech is smaller when both types of speech are artificially time-compressed than when they are played back at a normal rate.
- Time-compressed speech elicits slower detection times than speech which is presented at a normal rate.

To put these hypotheses to the test, a phoneme detection study was set up. This study is presented in section 3.3. Section 3.2 contains an experiment which was set up to compare the intelligibility of heavily time-compressed natural speech and synthetic speech. This intelligibility test was used as a first naïve attempt to find out whether synthetic speech has a higher intelligibility than natural speech after severe time compression. In section 3.4 the results of the pilot experiment and of the phoneme detection study will be discussed.

3.2 Experiment 1: Intelligibility of time-compressed natural and synthetic speech[5]

A pilot test was set up to compare the intelligibility of heavily time-compressed natural speech and synthetic speech. This pilot test can only indicate the difference in intelligibility between the two speech types after rather severe time compression. The hypothesis is that the intelligibility of synthetic speech is higher than that of natural speech because the high segmental redundancy of synthetic speech turns into an advantage when this type of speech is time-compressed.

3.2.1 Synthesis system

Speech synthesis must find a way to model phonetic transitions as well as the more stationary parts of speech. There are two major families of speech synthesis systems: rule-based systems and concatenation-based systems. In rule-based systems, the phonetic transitions and stationary parts of speech are modelled explicitly, in the form of rules that describe the influence of phonemes on one another. MITALK is an example of such a rule-based system (Allen, Hunnicutt & Klatt 1987). Rule-based

[5] This intelligibility test was carried out by two undergraduate students in phonetics, Fiona Sely and Eva Sittig, as part of a practical course.

synthesisers are always formant synthesisers which describe speech in terms of up to 60 parameters, mostly related to formant and antiformant frequencies and bandwidths (Klatt 1980). Exact modelling of the parameters is very time-consuming and error-prone, and achieving a high degree of naturalness is problematic (Dutoit 1997). Whereas rule-based synthesis requires much explicit knowledge, this type of phonetic knowledge is implicitly embedded in the stored segments in concatenation synthesis. Concatenation-based synthesis uses pre-recorded tokens of phonetic transitions and coarticulations into a database. The MBROLA system (Multi-Band Resynthesis Overlap-Add) is based on the concatenation of diphones (cf. section 1.2.2 on PSOLA and Dutoit (1997)). A diphone is a unit that begins in the middle of a phone and ends in the middle of the following one. Thus, diphones contain transitions between two speech sounds, recorded from natural speech. The standard Dutch diphone database stores over 2300 of such transition. The position of the boundary between the two phones is also stored, so that the duration of one half-phone can be modified without affecting the duration of the other half. To avoid amplitude mismatches at concatenation, the energy levels at the beginnings and at the end of segments are modified during an equalisation stage. This equalisation stage entails setting the energy of all phones of a given phoneme to their average value before storage (Dutoit 1997). The spectral envelopes, pitch, and phase of concatenated segments must somehow be adapted to one another to avoid audible discontinuities. The solution to this consists of resynthesising the original speech segments in the database. This is performed by resynthesising the voiced parts of all segments with constant synthesis pitch and fixed initial phases for each period, as performed by the MultiBand Resynthesis-PSOLA (MBR-PSOLA) algorithm (Dutoit & Leich 1993).

Spectral smoothing, or the attenuation of spectral mismatch, is solved only 'at run-time', by linear interpolation in the time-domain. Thus, the naturally introduced coarticulation is still maintained (Dutoit 1997). The MBR-PSOLA technique was turned into the more efficient MBROLA technique. The overlap-add technique can be applied during concatenation in order to provide the correct pitch and duration to the speech segments.

The input to the MBROLA synthesiser is a text file, containing a list of phonemes in SAMPA transcription (a machine-readable phonetic alphabet[6]), together with prosodic information (phoneme durations and a piecewise linear description of pitch).

[6] cf. http://www.phon.ucl.ac.uk/home/sampa/home.htm

3.2.2 Material

Forty sentence pairs were constructed for the present pilot experiment. The intelligibility test was set up such that the listeners had to fill in the missing word in a sentence. This word should therefore not be predictable from the sentence context. The sentences were selected randomly from a number of books, but were modified somewhat if necessary. Short sentence fragments were cut out of these sentences: these fragments were to be presented to the listeners. Two examples of sentence fragments are presented in (1) below (target words are in bold).

(1) Vorige week had de **juwelier** een grap gemaakt over dwergen[7]
Ik hoorde Robert zeggen dat hij nog een goede **kaart** zocht[8]

The target nouns contained one to three syllables and occurred in different positions in the sentence. The 40 sentences were read aloud by the same reference male speaker (Arthur Dirksen) whose diphones were used as the standard Dutch diphone database (NL2) for the MBROLA synthesiser (Dutoit 1997; Dutoit, Pagel, Pierret, Bataille & van der Vreeken 1996).

The natural speech material, as produced by the speaker, was recorded on DAT tape in a sound-treated booth with a Sennheiser ME30 microphone. The material was then fed as digital input to a computer, downsampled to 16 kHz, and then the sentence fragments were selected. These fragments were then segmented by hand. A close diphone copy was made based on the SAMPA transcription of the fragments, adjusted to the segment durations and pitch contours found in the natural sentences. All phoneme durations were exactly equal to those measured in the natural speech. The F_0 contour was a rough piece-wise close-copy version (time – logF_0 domain) of the F_0 contour found in the natural version. In this way the natural and synthetic conditions could be compared within a single speaker. Mean intensities of the synthetic and natural version of each sentence were made equal.

In Figure 3.1 below, a segmented waveform (plus SAMPA labels) of a natural sentence is shown, together with its close synthetic copy.

[7] 'Last week the **jeweller** had made a joke about dwarfs'
[8] 'I heard Robert say he was still looking for a good **map'**

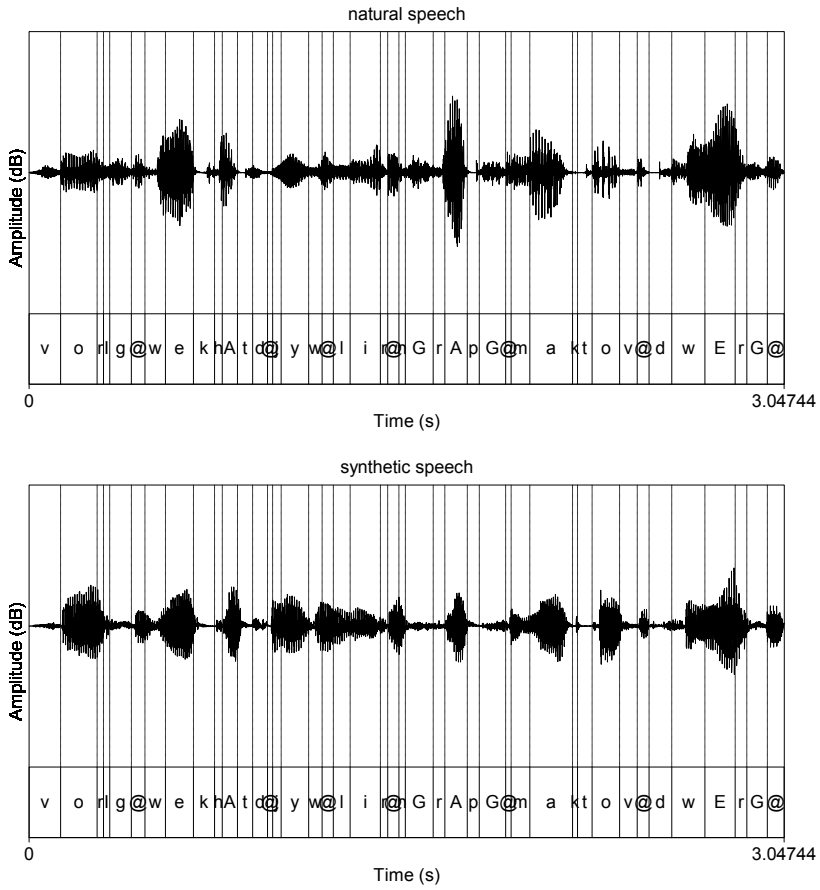

Figure 3.1. Waveforms and segment boundaries of natural and synthetic version of a test sentence (including SAMPA transcription, cf. footnote 6).

3.2.3 Design and Procedure

There were two experimental conditions (natural vs. diphone). The 40 items were rotated over the 2 conditions, yielding 20 items per condition. The design was a within-subject design and the two conditions were balanced (Latin square) over two experimental lists. Items were balanced over the two test conditions, in such a way that condition was not confounded with the monosyllabic/polysyllabic distinction.

First, by way of pre-tests, the degree of time compression was determined: intelligibility should not be too high to avoid ceiling effects and it should not be too low to avoid floor effects. After compression to 50% of the original duration,

intelligibility was still almost 100%. Therefore, a compression rate to 30% of the original duration was chosen. Time compression was carried out by means of PSOLA.

During the test session, subjects were seated in sound-treated booths, wearing closed-ear headphones. They were first presented with the sentence on a computer screen from which the target word was missing. Then the whole time-compressed sentence was presented to them over the headphones, including the target word. They were then asked to fill in the missing word. There was no time-pressure: only after they had hit the Enter key, would the following sentence appear on the screen.

3.2.4 Subjects

To each list, 18 subjects were assigned. The 36 subjects were all students at Utrecht University.

3.2.5 Results

A higher intelligibility score was expected for the synthetic speech than for the natural speech because of the higher segmental redundancy in synthetic speech. The overall raw correct recognition percentages (at compression to 30%) are shown in Table 3.1.

Table 3.1. Percentages of correct recognition for both speech types.

	Overall	Monosyllabic	Polysyllabic
Natural version	48%	51%	46%
Synthetic version	32%	34%	31%

The recognition percentages were computed for each item and for each subject in both conditions and (after arcsine transformation) were fed into ANOVAs in which either subjects or items were treated as repeated measures. First, the results do not support the hypothesis: the intelligibility of synthetic speech is lower than that of natural speech. The effect of Speech Type was significant in both the item and the subject analysis ($F_1(1,35)=36.5$, $p<0.001$; $F_2(1,39)=15.2$, $p<0.001$). Secondly, this lower intelligibility of synthetic speech holds for both monosyllabic and polysyllabic words. Although there were more polysyllabic words (N=28) than monosyllabic words (N=12) in the material, the effect of Syllable Number and the interaction between Syllable Number and Speech Type were analysed statistically in separate Repeated Measures ANOVAs. The effect of Syllable Number (monosyllabic vs. polysyllabic items) was not significant ($F_1(1,35)=4.31$, $p=.045$; $F_2(1,38)<1$, n.s.), and neither was the interaction between Speech Type and Syllable Number ($F_1(1,35)<1$, n.s.; $F_2(1,38)<1$, n.s.). Thus, our data

do not show different patterns for monosyllabic and polysyllabic words with respect to the intelligibility difference between the two types of speech.

A third observation is that the intelligibility of the polysyllabic words is overall lower than that of the monosyllabic words. Although this effect is not significant, it is rather surprising, because longer polysyllabic words are more redundant than shorter words. Another experiment (not described here) excluded the possibility that the monosyllabic words were more predictable from the sentence context. The very fast speech rate may have caused unstressed syllables to become too short to be perceived. Some of the unstressed syllables may have been extra vulnerable because of their segmental content. The segmental intelligibility results presented in Chapter 2 showed how some segments suffer more from time compression than others, in particular when they are part of the unstressed syllable. Thus, at this rate of speech, the higher lexical redundancy of disyllabic words is outweighed by unstressed syllables becoming perceptually obliterated, even though they were originally hyperarticulated.

Now that synthetic speech turns out to have a lower intelligibility than natural speech after severe time compression, the processing difference between the two speech types at a normal rate should also be established. Only if the processing differences at both normal rate and fast rate are known, do we know whether time compression does make the difference between the two smaller. Intelligibility at a normal rate is too high to find any differences in intelligibility by means of an intelligibility test, so a different type of test must be used to evaluate the differences between natural and synthetic speech at a normal rate and at a fast rate. For this purpose the phoneme detection task was selected, which has been shown to be a useful tool to compare the perception of highly intelligible speech types. The next section reports on a phoneme detection experiment, set up to test whether and how the processing advantage of natural over synthetic speech is affected by an increase in playback rate.

3.3 Experiment 2: Processing speed

The results of experiment 1 showed that, after heavy time compression, the intelligibility of synthetic speech is lower than that of natural speech. This means that the hypothesis for the intelligibility pilot was rather naïve: diphone speech does not only differ from natural speech in a positive, i.e., hyperarticulation sense. Diphone speech may be rich in acoustic cues, but it is also rich in false or misleading acoustic cues. Diphone /pɛ/ for the word *pen* may have been cut from the syllable *pet* and thus

still contains some cues for a coda /t/. Furthermore, it lacks cues for the actual nasal coda /n/, which may be equally disruptive for speech perception. Diphones can only account partially for the coarticulatory effects in speech because these often affect a whole segment rather than only its first or second half independently (O'Shaugnessy 1990). Concatenation of diphones also yields spectral discontinuities at the diphone edges. Although spectral smoothing is applied to make these discontinuities less audible, the signal is not as smooth as natural speech. Therefore, the processing difference at a normal rate should also be established. These negative aspects of synthetic diphone speech were expected to be equally harmful at a normal and at a fast playback rate. The hypothesis is not that synthetic speech is more intelligible than natural speech after time compression, but that the processing advantage of natural speech over synthetic speech decreases after time compression.

In Pisoni (1987; 1997) and Nix et al. (1993), reaction time measures are described as a tool for comparing perception of natural and synthetic speech. Nix et al. (1993) found longer response times for synthetic speech, relative to natural speech. Phoneme detection time is a good measure of the ease of processing, and thus of the speech communication quality of highly intelligible synthetic speech types. The difficulty of listening to synthetic speech has been argued to occur mainly at the lower phonetic level, and not at higher prosodic levels. Although improvements towards more appropriate and more natural prosody are certainly preferred by listeners (Terken & Lemeer 1988), the perceptual disadvantages at the lower phonetic level are assumed to demand the greater part of the extra processing capacity. Pisoni (1997) mentions several experiments with natural and synthetic speech. In an auditory lexical decision task subjects responded significantly faster to natural words and nonwords than to synthetic words and nonwords. The differences in response time between words and nonwords were equal for natural and synthetic speech. Thus, the "extra processing effort appears to be related to the initial analysis and perceptual encoding of the acoustic-phonetic information, and not to the process of accessing words from the lexicon" (Pisoni 1997: p.550). Similar results were obtained in a naming task using natural and synthetic words and nonwords. In this experiment subjects were asked to repeat the stimulus words which were presented to them auditorily. So, these two experiments suggest that early stages of perceptual encoding require more time for synthetic speech than for natural speech. Reaction time measures are assumed to give a better insight into processing difficulty than other measures taken after processing is complete (Levelt 1978).

Pisoni (1997) argues that if initial acoustic/phonetic analysis is slowed down, both pre-lexical processing, and consequently, lexical processing are slowed down as well. It is important to keep in mind that speech processing may work in this serial fashion, but phoneme detection responses can still be based on pre-lexical representations. Cutler &

Norris (1979) argued that phoneme detection can either be the result of a target detection procedure carried out on the pre-lexical representation or on the basis of phoneme information associated with a lexical representation. These two procedures run in parallel, and whichever is the fastest, wins the race. If the target is detected on the basis of pre-lexical information before lexical access is completed, the pre-lexical route wins. If lexical access is achieved before the target can be detected via the pre-lexical representation, the lexical route wins and, consequently, the response is based on the lexical representation. In the more recent Merge model (Norris, McQueen & Cutler 2000) phonemic decisions are argued to be based on the merging of pre-lexical and lexical information.

Note that the experiments of Pisoni (1997) and Nix et al. (1993) involved speech synthesis by rule. In the present study, diphone synthesis is used. This type of synthesis depends on the concatenation of naturally produced units. Nusbaum, Dedina & Pisoni (1984) argued that phonetic information is often redundantly specified in natural speech, but that in synthetic speech the cues are sparser, which implies harder work for the phonetic processor. Diphone speech is richer in phonetic cues than rule-based synthetic speech because of the inherent natural phoneme-to-phoneme transitions. In the present study the prosodic pattern (i.e., intonation pattern and durations) of the natural utterance is applied to its synthetic counterpart. This should help prosodic processing. In the Discussion section, the question at which level processing difficulties occur will be taken up again.

Phoneme detection can be used to evaluate the processing advantage of natural speech over synthetic diphone speech at a normal rate. The difference in response time can be investigated for highly intelligible time-compressed speech as well. By comparing the processing difference between the two types of speech at normal and fast rates, the main hypothesis can be tested that the processing advantage of natural over synthetic speech decreases with increasing rate. A second expectation is that time-compressed speech will elicit slower phoneme detection times than normal rate speech. Although speech time-compressed to 65% of the original duration is still highly intelligible, it is expected to put a higher processing load on the listeners. This higher processing load was expected to translate into slower phoneme detection times.

3.3.1 Method

Material

The material that was used in the two pilot tests did not contain enough suitable target phonemes. Those sentences formed only a small sub-set of the recording from which they had been taken. Therefore, a new sample of 100 sentences with suitable target

phonemes was selected from the recording of the same male speaker whose diphones are used as the standard MBROLA diphone set for Dutch. These sentences had all been taken from books, and had been modified slightly, if necessary. The target phonemes were all word-initial plosives: /p,t,k,d,b/. Because the speech material had not been designed for the purpose of a phoneme detection experiment, all possible target words were chosen. Consequently, the 100 sentences were not balanced over these five phonemes: there were many /b/s and only few /d/s in the material. The target word could be either a noun, a verb or an adverb. Target words were monosyllabic (32) or polysyllabic (68). If possible, the target phoneme did not occur elsewhere in the sentence, either word-initially, word-medially or word-finally. Three example sentences are presented in (2) below (target word in bold):

(2) target /p/ De **pater** en de non ruimden gezamenlijk de tafel op het terras af[9]
target /d/ Alleen één **ding** is in elke auto onzeker[10]
target /b/ Het mooie servies van oma was weer **b**ijna compleet[11]

A synthetic copy of the 100 sentences was made with the help of the Dutch text-to-speech conversion program Fluent Dutch (version 1.6). Fluent Dutch is based on the MBROLA synthesiser, and the same Dutch diphone set was used as in the experiment described above (NL2). Grapheme-to-phoneme conversion was done automatically, but was corrected manually if necessary. Fluent Dutch also computes a natural pitch contour and suitable durations. By default, the program assigns sentence accent to all content words, including main verbs. If certain words are to be accented or deaccented, this can be indicated in the orthographic input. The pitch contour was adapted manually to make it similar to that of the natural utterance. Secondly, the target word was made just as long as the natural version by means of linear time-scaling. Furthermore, the durations of the parts of the sentence preceding and following the target word were made equal to the natural version. This was done by means of PSOLA time scaling as implemented in the speech editing program GIPOS. Note that, unlike in the previous pilot tests, the natural sentences were not segmented manually phoneme-by-phoneme. Only the durations of the target words, and the parts preceding and following the target words were made equal in duration. All phoneme durations within the words were computed by the speech synthesis program Fluent Dutch.

For the fast condition, the natural and synthetic versions of the test sentences were time-compressed linearly to 65% of their original duration by PSOLA time scaling. This

9 'The priest and the nun together cleared the table on the terrace'
10 'Only one thing is in each car uncertain'
11 'Grandmother's beautiful crockery set was almost complete again'.

is about the fastest speech rate speakers can attain if they try very hard (cf. Chapter 4). Speech that is time-compressed to this rate is almost perfectly intelligible. The 100 test sentences were rotated over the 4 conditions (Natural-normal rate, Synthetic-normal rate, Natural-fast, Synthetic-fast), using a Latin square design. Because each subject could be presented with each item only once, there were 4 experimental lists. Apart from the 100 test sentences, there were 10 practice sentences, 10 warming-up sentences and 70 filler sentences. The filler sentences did not contain the phoneme the subjects were asked to detect, and were included in order to prevent subjects from pressing the button randomly. The fillers were rotated over the 4 test conditions and interspersed with the material.

Subjects

Ten subjects were assigned to each of the 4 experimental lists. The 40 subjects were all students at Utrecht University, and were paid a small amount of money for their participation. None of them reported any hearing or reading problems.

Procedure

Subjects were seated in a sound-treated booth and were tested individually. The speech material was presented to them over closed headphones. They first read instructions on the computer screen in front of them before they started the experiment. They were told to look at the screen because a letter would appear on the screen before the sentence was played to them. Once the sentence was playing, they were told to press the button as soon as they heard this sound in word-initial position. The spelling of all test words was regular: if subjects were asked to monitor for /k/, there were no target words which are spelled with 'c'. The onset of the target plosives was marked in the speech waveform by means of a time marker. The program then computed the phoneme detection time by measuring from that marker point in time to the moment that the button press was registered.

After the practice session, subjects could still ask questions if anything was unclear. Before subjects started with the actual test, 10 warming-up sentences were played to them after which they proceeded seamlessly with the actual test. All test and filler items were presented in random order. The experiment lasted 20 minutes.

3.3.2 Results

After subjects had finished the test, they were asked whether they thought that all speech conditions had been intelligible. Most subjects thought that all speech conditions were of good intelligibility. However, some subjects thought that the time-

compressed synthetic speech sounded a bit blurred and that it was difficult to detect word boundaries.

The raw mean phoneme detection times were computed, along with the percentage of missing observations. Missing observations were due to subjects missing the phoneme, or responding too early (i.e., to another phoneme). The raw detection times, plus the miss rates in each condition, are shown in Table 3.3.

Table 3.3. Raw mean phoneme detection time (in ms) plus standard error of mean, and miss rate for natural and synthetic conditions, broken down by speech rate.

	Normal rate			Time-compressed		
	mean	s.e.	miss rate	mean	s.e.	miss rate
Natural speech	598	8	3%	583	10	5%
Synthetic speech	677	11	7%	654	10	15%
Difference natural-synthetic	-79			-71		

The difference in detection time between synthetic and natural speech is quite large: 79 ms at normal rate. This difference is somewhat smaller after time compression. As a first (quick and dirty) analysis, all missing observations were replaced by the grand mean of 627 ms[12]. These results were entered into analyses of variance on items and on subjects (Repeated Measures) to test the effects of Speech Type and Rate. The effect of Speech Type was highly significant ($F_1(1,39)=77.0$, $p<0.001$; $F_2(1,99)=8.92$, $p=0.004$). This means that natural speech is easier to process than synthetic speech. A second hypothesis is that phoneme detection times are slowed down by time compression because of the higher processing load. The effect of Rate approached significance in the subject analysis, but was insignificant by items ($F_1(1,39)=4.02$, $p=0.052$; $F_2(1,99)<1$, n.s.). Thus, these data do not provide evidence for the idea that time compression makes speech more difficult to process. The main hypothesis, however, was that the processing advantage of natural speech over synthetic speech would decrease after both speech types had been time-compressed. The interaction between Speech Type and Rate was far from significant ($F_1(1,39)<1$, n.s.; $F_2(1,99)=1.1$, n.s.).

The miss rates (cf. Table 3.3) were also analysed to establish the effects of Speech Type and Rate on the number of missing observations. For both speech types, the miss rate increases as a result of time compression. The miss rates (after arcsine transformation) in all four conditions were analysed in ANOVAs treating either items or subjects as repeated measures. These analyses showed significant effects of Speech

[12] Repeated Measures ANOVAs in SPSS cannot cope with missing values. There are more sophisticated ways to replace missing values (Girden 1992). Particularly when missing values are not distributed equally over the conditions, replacing them by the grand mean is not very elegant.

Type ($F_1(1,39)=43.7$, $p<0.001$; $F_2(1,99)=35.1$, $p<0.001$), and of Rate ($F_1(1,39)=16.0$, $p<0.001$; $F_2(1,99)=10.2$, $p=0.002$). The interaction between Speech Type and Rate was not significant in either analysis ($F_1(1,39)=1.41$, n.s; $F_2(1,99)=2.27$, n.s.).

In a second analysis, several items were left out of the analysis. Although subjects thought that intelligibility of the speech material was generally very high, those items that elicited many missing values may have been lower in intelligibility than the rest. There should be 10 observations for each item in each condition (10 subjects on each list). If the number of valid observations (excluding missing data) was lower than 7 out of 10, the item was left out of the analysis. This was the case for 15 out of 100 items. In order to obtain equal numbers of observations for all conditions on each list, 9 more items had to be left out of the analysis. These 9 items were chosen at random. Thus, in total, 24 items were left out of the analysis. Overall, the percentage of missing observations in the remaining subset of 76 items was 3%. The mean raw detection times for this subset are shown in Table 3.4.

Table 3.4. Mean raw detection time (in msec) for subset of items, plus standard error of mean and miss rate for natural and synthetic speech, broken down by speech rate.

	Normal rate			Time-compressed		
	mean	s.e.	miss rate	mean	s.e.	miss rate
Natural speech	598	10	1%	554	9	2%
Synthetic speech	654	11	3%	644	11	6%
Difference natural-synthetic	-56			-90		

Because the missing observations are not distributed equally across the different conditions, the missing observations were replaced by the subject's mean for that condition in the subject analysis, and by the item mean in that condition in the item analysis (Girden 1992). Furthermore, after that, the reaction time data were transformed for the following reason. Reaction time data are not distributed in a normal (or Gaussian) way. Analyses of variances actually assume normally distributed data. In order to make the distributions more normal, the reaction time data were transformed to inverse reaction times (1/RT). A Kolmogorov-Smirnov test showed that even these transformed data were not entirely normally distributed ($z=5.63$, $p<0.001$). The statistical analyses were run again on the transformed data. As in the previous analysis, there was a significant main effect of Speech Type ($F_1(1,39)=60.9$, $p<0.001$; $F_2(1,75)=23.0$, $p<0.001$), but now, the effect of Rate was also significant ($F_1(1,39)=8.7$, $p=0.005$; $F_2(1,75)=6.2$, $p=0.015$). Note that this Rate effect is in the opposite direction from what was expected: generally, detection times turn out to be faster in the time-compressed condition than in the normal rate condition. The

interaction between Speech Type and Rate was significant as well ($F_1(1,39)=5.0$, $p=0.032$; $F_2(1,75)=4.8$, $p=0.031$).

Thus, it is clear that there is, at least, a tendency towards an increase of the processing advantage of natural speech over synthetic speech when both types of speech are time-compressed.

The miss rates (after arcsine transformation) in all four conditions in the selected subset (cf. Table 3.4) were analysed in ANOVAs treating either items or subjects as repeated measures. The analyses showed significant effects of Speech Type ($F_1(1,39)=38.2$, $p<0.001$; $F_2(1,75)=21.8$, $p<0.001$) and Rate ($F_1(1,39)=11.8$, $p=0.001$; $F_2(1,75)=7.06$, $p=0.01$). The interaction between Speech Type and Rate was not significant ($F_1(1,39)=2.77$, $p=0.104$; $F_2(1,75)=1.68$, $p=0.199$). So, also in this subset of items, miss rate increases after time compression.

Of the 76 remaining items, 25 items were monosyllabic and 51 were polysyllabic. Separate analyses were carried out for the monosyllabic and polysyllabic subsets. Note that each of the two subsets of data cannot be completely balanced over the 4 experimental conditions. The behaviour of the monosyllabic items differs considerably from that of the polysyllabic items. This is illustrated in Figure 3.2 below.

For the monosyllabic items, the processing advantage of natural over synthetic speech is substantial at both rates (122 ms at normal rate (529 natural vs. 651 synthetic); and 94 ms at fast rate (515 natural vs. 609 synthetic)). Univariate ANOVAs (which allows unequal numbers of observations over cells) of the inverse reaction time data (1/RT) showed a significant main effect of Speech Type ($F_1(1,39)=25.4$, $p<0.001$; $F_2(1,24)=10.6$, $p=0.003$). The effect of Rate was not significant ($F_1(1,39)<1$,n.s.; $F_2(1,24)<1$, n.s.), and neither was the interaction between Speech Type and Rate ($F_1(1,39)<1$, n.s.; $F_2(1,24)<1$, n.s.). For the polysyllabic items, the difference between the two speech conditions was 20 ms in the normal rate condition (633 natural vs. 653 synthetic), and 110 ms after time compression (575 natural vs. 685 ms. synthetic). For these items, the interaction between Speech Type and Rate was significant ($F_1(1,39)=6.24$, $p=0.017$; $F_2(1,50)=5.36$, $p=0.025$), and so were the main effects of Speech Type ($F_1(1,39)=36.7$, $p<0.001$; $F_2(1,50)=11.8$, $p=0.001$) and of Rate ($F_1(1,39)=13.4$, $p=0.001$; $F_2(1,50)=9.12$, $p=0.004$).

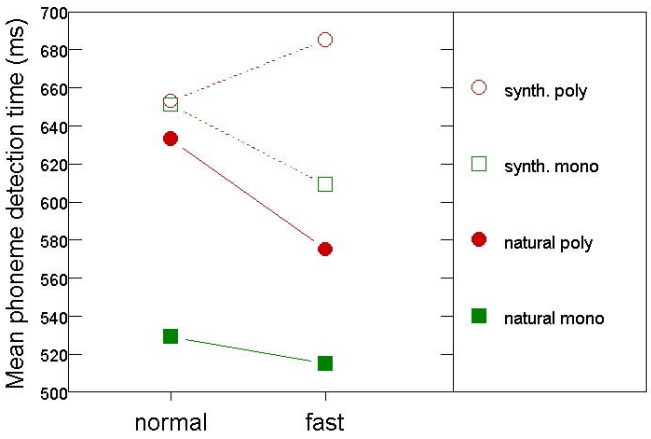

Figure 3.2. Mean phoneme detection time (in ms), for natural (solid lines) and synthetic conditions (dotted lines), broken down by speech rate and by monosyllabic (squares) vs. polysyllabic (circles) items.

Three main observations can be made from these data. First, the detection times are not slowed down by increased rate, which is against our prediction. For both speech types, detection times are even somewhat faster in the time-compressed than in the normal rate condition. This effect is significant only when the least intelligible items are left out of consideration. In other words, items are processed faster when speech is time-compressed, but, at the same time, miss rates in the fast conditions are significantly higher than in the normal rate conditions. The faster detection times in the time-compressed conditions are mainly due to the natural (polysyllabic) items. The second observation is that there is (at least a trend towards) an interaction between Speech Type and Rate (significant in the ANOVAs on the subset). Processing is sped up by increased playback rate for the natural items, but remains relatively unaffected for the synthetic items. Contrary to the first hypothesis, the processing advantage of natural over synthetic speech does not decrease when both speech types are time-compressed, but rather tends to increase. The third observation is that this mainly goes for the polysyllabic items. Time compression has a different effect on polysyllabic natural items than on monosyllabic natural items.

3.3.3 Further analysis of phoneme detection data

The phoneme detection literature reports several effects which may have blurred the expected differences between monosyllabic and polysyllabic items in the data. Because the sentences had not been designed for the purpose of a phoneme detection experiment, these effects were not systematically controlled. If, accidentally, all monosyllabic items score differently on one of these variables than the polysyllabic items, this might have influenced the data. Five such variables, the same as investigated in Nix et al. (1993), will be discussed below in relation to the present data set.

(1) *Transition probability.* Phoneme targets in contextually predictable words are detected faster than targets in unpredictable words (Dell & Newman 1980; Morton & Long 1976). By means of a paper-and-pen cloze test which was presented to 20 subjects, the predictability of the target items in our 100 sentences was established (following the procedure described in Nix et al. (1993). For both the monosyllabic words and the polysyllabic words, we checked whether targets in predictable items were detected faster than targets in unpredictable items. This appeared not to be the case: in all four conditions, detection times were slower for the predictable items.

(2) *Preceding word length.* If phoneme targets are preceded by long words, they are detected faster than when preceded by short words (Mehler, Segui & Carey 1978). The isolation or recognition point of longer words is often before the end of the word, whereas the isolation point of short words can be one or two syllables after the offset of the word (Grosjean 1985). However, in our material, targets which were preceded by monosyllabic words or by no word at all were detected faster than targets which were preceded by longer words.

(3) *Position of target bearing word in the sentence.* The later the target bearing item occurs in the sentence, the faster the RTs are (Cutler & Fodor 1979; Foss 1969). This is caused by the context being more restrained towards the end of the sentence. As in Nix et al. (1993) detection times were compared for targets in syllable positions 1 to 5 (early) with detection times in later positions. There was some evidence for this factor in our data: collapsed over all four conditions, targets were detected 22 ms faster in late positions than in early positions. There was no indication of an interaction between this factor Position and Speech Type or Rate.

(4) *Sentence accent.* Targets in words with sentence accent are detected faster than targets in unaccented words (Cutler 1976). There was no consistent effect of sentence accent on the detection times in the present study.

(5) *Lexical stress.* Taft (1984) showed that initial phonemes were detected faster when they were part of the stressed syllable than when part of an unstressed syllable. There is evidence that this effect plays a role only in spontaneous and not in read speech (Mehta

& Cutler 1988). For our polysyllabic items, we checked whether this was also the case in our data. Only in the natural-speech-normal-rate condition, were target items in initially stressed words detected faster (22 ms) than targets in non-initially stressed words.

Thus, of the five variables mentioned above, only the effect of position in the sentence appeared to play a role in our data. It is not clear how these five variables can shed any light on the results presented in Figure 3.2. What needs to be explained is not why there are differences between groups of items (i.e., monosyllabic vs. polysyllabic words), but rather how these groups are affected by an increase in rate. The key question remains why the difference between the natural and synthetic conditions remains stable for monosyllabic words, whereas the difference between the two speech conditions increases markedly for polysyllabic words. The variable 'Position in the Sentence' does not explain why this is the case. [13]

3.4 Discussion

Despite the high intelligibility and quality of synthetic diphone speech, listeners still find natural speech easier to process than synthetic speech. Now that natural speech was found to be more intelligible than synthetic speech after equal rates of time compression (in experiment 1), it was investigated whether the processing advantage of natural speech is affected by an increase in playback rate. The hypothesis was that the hyperarticulation that is present in synthetic diphone speech, consisting only of neatly articulated stressed syllables, might become advantageous in difficult listening conditions.

A second hypothesis was that processing would be slowed down for time-compressed speech, relative to normal rate speech because of the higher processing load of fast speech. A phoneme detection study (experiment 2) was set up to investigate these two hypotheses.

Contrary to what we expected, phonemes were detected faster in time-compressed speech than in the normal rate condition. There are at least two possible explanations

[13] One of the remaining factors that might have influenced the detection times could be the intelligibility or processing difficulty of the preceding word. Although it is assumed that all sentences were perfectly intelligible, at least at sentence-level, it is possible that subjects experienced difficulty in processing certain pre-target words in certain conditions. The outlier data condition in Figure 3.2, the polysyllabic synthetic condition, may have suffered from 'low-quality' words preceding the target words. This can only be checked by way of another phoneme detection experiment.

for this. Listeners may adapt to the higher rate of information, and phonetic and lexical processing may not become more difficult when speech is time-compressed to this moderately fast rate. The shorter detection times in the fast condition could then be caused by syllable and word durations being shorter. A target phoneme can only be detected after the word or the syllable has been processed. When speech is time-compressed, the isolation point at which the word can be recognised is reached earlier. Thus, the contribution of the target's duration to the total phoneme detection time is smaller after time compression. If this is the case, one would expect to find a significant correlation between detection time and length of the (monosyllabic) items in normal and fast condition, more specifically, between the difference in item duration and the difference in mean reaction time in the normal and fast condition. For all monosyllabic items, the difference in mean reaction time was computed between the normal and fast conditions, together with the difference in mean duration between the normal and fast condition. This correlation was not significant: neither when the results were collapsed for both speech types (Pearson correlation coefficient $r=0.06$, $p=0.76$), nor when only the natural item results were analysed ($r=0.26$, $p=0.22$). So, there is no linear correlation between the shorter duration of the monosyllabic words and the faster detection times.

An alternative explanation might be that listeners adjust their response deadline to the input rate. As a reaction to the fast rate of presentation, they adapt their processing rate. They succeed in reacting faster than normally, and thus in keeping up with the higher rate, but only at the expense of making more errors. This is witnessed by the higher miss rates in the time-compressed conditions, for both speech types.

Contrary to the hypothesis, the processing advantage of natural over synthetic speech did not become smaller after both speech types had been time-compressed. The phoneme detection study showed that the processing advantage of natural over synthetic speech was relatively large: phoneme detection time was 79 ms faster for natural speech than for synthetic speech when both speech types were presented at a normal rate. This supports the results of Nix et al. (1993) that phoneme detection can be used to test differences in speech communication quality. If the least intelligible items are left out, the processing advantage still amounts to 54 ms at the normal rate of presentation. This difference in quality must, at least partly, be attributed to phonetic processing difficulties. Pisoni (1997) and Nix. et al. (1993) relate the slower phoneme detection times in synthetic speech to the extra processing effort in the initial analysis and perceptual encoding of the acoustic-phonetic information. If this initial pre-lexical analysis of the speech signal requires more processing time, word recognition will consequently be delayed as well, relative to natural speech. Importantly, the processing advantage of natural over synthetic speech did not decrease after time compression.

There was even a tendency towards the opposite: in particular for the polysyllabic items, the advantage of natural speech even increased at fast playback.

A first explanation concerns the initial assumption that the negative aspects of synthetic speech would be independent of compression rate. It is conceivable, however, that these negative aspects (i.e., misleading coarticulatory cues) do become more harmful to speech processing when speech is time-compressed. Misleading spectral cues make initial pre-lexical processing more difficult, so both the pre-lexical and lexical route are slowed down. Even if segmental redundancy has a positive effect in difficult listening conditions, this may have been outweighed by these negative aspects becoming more problematic.

A second possible explanation bears on the different patterns found for monosyllabic and polysyllabic items. On the basis of the hyperarticulation-is-helpful-in-adverse-listening-conditions hypothesis, one would have expected the processing advantage of natural speech to decrease mainly for polysyllabic items because polysyllabic items consist only of stressed-and-hyperarticulated syllables. Hyperarticulation of unstressed syllables should make initial low-level analysis easier, and, consequently, phoneme detection via the pre-lexical route should benefit from the higher segmental intelligibility. For the polysyllabic items presented at a normal rate, natural speech has no robust processing advantage over synthetic speech. For weak-strong words (non-initial stress), targets in synthetic speech items were even detected 16 ms faster than in natural items in the normal rate condition. This was not the case at the fast rate of presentation: in the fast condition, targets in natural versions are detected faster (84 ms) than in synthetic versions. For strong-weak polysyllabic items, targets in natural versions were detected faster than those in synthetic conditions, both at the normal rate (natural speech advantage is 49 ms), and at the fast rate (advantage is 85 ms).

With respect to the fast rate of presentation, note that some subjects complained that the time-compressed synthetic speech sounded blurred to them, and that they found it difficult to detect word boundaries. When segmental intelligibility is affected because of time compression, it is conceivable that the prosodic template of a word, consisting of the speaking effort and duration pattern, becomes more important for the recognition of the word. Stress information is spread over longer chunks of the speech signal and is thus more robust against time compression than segmental information. Although stressed syllables and unstressed syllables differ in duration in the synthetic condition, as in natural speech, the natural speaking effort fluctuation due to different levels of stress is largely missing in synthetic speech. Speaking effort translates into loudness, but also into articulatory precision. The speaking effort contour may be an important suprasegmental characteristic of speech, which helps listeners to group weak

and strong syllables together. This grouping together is essential for the recognition of polysyllabic words and for the ease of processing of syntactic chunks. So, although we had expected hyperarticulation to work out positively, the absence of variation in speaking effort turns out to be harmful to the 'holistic' processing of e.g., polysyllabic words. Note that these prosodic cues make lexical access easier, not initial pre-lexical analysis. If we assume that the lexical route contributes most to the ultimate phoneme decision response, the lack of proper stress information should slow down the lexical route.

There is some further evidence that the lexical route contributes more to the phoneme detection results than the pre-lexical route. Phoneme detection times are slower overall for polysyllabic words than for monosyllabic words. This holds for natural speech at both rates, but for the synthetic speech only at the fast rate. At a normal rate, detection times of synthetic monosyllabic words are about equal to those of synthetic polysyllabic words. It is also important to note that, for the natural speech, the difference between the average detection time of monosyllabic and that of polysyllabic items decreases at faster playback rate: the difference between monosyllabic and polysyllabic is 114 ms at the normal rate, and 60 ms at the fast rate. This agrees with the fact that the difference in duration between monosyllabic and polysyllabic words after time compression is only 65% of what it was at the normal rate. Note that these results are in conflict with the results of Dupoux & Mehler (1990) who found that even after time compression, phoneme detection does not necessarily depend on the lexical code. However, Dupoux & Mehler's phoneme detection study (1990) was based on the presentation of single target words, which resulted in much faster detection times (about 430 ms) than we found in our present study (about 580 ms for the time-compressed natural speech). The design of their experiment may have caused subjects to rely more on the pre-lexical code than on the lexical code. When the targets are embedded in meaningful sentences, as was the case in the present experiment, the information from the *lexical*, rather than the *pre-lexical* route weighs more heavily (Cutler & Norris 1979).

The speaking-effort account for the increasing processing advantage of natural over synthetic polysyllabic words agrees with the metrical segmentation study of Cutler & Norris (1988), and the subsequent study by Young, Altmann, Cutler & Norris (1993). In Young et al. (1993) the question is raised whether speech is easier to recognise under difficult listening conditions when all strong syllables are word-initial. This question was based on Cutler & Norris (1988), who demonstrated that speakers of English segment speech input at the onset of strong syllables in the absence of explicitly marked cues to word boundaries. Young et al. (1993) tested whether time-compressed sentences in which all content words began with strong syllables proved

easier to recognise than time-compressed sentences in which all content words began with weak syllables. No difference was found between the two metrical conditions, but the idea that listeners are highly sensitive to metrical stress under difficult listening conditions is attractive. In the sense that speech recognition is pattern recognition, the speaking effort contour may be an important suprasegmental characteristic of polysyllabic words. Contrary to the hypothesis, hyperarticulation of diphone speech turns out to be harmful in difficult listening conditions.

Others have shown that there are ways of increasing intelligibility over normal intelligibility. Intelligibility of nonwords in noise can be improved by cue-enhancement of certain consonantal regions of rapid spectral change (Hazan & Simpson 1998). The tentative conclusion, however, is that the type of hyperarticulation present in diphone speech, i.e., having equal stress on all syllables, turns out to be harmful in adverse listening situations. For the recognition of polysyllabic words, a natural speaking effort contour is at least equally important.

3.5 Conclusion

Three conclusions can be drawn from our data. First, time-compressed speech is more difficult to process than speech presented at a normal rate, but this does not translate into slower detection times. Subjects adapt their response time in order to keep up with the higher input rate, but at the expense of making more errors.

Secondly, synthetic speech is more difficult to process than natural speech, both at a normal and at a fast rate. This is witnessed by lower scores in the intelligibility test, a higher miss rate in the phoneme detection experiment, and longer phoneme detection times. Misleading coarticulatory cues, consequent spectral discontinuities, and possibly, the lack of a natural prosodic pattern all contribute to this processing difficulty.

Thirdly, the data did not support the expectation that hyperarticulation that is found in synthetic speech is helpful at a faster playback rate. From a segmental intelligibility viewpoint, equal stress on all syllables might enhance intelligibility. However, the lack of speaking effort fluctuation, as an important suprasegmental characteristic of polysyllabic words, becomes more problematic for word recognition at a fast rate.

The results of this chapter show that segmental and suprasegmental factors both influence lexical processing. The question of how an increase in speech rate affects segmental and prosodic characteristics in natural speech plays an important role in the next chapter. The key question is again how segmental and prosodic factors contribute to speech intelligibility of artificially time-compressed speech.

4

Timing of Natural Fast Speech and Word-Level Intelligibility of Time-Compressed Speech

Abstract

In this study we investigate whether speakers, in line with the predictions of the Hyper- and Hypospeech theory, speed up most during the least informative parts and less during the more informative parts, when they are asked to speak faster. We expected listeners to benefit from these changes in word-level timing, and our main goal was to find out whether making the temporal organisation of artificially time-compressed speech more like that of natural fast speech would improve intelligibility over linear time compression. Our production study showed that speakers reduce unstressed syllables more than stressed syllables, thereby making the prosodic pattern more pronounced. However, both at very fast speech rates, and at moderately fast rates, applying fast speech timing worsens intelligibility or delays processing. It seems that the non-uniform way of speeding up may not be due to an underlying communicative principle, but may result from speakers' inability to speed up otherwise. As both prosodic and segmental information contribute to word recognition, we conclude that putting too much emphasis on either distorts the optimal balance between these two factors, and harms word perception.

This chapter is an extended version of an article by Janse, Nooteboom & Quené (in press) 'Word-level intelligibility of time-compressed speech: prosodic and segmental factors'; to appear in Speech Communication.

Parts of this chapter also appeared in conference proceedings papers: Janse, Sennema & Slis (2000), Janse (2000), and Janse (2001).

4.1 Introduction

Artificial time compression of speech, e.g., for the purpose of fast playback of long recordings, e-mails or voicemail messages, is normally performed in a linear way. This means that all segments are reduced by the same proportion. The relative timing pattern of speech played back at a fast rate is that of the original rate at which this speech was produced. It is a matter of debate whether speakers, when they are forced to speak faster than normal, also apply linear time compression. Kozhevnikov & Chistovich (1965) supported this notion of invariant timing patterns in speech movements. According to them, it is unrealistic to assume that there are separate motor programs for each rate at which an utterance can be produced. They therefore suggested that the rate of production may not be specified in the motor program but presents the "speed of realisation of the program" (Kozhevnikov & Chistovic 1965). Kozhevnikov and Chistovich found temporal invariance for the relative duration of words in a phrase: regardless of speech rate, the duration of each word was a constant proportion of the duration of the entire sentence. This temporal invariance was also found for the relative duration of the syllables in a word across different speech rates. However, they also found that the relative duration of the sounds in a syllable does vary as a function of speech rate.

Later studies showed that rate-dependent changes in timing in English are not confined to the within-syllable level, but also occur between syllables and between words. When people speak faster, consonant durations are reduced less, relatively, than vowel durations (Gay 1978; Lehiste 1970; Max & Caruso 1997). Furthermore, durations of sentence-stressed syllables are reduced less, relatively speaking, than durations of unstressed syllables (Peterson & Lehiste 1960; Port 1981). As a result, the relative difference in duration between stressed and unstressed syllables (i.e., the stressed/unstressed ratio) increases in faster speech, thereby making the prosodic pattern more prominent. This nonlinear way of increasing speech rate indicates that speakers are selective in their compression behaviour. The more prominent prosodic pattern might be the result of a strategic and communicative principle, namely that speakers tend to preserve the parts of information in the speech stream that are most informative. This assumption was laid down in the Hyper- and Hypospeech theory (H&H theory), which states that much of the variability of speech stems from the ways speakers adapt their speech to what they think that is needed by the listener to comprehend the message (Lindblom 1990). On the one hand, speakers want to be understood, and this output-oriented goal forces them to use hyperspeech. On the other hand, speakers do not want to spend too much energy on redundant parts of

speech, and this system-oriented, low-cost goal drives speakers to use hypospeech. In this way, speakers continuously estimate how much care of articulation is minimally needed or is permitted by the audience.

If speakers, for communicative reasons, do indeed speed up most during the least informative parts of the sentence, then lexically stressed syllables might be shortened less than unstressed syllables. In English the stressed syllable is the most informative syllable (Altmann & Carter 1989), and it is likely that the same goes for Dutch, cf. van Heuven & Hagman (1988). Furthermore, if the H&H principle of preserving the most informative parts holds, unaccented words might be affected more by an increase in speech rate than accented words. As unaccented words often refer to information already given, speakers might choose a higher speech rate during unaccented words than during new and highly informative accented words (but note that information value itself has no duration effect, only whether a word or a phrase is accented or not (Eefting 1991).

As noted above, artificial time compression of speech is normally performed in a linear way. One of the reasons why intelligibility and comprehension of artificially time-compressed speech breaks down at a certain playback rate may be its unnatural timing pattern. If we assume that the principle of nonlinear time compression introduced above is a strategic communicative principle, beneficial not only to the speaker but to the listener as well, then the intelligibility of artificially time-compressed speech might be improved if its temporal organisation is closer to that of natural fast speech. In other words, if speakers, or experimenters, i.e., by manipulation, deliberately assign a more prominent role to word-level prosody, this should be helpful to listeners.

Several studies have stressed the importance of prosody, and thus of temporal organisation, in word recognition and sentence processing in normal conditions. Cutler & Clifton (1984), van Heuven (1985) and Slowiaczek (1990) showed that word recognition is delayed when words in stress languages such as English and Dutch are deliberately mis-stressed. Cutler & Koster (2000) showed that stress information plays an important role in lexical activation in Dutch, and Cutler & van Donselaar (2001) also showed that listeners effectively use suprasegmental cues in Dutch. Under difficult listening conditions, prosodic factors seem to play an even more important role than they normally do (van Donselaar & Lentz 1994; Wingfield 1975; Wingfield et al. 1984). When the speech signal is degraded, prosodic information is usually preserved better than segmental information because it is spread over larger chunks of the speech signal. Furthermore, prosodic information is relatively well preserved in degraded speech also because the information (such as silence, pitch) is spread out over the entire spectrum. A degraded speech signal may therefore cause listeners to rely more on prosodic cues than when speech quality is high. Secondly, correct sentence-level phrasing is helpful in

the understanding of time-compressed speech (Wingfield et al. 1984). Wingfield (1975) showed that intelligibility of sentences with anomalous intonation and timing declined steeply as time compression increased, whereas the decline was much more gradual for sentences with a normal prosodic pattern. Wingfield explains this in terms of the correct intonation pattern adding redundancy to the speech signal: this redundancy can be exploited in difficult listening situations. Furthermore, van Donselaar & Lentz (1994) investigated the use of the interdependence between information and accent structure and how this is affected by speech intelligibility. Hearing-impaired subjects interpreted the accented words as being new, regardless of their information value. Only when speech quality was degraded, did the normal-hearing subjects switch to the default strategy of interpreting any accented word as being new: they also made use of the interaction between information and accentuation.

There is at least one study that seems to show that imitating natural fast speech timing leads to significant improvement over linear time compression at very heavy rates of time compression. The time-compression algorithm Mach1 (Covell, Withgott & Slaney 1998) is based on the compression strategies found in natural fast speech timing, such as compressing pauses most and compressing stressed (i.e., sentence-accented) vowels least. Covell et al. (1998) compared comprehension and preference for Mach1-compressed and linearly time-compressed speech. Mach1 not only offered significant improvement in comprehension over linear compression, but was also preferred by the listeners. However, it is not entirely clear what the contributions are of each of the nonlinear compression strategies. By compressing pauses most, the remaining sentence or paragraph of text can be compressed less, in order to be equal in total duration, than in the case of linear time compression. The lower articulation rate in the Mach1 compressed sentence is likely to make it more intelligible than the linearly time-compressed sentence. In other words, it is not clear what the separate contributions are of the word-level, sentence-level and paragraph-level changes in timing to the improvement in comprehension. In this study we will focus only on the word-level changes in timing between normal and fast speech rate. The main question is whether taking into account these changes in timing can improve the intelligibility of time-compressed speech.

For the present study, we assume that speakers behave in line with the H&H theory, and assign extra importance to the most informative parts when they are forced to speak fast. This leads us to the following hypotheses:

1. When speaking at increased rate, speakers will reduce lexically unstressed syllables more, relatively, than stressed syllables.

2. The durational correlate of pitch accent will become more prominent at faster speech rates because unaccented words (referring to 'given' information) are reduced more, relatively, than accented words (containing 'new' information).
3. Word-level intelligibility of time-compressed speech can be improved by taking into account the changes in temporal organisation going from normal to fast speech.

To investigate our hypotheses 1 and 2, we established how word-level timing is affected by an increase in speech rate in Dutch. This production study is described in section 4.2. Sections 4.3 through 4.6 describe perception experiments, that were set up to test the third hypothesis.

4.2 Experiment 1: Fast speech timing

4.2.1 Introduction

When asked to speed up their speech rate, speakers may have many ways to achieve this goal. It is assumed here that whatever speakers do when speaking fast, they will always choose a communicative strategy in accordance with the predictions we inferred from the H&H theory. Hence, speakers will speed up most during the least informative parts of their speech. For practical reasons, this study focuses on the shortening behaviour of stressed and unstressed vowels (and not syllables). This avoids the problem of resyllabification at syllable boundaries. Another practical advantage is that vowels are relatively easy to segment in the wave form. The main reason for measuring vowels is that increasing speech rate has its strongest effect on the duration of the vowels, as these are the most elastic segments. As the vowel's duration mainly determines the length of the syllable, we compared the difference between the shortening behaviour of stressed and unstressed syllables by looking at vowel durations.

In English, most unstressed vowels are reduced to schwa: lexical stress has a strong effect on the colour of the vowel. In Dutch, on the other hand, vowel quality is less dependent on the stress level of the syllable: the unstressed syllable may well contain a full vowel (van Bergem 1993; Kager 1989). In order to investigate whether there is a difference between the shortening behaviour of unstressed full vowels and schwa, we measured the durations of Dutch disyllabic words containing schwa and of words containing two 'full' vowels. This was done to ascertain that the syllables are reduced according to their stress level, and not because of their segmental quality.

4.2.2 Method

Material A set of 32 disyllabic nouns was selected in order to measure the durations of stressed and unstressed vowels: half of them with schwa, and half of them with 'full' unstressed vowels. Stress position and vowel length were balanced whenever possible. The target words are listed as Appendix E. The target words were embedded in long meaningful sentences, because pilot work had shown that it is easier to attain a high speech rate in longer sentences. The target words appeared at the beginning of the sentence to avoid final lenghtening. In order to make segmentation easier, the target words were selected such that the vowels were preceded and followed by plosive or fricative consonants where possible.

The sentences were recorded in two conditions: one in which the disyllabic target word had a pitch accent, and one in which the word was unaccented. A context sentence preceded the test sentence to indicate which words were to receive pitch accent in the following sentence. In the [-pitch accent] condition, the first adverb of the test sentence received pitch accent instead of the target word. The sentence structure was always the same, and so was the position of the target word.

Speakers Four female native speakers of Dutch were asked to read the test material aloud at normal and very fast speech rates. They were paid for their participation.

Procedure The recording session lasted about an hour and a half. First, the speakers were asked to read the material at a normal rate. If accentuation was not correct or if the sentence was not read out fluently (as judged by the experimenters), then the speaker was asked to repeat the sentence. After all material had been recorded at the normal rate, the speaker was asked to aim for a fast speech rate without abnormal slurring. Speakers were encouraged to use a stopwatch, so they got an impression of how fast they could speak, and they could try to outdo themselves in their speech rate. In order to increase the speech rate, they were asked to read out each sentence four or five times, and to keep an eye on the articulation time for each attempt. Again, the experimenters judged the speaker's performance. The material was recorded onto digital audiotape in a sound-treated cabin with a Sennheiser ME 10 microphone. The speech was then fed as digital input into a computer disk and downsampled to 16 kHz.

Duration measurements One version of each test sentence was selected on the basis of the correctness of the accent pattern. As there were about four versions of each

sentence pair in the fast rate condition, the fastest trial with a correct accent pattern was selected for analysis. The durations of the stressed and unstressed vowels were measured.

In Figure 4.1 below, two waveform graphs are displayed to show the difference between the normal (upper graph) and the fast version (lower graph) of a sentence fragment. The target word is *toffee* (/tɔfe/; 'toffee'), which is transcribed orthographically in the figure.

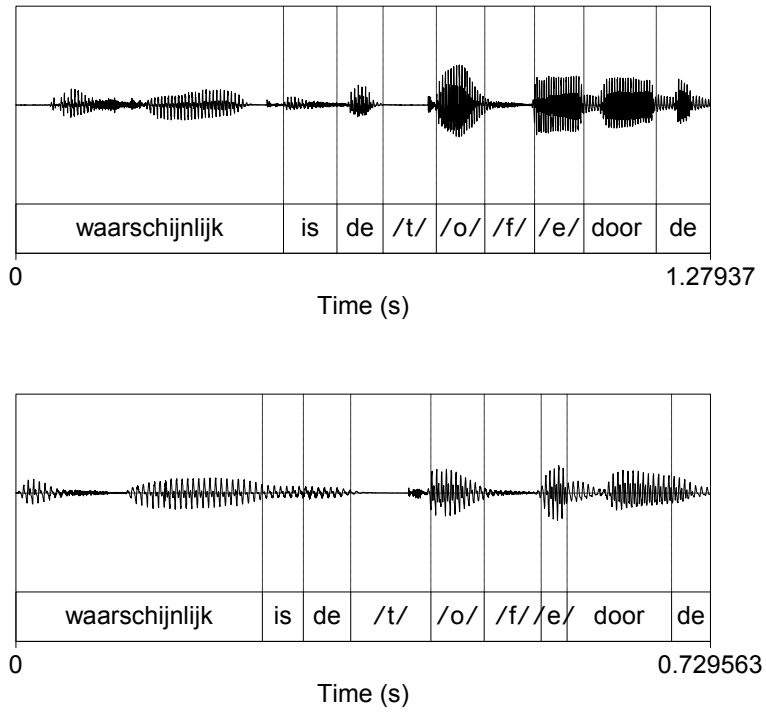

Figure 4.1. Waveform displays of sentence fragment including the target word *toffee* ('toffee') in the normal (upper graph) and fast-rate condition (lower graph), plus orthographic transcription.

Based on waveform and spectrogram displays, the criteria formulated below were used for the segmentation. In all cases, the segmentation was verified by auditory feedback. The vowel onset corresponded with the first positive zero crossing of the first periodic waveform at which an increased amplitude and a clearly visible change in the wave form due to a change in the harmonic structure occurred. The offset of the vowel corresponded with the positive zero crossing of the last periodic waveform before the following plosive or fricative started. If the target word ended in a vowel, the next word

always started with a plosive or a fricative. Some unstressed schwas were followed by a coda /r/ consonant (e.g., the target word *beker* 'beaker'). The durations of these vowels were very difficult to measure because of the short duration of the syllables containing schwa and because of the vowel-like articulation of /r/ in coda position. Segmentation was rather difficult in the fast speech because of heavy coarticulation. When no periodic vowel signal could be traced in the waveform and spectrogram, a minimum duration of 5 ms was postulated. This was established as a minimum duration because it corresponded to about one period (as the speakers were female with an average F_0 of about 200 Hz). Furthermore, this minimum duration of 5 ms enabled us to compute fast/normal ratios, which would have been impossible if we had assumed a duration of 0 ms.

All measurements were carried out by two undergraduate students in phonetics who checked each other's measurements. These measurements were then checked by the author. In most cases, the difference between the boundary locations was less than 10 ms. In case the measurements differed more than 10 ms (e.g., before /r/ or in the fast condition), the three judges decided on a 'compromise' duration. This procedure ensures a relatively high reliability of the vowel duration measurements.

Articulation rates were computed for the normal and fast speech rates by measuring the duration of the first part of the test sentence (containing the target word and up to the first major phrase boundary) and dividing it by the number of syllables (as counted in the canonical written version).

In Figure 4.2 the mean normal and fast rates are plotted for each speaker.

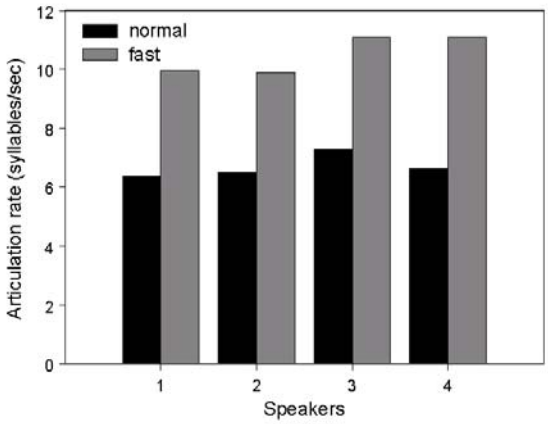

Figure 4.2. Mean articulation rates per speaker. Normal rate conditions are indicated by black bars, and fast rate conditions by grey bars.

The average articulation rate was 6.7 syllables/s in the normal speech rate condition, and 10.5 syllables/s in the fast condition. Paired-samples t-tests were carried out to investigate whether the mean fast articulation rate differed significantly from the mean normal rate, both in an analysis on speakers and on items. Speakers had a significantly higher articulation rate in the fast condition than in the normal condition ($t_1(3)=-16.6$, $p<0.001$); and all items were articulated significantly faster in the fast condition than in the normal rate condition ($t_2(31)=-49.0$, $p<0.001$).

4.2.3 Results

The mean durations (and standard error of mean) of stressed and unstressed vowels at normal and fast speech rate are shown in Table 4.1, together with the fast/normal ratios (vowel duration at fast rate / duration at normal rate).

Table 4.1. Mean durations (in ms) of stressed and unstressed vowels at normal and fast speech rate. Fast/normal ratios are also given.

	Normal rate (ms)		Fast rate (ms)		Fast/normal ratio
	Mean	S.E.	Mean	S.E.	
Stressed vowel	114	2.0	75	1.5	0.67
Unstressed vowel	55	1.5	21	1.0	0.42

At the fast speech rate, unstressed vowels were reduced to 42% of their normal rate duration, and stressed vowels were reduced to 67% of their normal rate duration.

We checked whether increasing speech rate had a similar nonlinear effect on the level of the entire syllable, and not only on the vowel durations. The syllable durations of the disyllabic target words in the [+pitch accent] conditions were measured of one speaker. In the fast rate condition, the stressed syllable was reduced to 64% of its normal rate duration, and the unstressed syllable was reduced to 45%. These data suggest that the entire syllable is reduced according to its stress level.

The fast/normal ratios for the vowel durations (within each item, per vowel) were analysed in two Repeated Measures ANOVAs on the 32 items and on the 4 speakers, with Stress (Stressed vs. Unstressed) and Accent (Accented vs. Unaccented) as fixed factors. The analyses show a significant effect of Stress on the fast/normal ratios ($F_1(1,3)=158.6$, $p=0.001$; $F_2(1,31)=64.0$, $p<0.001$).

Half of the disyllabic words contained two 'full' vowels, and these were balanced for vowel length. For this subset of items, the fast/normal ratios of the stressed and unstressed syllables show the same difference with ratios of 0.66 and 0.42 for the stressed and unstressed vowels, respectively. Thus, the first hypothesis is confirmed:

regardless of vowel length, unstressed vowels in fast speech are affected more, relatively, by an increase in speech rate than stressed vowels.

To make sure that syllables are reduced according to their stress level, and not because schwa may be more compressible than other 'full' unstressed vowels, the compression behaviour of the two types of unstressed vowels was checked. In Table 4.2 below the duration results are shown for the two types of unstressed vowels.

Table 4.2. Mean vowel duration of two types of unstressed vowel at normal and fast speech rate (plus fast/normal ratio).

	Normal rate (ms)		Fast rate (ms)		Fast/normal ratio
	Mean	S.E.	Mean	S.E.	
'Full' unstressed	66	2.1	24	1.5	0.42
Schwa	44	1.6	17	1.4	0.41

Obviously, the fast/normal ratio of unstressed 'full' vowels equals that of unstressed schwa vowels. The fast/normal ratios of the unstressed vowels were entered into Repeated Measures ANOVAs on items and on speakers, with Vowel Type and Accent as fixed factors. In the item analysis, the items were nested under Vowel Type (schwa vs. 'full' vowel). The effect of Vowel Type on the fast/normal ratios is not significant ($F_1(1,3)<1$, n.s.; $F_2(1,30)<1$, n.s.). Although the absolute duration of schwa was shorter on average than that of the 'full' unstressed vowels at both rates, schwa is not compressed more than 'full' unstressed vowels.

The second hypothesis was that vowels in words bearing a pitch accent reduce relatively less, in fast speech, than vowels in words without a pitch accent. In Figure 4.3 mean vowel durations of the stressed and unstressed vowels are shown, at both speech rates, and in [+pitch accent] and [-pitch accent] conditions.

The [-accent] vowels were expected to be compressed relatively more in fast speech than [+accent] vowels. This is not confirmed by the data: there even seems to be a trend in the opposite direction.

The analyses of variance with fast/normal ratios as the dependent variable, and Stress and Sentence Accent as fixed factors show that the main effect of Accent on the fast/normal ratios fails to reach significance ($F_1(1,3)=7.07$, $p=0.076$; $F_2(1,31)=5.62$, $p=0.024$).

Figure 4.3 shows a tendency for the lexically unstressed vowels to be affected more by the factor Accent, but this interaction between Stress and Accent was not significant in the item analysis ($F_1(1,3)=30.5$, $p=0.012$; $F_2(1,31)<1$, n.s.).

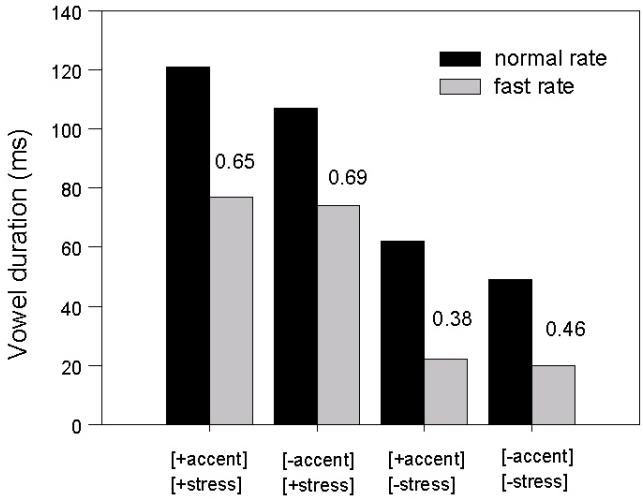

Figure 4.3. Mean vowel durations (in ms) of stressed ([+stress]) and unstressed vowel ([-stress]) in both [+pitch accent] and [-pitch accent] condition, at both speech rates. Fast/normal ratios are given above the paired bars.

In sub-analyses on the fast/normal ratios of the unstressed vowels only, with Accent and Vowel Type as fixed factor, the effect of Accent was not significant either ($F_1(1,3)=10.0$, $p=0.051$; $F_2(1,30)=2.71$, n.s.). Whereas other studies have found that accentual lengthening of vowels plays an important role at a normal rate, the present results show that this duration difference becomes relatively smaller at a fast speech rate. So, whereas the durational correlate of lexical stress becomes more prominent at faster speech rates, the durational correlate of sentence or pitch accent seems to become less prominent at increasing speech rate.

4.2.4 Discussion

As argued above, speakers tend to reduce the duration of unstressed syllables more than that of sentence-stressed syllables when speech rate is increased. Our aim was to find out whether speakers indeed show selective compression behaviour which is along the lines of the hyper- and hypospeech theory, namely that they speed up most during the parts which are least informative. Our first expectation was that speakers show a greater relative reduction in the duration of the unstressed syllable than of the stressed syllable. In a study on Dutch by Den Os (1988), increasing speech rate had a greater effect on stressed vowels than on unstressed vowels. However, Den Os (1988) may

have underestimated the relative shortening of unstressed vowels because the fastest unstressed vowels, which were too short to be measured, were disregarded in her study. The present results show that unstressed vowels are reduced more than stressed vowels, and that this is not an artefact of unstressed vowels often being schwa. The relative amount to which a vowel is reduced with increasing tempo depends mainly on the stress level of the syllable, and not on the quality of the vowel, nor on its pitch accent.

A similar non-uniform compression was expected for the sentence accented vs. unaccented words. Unaccented words were expected to shorten relatively more than accented words so that the most informative parts of the sentence are preserved. Vowel durations of disyllabic words were measured, which either did or did not have a pitch accent on the lexically stressed syllable. Contrary to the hypothesis, a trend was found for the relative duration difference between vowels in accented vs. unaccented conditions to become smaller at faster speech rate. In other words, the durational correlate of pitch accent becomes less prominent at faster speech rates. The durational correlate of pitch accent is sacrificed when the speaker is pressed for time. For lexical stress, the duration cue is the most important one. However, to indicate which words are accented in a sentence, the pitch excursion itself is a much more important cue than duration (Sluijter 1995; Sluijter, van Heuven & Pacilly 1997). One should note that the results concerning the durational aspect of accent are strongly linked to the design of the duration study: the time compression of disyllabic content words in [+pitch accent] and [-pitch accent] condition was compared. If the reduction of these content words had been compared with the reduction of function words, such as articles or auxiliary verbs in the same phrase, some important differences in phrase level timing might have been found.

In summary, increasing speech rate is accompanied by important changes in word-level timing in Dutch. The next section deals with the question whether taking these timing changes into account can improve the word-level intelligibility of time-compressed speech. In the Introduction section, several studies were mentioned that showed that the role of prosodic factors becomes more important under difficult listening conditions because prosodic information is preserved better than segmental information (van Donselaar & Lentz 1994; Wingfield 1975; Wingfield et al. 1984). Thus, in the next section a perception experiment is described to test the hypothesis that the more salient word-level prosodic pattern found in natural fast speech is helpful to listeners who are presented with artificially time-compressed speech.

4.3 Experiment 2: Intelligibility of words and nonwords after linear or nonlinear time compression

4.3.1 Introduction

The duration study described in the previous section shows that the prosodic pattern at word level is made more pronounced with increasing speech rate. These production data then lead to the expectation that the intelligibility of time-compressed speech will be improved if its temporal organisation is closer to that of natural fast speech. Experiments in our laboratory have shown that speech remains intelligible at rates that are much faster than can ever be attained in natural fast speech. Speech that is time-compressed to the fastest rate which human speakers can achieve is still almost perfectly intelligible. It would seem reasonable to evaluate the perceptual effects of applying fast speech timing to time-compressed speech at the fast rate which is produced by the speakers. However, the perceptual effects of more natural fast speech patterns will first be established for a much faster rate of speech. There are two reasons for this. First, a practical reason is that intelligibility of artificially time-compressed speech is very high, even at rates twice the normal rate. This ceiling effect would make any intelligibility differences between linearly time-compressed and nonlinearly time-compressed speech difficult to find. Second, a more fundamental reason is that the role of prosody is expected to become more important as the listening situation becomes more difficult. The information carried by the more salient prosodic pattern might be exploited in difficult listening situations. For these two reasons, the rules of fast speech timing were extrapolated to even faster rates.

As argued above, if the prosodic/temporal pattern is assigned a more prominent role in speech production at fast rates, such fast speech timing will also become more helpful in the perception of fast speech. Speakers are expected to speed up most during the parts that are least informative, and preserve the more important parts. Yet, on the other hand, at very fast rates of speech, prosody and segmental information play conflicting roles. Prosody requires that some syllables are longer and more prominent than others. Weak unstressed syllables will therefore be the first to become highly unintelligible after time compression, even more so when these syllables are compressed more than stressed syllables. Cutler & van Donselaar (2001) show that, although Dutch listeners make use of the suprasegmental cues in word recognition, the contribution of segmental information probably outweighs that of suprasegmental information. We should also consider the possibility that the speakers' nonlinear way of speeding up is not so much caused by a communicative strategy, but is rather caused by articulatory factors. Possibly, speakers simply cannot speed up in an approximately

linear fashion. Furthermore, there is some evidence that speakers tailor their utterances to internal representations of the listeners' needs, *except* under time or task pressure (Horton & Keysar 1996). This would mean that natural prosodic rules do not necessarily contribute to speech intelligibility. Thus, an alternative possibility is that segmental intelligibility plays such an important role that listeners are helped more, paradoxically, by an entirely unnatural compression strategy, namely by compressing the lexically stressed syllable relatively more than the lexically unstressed syllable (which is short already). This would preserve the segmental intelligibility of both syllables. Three strategies for time compression need to be considered to evaluate these possibilities. An experiment was set up to compare the intelligibility of speech after Linear Compression (compressing all syllables to the same degree); after Selective Compression based on natural fast timing (compressing unstressed syllables relatively more than stressed syllables); and after Unnatural Compression (the reverse of Selective Compression: compressing stressed syllables more than unstressed syllables).

The hypothesis is that the word-level intelligibility of strongly time-compressed speech can be improved by taking into account natural fast speech timing which assigns more importance to the most informative parts in the speech stream (i.e., Selective Compression). However, for the identification of nonwords, only segmental intelligibility counts. Thus, for nonwords, the hypothesis is that listeners are helped more by an entirely unnatural compression strategy, in order to preserve the segmental intelligibility of both syllables.

A competing hypothesis is that word-level intelligibility, both of real words and nonwords, is not improved by making its timing more like that of natural fast speech because the change in timing is not a communicative strategy, but due to articulatory restrictions.

4.3.2 Method

The intelligibility of words and nonwords was tested in the three compression conditions to study the effect of the compression conditions on word recognition and on non-word identification. The report of this experiment will be brief because the results did not show significant differences between the three time-compression conditions. The experiment was later rerun with different material; this experiment will be described more elaborately in the section 4.4 below.

To test the effect of the three ways of time compression on intelligibility, 48 disyllabic words were chosen: half of them with initial stress and half of them with final stress. The intelligibility of 48 phonotactically legal nonwords was also tested in the three compression conditions. The words and nonwords were embedded in two types

of carrier phrase: normally the carrier phrase *Je moet ... typen* ('You must ... type') was chosen. If the word ended with /t/ the carrier phrase *Je moet... schrijven* ('You must ... write') was chosen. A male native speaker of Dutch was asked to read the words and nonwords in the carrier phrases. Below the three compression conditions are given.

1. linear time compression (LC): compress lexically stressed and unstressed syllable to the same degree;
2. selective time compression (SC): global imitation of fast speech timing: compress stressed syllable less (to 65% of the normal rate duration) than unstressed syllable (to 40%);
3. unnatural time compression (UC): opposite to (SC): compress stressed syllable more (to 40%) than unstressed syllable (to 65%).

The PSOLA time-scaling technique (PSOLA), as implemented in the speech editing program GIPOS (version 2.3; http://www.ipo.tue.nl/ipo/gipos/), was used to time-compress the speech fragments. In GIPOS, selected parts of a speech waveform can be time-compressed, while, at the same time, the remainder of the speech signal remains unaffected. The sentence can thus be time-compressed fragment by fragment. The carrier phrase was compressed linearly (in all conditions) to 35% for the real words, and to 45% for the nonwords. After the selective or unnatural time compression was applied to the target word, the target words were compressed linearly even further to 35% of their original duration (for the LC condition speech was time-compressed linearly to 35% of its original duration). For the nonwords, the overall compression rate was 45% in all three compression conditions, because an earlier experiment had shown that identification scores were extremely low at compression to 35%.

First, all real words were presented to the listeners. To conceal the fact that all test items were disyllabic, 24 monosyllabic filler words were added to the test material. A practice set of 12 sentences preceded the actual test. Secondly, the nonwords (plus monosyllabic fillers) were presented, preceded by 12 practice sentences. Items were presented in random order to control for a possible adaptation effect during the experiment. The carrier phrase was first presented visually on the screen without the target word. Then the carrier phrase, including the target, was presented over closed headphones, and subjects were asked to fill in the missing word by typing on a keyboard. After they had pressed the Enter key, the next stimulus was presented. In both experiments, the three compression types were balanced over the 48 test words, and over the 2 stress positions (Latin square design). There were three experimental lists for both experiments. Each list was presented to 14 listeners, so that 42 subjects

participated in the listening experiments. The subjects, all students of Utrecht University, were given a small payment for their participation.

4.3.3 Results

The raw correct recognition percentages are shown in Table 4.3. Note that real words were compressed (overall) to 35% of their original duration and the nonwords to 45%.

Table 4.3. Percentages of correct identification in conditions Linear Compression (LC), Selective Compression (SC) and Unnatural Compression (UC).

	LC	SC	UC
Real words (compressed to 35%)	54	54	53
Nonwords (compressed to 45%)	37	25	28

The data in Table 4.3 suggest that there is no effect of the compression conditions on the recognition of the real words. When the real-word results are broken down by stress position, however, a different picture emerges (Figure 4.4).

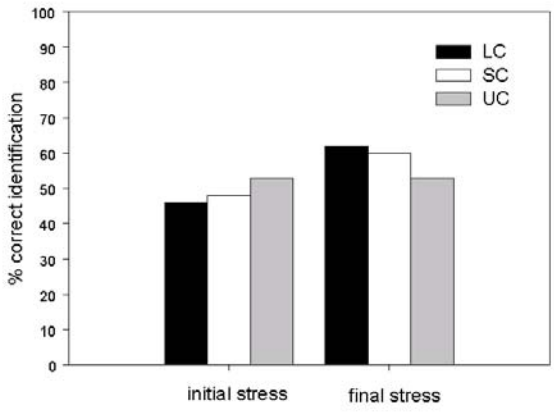

Figure 4.4. Percentage of correct recognition of real words in three time-compression conditions, broken down by stress position.

Repeated Measures analyses of variance were carried out on the percentages of correct recognition (after arcsine transformation), with Compression Type and Stress position as fixed factors, and either subjects or items as a random factor (items nested under Stress position). There was no main effect of Compression Type ($F_1(2,40)<1$, n.s.; $F_2(2,45)<1$, n.s.). The subject analysis showed a significant main effect of Stress

position ($F_1(1,41)=13.86$, $p=0.001$), but this was not significant in the item analysis ($F_2(1,46)=1.23$, n.s.). The interaction between Compression Type and Stress position, which can be seen in Figure 4.4, is only significant in the analysis on subjects ($F_1(2,40)=6.02$, $p=0.005$; $F_2(2,45)=2.44$, $p=0.098$).

Obviously, the differences in intelligibility in the three types of time compression were very small. An analysis was carried out to test whether intelligibility was affected by word frequency. However, an item analysis with Compression Type (all three levels) as the fixed factor, with items nested under Stress position, and word frequency (log) as a covariate did not show a significant effect of word frequency ($F_2(1,45)= 2.16$, n.s.).

The identification results of the nonwords, broken down by stress position, are shown in Figure 4.5.

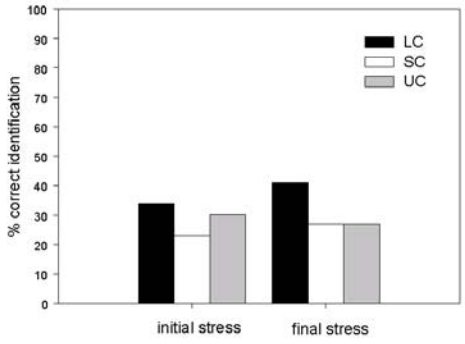

Figure 4.5. Percentages of correct identification (entire non-word correct) for nonwords with initial stress and final stress, in each of the Compression conditions.

This figure shows that Linear Compression yielded the best intelligibility for the nonwords, regardless of stress position. The non-word results (% of correct identification, after arcsine transformation) were fed into Repeated Measures ANOVAs, with Compression Type and Stress as fixed factors. The main effect of Compression Type was significant ($F_1(2,40)=14.5$, $p<0.001$; $F_2(2,45)=4.48$, $p=0.02$). The effect of Stress position was not significant ($F_1(1,41)=2.56$, $p=0.12$; $F_2(1,46)<1$, n.s.), and the interaction between Compression Type and Stress position was not significant either ($F_1(2,40)=1.7$, $p=0.19$; $F_2(2,45)<1$, n.s.).

In order to have a closer look at the effects of the three compression types, percentages of correct syllable recognition are shown in Table 4.4.

Table 4.4. Correct syllable identification (%) of both syllables of real words and nonwords with either initial or final stress, per Compression Type.

Lexical Status, Stress	Syllable	LC	SC	UC
Real words, initial stress	S1 [+stress]	59	76	57
	S2 [-stress]	61	52	74
Real words, final stress	S1 [-stress]	66	61	64
	S2 [+stress]	75	76	63
Nonwords, initial stress	S1 [+stress]	66	70	45
	S2 [-stress]	66	46	74
Nonwords, final stress	S1 [-stress]	66	47	83
	S2 [+stress]	71	77	55

Certain confusions within the syllable were allowed for the nonwords: place of articulation of nasals, and voicing value of fricatives and plosives. Table 4.4 shows that, for the nonwords in particular, Selective Compression improved the identification scores of the stressed syllable (albeit at the cost of the unstressed syllable), and Unnatural Compression improved the identification scores of the unstressed syllable (at the cost of the stressed syllable). The net result was that neither SC nor UC improved the intelligibility of the nonwords over LC: the positive effect on the one syllable is always outweighed by an equally large or even larger negative effect on the other syllable. Although the results for the real words are not entirely predictable from the length of the syllable due to lexical factors, the same overall trend applies: the longer the syllable duration remains after compression, the higher the segmental intelligibility.

The different ways of time compression did not appear to have important effects on the word-intelligibility of the real words. One possible reason for this unexpected result might be that the carrier phrase in which the target words were embedded provided a context which was prosodically not strong enough for prosodic manipulations to have an effect. First, the carrier phrase was compressed linearly in all three compression types. Secondly, words spoken in short carrier phrases almost behave as if they are spoken in isolation; the normal timing relations within a disyllabic word are strongly affected by final lenghthening. If the three compression conditions had been applied to normal meaningful sentences as a whole, the differences between the three types of conditions might have been clearer.

The results of the present experiment show that segmental intelligibility is affected by syllable duration (cf. Table 4.4). The absolute duration of a syllable after compression has a greater contribution than the prosodic pattern of normal rate speech (LC) or that of fast speech (SC). The data presented in Table 4.4 also showed that the intelligibility of the real words cannot be predicted from the results of the nonwords.

Apparently, this is due to lexical factors. Whereas the unnatural compression condition did not improve the identification rate of the nonwords and of the words with final stress, it slightly improved the recognition scores of real words with initial stress. This interaction between the effect of UC and Stress position for the real words might indicate that the segmental intelligibility of the unstressed syllable is more important for the recognition of words with initial stress than for words with final stress. Figure 4.4 shows that recognition scores for words with final stress were higher overall than for words with initial stress. As initial stress is the default stress pattern in Dutch, words with final stress have fewer neighbours. Following the definition of neighbourhood by Luce & Pisoni (1998), the mean number of neighbours for our set of words with initial stress was 1.5, whereas it was only 0.5 for the words with final stress (more than half of them did not have any neighbours at all). Identifying the stressed syllable alone may more often result in correct recognition in words with final stress than in words with initial stress. The ratio of the number of segments in the stressed syllable divided by the number of segments in the unstressed syllable is also higher in the words with final stress (mean ratio is 1.4 for words with final stress vs. 1 for initial stress). This might also give more weight to the identification of the stressed syllable for words with final stress than for words with initial stress.

Since the word-intelligibility results for the real words were about the same for the three time-compression conditions, they are rather inconclusive. The material used in this experiment provided a prosodically unviable environment. The three types of time compression will be tested again with different material in the next section.

4.4 Experiment 3: Three ways of time compression

In this section a second perception experiment is described, set up to test the hypothesis that the more salient word-level prosodic pattern found in natural fast speech improves word intelligibility of artificially time-compressed speech.

4.4.1 Method

Short sentences were constructed containing a target word which was to be identified in an intelligibility test. The short sentences were often the first clause of a longer sentence. The target words were of low semantic predictability in the sentence. The three types of compression were applied to the entire sentences: each syllable was assigned a plus or minus stress mark, and was then time-compressed accordingly. The

broad distinction between function words and content words was used as a criterion to assign a stress level to each syllable. Auxiliary verbs and articles were assigned [-stress], whereas the main verb and other content words were assigned [+stress]. For polysyllabic words, only the stressed syllable received a [+stress] mark. Example sentences are presented in (1): the target word is in bold.

(1) Hij+ had- de- **par-tij+** moe+ten- ver-nie+ti-gen-
('He should the **batch** have destroyed')
Het- pak-ket+ bleek+ **me-taal+** te- be-vat+ten-
('The package appeared **metal** to contain')
Ook+ is- er- een- **mo-del+** te- vin+den-
('Also is there a **model** to be found')

As in the previous experiment (cf. section 4.3), PSOLA was used to time-compress the speech fragments. For the selective compression condition, syllables with [+stress] marks were compressed less (i.e., to 65%) than [-stress] syllables, which were compressed to 45% of their original duration.[14] For the unnatural compression the compression strategy based on the plus and minus marks was reversed such that the [-stress] syllables were compressed less (i.e., to 65%) than the stressed syllables (i.e., to 45%). After the nonlinear compression, the entire word and sentence durations were measured and the word duration and the rest of the sentence were linearly compressed further to attain a compression rate of 35% (a pilot experiment with this material had shown that only at compression to 35% of the original duration the intelligibility scores would be around 50% correct). This was done separately for the target words, such that the target word duration would be the same in all three compression conditions. For the linear compression condition, all syllables were compressed to the same degree.

There were 144 monomorphemic disyllabic targets (all nouns); embedded in sentences. Half of them had initial stress, and half had final stress. Since each subject could be presented with the same item only once, there were three experimental lists. On each list, the three compression conditions were balanced over the 144 target words, and over the 2 stress positions (Latin square).

[14] The production study had shown that, in fast speech, stressed vowels were reduced to 67%, and unstressed vowels to 42% of their original normal-rate duration. In the first perception experiment, this nonlinear compression was translated into 65% vs. 40%. However, the measurements on the entire syllable durations indicated that stressed syllables were reduced to 64%, and unstressed syllables to 45% of their original duration. In the present experiment, therefore, the nonlinearity was somewhat less extreme: stressed syllables were reduced to 65% and unstressed syllables to 45% of their normal-rate duration.

Subjects To each of the three experimental lists, 11 subjects were assigned. The 33 subjects were tested individually in a sound-treated booth. The speech material was presented to them over closed earphones. They were all students at Utrecht University, and were paid a small amount of money for their participation.

Procedure A practice session of 20 items preceded the actual test session so that the subjects could adapt to the fast speech rate before the test began. The order of the items was randomised for each subject to cancel out a possible learning effect during the test. Monosyllabic fillers and filler targets with three syllables were interspersed in the material, so that subjects would not notice that all test words were disyllabic. First the entire sentence was presented on the screen, with a blank at the position of the target word plus its accompanying article. The article was also left out because the definite article in Dutch provides information about the grammatical gender of a word. Subjects were given sufficient time to read the visual presentation. After 3 seconds, the whole time-compressed sentence was presented to them auditorily, including the target word. Subjects had to fill in the missing word by typing on a keyboard. There was no time pressure: only after they had hit the Enter key, the following sentence would appear on the screen. The entire experiment lasted about 30 minutes.

4.4.2 Results

The percentages of correct responses per condition are shown in Figure 4.6.

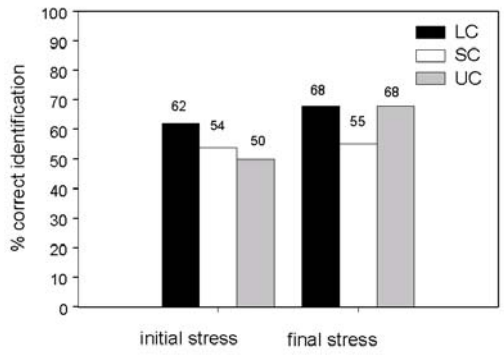

Figure 4.6. Percentages of correct word recognition, broken down by Type of Compression and Stress Position.

The analyses of variance (on percentages of correct word identification, after arcsine transformation) show that the main effect of Stress position does not reach significance

($F_1(1,32)=63.20$, $p<.001$; $F_2(1,142)=3.35$, $p=0.069$). Figure 4.6 suggests that words with final stress are, on the whole, somewhat easier to recognise. This insignificant difference may be due to a sparse neighbourhood effect: words with final stress tend to have fewer close neighbours in the Dutch lexicon and hence may be recognised more easily. The main effect of Compression Type is highly significant ($F_1(2,31)=15.4$, $p<.001$; $F_2(2,141)=13.2$, $p<0.001$). Figure 4.6 shows that Linear compression yields the highest intelligibility. Furthermore, there is a significant interaction between the effect of Compression Type and Stress position ($F_1(2,31)=10.33$, $p<.001$; $F_2(2,141)=6.89$, $p=0.001$). Making the prosodic pattern less pronounced (i.e., UC compression) decreases the intelligibility of words with initial stress, but it does not have a negative effect on the intelligibility of words with final stress. This pattern of results for finally stressed words can be explained as the outcome of two effects, working in opposite directions. The first effect is that of temporal alignment. The second effect is related to the duration of the stressed and most informative syllable. Regarding alignment, word recognition has a left-to-right aspect to it because speech unfolds over time. Misstressing initially-unstressed Dutch words is more disrupting than misstressing initially stressed words (van Leyden & van Heuven 1996). If the unstressed (word-onset) syllable of a finally stressed word is relatively long, it may be easier to start the correct alignment with possible word candidates at the start of the unstressed syllable because the unstressed syllable is relatively salient in the UC, compared to the LC condition. Secondly, the duration of the stressed syllable is shorter in the UC condition than in the other two conditions. So, the positive effect of the UC condition on the initial alignment of finally stressed words against word candidates is counterbalanced by the short duration of the stressed syllable.

The same two effects also explain the intelligibility pattern observed for initially stressed words. Initial alignment with possible word candidates is more difficult because the first syllable is shorter. As the first syllable is in this case also the stressed syllable, intelligibility is only affected negatively.

Further confirmation for these two tendencies can be found in the distribution of incorrect responses. A closer study of the error responses shows that in the majority of the false responses, either the correct stress pattern was reported, or subjects responded with a monosyllabic (stressed syllable) answer. The selective compression condition is expected to make correct alignment of word-onset rather difficult for words with final stress, because of the very short duration of the first unstressed syllable. The percentages of monosyllabic responses for the three types of compression are shown in Table 4.6.

Table 4.6. Percentages of monosyllabic responses to disyllabic stimuli, broken down by stress position and compression type.

	LC	SC	UC
Initial stress	15	17	14
Final stress	12	23	8

The percentages of monosyllabic responses, in all three compression conditions and in both stress conditions, were arcsine transformed. These transformed data were fed into analyses of variance with either item (nested under Stress position) or subject as random variable, and Compression Type and Stress position as fixed factors. There was no significant effect of Stress position on the percentage of monosyllabic responses ($F_1(1,32)=2.99$, $p=0.093$; $F_2(1,142)<1$, n.s.). The effect of Compression Type, however, was significant ($F_1(2,31)=22.1$, $p<0.001$; $F_2(2,141)=14.7$, $p<0.001$), and so was the interaction between Stress position and Compression Type ($F_1(2,31)=9.15$, $p=0.001$; $F_2(2,141)=7.56$, $p=0.001$). Separate post-hoc t-tests were carried out to compare the mean percentages of monosyllabic responses in the Linear and Selective Compression condition. For the initially stressed items, the difference between SC and LC was not significant ($t_1(32)=1.98$, $p=0.056$; $t_2(71)=1.04$, n.s.). For the finally stressed items, SC elicited significantly more monosyllabic responses than LC ($t_1(32)=4.30$, $p<0.001$; $t_2(71)=5.00$, $p<0.001$). In the SC condition, the unstressed syllable is reduced to such a short duration that, in some cases, it may be perceptually obliterated. Acoustically, there is something left of the syllable, but perceptually these very short syllables may almost 'fall out' of the signal. This is most often the case for words with final stress. Alignment with correct word candidates clearly fails here because of the very short duration of the unstressed first syllable.

For both words with initial and final stress, making the prosodic pattern of the disyllabic target words more like natural fast speech timing (SC) does not improve intelligibility over LC. Giving priority to the segmental intelligibility of the unstressed syllable (UC) does not improve intelligibility either.

The importance of correct initial alignment with word candidates agrees with some sort of word-beginning superiority effect (Nooteboom & van der Vlugt 1988). Further evidence for such a word beginning superiority effect comes from a study by Cutler & Clifton (1984) who studied the effect of misstressing of disyllabic words with two full vowels. It turned out that these misstressed items were only harder to recognise if their citation form pronunciation had initial stress. Thus, *nutMEG* was harder to recognise than *NUTmeg*, but *TYphoon* was not significantly more difficult than *tyPHOON*. A prosodic mismatch in the first syllable was more disruptive to the recognition process

than one in the second syllable. This can also be explained in terms of correct initial alignment.

Overall, one can conclude that linear compression wins. One possibible explanation for these results could be the extremely fast speech rate. It is possible that selective compression would in fact have improved intelligibility at speech rates humans can achieve, but not at the very fast rate employed in our study. We extrapolated the changes in timing from the moderately fast speech rate reached by our speakers to a much faster speech rate. Apart from the practical reason of avoiding ceiling effects in intelligibility, this was also done because we expected a degraded speech signal to cause listeners to rely more on prosodic cues than when speech quality is high. This very fast rate cannot be attained by human speakers, but listeners are still quite capable of filling in the missing words. In the next section the perception experiment is rerun with the speech-interference technique, to investigate whether selective compression can improve intelligibility at a moderately fast rate.

4.5 Experiment 4: The speech-interference technique

The previous section has shown that, at a very heavy rate of time compression, word-level intelligibility is not improved by either type of nonlinear time compression. In this section the three ways of time compression are compared at the fast speech rate speakers attained in our duration study (experiment 1; section 4.2): i.e., time compression to 65% of the normal duration. The prediction is that applying natural fast speech timing to time-compressed speech improves the speech quality at moderately fast speech rate. To test this, two experiments will be described. First, a speech-interference experiment is described in section 4.5. Section 4.6 reports on a phoneme detection experiment (experiment 5). Conclusions will be presented in section 4.7.

4.5.1 Method

The speech-interference technique, as developed by Nakatani & Dukes (1973), involves presenting speech at different S/N ratios in order to find out at which S/N ratio the speech is 50% intelligible. The technique was developed to be an indication of the quality of several types of synthetic speech, relative to a certain reference condition (i.e., natural speech). Nakatani & Dukes (1973) and Eggen (1992) used interfering speech as the masker, read by the speaker who also read out the test materials. Nakatani & Dukes (1973) argued that the sensitivity of the speech interference test increases by making the

speech masker and the test stimulus perceptually more similar. As the contributions of the word-level timing were the central issue in this research, the target word should, in all three compression conditions, be equally affected by the interfering noise. The masking sound should not fluctuate randomly in intensity as this might confuse the intelligibility in a random way. Therefore, in the present experiment USASI noise was used, which has the long-term spectrum of speech. The USASI noise was made by filtering white noise (by combining the effects of two filters: one highpass with a cut-off frequency of 100 Hz and one lowpass with a cut-off frequency of 320 Hz) with a –6 dB/octave slope. As the stressed and unstressed syllables are unequal in amplitude, the intensity curve of the speech signal was superposed onto the noise, such that the S/N ratio at each point in time would be constant. An intensity contour was computed for each time-compressed sentence, and this contour was then multiplied with the noise signal. These two signals (speech and noise) were to be presented to the subjects at a particular S/N ratio. Four S/N ratios were considered to be a minimum requirement for a relatively accurate estimation of the 50% intelligibility point.

The sentence material of section 4.4 was also used here: 144 sentences each containing a disyllabic target word. After the target word and the rest of the sentence were compressed to 65% or 45%, according to their stress level, the speech was expanded again to an overall compression rate of 65% of its original duration. This procedure did not result in audible artefacts.

Half of the items were to be presented as the reference condition (i.e., at the original rate), in four blocks of 18 sentences (9 items with initial and 9 with final stress). Each block had a particular S/N ratio. The other half of the items were presented in one of the three time-compressed conditions, also in four blocks of 18 sentences, each block at a particular S/N ratio. In a completely balanced within-subject design, each subject would have to be presented with time-compressed and uncompressed reference material, both at four S/N ratios, and with all three time-compression conditions for the fast material. This would amount to 24 conditions (2 x 4 x 3=24). Such a balanced design would then require much more sentence material than the 144 test sentences that were available here. Therefore, a between-subjects design was chosen with type of compression as the between-subjects factor.

There were three experimental lists, and for each list it could be computed at which S/N ratio the time-compressed and the reference speech is 50% intelligible. Quality measures are defined as the difference in 50% intelligibility S/N ratio (in dB) between the uncompressed reference speech and the compression condition of that particular list. The reference part was the same for all subjects: they all heard the same target items presented in the same four blocks of a particular S/N ratio. The test

(compressed) items were presented in four blocks, each with its own particular S/N ratios. These S/N ratios were the same for all three lists.

In a pilot experiment, the speech was presented at several S/N ratios in order to estimate at which S/N ratios identification would be around 50% correct.

Ten subjects were assigned to each of the 3 lists. The 30 subjects were all students from Utrecht University and were paid a small amount for their participation. They had not participated in the intelligibility test of section 4.4.

4.5.2 Results

For each subject, 50% intelligibility points were computed for the reference condition and for the test condition. First, the raw correct identification percentages for the four blocks of target items presented at different S/N ratios were computed. Then a linear trend line was computed to fit the four data points. The trend line equation could be used to compute the exact 50% intelligibility point. The mean 50% intelligibility points were computed for each list, together with their 95% confidence intervals. The Q measure for each experimental condition was computed by subtracting the intelligibility point in the test condition from that of the reference condition. The Q measures (with their 95% confidence intervals) for the three types of time compression are presented in Figure 4.7.

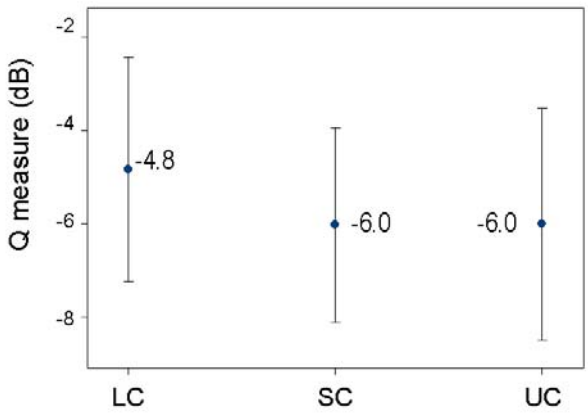

Figure 4.7. Quality measures (with 95% confidence intervals) for the three types of time compression.

Figure 4.7 shows that the LC condition has the smallest absolute Q measure: it has the smallest decrease in intelligibility from the reference to the test condition. Yet, given the small differences between the three Q measures (about 1 dB) and the relatively large confidence intervals, the three conditions are statistically indistinguishable. The intelligibility differences found at very fast speech rate cannot be demonstrated with the speech-interference technique. This leaves open two options. Either this technique is not sensitive enough to show the differences between our three conditions, or there are no quality differences between the three conditions at this moderately fast speech rate.

In the next section the question whether selective compression would in fact improve intelligibility at the fast rate which humans can achieve is addressed again, using a different experimental technique. As long as this question is unanswered, we do not know whether the speakers' 'strategy' of reducing stressed syllables less than unstressed syllables, is beneficial to perception. As noted before, the speakers' 'strategy' seems to be completely in line with the predictions of the Hyper- and Hypotheory. Still, if imitating the speakers' strategy appears to hamper perception, there may be other, perhaps articulatory, reasons underlying the speakers' way of speeding up.

4.6 Experiment 5: A phoneme detection study

Reaction time experiments can be used as a measure of the speed of processing of highly intelligible speech types. Now that synthetic speech often approaches the intelligibility of natural speech, it is useful to turn to more on-line speech processing measures. Pisoni (1981) found a disadvantage in naming time and auditory lexical decision time for synthesised words, relative to naturally produced words. Pisoni (1981; 1987; 1997) argued that this is due to a greater phonetic processing difficulty. Phoneme detection time is another on-line measure of the speech quality, or ease of processing, of highly intelligible speech types (Nix et al. 1993). Furthermore, phoneme detection has also been used to compare the processing speed of spontaneous speech and read speech (Mehta & Cutler 1988). These two types of speech mainly differ with respect to prosodic characteristics. The fact that phoneme detection time differences have been found for types of speech which are prosodically different leads to the expectation that by using this method we might also be able to find significant differences between word perception of linearly versus nonlinearly time-compressed speech. The hypothesis is that applying fast-speech timing to time-compressed speech improves word-perception.

4.6.1 Method

The sentence material was a selection of the material used in sections 4.4 and 4.5. Only 81 sentences of the original material (144 sentences) contained suitable target words for phoneme detection. The target words should have a word-initial plosive consonant, which, ideally, does not occur anywhere else in the sentence. The material did not contain enough target nouns which met these criteria. Thus, other word classes were used as well: the target items were nouns, verbs and adverbs.The compression rate was 65% (speech is compressed to 65% of the original duration), since this is an approximation of the speed-up factor that speakers can attain when they are asked to speak very fast.

The same three types of time compression, i.e., Linear Compression, Selective Compression, and Unnatural Compression, were used as in section 4.4. Details concerning these three types of compression can be found in section 4.4. The target words were equally long in all three conditions: the difference lies in the durations of the stressed and that of the unstressed syllables.

Apart from the 81 test sentences, 60 catch trials were interspersed with the material to keep subjects from pressing the button randomly. The 81 test sentences were rotated over the three time-compression condition conditions and were distributed over three lists in a Latin square design. Each subject could only be presented with the same sentence once. Time markers had been placed in the audiofiles at the start of the silent interval of the target plosive (or the start of the voice bar for voiced plosives). During the experiment, reaction times could be computed by subtracting this marker time from the time until the button press was registered.

Subjects were given written instructions on their task during the experiment. Subjects would first see a plosive phoneme on the computer screen in front of them. This was the sound they were supposed to monitor during the sentence which was presented to them over headphones. They were asked to press the button as soon as possible whenever the assigned target phoneme occurred word-initially. They were also told that there would be catch trials in which they were not supposed to push the button.

There were 10 practice items, after which subjects could still ask questions cocnerning the producedure if anything was unclear. After the subjects had resumed the test, 10 warming-up filler items were presented to make sure that subjects were warmed up before the actual test began. All test and filler items were presented in random order.

Ten subjects were assigned to each list, so that 30 subjects participated in the experiment. They were all students of Utrecht University and received a small payment for their participation

4.6.2 Results

There were a number of negative reaction times, indicating that subjects had responded to another initial plosive or to an earlier non-initial occurrence of the plosive. These were considered missing observations, together with those instances in which subjects had not pressed the button because they did not detect the target plosive (in time). All subjects agreed that the speech was highly intelligible. The average raw reaction times (i.e., computed over all valid observations only) in each of the three time-compression conditions are shown in Table 4.7, together with standard errors of the mean and miss rates.

Table 4.7. Mean raw reaction times plus standard errors in all three time-compression conditions. The miss rates are also shown.

Compression Type	Mean RT (msec)	s.e.	Miss rate (%)
Linear Compression	537	9	6
Selective Compression	576	11	8
Unnatural Compression	557	10	8

For the statistical analysis of the results, all missing observations were replaced by the grand mean (554 ms).[15] The data were then entered into analyses of variance (with either subjects or items as repeated measures) to establish the effects of Compression Type and Stress Position. The main effect of Compression Type was significant ($F_1(2,28)=6.7$, $p=0.004$; $F_2(2,78)=3.8$, $p=0.027$), and so was the interaction between Stress position and Compression Type ($F_1(2,28)=6.5$, $p=0.005$; $F_2(2,78)=5.3$, $p=0.007$). There was no main effect of Stress position ($F_1(1,29)<1$, n.s.; $F_2(1,79)<0$, n.s.).

An interaction between Compression Type and Stress Position was also found in the intelligibility results at a very fast speech rate (section 4.4): Selective Compression mainly lowered the intelligibility of the finally stressed items relative to LC, whereas Unnatural Compression only lowered the intelligibility scores for the initially stressed items. In Table 4.8 the phoneme detection results are broken down by Stress position to investigate whether indeed the same interaction is found here. There were 35 initially stressed items, and 46 items with non-initial stress.

[15] Although not very sophisticated, this is a practical solution. The reported effects, however, are large enough to warrant this approach (i.e., p values are small enough).

Table 4.8. Mean detection times in all three compression conditions, broken down by stress position. The schematic durations of first and second syllable are given as well.

	Initial stress (N=35)		Non-initial stress (N=46)	
	Detection time	Schematic duration	Detection time	Schematic duration
Linear Compression	532	____.__	542	__._____
Selective Compression	544	_____._	597	_._____
Unnatural Compression	586	___.___	535	___.____

The results in Table 4.8 are remarkably similar to the intelligibility results in Figure 4.6 (section 4.4): for the initially stressed words Linear Compression wins; for the finally stressed items, there is hardly any difference between LC and UC. Importantly, for the finally stressed items, UC is much better than SC.

The less refined Univariate ANOVA provides the possibility of doing post-hoc analyses to find out which conditions differed significantly from each other. This analysis allows unequal numbers of observations over cells so that the missing observations do not have to be replaced by the grand mean. In order to make the data distribution less skewed, the analyses are run on the inverse reaction time data (1/RT). In these Univariate ANOVAs, with inverse reaction time as the dependent variable, Compression Type as fixed factor, and either subject or item as random factors, the effect of Compression Type was significant as well ($F_1(2,28)=8.2$, $p=0.001$; $F_2(2,79)=5.0$, $p=0.008$). The results of the post-hoc test (Scheffé) are shown in Table 4.9 below.

Table 4.9. Results of post-hoc test (significance values)

	subject analysis	item analysis
LC vs. SC	p=0.008	p=0.011
LC vs. UC	p>0.1	p>0.1
SC vs. UC	p>0.1	p>0.1

The results in Table 4.9 show that the significant effect of Compression Type was mainly caused by the fact that Linear Compression and Selective Compression differed significantly from each other. The conditions LC and UC and SC and UC did not differ significantly from each other.

Separate post-hoc analyses were carried out for items with initial stress and for items with non-initial stress (inverse RT as dependent variable). For items with initial stress, the difference in phoneme detection time between the LC condition (532 ms)

and the SC condition (544 ms) was not significant (Scheffé analysis on subjects $p>0.1$; on items $p>0.1$). The considerable difference between the LC (532 ms) and UC condition (586 ms) was, strangely enough, not significant either (subjects and items $p>0.1$).

Conversely, for the non-initially stressed items, the difference between LC (542 ms) and SC (597 ms) is significant (Scheffé analysis on subjects $p=0.005$; on items $p=0.006$). The difference between SC (597 ms) and UC (535 ms) is also significant (subjects $p=0.006$; items $p=0.007$), but the difference between LC and UC is far from significant (both analyses $p>0.1$). So, neither type of nonlinear time compression (SC or UC) can improve phoneme detection speed over Linear Compression.

Before drawing conclusions from these data, one must take into account that phonemes can be detected via two routes. Cutler & Norris' race model (1979) states that phoneme monitoring can either be the result of a target detection procedure carried out on the pre-lexial representation, or on the basis of phoneme information associated with a lexical representation. These two procedures run in parallel, and whichever is the fastest, wins the race. Studies by Morton & Long (1976) and Dell & Newman (1980) showed that responses are faster when the target-bearing word is contextually predictable. These results evidently provide support for the lexical route, just as studies showing that phoneme targets in words are detected faster than in nonwords (Rubin, Turvey & van Gelder 1976). Whether phoneme detection responses are based on lexical or pre-lexical representations is also a matter of the experimental set-up (Cutler, Mehler, Norris & Segui 1987). Whereas stimulus monotony may induce subjects to focus on the *pre-lexical* route, responses are more likely to depend mainly on the *lexical* route when the targets are embedded in meaningful sentences. Note that Cutler & Norris' Race model has been challenged by several empirical studies showing e.g., lexical effects in phonemic decisions in nonwords. In answer to this, Norris, McQueen & Cutler (2000) came up with the Merge model in which information from the lexical and pre-lexical route can jointly lead to a phoneme detection response. The model allows pre-lexical processing to proceed independently of lexical processing, but merges the information of both processes at the decision stage. The pre-lexical processing feeds information to the lexical level to allow activation of lexical candidates. At the same time, this information is available for explicit phonemic decision making. The decision stage also continuously accepts input from the lexical level and can merge the two sources of information. Therefore, responses can no longer be said to be either a result of the lexical or of the pre-lexical route, they are always a result of both routes. However, the model still allows the possibility to shift attention between the two outlets. This feature must be maintained in the Merge model in order to explain why the experimental set-up can play such an important role (Cutler et al. 1987).

In the present experiment subjects had to monitor different phoneme targets, and the speech material consisted of meaningful sentences, instead of CVC word lists. These factors should all cause the subjects to focus their attention on the lexical level. Furthermore, the decreased quality of the speech as a result of moderate time compression may have caused the pre-lexical route to be rather inefficient, such that the information from the lexical route will have contributed most to the phoneme decision.

Further indications that the subjects may have focussed on the lexical level comes from the fact that the detection results are very similar to earlier word intelligibility results with respect to the same three compression conditions. As in section 4.4, the present detection results show that linear compression has a significant advantage over the selective compression condition, mainly for items with non-initial stress.

So, we tentatively assume that the detection results reflect speed of word processing. Consequently, these results, together with the intelligibility results in section 4.4, suggest that the initial syllable plays an important role in lexical access: making the initial syllable shorter than in the case of linear compression slows down speech processing. Remember that the Unnatural Compression condition lowered intelligibility and slowed down reaction times for initially stressed items but not for finally stressed items. This shows that the UC condition makes it easier to start the correct alignment with possible word candidates at the start of the unstressed syllable because the unstressed syllable is relatively salient, compared to the LC condition. This positive effect of initial alignment outweighs the fact that the stressed syllable is shorter. Thus, our results provide evidence for the importance of the stressed syllable and of the initial syllable.

In contrast to the hypothesis, the results of the present experiment confirmed the results of the intelligibility experiment presented in section 4.3. Even at the speech rate which human speakers can attain, perception of time-compressed speech is not improved by making the timing pattern more like that of natural fast speech. Intelligibility and processing speed are highest after linear time compression, even though this yields unnatural timing patterns.

4.7 General discussion

Given the results of our production study, namely, that speakers seem to speed up along the lines of the hyper- and hypotheory, we expected that imitating the more salient prosodic pattern found in natural fast speech would improve word-level

intelligibility over linear time compression. The alternative option was also investigated, namely, that intelligibility would be improved by a very unnatural timing: if segmental intelligibility of all stressed and unstressed segments outweighs the contribution of the prosodic pattern, intelligibility might be better if prosodic timing differences are made smaller than found in natural speech.

The results of the perception study with highly time-compressed material (Experiment 3) did not confirm our hypothesis: word-level intelligibility of time-compressed speech could not be improved by using either type of nonlinear time compression over linear compression.

To investigate whether selective compression would in fact improve intelligibility at speech rates humans can achieve, a speech-interference experiment (experiment 4) and a phoneme detection experiment (experiment 5) were set up. The phoneme detection results, however, pointed in the same direction as the intelligibility results for the very fast speech rate. Even at the speech rate which human speakers can attain, perception of time-compressed speech is not facilitated by making its timing pattern more like that of natural fast speech. Perceptually obliterated segments cannot be the underlying reason for the moderately fast rate results because the speech presented at this moderately fast rate was perfectly intelligible. Still, differences between the three conditions could be found in the speed with which this type of speech can be processed.

We are now left with two questions. First, how do our results fit in with the results obtained with the time-compression algorithm Mach1 (Covell et al. 1998)? Secondly, what about the H&H-based prediction that applying natural prosodic rules to artificially time-compressed speech would improve intelligibility or ease of processing?

The Mach 1 results (Covell et al. 1998) show that it is possible to obtain a significant increase in intelligibility over linear compression at heavy rates of time compression. Mach1 is based on the compression strategies found in natural fast speech timing, such as compressing pauses most and compressing stressed (i.e., sentence-accented) vowels least. Moreover, their algorithm was built so as to avoid overcompressing already rapid sections of speech. This suggests that intelligibility is helped if the prosodic pattern is not entirely at the expense of the segmental information. The fact that Mach1 could improve intelligibility by imitating fast speech timing whereas we could not may be due to at least the following. The increase in intelligibility of Mach1 could mainly be caused by those aspects of fast speech timing that exceed the word-level (such as pause reduction). In a study to be reported in section 5.5 it was found that removing pauses can significantly improve intelligibility of longer stretches of speech over linear compression. In order to achieve the same fragment duration, the remaining speech need not be time-compressed so much after

the pauses have been removed than in the case of linear time compression. The positive effect of the slower articulation rate on intelligibility outweighs the perceptual importance of speech pauses. Furthermore, at sentence-level one might find that applying the speaker's strategy of reducing function words more than content words improves sentence-level intelligibility, both at moderately fast, and at very fast rates of speech. Thus, the results obtained with Mach1 cannot really be compared with the present results. The positive effect of pause removal, and thus of preserving segmental information, may be so important that it outweighs all other possible negative effects of imitating natural fast speech timing.

There are two possible explanations why the natural way of speeding up in the present experiment did not lead to improved intelligibility for heavily time-compressed speech. The first explanation is that selectively time-compressed speech shows a mismatch between segmental intelligibility and prosodic salience. The unstressed syllables are compressed more than the stressed syllables, but segmentally they are still overspecified. In natural fast speech, there is a relation between articulatory precision and duration: extra reduction of a syllable is inevitably accompanied by extra coarticulation and slurring. Time-compressed speech does not involve this link between faster rate and decreased segmental intelligibility. Selective time compression emphasises this mismatch between prosodic pattern and segmental content even more: unstressed syllables are made very short, but their segmental content is all the more overspecified. Still, we expect that a type of spectral modification or reduction, combined with selective compression, might enhance the perceived naturalness of the speech, but does not improve intelligibility at moderately fast rates, and certainly not at the heavy rate of time compression used in sections 4.3 and 4.4. Earlier pilot experiments in our laboratory already showed that speech that was articulated fast was clearly less intelligible than speech spoken at a normal rate and later time-compressed to that same fast speech rate (cf. section 5.2.2). The increased assimilation and inevitable slurring make fast speech sound more natural, but not necessarily more intelligible. This makes the 'mismatch' explanation rather unlikely.

The second option is that our interpretation of the H&H theory is wrong. According to our particular interpretation, speakers make prosodic patterns more pronounced in order to help the listener. However, the nonlinear way in which speakers speed up at word-level may not be as strategic and communication-driven as we thought. It turns out that natural prosodic patterns do not contribute to speech intelligibility of fast speech. The attempt to preserve the segmental intelligibility of the unstressed syllable at the *expense* of the prosodic pattern even turned out to be more successful than enhancing the prosodic pattern. The more salient prosodic pattern is not helpful for listeners after all: it may just be easier for speakers to speed up in the

selective fashion, or it may perhaps be even impossible to speed up in any other way. Even though nonlinear speed-up is harmful for intelligibility, speakers are unable to speed up in such a way that it approaches linear time compression. Lexical stress is specified in the mental lexicon, and as a result of this specification, stressed syllables are produced with more articulatory precision. Stressed vowels are closer to their citation form (van Bergem 1993; Lehiste 1970). In the mental representation, the target values for stressed segments may be more strictly specified than for unstressed segments. De Jong (1995) argues that linguistic stress is localised hyperarticulation. Fowler (1981) found for English that lexically stressed vowels show less contextual variation than lexically unstressed ones: in other words, stressed vowels have a greater coarticulatory resistance than unstressed ones. Cho (2001) found that the same holds for sentence stress in English: accented vowels show a greater coarticulatory resistance than unaccented vowels. Cho (2001) also found that accented syllables were pronounced with greater articulatory strengthening, consisting of larger, longer and faster movements than unaccented syllables. Cho argues that both a change in articulatory stiffness (i.e., Moon & Lindblom's (1994) rate of change) and an increase in target are the most likely source for an increased displacement.

Consequently, if the target values of stressed syllables are more strictly specified than those of unstressed syllables, or if stressed targets are somehow 'increased' targets, the speaker is forced to spend more energy on coming close to the stressed syllable targets than for the unstressed syllable targets. A faster articulation rate is almost inevitably accompanied by undershoot of the pre-defined targets because of the inertia of the speech organs (Lindblom 1963; Moon & Lindblom 1994). Articulatory structures such as the jaw are relatively slow (cf., e.g., Perkell (1997) for some estimated minimal durations of articulatory movements). So, if more articulatory precision is required for the stressed syllables than for the unstressed syllables, the speaker simply cannot speed up that much during the production of stressed syllables.

We argue that the changes in timing that accompany faster articulation rates are not so much intended to make perception easier. They are rather the consequence of certain restrictions on articulation. Intelligibility of artificially time-compressed speech, on the other hand, is not improved by applying the temporal pattern of fast speech: time compression threatens the identifiability of unstressed segments, and selective time compression only makes this worse. Obviously, a natural prosodic pattern is not as helpful as we thought. Prosody should not be at the expense of the segmental intelligibility of the speech signal: both syllables contribute to the identification of polysyllabic words.

4.8 Conclusion

A production and perception experiment were set up to investigate what speakers do when they are forced to speak faster, and secondly, to test whether the way in which speakers speed up can improve intelligibility of time-compressed speech over linear time compression. Our first expectation was that speakers, in line with the H&H theory of speech, speed up most during the least informative parts of speech. This expectation was confirmed: lexically unstressed syllables were reduced more, relatively, than stressed syllables. The second expectation was that vowels in words bearing a pitch accent on the stressed syllable would be reduced relatively less, with increasing speech rate, than vowels in words without a pitch accent. The results did not confirm this hypothesis. This was attributed to the fact that duration is an important cue for lexical stress but not for sentence accent in Dutch (Sluijter 1995).

Because the nonlinear compression behaviour of human speakers was expected to be driven by a strategic communicative principle, the third expectation was that applying a more salient prosodic pattern to artificially time-compressed speech would improve its word-level intelligibility over linear time compression. The results of the experiments, both at a very fast rate and at a moderately fast rate, did not confirm this expectation. Instead, the reverse was found: making the temporal pattern of time-compressed speech more like that of natural fast speech worsens intelligibility and slows down speech processing. The attempt to preserve the segmental intelligibility of the unstressed syllable, at the *expense* of the prosodic pattern, even turned out to be more successful than enhancing the timing pattern. The 'selective' way of speeding up at word-level may not be the consequence of a strategic communication-oriented move, but seems to be caused by articulatory factors. Speeding up is inevitably at the expense of precision of articulation. Lexical stress requires a certain amount of precision in terms of the articulatory/acoustic targets. Hence, if precision is required, speakers cannot speed up that much.

The balance between segmental information and prosodic information turns out to be important in speech perception. Even though the stressed syllable is the most informative one, our results show that at a fast speech rate, perception is not helped by making the prosodic durational pattern more pronounced than at a normal rate. The role of prosody is not as crucial as we expected: natural prosodic rules of fast speech do not necessarily contribute to speech intelligibility. We conclude that prosody and segmental intelligibility cannot be treated as two separate factors. Putting too much emphasis on temporal prosody, at the expense of segmental intelligibility, distorts the optimal balance between these two factors, and harms word perception.

5

Word Perception in Fast Speech: Comparing Time-Compressed Speech and Natural Fast Speech

Abstract

The results of Chapter 4, indicating that word perception in artificially time-compressed speech is not improved by making its timing pattern more similar to that of natural fast speech, may have been due to a particular slurred fast pronunciation. In this chapter, the question is whether word perception is also slowed down in natural fast speech, relative to linearly time-compressed speech, when the natural fast speech is perfectly intelligible. The results of the present study confirm the earlier results: the more similar time-compressed speech is made to natural fast speech, the slower the processing time. Word perception in natural fast speech is hindered both by its changed timing pattern and by the inevitably reduced articulation. Furthermore, even when the natural fast speech is perfectly intelligible, listeners find artificially time-compressed speech more agreeable to listen to than naturally produced fast speech. The only aspect of the speaker's way of speeding up that can be imitated in order to improve perception of artificially time-compresed speech may be pause removal. At fairly heavy rates of time compression, intelligibility of artificially time-compressed speech can be improved over regular linear compression by removing pauses first.

Changes in temporal pattern and in segmental intelligibility that accompany fast speech do not occur because speakers want to help their listeners, but rather because speakers cannot speed up otherwise.

Part of this chapter also appeared as an abstract in the Journal of the Acoustical Society of America (Quené & Janse 2001).

A modified version of this chapter was submitted for publication in Speech Communication (Janse submitted).

5.1 Introduction

In the previous chapter, word-level timing in natural fast speech was shown to differ from that of normal rate speech. Speakers are selective in the way they speed up a sentence, in that unstressed syllables are reduced more than stressed syllables. Imitating this selective compression behaviour from natural fast speech was expected to improve word-level intelligibility, relative to linear time compression. This turned out not to be the case: neither at the very fast rate, nor at the moderately fast rate which speakers can attain was word perception improved by making the timing pattern of time-compressed speech more similar to that of natural fast speech. Hence, natural duration rules do not necessarily contribute to speech intelligibility. The nonlinear fashion in which speakers speed up must therefore be attributed to non-perceptual factors.

An alternative explanation would be that the speakers of Chapter 4 did not intend to be communicative at all. They were asked to speed up, and thus their focus may have been more on speed than on communication and intelligibility. The fact that they were asked to read the material out loud may have further reduced the success of imitating a real-life communicative situation. Had the speakers been confronted with a more realistic situation, e.g., describing a route to a tourist whilst they themselves were in a hurry, then they might not have lost the communicative intention out of sight.

In section 4.2, speakers were asked to produce sentences at a normal and a very fast rate. The intelligibility of the fast material can be compared with that of the normal rate material which is artificially time-compressed to that same fast rate afterwards. If the fast articulated speech turns out to be less intelligible than the artificially time-compressed speech, this may, at least partly, be attributed to the differences in word-level timing between the two types of speech. Yet, the slurring, coarticulation and assimilation processes that inevitably accompany a very fast speech rate will probably not contribute to the intelligibility of speech either, even though listeners might expect these processes to occur at such a fast rate.

The H&H theory (Lindblom 1990) claims that speakers continuously adapt their speech to what listeners need at that moment. More redundant parts of speech can be articulated in a less precise way (hypospeech) than parts of speech carrying new information (hyperarticulation). According to this theory, the listener's needs are always in the mind of the speaker. Speakers are thought to tailor their utterances to this internal representation of the listener's needs. Horton & Keysar (1996) found evidence that this may not be true when speakers are under time or task pressure. In their experiment, speakers carried out a referential communication task in which they had to describe objects. Horton & Keysar's data showed that common ground, or shared knowledge, was used in the descriptions without time pressure, but that common

ground was not used when speakers were under time pressure. Further evidence that speakers do not always take into account the needs of listeners comes from the study of intonation. Under time pressure, speakers of Dutch have been shown to make use of a smaller choice of pitch markers than in normal-rate speech (Caspers 1994; Caspers & van Heuven 1995). More marked pitch configurations were replaced by unmarked ones, such that shades of intonational meaning were lost. This may have some implications for the present study on natural fast speech. It is conceivable that the time pressure that was imposed on the speakers of the duration study reported in Chapter 4 may have made them lose sight of the listeners. In the previous chapter the nonlinear way of speeding up was argued to be due not so much to a communication-driven strategy, but to restrictions on articulation. However, if we are not sure that the speakers actually intended to be understood, we cannot exclude the possibility that speakers just chose the easiest, and not necessarily the only possible way to speed up. The question then becomes whether speakers can be asked to speak fast and intelligibly, without any negative consequences for the perception.

In Chapter 3 the importance of segmental intelligibility of fast speech was discussed. Listeners, when presented with fast speech, were assumed to find segmentally hyperarticulated speech easier to process than less redundant speech. Various studies by Marslen-Wilson, Nix & Gaskell (1995), and Gaskell & Marslen-Wilson (1996; 1998) suggested that, at a normal speech rate, there is no perceptual advantage for assimilated versions over unassimilated articulations of a word form, given the appropriate phonological context. Quené & Krull (1999) suggested that this may have been due to the rate and style of the experimental material in those three studies. They expected a perceptual advantage for assimilated over unassimilated versions when the speech rate was faster than normal. However, listeners turned out to detect assimilated word forms faster than unassimilated forms at normal speech rate, whereas the reverse was found for a fast speech rate (Quené & Krull 1999). Kohler (1990) describes assimilation as perceptually tolerated articulatory simplification. This agrees with the predictions of the H&H model (Lindblom 1990): speakers will try to economise on speaking effort as long as the communicative situation allows it. Whereas reduced redundancy in the form of assimilation is not problematic for listeners in normal conditions, it may be problematic for word perception in fast speech.

Summing up, it seems that word recognition and intelligibility in fast or time-compressed speech will be helped by segmental redundancy, even if that segmental redundancy is artificially high. In the present chapter, perception of naturally produced fast speech (assimilated) is compared with perception of speech that is articulated at a normal rate and is artificially time-compressed afterwards (hyperarticulated). This time, however, the question is whether the processing difference between natural fast and

time-compressed speech is due merely to the fact that the fast speakers did not care about the intelligibility of their fast speech. Or, in other words, is word perception also more difficult when the fast speech is produced at a more moderately fast rate and the speaker is pressed to remain intelligible?

If the conclusion of the previous chapter holds, namely that speakers speed up in a nonlinear fashion because of articulatory factors, then word perception is predicted to be more difficult, even though the speaker has communicative intentions. The nonlinear way of speeding up, combined with the extra, almost inevitable, reduced articulation, should make word perception more difficult in this type of speech. The present study was set up to investigate whether and how both factors, i.e., the reduced articulation factor and the prosodic timing factor, influence ease of processing of naturally produced fast speech, as compared with the processing of time-compressed speech.

First, two pilot experiments are presented in which the intelligibility and processing speed of the speech material of Chapter 4 (section 4.2) is evaluated. Both intelligibility and phoneme detection speed are compared for naturally produced very fast speech and linearly time-compressed speech. Then an experiment with naturally produced moderately fast speech is presented in which the following three conditions are compared:

1. linearly time-compressed speech
2. copy-fast-speech-timing (all segment durations of the normal condition are set to the segment durations found in the natural fast condition)
3. natural fast speech

In this way, the separate contributions can be investigated of increased segmental overlap and selective time compression (i.e., time-compressing the normal rate condition syllable by syllable in order to copy the syllable durations of the natural fast condition but to preserve the segmental quality of the normal rate condition).

On the basis of the results of the previous chapter, word-processing is expected to be more difficult when the timing pattern of natural fast speech is applied: nonlinear speed-up is expected to be due to articulatory factors, rather than being a communicative-oriented strategy. Perception is predicted to be even more difficult in the natural-fast condition, in spite of its naturalness, due to its reduced articulation.

Furthermore, not only is perception predicted to be more difficult in the natural-fast condition, but listeners may also find artificially time-compressed speech more agreeable to listen to than natural-fast speech. Even though natural-fast speech sounds more natural than artificially sped-up speech, it is conceivable that listeners find the

more neatly articulated time-compressed speech more agreeable because of the lower processing cost, even at a rate at which both types of speech are still perfectly intelligible. In a Comparative Mean Opinion Score test (ITU-P.800 1996; van Santen 1993), listeners' preference will be tested when they are presented with pairs of sentences of the three types of fast speech. The hypothesis here is similar to that of the phoneme detection experiment: both types of artificially compressed speech will be judged as 'more agreeable' over natural fast speech; and the linear type of compression will be judged as 'most agreeable'.

The predicted results concerning processing speed and subjective preference are indicated in Table 5.1.

Table 5.1. Predicted results for phoneme detection experiment and subjective preference test, for all three fast conditions.

	Phoneme detection time	Subjective Preference
Linear Compression	fastest	most agreeable
Copy-fast-speech-timing compression	intermediate	less agreeable
Natural fast speech	slowest	least agreeable

If the changed timing and increased segmental slurring only have a negative effect on intelligibility and ease of processing, is there anything left in the speaker's way of speeding up that could improve intelligibility over linear time compression? The results obtained with the Mach1 algorithm (Covell et al. 1998) have shown that intelligibility can be improved over linear time compression. One of the strategies included in the Mach1 algorithm is to strongly reduce the inter-phrasal pauses. By doing this, the remaining speech can be time-compressed to a lesser extent than in the case of linear time compression, which affects pauses to the same degree as the remaining speech. This pause-removal strategy may be the major factor in the ultimate advantage of Mach1 compression over linear time compression.

In this chapter, the following three general hypotheses are tested:

1. Processing of fast speech is hampered by a more natural speech signal: removing only the temporal or both the temporal and the segmental characteristics of natural fast speech (as in artificially time-compressed speech) will make processing easier. The more similar natural fast speech is to artificially (linearly) time-compressed speech, the shorter its processing times.
2. In a preference test, natural fast speech is expected to be judged as 'less agreeable' than the two types of artificial time compression. Linear time compression will be judged as 'most agreeable'.

3. Pause removal, combined with less linear time compression, can improve intelligibility of heavily time-compressed speech over strictly linear time compression.

The first two hypotheses will be addressed in sections 5.2 and 5.3. Section 5.5 will deal with the third hypothesis.

5.2 Two pilot tests

In this section, the intelligibility and processing speed of the natural speech material of Chapter 4 (section 4.2) is investigated. As noted in the Introduction section 5.1, the fast speech material may have been relatively slurred because subjects cared more about their ultimate fast speech rate than their intelligibility. In section 5.2.1 the intelligibility of the fastest speaker's material is studied, relative to the intelligibility of the normal rate material of that same speaker, which is time-compressed afterwards to the same fast rate. In the second pilot experiment, reported in section 5.2.2, phoneme detection time, as a measure of the ease with which speech can be processed, is evaluated for the natural fast speech of the most intelligible fast speaker of Chapter 4, and for her artificially time-compressed material. If linearly time-compressed speech turns out to have both an intelligibility and a speech processing advantage over naturally produced fast speech, further research can be set up to investigate whether this is mainly due to the fact that the speakers of Chapter 4 did not care about their intelligibility.

5.2.1 Experiment 1: Intelligibility of the sentence material of Chapter 4

The intelligibility of the speech material of one of the speakers of section 4.2 was tested. The fastest speaker was selected (cf. Figure 4.2: speaker 4). The mean fast/normal ratio of this speaker was 0.6: in the fast-rate condition, the duration of the sentence fragments was reduced to 60% of the normal-rate duration. Thirty sentences were selected: all with a pitch accent on the disyllabic target word. The intelligibility test was a cloze procedure in which the listeners were first presented with an incomplete sentence on the computer screen. Then the entire sentence was played to them and they were asked to fill in the missing word.

The first condition was the fast articulation condition. The second condition was Linear time compression. The target word's duration was measured in the normal and in the fast rate condition. The fast/normal ratio of this target word was applied to

linearly time-compress the entire sentence, so that the target word duration would be equal in the different test conditions. The third condition was Selective time compression. In the Selective compression condition, the duration of each segment of the target word (at normal speech rate) was reduced to the duration of that segment at fast speech rate on a segment-to-segment basis. The duration of each segment of the target word was measured in the normal and fast rate conditions so that for each segment, a fast/normal ratio could be computed. For the selective time compression condition, the target word segments of the normal rate conditions were time-compressed accordingly. The rest of the sentence was time-compressed linearly, according to the global fast/normal ratio of the entire target word, as in the Linear Compression condition.

The three experimental conditions were balanced over the 30 target items and were placed on 3 different experimental lists (Latin square design). Ten extra sentences, similar in length and complexity, were designed as a practice session. Subjects were 36 students at Utrecht University (12 for each of the 3 lists) who were paid for their participation.

Results

The intelligibility scores are shown in Figure 5.1 below.

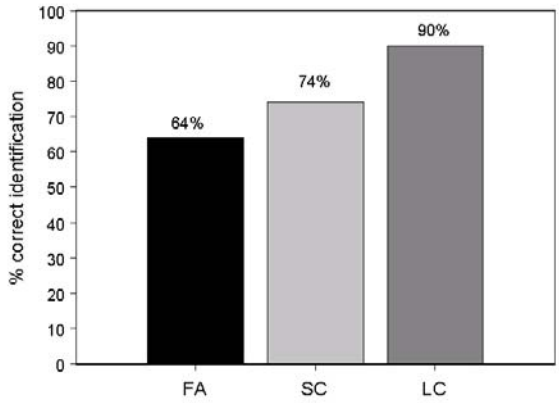

Figure 5.1. Intelligibility scores (% of correct word identification) for three fast conditions: Fast Articulation (FA), Selectively compressed speech (SC) and Linearly compressed speech (LC).

The correct identification scores (per subject or per item; after arcsine transformation) were entered into Repeated Measures analyses. The effect of Condition was highly significant ($F_1(2,34)=57.7$, $p<0.001$; $F_2(2,28)=12.6$, $p<0.001$). Separate pair-wise t-tests were carried out as post-hoc tests to investigate which conditions differed from each other. All three conditions differed significantly from each other (Fast Articulated vs. Selective Compression: ($t_1(35)=3.68$, $p=0.001$; $t_2(29)=2.19$, $p=0.036$); Fast Articulation vs. Linear Compression: ($t_1(35)=10.8$, $p<0.001$; $t_2(29)=5.10$, $p<0.001$); and Selective Compression vs. Linear Compression: ($t_1(35)=7.73$, $p<0.001$; $t_2(29)=2.73$, $p=0.011$)).

The intelligibility differences between the conditions are actually quite substantial. The difference between Fast Articulation and the time-compressed conditions suggests that the naturally produced fast speech was rather slurred.

5.2.2 Experiment 2: Processing speed of natural fast vs. time-compressed speech[16]

Reaction time measures can be used to compare speech quality of highly intelligible speech types, such as synthetic speech (Nix et al. 1993; Pisoni 1987). Phoneme detection time has been used to compare the processing speed of spontaneous speech and read speech (Mehta & Cutler 1988). These two types of speech mainly differ with respect to prosodic characteristics.

In this study, processing speed of fast articulated speech is compared with that of artificially time-compressed speech. Artificial time compression in this study concerns linear time compression. The two types of speech differ with respect to the temporal pattern, and with respect to segmental intelligibility. The results of Chapter 4 have shown that making the temporal pattern of artificially time-compressed speech more similar to that of natural fast speech slows down word processing, relative to linear time compression. The reduced segmental intelligibility of naturally produced fast speech is also expected to slow down speech processing, relative to the hyperarticulated artificially time-compressed speech. Hence, the prediction is that perception is more difficult in the natural-fast condition than in the linearly time-compressed condition.

In the duration study of the previous Chapter (section 4.2), a number of speakers were asked to read sentences both at a normal and at very fast speaking rate. In pilot experiment 1 above, the speech material of the *fastest* speaker was selected. For the present pilot experiment, the material of the *most intelligible* speaker was selected. Because this speech material was thought to be of relatively high intelligibility, phoneme detection was chosen, as a measure of ease of processing.

16 The experiment described in this section was carried out by two students, Fiona Sely and Eva Sittig, as part of a practical course.

In Chapter 3 several studies were mentioned which used phoneme detection to compare ease of processing of synthetic versus natural speech (Nix et al. 1993; Pisoni 1997). Even though both speech types are perfectly intelligible, natural speech can be shown to have a processing advantage over synthetic speech. Pisoni (1997) argues that the "extra processing effort appears to be related to the initial analysis and perceptual encoding of the acoustic-phonetic information, and not to the process of accessing words from the lexicon" (p.550). If initial acoustic-phonetic analysis is slowed down, both pre-lexical processing, and consequently, lexical processing are slowed down. The same might be said about the difference in segmental intelligibility between naturally produced fast speech and artificially time-compressed speech. The reduced articulation of the natural fast speech may make initial phonetic analysis more difficult for the listeners than in the case of artificially time-compressed speech.

Material A set of 30 sentences was selected, which all contained a word with a target plosive in word-initial position. The sentence material had not been constructed for the purpose of a phoneme detection experiment. Because of this limited set of material, selection criteria could not be very strict. Consequently, the target items were not uniform with respect to syllable number and syntactic class. There were 7 monosyllabic target words, 17 disyllabic target words, and 6 target words with three syllables. Of the polysyllabic items, 10 had initial stress, and 13 had non-initial stress. The target items were nouns, verbs and adjectives. There were 6 items with /t/, 6 with /p/, 6 with /k/, and 12 with /b/. Furthermore, there were 24 catch trials to prevent subjects from pressing the button randomly. The sentence material read at a normal rate was time-compressed linearly to the rate of the natural fast speech version. This was done in three steps. First, the part of the sentence up to the target word was measured in the normal and fast rate versions. The normal rate version was then time-compressed linearly to the fast rate. Then the target word itself was measured in the two speech rate conditions and the normal rate version was time-compressed linearly to the fast rate. Lastly, the remaining part of the sentence was time-compressed. By time-compressing the sentences in these three steps, the duration of the target word was made equally long in both conditions. Mean normal rate for this speaker was 6.5 syllables/second, and mean fast rate was 9.9 syllables/s. This means that the normal rate speech was time-compressed to 66% of its original duration (i.e., 1.5 times faster than the normal rate).

Design The 30 test items were distributed over two experimental lists because each subject could not be presented with the same item in both conditions. The 30 items were rotated over the two conditions in a Latin square design.

Subjects Twenty subjects were assigned to both experimental lists. All 40 subjects were students of Utrecht University. They received a small payment for their participation.

Procedure Subjects were seated in sound-treated booths with a computer screen and a button box in front of them. The speech material was presented to them over closed earphones. They were instructed to watch the computer screen in front of them, since a plosive phoneme would appear on the screen before each auditory trial. Subjects were asked to press a button (with their dominant hand) whenever they detected the assigned plosive in word-initial position in the auditorily presented sentence. They were asked to react as fast and as accurately as possible. Subjects were also told that there would be catch trials in which the target plosive did not occur in the sentence. Before the actual test session started, the subjects were presented with a practice session after which they could ask questions if anything was unclear. After the auditory presentation of each sentence, there was a 2 second period during which subjects could give their response. Normally, reaction times are measured from the onset of the silent interval or voice bar. In the present material, however, markers were placed in the audio files immediately after the burst of the plosive, instead of at the onset of the silent gap. This was done because in the natural fast speech condition, silent intervals were often absent or difficult to detect. Reaction times were computed during the experiment on the basis of the registered press of the button and the marker time. Two seconds after the sentence's offset, the test proceeded with the next test item. The experiment took approximately 10 minutes.

Results

The mean raw phoneme detection times are shown in Table 5.2 below. If subjects had failed to respond, or had responded too late, or if they had responded to an earlier non-initial occurrence of the sound, the responses were regarded as missing observations. The miss rates in both conditions are also reported in the table. The results in Table 5.2 show that mean phoneme detection is 87 ms faster in the linearly time-compressed condition than in the natural-fast condition.

Table 5.2. Raw mean phoneme detection times, plus standard error of the mean, for both conditions. Miss rates are also indicated.

	Mean Detection time (ms)	S.E.	Miss rate (%)
Fast articulation	625	13	16
Linear time compression	538	12	7

For the statistical analysis of the results, univariate analyses of variance are presented in which the missing observations do not have to be replaced. Either subjects or items are treated as random factors. The effect of Speech Condition was highly significant, both in the analysis by subjects ($F_1(1,39)=28.7$, $p<0.001$), and by items ($F_2(1,29)=18.1$, $p<0.001$).

Statistical analyses were also carried out on the percentages of missing observations (pairwise t-tests). The percentages of valid observations per condition were established for each subject and for each item, and were arcsine transformed. The difference between the two conditions with respect to the number of valid observations was highly significant ($t_1(39)=-5.92$, $p<0.001$; $t_2(29)=-3.36$, $p=0.002$).

Some sentences elicited quite a high number of missing observations, mainly in the natural fast condition. A selection of 'successful' items was made that had 15 or more valid observations (out of 20 per condition). This selection criterion yielded 25 'successful' items out of the 30 used in the experiment. The reaction time data were then analysed again. The raw mean detection time in the natural fast condition was 613 ms (miss rate 6%); versus 520 ms in the linear compression condition (miss rate 1%). This difference in detection time was still highly significant, both by subjects and by items ($F_1(1,39)=26.7$, $p<0.001$; $F_2(1,24)=16.4$, $p<0.001$).

Whether phoneme detection times reflect actual word processing is still a matter of debate. It is questionable whether the speech signal is continuously processed, such that even a few ms of speech can activate the lexicon as claimed by e.g., Marslen-Wilson & Tyler (1980) and McClelland & Elman (1986). Others believe that the signal is first parsed into large pre-lexical units, for instance, syllables, which are then used for lexical look-up, e.g., Massaro (1972), and Mehler, Dommergues, Frauenfelder & Segui (1981). Even though both lexical and pre-lexical processing proceed in order to come up with a phoneme decision, as modelled already in the Race model (Cutler & Norris 1979), the response is now considered to be always a combination of the information from both processing routes; as claimed in the Merge model (Norris et al. 2000). It is assumed here, as in the phoneme detection studies reported in Chapters 3 and 4, that phoneme detection via the pre-lexical route is rather inefficient, because both time compression and fast articulation deteriorate the segmental intelligibility of speech. Secondly, the fact that the target items were embedded in meaningful sentences, and were not presented as lists of isolated word items, may induce listeners to rely more on lexical rather than pre-lexical processing for their phoneme detection responses. Consequently, the information from the lexical route is assumed to contribute most to the ultimate phoneme decision.

The 87 ms difference in processing time in the present experiment is much larger than the 39 ms processing difference between linearly and selectively time-compressed

speech observed in the phoneme detection experiment in Chapter 4 (section 4.6). In the latter experiment, conditions only differed with respect to their temporal pattern. This suggests that linearly time-compressed speech is not only easier to process because of its temporal make-up, but also because it is segmentally more redundant than the natural fast speech. This will be explored further in the next experiment.

Note that the material that was used in the second pilot experiment was not tested in advance for intelligibility. In the Introduction, phoneme detection was introduced as a measure of the quality of *perfectly intelligible* speech. The large difference in response time, plus the high miss rate, raises doubts about the intelligibility of the natural fast condition and about the communicative intentions of the speaker. Even though this speaker may have been more intelligible than the fastest speaker (whose material was tested for intelligibility in the first pilot experiment), intelligibility may have been far from perfect. However, the same tendencies are expected for perfectly intelligible speech, produced at a more moderately fast rate.

5.3 Experiment 3: Word-perception in natural fast speech and time-compressed speech

In this section the question is addressed whether time-compressed speech still has a processing advantage over naturally produced fast speech even if the natural fast speech is perfectly intelligible. The question is addressed whether word-level timing is also different from normal rate for fast and intelligible speech. If this is the case, how does it influence perception, relative to linear compression? When speakers are asked to speak fast and intelligibly, is naturally produced fast speech easier to process than time-compressed speech? On the basis of the phoneme detection results reported in Chapter 4 (section 4.6), processing is expected to be slower in the natural fast condition, as compared to time-compressed conditions. If a change in timing from normal to fast rate is found, it is expected to result from articulatory restrictions, rather than from a listener-oriented strategy. Hence, this change in timing should then slow down speech processing. Secondly, the increased coarticulation and assimilation that almost inevitably accompany a faster speaking rate are expected to make word perception more difficult than in the case of artificial time compression. So, the question is whether and how the reduced segmental quality in natural fast speech, together with the changes in timing, contribute to slower speech processing than in the case of linear time compression of speech. To this end, word perception is compared in three conditions. First, processing speed will be established for the natural fast condition: this

is the condition in which the speaker articulates the sentences at a moderately fast rate. Secondly, this natural fast condition will be compared with a linearly time-compressed condition: the speech is articulated at a normal speaking rate and is time-compressed linearly afterwards to the same fast rate as in the former condition. Thirdly, all syllable durations, as measured in the normal rate condition, will be time-compressed to their respective durations as measured in the natural fast rate condition. This is what will be called the copy-fast-speech-timing condition. Note that this copy-fast-speech-timing condition is different from what we have called the Selective Compression condition in Chapter 4. Selective time compression is only a rough imitation of what speakers do. This global imitation was an extrapolation from what was found in our duration study (cf. section 4.2), and it may have been too coarse or too general. The copy-fast-speech-timing condition is an exact imitation of the actual changes in word- and sentence-level timing that the speaker has applied in speeding up. By comparing these three conditions, we hope to pull apart the respective contributions of changes in timing and of segmental slurring to the slower processing speed that we expect to find for the natural fast condition.

In the present experiment, the hypothesis is tested that word processing is hampered by a more natural speech signal: removing only the temporal characteristics or both the temporal and the segmental characteristics of natural fast speech (as in artificially time-compressed speech) should make processing easier.

5.3.1 Method

As in Chapters 3 and 4 (and section 5.2.2), phoneme detection is used to evaluate speech processing difficulty. Again, for the reasons mentioned in those previous sections, it is assumed that both information from the lexical and from the pre-lexical route are jointly responsible for a phoneme detection response (Norris et al. 2000). However, it is also assumed that the information from the lexical route weighs more heavily because segmental intellibility is decreased, and because subjects will focus on the lexical route when they are presented with meaningful sentences. Phoneme detection times are thus taken to reflect the ease of lexical processing in either of the three fast conditions.

Material News bulletin items (ANP news items) were collected and sentences were selected from those that had nouns in them starting with a plosive. Three examples are given below. The sentence fragments in italics were presented in a phoneme detection study. The letter in bold indicates the target plosive.

- *Een Duitse rechter heeft de echtgenoot van prinses Caroline van Monaco een **b**oete opgelegd van 1,1 miljoen gulden.* Hij wordt daarmee gestraft voor het herhaaldelijk beledigen van twee medewerkers van het Duitse boulevardblad Bild.[17]

- *Verf en **t**apijt brengen giftige stoffen in omloop.* Dat blijkt uit een chemische analyse van huisstof dat Greenpeace in Nederlandse huishoudens heeft opgezogen. De milieuorganisatie maakte de testresultaten maandag bekend.[18]

- De beslissing van de Britse premier Blair van maandag zichzelf een loonsverhoging toe te kennen is niet in goede aarde gevallen. *De **k**ritiek op de regeringsleider was dinsdag niet van de lucht*, omdat de Labour-leider voortdurend loonmatiging heeft gepredikt.[19]

In total, there were 82 of such news items. From those, 84 sentences or sentence fragments were chosen. In half of these the target noun had initial stress; in the other half, stress was non-initial. The target nouns had two to four syllables. The nouns were never compounds, but some were morphologically complex (e.g.*, **p**oging, **t**entoonstelling, **t**uinders).* The sentences or sentence fragments had a mean length of 23.4 syllables (s.d. 7.6).

One male speaker of Dutch was asked to read the sentence material at a normal and at a fast rate. It was stressed that the fast rate should be fast, but should still sound intelligible. The mean normal speaking rate was 6.1 syllables/second (s.e. 0.05), and this rate was increased to 8.5 syllables/second in the fast condition (s.e. 0.07). Thus, the overall fast/normal ratio was 0.72. In other words, the articulation rate was increased by a factor 1.4. Pair-wise comparison of the articulation rates in the normal and fast conditions showed that the mean fast articulation rate is significantly faster than the normal articulation rate ($t(83)=-41.6$, $p<0.001$).

It is important to note that the fast rate in the present study is not as fast as in the previous chapter (section 4.2). Speakers of that duration study increased their speech rate to a mean articulation rate of 10.5 syllables/second. At this rate, their speech

---

[17] *A German judge has imposed a1.1 million guilders fine on the husband of princess Caroline of Monaco.* He is punished for repeatedly offending two employees of the German tabloid Bild.

[18] *Paint and carpet spread toxic substances.* This is the result of chemical analyses by Greenpeace of dust hoovered in Dutch family homes. The environmental organisation published the test results on Monday.

[19] The decision of the British prime minister Blair of last Monday to give himself a pay rise did not go down well. *There was fierce criticism*, because the prime minister has so far been promoting wage restraint continually.

became relatively unintelligible, whereas the present speaker was pressed to speak fast, but to remain intelligible.

The 84 test sentences or fragments that were articulated at a normal rate were often somewhat louder than those that had been articulated at a fast rate. For each sentence, the mean intensity of the fast version was therefore amplified to equal that of the normal rate version.

The 84 test sentences were labelled manually: labels were placed in the waveform at all syllable boundaries. This was done for the normal and fast rate version of each sentence. An attempt was made to segment the normal and fast speech automatically at phoneme level. [20] This was done by way of an automatic phoneme alignment procedure for (American) English, developed at the Oregon Graduate Institute of Science and Technology by Hosom (2000). The automatic segmenter aligns hidden Markov model (HMM) states with the speech waveform, using a phonetic transcription of the utterance. One important feature of this alignment procedure is that phonemes are assumed to be products of distinctive (phonetic) features. This entails that the procedure can generalise quite easily to other languages. Performance of the automatic alignment procedure was fairly good on the TIMIT corpus (cf. Hosom (2000): agreement with manual alignment was 92.6% correct within 20 ms). However, many alignment errors were made for the Dutch fast speech condition. The heavy coarticulation within syllables at faster speech rates also makes hand-labelling, or manually correcting the automatically placed labels, fairly difficult. Consequently, again, only syllable boundary labels were placed. At this stage, it seemed more practical to place the syllable labels manually than automatically.

For the copy-fast-speech-timing time-compressed condition, all fast/normal ratios were computed by dividing the duration of each syllable in the fast condition by the respective duration of the same syllable in the normal rate condition. Then, durations were time-compressed on a syllable-by-syllable, rather than on a segment-by-segment basis (cf. section 4.4.1). In this way, the timing structure of the copy-fast-speech-timing time-compressed condition was a copy of that of the naturally produced fast version.

---

[20] Thanks are due to Johan Wouters at OGI (now at SVOX), for his attempt to align phoneme durations automatically. Initially, the idea was to not only time-compress normal rate material selectively, but to apply a type of spectral reduction as well, such that a more successful imitation of natural fast speech could be made by manipulating normal rate speech both durationally and spectrally (cf. (Wouters & Macon 2002a, 2002b). However, preliminary experiments showed that the spectral reduction technique proposed in (Wouters & Macon 2002b) did not produce very noticeable changes in the articulation of the Dutch normal rate speech. This could be because (a) the Dutch normal rate speech was already quite reduced spectrally (perhaps because it concerned running speech in long sentences and not short carrier phrases), (b) the technique is based on modifying the spectral rate of change of diphthong and liquid-vowel transitions that may be more specific to American English.

After the sentence had been time-compressed syllable-by-syllable, the end-result was somewhat shorter than the natural-fast condition. This is due to a PSOLA artefact: repetitive time compression of successive small windows of speech in the end always yields a slightly faster version than specified. In order to make the copy-condition and the natural-fast condition exactly equally long, the natural-fast condition was time-compressed somewhat (linearly to 98-99%). In this way, the two conditions were of exactly equal duration. Furthermore, potential differences between the three test conditions would not be due merely to the fact that two of them were PSOLA time-compressed, whereas one of them was not.

For the linear compression condition, the overall fast/normal ratio (sentence duration fast condition/sentence duration normal condition) was computed for each sentence. This overall fast/normal ratio was then applied by linear time compression to the normal rate version of each sentence. Lastly, the target word's duration was made equal to that in the natural fast condition. This was done to make sure that the word offset would not be reached earlier in either of the three conditions.

In addition to the 84 test sentences, 80 catch trials were interspersed with the material to prevent subjects from pressing the button randomly. The catch trial items did not contain the plosive the subjects were asked to detect. These 80 catch items were also taken from the news bulletin items.

In a pilot study with a small number of listeners, the intelligibility of the fast and time-compressed conditions was established. The intelligibility was very high; overall sentence-level intelligibility of the sentence material approached 100%.

Design and Procedure The 84 test sentences were rotated over the three experimental conditions and were distributed over three lists according to a Latin square design. Each subject could only be presented with the same sentence once. Subjects were seated in sound-treated booths and were tested individually. They listened to the speech material over headphones. The phoneme detection procedure was similar to that described in section 5.2.2 (and in Chapters 3 and 4).

There were 10 practice items, after which subjects could ask questions if anything was unclear. After the subjects had resumed the test, 4 warming-up filler items were presented to make sure that subjects were warmed up before the actual test began. All test and filler items were presented in random order.

Subjects Ten subjects were assigned to each list. The 30 subjects were all students of Utrecht University and received a small payment for their participation.

5.3.2 Results

Time markers had been placed in the audiofiles at the start of the silent interval of the target plosive (or at the start of the voice bar for voiced plosives). During the experiment, reaction times were computed by subtracting this marker time from the time until the button press was registered. The raw mean detection times are presented in Table 5.3.

Table 5.3. Raw mean phoneme detection times in three fast conditions; plus standard error of mean and miss rates.

	Mean detection time (ms)	S.E.	Miss rate (%)
Linear time compression	572	9	3
Copy-fast-timing time compression	600	13	3
Natural fast speech	624	14	3

The pattern of results is as predicted: phoneme detection is fastest in the linearly time-compressed condition, and slowest in the natural fast condition. Results for the copy-fast-speech-timing condition are in between. The miss rates are low in all three conditions, which provides some further evidence that the speech in all three conditions was of high intelligibility.

The phoneme detection results in section 4.6 were quite different for initially stressed vs. finally stressed target items. The present results are therefore also broken down by Stress position, to investigate whether the same applies here (cf. Table 5.4). The pattern of results is relatively similar for items with initial stress vs. items with non-initial stress.

Table 5.4. Raw mean phoneme detection time, plus standard error of mean and miss rates, in three fast conditions, broken down by Stress position.

	Initial stress (target stressed)			Non-initial stress (target unstressed)		
	Mean	S.E.	Miss rate	Mean	S.E.	Miss rate
Linear time compression	565	12	3%	578	14	3%
Copy-fast-timing time compression	588	16	3%	612	19	2%
Natural fast speech	614	19	3%	633	22	3%

The missing observations were replaced by the subject's mean in that condition for the subject analysis and by the item mean in that condition for the item analysis. Generally, reaction time data do not show normal, or Gaussian, distributions. Since analyses of

variance assume normally distributed data, reaction time data may pose a problem for ANOVA analysis. A cell variance test showed that the reaction time data differed significantly from a normal distribution (Kolmogorov-Smirnov One-Sample Test: $Z=11.7$, $p<0.001$). When reaction times are transformed to inverse reaction times (1/RT), the distributions are usually much less skewed. Although the transformed data are more normally distributed than the untransformed, another Kolmogorov-Smirnov test showed that the transformed data still differed from a normal distribution ($Z=2.3$, $p<0.001$).

The inverse detection time data were fed into analyses of variance with either items or subjects as repeated measures. Condition and Stress position were analysed as fixed factors (in the item analysis, items were nested under Stress position). The effect of Condition was significant in both analyses ($F_1(2,28)=7.4$, $p=0.002$; $F_2(2,81)=4.3$, $p=0.039$). The effect of Stress position was far from significant ($F_1(1,29)=1.0$, n.s.; $F_2(1,82)<1$, n.s.); and so was the interaction between Condition and Stress position ($F_1(2,28)<1$, n.s.; $F_2(2,81)<1$, n.s.).

It is impossible to carry out post-hoc tests in Repeated Measures ANOVAs in the statistics program used here (SPSS). The inverse RT data were therefore also analysed in classic univariate analyses of variance with Condition as a fixed factor, and either subjects or items as random factors (missing observations now replaced by either subject or item mean in that condition). The effect of Condition was significant in both analyses ($F_1(2,28)=5.5$, $p=0.007$; $F_2(2,82)=3.1$, $p=0.048$). The post-hoc analyses (Scheffé) results are shown in Table 5.5.

Table 5.5. Significance values of post-hoc pairwise comparisons

	Subject analysis	Item analysis
Linear vs. Copy-fast	p>0.1	p>0.1
Linear vs. Natural-fast	p=0.030	p=0.009
Copy-fast vs. Natural-fast	p>0.1	p>0.1

The post-hoc pairwise comparisons showed a significant difference between the conditions Linear Compression and Natural fast, but not for any other pair.

So, overall, the differences between the three experimental conditions are rather small. Only the 52 ms advantage of linear compression over natural fast speech (cf. Table 5.3) is significant. The fact that there is no robust significant difference between the two time-compressed conditions may be attributed to the relatively small difference in speaking rate between the normal and the fast speech conditions, which, in turn, induced only small changes in word- and sentence-level timing. The fast speech rate in the present study is much slower (8.5 syll./sec) than the fast rate observed in section 4.2

(10.5 syll./sec). The duration measurements presented in Chapter 4 illustrated the non-uniform way of speeding up at word-level: on average, stressed syllables were reduced to 65% of their normal rate duration, whereas unstressed syllables were reduced to 45% of their normal duration. In the present material, stressed syllables had a mean fast/normal ratio of 0.77; and the unstressed syllables had a mean fast/normal ratio of 0.71. It is therefore not surprising that listeners hardly show any processing difference between the word-level timing of the fast articulation rate (as in the Copy-fast-speech-timing condition) and linear time compression. Major shifts in word-level and sentence-level timing apparently take place only when the speech rate is sufficiently high.

To some extent then, the experimental set-up has failed to answer our question. We cannot quantify the separate effects of segmental slurring and changes in timing on phoneme detection times. However, the data do show a trend for detection times in the copy-fast-speech-timing condition to be slower, relative to linear compression. This provides further support for the findings of Chapter 4: although the Selective Time-compression condition used in Chapter 4 was an extrapolation of the nonlinear speed-up behaviour of vowels, an exact imitation of what speakers do also tends to slow down processing. The phoneme detection experiment in section 4.6 showed that, at even faster rates, imitating fast speech timing alone by way of selective time compression can slow down processing significantly. In the present experiment, however, only the combined effect of a changed timing and segmental slurring slows down speech processing significantly. The large processing advantage (87 ms) of linear time compression over natural fast speech in the pilot experiment presented in section 5.2.3 illustrated how the timing and segmental factors, although only weakly inhibitive in the present experiment, become really problematic at even faster rates. If speakers are pushed to speak faster than the speech rate of 8.5 syll./sec, slurring becomes more and more problematic for speech perception.

These experiments suggest that speakers cannot speed up their speech rate without making speech processing more difficult for the listener. Nevertheless, it would be interesting to find out whether listeners have any clear preference for either of the fast conditions. Although it seems reasonable to assume that the condition which is easiest to process may also be the most agreeable one to listen to, this does not necessarily have to be the case. Listeners' subjective preference was therefore tested separately in the next experiment.

5.4 Experiment 4: Subjective preference test

The three fast conditions of the latter experiment (Experiment 3) were evaluated perceptually using a subjective preference test. The question was whether listeners could actually indicate that one version (i.e., condition) of the same sentence sounded more 'agreeable' than an other. This dimension was chosen to evaluate listening effort or overall perceived quality of the three conditions (cf. van Bezooijen & van Heuven (1997) for a comprehensive chapter on evaluation of text-to-speech systems). Van Bezooijen & van Heuven distinguish between functional and judgment testing. Judgment (or opinion) testing is a procedure whereby a group of listeners is asked to judge the performance of a speech output system (often along a number of rating scales). Functional testing evaluates how well a speech system actually performs (e.g., in terms of intelligibility scores or in terms of successful task completion in an information-retrieval system). There is evidence that the results of judgment and functional evaluations converge: using the same group of listeners and stimuli, scaling results were found to highly correlate with the corresponding functional test scores (Pavlovic, Rossi & Espesser 1990). More importantly, high correlations were found between paired comparison results and reaction time data (word monitoring) by Delogu, Paolini & Sementina (1992), who evaluated the overall speech quality of synthesiser and vocoder systems and one human speaker.

The three fast conditions were evaluated by using the Comparative Mean Opinion Score (CMOS) test (ITU-P.800 1996). Listeners' subjective preference was tested by presenting pairs of utterances, and asking them whether version B sounded more agreeable than version A.

The hypothesis is that, in line with the previous results, listeners will judge the naturally produced fast version as less agreeable to listen to than the two artificially time-compressed versions. A competing hypothesis would be that, at a rate at which all conditions are still perfectly intelligible, listeners actually prefer to listen to naturally produced fast speech, simply because it sounds more natural and thus more agreeable to them. It is an open question whether listeners will weigh naturalness more heavily than ease of processing.

We hardly expect to find any preference for either of the two artificially time-compressed versions because the difference between these conditions is fairly small. Still, if anything, the linear compression condition is expected to be preferred over the copy-fast condition.

5.4.1 Material and Procedure

A selection of the speech material of experiment 3 (section 5.3) was used in the present experiment. In some of the 84 original utterance pairs, there were minor differences between the normal and fast rate versions of the sentence, with respect to their intonation patterns. Therefore, 45 test sentences (and 5 practice sentences) were selected in which the difference between the normal and natural-fast rate utterance was smallest. For each of the 45 test sentences, three pairs of fast conditions were evaluated (Linear vs. Copy-fast, Linear vs. Natural-fast, and Copy-fast vs. Natural-fast). A complementary (Latin square) design was set up in which these pairs were rotated over the sentence pairs and over three different lists. This was done to limit the duration of the test and to avoid training effects (van Santen 1993). In this design, each subject evaluates all pairs equally often, but he evaluates only one pair per sentence. Subsequent listener groups hear different pairs for each sentence.

The Comparative Mean Opinion Score procedure is as follows. Subjects are seated in front of a computer screen on which there are two buttons (one labelled 'version A' and one labelled 'version B'). Subjects listen to both members of the pair by first clicking on the version A button and then on the version B button (or in the reverse order). After listening to both members of each utterance pair, the subject is asked to indicate his or her preference by clicking on a 7-point scale, ranging from 'B is much more agreeable than A' (+3) to 'B is much less agreeable than A' (-3). In between are 'B is more agreeable than A' (+2), 'B is a little bit more agreeable than A' (+1), 'B and A are equally (un)agreeable' (0), and the reverse scale options (-1, -2).

It is conceivable that listeners will generally listen to sound A first and then to sound B. Consequently, listeners might have a bias towards perceiving sound B as more agreeable because they are by then familiar with the contents of the sentence. To avoid this effect, or any other unwanted bias effects, each condition within a pair appeared about equally often as A or B.

Subjects listened to the material over headphones while they were seated in a sound-treated booth. They were told that they could listen to the two members of the pair as often as they liked before giving their preference value. After they had indicated their CMOS value by clicking on one of the seven buttons on the scale, they could click a button 'Next' in order to hear the next sentence pair. The test lasted about 12 minutes. Before subjects started with the actual experiment, they were presented with 5 practice sentences, after which additional feedback or instruction was given, if necessary.

To each of the three experimental lists, 6 subjects were assigned. They were all students at Utrecht University, and were paid €5 for their participation.

5.4.2 Results

Each of the 18 listeners evaluated one condition pair per sentence, yielding 45 judgments per listener. The average perceptual scores for the three pairs, together with the standard errors, are given in Table 5.6.

Table 5.6. Mean perceptual scores of Comparative Mean Opinion Score (CMOS) test, plus standard errors, on a scale from +3 to –3. CMOS values are given for three pairs of conditions.

	Mean CMOS	S.E.
Linear vs. Copy-fast	-0.27	0.05
Linear vs. Natural-fast	-0.50	0.09
Copy-fast vs. Natural-fast	-0.02	0.09

A negative CMOS value indicates that the second member of the pair is judged as less agreeable than the first. Statistical analysis of these CMOS values takes the form of one-sample t-tests to test the hypothesis (H_1) that the mean CMOS value per pair differs significantly from zero ('0' equals the H_0 that there is no difference). The t-test for the first pair shows that the Copy-fast (nonlinear) time-compressed condition is judged as significantly less agreeable than the linearly time-compressed condition ($t(269)=-5.41$, $p<0.001$). Secondly, the Natural-fast condition is judged as less agreeable than the linearly time-compressed condition ($t(269)=-5.88$, $p<0.001$). Lastly, the difference between the Copy-fast time-compressed condition and the Natural-fast condition is not significant ($t(269)<1$, n.s.).

The results were also analysed by way of t-tests on item means and on subject means, yielding the same results (Linear vs. Copy-fast: ($t_1(17)=-3.4$, $p=0.003$), ($t_2(44)=-4.6$, $p<0.001$); Linear vs. natural fast: ($t_1(17)=-3.7$, $p=0.002$; $t_2(44)=-4.4$, $p<0.001$); and Copy-fast vs. Natural-fast: ($t_1(17)<1$, n.s.; $t_2(44)<1$, n.s.)).

The results confirm the hypothesis that listeners find the natural-fast condition less agreeable to listen to than the linearly time-compressed condition. A significant difference was also found between the two artificial time-compression conditions (in favour of linear compression), whereas this was not found in the phoneme detection experiment (section 5.3.2). Lastly, listeners did not prefer the copy-fast (nonlinear) time-compression condition over natural-fast speech. All in all, this means that even at a rate at which all three fast conditions are still perfectly intelligible, listeners have a slight preference for the condition which also proved easiest to process.

These results agree with the aforementioned study by Delogu et al. (1992) who also found that paired comparison results highly correlate with reaction time data. Furthermore, Delogu et al. found that the best discrimination between the systems (or conditions) was obtained with paired comparisons (of the four test methods that were

used, reaction time data showed the least discriminatory power). In our results, paired comparison also yields better discrimination between conditions than reaction time data.

5.5 Intermediate Discussion

The experiments described in this chapter have shown that speakers cannot speed up their speech rate without making speech processing more difficult for the listeners, even if the resulting speech is perfectly intelligible. Both changes in timing and segmental slurring slow down speech processing. Processing differences can be found between several types of perfectly intelligible speech. Even when speakers succeed in producing fast, yet intelligible speech, the inevitable reduced articulation makes processing more difficult for listeners.

Changes in timing alone had only a weak and non-significant effect on processing speed in experiment 3, whereas selective compression did slow down processing significantly in section 4.6. This discrepancy was attributed to the less fast rate in the present experiment 3. The nonlinearities at word-level reported in section 4.2 were not found so clearly in the present material: the reduction of stressed syllables (to 77% of their normal rate duration) did not differ much from the reduction of unstressed syllables (to 71% of their normal rate duration). One can conclude that the fast rate was not fast enough for such a strong nonlinear reduction 'strategy' to occur.

The 0.65/0.45 relation may be typical of very fast and slurred speech. The focus in the present chapter was on fast, yet intelligible speech. The present results have shown that even when the speaker is instructed to remain intelligible, the relatively small changes in timing already have a weakly negative effect on word processing. Secondly, at the fast, yet intelligible speech rate, listeners have a slight preference for linearly time-compressed speech, both over the natural-fast condition, and over the copy-fast condition. The linearly time-compressed condition, which also proved easiest to process, is judged as most agreeable to listen to. This strengthens our belief that changes in word-level timing are due to articulatory factors, and do not serve a communicative purpose.

This bring us to the remaining question of this study: whether there is anything at all in the speaker's nonlinear way of speeding up that might improve intelligibility or ease of processing over linear compression. This question is important for technological applications where artificial time compression may be desirable. So far, we have seen that imitating speakers' behaviour in speeding up, either with respect to

timing or to segmental content, only decreases intelligibility or slows down speech processing. Neither natural prosodic rules nor natural 'reduced articulation' contribute to perception of fast speech. This is in conflict with the claim of Covell et al. (1998) that imitating natural fast speech timing leads to significant improvement over linear time compression at very heavy rates of time compression. Their algorithm, Mach 1, is based on human strategies in speeding up, such as compressing pauses most and compressing stressed (i.e., sentence-accented) vowels least. In Chapter 4 it was already suggested that the improvement in comprehension, resulting from Mach1 compression over linear compression, may be due mainly to the fact that inter-phrasal pauses are compressed most. By doing this, the remaining speech can be time-compressed to a lesser extent in order to attain the same duration as in the linear time-compression condition. Thus, segmental intelligibility can be preserved better.

In the next section an experiment is described which investigates whether this aspect of natural fast speech timing does indeed improve processing of time-compressed speech.

5.6 Experiment 5: Compressing pauses more than speech[21]

The results obtained with the time-compression algorithm Mach 1, which is based on natural fast speech timing, did show a significant improvement in intelligibility and comprehension over linear compression (Covell et al. 1998). Several strategies underlie the algorithm: nonlinearities at word and at sentence level, and the nonlinearity that pauses are compressed more than the remaining speech. The separate contributions of these strategies were not evaluated in their paper (and never have been, Slaney personal communication). Covell et al. (1998) based their algorithm on duration studies of normal and fast rate speech. These studies reported that speakers either leave out, or strongly reduce, many of the inter-phrasal pauses. Goldman-Eisler (1968) and Trouvain & Grice (1999) also found that speech rate (including pauses) varies much more than articulation rate (without pauses).

Because the improvement of the Mach 1 algorithm might have been due mainly to this compress-pauses-most strategy (and less to smart compression elsewhere), we wanted to establish the intelligibility improvement by applying only the compress-pauses-most strategy. Note that we cannot quantify the separate contributions of each

21 The experiment described in this section was carried out by two students, Agnes Doorduin and Ritske Hermelink, as part of a practical course.

of the different strategies to the improvement found with Mach1. We can only establish what can be gained by merely applying the compress-pauses-most-strategy.

Several studies have shown that the presence of pauses at appropriate places improves comprehension and intelligibility of normal rate speech (Nooteboom, Scharpff & van Heuven 1990; Reich 1980; Sanderman & Collier 1997; Scharpff & van Heuven 1988). Furthermore, the intelligibility of speech of deaf speakers has also been reported to improve after speech pauses had been inserted at selected positions (Maassen 1985). Removing pauses from time-compressed speech should then only decrease its intelligibility. Still, listeners might rather listen to moderately time-compressed speech without pauses, than to speech which has pauses but also a higher compression factor. Henderson (1980) showed that an appropriate intonation fall at the end of a clause or sentence may help listeners more in perceptually segmenting speech than a pause. So, if the other boundary-marking cues are left intact (pre-boundary lengthening and intonation going down), the lack of pauses might not be a real problem for listeners.

The hypothesis is that a slower articulation rate is more important for the perception of heavily time-compressed speech than the presence of pauses. On the basis of the Mach 1 results, reducing speech pauses more than the remaining speech is expected to improve intelligibility over linear time compression.

5.6.1 Method

Material Short spoken news bulletin items were selected from the Corpus of Spoken Dutch.[22] There were 40 test fragments and 8 practice fragments. Each news bulletin fragment consisted of several sentences. Pauses were indicated in each waveform on the basis of visual and auditory inspection. As in the intelligibility experiments reported in Chapters 2 and 4, a compression ratio of 35% was used to avoid ceiling effects in the intelligibility test. So speech was time-compressed linearly to 35% of its original duration in the Linear Compression (LC) condition (i.e., speed-up factor 2.9). In condition 'Pauses Removed' (PR) all pauses were first removed. Note that this may not be entirely similar to what is done by the Mach 1 algorithm in which pauses are set to 150 ms and then the speech material is time-compressed further in a non-uniform way. It was found that long inter-phrase pauses can be reduced to 150 ms with little effect on comprehension. Below 100 to 150 ms, further inter-phrase pause compression may cause false pitch-reset percepts (Arons 1994). So, the Mach1 algorithm first reduces the pauses to 150 ms, but it is not entirely clear how much of

22 Thanks are due to Simo Goddijn, who assisted in making the news items available to us.

each pause remains in the ultimate time-compressed condition. We decided to remove the pauses completely. After that, the remaining speech fragment was reduced by way of linear compression to the same duration as in condition LC. By removing the pauses alone, the duration of the speech material had already been reduced to 90% of its original duration. After the pauses had been removed, the remaining speech was time-compressed to 39%, on average, in order to reach the same duration as in the LC condition (speed-up factor 2.6).

Design and procedure The two test versions were constructed for all 40 test fragments. Two experimental lists were made because each subject could be presented with the same fragment only once. The 40 test fragments were rotated over the two conditions on the two experimental lists in a Latin square design.

In previous intelligibility tests, a cloze procedure was used: after the auditory presentation, subjects had to fill in the missing word in a visually presented sentence. Now, the difference between the two test conditions was a higher-than-sentence-level factor: either the pauses between the sentences or phrases were extra heavily reduced or not. Although one would like to test the intelligibility or comprehension of the entire news fragment, subjects can obviously not be asked to type in an entire news item. Therefore, the entire news bulletin item was presented but only the intelligibility of the last sentence was tested. Intelligibility and comprehension of the first few sentences of the news fragment were assumed to improve the subject's comprehension of the last sentence of the news item.

The news items were played to the subjects over closed earphones. The subjects were tested individually and were seated in a sound-treated booth. A label was placed in the waveform at the start of the last sentence or phrase that the subjects were supposed to type in. Once the last fragment of the news item started, an exclamation mark appeared on the computer screen in front of the subject. Subjects were instructed that once the exclamation mark flashed, they were to memorise the sentence from that point. At the news item's offset, they had to type in the last sentence. When the exclamation mark had been placed at a clause boundary instead of at a sentence boundary, they had to type in the last clause.

Before the actual test started, subjects were presented with 8 practice items. In this way, subjects could get used to the task and could adapt to the very fast playback rate. Furthermore, both conditions were presented in the practice session. After the practice session, subjects could ask questions if anything was unclear. In the actual test, all test items were presented in random order. The experiment lasted about 25 minutes.

Subjects Ten subjects were assigned to both experimental lists. The 20 subjects were all students at Utrecht University and between 18 and 30 years of age. They were paid a small amount of money for their participation.

5.6.2 Results

The responses given by the subjects were scored for correct word identification per sentence (%). Only articles were left out of consideration because *de* ('the') might well be perceived as *een* ('a'), or vice versa. When all words of a sentence (excluding articles) had been transcribed correctly, the score was 100% correct. The mean correct recognition percentages in both conditions are given in Table 5.7.

Table 5.7. Mean percentages of correct word recognition for two compression methods, plus standard error of mean.

	% correct	s.e.
Linear Time compression (LC)	47	1
Compression after Pause Removal (PR)	56	1

Table 5.7 shows that there is an intelligibility difference between the two conditions. Because the data take the form of percentages of correct responses, and because there were only 2 conditions, paired t-tests were carried out for the statistical analysis of the data. In these paired t-tests, the mean identification percentages per subject or per item, in both conditions, were compared. The two conditions differ significantly, both by subjects and by items ($t_1(1,19)=-3.72$, $p=0.001$; $t_2(1,39)=-3.054$, $p=0.004$). The intelligibility of linearly time compression can be improved by removing pauses first. The perceptual disadvantage of not having pauses is outweighed by the perceptual advantage of preserving segmental intelligibility. Reducing the speech to either 35% of its original duration (in condition LC) or to 39% of its original duration (condition PR) yields a 9 percentpoint intelligibility advantage of PR over LC.

Although these results show that a significant improvement in intelligibility can be gained over linear time compression by removing pauses before time compression, this does not mean that the difference between Mach1-compressed speech and linearly time-compressed speech is due *only* to pause removal. The other nonlinearities that are integrated into the Mach1 algorithm could still also contribute to Mach1's intelligibility advantage. A study by He & Gupta (2001) might provide some insight into this issue. He & Gupta (2001) investigated the user benefits of nonlinear time compression by comparing linear time compression with two nonlinear compression systems. The first nonlinear compression system is a simple algorithm which combines pause removal

with linear time compression. The second is more sophisticated, and is based on the Mach1 algorithm as discussed earlier (Covell et al. 1998). Two main questions underly their study. First, what are the additional benefits of the two nonlinear time-compression algorithms over linear time compression at a comfortable speed-up rate? Secondly, how much better is the more sophisticated algorithm over the simpler nonlinear algorithm?

He & Gupta's Pause Removal plus Linear Time-Compression ('PR-Lin') method first detects pauses automatically and shortens them to 150 ms (pauses below 150 ms are left untouched). Arons (1994) found that long inter-phrase pauses can be reduced to 150 ms with little effect on comprehension. He & Gupta (2001) mention that pause removal typically shortens the speech by 10-25% before compression (cf. with the 10% reduction found in the present study). After pause removal, linear time compression is applied.

The more sophisticated nonlinear compression technique, or Adaptive time compression ('Adapt'), is a modification of the Mach 1 algorithm such that the achieved speed-up rate is the same as specified. The original Mach1 algorithm cannot guarantee a specific speed-up rate because it is 'open loop'. A preference study showed that listeners find the Adapt technique comparable to the original Mach1 technique.

The two nonlinear algorithms and the one linear time-compression algorithm were evaluated with respect to the following factors: highest intelligible speed, comprehension, subjective preference, and sustainable speed. The latter is defined as the speed-up factor that users will settle on when listening to long stretches of spoken text, yet still assuming some time pressure. Overall, the results of He & Gupta (2001) show that at moderately fast rates, comprehension is so high that no differences are found between the three types of compression. At the faster rates, the two nonlinear algorithms do significantly better than linear compression with respect to comprehension and subjective preference. Most importantly, however, is that there are hardly any differences between PR-Lin and Adapt.

He & Gupta (2001) conclude that the speed-up factor that listeners are most likely to use is about 1.6-1.7 times normal rate (i.e., reduction to about 60%). At this rate, there are hardly any differences between the three types of time-compression algorithms. This means that using a nonlinear technique is hardly worth the trouble. What is most interesting with respect to the present study is that, at the faster rates, the more complicated Adapt (Mach1) method yields no improvement over the simple PR-Lin method. This lends some support for the idea that the major gain of the Mach1 algorithm over linear compression is in pause-removal. This may be the only nonlinearity about fast speech timing that should be implemented when speech is to be artificially time-compresed to very fast rates. All the other nonlinear speed-up

characteristics which are to be found in natural fast speech either do not, or hardly improve intelligibility, and should therefore not be imitated.

5.7 General discussion

In this chapter the processing of natural fast speech was compared with that of artificially time-compressed speech which was originally produced at a normal speaking rate. For very fast and slurred speech (cf. experiment 2), the difference in processing time between these two types of speech is relatively large. However, this may have also been due to segmental intelligibility differences between the two speech conditions. In section 5.3 the difference between natural fast speech and time-compressed speech was evaluated again, by using natural fast speech that was found to be of perfect intelligibility. Furthermore, the changes in timing from normal to fast speech were evaluated separately from the segmental slurring that is involved in natural fast speech. The results of experiment 3 showed a small, but significant processing advantage of linearly time-compressed speech over naturally produced fast speech. This was caused by the joint contribution of segmental slurring and changed word-level timing. The difference between linearly time-compressed speech and speech to which the timing structure of natural fast speech was applied (copy-fast-speech-timing) was not significant, although there was a weak trend for slower speech processing in the condition with the changed timing.

The subjective listener's preference test served as a perceptual evaluation of the three different types of fast speech (experiment 4). Listeners judged the linear time-compression condition as slightly more agreeable to listen to than the natural-fast and the copy-fast-speech-timing condition. This confirms our earlier findings that the linearly compressed condition is easiest to process.

In section 5.6 we returned to very fast rates of speech to investigate the claim by Covell et al. (1998) that their nonlinear time-compression algorithm Mach1 improves identification over linear time compression. Their algorithm is based on several nonlinear compression strategies that speakers apply when they speak faster than normal. One of these strategies is removing or strongly reducing the pauses, and this pause-removal strategy is expected to be mainly responsible for the higher intelligibility of Mach1 compressed speech. Although pauses may be important for intelligibility and comprehension of speech, the slower articulation rate in the pauses-removed condition outweighed the advantage of having pauses. The experiment reported in 5.6, and study

by He & Gupta (2001), provide some further support for the idea that pause removal is the major factor in the improvement of Mach1 over Linear Compression.

It seems that speakers cannot speed up their speech rate beyond normal rate without making speech processing more difficult for listeners. The Horton & Keysar study (1996) suggested that speakers are less listener-oriented when they are asked to speed up. The speaker of the present study was instructed to remain intelligible, but the fact that he was also supposed to speak fast may still mean that the resulting speech is less tailored to the listener's needs than would normally be the case. Rather than claiming that the results of Chapters 4 and 5 provide evidence against the H&H theory, we would like to argue that the explanation for the segmental and prosodic characteristics of natural fast speech is not to be found in the assumption that speakers always try to help their listeners. Listeners just cannot speed up otherwise, and will then only choose to do this when the communicative situation allows it.

One must conclude that speeding up speech rate, globally or locally, is accompanied by a heavier processing load for the listeners. But this can still be functional, in that slower speech rate generally signals new and important information, and faster speech rate signals given or redundant information (Lindblom 1990). Or, in other words, when speakers speak faster during more redundant words, it seems rather unlikely that this should in the end be problematic for listeners. Bard and colleagues argue that natural variation in word pronunciation is not noise, but useful information (Bard, Sotillo & Aylett 2000; Bard, Sotillo, Kelly & Aylett 2001). Duration, prominence, and segmental reduction provide cues as to whether words are presented in isolation or in context, where the phrase boundaries are, whether the word is predictable or redundant etc. This information is mostly related to higher-level factors. In most theories of auditory word recognition, successful lexical access is dissociated from the recovery of the information contained in the variability in pronunciation. Bard et al. argue for a theory in which lower level lexical processes may suffer from variability in pronunciation, and may even fail to resolve lexical competition. This leaves room for higher-level information to aid the process of lexical competition. By means of a cross-modal identity priming experiment, they show that variation in pronunciation affects lexical access. Prime words were either taken from running spontaneous speech, or from clear list-read speech. Although robust priming was found for both conditions, the reduced tokens primed less than the list-read counterparts. The reduced tokens also proved less intelligible when they were presented in isolation. This means that both forms may access the intended lexical items, but that the reduced form may not be enough to resolve lexical competition. Consequently, lexical competition may be resolved only post-lexically through interaction with higher-level information.

So, even though lexical access may suffer from a faster speech rate at some points during the sentence, higher-level knowledge comes in later to resolve ambiguities and to make overall comprehension of the message faster.

The present results have shown that local word-recognition may be hindered by faster articulation, but we have not looked at more global levels of processing. It seems plausible that increased difficulty in word processing is in fact informative for higher-level processing of the message: the increased difficulty in itself signals the givenness or the redundancy of the word or phrase in question, and thus provides information on its role in the message as a whole.

5.8 Conclusion

Natural fast speech is more difficult to process than time-compressed speech. The results of the present chapter have shown that this also holds when the naturally produced fast speech is perfectly intelligible. The processing disadvantage of naturally produced fast speech is due to its changed timing, but also to its increased segmental slurring. Although research has shown that coarticulation and assimilation may help listeners in speech perception, as these provide cues to upcoming segments, increased coarticulation and assimilation seems to hamper speech processing. Therefore, the only aspect of naturally produced fast speech that should be imitated in order to make time-compressed speech more intelligible is shortening of pause duration. By compressing pauses more than the remaining speech, intelligibility is improved over linear time compression at very fast playback rates.

Natural fast articulation of a message makes processing more difficult for listeners than time compression of that same message, articulated at a normal rate. In everyday communication, faster articulation of words or phrases (i.e., a local increase in speech rate) is not just noise, but may be useful information. Even though lexical access may be hindered or delayed, the reduced pronunciation provides information on the word's role in the entire message. The care of pronunciation is often linked to the givenness or redundancy of a word. Thus, in the listener's head, higher-level information may interact with low-level lexical access processes in order to perceive and understand the message as a whole.

6

General Discussion and Conclusion

Abstract

The present study was set up to relate the perception of artificially time-compressed speech to that of naturally produced fast speech. In this chapter, a summary of the main findings is presented. The main findings are then discussed with respect to implications for theories on speech perception and production. Secondly, some practical conclusions were drawn with respect to applications of artificial time-compression methods. The chapter concludes with suggestions for future research and the general conclusions that can be drawn from this study.

6.1 Introduction

The main aim of the present study was to relate the perception of artificially time-compressed speech to that of naturally produced fast speech. Differences between the two types of fast speech were found both at the segmental and at the prosodic level. It was investigated whether one type of speech is more intelligible than the other, and, when both types of speech are perfectly intelligible, whether one type of speech is easier to process than the other. By looking at the contributions of segmental, lexical and prosodic factors, one might find an answer to the question of how listeners cope with speech that is presented to them at fast rates, perhaps even faster than they can produce themselves. This was worked out in a number of sub-issues. The main results of this study on the perception of fast speech are summarised below, each one headed by its respective question. In section 6.3 the implications of these findings are discussed in the light of theories on speech production and perception. Then the practical conclusions, with respect to how artificial time-compression of speech can be used in applications, are discussed in section 6.4. Suggestions for future research are given in section 6.5, followed by the main conclusions of this study on fast speech.

6.2 Summary of main results

6.2.1 Robustness and ease of processing

The experiments in which highly time-compressed speech was presented to listeners have disproved the idea that complete processing is possible as long as there is enough processing time available in between stretches of highly time-compressed speech (Foulke 1971). Instead, segments may become so short that they exceed the limits imposed by the temporal resolution of the hearing system. Even though an information handling limit may certainly play a role in the processing of longer stretches of time-compressed speech, the robustness against time-scale distortions of the speech signal was found to depend on the segmental make-up.

The experiments with highly time-compressed speech have also confirmed that speech perception is robust against time-scale distortion. Only when speech was presented to listeners at a rate that was almost three times faster than the normal rate (compression to 35%), did the recognition scores drop. Even though segmental intelligibility is severely affected by time compression, lexical redundancy helps to fill in

the difficult segments. The fact that speech remains intelligible at rates between two and three times the original rate shows that much of the speech signal is actually redundant and can be missed. At the same time, the faster rate is at the expense of ease of processing or listening effort. The results of Chapter 3 showed that faster playback of natural or synthetic speech leads to more errors in the phoneme detection task. Thus, faster playback speed increases the processing load for the listener. This illustrates the usefulness of this redundancy in speech: it makes speech processing easier and it makes it more robust against distortions from, e.g., interfering noise.

6.2.2 Adaptation to fast speech rates

Although the set-up of our experiments did not allow a computation of the exact number of sentences necessary to attain plateau performance, adaptation to fast rate takes only a limited amount of material. During the duration of the experiment, subjects approached plateau performance. This means that listeners do not have to be subjected to intensive training, or have to be as eager and motivated as the visually impaired mentioned in Chapter 1.

The results also suggested that the adaptation is not permanent. This type of adjustment is not like learning a 'trick' which is then stored in long-term memory, analogous to learning to read rotated letters and text (Kolers 1975). Kolers' results showed that subjects who had received extensive training in 'decoding' these letters were still able to apply this trick a year after they had been trained. The present results cannot tell us whether a permanent form of adaptation, or real learning, may occur when subjects are exposed to time-compressed speech on a regular basis, as the visually impaired mentioned in the Introduction chapter. However, the present data indicated that for our subjects, the adaptation effect had almost totally disappeared after five months. It was therefore assumed that adaptation or adjustment is not an explicit learning process, but is rather a gradual form of tuning in.

6.2.3 Higher speech rates in perception than in human speech production

In Chapters 4 and 5 fast speech rates were elicited from speakers. When speakers were pushed to speak as fast as they can, they attained articulation rates of about 10 syllables per second (their average normal speech rate being 6.7 syll./sec). This corresponds with the results of a study on German by Greisbach (1992), who found that 9 to 11 syllables per second was the maximal speed of reading aloud. In a subsequent study reported in Chapter 5 (section 3), more care was taken that the resulting fast speech should still be intelligible. In this study, the fast rate amounted to 8.5 syll./sec.

Even though speakers try very hard, they cannot even double their normal *articulation rate*. At the same time, perception of speech that is accelerated to a rate twice the normal rate is hardly problematic for listeners. This asymmetry between production and perception must then be caused by restrictions on speech production. Articulators need a certain minimum amount of time to reach articulatory or acoustic targets (Kiritani 1977; McClean 2000; Perkell 1997). The tongue tip may move fast, but the heavy jaw is a relatively slow articulator. These are restrictions on speed of articulation at the lowest physiological level. There may also be restrictions on the higher motor command level, or on higher speech planning levels. Some of the speakers in the present study made many more speech errors when they tried to attain fast rates: one speaker could not even do it because she continuously stumbled and had to start over again. Thus, the discrepancy between the rate of speech that humans can handle as listeners and the rate that they can produce is imposed by restrictions at several levels involved in speech production.

6.2.4 Naturally produced fast speech easier to process than artificially time-compressed speech?

This question was worked out in the segmental and in the timing domain. For both domains, the overall answer to this question is 'no'.

Segmental information

Even though some speakers, in line with the revised target undershoot model (Moon & Lindblom 1994), may succeed in attaining all acoustic/articulatory targets when they speed up their articulation rate, we assumed that the articulation rates in our study were so fast that 'reduced articulation' was inevitable. As expected, fast articulated speech turned out to be more difficult to process than artificially time-compressed speech, even at a rate at which both types of fast speech were still perfectly intelligble. This means that listeners cannot speed up their speech rate beyond normal rate without making processing more difficult for their listeners.

Timing information

Foulke (1971) raised the hypothesis that the temporal organisation of spoken language is relatively unimportant at a normal rate, but that it may become more critical to comprehension, the more the speech rate is increased. In line with previous studies (Gay 1978; Lehiste 1970; Max & Caruso 1997), speakers in the present study were found to speed up in a nonlinear way: some parts are reduced more than others. However, making the temporal organisation of artificially time-compressed speech

more like that of natural fast speech did not improve its intelligibility or ease of processing. Thus, there is a discrepancy between what speakers do when speeding up and what is beneficial to listeners.

Segmental and prosodic factors combined

A word's prosodic pattern is important for the mapping of the acoustic signal onto a stored word template. However, the prosodic pattern should not be so pronounced that the segmental content suffers from it. Reducing the unstressed syllable more than the stressed syllable, as speakers do, elicited quite a number of truncated (i.e., monosyllabic) responses to disyllabic target words at heavy rates of artificial time compression. At less extreme rates, applying the natural prosodic pattern of fast speech did not improve perception either, in particular for words with non-initial stress: presumably because it is easier to align the signal with word candidates when the word beginning is relatively salient.

The results of Chapter 3 also stressed this interaction between prosody and segmental content. In that study, the hyperarticulation of synthetic diphone speech, which consists only of initially stressed and hyperarticulated building blocks, did not turn out to be helpful when listeners are presented with artificially time-compressed speech. Rather, the difference between natural and synthetic speech even tended to become greater after the two speech conditions had been time-compressed. Fluctuation of speaking effort, which translates into intensity and into care of articulation, is important in speech perception. If all syllables are equally strong, speech is perceived as blurred in difficult listening conditions, and listeners find it difficult to group syllables together.

Segmental and prosodic factors both contribute to word recognition. Their relative contribution even seems to be time-scale independent: putting too much emphasis on either distorts the balance between a natural prosodic pattern and an intelligible speech signal. Or, in other words, the less the words deviate from their 'normal-rate form', the easier it is for the listener to map the incoming information onto the mental lexicon.

6.3 Implications for theories of speech perception and production

In this section the main results of this study are discussed in the light of theories of speech perception and speech production.

It is clear that speech perception is robust against distortions, e.g., time-scale distortions, exactly because speech is highly redundant. Segmental intelligibility, lexical

redundancy, prosodic patterns and higher-level context information generally provide the listener with a rich speech signal. Listeners make use of all possible types of information in order to derive the gist of what is being said. Furthermore, speech perception flexibly adapts to the listening situation. Adaptation to fast speech rates is not an explicit learning procedure, but mainly involves a gradual tuning in to the actual rate of speech.

Listeners are thus well able to cope with external noise, and can tune in to the speaker's characteristic, be it the speaker's rate of speech, dialect or accent or speech deficiency. In other words, speech perception is well adapted to real-life communication.

In the Introduction, the question was raised whether the asymmetry between the rate of speech that speakers can produce and the rate that they can handle as listeners is in conflict with the claims of the Motor theory (Liberman et al. 1967; Liberman & Mattingly 1985). It is difficult to infer testable predictions from the Motor theory because the authors claim that any type of speech (e.g., synthetic speech) will be treated as speech if it contains sufficiently coherent phonetic information. Consequently, the fact that people can listen to speech which is time-compressed to much faster rates than can be produced by human speakers is not a strong argument against the Motor theory because time-scaled speech is still sufficiently phonetically coherent.

However, naturally produced fast speech can be represented better in terms of articulatory gestures than artificially time-compressed speech. If listeners need an intermediate time-scaling operation in order to perceive time-compressed speech, one would expect that naturally produced fast speech is easier for listeners than time-compressed speech, assuming that all transformations take time.

Ohala's (1996) main argument against the claims of the Motor theory is that, for the purpose of maximum contrast between the units, the units of a language should be as different as possible in any signaling system. Because of the nonlinear relation between articulation and acoustics (i.e., the quantal nature of speech; Stevens (1989), it is the sounds of a language, rather than the underlying articulatory/gestural events, that are maximally contrastive. It makes sense that the reason why fast articulated speech becomes unintelligible is that segments get smeared or deleted, rather than that entire gestures become 'unrecoverable'. If listeners only attended to the gestures, they should have been able to cope with the increased smearing of segments. In that sense, the current fast speech data may provide some counterevidence against the claims of the Motor theory.

The major result of this study is that speakers cannot speed up their speech rate without making processing more difficult for the listener. The framework of the early and revised version of the target undershoot model can be applied to the point of

segmental intelligibility. In the early target undershoot model, as formulated in Lindblom (1963), Gay (1981), and Lindblom (1983), it was claimed that if a syllable is short, not all acoustic or articulatory targets can be reached. However, in the revised version of the target undershoot model (Moon & Lindblom 1994), it was acknowledged that undershoot in fast speech can be avoided by a change in articulatory effort (i.e., a change in speaking style) to compensate for the shorter segment durations. At relatively fast articulation rates, undershoot is practically inevitable because speakers are impeded by the inherent slowness of their articulators. Even though listeners may expect a certain amount of slurring when they are presented with fast speech, the increased articulatory overlap in very fast speech reduces segmental intelligibility and consequently, hinders the perception process.

How does this relate to perception theories of assimilation and coarticulation? With respect to *obligatory* assimilation processes, within-syllable nasal place assimilation was investigated by Weber (2001). In German (and in English and Dutch), in sequences such as *Bank* ('bank') the nasal's place of articulation assimilates to that of the following stop. Weber found that listeners were slower in detecting target stops /e.g., /k/) in items with illegal */nk/ clusters than in items with legal /ŋk/ clusters. The facilitatory effect in (obligatory) regressive assimilation is explained as 'high expectation': the assimilated segment reduces the set of possible following segments. On the other hand, with respect to *optional* assimilation processes, studies have shown, quite consistently, that there is no advantage for assimilated versions over unassimilated versions, given the appropriate phonological context. Optional regressive place assimilation in English neither speeds nor hinders speech processing: subjects were equally fast in detecting /g/ in *sweet girl* as in *sweek girl* (Koster 1987). The same was found by Gaskell and Marslen-Wilson and collegues (Gaskell & Marslen-Wilson 1996, 1998; Marslen-Wilson et al. 1995). At a normal speech rate, there is no perceptual advantage (or disadvantage!) for assimilated over unassimilated articulations of a word form. Quené & Krull (1999) found that listeners were faster in detecting the assimilated form than the unassimilated form at a normal rate, whereas the reverse was found when listeners were presented with a fast rate. In the fast rate condition, unassimilated forms were detected faster than the contextually more appropriate assimilated forms.

So, whereas perception is not hindered by coarticulation and assimilation at a normal rate, it becomes more difficult for listeners to map the assimilated acoustic signal onto stored underlying forms at faster rates of speech. Segmental intelligibility is then reduced to such an extent that recovering the intended word takes time. This may either indicate that these fast articulated versions are not stored in the mental lexicon, or are somehow less readily accessible than less reduced forms. Kohler (1990) describes assimilation as perceptually tolerated articulatory simplification. This agrees with the

predictions of the H&H model (Lindblom 1990): speakers will try to economise on speaking effort as long as the communicative situation allows it. Whereas reduced redundancy in the form of assimilation is not problematic for listeners in normal conditions, it does become problematic for word perception in fast speech.

The finding that speakers cannot speed up their speech rate without making processing more difficult for the listeners also holds at the prosodic level. The natural prosodic patterns of fast speech obviously do not contribute to speech intelligibility either. The more pronounced temporal pattern was argued to result from different specifications of stressed and unstressed syllables in the mental lexicon. It was argued, loosely on the basis of the target undershoot model, that the target values for stressed segments may be more strictly specified than for unstressed segments. Consequently, if more precision is required for the stressed syllables than for the unstressed syllables, the speaker simply cannot speed up that much during the production of stressed syllables.

Both segmental intelligibility and the natural *normal rate* temporal pattern contribute to the recognition of words and their relative contribution seems to be time-scale independent. This brings us to the notion of 'holistic listening'. This holistic processing supports the approach suggested by Wouters & Macon (2002a; 2002b). Wouters & Macon's acoustic analyses of natural speech showed that spectral rate of change of vowel transitions increases with linguistic prominence (Wouters & Macon 2002a). More prominent (i.e., lexically stressed or accented) syllables are produced with more articulatory effort, which translates into higher rates of spectral change. Wouters & Macon describe an approach for integrating this knowledge into a concatenative speech synthesis system in order to improve the perceived naturalness of synthetic speech (Wouters & Macon 2002a, 2002b). Their results show that controlling the articulation effort improves the perceived naturalness of the speech (Wouters & Macon 2002b). Thus, listeners are indeed sensitive to holistic characteristics of words and phrases.

Now, do the present production and perception results provide evidence against the H&H theory? Several studies have provided evidence that speakers are not as listener-oriented as some have thought them to be. Sotillo & Bard (1998) examined pronunciations of landmark names to investigate whether reductions in pronunciation are less where lexical competition is greater. They did not even find a trend towards less reduction for words with greater competitor sets. Stronger evidence against the H&H claims comes from Bard, Anderson et al. (2000), who found that listener's knowledge was irrelevant to the reductive effect of Givenness on duration and intelligibility of words in semi-spontaneous dialogues. Conversely, the Givenness effect on pronunciation was shown to depend only on what the speaker knew. This had also been found by Hawkins & Warren (1994): local phonetic variables (such as sentence accent and phonological and phonetic properties of individual segments) exert a greater

influence on intelligibility than whether or not the word had been used before in the conversation. Bard, Anderson et al. (2000) argue that the reason why speakers are so indifferent towards the listener's needs could be that the default case, in which the speaker proceeds from his or her own knowledge, is usually adequate. As long as listeners find the information sufficient, there is no pressure on the speaker to bother about computing the listener's needs.

These studies clearly demonstrate that speakers are not always as cooperative as the H&H theory claims them to be. When speakers are under time or task pressure, this will only become worse. Horton & Keysar (1996) observed that time pressure made speakers indifferent to what listeners knew. However, this indifference may be caused by restrictions on speech production. Under time or task pressure, speakers may not have the time to compute, nor address, the listeners' needs. Furthermore, the present results suggest that the way in which speakers speed up is the only possible way. Speakers are probably aware of the fact that the way in which they speed up a message is not beneficial to listeners, but they have no other option. The prediction that we inferred from the H&H theory, namely about the importance of the prosodic pattern found in natural fast speech, was proven to be wrong. Natural prosodic patterns do not contribute to speech intelligibility of fast speech. Contrary to our particular interpretation of the H&H theory, the explanation for the prosodic characteristics of natural fast speech is not to be found in the assumption that speakers always try to help their listeners. They rather result from the fact that speakers just cannot speed up in any other way. Speakers will therefore only choose to do this when the communicative situation allows it.

6.4 Practical conclusions

Although this study was not set up as a usability study, some practical conclusions may be drawn from it. First of all, there is the overall conclusion that the intelligibility of artificially time-compressed speech (Dutch, more specifically) cannot be improved by introducing some kind of nonlinear time compression based on natural fast speech behaviour. The other type of nonlinear time compression, which was set up to 'protect' the already short and unstressed syllables did not improve intelligibility either over linear compression. Chapter 5 demonstrates that the only nonlinear aspect of natural fast speech that does improve intelligibility over strictly linear compression is pause removal. Note, however, that this only becomes advantageous when compression rates are relatively high. He & Gupta's user benefit study (2001) showed that there was no

preference for nonlinear time compression at moderate compression rates. The compression rates that are most likely to be used, according to He & Gupta, are around 1.7 times faster than normal rate (compression to 60% of normal rate duration). Since there is no advantage yet of introducing nonlinear aspects (such as pause removal) at these moderate compression rates, He & Gupta conclude that the user benefits of nonlinear time compression are actually quite small, and do not outweigh the increased system complexity.

The listeners in the He & Gupta usability study were not presented with time-compressed speech on a regular basis, as were all the subjects who participated in the present study. One has to take into account, however, that more experienced listeners, such as the visually impaired, or perhaps people who would use time compression daily to listen to long audio documents while driving a car, might prefer higher playback rates. For those experienced listeners and for higher speed-up rates, linear time compression combined with pause removal might certainly be advantageous and 'worth the trouble'. This might be a topic for further usability research.

One study on nonlinear time compression that has not been mentioned before is that by Lee, Kim & Kim (1997). Their variable time-scale modification of speech is based on dividing the speech signal into transient and steady portions. This can be done by way of LPC cepstral distance measures or cross-correlation methods revealing the similarity between adjacent frames. After transient and steady portions have been identified, the technique proposed by Lee, Kim & Kim attains a certain target rate by modifying steady portions only. Subjective preference tests at different speech rates show that listeners prefer this type of nonlinear time compression over conventional linear time compression, especially at fast rates of speech. The results of Chapter 2 support the idea that intelligibility mainly collapses because certain rapid spectral transitions pose problems to listeners. The Lee, Kim & Kim paper does not provide details about the material that was used. One of the questions is what happens to rapid spectral transitions that occur in unaccented or unstressed syllables. Would the technique of not affecting rapid spectral transitions be helpful for the identification of polysyllabic words? Would this not create a rather unnatural timing pattern? Such signal processing methods might indeed be promising, but more detailed research needs to be done in order to investigate their merits for artificial time compression of longer stretches of running speech.

6.5 Suggestions for future research

As the focus of this thesis was mainly on the perception of fast speech, articulatory aspects of fast speech did not receive much attention. The idea of Kozhevnikov & Chistovich (1965) was that rate of speech production may not be specified in the motor program but presents the "speed of realisation of the program". Many later speech production studies have disproved this claim. First, there are duration studies which proved that speakers do not just change the overall speed of the motor program (Gay 1978; Lehiste 1970; Max & Caruso 1997). Secondly, there are many articulatory kinematics studies which have shown that patterns of muscle activity are not invariant across changes in speaking rate, e.g., Gay & Hirose (1973). Gay & Hirose demonstrated that changes in speaking rate are accompanied by complex reorganisation of motor activity patterns. Another finding from kinematic studies is that changes in speaking rate have differential effects for the movements corresponding to vowels and consonants: increasing rate causes increased velocities of movements corresponding to consonantal gestures, but less of an increase, or even a decrease, for vowel gesture movements (MacNeilage & Ladefoged 1976). In Fowler's coproduction model of coarticulation (Fowler 1980), these data have been used to support the existence of different underlying control structures for vowels and consonants. In the present study we have taken all articulatory reorganisation as either 'segmental reduction' or 'changed timing patterns' (at the syllable level only!). This is a gross oversimplification of the changes in articulation and timing that actually occur when speakers adapt their articulation rate.

Furthermore, it was already mentioned in the Introductory chapter that faster speech rate does not necessarily mean, for all speakers, that articulatory/acoustic targets will not be reached. Increases in speech rate are accompanied by different kinematic reorganisations across speakers (Kuehn & Moll 1976; Sonoda 1987). Some speakers show increases in articulatory velocity while maintaining the amount of articulator displacement; others maintain the movement velocity with a consequent decrease in displacement. This raises the question whether speakers who increase their articulatory velocity are indeed more intelligible than speakers who maintain their velocity and consequently decrease the amount of articulator displacement. Secondly, if the velocity-strategy speakers are more intelligible than the amount-of-displacement-strategy speakers, what causes this underlying difference in behaviour, and, perhaps more importantly, can speakers be trained to adopt the more efficient velocity strategy?

This ties in with studies that have attempted to define the acoustic characteristics of clear speech. We already mentioned the formant study by Moon & Lindblom (1994) which indicated that increased effort can make speech more intelligible, so that clear

speech is not merely louder and slower speech. People with larger vowel spaces had been shown to be more intelligible than people with small vowel spaces (Bond & Moore 1994; Bradlow, Torretta & Pisoni 1996). In a series of articles, Picheny, Durlach & Braida (1985; 1986; 1989) investigate the improvement in intelligibility associated with the attempt to speak more clearly when talking to hearing impaired listeners. In the third paper (Picheny et al. 1989) the authors attempt to assess which factors contribute to the enhanced intelligibility. The authors conclude that it is not possible to improve the intelligibility of conversational speech by uniform time expansion of the sentence durations. In the fourth paper, Uchanski, Choi, Braida, Reed & Durlach (1996) show that slowing down conversational speech (either uniformly or non-uniformly) did not improve intelligibility up to the level of the slower clear speech. Consequently, the conclusion is drawn that the high intelligibility of clear speech is due primarily to properties of words rather than suprasegmental characteristics. Cutler & Butterfield (1990), on the other hand, show that speakers may insert pauses before words in deliberately clear speech. Furthermore, syllables prior to certain 'difficult' words are lenghtened in order to cue the presence of a word boundary because speakers know that segmenting the continuous speech stream into words is difficult for listeners. Cutler & Butterfield also find evidence that prior syllable lenghtening occurs in particular before words beginning with a weak syllable. Cutler & Butterfield claim that this is a compensation strategy: normally, word boundaries coincide with the onset of strong syllables (in English). In order to mark word boundaries before weak syllables that might otherwise go unnoticed by English listeners, speakers apply explicit lengthening of the pre-boundary syllable.

Several studies have shown that background noise induces speakers to speak more clearly: this clear type of speech is called Lombard speech (Dreher & O'Neill 1957). Lombard speech shows systematic changes in encoding phonetic contrasts. Intelligibility of spoken digits was higher when the talkers had been exposed to wideband noise (Summers, Pisoni, Bernacki, Pedlow & Stokes 1988). Summers et al. found that the acoustic differences between speech produced in noise and speech produced in quiet (concerning word duration, F_0, rms amplitude, and spectral tilt) became more important as the signal-to-noise ratio decreased in the listener's environment. Bond & Moore (1994) show that native and non-native listeners agree on which speaker is difficult to understand. Moreover, Bond & Moore show that inadvertently clear speech shows similar acoustic-phonetic characteristics as deliberately produced clear speech.

Further differences between clear and conversational speech that are known to exist are consonant-to-vowel intensity ratio and consonant duration. Gordon-Salant (1986; 1987) found that elderly listeners with normal hearing and hearing impaired listeners

benefited from an increase in consonant-to-vowel intensity ratio, but not from the increased consonant duration, when these two properties were modified artificially. These findings concerning the amplification of consonantal regions were extended by Hazan & Simpson (1998; 2000). The effect of cue-enhancement also persisted in connected-speech material (Hazan & Simpson 1998); and cue-enhancement was also shown to be effective in improving speech intelligibility in non-native listeners (Hazan & Simpson 2000). Further research into the acoustic characteristics of clear speech may be desirable because of the wide range of practical applications: processing of speech that will ultimately be presented in degraded listening conditions, in the domain of second language acquisition, and in the pathological domain: for listeners with language or hearing disorders (cf. Tallal et al. (1996)).

One of the issues that was raised in the Introduction chapter was that the languages of the world can be categorised into broad classes on the basis of their rhythmic properties (Dauer 1983; Low et al. 2000; Ramus et al. 1999). The present study was carried out for Dutch, which is grouped into the stress-timed languages. Most of the literature cited in this study is on English, which is also a stress-timed language. Therefore, one might ask whether the prosodic production and perception results in this study are typical of stress-timed languages, rather than being language-universal. It seems in fact reasonable to assume that these results are heavily influenced by the language choice. One of the reasons why speakers shorten unstressed syllables more than stressed syllables when they speed up is that the stressed syllables have more strictly defined targets than the unstressed syllables. Vowel reduction, both spectral and durational, in unstressed syllables is a typical aspect of stress-timed, but not of syllable-timed languages. This would mean that the difference between stressed and unstressed syllables' target specifications may not be as large for the syllable-timed languages. Hence, there is no reason to expect such a nonlinear compression behaviour in speakers of syllable-timed languages. Further research on fast speech in syllable-timed languages would be necessary to validate this expectation.

One other aspect that did not receive much attention in the present study is that of intonation. Whenever prosodic differences between normal and fast speech were discussed, they mainly concerned duration. Speech rate has been shown to affect the choice of pitch markings: more marked pitch configurations may be replaced by more unmarked ones under time pressure, such that shades of intonational meaning are lost (Caspers 1994; Caspers & van Heuven 1995). When speakers are asked to speak as fast as they can, they may even produce rather monotonous speech. Thus, when the perception of artificially time-compressed speech is compared to that of naturally produced fast speech, one has to take into account that the two conditions may also differ with respect to the intonation contour. In line with the results of this thesis, it

seems that the intonational changes that speakers apply when they speed up mainly serve to make the speech production process easier, and not necessarily that of speech perception. To investigate whether, and to what extent, these intonational changes make perception more difficult is also an issue for further research.

Another suggestion for further research may concern the production and perception of slow speech. On the one hand, there is the practical type of research discussed above into whether (artificial) slowing down of speech could enhance intelligibility for certain listeners. A portable digital speech-rate converter was suggested as a useful tool for hearing impaired listeners to overcome their poor auditory temporal resolution (Matsushima et al. 1995). Their preliminary results suggest that the speech rate converter might be beneficial to elderly listeners as well, and for people trying to understand a foreign language. The latter listener group was also addressed in a French study which investigated the effect of selective enhancing and slowing down of certain spectral transitions (Colotte, Laprie & Bonneau 2001). When both modifications were applied simultaneously, intelligibility increased for students who were learning French as a foreign language. On the other hand, there is the more theoretical type of research into the characteristics of slow speech. As in speeding up articulation, slowing down articulation also involves all sorts of nonlinear segmental and prosodic effects. Whereas the prosodic pattern becomes more salient at fast rates of speech, the reverse might be true for slow speech. If this is the case, syllable length should become more uniform across stressed, unstressed, accented and unaccented syllables. What does this mean for the importance of the prosodic pattern in word recognition in slow speech? Within the segmental domain, Hertrich & Ackermann (1995) found that slowing of speaking rate resulted in a decrease of perseverative coarticulation, whereas the anticipatory effects remained unchanged. This corroborates the suggestion that different mechanisms underlie anticipatory and perseverative coarticulation. But how does this affect the perception or intelligibility of slow speech?

Another suggestion for further research is the aspect of resyllabification. Faster speech rate is often accompanied by resyllabification of consonants, cf. Stetson (1951). Articulatory studies have demonstrated that coda consonants tend to resyllabify as onset consonants, but only when this does not lead to phonotactically illegal onset clusters. Tuller & Kelso (1991) replicated Stetson's original experiment, and they also found a shift in perceived syllabic affiliation caused by important changes in syllabic organisation. MacKay (1974) found that the presence of coda consonants reduced the maximum rate at which words with various syllabic configurations could be produced. According to him, this indicates that syllables with codas are syntactically more complex than syllables without codas.

De Jong (2001) attempted to replicate and extend Stetson's and Tuller & Kelso's rate scaling experiments. In de Jong's experiments, speakers produced repetitions of simple CV and VC syllables in time to a metronome pacer which systematically changed in period. As had been found by Stetson (1951), the durational patterns of CV and VC syllables were different at slow rate, but converged at a faster rate. The VOT intervals of the plosives (in *eep* vs. *pee*) are consistently much longer for codas than onsets at slow rates. As rate increases, the coda VOT durations shorten until they are almost equal to interval durations consistent for onsets. CVs maintain their proportional durational structure (consisting of time to voice onset and time to vowel offset, and time to following consonant release). Still, importantly, the differences between fast CV and VC tokens were not completely neutralised, even though these were generally identified as CV tokens. It would be interesting to investigate this discrepancy between production and perception further. De Jong argues that the release of the (coda) consonant into the following vowel might produce a certain acoustic effect that overrides the production differences between CVs and VCs that remain at fast rates. It might also be interesting to find out whether listeners simply have a perceptual bias for CV structures, regardless of what they are presented with. Would listeners also be inclined to hear CV structures when they are presented with artificially time-compressed VC sequences that were originally articulated at a normal rate?

Furthermore, it might be interesting to look at resyllabification in normal sentences: the articulatory studies mentioned above have always used nonsense sequences. It is conceivable that the possibility of lexical confusion sometimes stops certain resyllabifications from occurring. It would be interesting to find out whether resyllabification is restricted to certain classes of words, and whether speakers actively decide at which points the listener would be able to deal with resyllabification or not. Vroomen & de Gelder investigated whether resyllabification hinders auditory word recognition (Vroomen & de Gelder 1999). They argue that listeners take the beginning of a syllable as the onset of a word, but that this strategy will fail in case of resyllabification. In the sequence *my bike is*, where /k/ may resyllabify as the onset of the last syllable, the last word does not begin at the onset of the last syllable. This is demonstrated for English by Cutler & Norris (1988), who found that the word *mint* is difficult to detect in the nonsense string *min.tayf*. The same was found for Dutch by Vroomen, van Zon & de Gelder (1996). The results of Vroomen & de Gelder's experiments (1999) show that the difference between resyllabified and non-resyllabified words results from a difference in how they are processed lexically. They rule out the possibility that it is the acoustic representation of the target phonemes themselves that makes detection more difficult in one case than in the other. Their results provide

support for the idea that words are segmented at the onset of a (strong) syllable (Cutler & Norris 1988), so a lexical access search is 'wrongfully' started at the onset of 'tis' of *de boot is*, and the system then has to backtrack. Even though there are some indications that such a Metrical Segmentation Strategy (Cutler & Norris 1988) might be less efficient in Dutch (Quené & Koster 1998), it is interesting to explore how resyllabification in fast speech may be one of the factors that make perception of fast articulated speech more difficult.

6.6 Conclusion

This thesis was set up to shed some light on the production and perception of fast speech. From the listener's point of view, the way in which speakers speed up a message is not the best way. Perception of fast speech is helped by a delicate balance between segmental intelligibility and a 'normal' prosodic pattern. Both aspects suffer when speakers are asked to speed up their articulation rate. At very fast rates of speech, artificially time-compressed speech has a higher intelligibility than naturally produced fast speech. At more moderately fast rates, and even when speakers succeed in producing perfectly intelligible fast speech, artificially time-compressed speech is easier to process, and is even judged to be slightly more agreeable to listen to than naturally produced fast speech. These results indicate that speakers are unable to speed up their speech otherwise: the segmental and prosodic changes that accompany naturally produced fast speech are inevitable, and do not serve to help the listener. The less words and phrases deviate from their 'normal-rate form', the easier it is for the listener to map the incoming information onto the mental lexicon.

The experiments have also shown that perception is quite flexible in handling less intelligible or degraded speech. Word perception may be somewhat delayed, but often higher-level information can help to resolve the lexical competition process. Even though the speech rates that were investigated in this study may have been beyond the range that is found in normal everyday communication, speech rate does vary continuously in normal running speech. Locally increased speech rate is not only disadvantageous in that faster fragments are more difficult to perceive. Although local word perception suffers, the faster speech rate signals that this specific fragment may be more redundant or less important. Generally, then, this variation in rate is not problematic, but is useful information for the more global levels of speech perception and understanding.

To conclude, we have also seen that speech perception is highly robust: when speech is presented at faster rates than one normally finds in connected speech, listeners are quite capable of dealing with this 'impoverished' speech. However, importantly, there is a cost involved, both in terms of increased effort and perhaps also in terms of accuracy. Once speech becomes less redundant because of its higher rate or increased smearing, listeners will have to put more effort into the perception process. Time compression of speech is therefore only attractive for those who wish to trade the increased effort against the obvious time-saving advantage.

NOTE

The following section is a methodological appendix on the technique of cross-modal semantic priming with partial primes. This part is unrelated to the dissertation as a whole. The original PhD proposal involved experiments with this technique. After we found that the experimental paradigm was not suited for research into multiple activation of word candidates, the proposal was changed completely into a study on the production and perception of fast speech. A section on cross-modal semantic priming is nevertheless included in this dissertation because to date, the negative findings on the suitability of the experimental paradigm have not been published elsewhere. Appendices and references, both those concerning the production and perception of fast speech, and those concerning the cross-modal semantic priming paradigm, are listed together at the back of this dissertation.

On Measuring Multiple Activation Using the Cross-modal Semantic Priming Technique

Abstract

Cross-modal semantic priming is often considered to be a valid technique for measuring the activation of multiple word candidates, particularly if used with auditory prime words cut off before their offset. In previous studies, the technique has been used to show that the activation of multiple candidates is modulated by preceding context. However, a recent model of word recognition has shown that semantic priming effects can at best be very weak when several words are active simultaneously. Previous research using this technique has indeed yielded inconsistent results with respect to priming of multiple words. Two priming experiments are reported here that specifically addresses the validity of this technique. Results show that consistent semantic priming is observed only after competition between multiple words has been resolved, with only one word remaining active. It is argued that the semantic priming technique cannot be used to investigate activation of multiple word candidates, and its use for that purpose should therefore be discontinued.

1 Introduction

In identifying spoken words, listeners use the sensory input to access representations in their mental lexicon. Current models of spoken-word recognition agree that word recognition involves two basic stages or processes. In the first or 'access' stage, multiple word candidates in the listeners' mental lexicon are activated that roughly match the available auditory input. In the second 'selection' or 'competition' stage, the number of lexical candidates narrows down when more auditory input becomes available, or when semantic context starts to influence the selection (Marslen-Wilson & Tyler 1980; McClelland & Elman 1986; Norris 1994; Norris et al. 2000). Recognition of a newly arriving word may be affected by the meanings of previous words or by the drift of the preceding context.

One of the issues in studying auditory word recognition models is the locus of the effect of higher-level sentence context. In other words, when during the word recognition process does sentence context have an effect? The more autonomous theories (Forster 1976; McQueen & Cutler 1997; Norris 1994) argue that lexical access and selection are based solely on the acoustic signal, and that sentence context only begins to have an effect during integration of the recognised word into a higher order semantic representation. In contrast, the more interactive theories (McClelland & Elman 1986; Morton 1969) argue that all types of information that can be used to restrict the candidate set are available immediately. The early version of the cohort model (Marslen-Wilson 1984; Marslen-Wilson & Welsh 1978) was a hybrid model: context was supposed to play a role during selection, after an autonomous lexical access stage during which a set of candidates had been activated solely on the basis of acoustic information.

In order to study the locus of the effect of sentential context on auditory word recognition, several studies have been carried out in which the cross-modal semantic priming technique was employed to measure the activation of multiple word candidates. The cross-modal semantic priming technique in its most common form was first introduced by Swinney (1979). The technique is based on spreading of activation from one lexical element to other semantically or associatively related elements (Collins & Loftus 1975). If listeners are required to make a lexical decision on a visually presented word (the visual probe, e.g., *money)* after hearing a prime word (e.g., *salary),* they react faster if the probe word and the auditory prime word are semantically or associatively related, than if these words are not related (e.g., *money* after unrelated control prime *piano).* Moreover, auditory prime words can be cut off before their offset, at a point where the acoustic information is not sufficient to identify the intended word uniquely. Supposedly, at that point in time there are more word candidates still activated. By presenting a visual probe related to one of these word candidates immediately following the presentation of a partial prime fragment, the activation of that word candidate can be measured. This seems an attractive technique, because it is assumed to tap directly into the ongoing process of spoken-word recognition. The technique is quite simple and non-intrusive, and it would allow us to follow the course of activation of competing word candidates. If so, the effect of sentence context on the activation of multiple word candidates can be established by embedding prime word fragments in either neutral or semantically

biasing sentences. This would help us in deciding an important issue in the study of the mental processes of spoken word recognition.

Robust priming effects have been reported with full primes (Meyer & Schvaneveldt 1971; Neely 1977; Swinney 1979). Obviously, it is important to know whether this technique, when applied to investigate activation of multiple word candidates, also gives reliable and robust results when partial primes are presented. The partial priming technique has already been criticised because the effects may be rather small and inconsistent (Jongenburger 1996), especially in sentence context (Gaskell & Marslen-Wilson 1996). Although some studies did obtain priming effects indicating activation of multiple word candidates, a number of other studies only found priming effects after the point at which the prime can be recognised on the basis of acoustic information alone. In those cases, as well as with full primes, only the activation of the prime itself can be established, because the competitor candidates have already been deactivated (an overview of these studies is provided in section 2). The recently developed Distributed Cohort Model (Gaskell & Marslen-Wilson 1997, 1999, 2002), which will henceforth be called the DC model, provides an explanation why semantic priming effects obtained with partial primes may not be as robust as some earlier studies suggests. In the DC model the process of speech perception is modelled as a recurrent neural network. In connectionist models such as the DC model, multiple representations must interfere with each other if they are active simultaneously. This was also modelled in two older models, notably, TRACE (McClelland & Elman 1986) and Shortlist (Norris 1994). These two models employ lateral inhibition between activated word candidates to reduce multiple activations. In the DC model, before the uniqueness point of a word, its semantic activation depends strongly on the number of candidates and their relative frequency that match the input so far.

In cross-modal *phonological* priming (also termed repetition priming, candidate priming or identity priming), the relation between the prime-probe pair is such that the probe is fully or partially identical to the prime in terms of its acoustic phonetic form. Subjects are presented with the auditory stimulus *The next word is por-* (/pɔr/), and then see the word *port* or *pork* (both are possible candidates for identification). Gaskell & Marslen-Wilson (1999) explain why the effects of phonological priming are much stronger than those of semantic priming if partial primes are presented. In their DC model, priming occurs if its lexical representation is more similar to the target representation than to an unrelated baseline. Phonologically, the word candidates are obviously coherent, but the semantic representations of the different candidates often have no meaning overlap at all. "In repetition priming, the target lexical representation is related to the prime representation in all dimensions, so recognition of the target can take advantage of overlap on both semantic and phonological nodes (…). In contrast, semantic priming relies on overlap in the semantic nodes alone." (Gaskell & Marslen-Wilson 1999, p.452).

Empirical results (described in Gaskell & Marslen-Wilson 2002) support their claim that the effects of phonological priming are much stronger than those of semantic priming if partial primes are presented. In their experiment, primes were presented either complete, or in two cut-off conditions. Semantic priming occurred only after the moment that the prime has become unambiguous onwards. By contrast, significant effects of phonological priming were found in all conditions and at all cut-off points.

These results make two important points. First, because activated candidates are co-represented in the same representational space, simultaneous activation of more than one candidate will necessarily create interference. Second, the extent of this interference will vary according to the coherence (phonological or semantic) of the representations being co-activated. Thus, phonological priming does occur because of the phonological coherence between the activated representations, but semantic priming is weak because the semantic representations are not coherent at all.

The present study was set up to find out whether cross-modal semantic priming using prime fragments is or is not a reliable technique to measure activation of multiple word candidates. Of course, if the technique is not sensitive enough to tap into the early stages of word recognition where more word candidates are still activated, then the effect of sentence context on the activation of those competing word candidates cannot be established either. First, a survey is presented of some relevant previous studies. Secondly, two experiments will be reported designed to test the validity of the technique. These experiments are partial replications of an important predecessor study, a cross-modal semantic priming experiment in which multiple activation of word candidates was observed. Some changes have been applied, to make sure that the results do not depend on an exact replication of the material and the design. Our reasoning is that if the latter would be the case, the chances are that any positive results do not reflect real effects of the experimental conditions, but rather are accidental artefacts of the way the experiment was set up. The results will show that significant priming effects can only be found after a single word candidate remains as the best-matching candidate.

2 Previous cross-modal priming studies

Tabossi (1996), describing and evaluating the cross-modal semantic technique, remarks that one of the advantages of the technique is that it "relies on a robust phenomenon" (p.573), namely semantic priming. This remark is based primarily on several studies using full primes to show priming effects. However, the use of partial primes is assumed to reveal more about the activation of candidates during spoken-word processing. Table 1 below summarises a number of such semantic priming studies using partial primes; the table lists the grand mean reaction time, and whether or not significant partial priming effects were found. Partial priming means that the prime is cut off before the isolation point (as determined in a gating task in a semantically neutral condition).

Table 1. Summary of the results of a number of semantic priming studies, with regard to partial priming effects and grand mean reaction time. All studies used lexical decision.

Study	Partial priming effects	significance level	Grand mean RT [23]
Zwitserlood (1989)	yes	p unknown	560 ms
Zwitserlood & Schriefers (1995)	no	n.s.	531 ms
Chwilla (1996)	no	n.s.	514 ms
Connine et al. (1994)	yes	$p<0.05$	690 ms
Moss et al. (1997)	raw: no normalised: yes	$p=0.06$ (raw); $p<0.05$ (normalised)	532 ms
Tabossi & Zardon (1993)	yes	$p<0.05$	585 ms
Jongenburger (1996)	no	n.s.	800 ms

Zwitserlood (1989) used the cross-modal semantic priming technique to study the activation of multiple word candidates by presenting fragments of prime words. She constructed pairs of words which were phonemically identical up to a certain point: the members of the Dutch word pair *salaris/salami* ('salary/salami') are identical up to the second syllable. Before the two word candidates start to diverge, the activation of both candidates was investigated by presenting a visual probe related to one of the word candidates, for example *geld* ('money') related to *salaris*; or *worst* ('sausage') related to *salami*. The results of the Zwitserlood (1989) study showed activation of both word candidates after the presentation of short prime fragments embedded in the carrier sentence and neutral sentence conditions. Once the phonetic information started to favour the actual word, the activation levels of the actual word and its competitor diverged. Secondly, at the position chosen to tap into the selection stage, sentence context had a positive effect on the activation of the actual word and a negative effect on the activation of the competitor compared to the activations of both candidates in semantically neutral sentences at that point. The conclusion was drawn that context can affect activations of word candidates before the auditory information is sufficient to isolate the actual word.

Other cross-modal priming studies have yielded inconsistent results, both with respect to the activation of multiple word candidates and with respect to the role of context. Chwilla (1996) carried out a semantic priming experiment to test a different experimental technique. Part of Chwilla's test material was the original Zwitserlood (1989) material, for which early priming effects were expected similar to those of Zwitserlood (1989). However, Chwilla's results (1996) showed that there were no priming effects for either word candidate before the recognition point. After the recognition point, there was only a significant priming effect for the actual word, which means that multiple activation of word candidates could not be shown.

[23] Grand mean RT is computed over test and control conditions

A study by Zwitserlood & Schriefers (1995) examined the role of processing time. In normal listening conditions, the more acoustic information is available, the more time is available to process earlier parts of the sensory input. Zwitserlood & Schriefers (1995) set up a cross-modal semantic priming experiment to study effects of increasing acoustic information and extra processing time separately. Their study showed that priming effects were obtained for two conditions which provided an extra amount of processing time: i.e., for a long prime fragment condition and for a short prime fragment condition with a certain time-lag of some 100 ms between the offset of the prime fragment and the presentation of the probe (delayed priming). There was no priming effect for the condition in which there was no time-lag between the offset of the short prime fragment and the presentation of the probe. The long prime position may have been actually at or near the isolation point, because no gating experiment was performed to establish isolation or recognition points. Furthermore, as Zwitserlood & Schriefers (1995) did not measure the activation of competitors, the study did not show activation of multiple word candidates. In a study by Connine, Blasko & Wang (1994) multiple activation was shown for prime words that were acoustically and lexically ambiguous with respect to the voicing value of the initial consonants (e.g., *dip/tip)*. However, the fact that lexical decision times were relatively long (mean RT of 690 ms) casts doubt on the on-line nature of these effects. Moss, McCormick & Tyler (1997) also used partial primes in their semantic priming study. The raw reaction time data were normalised, because of large differences between subject groups. Although the normalised data showed significant priming effects both at the full prime cut-off point and at the isolation point, the priming effect at the isolation point was only marginally significant in an analysis on the raw data ($p=0.06$).

Tabossi & Zardon (1993) studied the activation of the meanings of ambiguous words, in an experiment similar to that of Swinney (1979). Tabossi & Zardon (1993) showed significant facilitation at a point before the uniqueness point in a number of their context conditions. However, the prime was not cut off but remained audible during lexical decision. Hence, subjects receive additional acoustic information while they are processing the preceding information. This entails that one cannot be certain that the technique has actually tapped into early processing.

A study by Jongenburger (1996) focused on the role of lexical stress during spoken-word processing. Even at a point 750 ms after prime offset, consistent priming effects could not be found either for the prime word itself, or for the prime's stress partner, regardless of the context the primes were embedded in.

Moss & Marslen-Wilson (1993) examined activation of non-associated and associated words before and after prime offset. For the semantic property probes, they found facilitation in the biasing context conditions, but not in the neutral condition. Furthermore, there was no increase in facilitation with later cut-off point in the neutral condition: even 200 ms after prime offset, there was no facilitation in the neutral condition. For the associated probes they did find facilitation in the neutral condition, but still the facilitation did not increase from the early to the late cut-off point. This suggests that either the word was already selected from the set of candidates at the early position, or the technique does not allow us to follow the course of activation very accurately.

This literature survey shows that few studies have obtained early priming effects, and even fewer have shown activation of multiple word candidates. The results obtained with the cross-modal priming technique are far from consistent. Moreover, some studies that did obtain partial priming effects had relatively long reaction times, which makes the on-line nature of these effects questionable. The study of Zwitserlood & Schriefers (1995) suggested that extra processing time is needed for the early priming effects to occur. This would imply that the technique is not a reliable instrument to measure activation of word candidates in an on-line way. The next two sections report on two experiments designed to test whether cross-modal semantic priming with partial primes can be a useful and reliable instrument to study multiple activation.

3 Replication experiment I

Zwitserlood's study (1989) provides a solid framework to put the cross-modal semantic priming technique to a test. However, if we were to carry out a strict replication of her experiment, using exactly the same materials and design, the obtained effects might be due to the choice of material and design instead of the experimental conditions. Therefore, some changes were applied to the experimental set-up. If the priming effects are robust and generalisable, partial priming effects will be obtained with slightly different material as well. Furthermore, there are some aspects about Zwitserlood's experimental set-up and material which can be improved in order to yield a more severe test of the technique. These aspects will be discussed below in the Method section.

3.1 Method

Materials

Word materials selected for this experiment consisted of the following: 24 Actual Words to be used as primes, and 24 Control Words to be used as 'quasi' primes. Each Control Word matched its Actual Word in number of syllables, stress pattern, phonological pattern, and frequency of usage (based on Celex (Celex 1990)).

For each Actual Word, a Competitor Word was selected, such that there was considerable auditory overlap from word onset onward between Actual Word and Competitor. Unlike Zwitserlood (1989) we decided to base the choice of the Competitors not only on auditory overlap (via lexicon look-up), but the Competitors had to be given as frequent responses in the gating experiment as well in order to assure that these words really competed with the actual word (see below in the Gating study section).

For each Actual Word and each Competitor Word a semantically associated Visual Probe was selected, on the basis of an association test.

Sentence materials consisted of:

- The same carrier sentence for each of the 24 Actual Words and Control Words: namely Het volgende woord is — ('The next word is –').

- 24 biasing sentences, ending with the Actual Word (which primes the associated visual probe word). An example for the Actual Word schapen 'sheep' is the biasing context The farmer had negotiated a long time about the price. Finally he bought the sheep.
- 24 control sentences, ending in the control words. These sentences differed from the biasing sentences, although they were matched in informational value with respect to the sentence-final prime word in those sentences (Marslen-Wilson & Zwitserlood 1989; Zwitserlood & Schriefers 1995).

So far materials are basically the same as in Zwitserlood (1989). There were some differences in competitors and in visual probes. Some of Zwitserlood's competitors were replaced by ones that were more frequently mentioned in a gating experiment (see below). A list of all materials is presented in Appendix F, together with the original material used by Zwitserlood.

Recordings

The sentence material was recorded on DAT tape with a Sennheiser ME 30 microphone, in a sound-treated booth. A male native speaker of Standard Dutch read out the visually presented sentence material at a normal speaking rate. Distance between microphone and the speaker's mouth was approximately 40 cm.

Gating study and cut-off points

A gating study was carried out to determine isolation points (cut-off points) for the 24 prime words in both sentence contexts. The actual words spoken in the carrier sentences were copied and used to substitute the actual words excised from the biasing sentences. There were no audible consequences of this cut-and-paste operation. In this way we ensured that the same acoustic tokens were used in the two sentence conditions. The items in the two sentence conditions were presented to two groups of 12 listeners: one group heard the items in the carrier sentence condition; the other group heard the items in the biasing sentence condition. All listeners were students at Utrecht University and they were paid for their participation.

The listening material was presented over headphones to each subject individually. The subject was seated in a sound-treated booth. Subjects first heard the sentence with the final (actual) word cut off after the first 20 ms. On subsequent presentations, the fragment length of the actual word was increased in 35 ms steps until the entire word was made audible, and on each presentation subjects were asked to write down what they thought the word was going to be. The mean isolation point (defined as the mean point at which subjects first came up with the actual word without changing their response at later gates) in the biasing sentence condition served as cut-off point 1. At this point, the Actual Word, its Competitor and often other candidates are still compatible with the sensory input. The isolation point in the carrier sentence condition, of course always later in the word than cut-off point 1, was used as cut-off point 2. At this point the sensory input is only compatible with Actual Word and Competitor (with a preference in the gating study towards the former). Zwitserlood (1989) assumed that

the first cut-off point could be used to tap the lexical access stage, whereas the later cut-off point tapped the selection stage. The distinction between these stages is not drawn in the present experiment, because the main aim of the present experiment is to evaluate whether multiple activation can be found and secondly, to investigate whether context plays a role before the prime word can be isolated on the basis of acoustic information alone.

Cut-off point 3 was located at prime offset. This was done to compare the effect of the partial primes with the effect of full primes.[24]

The competitor choice was not based only on the criterion of auditory overlap, but also on the competitors named in the carrier sentence condition in the gating task. This was done to make it more likely that competitors were actually activated by the partial primes compatible with them. This led to the replacement of 5 of Zwitserlood's competitors (cf. Appendix F). In the gating task, the actual word kaas ('cheese') received more competition from *kaars* or *kaarsen* ('candle/candles') than from *kabel* ('cable'). Therefore, *kaarsen* was chosen instead of *kabel*.

Association test visual probes

An association test was carried out for three reasons. First, for competitors different from those used by Zwitserlood, different associates had to be selected as well. Secondly, for probe words which were associates of two primes, the second best associate had to be found, in order to avoid repetition of probes in the present within-subject design. A third reason to carry out an association test was that Zwitserlood relied on the association studies of Van der Made-van Bekkum (1973) and De Groot (1980). Some of these associates had now become dated and needed to be replaced.

The association test was carried out with 52 participants who were instructed to write down the three words that first came to mind on reading the words (in the order with which they came to mind). Whether a certain word was given as the first, second or third associate determined its weight factor. This weight factor, combined with the total number of subjects that named a particular associate, were used as criteria to select the best associate. All in all, 20 of the 48 visual probes differ from those used by Zwitserlood (cf. Appendix F).

Experimental design

In this experiment, three fixed factors were varied in the auditory prime word: Probe Type (i.e., related to Actual Word or Competitor), Prime Amount (cut-off point 1, 2 or 3), and Context (Carrier Sentence vs. Biasing Sentence). For a single set of items, there were 2x2 utterances involved, with each utterance cut off at 3 different points in the utterance. In the cross-modal priming task, activation of the auditory

[24] Note that there was no 'zero' cut-off point, at the onset of the prime. This was omitted because Zwitserlood never found any difference between the zero and control conditions. In addition, our cut-off point 2 combines Zwitserlood's positions 1 and 2. To sum up, the 0+4 cut-off points in the predecessor experiment were collapsed into 3 positions in the present experiment (her zero: now discarded; her 1 and 2: now 1; her 3: now 2; her 4: now discarded; new: 3 for full prime).

prime word in the spoken utterance (e.g., *salaris* 'salary') is operationalised as the reaction time to lexical decision on a visual probe word, which is semantically related to the prime word (e.g., *geld* 'money'). Table 2 presents a schematic design of the experiment, with the factor Cut-off point ignored. It shows that the utterances containing the actual word and its competitor were combined with two different visual probe words: the visual probe co-varied with the word type of the auditory prime. In order to determine baseline or control values for these visual probe words, each probe was also presented after a control auditory prime which was not related to the probe word (e.g., *piano* 'piano': piano - geld, and piano - worst). Hence, each test version of an item set had its matching control version, with the same visual probe, but with an auditory utterance not containing a semantic prime (cf. Table 2).

Table 2. Schematic design of the experiment, including the Test-vs-Control factor, but excluding the factor Cut-off point.

Prime Type	Context	Probe Type	Example utterance	Visual Probe	Block
Test	Carrier	Actual	Het volgende woord is salaris.[25]	geld	1
Control		Actual	Het volgende woord is piano.	geld	2
Test	Carrier	Comp.	Het volgende woord is salaris.	worst	2
Control		Comp.	Het volgende woord is piano.	worst	1
Test	Biasing	Actual	Marleen is zeer tevreden over haar nieuwe baan. Ze heeft ook een prima salaris.[26]	geld	3
Control		Actual	Het huis was sfeervol ingericht. In de woonkamer stond een oude piano.[27]	geld	4
Test	Biasing	Comp.	Marleen is zeer tevreden over haar nieuwe baan. Ze heeft ook een prima salaris.	worst	4
Control		Comp.	Het huis was sfeervol ingericht. In de woonkamer stond een oude piano.	worst	3

Ideally, the priming effect would be expressed as the difference (in ms of reaction time) within subjects between the unprimed (piano) and primed (salaris) presentations of the same visual probe (geld). In practice, this is impossible due to the long-term priming between the two successive presentations of the same probe word in an experiment. To avoid such long-term priming, the primed and unprimed versions must be presented to different subjects, with their intrinsic individual differences in reaction time. Hence, the priming effect could not be determined directly (within subject and within item), but only indirectly, as an additional fixed factor named Prime Type (Test versus Control; evaluated between subjects within item, or between items within subject). Semantic priming should

[25] 'The next word is ..' (carrier phrase)
[26] 'Marlene is very satisfied with her new job. She has indeed an excellent-'
[27] 'The house was nicely decorated. In the living room stood an old piano.'

manifest itself as a main effect of this latter factor. A full factorial design of these factors would have yielded 2x3x2x2=24 experimental versions. Counterbalancing these versions over listeners would require 24 groups of listeners. This design can be reduced. If listeners hear the auditory prime word salaris followed by the visual probe geld, they are still unprimed with regard to the auditory control word piano followed by worst. Hence, this latter version can also be presented to the same listener. Test-vs-Control and Word Type are thus combined into 2 rather than 2x2 experimental versions.

In the actual experiment, Table 2 was repeated three times for the three cut-off points, yielding 12 blocks in total. The main improvement of this design over Zwitserlood (1989) is that all listeners participated in all experimental versions, and not in a subset of versions. This enabled us to separate listener effects from (fixed) experimental effects, and to do this without the risk of introducing spurious differences while removing listener effects.

Subjects

To each of the 12 lists, 10 listeners were randomly assigned. These 120 listeners (all Utrecht University students) had not participated in the gating study and were paid a small sum for their participation.

Procedure

The visual probes were presented during a 50-ms interval, which started at the acoustic offset of the prime or at the cut-off point of the prime fragment. The short presentation interval was chosen in order to obtain fast reaction times with reduced intra-listener variance (Zwitserlood 1989). Subjects were instructed to listen carefully to the auditory material and to give a lexical decision response to the visual probe as fast and at the same time as accurately as possible, by pressing one of two buttons of a button box, one for 'yes' in case of a real word, and one for 'no' in case of a nonsense word. The 'yes' button was always under the index finger of the participant's dominant hand. The participant with the best performance (in terms of accuracy and speed) would receive a bonus reward of NLG 25. Listeners were also informed that they would be presented with a recall test after the priming experiment. For this recall test, they were asked to tick the sentences they had heard.

To make sure that subjects also paid attention to the prime fragments themselves, an extra task was added to the test. After approximately 15% of the items, after subjects had given their lexical decision response, a message appeared on the screen: *Repeat the last word you heard*. Subjects could choose to simply repeat the fragment or to repeat and finish the fragment. Because subjects might be distracted by the repeat command, at least one filler sentence (without the repeat command) followed before the next test or control stimulus was presented. Apart from this restriction, the order of the stimuli was randomised. Another way to avoid that subjects would not pay enough attention to the auditory information was to vary the cut-off point throughout the sentences. Therefore, a number of the filler items were cut off halfway or at the beginning of the sentences.

Reaction times were measured from the onset of the presentation of the visual probe until either of the response buttons was pressed. In the test part of the experiment, there were 58 filler items in

addition to the 48 test and control items. The entire experimental set was balanced for words and nonwords. Nonwords were always phonotactically and orthographically possible Dutch words. There was a practice session containing both real word probes and nonwords. If subjects had not responded within 3 seconds from the onset of the probe presentation, they proceeded with the next sentence. Reaction times were measured from the onset of the presentation of the visual probe until one of the response buttons was pressed.

3.2 Results

Figure 1 below presents the raw mean lexical decision times in all test and control conditions.

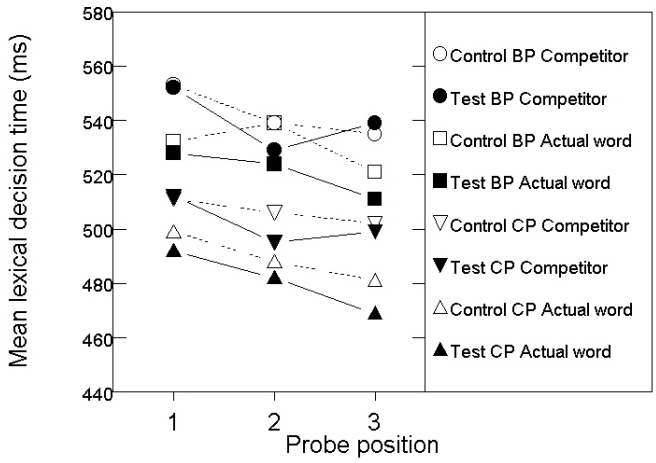

Figure 1. Mean lexical decision times (in ms) to Actual Word and Competitor probes in Carrier Sentence (CP) and Biasing Sentence (BP) conditions. Test conditions are represented by solid lines; control conditions by dotted lines.

Priming effects (each test condition – corresponding control) are listed in Table 3.

Table 3. Mean priming effect (test - control condition) for actual word and competitor in both sentence contexts (Carrier and Biasing Phrase), at three prime cut-off points.

		1	2	3
Carrier	Actual Word	-6	-6	-11
	Competitor	1	-10	-4
Biasing	Actual Word	-5	-16	-10
	Competitor	0	-10	4

In the carrier context, activation for both the Actual Word and the Competitor is expected at cut-off point 1, because the stimulus information up to that point is equally compatible with both. However, there is very little facilitation, if any, for the Actual Word, and none for its Competitor. At cut-off point 2, where more facilitation was expected for the Actual Word than for the Competitor, very small effects were found for both. Cut-off point 3, the full prime, seems to show a small facilitation effect for the Actual Word, and none for the Competitor, as predicted.

In the biasing context, the facilitation effect for the Actual word is expected to be greater than in the carrier context. Overall, this is not the case. Apart from priming effects, a main effect of sentence context was also predicted, with faster RTs in Biasing Sentence context as compared to Carrier Sentence context, for the Actual Word only. This prediction was clearly not supported by the results, which indicate a difference in the opposite direction. This unexpected difference may be related to a difference in processing strategies used by the listeners. Given the fixed carrier template, listeners need less attention to process the auditory stimulus. In addition, they can estimate the timing of the visual probe better in the fixed carrier context. Both effects may have led to faster RTs in carrier context.

The data were subjected to two analyses of variance, with either subjects or items as repeated measures. Missing observations (due to false rejections) were replaced by the mean of the subject mean in that experimental condition and the item mean for that experimental condition. The results of the statistical analyses are summed up in Table 4 below.

Table 4. Statistical results of subject and item ANOVAs

Effect	Analysis on subjects	Analysis on items
Prime Type (Test vs. Control)	$F_1(1,119)=3.95$, $p=0.049$	$F_2(1,23)=4.84$, $p=0.038$
Context (Biasing vs. Carrier Phrase)	$F_1(1,119)=148.3$, $p<0.001$	$F_2(1,23)=75.2$, $p<0.001$
Probe Type (Actual vs. Comp.)	$F_1(1,119)=31.99$, $p<0.001$	$F_2(1,23)=5.31$, $p=0.031$
Prime Amount (cut-off point)	$F_1(2,118)=8.65$, $p<0.001$	$F_2(2,22)=5.82$, $p=0.009$
Prime Type * Context	$F_1(1,119)<1$, n.s.	$F_2(1,23)<1$, n.s
Prime Type * Probe Type	$F_1(1,119)<1$, n.s.	$F_2(1,23)<1$, n.s.
Prime Type * Prime Amount	$F_1(2,118)<1$, n.s.	$F_2(2,22)<1$, n.s.
Context * Probe Type	$F_1(1,119)<1$, n.s.	$F_2(1,23)<1$, n.s.
Context * Prime Amount	$F_1(2,118)<1$, n.s.	$F_2(2,22)<1$, n.s.
Probe Type * Prime Amount	$F_1(2,118)=2.6$, $p=0.079$	$F_2(2,22)=4.49$, $p=0.023$
Prime Type * Context * Amount	$F_1(1,119)<1$, n.s.	$F_2(1,23)<1$, n.s.
Prime Type * Probe Type * Amount	$F_1(2,118)<1$, n.s.	$F_2(2,22)<1$, n.s.
Prime Type * Context * Amount	$F_1(2,118)<1$, n.s.	$F_2(2,22)<1$, n.s.
Prime Type * Context * Probe Type * Amount	$F_1(2,118)<1$, n.s.	$F_2(2,22)<1$, n.s.

Significant main effects were observed for Prime Type (Test-Control) ($F_1(1,119)=3.95$, $p=0.049$; $F_2(1,23)=4.84$, $p=0.038$), indicating that RTs in test conditions were generally faster than in control conditions. The significant effect of Context ($F_1(1,119)=148.3$, $p<0.001$; $F_2(1,23)=75.2$, $p<0.001$) shows that responses were faster in the Carrier than in the Biasing context. The significant main effect of Probe Type ($F_1(1,119)=31.99$, $p<0.001$; $F_2(1,23)=5.31$, $p=0.031$) shows that responses to actual word probes were faster than those to competitor probes, both in test and in control conditions. Lastly, the significant effect of Prime Amount ($F_1(2,118)=8.65$, $p<0.001$; $F_2(2,22)=5.82$, $p=0.009$) shows that responses were faster, the more of the prime was presented, again both in test and in control conditions. Of all interaction effects, that between Probe Type and Prime Amount was the only one which approached significance ($F_1(2,118)=2.6$, $p=0.079$; $F_2(2,22)=4.49$, $p=0.023$).

If activation for Actual Word and Competitor diverges only after cut-off point 1, as predicted, then this would yield a three-way interaction effect between Prime Type, Probe Type, and Prime Amount. This interaction was far from significant. Hence, our results do not support the idea that there is multiple activation in cut-off point 1, or that activation levels diverge between Actual Word and Competitor in later cut-off points. In the biasing context, priming was predicted for the Actual Word at cut-off point 1, with no or less priming for the Competitor, and with increasing divergence in priming between Actual Word and Competitor over later cut-off points. This should yield a four-way interaction between Prime Type, Context, Probe Type, and Prime Amount. This interaction was insignificant in both analyses.

Finally, in order to evaluate the facilitation by full primes (rather than by fragments of the prime words), the results were analysed for cut-off point 3 only, in two separate analyses of variance, with subjects and with items as repeated measures. The results of these analyses are summed up in Table 5 below.

Table 5. Statistical results (subject and item analysis) for priming data only after full presentation of the prime word

Effect	Analysis on subjects	Analysis on items
Prime Type	$F_1(1,119)=1.19$, n.s.	$F_2(1,23)=1.15$, n.s.
Context	$F_1(1,119)=51.3$, $p<0.001$	$F_2(1,23)=37.4$, $p<0.001$
Probe Type	$F_1(1,119)=36.7$, $p<0.001$	$F_2(1,23)=8.13$, $p=0.009$
Prime Type * Context	$F_1(1,119)<1$, n.s.	$F_2(1,23)<1$, n.s.
Prime Type * Probe Type	$F_1(1,119)=1.49$, n.s.	$F_2(1,23)<1$, n.s.
Context * Probe Type	$F_1(1,119)<1$, n.s.	$F_2(1,23)<1$, n.s.
Prime Type * Context * Probe Type	$F_1(1,119)<1$, n.s.	$F_2(1,23)<1$, n.s.

The main effect of Prime Type was not significant in either analysis ($F_1(1,119)=1.19$, n.s.; $F_2(1,23)=1.15$, n.s.). The expected two-way interaction between Prime Type and Probe Type was not significant ($F_1(1,118)=1.49$ and $F_2(1,23)<1$). There is no difference in facilitation by full primes for Actual Words or Competitors. Summing up, despite our efforts to set up the present experiment as neatly as possible,

our data do not support the idea that semantic priming with partial primes is robust. With this experimental technique, the course of activation of multiple word candidates could not be traced, contrary to previous results in other studies. At early positions, no significant facilitation was found for word candidates. Only the presentation of the full prime resulted in a weak but significant priming effect; for appropriate candidates and their competitors alike.

One obvious drawback of our neat within-subject design is that there were actually too few items with respect to the number of conditions. The choice of using most of the experimental conditions and the sentence material of Zwitserlood (1989) forced us into a design with few repetitions per subject: rotating 12 conditions over 24 items yields only 2 repetitions per subject. Given the enormous within- and between-subject variation in lexical decision time, two observations per condition may indeed be far too few to find robust results. The fact that 120 subjects participated in this experiment could not make up for this. The present results are not conclusive about the suitability of the cross-modal semantic priming paradigm because we have even failed to show robust priming effects after the presentation of full primes. Since priming with full primes is a rather robust phenomenon (Tabossi 1996), the present results cannot be taken as firm evidence against the suitability of the paradigm. The experiment was therefore rerun with a much less complicated design, set up to test just the one question of whether multiple activation of word candidates can be shown with this experimental technique. Since it was our goal to test whether the technique is sensitive enough to pick up activations of word candidates, maintaining the two sentence contexts was not necessary. This experiment will be described in the next section.

4 Replication experiment II

The classic study by Zwitserlood (1989) showed multiple activation of word candidates. Since these results were very promising, we set out to replicate these results as a way of getting to know the ins and outs of the experimental paradigm before starting our own research. In the previous section our failure was reported in replicating those results. In the Introduction the DC model of word recognition was mentioned which provides a plausible explanation for the fact that semantic priming effects with partial primes can at best be very weak. However, in the previous experiment, we not only failed to find partial priming results, but the results after the full presentation of the prime were not very convincing either. This is not in line with previous research, nor with the predictions of the DC model. The failure to find any robust effects at all was argued to result from the complicated design: rotating 12 conditions over 24 items yields only 2 repetitions per subject. We seek additional evidence to settle the discrepancy among the cross-modal partial priming studies cited in section 2. This may in fact be achieved by means of a relatively simple experiment focussing on just one question: does a partial auditory prime word yield early activation of multiple lexical candidates which are semantically related to that prime? Subsequent questions, e.g., about the modulating effect of sentence context, are ignored here. If multiple activation

cannot be demonstrated with this technique, the locus of the effect of sentence context cannot be studied either.

4.1 Method

Materials and design

The test material is a selection of the material described in the previous experiment. The main difference is in the experimental design. The number of cut-off points could be reduced to two (one late cut-off position as the partial prime, and prime off-set). Furthermore, the Carrier Phrase/Biasing Phrase distinction was left out because of the present focus on multiple activation.

Due to her large number of conditions - even exceeding the number of items - (Zwitserlood 1989) was forced to use an incomplete between-subjects design. In the former and in the present experiment, with fewer conditions, it was possible to use a within-subjects design. Three factors were varied in the present experiment. First, priming effects were obtained by comparing a visual probe that was semantically related to an auditory prime (in test items; e.g., 'salary' -'money') against the same visual probe that was not related to an auditorily control (e.g., quasi-prime 'piano' -'money'). This factor is called Prime Type. Secondly, both the auditory prime word salaris 'salary' and its competitor salami 'salami' were investigated via the visual probes GELD ('money') and WORST ('sausage'): this factor is called Probe Type. The third factor Prime Amount varies the amount of auditory information provided by the prime: either full (the visual probe is presented at the offset of the prime word, e.g., *salaris)* or partial (before the offset of the prime word, e.g., *sala-).*

A full factorial design of these three factors would have yielded 2×2×2=8 experimental versions. Counterbalancing these versions over listeners would require 8 groups of listeners. As in the previous experiment, this design can be reduced because two of these conditions can be presented to the same listener. In Table 6, a schematic design of the experiment is given.

Table 6. Schematic design of the experimental set-up, including the factors Prime Type, Probe Type, and Prime Amount (partial vs. full prime).

Prime Type	Probe type	Prime amount	Utterance: *Het volgende woord is..*	Visual probe	Listener group
Test	Actual	partial	… sala-	geld	1
Control	Comp.	partial	… pia-	worst	1
Test	Actual	full	… salaris	geld	2
Control	Comp.	full	… piano	worst	2
Test	Comp.	partial	… sala-	worst	3
Control	Actual	partial	… pia-	geld	3
Test	Comp.	full	… salaris	worst	4
Control	Actual	full	… piano	geld	4

The 4 test conditons (2 probe types, with partial vs. full primes) and the 4 corresponding control conditions were rotated over 4 stimulus lists. The 4 lists were presented to 4 groups of listeners. The main part of the experiment consisted of 48 test and control items, randomly mixed with 58 filler items. Before this main part, there was a warming-up part of 18 filler items, after which the experiment proceeded seamlessly to the main part. The total material set of 48+58+18 items was balanced for visual words and nonwords.

Procedure

Each run started with a separate practice session involving 12 items, after which additional instruction was possible.

Each presentation consisted of an auditory prime (or control) at the end of a carrier sentence. The visual probe was presented at the acoustic offset of the prime word or prime fragment and remained visible for 50 ms. Listeners were instructed to listen carefully and to give a lexical decision response to the visual probe as fast and at the same time as accurately as possible. The 'yes' button was always under the listener's dominant hand. Reaction times were measured from the onset of the presentation of the visual probe until one of the response buttons was pressed. The inter-stimulus interval was 4 seconds.

As in the previous experiment, an extra task was added to ensure that listeners paid attention to the auditory input. After about 1 in 7 items, after the subject's lexical decision response, a message appeared on the screen: *Repeat the last word you heard*. Because listeners might be distracted by the repeat command, at least one filler sentence (without repeat command) followed before the next test or control item was presented. Apart from this restriction, the order of items was randomised.

The accuracy and speed with which listeners performed the task was transformed into a score. During the experiment, participants were not informed about their score: the total score only appeared on the screen when the experiment was finished. Listeners were informed in advance that the person with the highest score would receive a bonus reward of NLG 25.

Subjects

To each of the 4 experimental lists, 15 subjects were assigned. The 60 subjects were students at Utrecht University and were paid NLG 10 for their participation. The subjects had not participated in any of the previous experiments or pretests.

4.2 Results

The mean raw lexical decision times are shown Figure 2.

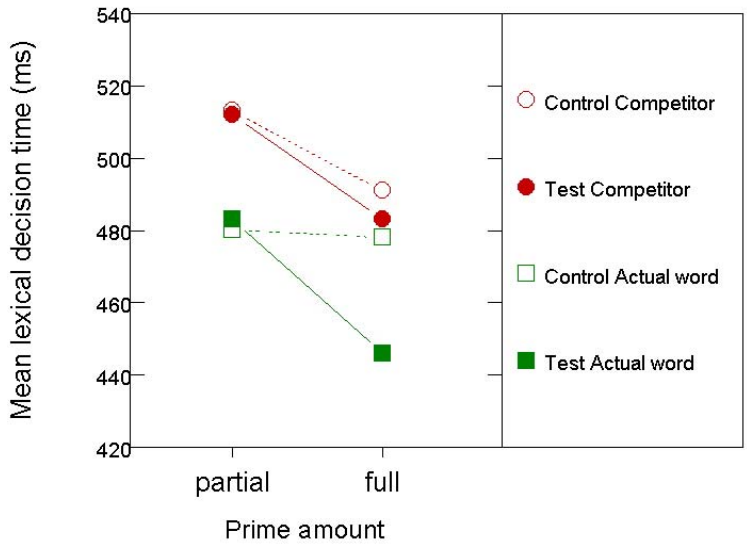

Figure 2. Mean lexical decision times (in ms) broken down by Prime Type (Test vs. Control; filled and open symbols), by Probe Type (Actual vs. Competitor; squares and circles) and by Prime Amount (horizontal axis).

All incorrect lexical decisions, non-responses, and responses with RTs exceeding 3000 ms were counted as misses. In total, 80 out of 2880 observations were missing (3%). The presentation of the full prime results in a 32 ms priming effect for the actual word probe, relative to the control condition. The 80 missing observations were replaced by the grand mean. These results were then entered into analyses of variance, with both items and subjects as repeated measures. The results of these two analyses are presented in Table 7.

Table 7. Results of the statistical analyses (missing observations replaced by grand mean)

Effect	Analysis on subjects	Analysis on items
Prime Type (test/control)	$F_1(1,59)=5.46$, $p=0.023$	$F_2(1,23)=5.29$, $p=0.031$
Probe Type (Actual/Comp.)	$F_1(1,59)=53.2$, $p<0.001$	$F_2(1,23)=9.13$, $p=0.006$
Prime Amount	$F_1(1,59)=23.0$, $p<0.001$	$F_2(1,23)=12.7$, $p=0.002$
Prime Type * Probe Type	$F_1(1,59)=1.78$, n.s.	$F_2(1,23)<1$, n.s.
Prime Type * Prime Amount	$F_1(1,59)=4.36$, $p=0.041$	$F_2(1,23)=8.92$, $p=0.007$
Probe Type * Prime Amount	$F_1(1,59)<1$, n.s.	$F_2(1,23)<1$, n.s.
PrimeType * ProbeType * Prime Amount	$F_1(1,59)=2.75$, $p=0.102$	$F_2(1,23)=2.81$, $p=0.107$

All three main effects were significant: average RTs were slightly faster in related test conditions (481 ms) than in unrelated control conditions (490 ms); main effect of Prime Type yields $F_1(1,59)=5.46$, $p=0.023$; $F_2(1,23)=5.29$, $p=0.031$. Average RTs were faster for actual words (472 ms) than for competitors (500 ms); $F_1(1,59)=53.2$, $p<0.001$; $F_2(1,23)=9.13$, $p=0.006$. As expected, conditions with full primes yielded significantly faster RTs (474 ms) than conditions with partial primes (497 ms); main effect of Prime Amount yields $F_1(1,59)=23.0$, $p<0.001$; $F_2(1,23)=12.7$, $p=0.002$. The two-way interaction of Prime Type by Prime Amount was also significant, $F_1(1,59)=4.36$, $p=0.041$; $F_2(1,23)=8.92$, $p=0.007$. The increment in priming effect for longer primes is larger for test conditions than for control conditions. In other words, the difference in RT between test and control conditions is larger in the fully primed condition (464–484= –20 ms) than in the partially primed condition (498–496= +2 ms), as is to be expected.

A three-way interaction was predicted between the factors Prime Type, Probe Type and Prime Amount because the activation of the actual word was expected to increase with increasing prime amount, and the activation of the competitor to decrease. The fact that only the two-way interaction between Prime Type and Prime Amount is significant suggests that the activation of the competitor also increases when more of the actual word becomes available. The analyses of variance were therefore carried out separately for actual word and competitor probes. The interaction between Prime Type and Prime Amount was significant in the subanalysis on the actual word data ($F_1(1,59)=9.26$, $p=0.003$; $F_2(1,23)=9.43$, $p=0.005$), but it was not significant in the analysis on the competitor data ($F_1(1,59)<1$, n.s.; $F_2(1,23)<1$, n.s.). The results of the analyses on the subsets suggests that only the priming effect caused by the actual word is responsible for the significant interaction between Prime Type and Prime Amount.

Lastly, the results were also analysed after inverse transformation. Statistical analyses assume that the data are normally distributed. However, reaction time data usually do not show a normal (Gaussian) distribution. When the data are transformed to inverse reaction times (1/RT) the distribution is more normal than before the transformation. In Table 8 the results of the analyses are shown for the transformed data.

Table 8. Results of the statistical analyses on inverse reaction time data

Effect	Analysis on subjects	Analysis on items
Prime Type (test/control)	$F_1(1,59)=13.95$, $p<0.001$	$F_2(1,23)=11.5$, $p=0.002$
Probe Type (Actual/Comp.)	$F_1(1,59)=61.2$, $p<0.001$	$F_2(1,23)=11.9$, $p=0.002$
Prime Amount	$F_1(1,59)=34.1$, $p<0.001$	$F_2(1,23)=14.8$, $p=0.001$
Prime Type * Probe Type	$F_1(1,59)=2.43$, n.s.	$F_2(1,23)<1$, n.s.
Prime Type * Prime Amount	$F_1(1,59)=5.33$, $p=0.024$	$F_2(1,23)=11.7$, $p=0.002$
Probe Type * Prime Amount	$F_1(1,59)<1$, n.s.	$F_2(1,23)<1$, n.s.
PrimeType * ProbeType * Prime Amount	$F_1(1,59)=3.41$, $p=0.070$	$F_2(1,23)=2.63$, $p=0.119$

Again, in these analyses, the three-way interaction between Prime Type, Probe Type and Prime Amount does not reach significance. The overall picture therefore remains unchanged: the separate analyses on actual word and competitor data remain the strongest evidence that significant priming is found for the actual word after full presentation of the prime, but there is no significant priming effect, neither for the actual word, nor for the competitor, after partial presentation of the prime.

These results are in line with the results of Chwilla (1996) who also only found a semantic priming effect after the presentation of the full prime, and not after the presentation of a prime fragment. Note that this cannot not be due to the prime fragment being too short: the "early" cut-off point used in the present study was the mean isolation point, as determined in an earlier gating study. The isolation point is defined as the average gate at which listeners first come up with the intended word, without changing their response at later gates (Grosjean 1980). This is Zwitserlood's cut-off point 3, at which a significant priming effect is found in her study (34 ms priming effect in carrier phrase condition). The priming effect after presentation of the full prime in the present results (32 ms) is comparable to Chwilla's results (27 ms after full prime), and to Zwitserlood's (1989; position 4 in carrier phrase condition: 40 ms).

These results show that the activation of multiple word candidates cannot be shown by the CMSP paradigm. The activation of a word candidate has to be sufficiently high to become measurable via semantic priming, and this probably means that subjects either need more processing time or more of the signal (Zwitserlood & Schriefers 1995) to show robust priming effects.

5 Discussion

The results obtained in this experiment clearly show strong and reliable priming effects, if the whole auditory prime is presented. This is in agreement with previous research using the same task (cf. Tabossi (1996) for references) which in turn lends credibility to the present experiment. The significant 32 ms priming effect after presentation of the full prime agrees with the effects in this condition as reported by Zwitserlood. This indicates that our amendments in the stimulus materials did not reduce the basic priming effects.

Second, our results clearly show that there is no semantic priming at all when partial primes are presented: RTs to partially primed probes are equal to those for (identical) probes in unprimed control conditions. If full primes are presented, then significant priming is observed for the actual test words, as mentioned above. But one would also have expected that the competitors of these actual test words would have been inhibited, yielding slower reaction times, as found by (Zwitserlood 1989). Our third main finding is that competitors were not inhibited, contrary to these expectations.

What then are the possible reasons for these differences in results between the present experiment and similar experiments? Given that we find reliable effects for full primes, it is highly unlikely that the changes in the experimental design and materials explain the absence of an early priming effect. We

suspect that the discrepancy between studies in which partial priming was found and our study can be due to differences in the experimental design or to differences in mean RT. As mentioned before, Zwitserlood (1989) used an incomplete between-subjects design, whereas we used a complete within-subjects design. In Zwitserlood's experiment, any interaction between listeners and main effects cannot be separated from those main effects, and such an interaction might have inflated the variance attributed to main effects (Janse & Quené unpublished manuscript).

In section 2 a table was presented in which a number of studies were listed which either did or did not obtain partial priming effects (cf. Table 1). This table also shows that the mean RTs vary enormously for the different studies. Some of the studies that did obtain early priming effects had much longer RTs than we found in our study. The point made in the Zwitserlood & Schriefers study (1995) was that two factors are involved in whether or not priming is obtained: more processing time, or more of the stimulus. It is not too far-fetched to assume that slow subjects will show more priming than faster subjects. In our study we could not find evidence that the slower subjects showed greater priming than the fast subjects. However, in a sense, regarding the mean detection times obtained in e.g., Tabossi & Zardon (1993) or Connine et al. (1994), all our subjects were fast. The short presentation of the visual probe (50 ms), combined with our emphasis on fast responses, may have caused our subjects to respond thus fast.

As said in the Introduction, theoretical back-up for these weak or absent partial priming effects may be found in the Distributed Cohort Model (Gaskell & Marslen-Wilson 1997, 1999, 2002). In the DCM model, before the uniqueness point of a word, its semantic activation depends strongly on the number of candidates and their relative frequency that match the input so far. In their 1999 article, Gaskell & Marslen-Wilson argue that priming in a distributed model depends on the similarity between the relevant words' representations: priming occurs if its lexical representation is more similar to the target representation than to an unrelated baseline. Phonologically, the word candidates are obviously coherent, but the semantic representations of the different candidates often have no meaning overlap at all. In a cross-modal priming experiment, using both repetition priming and semantic priming, they tested the prediction that the effects of repetition priming are greater than those of semantic priming by presenting the primes either complete, or in two cut-off conditions. This experiment is described in detail in Gaskell & Marslen-Wilson (2002). As expected, Gaskell & Marslen-Wilson (2002) found much greater effects of repetition priming, in all conditions and at all cut-off points, significant effects were found, relative to unrelated control conditions. The size of the effect, however, is related to the number of active competitors. Apparently, the system is unable to properly represent competing unrelated semantic representations.

The results of Gaskell & Marslen-Wilson (2002) provide some further interesting information concerning the effect of delayed priming. Since the authors argue that the system is unable to properly represent competing unrelated semantic representations, no priming effect is found, for none of the word candidates. However, when more processing time is provided by way of delayed priming, their results showed that the still ambiguous prime fragments even then did not yield any priming effect. Note

that these results are in stark contrast with those of Zwitserlood & Schriefers (1995). Providing the subjects with extra processing time after the presentation of a partial prime did increase the semantic priming effect in the Zwitserlood & Schriefers (1995) study, but this is not replicated in Gaskell & Marslen-Wilson (2002). In terms of their DCM model, this is only logical. Given the fact that the ambiguity is not resolved by allowing subjects more processing time, delayed priming cannot be expected to increase the semantic priming effect.

Two aspects of this study are relevant with respect to the present study. First, the authors argue that when activated candidates are co-represented in the same representational space, the simultaneous activation of more than one candidate will necessarily create interference. The second important finding of their study is that the extent of this interference will vary according to the coherence of the representations being co-activated. In other words, repetition priming does occur because of the phonological coherence between the activated representations, but semantic priming is weak because the semantic representations are not coherent at all. The system is unable to properly represent competing unrelated semantic representations. Gaskell & Marslen-Wilson (2002) argue that there is a cost involved in parallel activation of more than one lexical representation, in terms of reduced activation of the multiple lexical entries. This finding disproves models such as the original Cohort model (Marslen-Wilson & Welsh 1978) which did not reduce activation according to the number of active candidates, but it fits in with lateral inhibition models such as TRACE (McClelland & Elman 1986) and Shortlist (Norris 1994). The DC model provides a theoretical explanation for the failure to find consistent semantic priming effects indicating multiple activation of word candidates.

6 Conclusion

The experiments reported in sections 3 and 4 lend no empirical support for early activation of multiple word candidates. Priming results were obtained after the presentation of full primes where the word in question has been recognised. Before the recognition point, no semantic priming effects were obtained, neither for the word itself, nor for competitor word candidates.

There is abundant evidence for activation of multiple word candidates from studies which have used other experimental tasks. Evidence comes from phonological priming (Brown 1990; Slowiaczek, McQueen, Soltano & Lynch 2000; Vroomen & de Gelder 1995), word identification (Luce, Pisoni & Goldinger 1990), word spotting (Cutler & Norris 1988; Norris, McQueen & Cutler 1995), phoneme classification (Borsky, Tuller & Shapiro 1998), phoneme monitoring (Gaskell & Marslen-Wilson 1998; Vroomen & de Gelder 1995), and tracking of eye movement (Dahan, Magnuson, Tanenhaus & Hogan 2001). Competition among multiple candidates plays a role in many models of spoken word recognition, including TRACE (McClelland & Elman 1986), Shortlist (Norris 1994), and the Neighborhood Activation Model (Luce 1986). The body of evidence obtained with other experimental tasks certainly rules out the possibility that multiple activation would not be part of the recognition process.

Given the inconsistent results of previous experiments, the failure to show multiple activation in the present experiments, and the objections against the experimental designs of some of the previous studies that did find partial priming effects, the most logical explanation for the dubious results reported on multiple activation must indicate that the problem lies in the task itself. By way of semantic priming, the activations of word candidates cannot be tapped in an on-line way. The study by Gaskell & Marslen-Wilson (2002) even suggests that the activation of word candidates cannot be tapped at all. Their study provides a theoretical background as to why the semantic activation of candidates is so low, compared to the phonological activation of word candidates. Their results suggest that, even when subjects are given extra time after the presentation of the partial prime and before the visual presentation of the probe, multiple activation cannot be shown.

Summing up, only when the activation level of a word candidate rises above a certain threshold (i.e., after the recognition point), does the activation of the prime reliably affect processing of its semantically related visual probe. Consequently, cross-modal semantic priming cannot be used as an on-line measure of the activations of multiple word candidates. Hence, further questions concerning the way in which lexical access is influenced by semantic or syntactic context cannot be answered by using this paradigm. It is therefore advisable to discontinue its use in research into lexical access.

7 REFERENCES

Abercrombie, D. (1967). *Elements of general phonetics*. Edinburgh: Edinburgh University Press.

Ainsworth, W. A. (1974). The influence of precursive sequences on the perception of synthesized vowels. *Language and Speech, 17*, 103-109

Allen, J., Hunnicutt, S. & Klatt, D. H. (1987). *From text to speech, the MITALK system*. Cambridge: Cambridge University Press.

Altmann, G. & Carter, D. (1989). Lexical stress and lexical discriminability: stressed syllables are more informative, but why? *Computer Speech and Language, 3*, 265-275

Altmann, G. T. M. & Young, D. (1993). *Factors affecting adaptation to time-compressed speech*. Proceedings of the European Conference on Speech Communication and Language Technology Eurospeech, Berlin, 333-336.

Arons, B. (1994). *Interactively skimming recorded speech*. PhD dissertation, Massachussetts Institute of Technology, Cambridge.

Bard, E., Anderson, A. H., Sotillo, C., Aylett, M. P., Doherty-Sneddon, G. & Newlands, A. (2000). Controlling the intelligibility of referring expressions in dialogue. *Journal of Memory and Language, 42*, 1-22

Bard, E., Sotillo, C. & Aylett, M. P. (2000). *Taking the hit: Why lexical and phonological processing should not make lexical access too easy*. Proceedings of the Workshop on Spoken Word Access Processes, Nijmegen, the Netherlands, 3-6.

Bard, E. G., Sotillo, C., Kelly, M. L. & Aylett, M. P. (2001). Taking the hit: leaving some lexical competition to be resolved post-lexically. *Language and Cognitive Processes, 16*(5/6), 731-737

Beasley, D. S. & Maki, J. E. (1976). Time- and frequency-altered speech. In N. J. Lass (Ed.), *Contemporary issues in experimental phonetics*. London: Academic Press.

van Bergem, D. R. (1993). Acoustic vowel reduction as a function of sentence accent, word stress, and word class. *Speech Communication, 12*, 1-23

van Bezooijen, R. & van Heuven, V. J. (1997). Assessment of synthesis systems. In D. Gibbon & R. Moore & R. Winski (Eds.), *Handbook of standards and resources for spoken language systems* (pp. 481-563). Berlin: Mouton de Gruyter.

Bond, Z. S. & Moore, T. J. (1994). A note on the acoustic-phonetic characteristics of inadvertently clear speech. *Speech Communication, 14*, 325-337

Borsky, S., Tuller, B. & Shapiro, L. (1998). "How to milk a coat": The effects of semantic and acoustic information. *Journal of the Acoustical Society of America, 103*(5), 2670-2676

Bradlow, A. R., Torretta, G. M. & Pisoni, D. B. (1996). Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication, 20*, 255-272

Brown, C. M. (1990). *Spoken-word processing in context.* Doctoral dissertation, Catholic University Nijmegen.

Caspers, J. (1994). *Pitch movements under time pressure: Effects of speech rate on the melodic marking of accents and boundaries in Dutch.* Doctoral dissertation, Leiden University.

Caspers, J. & van Heuven, V. J. (1995). *Effects of time pressure on the choice of accent-lending and boundary-marking pitch configurations in Dutch.* Proceedings of the 4th European Conference on Speech Communication and Technology, Madrid, vol.2, 1001-1004.

Celex. (1990). *Celex Dutch database (release N.3.1.).* Celex Dutch Centre for Information, Available: http://www.kun.nl/celex/

Charpentier, F. & Stella, M. G. (1986). *Diphone synthesis using an overlap-add technique for speech waveforms concatenation.* Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, 2015-2018.

Cho, T. (2001). *Effects of prosody on articulation in English.* Doctoral dissertation, University of California, Los Angeles.

Chwilla, D. J. (1996). *Electrophysiology of word processing: the lexical processing nature of the N400 priming effect.* Doctoral dissertation, Nijmegen University.

Collins, A. M. & Loftus, E. F. (1975). A spreading activation theory of semantic processing. *Psychological review, 83*, 407-428

Colotte, V., Laprie, Y. & Bonneau, A. (2001). *Perceptual experiments on enhanced and slowed down speech sentences for second language acquisition.* Proceedings of the European Conference on Speech Communication and Technology Eurospeech, Aalborg (Denmark), 1, 469-472.

Connine, C. M., Blasko, D. G. & Wang, J. (1994). Vertical similarity in spoken word recognition: multiple lexical activation, individual differences, and the role of sentence context. *Perception & Psychophysics, 56*, 624-636

Covell, M., Withgott, M. & Slaney, M. (1998). *Mach1: Nonuniform time-scale modification of speech.* Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Seattle,

Cutler, A. (1976). Phoneme-monitoring reaction time as a function of preceding intonation contour. *Perception & Psychophysics, 20*, 55-60

Cutler, A. & Butterfield, S. (1990). Word boundary cues in clear speech. *Speech Communication, 9*, 485-495

Cutler, A. & Clifton, C. E. (1984). The use of prosodic information in word recognition. In H. Bouma & D. G. Bouwhuis (Eds.), *Attention and Performance X: control of language processes.* (pp. 183-196). Hillsdale, NJ: Erlbaum.

Cutler, A. & van Donselaar, W. (2001). Voornaam is not (really) a homophone: Lexical prosody and lexical access in Dutch. *Language and Speech, 44*(2), 171-195

Cutler, A. & Fodor, J. A. (1979). Semantic focus and sentence comprehension. *Cognition, 7*, 49-59

Cutler, A. & Koster, M. (2000). *Stress and lexical activation in Dutch.* Proceedings of the 6th International Conference on Spoken Language Processing, Beijing, vol.1, 593-596.

Cutler, A. & Mehler, J. (1993). The periodicity bias. *Journal of Phonetics, 21*, 103-108

Cutler, A., Mehler, J., Norris, D. & Segui, J. (1986). The syllable's different role in the segmentation of French and English. *Journal of Memory and Language, 25*, 385-400

Cutler, A., Mehler, J., Norris, D. & Segui, J. (1987). Phoneme identification and the lexicon. *Cognitive Psychology, 19*, 141-177

Cutler, A. & Norris, D. (1979). Monitoring sentence comprehension. In W. E. Cooper & E. C. T. Walker (Eds.), *Sentence processing: psycholinguic studies presented to Merrill Garrett* (pp. 113-134). Hillsdale, NJ: Erlbaum.

Cutler, A. & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance, 14*(1), 113-121

Dahan, D., Magnuson, J. S., Tanenhaus, M. K. & Hogan, E. M. (2001). Subcategorical mismatches and the time-course of lexical access: Evidence for lexical competition. *Language and Cognitive Processes, 16*, 507-534

Dauer, R. M. (1983). Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics, 11*, 51-62

Dell, G. S. & Newman, J. E. (1980). Detecting phonemes in fluent speech. *Journal of Verbal Learning and Verbal Behaviour, 19*, 608-623

Delogu, C., Paoloni, A. & Sementina, C. (1992). *Comprehension of natural and synthetic speech: Preliminary studies. ESPRIT Project 2589 (SAM)* Multilingual speech input/output assessment, methodology and standardisation; SAM internal Report II.c. London: University College London.

van Donselaar, W. & Lentz, J. (1994). The function of sentence accents and given/new information in speech processing: different strategies for normal-hearing and hearing-impaired listeners? *Language and Speech, 37*(4), 375-391

Dreher, J. J. & O'Neill, J. J. (1957). Effects of ambient noise on speaker intelligibility for words and phrases. *Journal of the Acoustical Society of America, 29*, 1320-1323

Drullman, R. & Collier, R. (1991). On the combined use of accented and unaccented diphones in speech synthesis. *Journal of the Acoustical Society of America, 90*(4), 1766-1775

Dupoux, E. & Green, K. (1997). Perceptual adjustment to highly compressed speech: effects of talker and rate changes. *Journal of Experimental Psychology, 23*, 914-927

Dupoux, E. & Mehler, J. (1990). Monitoring the lexicon with normal and compressed speech: frequency effects and the prelexical code. *Journal of Memory and Language, 29*, 316-335

Dutoit, T. (1997). *An introduction to text-to-speech synthesis*. Dordrecht: Kluwer Academic Publishers.

Dutoit, T. & Leich, H. (1993). MBR-PSOLA: Text-to-Speech synthesis based on an MBE Resynthesis of the segments database. *Speech Communication, 13*, 435-440

Dutoit, T., Pagel, V., Pierret, N., Bataille, F. & van der Vreeken, O. (1996). *The MBROLA Project: Towards a Set of High-Quality Speech Synthesizers Free of Use for Non-Commercial Purposes*. Proceedings of the International Conference on Spoken Language Processing, Philadelphia, 3, 1393-1396.

Eefting, W. (1991). The effect of "information value" and "accentuation" on the duration of Dutch words, syllables, and segments. *Journal of the Acoustical Society of America, 89*(1), 412-424

Eggen, J. H. (1992). *On the quality of synthetic speech.* Doctoral dissertation, Technical University Eindhoven, Eindhoven.

Fairbanks, G., Everitt, W. L. & Jaeger, R. P. (1954). Method for time or frequency compression-expansion of speech. *Transactions of the Institute of Radio Engineers Professional Group on Audio, AU2*, 7-12

Fairbanks, G., Guttman, N. & Miron, M. S. (1957). Auditory comprehension in relation to listening rate and selective verbal redundancy. *Journal of Speech and Hearing Disorders, 22*, 23-32

Fairbanks, G. & Kodman, F. (1957). Word intelligibility as a function of time compression. *Journal of the Acoustical Society of America, 29*, 636-641

Forster, K. I. (1976). Accessing the mental lexicon. In R. J. Wales & E. C. T. Walker (Eds.), *New approaches to language mechanisms* (pp. 257-287). Amsterdam: North-Holland.

Foss, D. J. (1969). Decision processes during sentence comprehension: Effects of lexical item difficulty and position upon decision times. *Journal of Verbal Learning and Verbal Behaviour, 8*, 457-462

Foulke, E. (1966). A survey of the acceptability of rapid speech. *New Outlook Blind, 60*, 261-265

Foulke, E. (1971). The perception of time-compressed speech. In D. L. Horton & J. J. Jenkins (Eds.), *The perception of language.* Columbus, Ohio: Charles E. Merrill publishing company.

Fowler, C. A. (1980). Coarticulation and theories of extrinsic timing. *Journal of Phonetics, 8*, 113-133

Fowler, C. A. (1981). Production and perception of coarticulation among stressed and unstressed vowels. *Journal of Speech and Hearing Research, 46*, 127-139

Garvey, W. D. (1953). The intelligibility of speeded speech. *Journal of Exceptional Psychology, 45*, 102-108

Gaskell, M. G. & Marslen-Wilson, W. D. (1996). Phonological variation and inference in lexical access. *Journal of Experimental Psychology: Human Perception and Performance, 22*, 144-158

Gaskell, M. G. & Marslen-Wilson, W. D. (1997). Integrating form and meaning: a distributed model of speech perception. *Language and Cognitive Processes, 12*(5/6), 613-656

Gaskell, M. G. & Marslen-Wilson, W. D. (1998). Mechanisms of phonological inference in speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 24*, 380-396

Gaskell, M. G. & Marslen-Wilson, W. D. (1999). Ambiguity, competition and blending in spoken word recognition. *Cognitive Science, 23*(4), 439-462

Gaskell, M. G. & Marslen-Wilson, W. D. (2002). Representation and competition in the perception of spoken words. *Cognitive Psychology, 45*, 220-266

Gay, T. (1978). Effect of speaking rate on vowel formant movements. *Journal of the Acoustical Society of America, 63*(1), 223-230

Gay, T. (1981). Mechanisms in the control of speech rate. *Phonetica, 38*, 148-158

Gay, T. & Hirose, H. (1973). Effect of speaking rate on labial consonant production: A combined electromyographic/high speed motion picture study. *Phonetica, 27*, 44-56

Girden, E. R. (1992). *ANOVA Repeated Measures*. Newbury Park (CA): Sage.

Goldman-Eisler, F. (1968). *Psycholinguistics: Experiments in spontaneous speech*. London: Academic Press.

Gordon-Salant, S. (1986). Recognition of natuarl and time/intensity altered CVs by young and elderly subjects with normal hearing. *Journal of the Acoustical Society of America, 80*, 1599-1607

Gordon-Salant, S. (1987). Effects of acoustic modification on consonant recognition by elderly hearing-impaired subjects. *Journal of the Acoustical Society of America, 81*, 1199-1202

Gottfried, T. L., Miller, J. L. & Payton, P. E. (1990). Effect of speaking rate on the perception of vowels. *Phonetica, 47*, 155-172

Greisbach, R. (1992). Reading aloud at maximal speed. *Speech Communication, 11*, 469-473

de Groot, A. M. B. (1980). *Mondelinge woordassociatienormen*. Lisse: Swets & Zeitlinger.

Grosjean, F. (1980). Spoken word recognition and the gating paradigm. *Perception & Psychophysics, 28*, 267-283

Grosjean, F. (1985). The recognition of words after their acoustic offset: evidence and implications. *Perception & Psychophysics, 38*(4), 299-310

de Haan, H. J. (1977). A speech-rate intelligibility threshold for speeded and time-compressed connected speech. *Perception & Psychophysics, 22*(4), 366-372

de Haan, H. J. (1982). The relationship of estimated comprehensibility to the rate of connected speech. *Perception & Psychophysics, 32*(1), 27-31

Hawkins, S. & Warren, P. (1994). Phonetic influences on the intelligibility of conversational speech. *Journal of Phonetics, 22*, 493-511

Hazan, V. & Simpson, A. (1998). The effect of cue-enhancement on the intelligibility of nonsense word and sentence materials presented in noise. *Speech Communication, 24*, 211-226

Hazan, V. & Simpson, A. (2000). The effect of cue-enhancement on consonant intelligibility in noise: speaker and listener effects. *Language and Speech, 43*(3), 273-294

He, L. & Gupta, A. (2001). *Exploring benefits of non-linear time-compression.* Proceedings of the Conference on Multimedia, Ottawa, 382-391.

Henderson, A. I. (1980). Juncture pause and intonation fall and the perceptual segmentation of speech. In H. W. Dechert & M. Raupach (Eds.), *Temporal variables in speech: Studies in honour of Frieda Goldman-Eisler*. The Hague: Mouton Publishers.

Hertrich, I. & Ackermann, H. (1995). Coarticulation in slow speech: durational and spectral analysis. *Language and Speech, 38*(2), 159-187

van Heuven, V. J. (1985). Perception of stress pattern and word recognition: recognition of Dutch words with incorrect stress position. *Journal of the Acoustical Society of America, 78*, s21

van Heuven, V. J. & Hagman, P. (1988). Lexical statistics and spoken word recognition in Dutch. In P. Coopmans & A. Hulk (Eds.), *Linguistics in the Netherlands 1988* (pp. 59-68). Dordrecht: Foris.

Hockett, C. (1955). *A manual of phonology (International Journal of American Linguistics, Memoir 11)*. Baltimore: Waverly Press.

Horton, W. S. & Keysar, B. (1996). When do speakers take into account common ground? *Cognition, 59*, 91-117

Hosom, J. P. (2000). *Automatic time alignment of phonemes using acoustic-phonetic information.* Doctoral dissertation, Oregon Graduate Institute of Science and Technology, Portland.

ITU-P.800. (1996). Methods for subjective determination of transmission quality, *International Telecommunication Union (ITU)*: Recommendation P.800.

Janse, E. (2000). *Intelligibility of time-compressed speech: three ways of time-compression.* Proceedings of the VIth International Congress of Speech and Language Processing, Beijing, vol.III, 786-789.

Janse, E. (2001). *Comparing word-level intelligibility after linear vs. non-linear time-compression.* Proceedings of the VIIth European Conference on Speech Communication and Technology Eurospeech, Aalborg (Denmark), vol.II, 1407-1410.

Janse, E. (2002). *Time-compressing natural and synthetic speech.* Proceedings of the 7th International Conference on Spoken Language Processing, Denver (CO, USA), 1645-1648.

Janse, E. (submitted). Word perception in fast speech: artificially time-compressed vs. naturally produced fast speech. *Speech Communication*

Janse, E., Nooteboom, S. & Quené, H. (in press). Word-level intelligibility of time-compressed speech: Prosodic and segmental factors. *Speech Communication*

Janse, E. & Quené, H. (unpublished manuscript). On measuring multiple lexical activation using the cross-modal semantic priming technique. *2002*

Janse, E., Sennema, A. & Slis, A. (2000). *Fast speech timing in Dutch: the durational correlates of lexical stress and pitch accent.* Proceedings of the VIth International Congress of Speech and Language Processing, Beijing, vol.III, 251-254.

de Jong, K. J. (1995). The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. *Journal of the Acoustical Society of America, 97*(1), 491-504

de Jong, K. J. (2001). Rate-induced resyllabification revisited. *Language and Speech, 44*(2), 197-216

Jongenburger, W. (1996). *The role of lexical stress during spoken word processing.* Doctoral dissertation, Leiden University.

Jongenburger, W. & van Bezooijen, R. (1992). *Evaluatie van ELK:attitudes van de gebruikers, verstaanbaarheid en acceptabiliteit van de spraaksynthese en bruikbaarheid van het zoeksysteem*: Stichting Spraaktechnologie.

Kager, R. (1989). *A metrical theory of stress and destressing in English and Dutch*. Dordrecht: Foris Publications.

Kidd, G. R. (1989). Articulatory-rate context effects in phoneme identification. *Journal of Experimental Psychology: Human Perception and Performance, 15*(4), 736-748

Kiritani, S. (1977). Articulatory studies by the X-ray microbeam system. In M. Sawashima & F. S. Cooper (Eds.), *Dynamic aspects of speech production*. Tokyo: University of Tokyo Press.

Klatt, D. H. (1980). Software for a cascade/parallel formant synthesiser. *Journal of the Acoustical Society of America, 67*, 971-995

Klatt, D. H. (1982). *Prediction of perceived phonetic distance from critical-band spectra: a first step*. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Paris, 2, 1278-1281.

Kohler, K. (1990). Segmental reduction in connected speech in German: Phonological facts and phonetic explanations. In W. J. Hardcastle & A. Marchal (Eds.), *Speech production and speech modeling* (pp. 69-92). Dordrecht: Kluwer Academic Publishers.

Kolers, P. A. (1975). Memorial consequences of automatized encoding. *Journal of Experimental Psychology: Human Learning and Memory, 1*, 689-701

Konkle, D. F., Beasley, D. S. & Bess, F. H. (1977). Intelligibility of time-altered speech in relation to chronological aging. *Journal of Speech and Hearing Research, 20*, 108-115

Kortekaas, R. (1997). *Physiological and psychoacoustical correlates of perceiving natural and modified speech*. Doctoral dissertation, Technical University Eindhoven.

Koster, C. (1987). *Word recognition in foreign and native language: effects of context and assimilation*. Dordrecht: Foris Publications.

Kozhevnikov, V. A. & Chistovic, L. A. (1965). *Speech articulation and perception*. Washington: Joint Publications Research Service.

Kuehn, D. & Moll, K. (1976). A cineradiographic study of VC and CV articulatory velocities. *Journal of Phonetics, 4*, 303-320

Lee, S., Kim, H. D. & Kim, H. S. (1997). *Variable time-scale modification of speech using transient information*. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 1319-1322.

Lehiste, I. (1970). *Suprasegmentals*. Cambridge, MA: MIT Press.

Levelt, W. J. M. (1978). A survey of studies in sentence perception. In W. J. M. Levelt & G. B. Flores d'Arcais (Eds.), *Studies in the perception of language* (pp. 1-74). New York: Wiley.

van Leyden, K. & van Heuven, V. J. (1996). Lexical stress and spoken word recognition: Dutch vs. English. In C. Cremers & M. d. Dikken (Eds.), *Linguistics in the Netherlands*. Amsterdam: John Benjamins.

Liberman, A. M., Cooper, F. S., Shankweiler, D. & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review, 74*, 431-461

Liberman, A. M. & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21*, 1-36

Lindblom, B. (1963). Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America, 35*, 1773-1781

Lindblom, B. (1983). Economy of speech gestures. In P. F. MacNeilage (Ed.), *The production of speech* (pp. 217-246). New York: Springer.

Lindblom, B. (1990). Explaining phonetic variation: a sketch of the H&H theory. In W. J. Hardcastle & A. Marchal (Eds.), *Speech production and speech modelling*. Dordrecht: Kluwer Academic Publishers.

Low, E. L., Grabe, E. & Nolan, F. (2000). Quantitative characterizations of speech rhythm: syllable-timing in Singapore English. *Language and Speech, 43*(4), 377-401

Luce, P. A. (1986). *Neighborhoods of words in the mental lexicon*. Doctoral dissertation, Indiana University, Bloomington.

Luce, P. A. & Pisoni, D. B. (1998). Recognizing spoken words: the Neighborhood Activation model. *Ear and Hearing, 19*, 1-36

Luce, P. A., Pisoni, D. B. & Goldinger, S. D. (1990). Similarity neighborhoods of spoken words. In G. T. M. Altmann (Ed.), *Cognitive models of speech perception*. Cambridge, MA: MIT Press.

Maassen, B. A. M. (1985). *Artificial corrections to deaf speech: Studies in intelligibility*. Doctoral dissertation, Catholic University of Nijmegen.

Maassen, B. A. M. & Povel, D.-J. (1984). The effect of correcting fundamental frequency on the intelligibility of deaf speech and its interaction with temporal aspects. *Journal of the Acoustical Society of America, 76*(6), 1673-1681

MacKay, D. G. (1974). Aspects of the syntax of behavior: Syllable structure and speech rate. *Quarterly Journal of Experimental Psychology, 26*, 642-657

MacNeilage, P. F. & Ladefoged, P. (1976). The production of speech and language. In E. C. Carterette & M. P. Friedman (Eds.), *Handbook of perception* (Vol. VII: Language and speech, pp. 76-120). New York: Academic Press.

van der Made-van Bekkum, I. J. (1973). *Nederlandse woordassociatienormen*. Amsterdam: Swets & Zeitlinger.

Markel, J. D. & Gray, A. H. (1976). *Linear prediction of speech*. Berlin: Springer.

Marslen-Wilson, W. D. (1984). Function and process in spoken word recognition. In H. Bouma & D. G. Bouwhuis (Eds.), *Attention and performance: control of language processes* (pp. 125-150). Hillsdale, NJ: Erlbaum.

Marslen-Wilson, W. D., Nix, A. J. & Gaskell, M. G. (1995). Phonological variation in lexical access: Abstractness, inference and English place assimilation. *Language and Cognitive Processes, 10*, 285-308

Marslen-Wilson, W. D. & Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition, 8*, 1-71

Marslen-Wilson, W. D. & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology, 10*, 29-63

Marslen-Wilson, W. D. & Zwitserlood, P. (1989). Accessing spoken words: the importance of word onsets. *Journal of Experimental Psychology: Human Perception and Performance, 15*, 576-585

Martin, J. G. & Bunnell, H. T. (1981). Perception of anticipatory coarticulation effects. *Journal of the Acoustical Society of America, 69*(2), 559-567

Massaro, D. W. (1972). Perceptual images, processing time, and perceptual units in auditory perception. *Psychological Review, 79*, 124-145

Matsushima, J., Sakai, N., Imamura, T., Ifukube, T., Hokimoto, T. & Nejime, Y. (1995). Assessment of a Portable Digital Speech-Rate Converter for Hearing-Impaired Listeners. *Annals-of-Otology,-Rhinology-and-Laryngology, 104*(9), 156-159

Max, L. & Caruso, A. J. (1997). Acoustic measures of temporal intervals across speaking rates: variability of syllable- and phrase-level relative timing. *Journal of Speech, Language and Hearing Research, 40*, 1097-1110

McClean, M. D. (2000). Patterns of orofacial movement velocity across variations in speech rate. *Journal of Speech, Language and Hearing Research, 43*, 205-216

McClelland, J. L. & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18*, 1-86

McQueen, J. M. & Cutler, A. (1997). Cognitive processes in speech perception. In W. J. Hardcastle & J. Laver (Eds.), *The handbook of phonetic sciences* (pp. 566-585). Oxford: Blackwell.

Mehler, J., Dommergues, J.-Y., Frauenfelder, U. & Segui, J. (1981). The syllable's role in speech segmentation. *Journal of Verbal Learning and Verbal Behaviour, 20*, 298-305

Mehler, J., Segui, J. & Carey, P. W. (1978). Tails of words: monitoring ambiguity. *Journal of Verbal Learning and Verbal Behaviour, 17*, 29-35

Mehta, G. & Cutler, A. (1988). Detection of target phonemes in spontaneous and read speech. *Language and Speech, 31*(2), 135-156

Meyer, D. E. & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: evidence of a dependence between retrieval operations. *Journal of Experimental Psychology, 90*, 227-234

Miller, G. A. (1951). *Language and communication*. New York: McGraw-Hill Book Co. Inc.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological review, 63*, 81-97

Miller, G. A. & Licklider, J. C. R. (1950). The intelligibility of interrupted speech. *Journal of the Acoustical Society of America, 22*, 167-173

Miller, G. A. & Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America, 27*(2), 338-352

Miller, J. L. (1981). Effects of speaking rate on segmental distinctions. In P. D. Eimas & J. L. Miller (Eds.), *Perspectives on the study of speech* (pp. 39-74). Hillsdale: Erlbaum.

Miller, J. L., Green, K. P. & Reeves, A. (1986). Speaking rate and segments: a look at the relation between speech production and speech perception for the voicing contrast. *Phonetica, 43*, 106-115

Miller, J. L., O'Rourke, T. B. & Volaitis, L. E. (1997). Internal structure of phonetic categories: effects of speaking rate. *Phonetica, 54*, 121-137

Moon, S.-J. & Lindblom, B. (1994). Interaction between duration, context, and speaking style in English stressed vowels. *Journal of the Acoustical Society of America, 96*(1), 40-55

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review, 76*, 165-178

Morton, J. & Long, J. (1976). Effect of word transitional probability on phoneme identification. *Journal of Verbal Learning and Verbal Behaviour, 15*, 43-51

Moss, H. E. & Marslen-Wilson, W. D. (1993). Access to word meanings during spoken language comprehension: effects of sentential semantic context. *Journal of Experimental Psychology: Learning, Memory and Cognition, 19*, 1254-1276

Moss, H. E., McCormick, S. F. & Tyler, L. K. (1997). The time-course of activation of semantic information during spoken word recognition. *Language and Cognitive Processes, 12*, 695-731

Moulines, E. & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication, 9*, 453-467

Nakatani, L. H. & Dukes, K. D. (1973). A sensitive test of speech communication quality. *Journal of the Acoustical Society of America, 53*(4), 1083-1092

Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: roles of inhibitionless spreading of activation and limited-capacity attention. *Journal of Experimental Psychology: General, 106*, 226-254

Nix, A. J., Mehta, G., Dye, J. & Cutler, A. (1993). Phoneme detection as a tool for comparing perception of natural and synthetic speech. *Computer Speech and Language, 7*, 211-228

Nooteboom, S. G. (1979). Perceptual adjustment to speech rate: a case of backward perceptual normalisation, *Anniversaries in Phonetics: Studia gratulatoria dedicated to Hendrik Mol* (pp. 255-269). Amsterdam: Institute of Phonetic Sciences.

Nooteboom, S. G. (1981). Speech rate and segmental perception or the role of words in phoneme identification. In T. Myers & J. Laver & J. Anderson (Eds.), *The cognitive representation of speech* (pp. 143-150): North Holland.

Nooteboom, S. G. & Eefting, W. (1994). Evidence for the adaptive nature of speech on the phrase level and below. *Phonetica, 51*, 92-98

Nooteboom, S. G., Scharpff, P. & van Heuven, V. J. (1990). *Effects of several pausing strategies on the recognizability of words in synthetic speech.* Proceedings of the 1st International Conference on Spoken Language Processing, Kobe (Japan), 1, 385-387.

Nooteboom, S. G. & van der Vlugt, M. J. (1988). A search for a word-beginning superiority effect. *Journal of the Acoustical Society of America, 84*(6), 2018-2032

Norris, D. (1994). Shortlist: a connectionist model of continuous speech recognition. *Cognition, 52*, 189-234

Norris, D., McQueen, J. M. & Cutler, A. (1995). Competition and segmentation in spoken-word recognition. *Journal of Experimental Psychology: Learning, Memory and Cognition, 21*(5), 1209-1228

Norris, D., McQueen, J. M. & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences, 23*(3), 299-370

Nusbaum, H. C., Dedina, M. J. & Pisoni, D. B. (1984). *Perceptual confusions of consonants in natural and synthetic CV syllables* Speech research laboratory technical note 84-02. Bloomington: Indiana University, Speech Research Laboratory.

Ohala, J. J. (1996). Speech perception is hearing sounds, not tongues. *Journal of the Acoustical Society of America, 99*(3), 1718-1725

den Os, E. A. (1988). *Rhythm and tempo of Dutch and Italian; a contrastive study.* Unpublished doctoral dissertation, Utrecht University, Utrecht.

O'Shaugnessy, D. (1990). *Spectral transitions in rule-based and diphone synthesis.* Proceedings of the First ESCA Workshop on Speech Synthesis, Autrans, 21-25.

Pallier, C., Sebastian-Galles, N., Dupoux, E., Christophe, A. & Mehler, J. (1998). Perceptual adjustment to time-compressed speech: a cross-linguistic study. *Memory and Cognition, 26*(4), 844-851

Pavlovic, C., Rossi, M. & Espesser, R. (1990). Use of the magnitude estimation technique for assessing the performance of text-to-speech synthesis system. *Journal of the Acoustical Society of America, 87*, 373-381

Perkell, J. S. (1997). Articulatory processes. In W. J. Hardcastle & J. Laver (Eds.), *The Handbook of phonetic sciences* (pp. 333-370). Oxford: Blackwell.

Peterson, G. E. & Lehiste, I. (1960). Duration of syllable nuclei in English. *Journal of the Acoustical Society of America, 32*(6), 693-703

Picheny, M. A., Durlach, N. I. & Braida, L. D. (1985). Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech. *Journal of Speech and Hearing Research, 28*, 96-103

Picheny, M. A., Durlach, N. I. & Braida, L. D. (1986). Speaking clearly for the hard of hearing: Acoustic characteristics of clear and conversational speech. *Journal of Speech and Hearing Research, 29*, 434-446

Picheny, M. A., Durlach, N. I. & Braida, L. D. (1989). Speaking clearly for the hard of hearing III: an attempt to determine the contribution of speaking rate to differences in intelligibility between clear and conversational speech. *Journal of Speech and Hearing Research, 32*, 600-603

Pisoni, D. B. (1981). *Speeded classification of natural and synthetic speech in a lexical decision task.* Proceedings of the Paper presented to the 102nd meeting Acoustical Society of America, Miami,

Pisoni, D. B. (1987). Some measures of intelligibility and comprehension. In J. Allen & S. Hunnicutt & D. H. Klatt (Eds.), *From Text to Speech: the MITALK System.* Cambridge: Cambridge University Press.

Pisoni, D. B. (1997). Perception of synthetic speech. In J. P. H. van Santen & R. W. Sproat & J. P. Olive & J. Hirschberg (Eds.), *Progress in Speech Synthesis.* New York: Springer Verlag.

Pols, L. C. W. (1983). Three-mode principal component analysis of confusion matrices, based on the identification of Dutch consonants, under various conditions of noise and reverberation. *Speech Communication, 2*, 275-293

Port, R. F. (1981). Linguistic timing factors in combination. *Journal of the Acoustical Society of America, 69*(1), 262-274

Quené, H. & Janse, E. (2001). Word perception in time-compressed speech. *Journal of the Acoustical Society of America, 110*(5), 2738, 4aSC11.

Quené, H. & Koster, M. L. (1998). Metrical segmentation in Dutch: Vowel quality or stress? *Language and Speech, 41*(2), 185-202

Quené, H. & Krull, J. (1999). *Recognition of assimilated words in normal and fast speech.* Proceedings of the 14th International Congress of the Phonetic Sciences, San Francisco, 1831-1834.

Ramus, F., Nespor, M. & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition, 73*, 265-292

Reich, S. S. (1980). Significance of pauses for speech perception. *Journal of Psycholinguistic Research, 9*(4), 379-389

Rubin, P., Turvey, M. T. & van Gelder, P. (1976). Initial phonemes are detected faster in spoken words than in non-words. *Perception & Psychophysics, 19*, 394-398

Sanderman, A. & Collier, R. (1997). Prosodic phrasing and comprehension. *Language and Speech, 40*(4), 391-409

van Santen, J. P. H. (1993). Perceptual experiments for diagnostic testing of text-to-speech systems. *Computer Speech and Language, 7*(1), 49-100

Scharpff, P. & van Heuven, V. J. (1988). *Effects of pause insertion on the intelligibility of low quality speech.* Proceedings of the 7th FASE Symposium, Edinburgh, 261-268.

Schwab, E. C., Nusbaum, H. C. & Pisoni, D. B. (1985). Effects of training on the perception of perception of synthetic speech. *Human Factors, 27*, 395-408

Scott, R. J. (1965). *Temporal effects in speech analysis and synthesis.* Unpublished doctoral dissertation, University of Michigan.

Sebastian-Galles, N., Dupoux, E., Costa, A. & Mehler, J. (2000). Adaptation to time-compressed speech: phonological determinants. *Perception & Psychophysics, 64*(2), 834-842

Slowiaczek, L. M. (1990). Effects of lexical stress in auditory word recognition. *Language and Speech, 33*(1), 47-68

Slowiaczek, L. M., McQueen, J. M., Soltano, E. G. & Lynch, M. (2000). Phonological representations in prelexical processing: evidence from form-based priming. *Journal of Memory and Language, 43*, 530-560

Sluijter, A. M. C. (1995). *Phonetic correlates of stress and accent.* Doctoral dissertation, Leiden University.

Sluijter, A. M. C., van Heuven, V. J. & Pacilly, J. J. A. (1997). Spectral balance as a cue in the perception of linguistic stress. *Journal of the Acoustical Society of America, 101*(1), 503-513

Soli, S. D. & Arabie, P. (1979). Auditory versus phonetic accounts of observed confusions between consonant phonemes. *Journal of the Acoustical Society of America, 66*, 46-59

van Son, R. J. J. H. & Pols, L. C. W. (1990). Formant frequencies of Dutch vowels in a text, read at normal and fast rate. *Journal of the Acoustical Society of America, 88*(4), 1683-1693

van Son, R. J. J. H. & Pols, L. C. W. (1992). Formant movements of Dutch vowels in a text, read at normal and fast rate. *Journal of the Acoustical Society of America, 92*(1), 121-127

Sonoda, Y. (1987). Effect of speaking rate on articulatory dynamics and motor event. *Journal of Phonetics, 15*, 145-156

Sotillo, C. & Bard, E. G. (1998). *Is hypo-articulation lexically constrained?* Proceedings of the SPoSS, Aix-en-Provence, 109-112.

Stetson, R. H. (1951). *Motor phonetics (2nd edition)*. Amsterdam: North-Holland.

Stevens, K. N. (1989). On the quantal nature of speech. *Journal of Phonetics, 17*, 3-45

Summerfield, A. Q. (1975). How a full account of segmental perception depends on speech prosody and vice versa. In A. Cohen & S. G. Nooteboom (Eds.), *Structure and Process in Speech Perception* (pp. 51-66). Heidelberg: Springer Verlag.

Summerfield, Q., Bailey, P. J., Seton, J. & Dorman, M. F. (1981). Fricative envelope parameters and silent intervals in distinguishing 'slit' and 'split'. *Phonetica, 38*, 181-192

Summers, W. V., Pisoni, D. B., Bernacki, R. H., Pedlow, R. I. & Stokes, M. A. (1988). Effects of noise on speech productions: Acoustic and perceptual analyses. *Journal of the Acoustical Society of America, 84*, 917-928

Swinney, D. A. (1979). Lexical access during sentence comprehension: (Re)consideration of context effects. *Journal of Verbal Learning and Verbal Behaviour, 18*, 645-659

Tabossi, P. (1996). Cross-modal semantic priming. *Language and Cognitive Processes, 11*(6), 569-576

Tabossi, P. & Zardon, F. (1993). Processing ambiguous words in context. *Journal of Memory and Language, 32*, 359-372

Taft, L. (1984). *Prosodic constraints and lexical parsing strategies*. Ph.D. thesis, University of Massachusetts.

Tallal, P., Miller, S. L., Bedi, G., Byma, G., Wang, X., Nagarajan, S., Schreiner, C., Jenkins, W. & Merzenich, M. (1996). Language comprehension in language-learning impaired children improved with acoustically modified speech. *Science, 271*, 81-84

Terken, J. & Lemeer, G. (1988). Effects of segmental quality and intonation on quality judgments for texts and utterances. *Journal of Phonetics, 16*, 453-457

Trouvain, J. & Grice, M. (1999). *The effect of tempo on prosodic structure*. Proceedings of the International Congress of the Phonetic Sciences, San Francisco, 1067-1070.

Tuller, B. & Kelso, J. A. S. (1991). The production and perception of syllable structure. *Journal of Speech and Hearing Research, 34*, 501-508

Uchanski, R. M., Choi, S. S., Braida, L. D., Reed, C. M. & Durlach, N. I. (1996). Speaking clearly for the hard of hearing IV: further studies of the role of speaking rate. *Journal of Speech and Hearing Research, 39*, 494-509

Versfeld, N. J. & Dreschler, W. A. (2002). The relationship between the intelligibility of time-compressed speech and speech in noise in young and elderly listeners. *Journal of the Acoustical Society of America, 111*(1), 401-408

Vroomen, J. & de Gelder, B. (1995). Metrical segmentation and lexical inhibition in spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 21*, 98-108

Vroomen, J. & de Gelder, B. (1999). Lexical access of resyllabified words: Evidence from phoneme monitoring. *Memory and Cognition, 27*(3), 413-421

Vroomen, J., van Zon, M. & de Gelder, B. (1996). Cues to speech segmentation: Evidence from juncture misperceptions and word spotting. *Memory and Cognition, 24*, 744-755

Waals, J. (1999). *An experimental view of the Dutch syllable.* Doctoral dissertation, Utrecht University, Utrecht.

Wayland, S. C., Miller, J. L. & Volaitis, L. E. (1994). The influence of sentential speaking rate on the internal structure of phonetic categories. *Journal of the Acoustical Society of America, 95*(5), 2694-2701

Weber, A. (2001). Help or hindrance: how violation of different assimilation rules affects spoken-language processing. *Language and Speech, 44*(1), 95-118

Whalen, D. H. (1991). Subcategorical phonetic mismatches and lexical access. *Perception & Psychophysics, 50*, 351-360

van Wieringen, A. (1995). *Perceiving dynamic speech sounds: psychoacoustics and perception.* Doctoral dissertation, University of Amsterdam.

van Wieringen, A. & Pols, L. C. W. (1995). Discrimination of single and complex consonant-vowel and vowel-consonant-like formant transitions. *Journal of the Acoustical Society of America, 98*(3), 1304-1312

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

Wingfield, A., Lombardi, L. & Sokol, S. (1984). Prosodic features and the intelligibility of accelerated speech: syntactic versus periodic segmentation. *Journal of Speech and Hearing Research, 27*, 128-134

Wouters, J. & Macon, M. W. (2002a). Effects of prosodic factors on spectral dynamics. I. Analysis. *Journal of the Acoustical Society of America, 111*(1), 417-427

Wouters, J. & Macon, M. W. (2002b). Effects of prosodic factors on spectral dynamics. II. Synthesis. *Journal of the Acoustical Society of America, 111*(1), 428-438

Young, D., Altmann, G. T. M., Cutler, A. & Norris, D. (1993). *Metrical structure and the perception of time-compressed speech*. Proceedings of the European Conference on Speech Communication and Technology Eurospeech, Berlin, 771-774.

Zemlin, W. R., Daniloff, R. G. & Shriner, T. H. (1968). The difficulty of listening to time-compressed speech. *Journal of Speech and Hearing Research, 11*, 875-881

Zwitserlood, P. (1989). The locus of the effects of sentential-semantic context in spoken-word processing. *Cognition, 32*, 25-64

Zwitserlood, P. & Schriefers, H. (1995). Effects of sensory information and processing time in spoken-word recognition. *Language and Cognitive Processes, 10*, 121-136

Appendix A

Below an overview table is provided of all the experiments dealing with the production or perception of fast speech.

Chapter	Experiment (nr)	Topic	Technique
2	1	phoneme and word intelligibility of normal-rate and time-compressed speech (words and nonwords)	intelligibility test (cloze procedure)
3	1	intelligibility of time-compressed synthetic and natural speech	intelligibility test (cloze procedure)
3	2	processing speed of normal-rate and time-compressed natural vs. synthetic speech	phoneme detection
4	1	speech timing in normal-rate and fast speech production	analysis of natural read speech
4	2	intelligibility of artificially time-compressed speech after three ways of time-compression: words and nonwords in carrier phrases (compressed to 35% (words) or 45% (nonwords) of normal-rate duration)	intelligibility test (cloze procedure)
4	3	intelligibility of artificially time-compressed speech after three ways of time-compression: words embedded in normal sentences (compressed to 35% of normal-rate duration)	intelligibility test (cloze procedure)
4	4	intelligibility of artificially time-compressed speech in noise, before and after three ways of time-compression: words embedded in normal sentences (compressed to 65% of normal-rate duration)	speech-interference technique
4	5	processing speed of artificially time-compressed speech after three ways of time-compression (compression to 65% of original duration)	phoneme detection
5	1	intelligibility fast-rate sentence material vs. artificially time-compressed material of fastest speaker of experiment 1 (Chapter 4) (compression to 60%)	intelligibility test (cloze procedure)
5	2	processing speed of fast-rate vs. artificially time-compressed sentence material of most intelligible speaker of experiment 1 (Chapter 4) (compression rate 66%)	phoneme detection
5	3	word-perception speed in natural-fast vs. artificially time-compressed speech (compression rate 72%)	phoneme detection
5	4	subjective preference test natural-fast vs. artificially (compression rate 72%)	comparative mean opinion score test
5	5	intelligibility of speech after linear time compression vs. nonlinear 'remove-pauses-first' time compression (compression to 35% of original duration)	intelligibility test

Appendix B

Below the nonwords of Chapter 2 are listed, embedded in their carrier phrase. Target words are in bold. Between brackets are the two real words the non-word has been composed from. These real words were presented in the real-word experiment in Chapter 2.

Set 1: sonorant onset; CCC- coda

Je moet **runst** schrijven	(rust/kunst)
Je moet **jurcht** schrijven	(jurk/burcht)
Je moet **lorst** schrijven	(lok/korst)
Je moet **wekst** schrijven	(web/tekst)
Je moet **narts** typen	(nacht/arts)
Je moet **mangst** schrijven	(mat/angst)
Je moet **munst** schrijven	(mug/gunst)
Je moet **rengst** schrijven	(rek/hengst)
Je moet **werts** typen	(wet/erts)
Je moet **nurcht** schrijven	(nul/wurgt)
Je moet **jarst** schrijven	(jas/barst)
Je moet **lekst** schrijven	(lef/gekst)

Set 2: obstruent onset; sonorant-obstruent coda

Je moet **kons** typen	(kop/dons)
Je moet **bunt** schrijven	(bult/munt)
Je moet **zamp** typen	(zand/kamp)
Je moet **poms** typen	(pols/soms)
Je moet **delt** schrijven	(den/speld)
Je moet **verp** typen	(velg/scherp)
Je moet **gink** typen	(gif/zink)
Dit als **taars** typen	(taak/laars)
Je moet **kerg** typen	(kern/dwerg)
Je moet **pals** typen	(palm/hals)
Je moet **sork** typen	(sop/vork)
Je moet **felk** typen	(ferm/melk)

Set 3: /s/-plosive onset; obstruent-obstruent coda

Je moet **spogt** schrijven	(spons/bocht)
Je moet **stups** typen	(stuk/rups)
Je moet **schits** typen	(schip/flits)
Je moet **steks** typen	(stem/heks)
Je moet **spaft** schrijven	(spat/kaft)
Je moet **schoost** schrijven	(school/troost)
Je moet **spoest** schrijven	(spoed/hoest)
Je moet **steeft** schrijven	(steek/kreeft)
Je moet **scheeks** typen	(scheef/reeks)
Je moet **spiets** typen	(spier/fiets)
Je moet **stoops** typen	(stoof/loops)
Je moet **schugt** schrijven	(schub/zucht)

Appendix B (cont.)

Set 4

Je moet **kwaag** typen	(kwaal/zaag)
Je moet **knoot** schrijven	(knoop/rood)
Je moet **kleip** typen	(klein/pijp)
Je moet **briek** typen	(brief/piek)
Je moet **pluif** typen	(pluis/kuif)
Dit als **traus** typen	(trouw/kous)
Dit als **twig** typen	(twist/big)
Je moet **vlat** schrijven	(vlag/blad)
Je moet **frop** typen	(front/krop)
Je moet **gleuk** typen	(gleuf/deuk)
Je moet **sleef** typen	(sleep/neef)
Je moet **smes** typen	(smet/fles)

Set 5

Je moet **spleem** typen	(spleet/zeem)
Je moet **sprein** typen	(sprei/lijn)
Je moet **schroel** typen	(schroef/doel)
Je moet **straai** typen	(straal/haai)
Je moet **spleen** typen	(splijt/been)
Je moet **spruil** typen	(spruit/kuil)
Je moet **schring** typen	(schrik/ring)
Je moet **strier** typen	(striem/pier)
Je moet **splim** typen	(split/gym)
Je moet **stroei** typen	(stroef/boei)
Je moet **schraar** typen	(schraal/blaar)
Je moet **sprieuw** typen	(spriet/nieuw)

Set 6-10: disyllabic words

Set 6

Je moet **'stader** typen	(stapel/ader)
Je moet **'schager** typen	(schade/hagel)
Je moet **'spiekus** typen	(spikkel/stiekem)
Je moet **'spiezel** typen	(spiegel/kiezel)
Je moet **'schapel** typen	(schakel/kaper)
Je moet **'stafer** typen	(staking/wafel)

Je moet **schap'piel** typen	(scharnier/papier)
Je moet **spi'dool** typen	(spiraal/idool)
Je moet **stak'ket** schrijven	(statief/pakket)
Je moet **spi'foom** typen	(spion/atoom)
Je moet **scha'zant** schrijven	(schavot/fazant)
Je moet **stag'gon** typen	(statuut/wagon)

Set 7

Je moet **'klappus** typen	(klamboe/lobbes)
Je moet **'blartif** typen	(blanco/kalief)
Je moet **'griedak**typen	(griezel/bivak)
Je moet **'florrog** typen	(flodder/oorlog)
Je moet **'kreelup**typen	(krekel/hennep)
Je moet **'dresset**schrijven	(drempel/debet)

Appendix B (cont.)

Je moet **klar'rug** typen (klassiek/terug)
Je moet **bla'pet** schrijven (blazoen/ballet)
Je moet **griek'tes** typen (grimas/succes)
Je moet **flor'rak** typen (florijn/tabak)
Je moet **krum'mop** typen (krediet/galop)
Je moet **drep'pief** typen (dressuur/motief)

Set 8
Je moet **'golder** typen (gordel/zolder)
Je moet **'darken** typen (dartel/Parker)
Je moet '**kormel** typen (korrel/mormel)
Je moet **'sompel** typen (somber/dompel)
Je moet '**pantel** typen (panter/mantel)
Je moet '**vartel** typen (varken/marter)

Je moet **kol'daar** typen (kolom/soldaat)
Je moet **var'ket** schrijven (vandaal/parket)
Je moet **gor'mool** typen (gordijn/hormoon)
Je moet **pom'pat** schrijven (pompoen/kompas)
Je moet **san'tiel** typen (sandaal/antiek)
Je moet **dar'tiek** typen (damast/artiest)

Set 9
Je moet **'pammuk** typen (passie/jammer)
Je moet **'dierem** typen (diva/sire)
Je moet **'kallut** schrijven (kachel/alles)
Je moet **'gienel** typen (giro/sinus)
Je moet **'vajes** typen (vader/bajes)
Je moet **'suwor** typen (super/fluor)

Je moet **pa'meek** typen (paniek/kameel)
Je moet **di'room** typen (direct/siroop)
Je moet **kal'leet** schrijven (kado/alleen)
Je moet **gie'neel** typen (gitaar/diner)
Je moet **va'joor** typen (vallei/majoor)
Je moet **su'week** typen (subiet/fluweel)

Set 10
Je moet **'rakel** typen (radar/lepel)
Je moet **'liestem** typen (linie/bodem)
Je moet **'morkuw** typen (monnik/schaduw)
Je moet **'natum** typen (nagel/datum)
Je moet **'juuker** typen (judo/beker)
Je moet **'woeren** typen (woede/baken)

Je moet **ra'teel** typen (raket/gareel)
Je moet **li'puum** typen (limiet/kostuum)
Je moet **mop'paai** typen (montuur/lawaai)
Je moet **na'feem** typen (natuur/probleem)
Je moet **ju'dier** typen (juweel/manier)
Je moet **woe'gaan** typen (woestijn/banaan)

Appendix C

Confusion matrix of onset consonants (monosyllabic nonwords of Chapter 2). Down: presented consonants. Across are the responses (percentages of all responses; based on 32 observations). Correct identification scores (in percentages) are in bold. The data are collapsed over the two time-compressed conditions (i.e., compression to 35% and 40%).

	Plosive responses					Fricative responses						Nasals		Liquids		Semi-vowel		Rest	
Target ↓	/k/	/b/	/p/	/d/	/t/	/z/	/s/	/ʃ/	/v/	/f/	/χ/	/m/	/n/	/l/	/r/	/ʊ/	/j/	/h/	/-/
/k/	**76**	1		2	13								1			1			6
/b/		**81**	3									8	3						5
/p/	2		**48**		2							3				1			44
/d/				**89**								1	3				3		4
/t/				2	**0**							6	20	20		2			50
/z/						**39**	59	2											
/s/						38	**55**	5											2
/v/									**86**	6	3				2	2			1
/f/									86	**14**									
/χ/					2		1			2	**88**								7
/m/												**94**	2	4					
/n/												12	**88**						
/l/													2	**89**	3	2			4
/r/													2	23	**59**			8	7
/ʊ/												4	4	30	10	**39**		3	10
/j/													3	2			**95**		

Appendix D

Confusion matrix of coda consonants (monosyllabic nonwords of Chapter 2). Down: presented consonants. Across are the responses (percentages of all responses, based on 32 observations). Correct identification scores (in percentages) are in bold. The data are collapsed over the two time-compressed conditions (i.e., compression to 35% and 40%).

	/t/	/p/	/k/	/s/	/f/	/χ/	/n/	/m/	/ŋ/	/l/	/r/	/ʊ/	/-/
/t/	**87**	9											4
/p/	1	**59**	24				3	6					7
/k/	2	15	**77**										6
/s/				**93**									7
/f/					**100**								
/χ/						**99**							1
/n/							**70**	12	16	2			
/m/							51	**22**	27				
/ŋ/							22	5	**73**				
/l/							4	1		**90**			5
/r/		1	3		2	34	2	1		2	**52**		3
/ʊ/							3	2		81		**14**	

Appendix E

List of 32 target nouns in production experiment (Chapter 4; experiment 1)

	Unstressed vowel schwa		**'Full' unstressed vowel**	
	initial stress	*final stress*	*initial stress*	*final stress*
Long vowel in stressed syllable	beker	bedrijf	specie	kopij
	'beaker'	'company'	'mortar'	'copy'
	schade	betoog	havik	saucijs
	'damage'	'argumentation'	'hawk'	'sausage'
	code	getij	sofa	octaaf
	'code'	'tide'	'sofa'	'octave'
	pater	bereik	foto	pastei
	'father'	'reach'	'photo'	'pie'
Short vowel in stressed syllable	stekker	gebod	ghetto	kopie
	'plug'	'command'	'ghetto'	'copy'
	ticket	bestek	toffee	schavot
	'ticket'	'cutlery'	'toffee'	'scaffold'
	stakker	gezag	asbest	pakket
	'poor wretch'	'authority'	'asbestos'	'parcel'
	bakkes	gebit	sabbat	effect
	'mug'	'set of teeth'	'sabbath'	'effect'

APPENDIX F

Word and sentence material of the present cross-modal semantic priming study, based on Zwitserlood (1989), with some changes in the choice of competitors and associates. First, the biasing sentences and their control sentences (between brackets) are shown. The target words, control words, competitor words and their respective visual probes are presented in a table below that. Below this table, both Zwitserlood's materials and the present materials are listed, in order to show the differences.

1. Deze loodgieter is erg vakbekwaam, goedkoop en snel. Hij krijgt daardoor veel *klanten*. (Zoals iedere zomer namen ze een lange vakantie. Dit jaar gingen ze naar de *bergen.)*
2. De boer had lange tijd over de prijs onderhandeld. Uiteindelijk kocht hij de *schapen*. (Af en toe neem ik wel eens een lekkere bak friet. Maar verder eet ik vooral veel *fruit*.)
3. Moe en stoffig zochten de pelgrims een onderkomen. Gelukkig kwamen ze spoedig bij de *herberg*. (Piet is niet zo'n handige klusser. Nu heeft hij weer een pleister om zijn *pink*.)
4. De kinderen van de derde klas waren op schoolreis in Antwerpen. Ze liepen langs de *haven.* (Jan kon het niet hebben dat zijn zusje chocolademelk had gekregen van tante Sien. Stiekem verstopte hij haar *beker*.)
5. De zakenman ging met de ondergrondse naar zijn werk. Onder zijn arm hield hij een *krant.* (Kun je deze mappen even voor me opbergen? Er is op Harry's kamer nog een lege *plank*.)
6. Willem struikelde over een steen en viel languit op het pad. Zijn haren zaten vol met *bloed. (*Morgen gaan we naar München. Maar als het zo mistig blijft neem ik liever de *trein.)*
7. Dit bedrijfje gebruikt uitsluitend natuurlijke grondstoffen voor haar producten. Ze maken er verschillende soorten *kaas.* (De dokter had weer veel verschillende klachten gehoord op het spreekuur. Eén dame had een afspraak voor haar *heup.)*
8. Marlene is zeer tevreden over haar nieuwe baan. Ze heeft ook een prima *salaris.* (Het huis was sfeervol ingericht. In de woonkamer stond een oude *piano.)*
9. De bewoners van het eiland zaten vol angstige spanning rond de radio. Ze luisterden naar de aankondiging van een *orkaan.* (Nieuwsgierig liepen de kinderen over het terrein om het vreemdsoortige bouwsel te bekijken. De hele constructie was van *beton*.)
10. Voor het eerst van haar leven bezocht Christine Athene. Ze fotografeerde een paar oude *pilaren.* (Meestal waren er meer dan genoeg frambozen aan de struik. Lisa had haar diepvries vol na zo'n *seizoen.)*
11. Pamela stond al in de gang toen de taxi voorreed. Ze pakte snel haar *koffer.*
(Jan vroeg of Suzan een boek voor hem mee wilde nemen uit de bibliotheek. Onderweg vergat ze alleen helaas de *titel.)*
12. De gevangene kreeg een stuk brood voorgezet. Hij durfde niet te klagen over zijn *dorst.* (Kennelijk hadden ze ruzie gehad. Hij keek snel de andere kant op als hij haar tegenkwam op de *gang.)*

Appendix F (cont.)

13. Oma bewaart alles waarvan ze denkt dat het later ooit nog van pas kan komen. Ze heeft bijvoorbeeld een doos vol *knopen*. (Het rommelhok van Peter stond bomvol. Allereerst stond er nog voor jaren *verf.)*

14. In bedrukte stemming stonden de mannen rond het graf. Ze treurden om het verlies van hun *kapitein.* (Ik zag een man op de hoek van de straat. Hij zocht zenuwachtig naar een *sigaret.)*

15. Vader was druk in de weer met zijn gereedschapskist. Hij repareerde de kapotte *kraan*. (Piet en Klazien hebben een heel mooi huis. Ze hebben ook zo'n mooie *trap.)*

16. De stervende vrouw geloofde rotsvast in het bestaan van een hiernamaals. Haar hart was vol *vrede.* (Kees vond de scheikundelessen meestal wel interessant. Op dinsdag was het in ieder geval leuk, want dan deden ze altijd een *proef.)*

17. De kogel had de misdadiger blijkbaar toch getroffen. Snel keek hij naar zijn b*een.* (Gisteren mocht ik zijn verzameling bekijken. Hij had één munt met een heel bijzondere *kleur.)*

18. Meneer Willems was zeer onder de indruk van het geheel gerenoveerde grachtenpand. Hij bewonderde vooral de prachtige *kozijnen.* (Jan vertelde enthousiast over zijn nieuwe hobby diepzeeduiken. Hij had prachtige koraalriffen gezien en *dolfijnen).*

19. De bezoekers probeerden zo snel mogelijk het terrein te verlaten. Overal om zich heen zagen ze de *vlammen.* (Jan had de hele middag nauwelijks omkijken naar zijn dochtertje Marleen. De kleuter speelde rustig met de *blokken.)*

20. Toen moeder de kleintjes onder de wol had gestopt, vertelde ze hen nog een verhaaltje. Het ging over een *draak*. (Als je een CD-zaak zoekt weet ik nog wel een goeie. Er zit een heel goedkope zaak op dat grote *plein.)*

21. Na een denderende ruzie met haar vriend bleef José alleen achter. Ontmoedigd keek ze naar de *scherven.* (Het was niet duidelijk of er wel genoeg zitplaats zou zijn voor de vergadering. Ze liepen af en aan met stoelen en *krukken*.)

22. Op de hacienda's in Brazilie komt nog steeds veel kinderarbeid voor. De kinderen moeten vaak werken in de *stallen.* (Op de camping waren altijd wel wat schoonmaakklusjes te doen. Meestal begon ze met de *douche*.)

23. Fransje was weer eens in de modder gevallen. Zuchtend waste moeder zijn *broek.* (De zomer vind ik zo'n heerlijk jaargetijde. Ik kan nu alweer verlangen naar de zon en de geur van pas gemaaid *gras*.)

24. Die eeuwige slordigheid van Maria heeft af en toe gevaarlijke consequenties. Nu ligt de hele vloer vol met *spelden.* (De benzinemeter stond al aardig in het rood. Opeens hoorden ze een harde *knal.)*

Appendix F (cont.)

The table below lists the prime words, competitor words, their respective visual probes, and auditory control words.

nr	auditory prime		visual probe		auditory control
	prime	competitor	prime's probe	competitor's probe	
1	klanten	klappen	winkel	slaan	bergen
	'customers'	'slaps'	'shop'	'smack'	'mountains'
2	schapen	schaar	wol	knippen	fruit
	'sheep'	'scissors'	'wool'	'cut'	'fruit'
3	herberg	hert	slapen	bos	pink
	'inn'	'deer'	'sleep'	'forest'	'little finger'
4	haven	hamer	boot	spijker	beker
	'harbour'	'hammer'	'boat'	'nail'	'cup'
5	krant	kramp	nieuws	pijn	plank
	'newspaper'	'cramp'	'news'	'pain'	'shelf'
6	bloed	bloem	rood	geur	trein
	'blood'	'flower'	'red'	'smell'	'train'
7	kaas	kaarsen	brood	licht	heup
	'cheese'	'candles'	'bread'	'light'	'hip'
8	salaris	salami	geld	worst	piano
	'salary'	'salami'	'money'	'sausage'	piano'
9	orkaan	orkest	wind	muziek	beton
	'hurricane'	'orchestra'	'wind'	'music'	'concrete'
10	pilaren	piloten	kerk	vliegtuig	seizoen
	'pillars'	'pilots'	'church'	'plane'	'season'
11	koffer	koffie	reis	thee	titel
	'suitcase'	'coffee'	'trip'	'tea'	'title'
12	dorst	dorp	drinken	klein	gang
	'thirst'	'village'	'drink'	'small'	'corridor'
13	knopen	knoken	gat	bot	verf
	'buttons'	'knuckles'	'hole'	'bone'	'paint'
14	salaris	salami	zee	rijk	sigaret
	'captain'	'capital'	'sea'	'rich'	'cigarette'
15	kraan	kraag	water	jas	trap
	'tap'	'collar'	'water'	'coat'	'stairs'

16	vrede	vrees	oorlog	angst	proef
	'peace'	'fear'	'war'	'fear'	'test'
17	been	beest	lopen	dier	kleur
	'leg'	'animal'	'walk'	'animal'	'colour'
18	kozijnen	kozakken	raam	Rus	dolfijnen
	'frame'	'cossacks'	'window'	'russian'	'dolphins'
19	vlammen	vlaggen	brand	wimpel	blokken
	'flames'	'flags'	'fire'	'banner'	'blocks'
20	draak	draad	vuur	naald	plein
	'dragon'	'thread'	'fire'	'needle'	'square'
21	scherven	schelpen	glas	strand	krukken
	'splinters'	'shells'	'glass'	'beach'	'stools'
22	stallen	stad	paard	druk	douche
	'stables'	'city'	'horse'	'busy'	'shower'
23	broek	broer	pijp	zus	gras
	'pants'	'brother'	'trouser leg'	'sister'	'grass'
24	spelden	spek	naaien	varken	knal
	'pins'	'bacon'	'sew'	'pig'	'crack'

Note: Below the actual prime word and its competitor are listed, respectively, each followed by its related visual probe in uppercase. The left-hand prime/probe pairs are from Zwitserlood's material; the right-hand prime/probe pairs were used in the present replication experiment.

Zwitserlood (1989) material	**Material present study**
1 klanten/KONING, klappen/PIJN	klanten/WINKEL, klappen/SLAAN
2 schapen/WOL, schaar/MES	schapen/WOL, schaar/KNIPPEN
3 herberg/BIER, hertog/GRAAF	herberg/SLAPEN, hert/BOS
4 haven/BOOT, haver/GORT	haven/BOOT, hamer/SPIJKER
5 krant/LEZEN, krans/DOOD	krant/NIEUWS, kramp/PIJN
6 bloed/ROOD, bloesem/LENTE	bloed/ROOD, bloem/GEUR
7 kaas/MELK, kabel/TOUW	kaas/BROOD, kaarsen/LICHT
8 salaris/GELD, salami/WORST	salaris/GELD, salami/WORST
9 orkaan/WIND, orkest/MUZIEK	orkaan/WIND, orkest/MUZIEK
10 pilaren/KERK, piloten/VLIEGTUIG	pilaren/KERK, piloten/ VLIEGTUIG
11 koffer/REIS, koffie/THEE	koffer/REIS, koffie/THEE
12 dorst/DRINKEN, dorp/STAD	dorst/DRINKEN, dorp/KLEIN
13 knopen/JAS, knoken/BOT	knopen/GAT, knoken/BOT
14 kapitein/SCHIP, kapitaal/GELD	kapitein/ZEE, kapitaal/RIJK

Appendix F (cont.)

15 kraan/WATER, kraag/BOORD	kraan/WATER, kraag/JAS
16 vrede/OORLOG, vrees/ANGST	vrede/OORLOG, vrees/ANGST
17 been/BOT, beest/DIER	been/LOPEN, beest/DIER
18 kozijnen/RAAM, kozakken/RUS	kozijnen/RAAM, kozakken/RUS
19 vlammen/VUUR, vlaggen/WIMPEL	vlammen/BRAND,vlaggen/ WIMPEL
20 draak/VUUR, draad/NAALD	draak/VUUR, draad/NAALD
21 scherven/GLAS, schelpen/STRAND	scherven/GLAS, schelpen/ STRAND
22 stallen/PAARD, stad/DORP	stallen/PAARD, stad/DRUK
23 broek/RIEM, broer/ZUS	broek/PIJP, broer/ZUS
24 spelden/NAAIEN, spek/VARKEN	spelden/NAAIEN, spek/VARKEN

Samenvatting in het Nederlands

Er zijn vele applicaties denkbaar waarbij kunstmatige versnelling van spraak nuttig zou kunnen zijn. Versnelde spraak wordt bijvoorbeeld gebruikt om voicemail-berichten versneld af te luisteren, maar zou ook handig kunnen zijn om lange opnames snel door te kunnen luisteren. Bij de meeste tekst-naar-spraaksystemen kunnen de gebruikers zelf het gewenste voorleestempo instellen.

In 1991 werd in Nederland het *Electronisch Lezen van een Krant* (ELK) project gestart dat als doel had het voor visueel gehandicapten makkelijker te maken om snel kennis te nemen van het nieuws (in dit geval het dagblad Trouw). De digitale tekst van deze krant kon hoorbaar gemaakt worden met behulp van een tekst-naar-spraaksysteem. De spraakkwaliteit en segmentele verstaanbaarheid van het spraak-synthesesysteem waren niet al te best (zie evaluatierapport Jongenburger & van Bezooijen 1992), maar desondanks gaven sommige luisteraars aan dat ze de voorkeur gaven aan een versneld afspeeltempo. Onderzoek uit de jaren '60 (Zemlin, Daniloff & Shriner 1968) had ook al laten zien dat het begrip van versnelde spraak relatief intact blijft bij versnelling tot twee keer het normale tempo, maar dat het luisteren dan wel meer moeite kost. Het ELK evaluatierapport liet ook zien dat de luisteraars snel aan de spraaksynthese gewend raakten.

De bevindingen van het ELK-project hebben geleid tot de huidige studie naar de productie en perceptie van snelle spraak. Het hoofddoel van deze studie is om een vergelijking te maken tussen de waarneming van natuurlijke snelle spraak en kunstmatig versnelde spraak. De studie valt uiteen in 4 thema's, te weten:

1. Robuustheid en gemak van verwerking
2. Adaptatie aan snelle spreektempo's
3. Hogere spreektempo's in spraakperceptie dan in spraakproductie
4. Natuurlijk geproduceerde snelle spraak makkelijker te verwerken dan kunstmatig versnelde spraak?

Per thema zal hieronder worden uitgelegd om wat onderzocht werd en waarom, wat de experimenten lieten zien en wat hieruit geconcludeerd mag worden met betrekking tot het proces van spraakproductie en spraakperceptie.

1 Robuustheid en gemak van verwerking

Het feit dat luisteraars sterk versnelde spraak nog goed kunnen verstaan betekent dat veel van het spraaksignaal kennelijk redundant is, d.w.z. gemist kan worden. In hoofdstuk 2 van deze dissertatie werd de robuustheid van spraakperceptie tegen sterke

versnelling onderzocht, onder andere door het effect van lexicale redundantie te bekijken. Lexicale redundantie houdt in dat je bij een echt woord niet elke klank hoeft te verstaan om het woord te kunnen herkennen. De verwachting was dat het effect van lexicale redundantie groter zou zijn na tijdscompressie en dit werd onderzocht aan de hand van het verschil in verstaanbaarheid tussen echte woorden en onzinwoorden. De identificatie van echte woorden heeft inderdaad relatief weinig te lijden onder sterke versnelling (ten opzichte van het originele tempo), maar de identificatie van niet-bestaande woorden wel. Zo helpt lexicale redundantie om niet-herkende segmenten aan te vullen.

Voor de niet-bestaande woorden kon onderzocht worden of de robuustheid tegen versnelling mede afhangt van de segmenten zelf: klanken met een langer steady-state (stabiel) stuk, zoals klinkers en fricatieven, zouden beter bestand moeten zijn tegen tijdscompressie dan klanken met een korter of zelfs geen steady-state stuk (zoals plosieven). De identificatie van klinkers en fricatieven bleek inderdaad het minst te lijden onder sterke versnelling en de identificatie van plosieven het meest.

Hoewel sterk versnelde spraak nog goed verstaanbaar kan zijn kost het luisteren ernaar meer moeite dan het luisteren naar spraak op een gewoon tempo. In hoofdstuk 3 werd onderzocht in hoeverre die extra moeite meetbaar is: het verwerken van 'normale' spraak werd vergeleken met het verwerken van versnelde maar nog goed verstaanbare spraak (ongeveer 1.5 keer sneller dan normaal). De toegenomen verwerkingsdruk vertaalde zich niet in langere reactietijden (die waren juist iets korter in de versnelde dan in de normale conditie), maar wel in een groter foutpercentage. Luisteraars slagen er dus redelijk in hun verwerkingssnelheid aan te passen aan het snelle afspeeltempo, maar wel ten koste van de nauwkeurigheid. De redundantie van normaal-tempo spraak maakt de verwerking ervan makkelijker en daardoor ook beter bestand tegen verstoringen.

2 Adaptatie aan snelle spreektempo's

In hoofdstuk 2 werd ook het proces van adaptatie (of 'gewenning') aan snelle spraak onderzocht. Voor adaptatie is slechts een beperkte hoeveelheid materiaal nodig: aan het eind van het experiment (na ongeveer 30 minuten) leek het erop dat de identificatiepercentages niet meer verder stegen. De gewenning die optrad bij de proefpersonen uit het eerste experiment was echter na een paar maanden weer verdwenen: luisteraars die niet hadden meegedaan aan het eerste experiment scoorden nauwelijks slechter dan luisteraars die voor de tweede keer naar sterk versnelde spraak luisterden. Dit betekent dat deze gewenning geen expliciet leerproces is, maar een flexibel aanpassen aan de huidige luisteromstandigheden.

3 Hogere spreektempo's in spraakperceptie dan in spraakproductie

In hoofdstukken 4 en 5 werd snelle spraak ontlokt aan sprekers. Als de sprekers gedwongen werden zeer snel te spreken (zonder dat het absoluut onverstaanbaar werd), haalden ze articulatiesnelheden van ongeveer 10 lettergrepen per seconde (vgl. hun normale spreektempo van gemiddeld 6.7 lettergrepen/sec). Hoewel sprekers erg hard hun best doen slagen ze er niet in hun articulatiesnelheid te verdubbelen. En dat terwijl de verstaanbaarheid van spraak die kunstmatig versneld is tot tweemaal de originele snelheid nauwelijks problemen oplevert voor luisteraars. Deze discrepantie tussen productie en perceptie moet veroorzaakt worden door beperkingen aan de spraakproductie. Articulatoren hebben een minimum duur nodig om hun articulatorisch/akoestisch gedefinieerde doel te bereiken (Kiritani 1977; McClean 2000; Perkell 1997). Zo is de kaak is een relatief langzame articulator. Afgezien van beperkingen aan de maximale spreeksnelheid op het laagste fysische niveau, zijn er wellicht ook beperkingen op het motorcommando-niveau of op hogere spraakplanningsniveaus. De discrepantie tussen wat mensen aankunnen als luisteraars en wat ze zelf kunnen produceren wordt dus veroorzaakt door beperkingen op verschillende niveaus van spraakproductie.

4 Natuurlijk geproduceerde snelle spraak makkelijker te verwerken dan kunstmatig versnelde spraak?

Hoewel sommige sprekers erin zullen slagen om sneller te spreken en tegelijkertijd nauwkeurig te blijven articuleren, werd voor deze studie aangenomen dat de articulatietempo's dermate hoog waren dat 'gereduceerde articulatie' onvermijdelijk was. Snel uitgesproken spraak bleek dan ook minder makkelijk te verwerken dan kunstmatig versnelde spraak, zelfs op een tempo waarbij beide typen spraak nog perfect verstaanbaar waren (hoofdstuk 5). Hoewel luisteraars geen problemen hebben met bijv. het ongedaan maken van assimilatie bij normaal spreektempo (vgl. studies van Gaskell & Marslen-Wilson 1996, 1998) zijn ze sneller met het verwerken van redundante of niet-gereduceerde woordvormen als ze snelle spraak aangeboden krijgen. Kohler (1990) noemde assimilatie 'perceptually tolerated articulatory simplification'. Luisteraars tolereren de gereduceerde vorm van articulatie in normale luisteromstandigheden, maar in moeilijke luisteromstandigheden hebben ze liever spraak die zo netjes mogelijk is. Sprekers kunnen hun spreektempo kennelijk niet verhogen zonder het verwerkingsproces voor de luisteraar moeilijker te maken.

De tweede vraag is of dit laatste ook geldt voor de prosodische (meer specifiek, temporele) veranderingen die de spreker toepast bij het versnellen van spraak. Uit eerder onderzoek was al gebleken dat er verschillen zijn tussen de temporele organisatie van normale spraak en spraak die snel wordt uitgesproken. Eén van de meest in het oog

lopende verschillen is het pauzegedrag: als sprekers hun spreektempo verhogen zullen ze dat onder meer doen aan de hand van het weglaten en/of sterk verkorten van spreekpauzes. Afgezien daarvan blijkt de spreker sommige stukken spraak sterker te versnellen dan andere stukken. De verwachting voor de productiestudie uit hoofdstuk 4 was dat onbeklemtoonde lettergrepen (in meerlettergrepige woorden) sterker verkort zouden worden dan beklemtoonde lettergrepen omdat de spreker er bij het sneller spreken rekening mee houdt dat de beklemtoonde lettergrepen het meest informatief zijn. Deze verwachting was afgeleid van de Hyper- en Hypoarticulatietheorie van Lindblom (1990). Deze theorie komt erop neer dat de spreker voortdurend een afweging maakt tussen wat wat hij/zij zichzelf kan permitteren en wat de luisteraar nodig heeft. De resultaten van de productiestudie leken deze hypothese te bevestigen: onbeklemtoonde ('onbelangrijke') lettergrepen werden sterker verkort dan beklemtoonde ('belangrijke'). In daarop volgende luisterexperimenten werd onderzocht of de verstaanbaarheid van kunstmatig versnelde spraak verbeterd zou kunnen worden door niet *linear* te versnellen, maar door de onbeklemtoonde lettergrepen sterker te verkorten dan de beklemtoonde om zodoende het temporeel patroon van kunstmatig versnelde spraak meer te laten lijken op dat van natuurlijke snelle spraak. Uit de luisterexperimenten bleek echter dat verstaanbaarheid en verwerkingsgemak juist verslechterden als de temporele structuur van snel uitgesproken spraak werd aangebracht, ten opzichte van lineaire versnelling. Het lijkt erop dat het sterk verkorten van pauzes (sterker dan de resterende spraak) het enige aspect is van natuurlijk geproduceerde snelle spraak dat de verstaanbaarheid van kunstmatig versnelde spraak zou kunnen verbeteren t.o.v. lineaire versnelling (zie resultaten met Mach1 algoritme; Covell, Withgott & Slaney 1998): de resterende spraak kan dan immers minder sterk versneld worden. Niet-lineaire aanpassingen onder het fraseniveau (dus afgezien van die pauzes) die de spreker toepast bij sneller spreken leveren alleen een verslechtering van de verstaanbaarheid op. De natuurlijke niet-lineaire manier van versnellen is dus niet bedoeld om het de luisteraar makkelijker te maken, maar sprekers zijn waarschijnlijk niet in staat om min of meer lineair te versnellen. Dit zou het gevolg kunnen zijn van de manier waarop meerlettergrepige woorden gespecificeerd zijn in het mentale lexicon. Lexicale klemtoon is onderdeel van die uitspraakspecificatie. Als gevolg hiervan worden beklemtoonde lettergrepen uitgesproken met meer articulatorische precisie dan onbeklemtoonde (van Bergem 1993, Lehiste 1970). De articulatorische of akoestische doelen in de opgeslagen uitspraakrepresentatie zijn nauwkeuriger gespecificeerd voor beklemtoonde dan voor onbeklemtoonde lettergrepen. De Jong (1995) noemt klemtoon dan ook 'locale hyperarticulatie'. Als nu deze doelen voor beklemtoonde lettergrepen nauwkeuriger gespecificeerd zijn voor beklemtoonde dan voor

onbeklemtoonde lettergrepen, moet de spreker wel meer moeite doen om die doelen ook te bereiken en zal daarom tijdelijk minder goed kunnen versnellen.

Een sterker prosodisch patroon (meer uitgesproken afwisseling van sterke en zwakke lettergrepen) helpt de luisteraar kennelijk niet bij de woordherkenning. Er lijkt een optimale balans te bestaan tussen de bijdrage van prosodie en segmentele inhoud aan woordherkenning. In hoofdstuk 3 werd deze interactie tussen prosodie en segmentele inhoud ook nog eens benadrukt. In die studie werd onderzocht wat het effect van tijdscompressie was op de verwerking van natuurlijke spraak en op die van synthetische difoonspraak. Difoonspraak bestaat uit aan elkaar geregen stukjes spraak waarbij de bouwstenen (in dit geval) allemaal oorspronkelijk beklemtoond en gehyperarticuleerd zijn. Normaal gesproken is natuurlijke spraak gemakkelijker te verwerken dan synthetische spraak. Dit verwerkingsvoordeel werd echter nog groter na tijdscompressie. Dit zou te maken kunnen hebben met het feit dat beklemtoonde en onbeklemtoonde lettergrepen wel verschillen in duur in de synthetische spraak, maar er is nauwelijks afwisseling tussen sterke (luide/nauwkeurig uitgesproken) en zwakkere (minder luide/minder nauwkeurig uitgesproken) lettergrepen. Hoewel hyperarticulatie wellicht de fonetische verwerking van klanken vergemakkelijkt, levert het gebrek aan variatie in sprekerinspanning een probleem op bij het groeperen van zwakke en sterke lettergrepen tot één woord of frase, juist in moeilijke luisteromstandigheden. Segmentele en prosodische factoren dragen beide bij aan het proces van woordherkenning. Hun relatieve bijdrage lijkt onafhankelijk te zijn van tempo: teveel nadruk op de één verstoort de balans tussen een natuurlijk prosodisch patroon en een verstaanbaar spraaksignaal. Hoe meer een spraaksignaal afwijkt van de 'normale' vorm, des te meer moeite zal de luisteraar hebben om dit te 'mappen' op het mentale lexicon.

De experimenten hebben ook laten zien dat spraakwaarneming erg flexibel is in het omgaan met minder verstaanbare spraak. Woordherkenning kan eventueel wat vertraagd worden, maar vaak kan hogere-orde-informatie helpen om het lexicale-competitieproces op te lossen. In normale lopende spraak varieert het spreektempo ook continu. Een locaal hoger spreektempo is niet alleen nadelig voor spraakwaarneming: hoewel de locale woordherkenning misschien even moeilijker is geeft het hogere tempo ook gelijk weer aan dat het om een redundant of minder belangrijk onderdeel gaat. Over het geheel genomen is variatie in spreektempo dus niet problematisch, maar is juist nuttige informatie voor de meer globale niveaus van spraakwaarneming en taalbegrip.

CURRICULUM VITAE

Esther Janse was born on 23 November 1973 in Grijpskerke. She attended the Christelijke Scholengemeenschap Walcheren in Middelburg where she obtained her VWO diploma in 1992. As from 1993, she studied at the University of Leiden, and obtained her Master's degree in English (with a specialisation in phonetics and linguistics) in 1997. From 1997 to 2002 she was employed as a PhD student by the Utrecht institute of Linguistics OTS. This thesis is the result of the research carried out during that period. In October 2002, she started working as a post-doc researcher at the UiL OTS for the NWO project 'Auditory perception in healthy speakers and in speakers with acquired or developmental language impairments'.