On the quality of synthetic speech : evaluation and improvements

On the Quality of Synthetic Speech
Evaluation and Improvements

J.H. Eggen

On the Quality of Synthetic Speech

Evaluation and Improvements

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Technische Universiteit Eindhoven,
op gezag van de Rector Magnificus, prof. dr. J.H. van Lint,
voor een commissie aangewezen door het College van Dekanen
in het openbaar te verdedigen
op dinsdag 17 november 1992 om 16:00 uur

door

Josephus Hubertus Eggen

geboren te Vlaardingen

Dit proefschrift is goedgekeurd door de promotoren:

prof. dr. S.G. Nooteboom
prof. dr. A.J.M. Houtsma

The work described in this thesis has been carried out at the Philips Research Laboratories as part of the Philips Research programme.

voor Kitty
en mijn ouders

Het oort, oogt, lipt;
waar het wil plant het zin,
en kleedt het aan met zien.

Sybren Polet

Voorwoord

HET is woensdagavond, 23 september. Morgen gaat het proefschrift naar de drukker. De hoogste tijd om de laatste tekst, die de eerste zal zijn, aan het proefschrift toe te voegen. Dadelijk Sieb bellen of ik niks vergeten ben, en thuis Kit nog even dit stukje laten lezen. Zouden de jongens trouwens al slapen?

Dank aan de volgende mensen, die elk op hun eigen wijze hebben bijgedragen aan de totstandkoming van dit proefschrift.

- Sieb Nooteboom, die er op belangrijke momenten was.
- Aad Houtsma voor zijn enthousiaste begeleiding.
- De medewerkers van het Instituut voor Perceptie Onderzoek, die het instituut maken tot wat het is: een stimulerende en bijzonder plezierige omgeving voor het doen van onderzoek.
- De uitvoerders en stuurgroepleden van het SPIN-programma "Analyse en Synthese van Spraak" voor de prettige samenwerking.
- Hugo van Leeuwen, de LaTeX-expert.
- René Collier, iemand met gevoel voor mensen en muziek.
- Mijn (koffie-)drinkebroeders: Marc, Niek, Roel en Wil.
- De thuisbasis: Kit, Daan en Niels.

Eindhoven, 23 september 1992
Berry Eggen

Contents

List of Figures x
List of Tables xi

1 Introduction 1

2 Intelligibility of synthetic speech in the presence of interfering speech 5
2.1 Introduction 6
2.2 Monosyllabic Adaptive Speech Interference Test (MASIT) 8
2.3 Evaluation study 9
2.3.1 Speech types 9
2.3.2 Speech materials 11
2.3.3 Interfering speech 12
2.3.4 Subjects 13
2.3.5 Procedure 13
2.4 Results 13
2.4.1 Articulation scores 14
2.4.2 Q measures 15
2.5 Discussion 15

3 On the role of amplitude and phase in the synthesis of male and female voices 23
3.1 Introduction 24
3.2 Speech processing 27
3.2.1 Objectives and general concept 28
3.2.2 Speech analysis 30
3.2.3 System overview 32
3.2.4 System performance evaluation 34
3.3 Experiment 35

3.3.1 Stimuli . . . 37
Speech material . . . 37
Stimulus generation . . . 38
3.3.2 Subjects . . . 41
3.3.3 Procedure . . . 42
3.3.4 Results . . . 42
Statistical analysis . . . 43
Quality judgments . . . 44
Male/Female differences . . . 47
Remarks on the experimental method . . . 48
3.4 Discussion . . . 50
3.4.1 Speech processing . . . 50
3.4.2 Synthesis of high-quality natural-sounding speech 53
3.5 Conclusions . . . 55

4 Contributions of voice-source and vocal-tract characteristics to speaker identity 57
4.1 Introduction . . . 58
4.2 Speech processing . . . 60
4.2.1 Modeling voiced speech sounds . . . 60
4.2.2 The glottal-excited (GE) speech synthesizer . . . 63
The vocal-tract model . . . 64
The voice-source model . . . 65
4.2.3 Speech analysis . . . 67
Correction of low-frequency phase distortion . 67
Pitch-synchronous segmentation . . . 67
Inverse filtering . . . 69
Stylization of the radiated glottal pulse . . . 71
4.3 Experiment . . . 72
4.3.1 Earlier studies on speaker identity . . . 73
Production differences between speakers . . . 73
Perceptual differences between speakers . . . 74
Hybrid voice samples . . . 75
4.3.2 Stimuli . . . 78
Speakers . . . 78
Recordings . . . 78
Natural vowels . . . 78
Resynthesized vowels . . . 79

Hybrid vowels 79
4.3.3 Subjects 79
4.3.4 Procedure 79
Familiarization 80
Training 81
Testing: natural vowels 81
Testing: resynthesized and hybrid vowels . . . 82
4.3.5 Results 82
Acoustic data 82
Natural vowels 86
Resynthesized vowels 88
Hybrid vowels 88
Informal comments made by the subjects . . . 94
4.4 Discussion 95
4.4.1 Speech processing 95
4.4.2 Perceptual cues for speaker identity 99
4.5 Conclusions 105

5 Concluding remarks **107**
5.1 Limitations of the present research 107
5.2 Improvements 110

References **113**
Summary **125**
Samenvatting **127**
Curriculum Vitae **130**

List of Figures

2.1 Results articulation test 14
2.2 Results MASIT 16
2.3 Coding and masking of acoustic-phonetic properties . . . 19

3.1 Speech production model 29
3.2 Segmentation strategy 32
3.3 Signal processing overview 33
3.4 Pitch-synchronous marking of the speech waveform . . . 39
3.5 Quality scale of the pooled data 46
3.6 Male and female quality scales 48
3.7 Waveforms of four versions of a male utterance 54

4.1 Inverse filtering: vowel /a/ 62
4.2 The glottal-excited (GE) speech synthesizer 64
4.3 The Liljencrants-Fant (LF) model 66
4.4 Low-frequency phase correction 68
4.5 Pitch-synchronous segmentation of voiced speech 70
4.6 Inverse filtering: vowel /i/ 71
4.7 Stylization of the radiated glottal pulse 72
4.8 Speaker-identification scores for the natural vowels 87
4.9 Speaker-identification scores for the resynthesized vowels 89
4.10 Speaker-identification scores for the hybrid vowels 90
4.11 Formant-based accuracy measure 92
4.12 Glottal-based accuracy measure 93
4.13 Speaker-identification scores for the LPC stimuli 97

List of Tables

2.I	Speech-coding types used in the evaluation study	10
3.I	Dutch sentences used in the experiment	38
3.II	Speech utterances used in the experiment	39
3.III	Amplitude/phase conditions	41
3.IV	Quality scales	45
4.I	Acoustic voice-source parameters	83
4.II	Acoustic vocal-tract parameters	84

Chapter 1

Introduction

IN a text-to-speech system we try to model the human letter-to-sound conversion. This is a complex process which involves many steps. If we compare the generation of speech by reading aloud a written text with the generation of music by performing a written piece of music, text-to-speech conversion implies that we have to model both the "instrument" and the performer. The latter is modeled by a set of routines which analyse the semantic, syntactic and lexical structure of the text to provide an abstract underlying linguistic representation (Klatt, 1987). Next, synthesis-by-rule systems are applied on this representation to generate the control parameters for the "instrument". In the case of text-to-speech systems, a speech synthesizer or a speech-coding algorithm plays the role of the "instrument". The research reported on in this dissertation dealt with the evaluation and improvement of this speech "instrument".

Many of the speech-coding algorithms used for the generation of the output speech of text-to-speech systems are based on the source-filter theory of human speech production. According to this theory, speech results from the excitation of the vocal-tract by a sound source. For voiced sounds the source is formed by the air-flow through the glottis which is modulated by the vibrating vocal folds. For unvoiced sounds the source consists of noise generated at constrictions along the vocal-tract. The Linear-Predictive-Coding (LPC) synthesizer is closely related to this source-filter theory of human speech production. The LPC coefficients describe an all-pole filter which models the vocal tract. In the case of voiced sounds, this filter is excited by a quasi-periodic series of delta pulses whereas white noise is used as a source for the generation of unvoiced speech sounds. This approximation of the human speech production process has proved to be very powerful.

Despite its many advantages such as the capability to resynthesize highly intelligible speech, the possibility to manipulate perceived aspects of speech, and the power to provide accurate estimates of speech parameters, LPC also has its shortcomings. LPC speech lacks naturalness and speaker characteristics are degraded. The research reported on in this dissertation was aiming for two things. One was the assessment of some limitations of LPC as a scheme for speech analysis, manipulation and synthesis, the other was the exploration of ways to remove some of the drawbacks of LPC.

If we want to evaluate and improve the quality of synthetic speech we should keep in mind that the term speech quality refers to the total auditory impression of speech on a listener. This means that, besides intelligibility, factors like naturalness, speaker identity, loudness, voice quality, prosodic structure, and many others, contribute to the quality of synthetic speech. In this dissertation we present research on different aspects of speech quality.

In chapter 2 we focus on the intelligibility of synthetic speech, which is an important attribute of speech quality. In general, LPC-based speech-coding schemes are capable of synthesizing highly intelligible speech which cannot easily be discriminated from natural speech by traditional articulation tests like the Modified Rhyme Test (MRT) and the Diagnostic Rhyme Test (DRT) (House, Williams, Hecker & Kryter, 1965; Voiers, 1977). Recently, assessment of synthetic speech has gained much interest in both national and international speech research programs (Fourcin, Harland, Barry & Hazan, 1989; Van Bezooijen & Pols, 1990). This has led to a number of new tests which provide greater sensitivity (Benoit, 1990; Spiegel, Altom & Macchi, 1990; Van Bezooijen & Pols, 1990; Carlson, Granström & Nord, 1992; Steeneken, 1992). In chapter 2 we tried to increase the sensitivity of traditional articulation tests by measuring the intelligibility in the presence of interfering speech. In fact, this technique was borrowed from Nakatani & Dukes (1973) who presented their test stimuli under more difficult listening conditions in order to turn small differences in intelligibility into large differences. Our main goal was to see if, in the case of synthetic speech, small differences in intelligibility can also be magnified into large differences by adding interfering speech. We conducted a perception experiment in which we used both a traditional articula-

tion test without noise and a newly developed "Monosyllabic Adaptive Speech Interference Test (MASIT)" to evaluate the intelligibility of nine different speech-coding schemes.

As said earlier, intelligibility of synthetic speech is just one attribute of speech quality. In chapter 3 we present a study on the naturalness of synthetic speech. As LPC speech lacks naturalness, we tried to determine which requirements are needed for the generation of natural-sounding speech. The LPC residue seems the obvious choice to study in more detail if we want to improve naturalness of LPC speech. It is defined as the difference between the actual speech samples and the linearly predicted ones. Theoretically, this means that the residue contains all information necessary to give LPC speech a natural-sounding quality. The importance of the LPC residue also becomes apparent if we listen to it. In many cases, large parts of the LPC residue are intelligible, and the speaker of the original utterance can be identified. The work presented in chapter 3 determines the relative importance of amplitude and phase information of the residue for the synthesis of natural-sounding male and female speech.

Besides intelligibility and naturalness, preservation of the identity of a speaker may also be an important feature of a high quality speech-coding scheme. In chapter 4 we present a study in which we investigated this attribute of speech quality. It was felt that we needed a more sophisticated tool for the systematic manipulation of speaker identity. Various reasons made us decide to implement a glottal-excited (GE) LPC synthesizer which incorporates a more detailed model of the human voice source. Firstly, the study described in chapter 3 showed that both amplitude and phase information of the LPC residue can improve naturalness of synthetic speech. One way to code this information could be the use of a model of the human voice source. Secondly, pilot experiments on the identification of persons by their LPC resynthesized speech, showed that prosodic information alone is not sufficient to code speaker identity (Eggen & Vogten, 1990). This implies that the LPC model of human speech production should also be improved in order not to degrade speaker characteristics. In chapter 4 we describe a perception experiment in which we investigated the relative importance of coded vocal-tract and voice-source information for perceived speaker identity.

In the last chapter of this dissertation, chapter 5, we discuss the limitations of the present research. We also make some general remarks on the way we explored possible improvements to existing schemes for the analysis, manipulation and synthesis of speech.

Chapter 2

Intelligibility of synthetic speech in the presence of interfering speech[1]

Abstract

Traditional articulation tests are not always sensitive enough to discriminate between speech samples which are of high intelligibility. One can increase the sensitivity of such tests by presenting the test materials in noise. In this way, small differences in intelligibility can be magnified into large differences in articulation scores. We used both a more conventional articulation test and a monosyllabic adaptive speech interference test (MASIT) to evaluate the intelligibility of nine different speech-coding techniques. We found different patterns of responses for the articulation test and MASIT. These differences can be explained by the fact that different speech-coding schemes code different acoustic-phonetic properties of the speech signal. Some of these properties are more liable to masking by interfering noise than others. Our results show that, in the case of synthetic speech, differences in intelligibility are not always magnified by adding interfering noise; they may even disappear.

[1]This chapter is a slightly modified version of a previously published article: Eggen, J.H. (**1989**). "Intelligibility of synthetic speech in the presence of interfering speech", Speech Communication **8**, 319–327.

2.1 Introduction

THIS chapter deals with the evaluation of speech intelligibility resulting from differences in speech-coding schemes. According to Nakatani & Dukes (1973): "*Quality* refers to a conglomeration of attributes which determines the suitability of a speech sample for communication". On the one hand, intelligibility is a necessary feature of high-quality speech. Therefore, intelligibility scores are among the quality measures most often used to quantify the performance of speech-coding schemes. On the other hand, intelligibility of the synthetic speech is just one factor which determines its quality. Other features of a speech sample, such as naturalness, also determine the ease with which speech communication takes place.

Segmental intelligibility can be measured with an articulation test (French & Steinberg, 1947). In such a test, a list of usually monosyllabic words is presented to the listener. The listener's task is to identify the word presented. The percentage of phonemes that are correctly identified is called the articulation score. The best-known examples of articulation tests are the Modified Rhyme Test (MRT) and the Diagnostic Rhyme Test (DRT) (House, Williams, Hecker & Kryter, 1965; Voiers, 1977). These tests present the subject with a closed set of rhyming words from which a selection is to be made. Segmental intelligibility, measured with monosyllabic words, does not predict performance with sentences or passages. It has been found that the contribution of semantic and syntactic information, only leads to higher levels of performance (Pisoni, Nusbaum & Greene, 1985).

Because of technological progress the quality of synthetic speech has improved so greatly, that traditional articulation tests fail to discriminate between various sorts of highly intelligible synthetic speech (Pisoni, 1982; Pratt, 1987; Mackie, Dermody & Katsch, 1987). Recently, Mackie *et al.* (1987) studied the assessment of evaluation measures for processed speech. Their results confirmed Pisoni's findings that tests, which measure how fast subjects can decode the acoustic-phonetic information of the speech signal, can be used to discriminate between highly intelligible speech samples (Pisoni *et al.*, 1985). They also found that intelligibility tests can discriminate between highly intelligible processed speech, only if the speech material is presented to

the subjects under more difficult listening conditions.

In the areas of telecommunications and audiology it has been known for a long time that small differences in intelligibility can be magnified into large differences in articulation scores by adding noise to the test speech. Several kinds of noise have been used: white noise (House *et al.*, 1965), noise with a spectrum equal to the long-term average spectrum of speech (Plomp & Mimpen, 1979), and interfering speech (Nakatani & Dukes, 1973; Kalikow & Stevens, 1977). It should be noted that most of this research on the perception of speech in noise concerned natural, i.e. unprocessed, speech.

Pisoni & Koen (1982), Pratt (1987), and Vogten (1980) found that in the presence of noise, synthetic speech is less intelligible than natural speech. Pratt (1987) found that the effect of masking noise is not uniform with respect to different acoustic features, and that this effect is different for different speech synthesis systems. Pisoni *et al.* (1985) argued that, as compared to natural speech, synthetic speech is a phonetically impoverished signal. Therefore, decoding the acoustic-phonetic structure of synthetic speech requires more cognitive effort and capacity than decoding natural speech.

In this chapter we present an evaluation study, in which we measured the segmental intelligibility of synthetic speech. We used both an articulation test and a speech interference test. The main purpose of this study was to see whether small differences in intelligibility, due to different speech-coding schemes, could be magnified by adding noise to the synthetic speech.

We adopted the speech interference test developed by Nakatani & Dukes (1973). The main difference between the speech interference test and a more traditional articulation test is that, in the former case, the target words are presented in a background of interfering speech. We chose speech as a masker, because the sound of competing voices is frequently a source of interference in everyday listening situations. Besides, it has been shown that this type of noise interferes with speech intelligibility more than random nonspeech noise (Carhart, Johnson & Goodman, 1975). It has also been shown that attributes of consonants are more uniformly masked by speech than by white noise (Voiers, 1969).

In the next sections we will describe the speech interference test in more detail and present some modifications we made to it. Then we will discuss the evaluation study which consisted of two parts. In the first part a conventional articulation test was performed, while the speech interference thresholds were determined in the second part of the experiment. The results of the experiments are presented and discussed at the end of the chapter.

2.2 Monosyllabic Adaptive Speech Interference Test (MASIT)

The speech interference test was developed by Nakatani & Dukes (1973). Speech fragments are presented to the listener in the presence of interfering speech. The identification score as a function of the signal-to-noise (S/N) ratio of the speech and the interfering speech follows a psychometric curve. The S/N ratio, where 50% of the stimuli are correctly identified, is called the speech interference threshold. The Q measure of a degraded speech sample is defined as the difference between the threshold of the degraded speech and the threshold of the reference speech. We will now discuss some modifications we made to the speech interference test.

Nakatani and Dukes determined the speech interference threshold using the method of constant stimuli. This method samples the complete psychometric curve, which is time-consuming and not very economical. One also has to make parametric assumptions about the psychometric function. In order to avoid these problems, we chose to determine the speech interference threshold by means of a simple up-and-down adaptive procedure (Levitt & Rabiner, 1967; Bode & Carhart, 1973). The intensity of a test sentence is decreased stepwise with a fixed step in dBs if the preceding test sentence is correctly identified. An incorrect response causes a stepwise increment of the intensity of the next test sentence. In this way most of the speech stimuli are presented at a S/N level near the 50%-threshold. There is no need to make parametric assumptions about the psychometric curve. The only restriction is that intelligibility should increase monotonically with S/N ratio (Levitt & Rabiner, 1967; Levitt, 1971).

Nakatani and Dukes used nonsense sentences as speech materials. These sentences show the following structure : "The (adjective) (noun) (past tense verb) the (noun)". For example: " The blue tire held the king". The listener has to identify the four target words between brackets. In our case we wanted to determine the speech interference threshold completely automatically. To achieve this with the Nakatani and Dukes sentences subjects have to respond by typing the four target words of the test sentence on a computer terminal. This "whole response" mode is a time-consuming and difficult task. Therefore, the nonsense sentences were replaced by short, semantically neutral, carrier phrases each containing one monosyllabic CVC word receiving main stress. Both the carrier sentence and the test word were presented with interfering speech. The main purpose of the carrier phrase was to facilitate (on syntactic grounds) the detection of the target CVC word in the continuous presence of interfering speech.

2.3 Evaluation study

MASIT was used to evaluate the intelligibility of nine different speech-coding techniques. In order to compare the performance of MASIT, the same speech materials were also evaluated with an articulation test. The articulation test was run before MASIT.

2.3.1 Speech types

Nine different speech types were used in the experiments. The speech types are described in table 2.I. They can be categorised roughly into four groups with different bit rates.

The first group contained speech with a low bit rate of 4 kbit/s. The *MEA*-speech type was a software simulation of the hardware synthesiser MEA8000 (Philips Export BV, 1983). The *RFC*-speech type featured reflection coefficients which were quantised using a quantisation algorithm described by Markel & Gray, Jr. (1974). The third speech type in this group (*TDC*) was generated by the analysis-synthesis technique called temporal decomposition (Atal, 1983; Van Dijk-Kappers & Marcus, 1989).

The second group comprised three 12 kbit/s speech types. They were all synthesised on the basis of a linear predictive analysis with ten coefficients. Two of them used different formant-picking methods to estimate formants on the basis of the LPC-coefficients. The third speech type in this group (LPC_{10}) was synthesised using the LPC-coefficients directly. The $LPC_{10}F_1$-speech type used a formant-coding method developed by Willems (1986). The $LPC_{10}F_2$-speech type used a formant analysis described by Vogten (1983). The pitch and the voiced/unvoiced parameters were estimated using a pitch-detection algorithm developed by Hermes (1988).

The third group contained two speech types with a bit rate of 20 kbit/s. The LPC_{18}-speech type was generated using a linear predictive

Table 2.I: Speech-coding types used in the evaluation study

Speech type	Description	Bit rate (kbit/s)
1. *TDC*	Temporal Decomposition	
2. *MEA*	LPC, software simulation MEA8000-chip	4
3. *RFC*	LPC, 10 reflection coefficients	
4. LPC_{10}	LPC, 10 coefficients, no formants	
5. $LPC_{10}F_1$	LPC, 10 coefficients, formant-coding method 1	12
6. $LPC_{10}F_2$	LPC, 10 coefficients, formant-coding method 2	
7. *MPE*	Multi Pulse Excitation, 12 pulses, no formants	
8. LPC_{18}	LPC, 18 coefficients, no formants	20
9. *PCM*	Pulse Code Modulation	120

coding scheme (Atal & Hanauer, 1971) with eighteen linear predictive coefficients per 10-ms frame. Pitch and voiced/unvoiced parameters were estimated as in the 12 kbit/s group. The second 20 kbit/s speech type was generated by means of a multipulse excitation (MPE) coding scheme (Atal & Remde, 1982) using twelve pulses and ten linear predictive coefficients per 10-ms frame.

We used 120 kbit/s pulse code modulated speech (PCM) as reference speech (12 bit; 10kHz). This speech type served as an input signal for analysing and synthesising the other speech types.

2.3.2 Speech materials

All speech materials were spoken by the same male Dutch talker. The test sentences consisted of a neutral carrier phrase with a CVC test word receiving main stress. For example : "Nu volgt het woord BAP" ("The next word is BAP"). There were five different carrier phrases having a similar syntactic structure. The carrier phrases were not used in the articulation test.

Pisoni *et al.* (1985) showed that, as speech becomes less intelligible, listeners rely more heavily on response-set constraints to aid performance. They showed that an intelligibility test with an open-response format is much more sensitive in discriminating between synthetic speech samples than a test using a closed-response format. We decided to include both nonsense and meaningful CVCs in the test sets. Excluding the meaningful CVC words from the stimulus sets would have provided the subjects with an additional cue and would have restricted the open-response set.

The articulation score and the speech interference threshold were based on lists consisting of fifty CVC words. The lists were not phonetically balanced, because we wanted all phonemes to receive equal weight in the evaluation of a particular type of coded speech. The number of times a particular phoneme occurred in a CVC list was approximately the same for all lists. The order of the phonemes complied with the phonotactic rules of Dutch (Cohen, Ebeling, Fokkema & Van Holk, 1961).

Segmental intelligibility may depend on segmental context. There-

fore, one would like to include all possible combinations of consonants and vowels in the CVC word lists. In practice, this means that the number of CVCs in a list would increase considerably. Such long stimulus lists cannot be divided into smaller ones, because the speech interference threshold is expressed as a single number, which reflects the overall intelligibility of the test list. Therefore, segmental context was not addressed in this study.

Nine word lists were generated. Each of the nine word lists was processed with one of the nine speech-coding algorithms mentioned in the previous section. The speech levels of the nine stimulus lists were equalised, using the equivalent-peak-level method (EPL) of Brady (1968). From pilot experiments we concluded that the results for different word lists can be compared directly. However, direct comparison of intelligibility scores is not the main purpose of this study. Rather, we want to compare the patterns of responses as determined with an articulation test and a speech interference test. In particular, we want to see whether small differences in the intelligibility of coded speech can be magnified by adding extraneous speech.

2.3.3 Interfering speech

The carrier phrases, which were spoken by the same male Dutch talker who produced the test words, also served as a basis for generating the interference speech. By using the same speaker for the test words and the interference speech, the long-term spectral properties of the speech and noise were made as similar as possible. The carrier phrases were concatenated in random order. This resulted in a long speech fragment. Five such long fragments were added together and reversed in time.

Before addition, the carrier phrases were processed with the appropriate speech-coding scheme. In this way, both the test speech and the interfering speech were processed with the same speech-coding algorithm. As a consequence, there were nine different versions of interfering speech corresponding to the nine speech-coding schemes used in the evaluation study. The speech levels of these nine were equalised using the EPL-method. By making the speech masker and the test stimulus perceptually more similar, the sensitivity of the speech interference test increases (Nakatani & Dukes, 1973).

2.3.4 Subjects

Eight listeners participated in the experiments. They all had Dutch as their native language and no one reported hearing or reading deficiencies. All listeners had been previously exposed to synthetic speech and had participated in the pilot experiments preceding the present study. Therefore they were all thoroughly acquainted with the type of speech stimuli and the measuring procedure at the time the experiment started.

2.3.5 Procedure

The experiment consisted of two parts. In the first part the articulation tests were run. The speech interference thresholds were determined in the second part of the experiment. In both parts the speech types were played in real-time from the computer. In the case of MASIT, the interference speech was played back from a tape recorder. After passing through an attenuator, the interfering speech was mixed with the test sentences and fed into an amplifier. The listener was seated in a quiet room and could hear the speech stimuli through headphones. The test sentences were presented diotically (both ears received the same acoustical information) at 70 dB SL. The level of the interfering speech varied depending on the listener's response to the preceding stimulus.

The subject was sitting behind a computer terminal. An experimental run was started by the subject. The next test sentence was announced by a message on the terminal screen. The subject responded by typing the word he/she heard. There was no limit to the response time needed by the subject.

2.4 Results

First, we will describe the results obtained with the articulation test. The Q measures determined with MASIT are presented in the second part of this section.

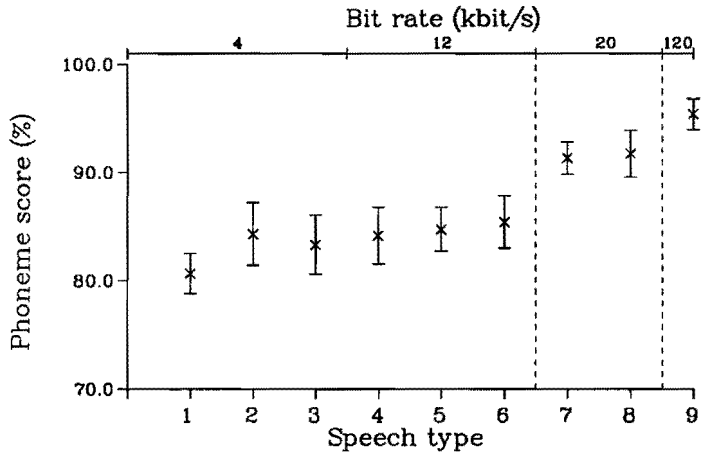

Figure 2.1: Results articulation test: mean percentage of phonemes correctly identified, plotted for the various speech types. The averages of 8 subjects are shown. The vertical bars represent the 95% confidence intervals. The parameter bit-rate is indicated at the top of the figure. The sub-division of the speech-coding algorithms into different groups is indicated by the dashed lines. (Speech types: 1=TDC, 2=MEA, 3=RFC, 4=LPC_{10}, 5=$LPC_{10}F_1$, 6=$LPC_{10}F_2$, 7=MPE, 8=LPC_{18}, 9=PCM).

2.4.1 Articulation scores

The mean articulation scores as measured for the nine different speech-coding techniques are shown in figure 2.1. The mean percentage of phonemes correctly identified is plotted for the various speech types. A description of the speech types is given in table 2.I. Each data point represents the mean articulation score, averaged over eight subjects. The articulation score for any individual subject is based on a list of 150 phonemes, i.e. fifty CVC words. The vertical bars show the widths of the 95% confidence intervals.

A one-way analysis of variance was performed on the arcsin-transformed articulation scores (Studebaker, 1985). The analysis showed significant differences between the nine speech-coding techniques ($F(7,71)=31.56$; $p<0.001$). *Post hoc* comparisons were carried out on the performance data using a Student-Newman-Keuls multiple range test with a 0.05 level of significance. According to this test, the nine different coded speech types can be divided into three different subsets, as indicated by the dashed lines in figure 2.1.

2.4.2 Q measures

Figure 2.2 shows the mean Q measures in dB for the various speech types. The Q measure is defined as the difference between the speech interference thresholds for the test-speech type and the reference-speech type. We used the PCM-speech type as the reference speech type. Therefore the Q measure of the PCM-speech type is zero by definition. Separate Q measures were computed for each subject. The mean Q measures are averaged over eight subjects. The 95% confidence intervals are shown as vertical bars.

A one-way analysis of variance was performed on the mean Q measures. The analysis showed significant differences between the nine speech-coding techniques ($F(7,71)=28.56$; $p<0.001$). *Post hoc* comparisons were carried out on the performance data using a Student-Newman-Keuls multiple range test with a 0.05 level of significance. According to this test, the nine speech types can be categorized into four different subsets, as indicated by the dashed lines in figure 2.2.

2.5 Discussion

According to the physical parameter of bit rate, the nine different speech-coding techniques can be divided into four categories with bit rates of 4 kbit/s, 12 kbit/s, 20 kbit/s and 120 kbit/s, respectively. Quality differences between these groups are clearly audible. One expects a good measure of speech quality to discriminate at least between these four groups.

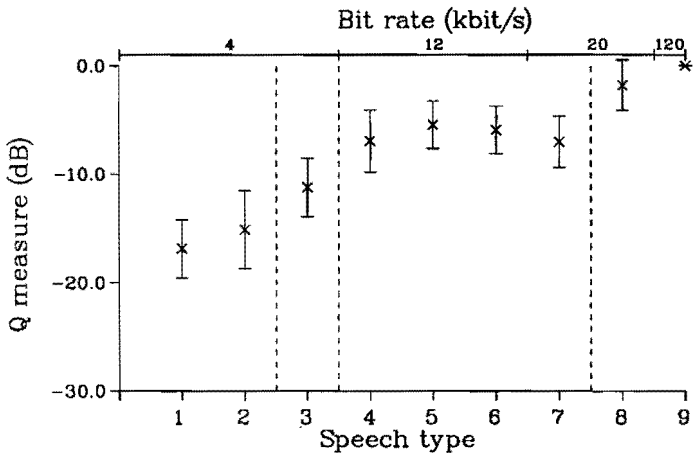

Figure 2.2: Results MASIT: Q measure in dB as determined for the various speech types. The averages of 8 subjects are shown. The vertical bars represent the 95% confidence intervals. The parameter bit-rate is indicated at the top of the figure. The sub-division of the speech-coding algorithms into different groups is indicated by the dashed lines. (Speech types: 1=TDC, 2=MEA, 3=RFC, 4=LPC_{10}, 5=$LPC_{10}F_1$, 6=$LPC_{10}F_2$, 7=MPE, 8=LPC_{18}, 9=PCM).

The articulation test discriminates between the last three groups. However, the mean articulation scores of the 4 kbit/s speech types do not differ significantly from those of the 12 kbit/s speech types. These results clearly demonstrate the need for a more sensitive test. Therefore, the speech reception task was made more difficult by adding extraneous speech and, as a consequence, we expected small differences in intelligibility to be magnified into large differences in articulation scores.

The results show that MASIT can discriminate between the 4 kbit/s speech and the 12 kbit/s speech types, while the articulation test fails to do so. But neither MASIT nor the articulation test isolates the

LPC_{10}, $LPC_{10}F_1$ or the $LPC_{10}F_2$-speech type as a separate 12 kbit/s class.

MASIT differentiates within the 4 kbit/s speech types. The RFC-speech type shows a higher Q measure than the other two 4 kbit/s speech types. This can be explained by the property of reflection coefficients to be more robust with respect to quantisation than LPC-coefficients or formants and bandwidths (Markel & Gray, Jr., 1974). It should also be noted that in the case of the MEA-speech type only four formants are coded. The relatively low Q measure for the TDC-speech type is probably due to the fact that the temporal-decomposition parameters were quantised in a rather ad-hoc manner. In this study, no systematic research was carried out on efficient coding of TDC parameters.

The Q measure of the MPE-speech type was unexpectedly low. As said before, when listening to speech samples belonging to the 12 kbit/s and 20 kbit/s speech types, respectively, clear quality differences can be heard. This subjective impression is confirmed by the articulation test: the mean articulation scores for the 12 kbit/s speech types differ significantly from those of the 20 kbit/s speech types. However, in the case of MASIT the Q measure of the MPE-speech type has dropped to that of the 12 kbit/s speech types. The same effect can be seen with the PCM-speech type and the LPC_{18}-speech type: the articulation test shows that both speech types differ significantly in intelligibility, but this difference has disappeared in the case of MASIT.

The different response patterns for the 12 kbit/s, 20 kbit/s and the 120 kbit/s speech types, as measured with the articulation test and MASIT need an explanation. In figure 2.3 we see some hypothetical speech spectra. The upper row shows part of the spectrum of the original speech (PCM). Roughly, this part of the spectrum can be described by two resonances (peaks) and one anti-resonance (dip). The two resonances show some acoustic fine structure. In this example we will assume that the exact location of the peaks and dip, as well as the fine structure, represent acoustic-phonetic information which is important with respect to the intelligibility of the speech. The second row shows the corresponding part of the LPC_{10} spectrum. Although LPC_{10} codes the location of the resonances correctly, it fails to describe the fine structure and the anti-resonance. More details of the speech spec-

trum can be captured by increasing the number of parameters used in the coding algorithm. In the MPE algorithm, extra pulses are added to the excitation function of the linear predictive filter. (This can be viewed as being equivalent to the introduction of zeros in the transfer function of a system with monopulse excitation (Atal & Remde, 1982). The effect of adding pulses can be seen in the third row of figure 2.3: MPE has coded the anti-resonance. In the case of LPC_{18}, more LPC-coefficients are used to code resonances. As can be seen in the bottom row of figure 2.3, LPC_{18} has coded the fine structure of the spectral peaks. Both MPE and LPC_{18} increase intelligibility by resolving spectral fine structure. However, they code different aspects of the speech spectrum.

For the articulation test, the interaction between noise and speech-coding scheme is indicated in the left column of figure 2.3. As interfering speech is not present in the case of the articulation test, the dashed lines in the left column of figure 2.3 indicate the noise level of the D/A-convertor (S/N ratio of 72 dB). The right column shows the interaction between noise and speech-coding algorithm for MASIT. The dashed lines in the right column of figure 2.3 indicate the level of the interfering speech. Better resolution of the acoustic fine structure, as achieved by increasing the number of coding-parameters, results in a higher articulation score for both MPE and LPC_{18}. With MASIT, however, the coding of the anti-resonance is masked by the interfering noise, whereas the fine structure of the peaks is not. Therefore, only the articulation score of LPC_{18} increases. The intelligibility of MPE stays at the LPC_{10} level. The anti-resonance present in the PCM spectrum is also masked by the interfering noise. This decreases intelligibility. The intelligibility of PCM can be compared with LPC_{18}.

According to Nakatani & Dukes (1973) processed speech can be viewed as being degraded in comparison with natural speech. Because the degraded speech is supposed to be on the same continuum as natural speech, small differences in intelligibility can be magnified into large differences in articulation scores by adding interfering speech. Maybe this is true of telephone-like speech (filtered speech), but, apparently, it is generally not true of synthetic speech. Indeed, from our findings we conclude that, in the case of synthetic speech, differences in intelligibility are not always magnified by adding interfering noise. Differences

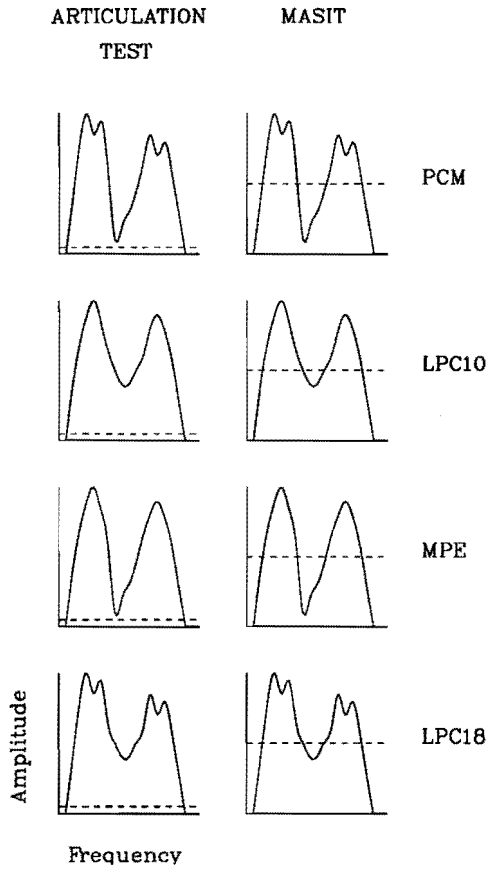

Figure 2.3: Coding and masking of acoustic-phonetic properties of the speech signal. Each panel shows a spectral representation (amplitude versus frequency) of a hypothetical speech sample which is coded with four different speech-coding algorithms (the four rows). The results for the articulation test and MASIT are shown in the left and the right column, respectively. In the case of the articulation test, the dashed lines represent the noise level of the D/A convertor (S/N ratio of 72 dB). The dashed lines represent the level of the interfering speech in the case of MASIT.

in intelligibility may even disappear. Our results show that the inter-

action between the masker and the test speech strongly varies with the speech-coding scheme used. Different speech-coding algorithms may code different acoustic-phonetic properties of the speech signal. Some of these acoustic-phonetic properties (anti-resonances) are more easily masked by the interfering speech than others (resonances). Our results are in agreement with the findings of Pisoni *et al.* (1985) and Pratt (1987). According to Pisoni *et al.* (1985), "synthetic speech may be thought of as perceptually impoverished relative to natural speech. Synthetic speech is fundamentally different from natural speech in both degree and kind because many of the important critical acoustic cues are either poorly represented or not represented at all" (Pisoni *et al.*, 1985, page 1670). Pratt (1987) also performed intelligibility tests in the presence of noise. He found that the effect of masking noise is not uniform with respect to different acoustic features, and that this effect is different for different speech synthesis systems. Although Pratt does not mention it explicitly, his results also show that differences in intelligibility may disappear. In some cases, the rank-order of the tested text-to-speech systems is even reversed due to the presence of interfering noise.

Pisoni *et al.* (1985) point out that perceptual tests of synthetic speech should address the environmental conditions in which synthetic speech will be used. Speech is a source of interference which is frequently present in everyday listening situations. Therefore, MASIT can suggest which synthesis technique to choose in practical voice-response applications. For example, if one has to implement a 20 kbit/s speech-coding scheme in a speech communication aid for the speech impaired, MASIT indicates that the LPC_{18} speech-coding algorithm is better suited for the job than the MPE algorithm. Indeed, LPC_{18} is more robust with respect to the everyday presence of interfering speech.

Our results show that MASIT is a valuable tool for evaluating synthetic speech. MASIT can be used to assess the performance of speech-coding algorithms in noisy environments. We have shown that the masking effect of the interfering speech on the test speech differs for various speech-coding schemes. This is not surprising since different speech-coding algorithms code different acoustic-phonetic properties of the speech signal. Some of these acoustic-phonetic properties are more liable to masking by interfering noise than others. As a consequence,

the evaluation of nine speech-coding schemes reveals different patterns of responses for the articulation test and MASIT. Selection of the appropriate test should depend on the environmental conditions in which the synthetic speech will be used.

Chapter 3

On the role of amplitude and phase in the synthesis of male and female voices [1]

Abstract

A pitch-synchronous segmentation was used to obtain a short-time Fourier representation of the LPC residue. After selected amplitude and phase manipulations of voiced segments, a residue was reconstructed, which was used to drive the LPC synthesis filter. Twenty utterances (10 male, 10 female) were investigated under two amplitude (original/flat) and two phase conditions (original/zero), yielding four versions for each utterance. The quality of these versions was judged by 12 subjects in a paired-comparison experiment. Original amplitude information was consistently preferred over original phase information. For female voices, there were significant quality differences between any of the four versions. However, for male voices the original amplitude information alone proved to be sufficient to make the synthetic speech almost indistinguishable from natural speech.

[1] This chapter is co-authored by W.D.E Verhelst, and will be submitted to J.Acoust.Soc.Am.

3.1 Introduction

LINEAR predictive coding (LPC) is still one of the most powerful techniques for the analysis, manipulation and resynthesis of speech (Atal & Hanauer, 1971; Markel & Gray, Jr., 1976). It is applied in many different areas of speech processing. For instance, as a research tool, LPC is widely used to study the perceptual effects of systematic manipulations of various aspects of speech, such as pitch, duration, loudness and timbre. Also, in many text-to-speech systems, rule-generated artificial parameter contours are used to drive an LPC synthesizer.

Despite its popularity, the LPC method also has its limitations. For instance, as far as speech quality is concerned, it is well known that LPC speech lacks naturalness. Although intelligibility can be very high (see, for instance, chapter 2), LPC speech can be easily recognized by its synthetic ("buzzy") sound quality. There have been various attempts to improve the naturalness of LPC speech (Rosenberg, 1971; Sambur, Rosenberg, Rabiner & McGonegal, 1978; Atal & David, 1979; Atal & Remde, 1982). Recently, the increasing number of practical applications of synthetic speech, has only urged the need for more natural-sounding speech (Pinto, Childers & Lalwani, 1989; Childers & Wu, 1990).

To find ways for improving the naturalness of LPC speech, it is reasonable to look in more detail at the relation between the LPC model and the model of human speech production (Fant, 1960; Rabiner & Schafer, 1978). The LPC method approximates a speech sample as a linear combination of past speech samples. The predictor coefficients, i.e. the weighting coefficients which are used in the linear combination, define an all-pole filter. When excited by either quasi-periodic pulses (during voiced speech) or random noise (during unvoiced speech), this linear system was shown to model the speech production process adequately (Rabiner & Schafer, 1978). Since the excitation has a flat spectrum, the LPC filter models the spectral characteristics of the glottis, vocal tract, and lip radiation. In practice, the LPC method cannot perfectly estimate these composite spectral effects. For example, correct modeling of nasals and fricative sounds requires at least a system function which features both poles and zeros. Furthermore, information of the relative phase of the spectral components of the speech signal is

not coded by the LPC method. These modeling errors are reflected in the so-called linear prediction error signal.

This error signal (or residue) is defined as the difference between the actual speech samples and the predicted samples. Since the original speech wave can be exactly reproduced by exciting the LPC filter with the residue, the residue contains all information necessary for synthesizing natural-sounding speech by linear prediction. One can get a qualitative impression of this information by listening to the residue. In this way, one can hear that residues of LPC filters up to orders of fifteen, can contain sufficient information to be intelligible. Also, by listening to the residue one can often recognize the speaker who produced the original utterance. In fact, it has been shown by Feustel, Logan & Velius (1988) that the residue indeed provides information which can be used by listeners to distinguish between speakers.

Based on these observations, and on the previously mentioned fact that the residue reflects the LPC modeling errors, we decided to investigate the perceptual relevance of the information contained in the LPC residue in a more formal way. For this purpose, we used a residual-excited LPC (RELP) analysis-resynthesis scheme (Un & Magill, 1975). This scheme was also used by Atal & David (1979), and Gautherot, Mason & Corney (1989). They showed that systematic manipulations of the LPC residue can vary the quality of synthetic speech almost continuously from that of LPC speech to that of natural speech. By presenting these various synthetic speech qualities to listeners and asking them about their preferences we are able to study the perceptual correlates of natural-sounding speech.

Both Atal & David (1979), and Gautherot *et al.* (1989), applied a Fourier series expansion on the LPC residue. In this way, it is possible to systematically modify the amplitude and phase characteristics of the LPC excitation. Atal & David restricted the manipulations to voiced speech segments, whereas Gautherot *et al.* did not distinguish between voiced and unvoiced speech. Atal & David found that correct amplitude information is of greater importance for natural-sounding speech than correct phase information. Unlike Atal & David , Gautherot *et al.* found that most of the important information within the residue is contained in the phase spectrum. The different outcomes of these studies could be caused by the different ways in which the residue was

segmented. Atal & David used a pitch-synchronous segmentation, in contrast with Gautherot *et al.* who used an asynchronous segmentation. In the first case, phase manipulations do not distort pitch information, whereas in the second case they do.

It was a *first objective* of this study to develop a tool to manipulate the quality of voiced speech sounds. The examples discussed in the previous paragraph demonstrate that, in designing such a tool, one should be careful about choosing a particular segmentation strategy, because the way the residue is segmented greatly influences the experimental results and their interpretation. In this chapter we introduce a new RELP-based analysis-resynthesis scheme which features a segmentation strategy that is based on a speech production model. As will be explained in section 3.2, this decision enables us to interpret the experimental findings in terms of a speech synthesis filter which has physical meaning. This approach has the additional advantage that the residue information, which would turn out to be necessary or sufficient for the synthesis of natural-sounding speech, directly indicates the flaws in LPC systems. This implies that the experimental results can be used to inspire design criteria for improving current LPC systems.

A *second objective* of this study was to perform a perception experiment to determine the relative importance of amplitude and phase information for the synthesis of natural-sounding speech. In particular, we wanted to examine for a larger group of speakers, including both males and females, whether the perceptual importance of phase information was really as small as suggested by the experiments of Atal & David (1979).

If we look at the literature on monaural phase perception we see that most of it supports the finding of Von Helmholtz (1877) that the quality (timbre) of a complex sound depends solely on the number and relative strength of its components and not on their phase differences. However, it also has been shown that with respect to the quality of a sound the ear certainly is not "phase deaf". For instance, Mathes & Miller (1947) showed that phase relations of harmonic components within a single critical band may affect sound perception. Plomp & Steeneken (1969) also demonstrated that the effect on timbre of varying the phase spectrum of a complex tone can be heard, although they found this effect to be small compared with the effect of varying the am-

plitude spectrum. Carlson, Granström & Klatt (1979) reported large perceptual effects for synthetic vowels if the phase relations of the spectral components of the voice source were manipulated. Schroeder & Strube (1986) also demonstrated the perceptual importance of phase relations between the harmonics of a complex sound by showing that it was possible to produce vowel-like timbre sensations with stimuli consisting of many equal-amplitude harmonics (flat-spectrum stimuli) if the phase angles of individual harmonics were properly manipulated. Recently, Patterson (1987) investigated the ability to discriminate between changes in the phase spectra of wideband periodic sounds. He showed that for complex tones with repetition rates up to 400 Hz phase changes are perceptible. From this finding he concludes that the quality of most men's voices and many women's voices depends on phase relations. In summary, these recent studies all suggest that the quality of speech sounds can be manipulated by changing the phase relations between the spectral components of the speech signal. This evidence, and the fact that Atal & David (1979) found small but audible effects for phase-manipulated speech signals, formed the main motivation to conduct the perception experiment described in this chapter.

This chapter is organized as follows. In section 3.2 we discuss the signal processing aspects involved in the development of the analysis-resynthesis system. The perception experiment and the results are presented in section 3.3. Section 3.4 contains a general discussion. Section 3.5 concludes this chapter by summarizing the main findings.

3.2 Speech processing

This section deals with the signal processing aspects involved in the development of the analysis-resynthesis system. The first part of the section discusses the objectives and general concept of the system. Next, a detailed description of the analysis and synthesis strategy is presented. A qualitative evaluation of the system concludes this section.

3.2.1 Objectives and general concept

As mentioned in the introduction, a first objective of this work was to develop a system for the study of perceptual correlates of spectral characteristics of voiced speech. The general strategy was to provide means for the computation of the spectral characteristics of a natural speech utterance, for the selective manipulation of these characteristics, and for the construction of a correspondingly modified signal. Listening experiments would then be conducted on original and modified signals to reveal perceptual correlates of the manipulated spectral characteristics.

Because speech is essentially a time-varying signal, the processing has to be based on a carefully chosen time-frequency representation. The choice of the particular time-frequency representation is important because it obviously determines the relation between the (manipulated) spectra and the actual speech signal. Therefore, it is also one of the factors that determine what is actually to be learned from the experiments.

In order for our results to be directly applicable in speech (re-) synthesis, as well as to provide more fundamental insights in speech perception, we decided to base our time-frequency representation on the simplified speech production model shown in figure3.1.I (Rabiner & Schafer, 1978).

In this model, the simplification mainly consists of speech being considered as the output of a slowly time-varying linear system. Specifically, the production process is approximated by the following synthesis equations:

$$i(n) = \sum_k \delta(n - p(k)) \tag{3.1}$$

$$s(n) = \sum_k f_{opt}(n, p(k)) \tag{3.2}$$

$$f_{opt}(n, p(k)) = \sum_m g(m, p(k)) \cdot v(n - m, p(k)) \tag{3.3}$$

where $i(n)$ is the input to a linear time-varying system and consists of an impulse train with impulses at time instants $p(k)$ correspond-

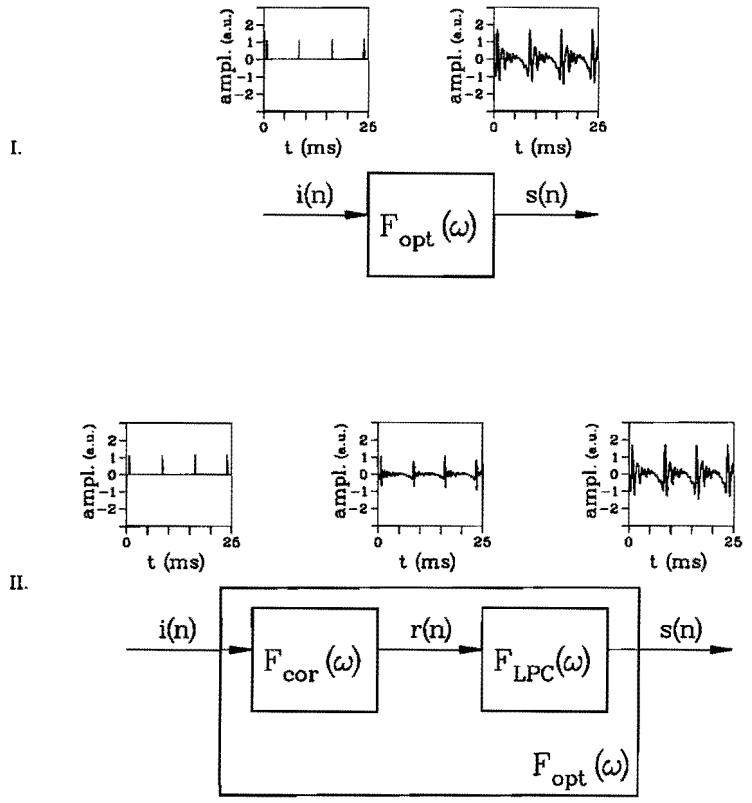

Figure 3.1: Speech production model: *I.* simplified speech production model, *II.* cascade of the correction filter with the LPC filter. The amplitude of the waveforms is expressed in arbitrary units (a.u.).

ing to the successive moments of glottal closure (indexed by k), and $f_{opt}(n, p(k))$ is the impulse response of the system at time instant $p(k)$, and represents the combined contribution of the radiated glottal wave $g(n, p(k))$ and the vocal-tract response $v(n, p(k))$.

As will be shown, a good approximation to the idealised ("optimal") time-varying filter $f_{opt}(n, p(k))$ can be obtained from the anal-

ysis of natural voiced speech. This makes it interesting to choose the time-varying transfer function $F_{opt}(\omega, p(k))$ as a time-frequency representation for voiced speech, because it allows a direct evaluation of the perceptual effects of manipulated parameters of the speech production model. At the same time, it would enable us to explore which aspects of the source-filter model are minimally needed in order to synthesize natural-sounding speech.

In the next subsection, we will first derive an analysis procedure for $f_{opt}(n, p(k))$ under the assumption that voiced speech is actually produced according to the simplified production model of figure 3.1.I.

3.2.2 Speech analysis

Having decided that the speech processing system should be based on a time-frequency representation which corresponds to the synthesis model of figure 3.1.I, the critical problem is to derive an appropriate analysis procedure.

From observation of natural speech signals, it is clear that the time-varying impulse response $f_{opt}(n, p(k))$ typically lasts longer than a pitch period, and some form of explicit deconvolution will have to be applied. As discussed in the introduction, LPC is a popular and well studied parametric deconvolution method. Furthermore, the LPC modeling error (residue) can be easily computed and, to some extent, explained in terms of speech production. We therefore decided to use an LPC-based deconvolution approximation. Consequently, the optimal time-varying synthesis filter can be rewritten as the cascade connection of an, as yet unknown, correction filter $f_{cor}(n, p(k))$ with the LPC synthesis filter, as illustrated in figure 3.1.II.

Thus, the problem is now reduced to finding a good approximation for the correction filter $f_{cor}(n, p(k))$ which shows the LPC residue $r(n)$ at its output in response to the input impulse train $i(n)$. From the model of figure 3.1.II, and equation 3.1, we have

$$r(n) = \sum_k f_{cor}(n, p(k)). \tag{3.4}$$

On the other hand, from the linearity of the proposed synthesis model (equation 3.2), and using equation 3.3, it follows that

$$f_{cor}(n, p(k)) = g(n, p(k)) * v(n, p(k)) * lpc^{-1}(n), \tag{3.5}$$

where $lpc^{-1}(n)$ represents the impulse response of the inverse LPC filter, and $*$ denotes convolution in n.

Because the residue results from a short-term deconvolution of voiced speech by means of LPC, the spectral characteristics of the correction filter (equation 3.5) are expected to be globally flat. If we further assume that the group delay characteristics of equation 3.5 indicate no important time spread, the effective duration of $f_{cor}(n, p(k))$ can expected to be short. The input to the correction filter being an impulse train with a spacing equal to the fundamental period of speech, it then follows that $f_{cor}(n, p(k))$ could be reasonably well approximated by pitch-synchronously segmenting the LPC residue, using a time-varying window function $w(n, p(k))$:

$$f_{cor}(n, p(k)) \simeq r(n) \cdot w(n, p(k)) \tag{3.6}$$

In equation 3.6, the location and length of the time-varying window function $w(n, p(k))$ should be chosen such that the corresponding synthesis model of figure 3.1.II finds a natural interpretation in terms of speech production. The exact shape of the window is less important in this respect, and was chosen to be trapezoidal.

As illustrated in figure 3.2.I, our choice for $w(n, p(k))$ is such that $f_{cor}(n, p(k))$ corresponds to that segment of the residue that lies *around* the moment of main excitation (this moment usually coincides with the instant of glottal closure[2]). In the synthesis model of figure 3.1.II, the portion of $f_{cor}(n, p(k))$ which precedes this moment can then be interpreted to mainly contribute a correction due to the radiated glottal wave: it is well known that LPC can not accurately model this speech component, and that the residue before glottal closure is largely determined by it (a fact which is used in numerous approaches to glottal

[2]The method for determining the location of the moments of glottal closure will be discussed in more detail in chapter 4.

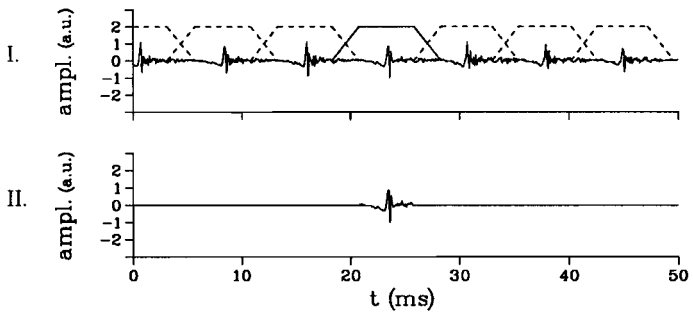

Figure 3.2: *I.* Segmentation of the LPC residue, *II.* windowed segment.

wave recovery; see, for instance, chapter 4). Because LPC will not perfectly model the vocal-tract impulse response either, the response of the LPC filter during glottal closure would still be different from that of the true vocal tract. The part of $f_{cor}(n, p(k))$ which lies in the closed glottis interval can then be considered to contribute adjustments which compensate for this different response (see figure 3.2.II).

In this subsection we derived an analysis procedure for $f_{opt}(n, p(k))$. This analysis procedure will be used in the next subsection to build a system for the analysis, manipulation, and resynthesis of voiced speech.

3.2.3 System overview

Figure 3.3 gives an overview of the signal processing steps involved in analysing, modifying, and reconstructing a segment of voiced speech[3]. In a first step, a sequence of LPC filters is obtained from a standard LPC analysis procedure, and used to compute the LPC residue. Subsequently, a pitch-synchronous segmentation of the residue is performed using the windowing function $w(n, p(k))$, and each windowed segment

[3]A description of specific implementation details follows in the experimental section of this chapter.

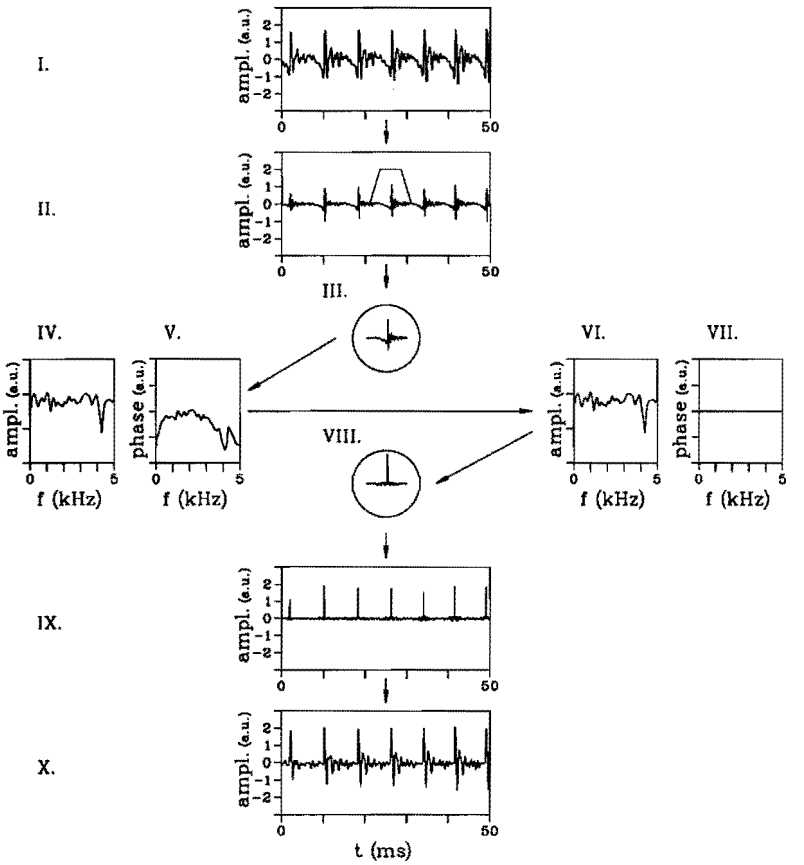

Figure 3.3: Signal processing overview. $I.$ original speech signal $s(n)$, $II.$ LPC residue signal $r(n)$, $III.$ filter impulse response $f_{cor}(n, p(k))$, $IV.$ amplitude spectrum of $F_{cor}(\omega, p(k))$, $V.$ phase spectrum of $F_{cor}(\omega, p(k))$, $VI.$ modified amplitude spectrum of $F^*_{cor}(\omega, p(k))$, $VII.$ modified phase spectrum of $F^*_{cor}(\omega, p(k))$, $VIII.$ modified filter impulse response $f^*_{cor}(n, p(k))$, $IX.$ modified LPC residue signal $r^*(n)$, $X.$ modified speech signal $s^*(n)$, (for explanation see text).

is used as an approximation to successive impulse responses of the time-varying correction filter $f_{cor}(n, p(k))$. After transforming $f_{cor}(n, p(k))$

to the frequency domain, the amplitude and phase of $F_{cor}(\omega, p(k))$ can be modified in selected frequency regions. The modified filter $F^*_{cor}(\omega, p(k))$ is then transformed back to the time domain to obtain a sequence of manipulated filter impulse responses $f^*_{cor}(n, p(k))$, which are used in the synthesis scheme of figure 3.1.II to construct a modified residue

$$r^*(n) = \sum_k f^*_{cor}(n, p(k)). \tag{3.7}$$

This modified residue is then used to drive the sequence of LPC synthesis filters, yielding the correspondingly modified speech $s^*(n)$.

In the next subsection we will present a qualitative evaluation of the performance of the system.

3.2.4 System performance evaluation

As illustrated in figure 3.2.I, segmentation windows are such that

$$r(n) = \sum_k f_{cor}(n, p(k)). \tag{3.8}$$

Therefore, if amplitude and phase responses of the correction filters are left unchanged, the output of the system will be identical to the original speech waveform. From previous arguments, it also follows that the cascade of $f_{cor}(n, p(k))$ with the LPC filters forms a more realistic approximation to deconvolution than could for example be obtained by applying a short-time Fourier transform directly to the speech signal itself.

Nevertheless, in designing the system, it was assumed that speech is actually produced according to the model of figure 3.1.I, and that as far as the LPC residue is concerned, a pitch-synchronous segmentation constitutes a sufficiently close approximation to deconvolution. In order to evaluate how effective our analysis results are in terms of speech production, informal listening experiments were performed. In these experiments, an impulse train with a modified pitch period was used as an input to the synthesis scheme of figure 3.1.I. From the fact that the resulting speech quality was very high, it was concluded that for auditory perception the proposed deconvolution approximation is sufficiently accurate.

More evidence that our procedure provides a good approximation to $f_{opt}(n, p(k))$ is emerging in the recent literature on waveform manipulation (Charpentier & Moulines, 1989; Lent, 1989). As one example, Hunt, Zwierzynski & Carr (1989) deleted samples, or inserted zero samples, in the residue at a point 80 % into the glottal cycle, and reported no degradation in the speech was perceived over a range of pitch changes. In a recent study by Hamon, Moulines & Charpentier (1989) it was shown that a segmentation similar to the one in figure 3.2 allows for high quality prosodic modifications of speech, even when applied directly on the speech wave

While the above indicates that our analysis provides a satisfactory deconvolution, it is not possible to evaluate directly how close $f_{opt}(n, p(k))$ approximates the output of the physical speech production system for a single glottal cycle. Nevertheless, by choosing an overlap-add technique for constructing the modified residue $r^*(n)$, we implicitly assumed that $f_{opt}(n, p(k))$ has physical meaning. Otherwise, the sequence $F_{cor}(\omega, p(k))$ should have been interpreted to only represent a short-time Fourier transform of the LPC residue. In that case, the modified residue should be constructed from $F^*_{cor}(\omega, p(k))$ by short-time Fourier synthesis techniques, such as the least-squares approach by Griffin & Lim (1984). Informal listening showed that, with this alternative method, perceptual effects of amplitude and phase manipulations are of a same nature, but much less pronounced than with our original approach.

This qualitative evaluation of the performance of the system concludes this section on speech processing. In the next section, a perception experiment is described in which we use the analysis-resynthesis system to investigate the perceptual importance of amplitude and phase information for the synthesis of natural-sounding speech.

3.3 Experiment

In this section we describe a perception experiment in which listeners had to judge the quality of speech utterances for which the amplitude and phase characteristics were manipulated. As said earlier, one source of inspiration for our experiment was the work of Atal & David

(1979) who also investigated the perceptual importance of amplitude and phase for the synthesis of natural-sounding speech. Before we describe the differences between the present study and the work of Atal & David in more detail, we have to be more specific about what we mean by "natural-sounding" speech.

In the literature on speech quality, the term "naturalness" is often attributed to different aspects of speech quality. For instance, one can link naturalness to speech attributes like intonation or duration, but, at the same time, one can also speak of a natural (human-like) voice quality. In practice, naturalness is often used as a catch-all term to address speech quality attributes other than intelligibility or speaker identity (Klatt, 1987). As in this chapter we did not want to investigate speech quality attributes like intelligibility (see chapter 2) or speaker identity (see chapter 4), the conditions of the present experiment were chosen in such a way that the level of intelligibility was equal to that of the original speech and the same for all speech stimuli, and the recognizability of the speakers was of no interest to the listeners. Under these conditions, and under the assumption that loudness has no influence on the quality of the speech signal to be evaluated, it is was argued by Rothauser, Urbanek & Pachl (1968) that over-all speech quality can be assessed by asking listeners to state their preference for one of two speech signals to be compared. We therefore decided to investigate the role of amplitude and phase in speech synthesis by means of a preference test.

This approach is different from the way Atal & David (1979) performed their experiments. In the case of Atal & David, four listeners had to sort the different versions of one utterance according to their naturalness along a one-dimensional scale. We did not use such a sorting task because we thought that the quality differences between the stimuli was not always fairly apparent. A simple ranking would have been too time consuming because small quality differences require in practice many pairwise comparisons of tentative neighbours before a reasonable ordering becomes established. Nor is it necessarily possible to achieve a wholly satisfactory ranking (David, 1963). On the one hand, as the differences between the stimuli in our experiment were small, we thought that it would be desirable to make the comparisons between two of them as free as possible from any extraneous influences

caused by the presence of others. We therefore decided to use the method of paired comparisons where the stimuli to be compared are presented to the listeners in pairs (David, 1963). On the other hand, we thought the existence of differences clear enough to be classified on some finer scale. Therefore, listeners had to indicate which one of the two stimuli of a pair they preferred on a 5-point scale. An important difference between our choice of paradigm and Atal & David's, is that their method depends on the one-dimensionality of naturalness, which is not a-priori obvious and maybe even false. Our method of paired-comparisons does not depend on dimensionality (at least not until we start processing the data).

Another difference between the experiments of Atal & David (1979) and the present study is the number of different speakers which produced the utterances. Atal & David determined the relative importance of amplitude and phase for the synthesis of natural-sounding speech on the basis of three sentence-length utterances. Unfortunately, they did not mention whether the three sentences were produced by one speaker, or by different speakers. In this study we want to find out whether the findings of Atal & David hold for a larger group of speakers, including both males and females. In particular, we want to see if the perceptual importance of phase information is really as small as suggested by the experiments of Atal & David.

This section starts with a description of the speech stimuli. After the subjects and the experimental procedure have been described, the results are presented.

3.3.1 Stimuli

Speech material

Five short Dutch sentences of the Plomp & Mimpen (1979) corpus were read by four male and four female speakers. The five sentences are shown in table 3.I. A one-letter code is used to refer to a particular sentence.

A subset of twenty utterances was used in the experiment. The selected speaker-sentence combinations are indicated in table 3.II. The

remaining twenty utterances (the empty slots in table 3.II) were not used in the experiment, because the recordings of these utterances contained unwanted by-sounds.

The amplitude and phase information of the twenty utterances of table 3.II was systematically manipulated with the analysis-resynthesis system described in section 3.2.

Stimulus generation

We first list the implementation details of the speech processing system which was introduced in the previous section. Next, the specific amplitude and phase manipulations are described.

The sequence of pitch markers $p(k)$ contains the sample indices at, or immediately after, the zero crossings which mark the beginning of successive pitch periods (figure 3.4).

This sequence was estimated using an automatic procedure (Eggen, 1989b, see also chapter 4), and corrected manually. A manual correction was found to be important in order to avoid distortion noises in the output speech. For the same reason, a manual correction of voiced/unvoiced decisions was sometimes necessary (mostly at transitions between voiced and unvoiced sounds).

The sequence of LPC filters was computed using a 12^{th} order

Table 3.I: Dutch sentences used in the experiment. The second column shows the one-letter code of each sentence.

Sentence	Code
De bomen waren helemaal kaal.	B
Hij probeerde het nog een keer.	K
Rijden onder invloed is strafbaar.	R
Toch lijkt me dat een goed voorstel.	V
De zak zat vol oude rommel.	Z

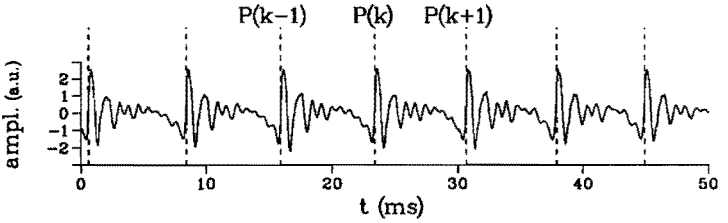

Figure 3.4: Pitch-synchronous marking of the speech waveform. The markers are indicated at the top of the figure.

pitch-asynchronous autocorrelation type of LPC analysis (LeRoux &

Table 3.II: Speech utterances used in the experiment. The one-letter sentence code is explained in table 3.I. The utterances printed in *italics* form the largest subset of balanced male/female utterances (the construction of this balanced male/female set will be discussed in subsection 3.3.4).

Speaker		Sentence				
		B	K	R	V	Z
Females:	F1	F1B	F1K		*F1V*	
	F2			*F2R*		*F2Z*
	F3		*F3K*			*F3Z*
	F4	*F4B*		*F4R*	F4V	
Males:	M1	M1B		*M1R*		*M1Z*
	M2			M2R	*M2V*	
	M3	*M3B*		*M3R*		
	M4		*M4K*		M4V	*M4Z*

Gueguen, 1977), performed on the 10–kHz sampled speech signal, using a segmentation with 10–ms frame repetition, and a 25–ms hamming window. We found it preferable not to use a pre-emphasis, in view of the spectral manipulations which are to be performed between the analysis and resynthesis.

After inverse filtering the speech wave with the sequence of LPC filters, the resulting residue $r(n)$ was segmented to approximate the time-varying correction filter impulse response:

$$f_{cor}(n, p(k)) := r(n) \cdot w(n, p(k)), \tag{3.9}$$

with

$$w(n, p(k)) = \begin{cases} 0 & & & n \leq n_b(k) \\ \frac{n-n_b(k)}{n_e(k)-n_b(k)} & n_b(k) < & n & < n_e(k) \\ 1 & n_e(k) \leq & n & \leq n_b(k+1) \\ 1 - \frac{n-n_b(k+1)}{n_e(k+1)-n_b(k+1)} & n_b(k+1) < & n & < n_e(k+1) \\ 0 & n_e(k+1) \leq & n & , \end{cases} \tag{3.10}$$

and

$$\begin{aligned} n_b(k) &= p(k-1) + \lfloor (p(k) - p(k-1) - 1)/3 \rfloor \\ n_e(k) &= p(k-1) + 2 \cdot \lfloor (p(k) - p(k-1) - 1)/3 \rfloor + 1. \end{aligned} \tag{3.11}$$

In order to manipulate the characteristics of the correction filter, samples of the transfer function $F_{cor}(\omega, p(k))$ were computed using a 1024 point FFT algorithm (zero padding has to be applied in order to reduce time alias distortions in the modified filter response, and prefolding was used to align $p(k)$ with the local time origin). After selected spectral modifications, the inverse FFT was computed. From inspection of the resulting sequences $\widetilde{f^*_{cor}}(n, p(k))$, it was verified that no serious time aliasing had occurred, and $f^*_{cor}(n, p(k))$ was obtained by unfolding the inverse FFT results:

$$\begin{aligned} f^*_{cor}(n, p(k)) &= \widetilde{f^*_{cor}}(n, p(k)) && n : 0..512 \\ &= \widetilde{f^*_{cor}}(n + 1024, p(k)) && n : -511..-1 \end{aligned} \tag{3.12}$$

Finally, the modified residue $r^*(n)$ was constructed by overlap-adding the individual $f^*_{cor}(n, p(k))$ according to equation 3.7, and used to drive the sequence of LPC synthesis filters.

The amplitude and phase manipulations of $F_{cor}(\omega, p(k))$ are summarized in table 3.III. The amplitude and phase spectra were not manipulated for the original-amplitude, and original-phase conditions, respectively. For the flat-amplitude condition, the amplitudes of the spectral components of $F_{cor}(\omega, p(k))$ were set to a constant value equal to the RMS value of the amplitudes of the Fourier spectrum. For the zero-phase condition, the phases of all spectral components were set to zero. As there are two amplitude and two phase conditions, this means that there are four different versions for each utterance of table 3.II. In the remainder of this chapter we refer to the four different versions either by their numbers 1, 2, 3 and 4, or by their corresponding mnemonics $A_o\Phi_o$, $A_f\Phi_o$, $A_o\Phi_z$, and $A_f\Phi_z$ (A = amplitude, Φ = phase, $_o$ = original, $_z$ = zero, $_f$ = flat; see also table 3.III).

3.3.2 Subjects

Twelve male subjects participated in the listening experiments. All subjects were members of the Speech and Hearing group of the Institute for Perception Research. They all had been previously exposed to synthetic speech and no one reported hearing deficiencies.

Table 3.III: Amplitude (A) and phase (Φ) conditions of the four stimulus versions. The version numbers and their corresponding mnemonics are indicated.

Phase	Amplitude	
	Original (A_o)	*Flat* (A_f)
Original (Φ_o)	1 ($A_o\Phi_o$)	2 ($A_f\Phi_o$)
Zero (Φ_z)	3 ($A_o\Phi_z$)	4 ($A_f\Phi_z$)

3.3.3 Procedure

The total number of paired comparisons for four versions is six. These six different pairs were presented to all subjects in both orders. The presentation order of the twelve pairs corresponded to the optimal presentation order described by Phillips (1964). As all subjects performed every possible paired comparison, we have a "balanced paired-comparison experiment".

Twenty sets of twelve pairs were recorded on a digital audio recorder (Sony PCM-501 ES) in a random order. At the beginning of each set, the four different versions were presented to give the subject an impression of the quality differences to be expected for that particular utterance.

All subjects heard the same sets of stimuli in two sessions of half an hour each. Subjects were seated in a quiet room and listened to the stimuli through headphones (Pioneer, monitor 10) at a comfortable listening level. The number of trials per subject was 240 (20 utterances, 12 pairs), and the total number of trials 2880 (12 subjects).

Subjects had to state their amount of preference on a 5-point scale. For each ordered pair (i, j) the following statements could be made:

```
(-2) :  I prefer i to j strongly
(-1) :  I prefer i to j slightly
( 0) :  No preference
( 1) :  I prefer j to i slightly
( 2) :  I prefer j to i strongly
```

The subjects had three seconds to indicate one of the five alternatives before the next pair of stimuli was presented. At the end of the experiment, subjects were asked if they had been able to identify any of the four versions of the utterances, explicitly.

3.3.4 Results

In this subsection we present the experimental results. First we discuss the statistical model used to analyze the data. Next, the speech quality

judgments of the listeners are presented for the twenty utterances. After we have compared the results for the male and the female utterances, we conclude this subsection with some remarks on the experimental method.

Statistical analysis

We adopted a method described by Scheffé (1952) which was specifically developed for the analysis of paired-comparison experiments. In this method, an analysis of variance is applied on the preference scores which are expressed on an n-point scale. Scheffé uses the following analysis model:

$$x_{ijk} = (\alpha_i - \alpha_j) + \gamma_{ij} + \delta_{ij} + e_{ijk}. \tag{3.13}$$

When the stimulus pair is presented in the order (i, j), x_{ijk} represents the amount of preference for stimulus i over stimulus j of the k^{th} of r judges. The parameters α_i and α_j characterize an inherent quality of the i^{th} and j^{th} stimuli. If the *hypothesis of subtractivity* is not rejected, i.e. all γ_{ij} are zero, the average preference for i over j equals the difference $\alpha_i - \alpha_j$. This means that the α's can be compared on a one-dimensional scale. $2\delta_{ij}$ represents the difference due to the order of presentation. The error term e_{ijk} is the only random variable on the right-hand side of equation 3.13. In the terminology of the analysis of variance of a two-way layout, the α's are analogous to the main effects, and the γ_{ij} to the interactions.

The Scheffé analysis provides a "yardstick" Y_ϵ. With the "yardstick" we can make inferences about the significance of differences among the α's. For instance, α_i and α_j differ significantly at the ϵ level if and only if their difference is greater than Y_ϵ.

The Scheffé model uses the following underlying assumptions: all x_{ijk} are independent random variables, the x_{ijk} are normal, and for a fixed ordered pair (i, j) all r variables x_{ijk} have the same mean μ_{ij} and the same variance σ^2. As none of the subjects was able to identify any of the four versions of the utterances explicitly, the x_{ijk} are assumed to be independent random variables. We did not check the normality assumption because the analysis of variance for balanced experiments

has shown a great robustness for even wide departures from the normal distribution (Hays, 1988). Cochran's test was used to test for homogeneous variance (Cochran, 1947). It can be noted that the Scheffé method gives least-squares estimates of the model parameters α_i.

Next we present, for each of the twenty different utterances, how the inherent qualities α_i $(i = 1, 4)$ of the four different stimulus conditions are ordered on a one-dimensional quality scale.

Quality judgments

Table 3.IV shows the scale values α_i $(i = 1, 4)$ for the twenty utterances as obtained with the Scheffé analysis. The hypothesis of homogeneous variance has to be rejected[4] only for utterance M2V. This means that strictly speaking the Scheffé model does not hold for this utterance. The hypothesis of subtractivity has to be rejected for utterance M4V. This means that for this utterance the stronger preference of stimulus i compared to stimulus j is, statistically speaking, only true in the average sense when i and j are compared with the $m-2$ other stimuli as well as with each other. Five of the twenty utterances show significant order effects.

Scheffé suggests to declare the main effects significant at the ϵ level if and only if the largest and the smallest of the estimated main effects α_i $(i = 1, 4)$ differ by more than the "yardstick" Y_ϵ. As this is true for all utterances, the experiment shows an over-all difference in preference among the four versions for each utterance. We also applied F-tests to check for the significance of the main effects α_i $(i = 1, 4)$. The γ_{ij} and/or the δ_{ij} were pooled with the error term e_{ijk} if they were not significant. This means that we simplified the Scheffé model for these cases. F-tests showed significant main effects α_i $(i = 1, 4)$ for all utterances ($p < 0.002$, df 1,2 = 3,132).

Column 7 of table 3.IV shows the order of the α's, where the α's are indicated by their index i. From left to right the inherent quality of the particular α_i decreases. For instance, the order of the four signal conditions of utterance $F1B$ is α_1, α_3, α_2, α_4, which means that the

[4] If not specified explicitly, hypotheses in this chapter are tested at the 0.05 level of significance.

Table 3.IV: Estimated α's (columns 2,3,4,5) for the four signal conditions of each utterance (column 1). Column 7 shows the order of the α's (indicated by their index i), where any two α's not underlined by the same line may be considered distinguishably different at the 5% level, i.e. the α's differ by at least the "yardstick" $Y_{0.05}$ (column 6).

	$\alpha 1$	$\alpha 2$	$\alpha 3$	$\alpha 4$	$Y_{0.05}$	order α_i
F1B	0.76	-0.25	0.23	-0.74	0.28	1 3 2 4
F4B*	1.01	-0.52	0.29	-0.78	0.28	1 3 2 4
F1K	0.35	-0.11	0.25	-0.49	0.30	1 3 2 4
F3K*	0.75	-0.40	0.32	-0.68	0.25	1 3 2 4
F2R	0.97	-0.38	0.20	-0.79	0.31	1 3 2 4
F4R	0.74	-0.34	0.20	-0.59	0.30	1 3 2 4
F1V	0.65	-0.16	-0.09	-0.40	0.31	1 3 2 4
F4V	0.61	-0.23	0.26	-0.65	0.29	1 3 2 4
F2Z	0.56	-0.07	0.19	-0.68	0.29	1 3 2 4
F3Z*	0.56	-0.18	0.29	-0.68	0.28	1 3 2 4
M1B	0.69	-0.10	0.42	-1.00	0.29	1 3 2 4
M3B	0.44	-0.22	0.33	-0.55	0.28	1 3 2 4
M4K	0.35	0.20	0.19	-0.74	0.30	1 2 3 4
M1R	0.45	-0.05	0.30	-0.70	0.29	1 3 2 4
M2R	0.54	-0.23	0.54	-0.85	0.35	1 3 2 4
M3R	0.56	-0.10	0.27	-0.73	0.24	1 3 2 4
M2V*‡	0.66	-0.38	0.60	-0.89	0.33	1 3 2 4
M4V†	0.43	-0.15	0.44	-0.72	0.20	3 1 2 4
M1Z	0.21	-0.05	0.09	-0.25	0.31	1 3 2 4
M4Z*	0.27	-0.09	0.21	-0.39	0.27	1 3 2 4

* Significant order effect

† Rejection of the hypothesis of subtractivity

‡ Rejection of the hypothesis of homogeneous variances

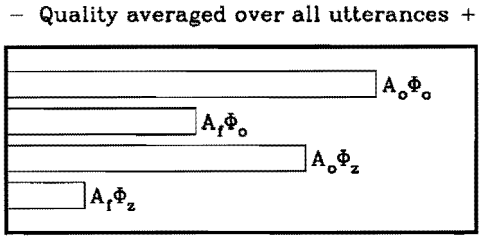

Figure 3.5: Quality scale of the pooled data.

inherent quality of stimulus condition 3 ($A_o\Phi_z$) is higher than the quality of stimulus condition 2 ($A_f\Phi_o$). If any two α_i are not underlined by the same line, they may be considered distinguishably different at the 0.05 level, i.e. the α_i differ by more than the "yardstick" $Y_{0.05}$.

If we rank the α_i for each utterance, we can determine the association among the twenty rankings by using the Kendall coefficient of concordance W (Siegel, 1956). The coefficient of concordance is significant at the 0.05 level of significance (W = 0.96, s = 1903.5). This may be interpreted as meaning that the pooled ordering, which corresponds to the α-order α_1, α_3, α_2, α_4, may serve as the best estimate of the "true" ranking of the four versions.

We can get an average scale by pooling the data for all utterances. Unfortunately, it is not possible to apply the Scheffé analysis on the pooled data, because the hypothesis of homogeneous variances has to be rejected. As an alternative, we can analyze the pooled data with the following linear model:

$$x_{ijklm} = (\alpha_i - \alpha_j) + \beta_{ijklm}. \tag{3.14}$$

In this equation, the indices l and m indicate the l^{th} sentence uttered by the m^{th} speaker. The four unknown α's can be determined as least-squares estimates of equation 3.14. We solve equation 3.14 by minimizing the RMS value of the error term β_{ijklm} under the constraint that $\sum_i \alpha_i = 0$. The resulting quality scale is shown in figure 3.5.

In order to check the results obtained by the Scheffé method, we

also applied Thurstone's one-dimensional scaling technique (Torgerson, 1958). In order to construct the Thurstone scales we had to apply a binary transformation to the raw data (David, 1963; Rietveld & Gussenhoven, 1985). The 5-point scale was contracted by ignoring the degree of preference and by assigning ties randomly to one of the two members of each tied pair. The goodness of fit of the Thurstone model to the transformed data was tested with a single over-all test given by Mosteller (1951). According to this test the model fits adequately for all utterances. For three scales the order of two α_i's on the Thurstone scale is reversed on the Scheffé scale. However, these scale values do not differ significantly on the Scheffé scales.

Male/Female differences

Unfortunately, the number of times a sentence or speaker occurs in the set of twenty utterances is not equal for all sentences and speakers. We therefore extracted two subsets, A and B respectively, according to the following criteria:

- subset A contains *female* speakers only, subset B consists of *male* speakers only,
- for each speaker of a subset, there is a corresponding speaker in the other subset who uttered exactly the same sentences,
- subset A and subset B contain a maximum number of utterances.

The utterances which belong to these "balanced" female and male subsets are printed in *italics* in table 3.II. If we pool the data for each subset and apply the Scheffé analysis on the pooled data we get the results shown in figure 3.6.I and figure 3.6.II. For both sets the main effects α_i differ significantly ($p < 0.001$, df $1,2 = 3,996$), the hypothesis of subtractivity is not rejected, and there are significant order effects. Although, according to Cochran's test, we had to reject the hypothesis of homogeneous variances for the male subset, we still think the results of the Scheffé analysis to be valid.

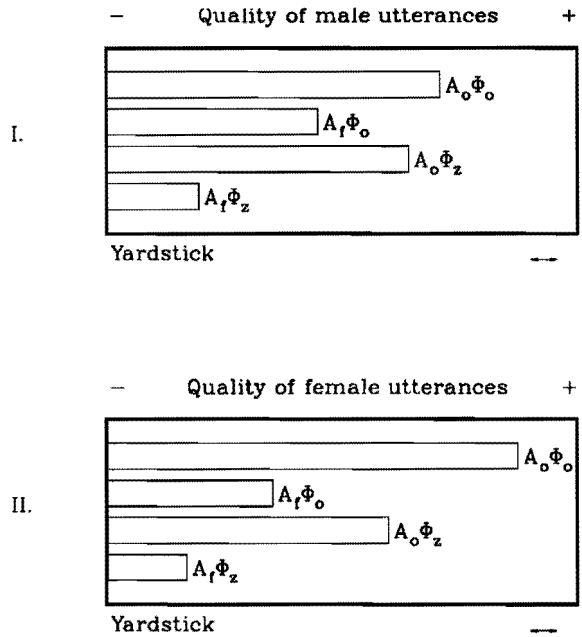

Figure 3.6: Male and female quality scales. *I*. Scheffé scale for male utterances, *II*. Scheffé scale for female utterances.

Remarks on the experimental method

At the end of this subsection on the results, we will make some remarks on the experimental method. In the original Scheffé method the judges' preferences are expressed on a 7 or 9-point scale. If the number of scale divisions is too small, the danger arises that too many scores are jammed up against the ends of the scale. This may cause a biasing effect in the α estimates as well as non-homogeneity of variances. Although we used a 5-point scale, inspection of the scales showed no serious jamming effects. The 5-point scale seems fine enough for subjects to indicate their preferences for the stimuli used in this study.

A detailed inspection of how each subject used the 5-point scale to state his preference for a particular member of the ordered pair (i, j),

averaged over utterances, showed that for some subjects the probability of a judgment of a slight preference one way or the other (-1 or 1) seemed to be greater than that of a judgment of no preference (0). In general, there was however no scarcity of zeros. In order to see whether the fact that some judges may have been declaring ties more readily than others, could have influenced the ordering of the α_i on the Scheffé scales, we randomly broke the ties in a binary transformation of the data. As the α order on the resulting Thurstone scales was the same as on the Scheffé scales, we can conclude that the seemingly subject-dependent assignment of ties can be ignored. We could not simply ignore all ties in the experimental results because they are essential in the estimation of the α_i; with the same numbers of clear preferences, two stimuli are more equal in quality the more ties there are between them.

We also examined how consistent listeners were in assigning their preference by comparing the scores for the pairs (i, j) and (j, i). The following cases were considered to be consistent: ($x_{ij} = 0$ AND $x_{ji} = 0$), ($x_{ij} < 0$ AND $x_{ji} > 0$), ($x_{ij} > 0$ AND $x_{ji} < 0$). The pairs x_{ij} and x_{ji} were ignored in the counting of consistent scores if only one of the pairs was zero. The cases ($x_{ij} > 0$ AND $x_{ji} > 0$) and ($x_{ij} < 0$ AND $x_{ji} < 0$) were considered to be inconsistent. If we examine the data in this way, the percentage of preferences which is stated consistently is lower than 75% for only one of the twelve listeners.

The Scheffé model assumes that for a fixed ordered pair (i, j) the score x_{ijk} may be thought of as the sum of two components, one representing the mean preference of the "average"-listeners, the other being the chance deviation from this mean. The Scheffé scales should therefore be interpreted to represent a quality ordering of the four versions according to the "average" listener. We can also perform the analysis with utterances as replication factor. In this way we can construct a Scheffé scale for each listener. The α ordering on these scales was the same for all twelve listeners: α_1, α_3, α_2, α_4.

As a last remark on the experimental method, we want to say something about the suitability of the Scheffé model for our data. Inspection of the raw data showed order effects for all pairs, that is to say, there was always a slight advantage for the second stimulus of a pair. The order effects were not equal for all pairs; order effects seemed more prominent with respect to pairs for which the quality difference be-

tween the stimuli was relatively small. From this we concluded that the Scheffé model seems to be an appropriate model to fit to our data.

In the next section we will discuss the experimental results of the perception experiment in more detail.

3.4 Discussion

In this section we start with a discussion of some topics related to the speech processing applied in this study. Next, we will discuss the experimental results in more detail.

3.4.1 Speech processing

The analysis-resynthesis system we have developed and used in this study makes it possible to interpret the results directly in terms of a speech synthesis filter. As mentioned earlier in this chapter, the results of our experiment can be seen as direct indicators of the flaws in LPC systems and can be used to inspire design criteria for improving them. In this view, the experimental findings indicate that it is important that the amplitude spectrum of the synthesis filter should be modeled very accurately. Correct modeling of the phase characteristics of the synthesis filter will also improve speech quality. The results of the experiments however do not say anything about the way we should code this information.

The multi-pulse model for LPC excitation is one way to achieve this goal (Atal & Remde, 1982). The amplitude and phase characteristics of the multi-pulse excitation patterns contain information which significantly improves the quality of LPC speech. In a study of Caspers & Atal (1987), which aimed at creating a proper understanding of the role of multi-pulse excitation in the synthesis of natural-sounding voiced speech, it was reported that fixed multi-pulse patterns introduced only small degradations in synthetic speech. The fact that the multi-pulse patterns are fixed for all voiced speech parts of an utterance, means that locally there can be big spectral distortions which are not audible. Based on these findings, Caspers & Atal (1987) suggest that it is the combination of amplitude spectrum and phase spectrum which

is important for the synthesis of natural-sounding speech, and not the amplitude or phase spectra by themselves.

A model of the human voice source can also be used to code amplitude and phase information contained in the LPC residue. A source model combines amplitude and phase characteristics of the LPC residue in a way that the resulting voice-source waveform clearly reflects phonatory events like glottal opening and closing. Our results indicate that speech quality would improve for such a coding scheme. In the next chapter we will introduce an LPC synthesis scheme which incorporates a model of the human voice source. This scheme will be used to investigate whether the coded voice-source information is used by listeners to identify speakers by their voice.

From the discussion above we conclude that, as a next step, it would be interesting to work out our research question concerning the relative importance of amplitude and phase information for the synthesis of high quality natural-sounding speech in more detail. For instance, we can use our analysis-resynthesis system to study the effect of other possible spectral manipulations. As an example, we mention two different manipulations which seem to be of particular interest for the development of coding schemes which generate natural-sounding synthetic speech. As a first possibility we can keep original amplitude in the frequency region below 1 kHz, and original phase for frequency components above 1 kHz. In this case, amplitude and phase are kept original in those regions where they are supposed to be perceptually the most relevant. Recently, Gupta & Atal (1991) reported that this strategy yields reconstructed speech with no perceptually significant distortion. Secondly, it would be interesting to maintain the original amplitude and phase in the low frequency region. This could give an impression of the quality we could get with a "glottal-excited" LPC analysis-resynthesis scheme. By keeping only the original amplitude in this region we could say something about the necessity of correct phase behaviour of glottal-excited vocoders.

For practical reasons, we used an asynchronous LPC analysis as deconvolution method. From a signal processing standpoint, a careful synchronous LPC analysis could have provided a more accurate vocal tract approximation. In that way, it might have been possible to obtain an even shorter effective impulse response from the formula

$f_{cor}(n) = g(n) * v(n) * lpc^{-1}(n)$, which in turn could result in a more accurate deconvolution approximation (in particular for the phase characteristics when the pitch period becomes relatively short). It should be noted that female spectra have wider-spaced harmonics, such that, even with a synchronous LPC analysis, formants and bandwidths will be less accurately modelled than for male voices. This could be one of the reasons for the relative importance of phase information (or equivalently the lack of effectiveness of amplitude information alone) which we found for female voices: if phase deconvolution becomes worse for higher pitches, the original phase could become more important. If this is true, it could be the case that with a perfect deconvolution we would have found that phase is not at all perceptually relevant for any type of voice.

At the end of this discussion on speech processing aspects, we want to indicate how amplitude and phase manipulations for unvoiced speech sounds can be incorporated in our analysis-resynthesis system. Only voiced speech sounds were manipulated in this study because these segments contribute the most to the over-all quality of speech under the assumption that intelligible is already very high. If it is also required to modify unvoiced segments, the model of figure 3.1.II can obviously not be used as such. However, for unvoiced speech, short segments of the residue can be modeled as stationary noise segments with, again, a globally flat amplitude spectrum. Therefore, one could study unvoiced amplitude and phase effects by short-time Fourier analysis and synthesis techniques. For the segmentation, we propose to use the same strategy as we used for voiced speech, but with a regular 10 ms spacing between successive windows. The manipulation of amplitude spectra can be done in the same way as for the voiced segments; for phase manipulation, however, the modified phase should be a random function in order for the modified segments to maintain their random character (a 100 Hz pitch percept would be introduced if we would use a zero-phase function). For the reconstruction of a modified residue, the least squares short-time Fourier synthesis method of Griffin & Lim (1984) can be used.

In the next subsection we come back to the results of the perception experiment we performed, and discuss the implications of these results for speech synthesis.

3.4.2 Synthesis of high-quality natural-sounding speech

The experimental results show that version $A_o\Phi_o$ has the highest overall quality whereas version $A_f\Phi_z$ has the lowest. This is to be expected as $A_o\Phi_o$ corresponds to the original speech and $A_f\Phi_z$ is equivalent to a pitch-synchronous mono-pulse (i.e. flat amplitude spectrum, zero phase spectrum) excited LPC analysis-resynthesis. Version $A_o\Phi_z$ is preferred over version $A_f\Phi_o$ and both have inherent qualities which are lower than $A_o\Phi_o$ but higher than $A_f\Phi_z$. This means that amplitude information of the LPC residue is more important with respect to speech quality than phase information. This finding corresponds to the general view of the relative importance of amplitude and phase for the perception of complex sounds, as discussed in the introduction of this chapter.

Our results are also in line with the findings of Atal & David (1979). They found that errors in spectral amplitude are a major source of unnatural sound quality in LPC speech, but zero-phase can also cause small audible degradations in synthetic speech. The results of the present study confirm these findings. In addition, our results show that amplitude is preferred over phase information for *various* speakers, both males and females. At the same time clear differences between male and female utterances can be seen (figure 3.6). It should be noted that we cannot compare the absolute α values for the male and female scales directly because the Scheffé analysis uses the assumption that the sum of the α_i is zero. However, if we compare differences, we can draw the following conclusions. Figure 3.6 shows that the distance between $A_o\Phi_o$ and $A_o\Phi_z$ is much smaller on the scale for male speech than it is on the scale for female speech. This means that for the male speakers original amplitude alone provides near naturalness, whereas for the female speakers correct phases are also needed. Moreover, the distance between $A_o\Phi_o$ and $A_f\Phi_z$ is larger on the female scale than it is on the male scale. This finding confirms the general experience that LPC (featuring an excitation function which has a flat spectral envelope and a zero phase spectrum) causes a greater degradation of speech quality for females than for males.

At the end of this section we want to say something about the argument often heard that, in the ideal case, the envelope of the magnitude of the LPC residue spectrum is flat, suggesting that phase would carry

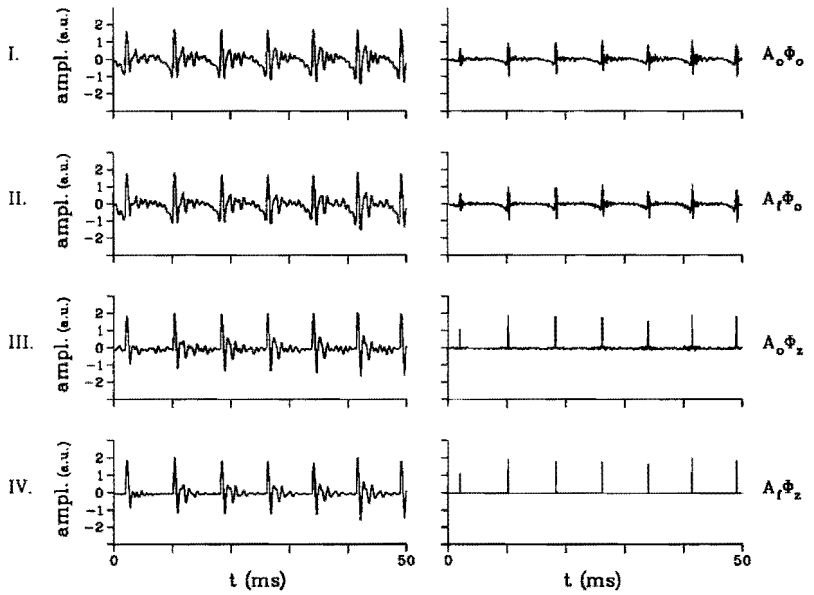

Figure 3.7: Four stimulus versions of a male speech signal and its corresponding LPC residue. The left column shows the speech waveforms, $s(n)$, of the four stimulus versions.The right column shows the corresponding residues $r(n)$. The four stimulus conditions $A_o\Phi_o$, $A_f\Phi_o$, $A_o\Phi_z$, and $A_f\Phi_z$ are shown from top to bottom, respectively.

the principal information. This suggestion is strengthened if we look at the speech waveforms and their corresponding LPC residues (figure 3.7). From figure 3.7 we can see that the waveform of version $A_f\Phi_o$ resembles the original waveform $A_o\Phi_o$ much more than the waveform of version $A_o\Phi_z$. Nevertheless, the results of the perception experiment show that $A_o\Phi_z$ is perceptually almost indistinguishable from $A_o\Phi_o$ (see figure 3.6). This clearly demonstrates that it can be dangerous to evaluate speech quality on the basis of visual appearance of speech waveforms alone.

3.5 Conclusions

We developed an analysis-resynthesis system which can be used as a tool to study the acoustic correlates of perceived aspects of high quality natural-sounding speech. This system was used to systematically manipulate amplitude and phase characteristics of twenty utterances. The manipulated speech stimuli were used in an experiment in which listeners had to state their preference for the different stimuli. The newly developed analysis-resynthesis system makes it possible to interpret experimental findings in terms of a speech synthesis filter. We find that amplitude information contained in the LPC residue is more important with respect to the synthesis of high quality natural-sounding speech than phase information. Both for male and female speech, incorrect phase information causes audible degradations. If we compare female speech with male speech we see that female speech is degraded more by linear predictive coding than male speech. For male voices the original amplitude information alone proved to be sufficient to make the synthetic speech almost indistinguishable from natural speech.

Chapter 4

Contributions of voice-source and vocal-tract characteristics to speaker identity

Abstract

In this chapter the perceptual importance of voice-source and vocal-tract characteristics for speaker identity is investigated. By means of a perception experiment we determined whether a voice-source model codes information which is used by listeners to identify speakers by their voices. Four male speakers produced 4 different versions of a sustained vowel /a/ (100 Hz, normal voice). For each of the 16 vowels, voice-source and vocal-tract functions were derived from the speech wave by means of a closed-phase covariance LPC analysis. The voice-source waveforms were modeled by the Liljencrants-Fant model. From the 16 natural /a/'s we selected one for every speaker. For these /a/'s, all possible combinations of voice-source and vocal-tract functions were synthesized, yielding 16 hybrid stimuli. These 16 hybrids were mixed with the resynthesized 12 remaining /a/'s. Four subjects were extensively trained to recognize the 4 speakers by their natural /a/'s. After each subject reached a score of 87.5 % correct or higher, the synthetic stimuli were presented to the subjects over headphones. The experimental results clearly show that listeners use both voice-source and vocal-tract information to perform the identification task. After the experiments were finished, subjects were asked which criteria they had used to identify the speakers. Their answers were remarkably similar, and could be related to physical aspects of the stimuli. Besides the expected importance of the vocal tract filter, the spectral balance between high and low-frequency components of the voice-source spectrum and the flutter of F_0 proved to be important perceptual cues for speaker identity.

4.1 Introduction

As text-to-speech (TTS) systems presently can produce reasonably intelligible speech, research efforts are more and more directed towards the improvement of other aspects of speech quality (Klatt, 1987). In general, it is felt that improved models of the complex processes of human speech production and perception, as well as better rule systems for the letter-to-sound conversion are needed to give synthetic speech a more natural sounding quality. However, "lack of naturalness" is not the only shortcoming of current TTS systems. Practical applications also have demonstrated the need to synthesize different voice qualities (Carlson, Galyas, Granström, Pettersson & Zachrisson, 1980; Waterham, 1989). Qualities like breathy, creaky, and pressed voice not only characterize a particular speaker, they also belong to the prosodic repertory of speakers used to add meaning to utterances (Laver, 1980). An even more dramatic example of the limited power of current TTS systems is the fact that it is still very hard to synthesize a convincing female voice (Karlsson, 1989; Karlsson, 1991; Klatt & Klatt, 1990). In general, one would like a TTS system to have some sort of provision for the conversion of a more or less standard voice to any other natural sounding voice (Childers, Wu, Hicks & Yegnanarayana, 1989).

A number of solutions are presented in the literature for improving the quality of synthetic speech. Most of these solutions deal with resynthesis problems, as for instance in the case of transmission or storage of coded speech (Atal & Remde, 1982; Kroon & Deprettere, 1988). As a consequence, many of these techniques do not allow for the independent manipulation of basic speech parameters like pitch and timbre. This makes them less suited for TTS applications. There are three basic types of synthesis used to generate the output of TTS systems (Pols, 1992). In articulatory-based synthesis the control parameters are specified in terms of the voice and articulation mechanism itself. In allophone-based synthesis the allophones are described in parametric form and the transitions are controlled by rules. In diphone-based systems, the output is generated by concatenating brief prestored speech fragments. These fragments are coded by a set of parameters which control a speech synthesizer. In order to generate speech with an acceptable quality, the parameters of the fragments need to be adjusted.

For instance, the original pitch information of the fragments is replaced by a rule-generated pitch contour. This example clearly shows that it is essential for TTS systems that the applied coding scheme provides means for the independent manipulation of basic speech parameters.

One specific way to improve the quality of synthetic speech is to take into account not only articulatory characteristics of speech production, but also aspects of vocal fold vibration. This can be realized by incorporating a model for the human voice source into the speech coding algorithm. This approach has the advantage that speech quality can be improved without losing the possibility of independent manipulation of speech parameters like pitch and timbre. It has been shown that such a coding scheme leads to high quality resynthesized speech (Sambur, Rosenberg, Rabiner & McGonegal, 1978; Hedelin, 1986; Fujisaki & Ljungqvist, 1986). Also, for TTS applications it is believed that the incorporation of more advanced voice-source models will improve the naturalness of synthetic speech, and will provide means for manipulating voice quality (Carlson, Fant, Gobl, Granström, Karlsson & Lin, 1989; Klatt & Klatt, 1990).

There have been various attempts to estimate the perceptual relevance of voice-source models with respect to different attributes of speech, in particular naturalness and voice quality. In general, however, there is still a lack of knowledge on the perceptual importance of current voice-source models. Moreover, according to Klatt, the present state of speech research is characterized by: "the absence of a satisfactory perceptual theory to account for listeners' behaviour in terms of observable spectral or waveform details" (Klatt, 1987, page 781). We still do not have satisfactory answers to questions like: how naturally and how variously can different voice qualities be synthesized by a voice-source model, or what is the perceptual importance of a voice-source model with respect to the perceived characteristics or identity of a speaker?

This chapter describes an attempt to bring such questions closer to an answer. We conducted an experiment in which listeners had to identify speakers from their voices. Four male speakers produced four different versions of a sustained vowel /a/. For each of these sixteen vowels, voice-source and vocal-tract functions were derived from the speech wave. For a subset of four /a/'s, i.e. one /a/ for every speaker,

all possible combinations of voice-source and vocal-tract functions were synthesized. Listeners who were extensively trained to recognize the four speakers by their natural voice had to identify the sixteen hybrid stimuli. From the confusions made by the listeners we can learn to what extent voice-source and vocal-tract information are used to perform the voice identification task. However, it should be stressed that the results of the perception experiment strongly depend on the way in which we separate the speech signal into voice-source and vocal-tract information. In fact, the contributions of voice-source and vocal-tract to speaker identity can only be interpreted in terms of this source-tract separation.

This remainder of this chapter is organized as follows. Section 4.2 discusses the speech processing techniques used to generate the synthetic speech stimuli. In section 4.3, the perception experiment is described and the results are presented. After a general discussion in section 4.4, the main findings of this study are summarized in section 4.5.

4.2 Speech processing

The first part of this section discusses the modeling of voiced speech sounds. Next, one particular way of modeling voiced speech, the glottal-excited (GE) speech synthesizer, is described in more detail. This synthesizer was used to generate the synthetic speech stimuli for the perception experiment which will be described in section 4.3. The last part of this section describes the speech-analysis techniques used to derive the parameters of the GE speech synthesizer from the speech waveform.

4.2.1 Modeling voiced speech sounds

For voiced speech sounds, the human voice source generates a quasi-periodic series of air pulses which excite the vocal tract. The modulation of the air stream from the lungs is caused by the vibration of the vocal folds. When the glottis, i.e. the orifice between the vocal folds, is closed, the sub-glottal pressure builds up. If this pressure is high enough, the vocal cords separate and a jet of air can escape through

the glottis into the vocal tract. Due to the Bernoulli effect, the pressure in the glottis decreases, and the resulting force together with the myoelastic tensions acting on and in the vocal folds "suck" the vocal folds together. The sub-glottal pressure starts to build up again, and the cycle is repeated. The way in which these interactions of aerodynamic and muscular forces set the vocal folds into vibration is far too complex to be modeled in every detail. Therefore, simplified models are used which only describe the most important aspects of vocal fold vibration (i.e. phonation). Which aspects of phonation are considered to be important, strongly depends on the context in which the model is to be used. For instance, models which aim at a correct description of the physical details of phonation differ from models developed for technological applications.

Voice-source models belonging to the first category describe the phonation process by means of a mechanical model which is controlled by a set of physical and physiological parameters like sub-glottal pressure, tissue characteristics of the vocal folds, and glottal rest area (Ishizaka & Flanagan, 1972; Cranen, 1987; Titze, 1989). It has been shown that these articulatory models are capable of generating realistic glottal flow patterns. However, knowledge about the relative perceptual importance of the various physical aspects of phonation which are captured by these models is still lacking (Klatt, 1987). One reason for this lack of knowledge is the laborious procedures one has to go through in order to obtain reliable measurements of various physical and physiological parameters (see for instance Cranen (1987)). This makes it difficult to extract rules from natural speech which are needed in TTS systems to control the dynamically changing voice source. High computational costs also prevent practical use of articulatory synthesis. Despite these disadvantages, it is believed that articulatory models will be the most powerful in the end (Klatt, 1987; Sondhi, 1990).

Another class of voice-source models directly describes the shape of the glottal flow pulse (Fujisaki & Ljungqvist, 1986; Fant, Liljencrants & Lin, 1985; Klatt & Klatt, 1990). The parameters of these models are not neccesarily related to physiological processes. They are chosen in such a way as to give the model optimal flexibility in synthesizing a wide variety of realistic glottal pulse shapes while keeping the number of parameters low.

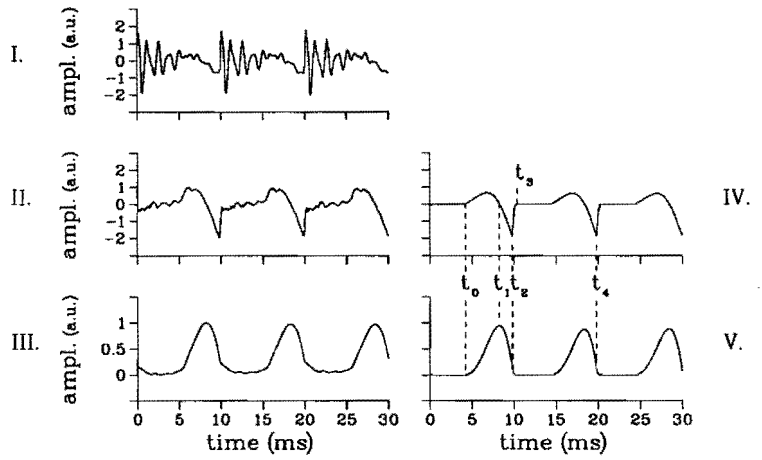

Figure 4.1: Inverse filtering: vowel /a/. *I.* Speech signal, *II.* radiated glottal-pulse waveform, *III.* glottal-pulse waveform, *IV.* modeled radiated glottal-pulse waveform, *V.* modeled glottal-pulse waveform. t_0: glottal onset time, t_1: moment of maximum flow, t_2: moment of glottal closure, t_3: moment of complete glottal closure, t_4: next moment of glottal closure. The amplitude of the waveforms is expressed in arbitrary units (a.u.).

Figure 4.1 shows a typical example of a glottal-pulse waveform (figure 4.1.III) and its corresponding speech waveform (figure 4.1.I). An inverse filtering technique was used to remove the effects of vocal tract and lip radiation, respectively (Wong, Markel & Gray, Jr., 1979). If only the vocal-tract effect is cancelled, we get the so-called radiated glottal-pulse waveform (figure 4.1.II). If both the vocal-tract and the lip-radiation effects are removed, the glottal-pulse waveform is revealed (figure 4.1.III). The main characteristics of these excitation waveforms are usually described by a set of time-based and amplitude-based parameters indicated in panels IV and V of figure 4.1: t_0 indicates the glottal onset, t_1 refers to the moment of maximum flow, t_2 corresponds to the moment of major discontinuity of the radiated glottal pulse, also

called the moment of glottal closure, and t_3 indicates the moment of complete glottal closure. From these timing parameters we can define the pitch period $T_0 = t_4 - t_2$, the open quotient $OQ = (t_2 - t_0)/T_0$[1], the speed quotient $SQ = (t_1 - t_0)/(t_2 - t_1)$, and the closing quotient $CQ = (t_3 - t_1)/T_0$. The amplitude parameters are the peak flow, the dc flow, and the ac flow. A large number of glottal waveform models have been proposed which implicitly or explicitly make use of these parameters. (Rosenberg, 1971; Fant, 1979; Klatt, 1980; Hedelin, 1984; Fant *et al.*, 1985; Fujisaki & Ljungqvist, 1986; Imaizumi, Kiritani, Fukawa & Saito, 1989; Price, 1989; Klatt & Klatt, 1990; Schoentgen, 1990; Tenpaku & Hirahara, 1990).

The parameters of almost all of these source models are derived under the assumption that speech can be described by a non-interactive source-filter model. This means that the vocal-tract transfer function cannot be affected by changes in glottal state, and vice versa, that the source waveform is unaffected by changes of the vocal tract. It has been shown that in general these assumptions are not valid for real speech (Fant & Lin, 1987, Lin, 1990). However, despite the fact that an interactive model can indeed reproduce certain spectral and temporal details of speech waveforms which are observed in real speech, the perceptual importance of these details has not yet been established (Nord, Ananthapadmanabha & Fant, 1984, Lin, 1990). Besides, some of these interaction phenomena can be simulated within the framework of a non-interactive model (Klatt & Klatt, 1990, Lin, 1990, Koizumi & Taniguchi, 1988). In this study we therefore adopt a non-interactive model instead of a more complex interactive model. The non-interactive model we used will be described in the next section.

4.2.2 The glottal-excited (GE) speech synthesizer

Figure 4.2 shows the GE speech synthesizer. $G(z)$ describes the characteristics of the glottal waveform. As a first approximation, the radiation effect is modeled by a differencing filter $R(z)$ that is constant in time and independent of the glottal and vocal-tract filtering. Therefore, we

[1]Sometimes, other definitions of OQ are used, for instance $OQ = (t_3 - t_0)/T_0$. As for the stimuli in this study t_3 nearly equals t_2, there is almost no difference between the various definitions.

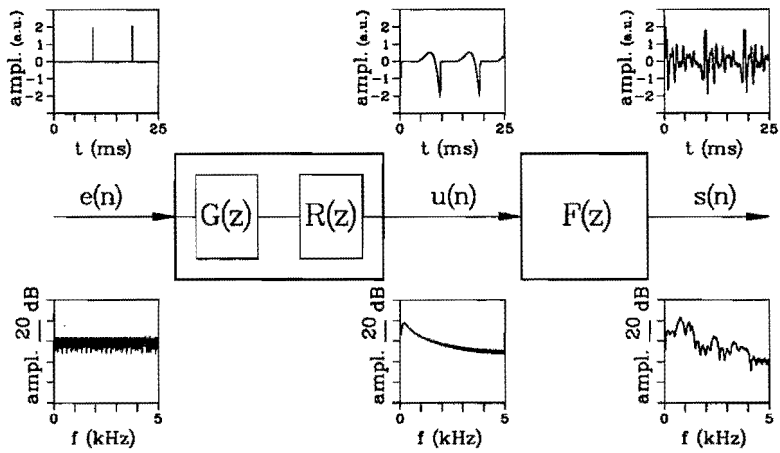

Figure 4.2: The glottal-excited (GE) speech synthesizer. The panels in the upper row show from left to right the waveforms of the delta pulse train $e(n)$, the radiated glottal pulse train $u(n)$, and the speech waveform $s(n)$. The panels in the lower row show the corresponding spectra. $G(z)$, $R(z)$, and $F(z)$ represent the glottal-pulse, the radiation, and the vocal-tract filter, respectively.

can treat the cascade of $G(z)$ and $R(z)$ as one filter. The vocal-tract filter $F(z)$ is excited by the radiated glottal pulse train $u(n)$.

In the case of the LPC model, the cascade of $G(z)$, $F(z)$ and $R(z)$ is treated as one filter. Within the framework of the GE synthesizer, however, we can control the characteristics of the glottal waveform $G(z)$ and the vocal-tract filter $F(z)$ independently.

The vocal-tract model

The vocal tract is modeled as an all-pole filter consisting of a cascade of two-pole resonators. Each resonator corresponds to a formant and is characterized by a centre frequency and a bandwidth. It should be

noted that for nasals, fricatives, and plosives the vocal-tract transfer function should also contain zeros. As we only use vowels in this study we will not consider pole-zero models for the vocal tract.

The voice-source model

We have chosen the Liljencrants-Fant (LF) model for the voice source, because it is mathematically well developed and can generate a wide variety of realistic pulse shapes (Fant *et al.*, 1985, Lin, 1990). Recently, a number of different studies have demonstrated the suitability of the LF model for speech synthesis. Fujisaki & Ljungqvist (1986) have given an overview of parametric models that are used to describe the glottal waveform. They used the minimum rms error as a measure to evaluate how well the voice-source models performed with respect to high-quality resynthesis of natural speech. They found that their model together with the LF model outperformed the other models, among which were the Rosenberg model (Rosenberg, 1971) and the old Fant model (Fant, 1979). Tenpaku & Hirahara (1990) also compared their model to the Rosenberg model, the old Klatt model (Klatt, 1980), the LF model and the Fujisaki-Ljungqvist model. The naturalness of male and female vowel stimuli synthesized with the different voice-source models was evaluated by listeners in a preference test. For both male and female stimuli natural speech was preferred over synthesized speech. For the male stimuli there was little difference between the five models. For female stimuli the Rosenberg and the LF model performed best. The LF model has also been used to study voice-source variations in connected speech (Gobl, 1988, Gobl & Ní Chasaide, 1988), the synthesis of female voices (Karlsson, 1989), and the acoustic correlates of different voice qualities (Gobl, 1989). Based on these results, rules are being developed to control the LF model in a text-to-speech system (Carlson *et al.*, 1989).

The LF model is shown in figure 4.3. The parameter T_p indicates the moment of maximum flow and T_e corresponds to the moment of glottal closure. With the parameter T_a we can describe incomplete glottal closure. The return-time T_a models the residual phase of progressing closure after the major discontinuity at T_e. The frequency-domain correspondence of T_a is a first-order low-pass filter. The high-frequency

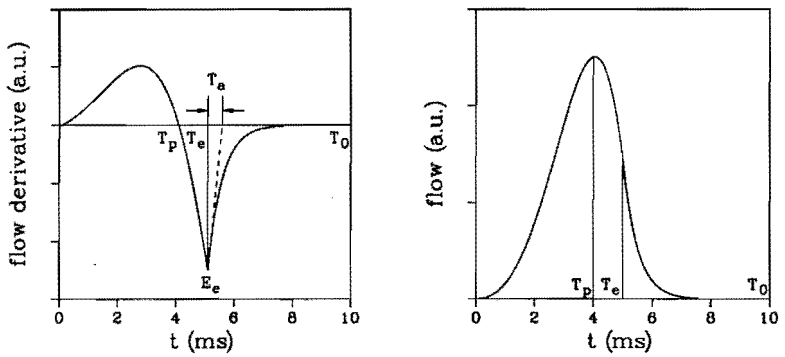

Figure 4.3: The Liljencrants-Fant (LF) model. The right panel shows the glottal-pulse waveform and the left panel its derivative. The parameters of the model, T_p, T_e, T_a, T_0, and E_e, are explained in the text.

spectral slope can hence be manipulated by varying T_a. The fourth parameter E_e represents the strength of the excitation at T_e.

The set which contains the five waveform parameters T_0, T_p, T_e, T_a, and E_e, is derived directly from the inverse filtered speech wave. Next, the waveform parameter set is converted to a parameter set which is used to synthesize the LF waveform (Fant *et al.*, 1985; Lin, 1990). This synthesis parameter set comprises the parameters T_0, α, ω_g, T_a, E_0. The flow derivative curve U'_g of figure 4.3 is generated according to the following equations:

$$\begin{aligned} U'_g &= E_0 \cdot e^{\alpha \cdot t} \cdot \sin(\omega_g \cdot t), && 0 \le t \le T_e \\ U'_g &= -\frac{|E_e|}{\varepsilon \cdot T_a} \cdot [e^{-\varepsilon \cdot (t - T_e)} - e^{-\varepsilon \cdot (T_0 - T_e)}], && T_e < t \le T_0 \end{aligned} \tag{4.1}$$

with

$$E_e = E_0 \cdot e^{\alpha \cdot T_e} \cdot \sin(\omega_g \cdot T_e). \tag{4.2}$$

4.2.3 Speech analysis

This subsection describes the various steps involved in the estimation of the parameters of the GE synthesizer from the speech waveform.

Correction of low-frequency phase distortion

The parameters of the LF model are estimated in the time domain. It is therefore important to have a linear phase response of the recording system, especially for frequencies up to the fundamental frequency. Phase distortions introduced by amplifiers, mixers and filters must be corrected.

Hunt (1978) describes a method for automatic correction of low-frequency phase distortions. A 20-Hz reference square wave is recorded. Due to the recording process, low-frequency phase distortions are introduced. The phase spectrum of the distorted wave is compared with that of an ideal undistorted 20-Hz square wave. The phase distortions are modeled by the filter

$$D(z) = \frac{z^{-1} - a}{1 - az^{-1}}. \tag{4.3}$$

Hunt uses a second-order all-pass filter with two variables as a correction filter. A first-order all-pass filter with only one variable is sufficient for our case, as the phase distortions which are introduced in our recording system are relatively small compared to the ones reported by Hunt.

The parameter a of equation 4.3 is automatically estimated by minimizing the r.m.s. error between the phase spectra of the model filter $D(z)$ and the experimentally determined transfer function. The corrected square wave is obtained by filtering the distorted signal with $D^{-1}(z)$. Figure 4.4. shows an example of the distorted and the corrected square waves as they are measured for the actual recording system.

Pitch-synchronous segmentation

The glottal waveform has to be derived pitch synchronously. To this end we determine the length and the *time location* of a pitch period by

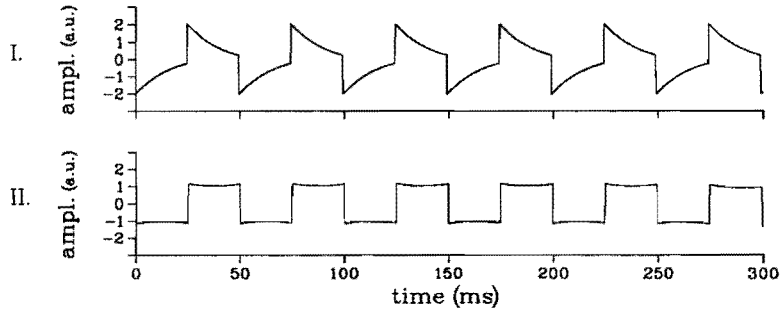

Figure 4.4: Low-frequency phase correction of a 20-Hz reference square wave. *I*. Distorted square wave, *II*. phase-corrected square wave.

using the moment of glottal closure to mark a pitch period. Markers are determined from the speech waveform.

Many methods for the pitch-synchronous segmentation of the speech waveform have been proposed in the literature (Strube, 1974; Ananthapadmanabha & Yegnanarayana, 1979; Wong *et al.*, 1979; Cheng & O'Shaughnessy, 1989; Dologlou & Carayannis, 1989; Funada, 1989; Moulines & Di Francesco, 1990; Ma, Kamp & Willems, 1992). Within the context of the present study, we developed our own method which is based on an idea by Wong *et al.* (1979). A linear prediction analysis using the covariance method is applied with a very small time window (3 ms). The analysis window is shifted sample by sample through the speech signal. In this way we calculate the total squared error as a function of time. This strategy was proposed by Wong *et al.* (1979), who showed that the moment of glottal closure can be accurately determined from the LPC normalized total squared error. They made three basic assumptions:

1. speech can be described by a linear model of speech production,
2. the vocal tract can be modeled by an all-pole filter,
3. the effective driving function shows a stable closed-glottis interval.

In general, these assumptions do not hold for natural speech. In continuous speech, most of the time, the moment of glottal closure cannot be determined exactly, owing to inaccurately estimated vocal-tract resonances. Also, the glottis may close more gradually or may not even close at all. In practice, we can only approximate the moment of glottal closure. As we had to determine the moments of glottal closure from natural speech, we modified the algorithm proposed by Wong *et al.* (1979) in a number of ways. In the next two paragraphs we describe the modifications in more detail.

Contrary to Wong *et al.* (1979), we do not normalize the total squared error. We use the fact that the total-squared-error signal is proportional to the input-signal energy. This means that the input-signal energy peaks near the moment of glottal closure. As we do not normalize the error signal, we use the input-signal energy as additional information in estimating the moment of glottal closure.

In order to reduce pitch jitter, the total squared error is smoothed with a second-order low-pass filter and local maxima are searched for in the error signal. These maxima indicate the region of glottal closure. A pitch marker is defined as that positive zero-crossing of the speech wave which is nearest in time to the left of a maximum of the error signal. Figure 4.5 shows an example of the error signal from which the pitch markers are derived.

Sometimes secondary maxima occur (see figure 4.5). These maxima correspond to moments of glottal opening. For each speech utterance we chose a fixed ratio which indicates the closed phase of a pitch period. The secondary maxima are used to check the validity of this choice.

Inverse filtering

Within the closed-glottis interval (CGI), the speech waveform is a freely decaying oscillation that is determined by the resonances of the vocal tract. It has been shown that CGI analysis is superior over other methods, such as pitch-synchronous and fixed-frame formant analysis, in deriving formant parameters from natural speech signals (Krishnamurthy & Childers, 1986, Pinto, Childers & Lalwani, 1989, Wood & Pearce, 1989). De Veth, Van Golstein-Brouwers, Boves & Van Heugten (1991) compared different inverse filtering techniques for the estimation of the

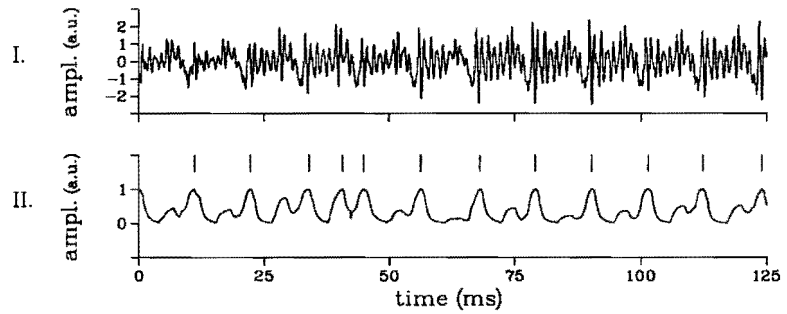

Figure 4.5: Pitch-synchronous segmentation of voiced speech. *I.* Speech signal of the vowel /a/, *II.* total-squared-error signal (maxima are indicated on top of this figure).

voice-source signal and found that CGI analyses performed best. In this study we estimate the vocal-tract filter by means of a CGI covariance analysis.

As mentioned earlier, we assume that the vocal tract can be modeled as a cascade of formant filters. The formant filters are estimated by means of an LPC covariance analysis applied to the closed-glottis interval. In this study, we use a fixed analysis order of 10. Next, the LPC filter is decomposed into second-order all-pole filters. A pole pair has to be complex conjugated in order to describe a formant filter. Therefore, only complex pole pairs are involved in the construction of the vocal-tract filter; real poles are excluded. The formant resonances were estimated by solving the roots of the LPC polynomial.

Once the vocal-tract filter $F(z)$ is known, we can calculate the source signal $u(n)$ by inverse filtering the speech signal $s(n)$ with $F^{-1}(z)$. The inverse-filter parameters are updated at the moments of glottal closure. The glottal-pulse waveform can be derived by integrating the radiated glottal-pulse waveform. Figure 4.6 shows some inverse filtering results.

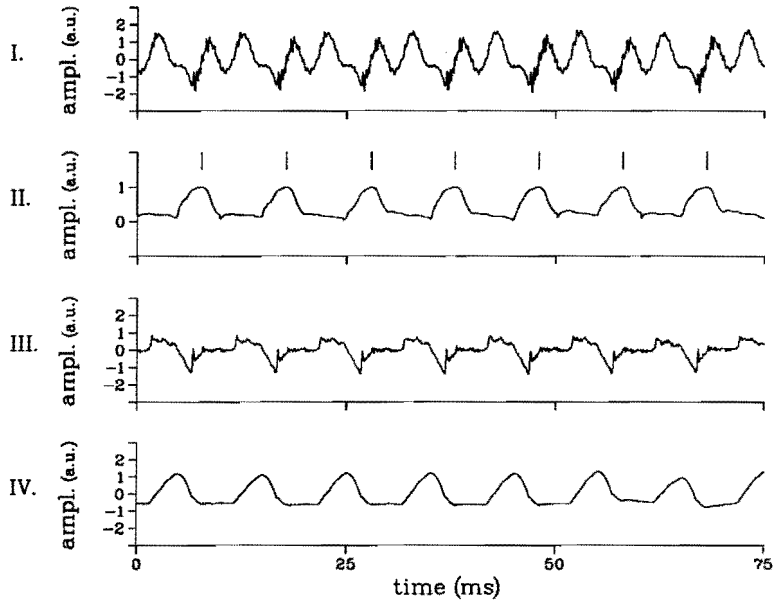

Figure 4.6: Inverse filtering results for the vowel /i/. *I*. Speech signal, *II*. total-squared-error signal, *III*. radiated-glottal-pulse waveform, *IV*. glottal-pulse waveform.

Stylization of the radiated glottal pulse

The measured radiated-glottal-pulse waveform is modeled with the Liljencrants-Fant model (Fant *et al.*, 1985). The program which fits the LF model to the calculated radiated-glottal-pulse waveform has two modes: an interactive mode and an automatic mode.

In the interactive mode the LF parameters can be adapted manually. The following four LF parameters can be adjusted: T_p (moment of maximum flow), T_e (moment of glottal closure), T_a (return time) and E_e (main excitation strength) (see figure 4.3). The starting point T_0 of the next LF pulse is also indicated. Whenever a parameter is changed, the LF curve and the corresponding spectrum are updated. At all times, the model curve and the measured radiated-glottal-pulse waveform can be graphically compared in both the time and frequency domains. If,

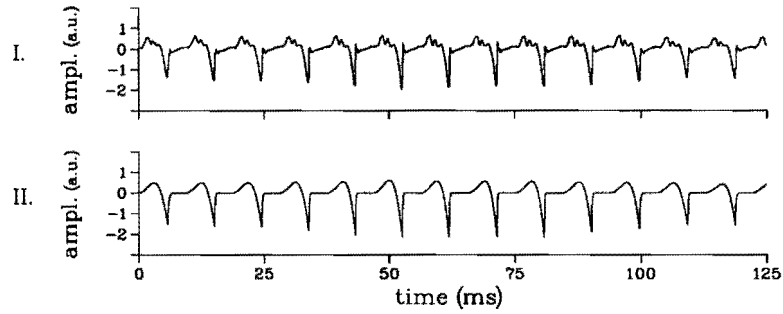

Figure 4.7: Stylization of the radiated glottal pulse with the Liljencrants-Fant model. *I*. Radiated glottal pulse, *II*. stylized source signal.

after visual inspection, the user of the program is satisfied about the match between the measured signal and its modeled equivalent, he can choose to continue the fitting procedure for the next radiated glottal pulse.

In the automatic mode the LF model is positioned on the measured source signal by means of a least-squares fit in the time-domain. Next, the spectral slope of the stylized signal is optimized by performing a frequency-domain least-squares fit. In practice, the first few pitch periods were fitted manually in order to get good initial conditions for the automatic fitting procedures. Figure 4.7 shows an example of a stylized source signal. In the next section we use the speech processing techniques to synthesize speech stimuli which are presented to subjects in a listening experiment.

4.3 Experiment

In this section a speaker identification experiment is described. According to Atal, speaker identification refers to "any decision-making process that uses some features of the speech signal to determine if a particular person is the speaker of a given utterance" (Atal, 1976, page 460, fn. 1). By using the analysis and synthesis techniques described in

the previous section we tried to determine which features of the speech signal a listener uses to perform the identification task. In particular, we wanted to find out whether the applied voice-source model codes speaker-specific information, and if so, what the relative importance is of voice-source and vocal-tract information used by listeners to identify a speaker.

This section starts with a discussion of earlier studies on speaker identity which are relevant to the present study. Next, speech stimuli, subjects, and experimental procedure are described. This section ends with a presentation of the results.

4.3.1 Earlier studies on speaker identity

There has been much research on speaker recognition (for overviews see: Hecker, 1971; Nolan, 1983; Doddington, 1985). Both from the viewpoints of speech production and speech perception there is evidence that voice-source as well as vocal-tract characteristics contribute to speaker identity.

Production differences between different speakers

Production differences between individual speakers have been found in several studies. Peterson & Barney (1952) determined the formant frequencies of 10 vowels produced by 76 men, women, and children. They showed that formant frequency patterns for the same vowels produced by different talkers differed considerably. The same holds for the glottal source characteristics. Monsen & Engebretson (1977) collected glottal waveforms produced by 10 male and female subjects. Analysis of these waveforms showed that glottal waveform characteristics, like shape, intensity, fundamental frequency, and phase and intensity spectrum vary over a wide range. Besides voice register and linguistic context, the sex of a speaker was one of the factors needed to account for these variations. Price (1989) also found clear male-female differences in glottal waveforms. These differences were much more pronounced than the variability measured among subjects of the same sex. Karlsson (1988) investigated glottal waveform differences for seven

normal female voices. She compared the dynamic variations of the glottal waveform parameters to a classification of the speakers by a speech therapist and found correlations between them. Klatt & Klatt (1990) investigated the acoustic cues to breathy voice quality. They analyzed utterances of ten female and six male talkers and found very large differences between subjects within each gender. The studies mentioned above indicate that both voice-source and vocal-tract characteristics differ from speaker to speaker. However, on the basis of these production studies, it is not clear how differences are perceived, and how, or even, if, they are used by humans to identify speakers from their voice. Next, we discuss some other studies which focus on the perception of acoustic differences between speakers of one gender.

Perceptual differences between different speakers

In an experiment by LaRiviere (1975), eight male speakers produced four isolated vowels. For each vowel there were three versions: a voiced, a whispered, and a low-pass filtered (200 Hz) vowel. The filtered vowel was meant to represent a "fundamental-frequency-only" condition, the whispered vowel represented a "formant-frequency-only" condition, and the voiced vowel contained both cues. Listeners were asked to perform a speaker identification on these stimuli. The whispered and the filtered vowels were identified equally well (22% and 21%, respectively), whereas the voiced vowels were identified best (40%). From this it was concluded that both voice-source (fundamental frequency) and vocal-tract information contribute equally to speaker identification judgments. Kuwabara & Ohgushi (1987) showed that the perception of speaker identity can be strongly affected by shifting the frequency of the lower three formants. They showed that the identity of a speaker is completely lost if all formant frequencies are shifted five percent. Formant bandwidths proved to be less important to perceived speaker identity. Imaizumi *et al.* (1989) showed that the resemblance between an original vowel and a synthetic one can be very high if a glottal source model is used. However, some aspects of the different voice qualities were not captured by the voice-source model. Klatt & Klatt (1990) showed that the use of a voice-source model enables the synthesis of a breathy voice quality. They also demonstrated that it was possible to synthesize female voices that were almost indistinguishable from their

original recordings by using a synthesizer which incorporates a voice-source model. Childers *et al.* (1989) showed that in order to achieve a high-quality conversion from one voice to another voice, the glottal excitation parameters should be taken into account. In conclusion, the studies mentioned in this section indicate that both voice-source and vocal-tract information are perceptually relevant for the synthesis of more natural male and female voices.

Hybrid voice samples

In the present study we want to determine whether voice-source characteristics which are coded by a model are used by listeners to identify speakers from their synthetic voices. To achieve this goal we adopted an experimental technique first introduced by Miller (1964) and later used by Matsumoto, Hiki, Sone & Nimura (1973), and Carrell (1984). This technique uses *hybrid voice samples*, i.e. voice samples for which the voice-source waveform of one speaker is modulated by the vocal-tract transfer function of another speaker. In this way the relative contributions of vocal-tract and voice-source characteristics to speaker identity can be investigated in a direct way.

Miller (1964) used an inverse filtering technique to separate the glottal-source characteristics from the vocal-tract transfer function. As explained in Hecker (1971), Miller used these voice-source and vocal-tract functions in four different experiments to determine their relative contribution to speaker identity. In the first experiment, the word /hod/ was uttered by two different speakers at the same fundamental frequency and the same duration. The corresponding two hybrid speech samples sounded more like the speaker whose vocal tract was represented. In the second experiment, the vocal-tract transfer function of one speaker corresponding to the word /hod/ was excited by various synthetic voice-source waveforms ranging from sinusoids and pulses to waveforms which were intended to be more realistic. Listeners reported that these stimuli sounded as produced by one speaker, although speech quality clearly differed among stimuli. The third experiment used two artificial but realistic glottal waveforms which excited the vocal-tract transfer functions of six speakers who uttered a sustained isolated vowel /a/. Again, the perceptual differences were mainly due to different

vocal-tract transfer functions. In the last experiment, Miller presented two natural and two hybrid voice samples to listeners in a two-choice identification test. The voice samples consisted of many repetitions of a 10-ms interval of a sustained isolated vowel /a/. The reference samples were the two natural voice samples and the test sample could be either a natural or a hybrid voice sample. The results show that the hybrid samples tend to be matched with the natural sample having the same vocal-tract transfer function. In summary, the experiments by Miller suggest that vocal-tract characteristics are much more important for the identity of a speaker than voice-source characteristics. As said in the introduction of this chapter we should keep in mind that the results of a study like Miller's depend on the inverse filtering technique used to separate voice-source and vocal-tract information.

Matsumoto *et al.* (1973) also found that vocal-tract characteristics contribute more to perceived speaker identity than voice-source characteristics. Although they also used hybrid voice samples, their experimental paradigm differed from Millers. In a first experiment, eight speakers uttered a sustained isolated vowel /a/ at three different pitches (120, 140, 160 Hz). These natural voice samples were presented in pairs to six listeners. The listeners had to indicate if the members of a pair were produced by the same or by a different talker. Multidimensional scaling techniques were used to construct a multidimensional representation of personal quality. In a second experiment, a subset of hybrid voice samples was synthesized to investigate the relation between voice-source and vocal-tract characteristics more directly. Matsumoto *et al.* used the same technique of inverse filtering as Miller. Like in Millers experiment, the stimuli also consisted of many repetitions of one pitch period. The hybrid samples were mixed with some of the original voice samples and presented to the listeners in a same-different experiment. From the analysis of these experimental data, the hybrid samples could be placed in the psychological auditory space which was determined in the first experiment. The results show that the hybrids tend to be closer to the original voice samples having the same vocal-tract configuration. As Matsumoto *et al.* and Miller used the same inverse filtering technique to generate the hybrid speech samples, we conclude that the work of Matsumoto *et al.* confirmed the findings of Miller that the relative contribution of the vocal-tract characteristics is greater than

the voice-source characteristics.

Hybrid voice samples were also used by Carrell (1984) who determined the contributions of fundamental frequency, formant spacing, and glottal waveform to talker identity. Carrell performed four different experiments of which two are of importance here. In his second experiment, Carrell presented glottal waveforms produced by six known talkers to a listener group. These waveforms were determined with a 'pseudo infinite length tube (PILT)' (Sondhi, 1975) and were not modulated with a vocal-tract transfer function. The experimental results showed that cues to talker identification were preserved in the glottal waveform, and that for voices that were well learned, this information is sufficient to identify talkers. In his fourth experiment Carrel used hybrid stimuli. This experiment was designed to study the interaction between fundamental frequency, formant spacing, and glottal waveform. Although each factor was shown to be important for speaker identification, the interactions turned out to be rather unexpected. It was found that the factorial combination of fundamental frequency, formant spacing, and glottal waveform of one and the same speaker did not produce the highest correct identification score. Carrell found that the contribution of the vocal-tract characteristics to speaker identity depended on the particular glottal waveform it was combined with. In should be noted that stimuli of Carrell consisted of CVC words which were synthesized by exciting the vocal-tract transfer functions with glottal waveforms derived from neutral sustained isolated vowels using the PILT device. The fact that this PILT device possibly did not remove all vocal-tract characteristics of the neutral vowel from the speech wave may account for the surprising interactions between glottal waveform and formant spacing.

This last example, once again, clearly shows that the experimental results can be strongly influenced by the speech processing techniques used to separate voice-source and vocal-tract information. Keeping this in mind, the experiments with hybrid voice samples seem to indicate that vocal-tract characteristics dominate speaker recognition by listening. However, the experiments by Carrell (1984) suggest that glottal waveform characteristics may also play a role in speaker identification. In this study we wanted to verify this last suggestion by means of a speaker identification experiment using hybrid voice sam-

ples. Our experiment differed in two respects from the previous experiments. Firstly, we used modern digital signal processing techniques to extract the voice-source and vocal-tract functions from the speech wave (see previous section). These techniques differed from the methods used in the previously reported experiments on hybrid voice samples. As a consequence, it is difficult to make a direct comparison between our results and the results of earlier studies. Secondly, in this study the glottal waveforms were parameterized by a voice-source model, whereas in the previous studies the measured glottal waveforms were used directly to excite the vocal-tract transfer functions.

4.3.2 Stimuli

Speakers

Four male speakers produced the speech stimuli used in the experiment. They all had normal voice quality.

Recordings

The speech recordings were made on a PCM recorder using a condensor microphone. During the recording sessions, samples of different vowels (/u/, /i/, /a/) were presented to the talkers over loudspeakers. These vowels had constant pitch, lasted 500 ms and were followed by a 2-second interval of silence. During this interval the speakers had to produce the same vowel at the same pitch, and at a certain intensity level. This task was repeated many times for nine different pitch–intensity conditions: three pitches (100, 120 and 150 Hz) and three intensity levels (low, medium, and high). In this way we got a small data base from which we selected the vowels used in the experiment.

Natural vowels

For every speaker four realizations of the vowel /a/ were selected from the data base. These realizations were produced at medium intensity level, and at a 100 Hz pitch. The following notation will be used to refer to the sixteen natural vowels: $(P_i R_j)_{natural}$, where P_i represents the i^{th} sPeaker, and R_j the j^{th} Realization ($i = 1,4$; $j = 1,4$).

Resynthesized vowels

The analysis strategy described in the previous section was used to derive the voice-source waveforms and the vocal-tract transfer functions from the sixteen natural vowels. Part of the analysis results will be presented in more detail in section 4.3.5. If a voice-source waveform is used to excite the vocal-tract transfer function which stems from the same analysis, the result will be a resynthesized version of the natural vowel. The following notation will be used to refer to the resynthesized vowels: $(P_i R_j)_{resyn}$, where P_i represents the i^{th} sPeaker, and R_j the j^{th} Realization $(i = 1,4\ ;\ j = 1,4)$.

Hybrid vowels

If the voice-source waveform of one speaker is modulated by the vocal-tract transfer function of another speaker, the result will be a so-called hybrid vowel. For every speaker, the voice-source waveform and the vocal-tract transfer function of the fourth realization of the vowel /a/ (R_4) was used to generate the hybrid stimuli. In this way, twelve hybrid vowels were synthesized: $(S_m T_n)_{hybrid}$, where S_m represents the voice-Source waveform of speaker m $(m = 1,4)$, and T_n the vocal-Tract transfer function of speaker n $(n = 1,4)$. If $n = m$ we get the four resynthesized vowels $(P_m R_4)_{resyn}$ (see the previous subsection on resynthesized vowels).

4.3.3 Subjects

Four subjects participated in the experiments (MS, MH, RS, and NV). Two of them (RS, NV) also served as speakers. No one reported hearing deficiencies. All subjects know the speakers very well.

4.3.4 Procedure

Subjects were seated at a computer terminal in a sound-proof room. All stimuli were presented diotically through insert earphones (etymōtic research ER-2) at a level of 70 dB SPL. All stimuli had a duration of 691.2 ms, started at a zero crossing, and had 25.6 ms offsets. Subjects

had to indicate which speaker they thought had produced the vowel stimulus by typing keys 1, 2, 3, or 4 (the numbers corresponding to the four speakers) on the computer keyboard.

Before the experimental runs started, all subjects were trained to identify the speakers at percentage-correct levels of 87.5% or higher. The natural vowel stimuli were used for this training. After the subjects passed the initial training procedure, they participated in four 45-minute sessions.

Each 45-minute session comprised three experimental runs. After the last session subjects were asked informally about the cues they had used to perform the identification task.

One experimental run contained four parts: 1. familiarization of natural vowels, 2. training of natural vowels, 3. testing of natural vowels, 4. testing of resynthesized and hybrid vowels.

Familiarization

The goal of the familiarization phase was to give the subjects an auditory impression of the vowels which were produced by the four speakers, and which were used in the experiment. This was realized by presenting the sixteen natural vowels (four speakers, four realizations) to the subjects two times. The first time started with the presentation of the first realizations of speaker 1, 2, 3, and 4, respectively. Next, the second, third, and fourth realizations were presented in the same order. The second time, the sixteen natural vowels were presented in a random order.

During the familiarization phase the numbers 1 through 4, corresponding to speaker 1 through 4, respectively, were displayed on the computer terminal. 500 ms before the presentation of a stimulus, an arrow head pointed to the number of the speaker which would utter the vowel. 500 ms after the presentation, the arrow head disappeared and a 1-second interval of silence preceded the next stimulus presentation.

Training

The training phase was meant to give subjects the opportunity to learn to recognize the speakers by their natural vowels. During the familiarization phase subjects only had to listen to the stimuli, whereas this time they also had to indicate the speaker who produced the natural vowel. Both for the familiarization and the training phase, feedback was given on the speaker who uttered the vowel. The next two phases of the experimental run were tests which provided no feedback on the correct speaker of the utterance. At the same time, the training phase was used by the experimenter to check the number of correct identifications. If this score was lower than 87.5%, the training phase was repeated. However, due to the training sessions which preceded the experiment, most of the time the percentage-correct scores were above this threshold.

All sixteen natural vowels were presented three times. The first time, the sixteen stimuli were presented in a random order. A new random order of the sixteen stimuli was determined for the second and third time, respectively.

During the training phase the numbers 1 through 4, corresponding to speaker 1 through 4, respectively, were displayed on the computer screen. After the subject had identified the speaker of the stimulus, feedback on the correct answer was given by means of an arrow head pointing to the number of the speaker which had uttered the stimulus. This arrow head remained on the display for 500 ms, after which a 1-second interval of silence preceded the next stimulus presentation.

Testing: natural vowels

After the training phase the subjects performed a test in which they had to identify the speakers by their natural vowels. This test much resembled the training phase, only this time no feedback was given. The presentation order of the stimuli was determined the same way as for the training phase.

Testing: resynthesized and hybrid vowels

In this test the resynthesized stimuli $(P_iR_j)_{resyn}$, where $i = 1, 4$, and $j = 1, 3$ were mixed with the hybrid stimuli $(S_mT_n)_{hybrid}$, where $m = 1, 4$, and $n = 1, 4$.[2]

Each resynthesized stimulus was presented two times, whereas each hybrid stimulus was presented three times. All stimuli were presented in a "quasi" random order. This order was determined in such a way that both the voice-source waveform and the vocal-tract transfer function of the current stimulus differed from the voice-source waveform and vocal-tract transfer function of the previous stimulus. The seventy-two stimuli were preceded by ten 'dummy' stimuli for which the subject responses were discarded from the data analysis.

During this test, the numbers 1 through 4, corresponding to speaker 1 through 4, respectively, were displayed on the computer screen. After the subject had responded by typing 1, 2, 3, or 4 on the computer keyboard, the next stimulus was presented after a 1-second interval of silence. No feedback was provided.

4.3.5 Results

In this section we start with a presentation of the results of the acoustic analysis of the speech utterances. Next, the results of the listening experiments will be presented.

Acoustic data

The acoustic data of the voice-source functions are shown in Table 4.I. All parameter values in this table are presented per speaker. Each parameter value is the result of averaging over the four different realizations of the vowel produced by that particular speaker. This means

[2] Four of the sixteen hybrids, $(S_mT_m)_{hybrid}$ $(m = 1, 4)$, are in fact resynthesized stimuli, $(P_iR_4)_{resyn}$ $(i = 1, 4)$. However, in this test we refer to them as hybrid stimuli as they also result from the analysis of the fourth realizations of the vowel /a/. Besides, like the other hybrids, they were also presented three times in this test, whereas the other resynthesized stimuli were presented only two times.

that every parameter value is based on 224 measurements (for each of the 4 realizations 56 glottal pulses were involved in the measurements).

Table 4.I: Acoustic voice-source parameters. The four columns correspond to speakers P_1, P_2, P_3, and P_4, respectively. For an explanation of the parameters see text.

Parameter	P_1	P_2	P_3	P_4
Fundamental frequency				
F_0 (Hz)	98.7	98.9	99.9	99.2
se (Hz)	1.3	0.8	1.7	0.8
range (Hz)	2.6	2.1	4.1	1.6
jitter (ms)	0.076	0.061	0.070	0.073
se (ms)	0.008	0.004	0.010	0.007
σ_{DF_0}	0.009	0.007	0.009	0.009
Open quotient				
OQ	0.57	0.66	0.64	0.85
se	0.02	0.08	0.06	0.03
Closing quotient				
CQ	0.161	0.160	0.167	0.225
se	0.007	0.011	0.003	0.014
Normalized return time				
R_a	0.013	0.003	0.008	0.007
se	0.002	0.001	0.002	0.002
Excitation strength				
E_e (dB)	-3.8	-3.6	-3.0	-3.6
se (dB)	1.1	0.8	0.6	1.1
shimmer (dB)	0.50	0.40	0.45	0.49
se (dB)	0.06	0.04	0.06	0.08

The Liljencrants-Fant model can be described by different sets of parameters (Fant, Liljencrants & Lin, 1988). The set consisting of four timing parameters (T_p, T_e, T_a, T_0) and amplitude parameter (E_e),

Table 4.II: Acoustic vocal-tract parameters. The four columns correspond to speakers P_1, P_2, P_3, and P_4, respectively. For an explanation of the parameters see text.

Parameter	P_1	P_2	P_3	P_4
F_1 (Hz)	823.7	788.4	775.7	752.0
se (Hz)	8.7	7.3	12.4	7.5
B_1 (Hz)	20.9	73.7	65.9	68.3
se (Hz)	2.6	5.9	4.0	14.0
F_2 (Hz)	1233.0	1251.5	1204.1	1178.2
se (Hz)	23.8	39.6	29.4	7.0
B_2 (Hz)	20.7	55.9	31.7	29.5
se (Hz)	1.3	5.4	1.5	7.4
F_3 (Hz)	2612.2	2419.7	2433.9	2463.4
se (Hz)	31.9	25.5	19.6	20.0
B_3 (Hz)	95.1	60.0	51.2	60.9
se (Hz)	27.8	8.2	3.7	15.3
F_4 (Hz)	3156.7	3268.4	3342.6	3227.1
se (Hz)	126.0	60.5	58.9	23.3
B_4 (Hz)	174.5	168.3	189.6	61.2
se (Hz)	6.7	38.5	33.6	4.9
F_5 (Hz)	3998.3	3891.8	3949.4	3950.3
se (Hz)	77.5	20.1	60.2	128.0
B_5 (Hz)	176.3	124.9	97.9	239.0
se (Hz)	39.1	25.7	18.9	87.6

which was used in section 4.2.2 to introduce the LF model, is just one example of such a set. In Table 4.I we use another set comprising the following five parameters: fundamental frequency F_0, open quotient OQ, closing quotient CQ, normalized return time R_a, and excitation strength E_e. This set is used because the parameters have a closer relation to perceptually important frequency-domain properties of the voice-source function.

Besides the mean fundamental frequency F_0, the corresponding standard error (se), and the F_0 range, two F_0 perturbation measures are also tabulated. Horii (1985) defines jitter as the mean of the differences between the T_0's of consecutive pitch periods. Another perturbation measure is σ_{DF_0} which represents the standard deviation of the distribution of the relative F_0 frequency differences. Askenfelt & Hammarberg (1986) found that, out of a set of seven different waveform perturbation measures, this measure performed best with regard to acoustic–perceptual correlation and the ability to discriminate between normal and pathological voice status.

The open quotient OQ is defined as T_e/T_0. It measures the fraction of the pitch period the glottis is open. In the frequency domain this means that the spectral component having a duty-cycle nearest to the open time of the glottis is favoured in the voice-source spectrum (Pabon, 1991). In this way, the OQ determines the level balance of the lower harmonics of the voice-source spectrum. In particular, the amplitude of the first harmonic relative to adjacent harmonics is increased or decreased due to changes in OQ (Klatt & Klatt, 1990).

The closing quotient CQ is defined as $(T_e + T_a - T_p)/T_0$ (Tenpaku & Hirahara, 1990). It determines the transient character of the glottal pulse. The smaller the CQ, the stronger the high frequency partials in the spectrum.

The normalized return-time parameter R_a is defined as T_a/T_0. The effect of the return phase of the LF pulse on the voice-source spectrum can be modeled as a first-order low-pass filter with cut-off frequency $F_a = 1/(2\pi T_a)$ (Fant *et al.*, 1985, Fant *et al.*, 1988). In this way, R_a influences the spectral slope of the voice-source spectrum. An increasing R_a causes an increased high frequency deemphasis.

Other parameters often found in the literature, like the speed quo-

tient SQ, the pulse-asymmetry factor R_k, and the normalized glottal frequency R_g can be derived easily from OQ, CQ, and R_a[3].

The excitation strength E_e is expressed in dBs relative to the minimum E_e (in our case, E_e has a minimal integer value of -2048, according to the 12-bit signed-integer format used to store the samples of the glottal pulse). The perturbation of E_e is expressed in Table 4.I by the shimmer parameter. This parameter is calculated according to the definition of shimmer given by Horii (1985).

The acoustic data for the vocal-tract functions are shown in Table 4.II. The formant frequencies F_i $(i = 1,5)$, and bandwidths B_i $(i = 1,5)$ are estimated by solving the roots of the LPC polynomial. Again, each parameter value is based on 224 measurements (4 realizations times 56 vocal-tract functions).

Natural vowels

In this subsection we present the speaker-identification scores for the natural vowels. Speaker identifications made by the four subjects were pooled because the intersubject consistency proved to be good. Figure 4.8 shows sixteen panels each of them corresponding to one of the sixteen natural vowels $(P_iR_j)_{natural}$ used in the experiment (P_i represents the i^{th} sPeaker, and R_j the j^{th} Realization, $i = 1,4$; $j = 1,4$). Each panel shows which percentage of the total number of responses for that stimulus is attributed to each of the four response alternatives. A response alternative corresponds to one of the speakers who uttered the natural vowels. Each panel is based on 144 responses (4 subjects, 12 runs, each stimulus occurred 3 times in a run). As an example, the upper right panel of figure 4.8 shows that whenever the fourth realization R_4 of the vowel uttered by speaker four P_4 was presented to the subjects, they responded by typing 4 (i.e. speaker 4) on the computer keyboard.

The overall percentage correct identifications is 91.9%. The percentage-correct scores per speaker are 97.2%, 87.2%, 87.8%, and 95.5% for speakers 1 through 4, respectively. Most of the confusions

[3]Speed quotient $SQ = \frac{T_p}{T_e - T_p} = \frac{OQ}{CQ - R_a} - 1$, pulse-asymmetry factor $R_k = \frac{T_e - T_p}{T_p} = \frac{CQ - R_a}{OQ - CQ + R_a}$, normalized glottal frequency $R_g = \frac{T_0}{2T_p} = \frac{1}{2(OQ - CQ + R_a)}$.

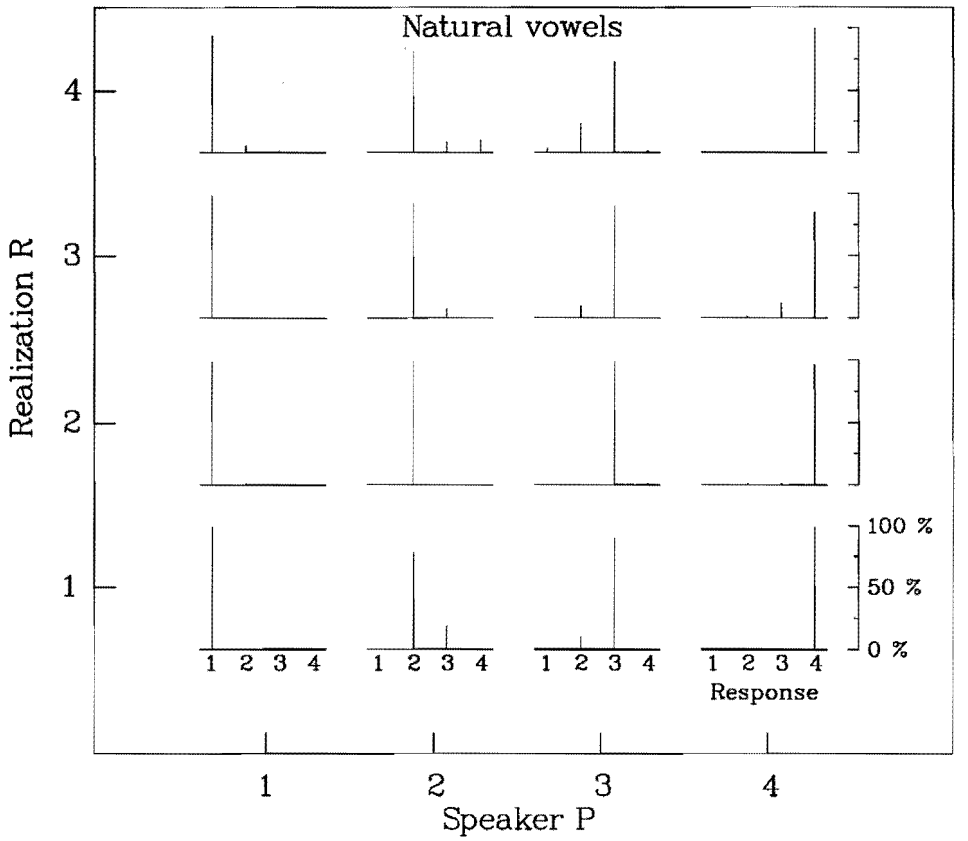

Figure 4.8: Speaker-identification scores for the natural vowels. This figure shows 16 panels each corresponding to one the 16 natural vowels used in the experiment (4 speakers, 4 realizations per speaker). Each panel shows which percentage of the total number of responses for that particular stimulus is attributed to each of the 4 possible response categories (speakers 1, 2, 3, 4, respectively).

are made between speakers 2 and 3. In 9.2% of the cases an utterance of speaker 2 was presented to the subjects, they responded by typing speaker 3. Vice versa, 10.9% of the speaker–3 stimuli were identified as

being produced by speaker 2.

Resynthesized vowels

In this subsection we present the speaker-identification scores for the resynthesized vowels $(P_iR_j)_{resyn}$ (P_i represents the i^{th} sPeaker, and R_j the j^{th} Realization, $i = 1,4$; $j = 1,4$). The percentage-correct identification scores are plotted in figure 4.9. Speaker identifications made by the four subjects were pooled. Again, each panel shows the distribution of the responses of the subjects to the resynthesized version of the j^{th} vowel realization uttered by the i^{th} speaker. Each of the four panels $(P_iR_4)_{resyn}$ of the upper row of figure 4.9 is based on 144 responses (4 subjects, 12 runs, each stimulus occurred 3 times in a run). All other panels are based on 96 responses (4 subjects, 12 runs, each stimulus occurred 2 times in a run).

The overall percentage correct identifications is 81.8%. The percentage-correct scores per speaker are 94.7%, 78.5%, 80.6%, and 73.4% for speakers 1 through 4, respectively. If we compare these results to the results for the natural vowels, we see that the identification score for speaker 4 has dropped more than 20%. In particular, the second and third realizations are not very well resynthesized. The fourth realization which was used for generating the hybrid stimuli is resynthesized very well. If we compare figure 4.9 with figure 4.8 we see that in general the confusions are somewhat more pronounced for the resynthesized stimuli. The distribution patterns of the corresponding panels, however, much resemble each other with the exception of panels $(P_4R_2)_{resyn}$ and $(P_4R_3)_{resyn}$.

Hybrid vowels

The results for the hybrid vowels are shown in figure 4.10. Again, speaker identifications made by the four subjects were pooled. Each hybrid vowel $(S_mT_n)_{hybrid}$, synthesized by using the voice-Source waveform S_m of speaker m and the vocal-Tract transfer function T_n of speaker n, was presented 144 times to the subjects (4 subjects, 12 runs, each

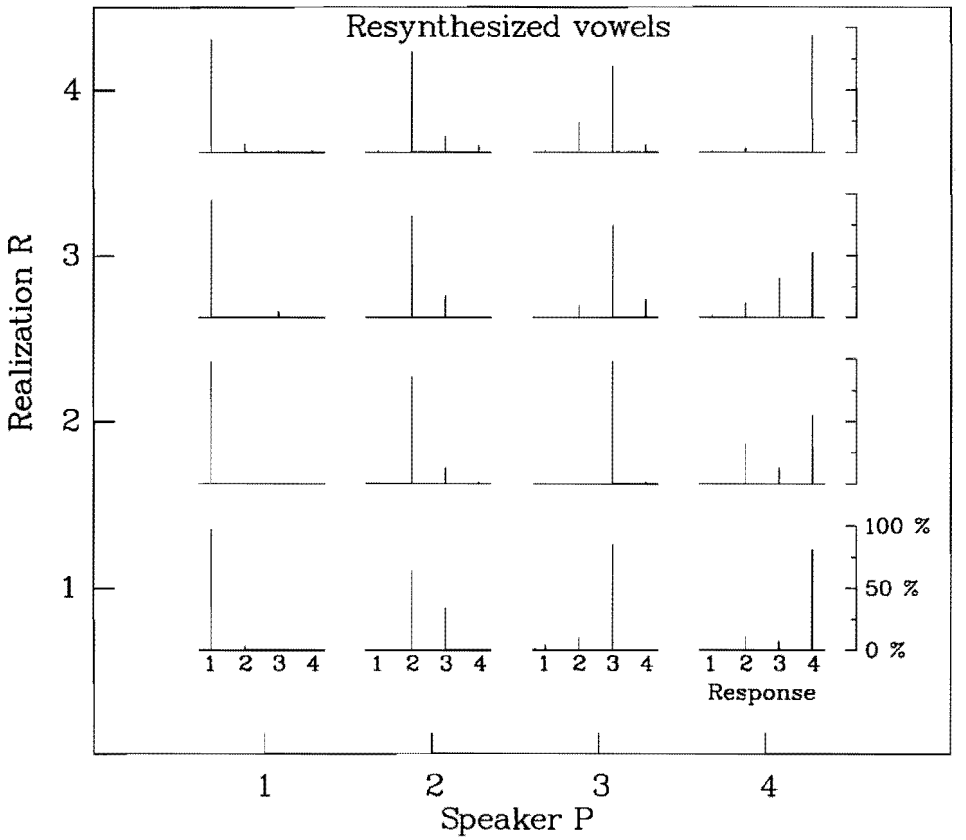

Figure 4.9: Speaker-identification scores for the resynthesized vowels. This figure shows 16 panels each corresponding to one the 16 resynthesized vowels used in the experiment. Each panel shows which percentage of the total number of responses for that particular stimulus is attributed to each of the 4 possible response categories (speakers 1, 2, 3, 4, respectively).

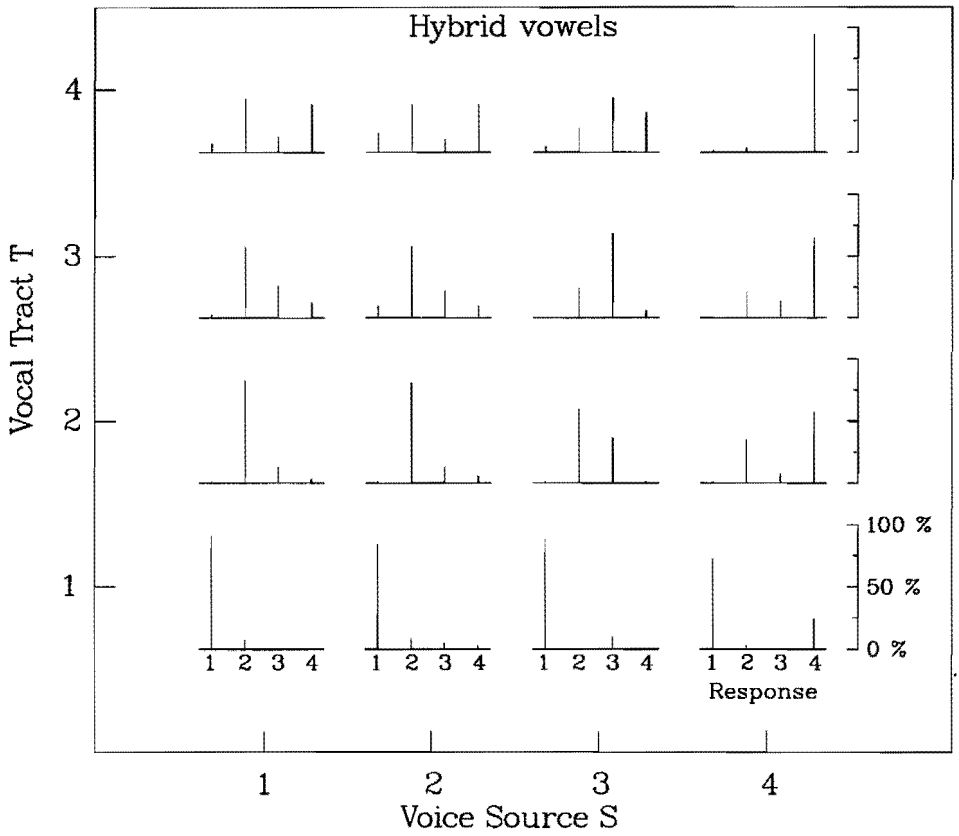

Figure 4.10: Speaker-identification scores for the hybrid vowels. This figure shows 16 panels each corresponding to one the 16 hybrid vowels used in the experiment. Each panel shows which percentage of the total number of responses for that particular stimulus is attributed to each of the 4 possible response categories (speakers 1, 2, 3, 4, respectively).

stimulus occurred 3 times in a run). As an example, the lower right panel of figure 4.10 shows the distribution of the subjects responses for the hybrid vowel $(S_4T_1)_{hybrid}$. In 104 cases (72%) subjects responded by typing 1 on the computer keyboard (i.e. speaker 1). The hybrid

vowel was identified 34 times (24%) as being produced by speaker 4. Speakers 2 was chosen 5 times (3%) and speaker 3 only once (1%).

From figure 4.10 it can be seen that for the stimuli on the diagonal S_mT_m $(m = 1, 4)$, the number of responses which corresponds to speaker m is by far the strongest. Figure 4.10 also shows that for most panels the response alternatives corresponding to either the speaker who's voice-source function was used to synthesize the hybrid vowel, or the speaker who's vocal-tract function was applied, are favoured over the other response alternatives. Panels $(S_1T_3)_{hybrid}$ and $(S_1T_4)_{hybrid}$ form an exception.

In the case of the hybrid stimuli there are in fact no correct responses[4]. Therefore, we adopted two independent definitions of correct responses from Carrell (1984): a formant-based accuracy measure, and a glottal-based accuracy measure.

When calculating the formant-based accuracy measure, a response is considered to be correct if the identified speaker corresponds to the speaker who's vocal-tract function was used to generate the hybrid vowel (Carrell, 1984). Figure 4.11 shows the percentage correct scores according to the formant-based accuracy measure. This figure shows four panels corresponding to the four vocal-tract functions used to generate the hybrid stimuli. Each panel contains 4 bars indicating the percentage correct responses. Each bar represents 144 responses and corresponds to one of the four possible voice-source functions S_1, S_2, S_3, S_4, respectively. On top of the percentage-correct bars the actual score is indicated (in percents). If a percentage correct score does not differ significantly $(n.s)$ at the 5% level from chance (25%), this is indicated at the bottom of the bar.

The overall percentage correct formant-based identifications is 58%. The overall percentage correct scores for vocal-tract 1 through 4 are 84%, 65%, 33%, and 50%, respectively. As expected, the bars in figure 4.11 which correspond to stimuli for which the voice-source and vocal-tract functions belong to the same speaker are highest (the difference between $(S_1T_2)_{hybrid}$, $(S_2T_2)_{hybrid}$ being not significant).

In the case of the glottal-based accuracy measure, a response is

[4]Of course, this is not true for the hybrids $(S_mT_n)_{hybrid}$, where $m = n$, as these stimuli are resynthesized versions of original vowels.

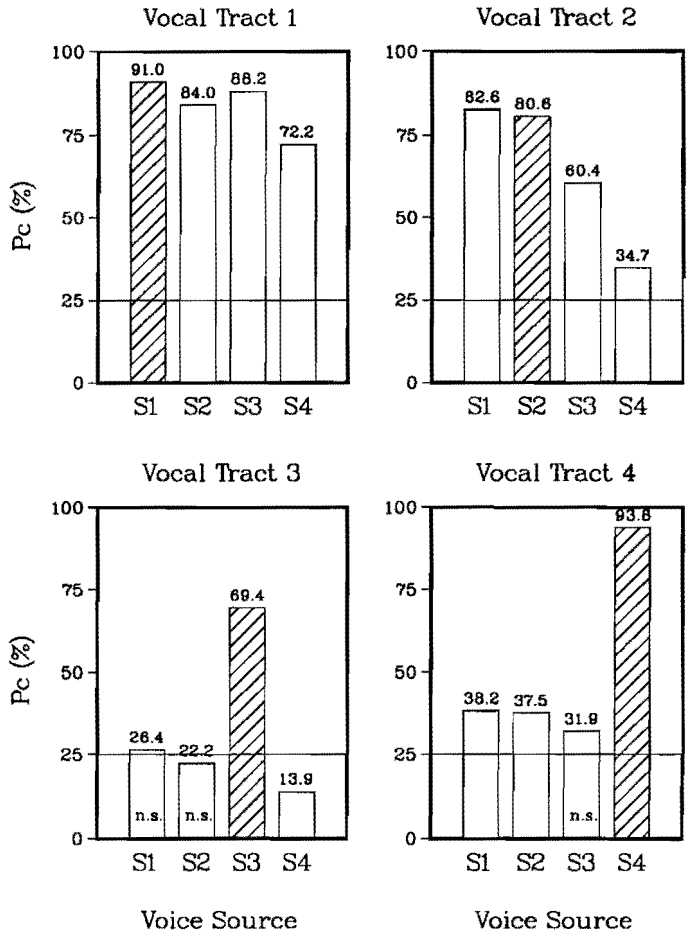

Figure 4.11: Formant-based accuracy measure. Each of the four panels corresponds to one of the four vocal-tract functions used to generate the hybrid stimuli. The height of a vertical bar indicates the percentage correct responses as defined by the formant-based accuracy measure (see text). Each panel contains 4 bars corresponding to the four possible voice-source functions S_1, S_2, S_3, S_4, respectively. The shaded bars correspond to the stimuli for which the voice-source and vocal-tract functions belong to the same speaker. In each panel, chance level (25%) is indicated by a horizontal line. If a percentage correct score does not differ significantly ($n.s$) at the 5% level from chance, this is indicated at the bottom of the bar.

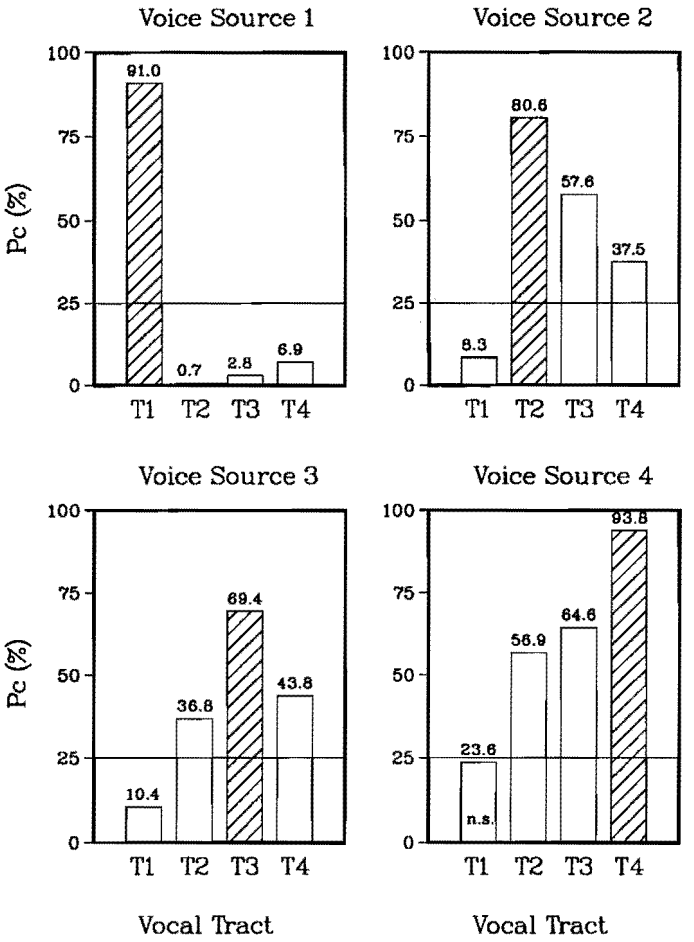

Figure 4.12: Glottal-based accuracy measure. Each of the four panels corresponds to one of the four voice-source functions used to generate the hybrid stimuli. The height of a vertical bar indicates the percentage correct responses as defined by the glottal-based accuracy measure (see text). Each panel contains 4 bars corresponding to the four possible vocal-tract functions T_1, T_2, T_3, T_4, respectively. The shaded bars correspond to stimuli for which the voice-source and vocal-tract functions belong to the same speaker. In each panel, chance level (25%) is indicated by a horizontal line. If a percentage correct score does not differ significantly ($n.s$) at the 5% level from chance, this is indicated at the bottom of the bar.

counted as a correct answer if the identified speaker corresponds to the speaker who's voice-source function was used to generate the hybrid vowel (Carrell, 1984). Figure 4.12 shows the percentage correct scores according to the glottal-based accuracy measure. This figure shows four panels corresponding to the four voice-source functions used to generate the hybrid stimuli. Each panel contains 4 bars indicating the percentage correct responses. Each bar represents 144 responses and corresponds to one of the four possible vocal-tract functions T_1, T_2, T_3, T_4, respectively. On top of the percentage-correct bars the actual score is indicated (in percents). If a percentage correct score does not differ significantly ($n.s$) at the 5% level from chance (25%), this is indicated at the bottom of the bar.

The overall percentage correct glottal-based identifications is 43%. The overall percentage correct scores for voice-source 1 through 4 are 25%, 46%, 40%, and 60%, respectively. Like in figure 4.11 , the bars in figure 4.12 are highest when the voice-source and vocal-tract functions used to generate the stimulus belong to the same speaker. The data clearly show that listeners use both source and tract information to identify the speakers. If we compare figure 4.11 with figure 4.12 we see that speaker 1 is dominated by his vocal tract, whereas speakers 3 and 4 mostly by their source. For speaker 2, tract and source seem to play an equally important role.

Informal comments made by the subjects

As mentioned before, after the last experimental run subjects were asked informally about the cues they had used to perform the identification task. Every subject was able to explicitly tell something about the cues he had used to choose between the four speakers. There was a remarkable correspondence in terminology used to describe these cues. We will now give an impression of the four speakers based on the informal comments made by the subjects.

Speaker 1 : "this one knows how to produce an /a/: from the toes", "the /a/ of speaker 1 is produced with a wide open mouth".

Speaker 2 : "the vowels of speaker 2 have an LPC-like quality".

Speaker 3 : "this speaker has difficulty in producing a vowel with stable pitch".

Speaker 4 : "this speaker has lots of bass in his voice".

Once again, it should be noted that these remarks were made informally. Nevertheless, they are representative for the way the subjects characterized the four speakers. It is therefore interesting to see if the results of the acoustic analysis can support these subjective statements in a more objective way. We will come back to this issue at the end of the next section where the experimental results will be discussed in more detail.

4.4 Discussion

In this section we start with a discussion of the techniques we used to analyse and synthesize voiced speech sounds. Next, we will return to the main question of this study: Does a voice-source model code speaker-specific information if we separate voice-source and vocal-tract information the way we did in the present study, and if so, what is the relative importance of coded voice-source and vocal-tract information for perceived speaker identity?

4.4.1 Speech processing: analysis and synthesis of voiced speech

If we compare the identification scores for the natural vowels with the scores for the resynthesized vowels, we see that the overall percentage correct responses has decreased from 91.9% to 81.8%. Comparison of figure 4.9 with figure 4.8 shows that the confusion patterns of the corresponding panels much resemble each other with the exception of panels $(P_4R_2)_{resyn}$ and $(P_4R_3)_{resyn}$. These results indicate that the analysis/resynthesis technique captured most of the speaker-specific characteristics used by the listeners to identify the speakers.

To see how our analysis/resynthesis techniques compares to a more standard LPC analysis/resynthesis technique, we performed an additional experiment. The same subjects participated in this experiment.

The experimental procedure was almost identical to the experimental procedure described in section 4.3. Only this time, an experimental run contained three instead of four parts, because the testing of natural vowels was skipped. The LPC stimuli were presented in the last part of each experimental run. The following notation will be used to refer to the LPC stimuli: $(P_iR_j)_{LPC}$, where P_i represents the i^{th} sPeaker, and R_j the j^{th} Realization ($i = 1,4$; $j = 1,4$). A pitch-asynchronous autocorrelation LPC analysis was applied on the same sixteen vowels used in the main experiment. Next, the LPC stimuli were synthesized using a 10^{th}-order LPC resynthesis (Vogten, 1983). Each of the sixteen LPC stimuli was presented 24 times to the listeners. The results for the speaker identification of the LPC stimuli are shown in figure 4.13.

Speaker identifications made by the four subjects were pooled, because the results were reasonable uniform across subjects. Each panel of figure 4.13 shows the distribution of the responses of the subjects to the LPC version of the j^{th} vowel realization uttered by the i^{th} speaker. Each panel is based on 96 responses (4 subjects, 6 runs, each stimulus occurred 4 times in a run). The overall percentage correct identifications is 75.5%. The percentage-correct scores per speaker are 94.5%, 64.3%, 81.0%, and 62.0% for speakers 1 through 4, respectively. If we compare the overall percentage correct responses for the resynthesized stimuli (see section 4.3.5) with the overall score for the LPC stimuli, we see that our analysis/resynthesis technique performs better than the standard LPC method (81.8% versus 75.5%). By a better performance we mean a better preservation of the speaker-specific features contained in the speech signal. A comparison between figure 4.13, figure 4.9, and figure 4.8 shows that the confusion patterns for the LPC stimuli differ more from the patterns for the natural vowels than the confusion patterns for the resynthesized stimuli do. This also indicates a better performance of our analysis/resynthesis scheme.

In case of the LPC analysis/resynthesis scheme, the excitation waveform consists of a quasi-periodic train of delta pulses. The F_0 parameter controls the repetition frequency of the pulses, and a gain factor determines the amplitude of the pulses. This means that the LPC filter codes spectral characteristics of the human voice source as well as spectral characteristics of the vocal-tract. Therefore, it is not possible to synthesize hybrid stimuli with this technique. In this study we made

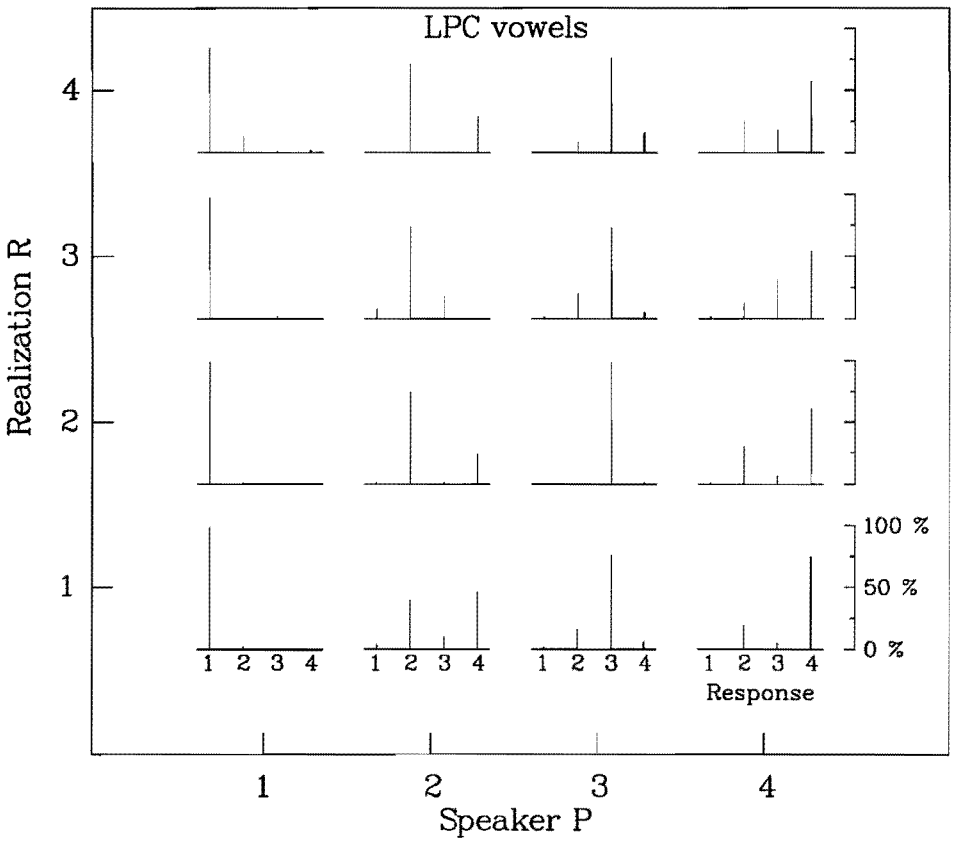

Figure 4.13: Speaker-identification scores for the LPC stimuli. This figure shows 16 panels each corresponding to one the 16 LPC vowels used in the additional experiment. Each panel shows which percentage of the total number of responses for that particular stimulus is attributed to each of the 4 possible response categories (speakers 1, 2, 3, 4, respectively).

some assumptions about how humans produce voiced speech sounds (see section 4.2.3). Under these assumptions we could separate the speech signal into a voice-source waveform which reflects phonatory aspects, and a vocal-tract transfer function which represents articulatory

characteristics. If the assumptions are not valid for the speech stimuli used, this degrades the quality of the source-filter separation, and, as a consequence, it might have an influence on the experimental results.

We determined a closed-glottis phase for all vowels used in this study. On this closed interval we determined the formants and bandwidths of an all-pole filter representing the vocal tract. The quality of the automatically derived vocal-tract functions was checked by visual inspection of the inverse filtering results. As a criterium for a good vocal-tract function we looked for minimum ripple in the closed phase of the inverse filtering result. After the vocal-tract functions were derived, the LF model was fitted to the inverse filter output. In general, we found that the model of voiced speech sounds, which was adopted in this study, and which was described in section 4.2.3, could be fitted very well to the isolated sustained vowels /a/ used in the experiments.

One aspect of voiced speech which was not modeled by the LF-model of the human voice source is the presence of noise. Recently, there have been attempts to incorporate a noise component in voice-source models (Klatt & Klatt, 1990; Childers & Lee, 1991; Granström, 1991). It has been reported that this noise component is needed to synthesize breathy voice qualities (Klatt & Klatt, 1990). We think that in this study, which only involved speakers with a normal voice quality, the ommission of a noise component in the source model does not have severe consequences for the generalizability of the results. A serious drawback for the incorporation of a noise component in a voice-source model is the fact that, at this moment, there are no analysis methods available to estimate the parameters of the noise source from the speech wave. Also, when a noise source is incorporated in a voice-source model, one can encounter the problem that the harmonic components and the noise components do not perceptually integrate, but heard as coming from different sound sources (Hermes, 1991).

We want to compare the automatic methods we used to analyze and synthesize voiced speech sounds to methods which involve manual adjustments. It has been shown that methods of the last category can provide parameters values which can be used to generate synthetic speech which is almost indistinguishable from natural speech (Holmes, 1973; Klatt & Klatt, 1990). On the other hand, it sometimes can be difficult to reproduce the analysis results of manual methods. Analysis

performed by different researchers may yield different results. Also, manual methods are time consuming, so that only limited amounts of speech data can be analyzed. These drawbacks can be removed by using automatic methods. However, at this moment, automatic methods also have their drawbacks. The methods which were applied in this study were only tested on a very small class of speech sounds: sustained isolated vowels /a/. Although for this class of speech stimuli the methods performed well, problems might be expected for other speech sounds, like nasals and voiced fricatives.

4.4.2 Perceptual cues for speaker identity

The main conclusions reported in the literature on experiments using hybrid stimuli state that speaker identification is predominantly determined by vocal-tract characteristics (Miller, 1964; Matsumoto *et al.*, 1973; Carrell, 1984), although Carrell found some evidence that glottal waveforms can indeed carry information which can be used by listeners to identify talkers. However, this finding of Carrell was not confirmed in his experiments using hybrid stimuli. Our results clearly show that besides vocal-tract information listeners also use voice-source information to identify speakers. For instance, figure 4.10 shows that, for most panels, the response alternatives corresponding to either the speaker who's voice-source function was used to synthesize the hybrid vowel, or the speaker who's vocal-tract function was applied, are favoured over the other response alternatives. The perceptual importance of voice-source information is also demonstrated in the panels of figure 4.12 which belong to voice source 2, voice source 3, and voice source 4, respectively. A more detailed inspection of the experimental differences between our study and the ones reported on in literature may explain part of the contradictory findings.

Miller (1964) conducted four experiments to determine the contribution of vocal-tract and voice-source characteristics to speaker identifiability (for a description of Miller's experiments, see 4.3.1). In particular, the last experiment, in which listeners had to identify hybrid stimuli, can be compared with our experiment. The main differences between this experiment[5] and our experiment concern the dynamic variations

[5]The experiments of Miller were originally published only as an abstract in the Jour-

of speech attributes within a vowel, the number of stimuli presented, and the task listeners had to perform. Miller's stimuli consisted of many repetitions of a single, 10-ms fundamental period (Hecker, 1971). In this way, small dynamic variations of vocal-tract and voice-source characteristics which occur in natural vowels were not taken into account. Our stimuli captured these natural vowel irregularities much better, as the speech samples were generated by means of a pitch-synchronous analysis/resynthesis technique. Miller used a relatively small number of 4 stimuli (two speakers both produced one sustained isolated vowel /a/, from which two hybrids were constructed), whereas we used a total of 28 different stimuli (16 resynthesized vowels and 12 hybrids). This means that in our case the stimulus set contained a much greater variation of vocal-tract and voice-source characteristics, making the identification task inherently more difficult. Miller presented the stimuli to the listeners in a two-choice identification test (Hecker, 1971). Every time the listeners had to make an identification they could compare the test stimulus with the two natural reference samples. In this paradigm listeners do not have to rely on long-time memory as they can make direct comparisons of speech samples each time a test stimulus is presented. In fact, listeners have to decide which reference sample is most similar to the test sample. This means that listeners estimate the differences between the test sample and the two reference samples, and choose the reference sample which corresponds to the smallest difference. If this strategy is used by listeners, it is not clear what criteria they use to estimate the differences between the speech samples. These criteria could have been related to the identity of the speaker, but they could have been related equally well to other information contained in the speech signal, like, for instance, phonetic quality. From this argument we can raise the questions whether Miller really determined the contributions of vocal-tract and voice-source characteristics to speaker identity, or whether Miller estimated to what extent vocal-tract and voice-source characteristics can be used by listeners to just discriminate between speech samples. In this study we tried to create experimental conditions for which it was favourable for the listeners to use speaker-dependent features of the speech signal to perform the

nal of the Acoustical Society (Miller, 1964). Unfortunately, this abstract contains not much information about the experimental details. However, more information about the experiments was provided in Hecker (1971).

identification task, instead of using phonetic information. First, listeners had to rely on their long-term memory to perform the identification task as no reference samples were presented. Second, each of the four speakers produced four different vowel realizations which resulted in a stimulus set containing sixteen different vowels. Third, the resynthesized stimuli and the hybrids were mixed during presentation. These three conditions made it difficult for the listeners to use phonetic information instead of speaker characteristics to perform the identification task.

Matsumoto *et al.* (1973) also found that the relative contribution of the vocal-tract characteristics is greater than the voice-source characteristics. Like Miller, Matsumoto et al. also used hybrid stimuli which consisted of many repetitions of one pitch period. Instead of an identification task, subjects had to perform a same-different experiment (see 4.3.1). The conclusions of Matsumoto et al. were based on distances between the stimuli as they were measured in a three-dimensional space representing personal quality. By representing the data in this way, it is difficult to compare distances between stimuli. Also, for the same-different paradigm one can raise the question to what extent the listeners were actually discriminating on the basis of speaker-dependent characteristics of the speech signal and to what extent just differences in sound quality.

As far as the experimental procedure is concerned, the fourth experiment of Carrell (1984) (see 4.3.1) is almost identical to our experiment. The main difference is that, instead of sustained isolated vowels, Carrell used monosyllabic words. His stimulus set consisted of three different words ("dish", "bar", "fuss"), spoken by four speakers (two male, two female), and synthesized at five different fundamental frequencies, yielding a total of 240 different stimuli (3 words, 4 voice-source waveforms, 4 vocal-tract functions, 5 fundamental frequencies). Carrell found some unexpected interactions between glottal waveform and formant spacing. The stimuli which were generated using the voice-source and vocal-tract function of the same speaker did not yield the highest correct formant-based identification scores. The same was found for the glottal-based identification scores. These findings may indicate that the analysis and synthesis techniques used by Carrell did not fully capture the speaker-specific characteristics of the speech samples, or

even worse, that the techniques introduced artifacts. We could not reproduce the phenomena observed by Carrell. If we look at figure 4.11 and figure 4.12 we see, as one might expect, that stimuli which share the voice-source and vocal-tract function of the same speaker have the highest percentage correct identification scores[6].

In the discussion above we have indicated in what way earlier studies using hybrid speech stimuli differed from the current study. We think that these differences can account for the fact that those earlier studies found vocal-tract characteristics to dominate speaker identification by listening whereas our results clearly show that voice-source information also contributes to speaker identity. We will now discuss to what extent the speech stimuli used in this study restrict the generalizability of the results.

Sustained isolated vowels belong to the most simple speech utterances which can be produced by humans. Although they can occur as exclamations, fillers, or words, they are far removed from connected speech (Repp & Crowder, 1990). This means that the general applicability of the results obtained from experiments using isolated vowels is restricted. This is, for instance, demonstrated by Bricker & Pruzansky (1966), and LaRiviere (1975) who showed that with respect to speaker identity different vowels generate different confusion patterns. In our case, we wanted to find out whether coded voice-source information is used by listeners to identify speakers by their voice. As far as we know, this has, until now, not been demonstrated by means of a formal perception experiment. Therefore, it seemed appropriate for us to use a "simple" sustained isolated vowel as a starting point for investigating the perceptual importance of voice-source information for speaker identity. Now we have found that listeners indeed use voice-source information to identify speakers by their voice, it would be interesting to conduct future research to see whether the results hold for other isolated vowels, monosyllabic words, or sentences.

It can be argued that hybrid stimuli represent speech utterances which might not be encountered in daily life, because it is physiologically impossible to interchange either the vocal tract or the larynxes

[6]The only exception is the percentage correct formant-based identification score of stimulus S_1T_2. However, this scores does not differ significantly from the score of stimulus S_2T_2.

of two speakers (Hecker, 1971). This is reflected by the fact that it is not possible to define whether the response of a subject is correct or incorrect. We therefore adopted the formant-based and glottal-based accuracy measures introduced by Carrell (1984) (for an explanation of these measures, see 4.3.5). If one assumes that the separation of voice-source and vocal-tract information is perfect, which means that the voice-source waveform only captures phonatory information whereas the vocal-tract transfer function only models articulatory information, we think that hybrid speech samples provide a means to investigate the contributions of voice-source and vocal-tract characteristics to speaker identity in a direct manner.

The speech stimuli in this study were all produced by male speakers. This means that we cannot extrapolate our findings to female speakers. However, speech synthesis experiments by Klatt & Klatt (1990) and Karlsson (1991) showed that a model of the voice source is needed to synthesize a convincing female voice.

Below we compare the results of the acoustic analysis with the informal comments made by the subjects after the last experimental run. Although we did not systematically investigate how the subjective cues reported by the listeners relate to physical parameters of the speech synthesizer, inspection of the acoustic and experimental data makes it possible to support the remarks made by the subjects in a more objective way.

Subjects reported that they recognized speaker 1 if a vowel gave the impression of having been produced with a wide open mouth. Phoneticians use the term "openness" to describe a perceptual dimension referring to the degree of mouth opening. It has been shown by Traunmüller (1981) that the frequency distance between the first formant F_1 and the fundamental frequency F_0 is decisive for perceived openness. If the distance between F_1 and F_0 increases a higher degree of openness is perceived. From Table 4.II it can be seen that the frequency distance between F_1 and F_0 is largest for speaker 1.

Speaker 2 was identified as being the one who produced vowels which had an LPC-like quality. From Table 4.I we can see that, in general, the speech waveform perturbation measures, jitter, σ_{DF_0}, and shimmer are smallest for speaker 2. As it is known from speech synthesis experiments

that the absence of jitter and shimmer might result in a mechanical sound quality (Klatt & Klatt, 1990), one could expect speaker 2 to have the most mechanical sounding voice quality. However, all jitter and shimmer values of Table 4.I are below the detectability thresholds for jitter and shimmer, as they were reported by Klatt & Klatt (1990)[7]. It should be noted that our measures are means, and that we did not look at local significant changes which might have been perceived by the subjects. The normalized return-time parameter R_a has its lowest value for speaker 2. The effect of R_a on the voice-source spectrum is an additional high-frequency deemphasis. The lower the R_a value, the less the high-frequency slope of the voice-source spectrum deviates from the the high-frequency slope of the LPC excitation (deemphasized delta-pulse). This means that the high-frequency slope of the voice-source spectrum of speaker 2 comes closest to its LPC equivalent.

Subjects reported that speaker 3 produced vowels which had an unstable pitch. Besides the fact that the waveform perturbation measures of Table 4.I are below the detection thresholds reported by Klatt & Klatt (1990), we also see that the jitter and σ_{DF_0} values for speaker 3 do not differ significantly from the values measured for the other speakers. The F_0 variations observed in the vowels of speaker 3 much more resemble the slow quasirandom drift of the F_0 contour as described by Klatt & Klatt (1990). Klatt & Klatt (1990) use the term "flutter" to describe these slow variations of F_0[8]. The flutter present in the F_0 contours of speaker 3 is reflected in Table 4.I by the larger values of the standard error and the range of F_0.

Listeners reported that the vowels of speaker 4 had a certain "bass" quality. Two parameters of Table 4.I support this subjective impression. Speaker 4 clearly has the largest open quotient OQ. This means that the amplitude of the first harmonic relative to adjacent harmonics is largest for the voice-source spectrum of speaker 4 (see section 4.3.5). The closing quotient CQ of speaker 4 is much higher than the parameter values for the other speakers. As explained in section 4.3.5 this

[7] A threshold for jitter of about 2% and for shimmer of about 1 dB (Klatt & Klatt, 1990).

[8] We implemented the synthesis strategy described by Klatt & Klatt (1990) to add flutter to a synthetic vowel and found that the resulting pitch instability was perceptually very similar to the F_0 variations of the vowels produced by speaker 3.

parameter determines the transient character of the glottal pulse. A higher value of CQ corresponds to a lesser high-frequency content of the voice-source spectrum.

In the previous paragraphs we argued that the subjective cues for speaker 1 relate to spectral characteristics of the vocal tract, whereas the cues reported for speakers 2, 3, and 4 relate to voice-source characteristics. This observation is supported by the experimental data (section 4.3.5). It was found that speaker 1 is dominated by his vocal tract, whereas speakers 3 and 4 mostly by their source. For speaker 2, both tract and source seems to be important. This means that the results fit the subjective cue descriptions of the speakers quite well.

From the discussion above we conclude that for our experiment the subjective cues reported by the listeners can be related to physical aspects of the stimuli, and are supported by the experimental results. However, it should be stressed that the acoustical correlates of speaker identity, as they are indicated by our data, are only based on comparisons between informally reported speaker identification cues and data resulting from acoustic analysis. Further research should be conducted to determine the exact relationships between the physical parameters of the speech production model and the subjective attributes of speaker identity. The results that were found in the current study may serve as a possible starting point for such research.

4.5 Conclusions

Automatic analysis/resynthesis techniques which were used in this study can capture speaker-specific features of the speech signal. These techniques made it possible to generate so-called hybrid speech samples for which the voice-source characteristics of one speaker are combined with the vocal-tract characteristics of another speaker. An important point is that the source information used in this experiment was coded in a model of the voice source. This entails that our experimental results are of importance for speech synthesis and text-to-speech systems. We conclude from our experiments that there is no general rule that vocal-tract information contributes more to perceived speaker identity than voice-source information. Sometimes vocal-tract information is

more important, sometimes voice-source information. The results fit the subjective cues which listeners reported they had been using to perform the speaker identification task quite well. We showed that is also possible to relate physical aspects of the speech stimuli to the subjective cue descriptions.

Chapter 5

Concluding remarks

IN the introductory chapter of this dissertation we formulated two research aims. One was the evaluation of the quality of LPC as a scheme for speech analysis, manipulation and synthesis, the other was the exploration of ways to improve this quality. In this final chapter we discuss how these aims were dealt with. We start with a discussion on the limitations of the present research. Next, we discuss our contributions to the improvement of schemes for speech analysis and synthesis.

5.1 Limitations of the present research

The conversion from text to speech involves many steps. This means that if we want to evaluate the quality of a TTS system we can focus our attention on many different aspects of the text-to-speech conversion process (Pols, 1988; Van Bezooijen & Pols, 1990). In the present research we concentrated on that part of the TTS system which is actually generating the output speech of the system, the speech-coding algorithm, or the speech synthesizer. In chapter 1 of this dissertation we called this the "instrument", as opposed to the "performer" which, in a TTS system, is modeled by a set of rules. Even if we try to assess the quality of only one component of a TTS system, we can still evaluate different speech quality attributes of this component. In this research we made the term speech quality operational by studying the following attributes: intelligibility, naturalness, and speaker identity. Although these attributes are generally accepted as being important aspects of speech quality, they certainly do not cover the term speech quality completely. This fact, and the fact that we restricted ourselves

to the evaluation of the speech "instrument", limits the impact of our experimental results for TTS synthesis in general. Below we will discuss the limitations of our approach in more detail.

In chapter 2 we studied the intelligibility of synthetic speech produced by different speech-coding schemes. The speech stimuli which were presented to the listeners were resynthesized versions of natural speech. By this approach we were able to concentrate on the speech synthesis algorithm. However, in a TTS system other components of the system also contribute to the overall intelligibility. Pisoni, Nusbaum & Greene (1985) distinguish three areas which are important with respect to the intelligibility of synthetic speech produced by a TTS system: 1. rules for the letter-to-sound conversion, 2. rules for the generation of prosodic information, 3. rules that convert the internal representation of basic speech fragments into a speech waveform. By using resynthesis instead of TTS synthesis, we only addressed part of the third area. We evaluated how well different speech-coding algorithms captured those aspects of the speech signal which are important with respect to intelligibility. Knowledge obtained by this evaluation study is nevertheless important for the development of high-quality TTS systems: if a speech-coding algorithm can not produce intelligible speech, in most cases, it is not possible to improve intelligibility by means of better rules for letter-to-sound conversion or prosody generation.

In practice, the perception of synthetic speech generated by a TTS system is also influenced by factors which do not directly relate to the text-to-speech conversion process itself (Pisoni *et al.*, 1985). For instance, the conditions under which the speech is perceived can seriously affect the intelligibility of synthetic speech. In chapter 2 we demonstrated that the intelligibility of synthetic speech in a noisy environment is degraded in different ways for different speech-coding algorithms. It would be interesting to investigate whether this is also true for situations in which people are engaged in other tasks which require attention. Other factors which may influence the perception of synthetic speech are the degree to which listeners have previously been exposed to synthetic speech, and the structure of the speech material. Our subjects were quite familiar with the quality of the synthetic speech. The monosyllabic words we used in our test do not predict the intelligibility of sentences, passages, or fluent continuous speech. Nev-

ertheless, the semantic and syntactic information which is present in these cases will only increase the overall intelligibility of the synthetic speech.

In chapter 3 we studied the naturalness of synthetic speech. This attribute is a kind of catch-all term which is used with respect to many different aspects of the TTS conversion process. For instance, people talk about a natural intonation contour, a natural speech rhythm, or a natural voice. In chapter 3 we chose the experimental conditions in such a way that naturalness was not influenced by factors like intelligibility, loudness, and speaker identity. Under these conditions naturalness can be interpreted as the overall speech quality. Like in chapter 2, we used speech resynthesis techniques to concentrate on the speech-coding algorithm. As a consequence, the same remarks can be made with respect to the importance of rules for the letter-to-sound conversion and the synthesis of prosodic information. They all contribute to the naturalness of synthetic speech, and therefore they should be studied to develop TTS systems which can produce high-quality natural-sounding speech.

The speech-coding schemes which are currently used for TTS synthesis are deficient in preserving speaker identity and speaker characteristics. In a pilot study we found that it was difficult, and sometimes even impossible, for listeners to recognize familiar speakers by their resynthesized speech (Eggen, 1987; Eggen & Vogten, 1990). We concluded that the proper reproduction of the prosodic structure of natural speech by means of LPC resynthesis is not sufficient to preserve reliably the identity of different speakers. Obviously, other aspects of the speech signal are also important for perceived speaker identity. One of those aspects is the way in which a speech-coding algorithm codes speaker specific information contained in the speech signal. In chapter 4 we included a more detailed model of the human voice source in the LPC synthesizer. By means of a perception experiment we determined whether listeners use information coded by the source model to identify speakers. We found that listeners use both coded voice-source and vocal-tract information to perform the identification task. The results of chapter 4 show that speech-coding algorithms can code speaker characteristics of the speech signal. However, at present we are not able to estimate the relative importance of the ability of the speech-coding

algorithm to code speaker specific information and other suprasegmental speaker characteristics of the speech signal. On the one hand, we do not know how well the glottal-excited LPC scheme can synthesize different voice qualities. On the other hand, it has been known for a long time that speakers can be recognized, and even sometimes convincingly imitated, by their speaking style (Hecker, 1971; Laver, 1980, Nolan, 1983). Future research is needed to better understand how different speakers, consciously or unconsciously, use intonation, rhythm, the insertion of pauses, voice quality variations, and possibly many other speech attributes to personalize their speech. It is interesting to note that in much of the recent research on speaker characteristics speech synthesizers are used which model the human voice source in more detail (Childers, Wu, Hicks & Yegnanarayana, 1989; Gobl, 1989; Cummings & Clements, 1990; Klatt & Klatt, 1990; Carlson, Granström & Karlsson, 1991; Childers & Lee, 1991; Granström, 1991).

From the discussion above it is clear that the present research has its limitations with respect to the overall quality of TTS synthesis. Nevertheless, we have also seen that improvements of existing speech-coding schemes can increase the quality of synthetic speech significantly. We even think that many of the speech quality aspects which were not addressed in this study can only be investigated properly if we have better schemes for the analysis, manipulation and synthesis of speech. The present research tries to indicate ways for improving these tools. In the next section we look in more detail at the particular improvements which were explored in this dissertation.

5.2 Improvements

In chapter 3 we investigated the perceptual correlates of spectral characteristics of voiced speech. In particular we wanted to determine the relative importance of amplitude and phase information contained in the LPC residue. We thought that this knowledge could help us to come up with a solution which could improve the speech quality of a traditional LPC speech synthesizer. As argued in chapter 3, the results of a perception experiment can be greatly influenced by the way in which we process the speech signal. Therefore, we developed a new analysis-resynthesis system which is based on a model for human speech

production. This system enables us to interpret the concepts of amplitude and phase in terms of a speech synthesis filter. As a consequence, the experimental results of chapter 3 are directly applicable to speech synthesis.

Based on the results of the study on the relative importance of amplitude and phase, we decided to implement a glottal-excited LPC synthesizer. The idea was that a more detailed model of the human voice source was a good way to capture some of the amplitude and phase information contained in the LPC residue. In chapter 4 we developed automatic procedures for the estimation of the parameters of the voice-source model from the speech signal. In order to determine the voice-source and vocal-tract parameters we had to make assumptions about the speech production model. For instance, we had to assume that the speech production model can be approximated by a linear system, that the glottis is closed during a certain part of a pitch period, and that the vocal-tract can be modeled as an all-pole filter. Only under these assumptions we were able to derive the voice-source waveform from the speech signal and to estimate the parameters of the voice-source model. This means that, like in chapter 3, the experimental results depend on speech processing algorithms which were applied to the speech signal.

Recently, there has been a renewed interest in research on the human voice source. It is believed that the next generation of TTS systems can only provide means for synthesizing different voice qualities and speaking styles, if the applied speech synthesis techniques incorporate a more detailed model of the human voice source (Carlson *et al.*, 1991). Currently, there is a lack of knowledge on how to adjust the parameters of such a voice-source model to synthesize the wide range of voice variations encountered in natural speech. Although data on speaker variability is now being accumulated (see for instance, Carlson *et al.*, 1991), much more data is needed for the development of rules which control the voice-source model of a TTS synthesizer. In this view, reliable automatic techniques for voice-source parameter estimation, like the ones we developed in chapter 4, could be of great help.

With respect to automatic methods for the derivation of voice-source parameters from real speech, we would like to make two final remarks. Firstly, we think that it is important that speech researchers from different laboratories agree on what is the best way to separate voice-source

and vocal-tract characteristics from the speech signal. Only then it is possible to compare and combine voice-source data measured by different researchers. Secondly, we would like to stress the important role perception experiments can play in the development of good voice-source models for the synthesis of voice variations in TTS systems. By means of perception experiments we can determine the importance of the various spectral and waveform details of the human voice-source signal. Voice-source models should only capture those details which are perceptually relevant for the generation of different voice qualities and speaking styles.

References

Ananthapadmanabha, T.V. & Yegnanarayana, B. (**1979**). "Epoch extraction from linear prediction residual for identification of closed glottis interval", IEEE Trans. Acoust., Speech, Signal Processing **27**, 309–319.

Askenfelt, A.G. & Hammarberg, B. (**1986**). "Speech waveform perturbation analysis: a perceptual–acoustical comparison of seven measures", J. Speech Hearing Res. **29**, 50–64.

Atal, B.S. & Hanauer, S.L. (**1971**). "Speech analysis and synthesis by linear prediction of the speech wave", J. Acoust. Soc. Am. **50**, 637–655.

Atal, B.S. (**1976**). "Automatic recognition of speakers from their voices", Proc. IEEE **64**, 460–475.

Atal, B.S. & David, N. (**1979**). "On synthesizing natural-sounding speech by linear prediction", Proc. Int. Conf. Acoust. Speech Signal Process., 44–47.

Atal, B.S. & Remde, J.R. (**1982**). "A new model of LPC excitation for producing natural-sounding speech at low bit rates", Proc. Int. Conf. Acoust. Speech Signal Process., 614–617.

Atal, B.S. (**1983**). "Efficient coding of LPC parameters by temporal decomposition", Proc. Int. Conf. Acoust. Speech Signal Process., 81–84.

Benoit, C. (**1990**). "An intelligibility test using semantically unpredictable sentences: towards the quantification of linguistic complexity", Speech Communication **9**, 293–304.

Bode, D. & Carhart, R. (**1973**). "Measurements of articulation functions using adaptive test procedures", IEEE Trans. Audio Electroacoust. **AU-21**, 196–201.

Brady, P.T. (**1968**). "Equivalent Peak Level : a threshold-independent speech-level measure", J. Acoust. Soc. Am. **44**, 695–699.

Bricker, P.D. & Pruzansky, S. (**1966**). "Effects of stimulus content and duration on talker identification", J. Acoust. Soc. Am. **40**, 1441–1449.

Carhart, R., Johnson, C. & Goodman, J. (**1975**). "Perceptual masking of spondees by combinations of talkers", J. Acoust. Soc. Am. **58**, S35(A).

Carlson, R., Granström, B. & Klatt, D.H. (**1979**). "Vowel perception: the relative perceptual salience of selected acoustic manipulations", Speech Trans. Lab. Q. Prog. Stat. Rep. **3–4**, Royal Institute of Technology, Stockholm, 73–83.

Carlson, R., Galyas, K., Granström, B., Pettersson, M. & Zachrisson, G. (**1980**). "Speech synthesis for the non vocal in training and communication", Speech Trans. Lab. Q. Prog. Stat. Rep. **1**, Royal Institute of Technology, Stockholm, 13–27.

Carlson, R., Fant, G., Gobl, C., Granström, B., Karlsson, I. & Lin, Q. (**1989**). "Voice source rules for text-to-speech synthesis", Proc. Int. Conf. Acoust. Speech Signal Process., 223–226.

Carlson, R., Granström, B. & Karlsson, I. (**1991**). "Experiments with voice modelling in speech synthesis", Speech Communication **10**, 481–489.

Carlson, R., Granström, B. & Nord, L. (**1992**). "Segmental evaluation using the Esprit/SAM test procedures and mon-syllabic words", in *Talking Machines: Theories, Models, and Designs*, edited by G. Bailly, C. Benoit, and T.R. Sawallis (North Holland Elsevier Science Publishers B.V., Amsterdam), 443–453.

Carrell, T.D. (**1984**). "Contributions of fundamental frequency, formant spacing, and glottal waveform to talker identification", Speech Res. Lab. Prog. Rep. **5**, Indiana Univ., Bloomington, IN.

Caspers, B. & Atal, B.S. (**1987**). "Role of multi-pulse excitation in synthesis of natural-sounding voiced speech", Proc. Int. Conf. Acoust. Speech Signal Process., 2388–2391.

Charpentier, F. & Moulines, E. (**1989**). "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using di-

phones", Proc. Eurospeech, 13–19.

Cheng, Y.M. & O'Shaughnessy, D. (**1989**). "Automatic and reliable estimation of glottal closure instant and period", IEEE Trans. Acoust., Speech, Signal Processing **37**, 1805–1815.

Childers, D.G., Wu, K., Hicks, D.M. & Yegnanarayana, B. (**1989**). "Voice Conversion", Speech Communication **8**, 147–158.

Childers, D.G. & Wu, K. (**1990**). "Quality of speech produced by analysis-synthesis", Speech Communication **9**, 97–117.

Childers, D.G. & Lee, C.K. (**1991**). "Vocal quality factors: analysis, synthesis and perception", J. Acoust. Soc. Am. **90**, 2394–2410.

Cochran, W.G. (**1947**). "Some consequences when assumptions for the analysis of variance are not satisfied", Biometrics **3**, 22–38.

Cohen, A., Ebeling, C.L., Fokkema, K. & Van Holk, A.G.F. (**1961**). *Fonologie van het Nederlands en het Fries: inleiding tot de moderne klankleer* (Martinus Nijhoff, The Hague).

Cranen, L.I.J. (**1987**). *The acoustic impedance of the glottis. Measurements and modeling* (Doctoral thesis, Nijmegen Catholic University, Sneldruk, Enschede).

Cummings, K.E. & Clements, M.A. (**1990**). "Analysis of glottal waveforms across stress styles", Proc. Int. Conf. Acoust. Speech Signal Process., 369–372.

David, H.A. (**1963**). *The method of paired comparisons* (Charles Griffin & Company Limited, London).

Dologlou, I. & Carayannis, G. (**1989**). "Pitch detection based on zero-phase filtering", Speech Communication **8**, 309–318.

De Veth, J., Van Golstein-Brouwers, W.G., Boves, L. & Van Heugten, L.J.P. (**1991**). "Research on the human speech production system by means of robust ARMA-analysis (partly in Dutch)", SPIN/ASSP report **38**, Speech Technology Foundation, Utrecht.

Doddington, G.R. (**1985**). "Speaker recognition - identifying people by their voices", Proc. IEEE **73**, 1651–1664.

Eggen, J.H. (**1987**). "Start evaluation of available methods for analysis, manipulation, and resynthesis of speech", SPIN/ASSP report **2**, Speech Technology Foundation, Utrecht.

Eggen, J.H. (**1989a**). "Intelligibility of synthetic speech in the presence of interfering speech", Speech Communication **8**, 319–327.

Eggen, J.H. (**1989b**). "A glottal-excited speech synthesizer", Ann. Prog. Rep. **24**, Institute for Perception Research, Eindhoven, 25–32.

Eggen, J.H. & Vogten, L.L.M. (**1990**). "Degradation of speaker identification for LPC fromant-coded speech", J. Acoust. Soc. Am. **87**, S109.

Fant, G. (**1960**). *Acoustic theory of speech production* (Mouton, The Hague).

Fant, G. (**1979**). "Glottal source and excitation analysis", Speech Trans. Lab. Q. Prog. Stat. Rep. **1**, Royal Institute of Technology, Stockholm, 85–107.

Fant, G., Liljencrants, J. & Lin, Q. (**1985**). "A four-parameter model of glottal flow", Speech Trans. Lab. Q. Prog. Stat. Rep. **4**, Royal Institute of Technology, Stockholm, 1–13.

Fant, G. & Lin, Q. (**1987**). "Glottal source - vocal tract interaction" Speech Trans. Lab. Q. Prog. Stat. Rep. **1**, Royal Institute of Technology, Stockholm, 13–27.

Fant, G., Liljencrants, J. & Lin, Q. (**1988**). "Frequency domain interpretation and derivation of glottal flow parameters", Speech Trans. Lab. Q. Prog. Stat. Rep. **2–3**, Royal Institute of Technology, Stockholm, 1–21.

Feustel, T.C., Logan, R.J. & Velius, G.A. (**1988**). "Human and machine performance on speaker identity verification", J. Acoust. Soc. Am. **83**, S55.

Flanagan, J.L. (**1972**). *Speech Analysis Synthesis and Perception* (Second Edition, Springer-Verlag, Berlin).

Fourcin, A.J., Harland, G., Barry, W. & Hazan, V. (**1989**). *Speech input and output assessment – multilingual methods and standards* (Ellis Horwood Limited, Chichester).

French, N.R. & Steinberg, J.C. (**1947**). "Factors governing the intelligibility of speech sounds", J. Acoust. Soc. Am. **19**, 90–119.

Fujisaki, H. & Ljungqvist, M. (**1986**). "Proposal and evaluation of models for the glottal source waveform", Proc. Int. Conf. Acoust. Speech Signal Process., 1605–1608.

Funada, T. (**1989**). "A new method for extracting the glottal closure intervals of voiced speech", J. Acoust. Soc. Jpn. (E) **10**, 349–355.

Gautherot, O., Mason, J.S. & Corney, P. (**1989**). "LPC residual phase investigation", Proc. Eurospeech, 35–38.

Gobl, C. (**1988**). "Voice source dynamics in connected speech", Speech Trans. Lab. Q. Prog. Stat. Rep. **1**, Royal Institute of Technology, Stockholm, 123-159.

Gobl, C. & Ní Chasaide, A. (**1988**). "The effects of adjacent voiced/voiceless consonants on the vowel voice source: a cross language study", Speech Trans. Lab. Q. Prog. Stat. Rep. **2–3**, Royal Institute of Technology, Stockholm, 23–59.

Gobl, C. (**1989**). "A preliminary study of acoustic voice quality correlates", Speech Trans. Lab. Q. Prog. Stat. Rep. **4**, Royal Institute of Technology, Stockholm, 9–22.

Granström, B. (**1991**). "The use of speech synthesis in exploring different speaking styles", Speech Trans. Lab. Q. Prog. Stat. Rep. **2–3**, Royal Institute of Technology, Stockholm, 1–10.

Griffin, D.W. & Lim, J.S. (**1984**). "Signal estimation from modified short-time Fourier transform", IEEE Trans. Acoust., Speech, Signal Processing **32**, 236–243.

Gupta, S.K. & Atal, B.S. (**1991**). "Efficient frequency-domain representation of LPC-excitation", IEEE Workshop on Speech Coding for Telecommunications Digital Voice for the Nineties, 64–65.

Hamon, C., Moulines, E. & Charpentier, F. (**1989**). "A diphone synthesis system based on time-domain prosodic modifications of speech", Proc. Int. Conf. Acoust. Speech Signal Process., 238–241.

Hays, W.L. (**1988**). *Statistics* (Holt, Rinehart and Winston, Inc., New York).

Hecker, M.H.L. (**1971**). *Speaker recognition: an interpretive survey of the literature* (American Speech and Hearing Association Monographs **16**).

Hedelin, P. (**1984**). "A glottal LPC-vocoder", Proc. Int. Conf. Acoust. Speech Signal Process., 1.6.1–1.6.4.

Hedelin, P. (**1986**). "High quality glottal LPC-vocoding", Proc. Int. Conf. Acoust. Speech Signal Process., 465–468.

Hermes, D.J. (**1988**). "Measurement of pitch by subharmonic summation", J. Acoust. Soc. Am. **83**, 257–264.

Hermes, D.J. (**1991**). "Synthesis of breathy vowels: some research methods", Speech Communication **10**, 497–502.

Holmes, J.N. (**1973**). "The influence of glottal waveform on the naturalness of speech from a parallel formant synthesizer", IEEE Trans. Audio Electroacoust. **AU-21**, 298–305.

Horii, Y. (**1985**). "Jitter and shimmer in sustained vocal fry phonation", Folia phoniat. **37**, 81–86.

House, A.S., Williams, C.E., Hecker, M.H.L. & Kryter, K.D. (**1965**). "Articulation-testing methods: consonantal differentiation with a closed-response set", J. Acoust. Soc. Am. **37**, 158–166.

Hunt, M.J. (**1978**). "Automatic correction of low-frequency phase distortion in analogue magnetic recordings", Acoustic Letters **2**, 6–10.

Hunt, M.J., Zwierzynski, D.A. & Carr, R.C. (**1989**). "Issues in high quality LPC analysis and synthesis", Proc. Eurospeech, 348–351.

Imaizumi, S., Kiritani, S., Fukawa, H. & Saito, S. (**1989**). "A polynomial glottal source model for the synthesis of various voice qualities", Ann. Bull. University of Tokyo **23**, 109–118.

Ishizaka, K. & Flanagan, J.L. (**1972**). "Synthesis of voiced sounds from a two mass model of the vocal cords", Bell Syst. Tech. J. **51**, 1233–1268.

Kalikow, D.N. & Stevens, K.N. (**1977**). "Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability", J. Acoust. Soc. Am. **61**, 1337–1351.

Karlsson, I. (**1988**). "Glottal waveform parameters for different speaker types", Speech Trans. Lab. Q. Prog. Stat. Rep. **2–3**, Royal Institute of Technology, Stockholm, 61–67.

Karlsson, I. (**1989**). "A female voice for a text-to-speech system", Proc. Eurospeech, 349–352.

Karlsson, I. (**1991**). "Female voices in speech synthesis", J. Phonetics **19**, 111–120.

Klatt, D.H. (**1980**). "Software for a cascade/parallel formant synthesizer", J. Acoust. Soc. Am. **67**, 971–995.

Klatt, D.H. (**1987**). "Review of text-to-speech conversion for English", J. Acoust. Soc. Am. **82**, 737–793.

Klatt, D.H. & Klatt, L.C. (**1990**). "Analysis, synthesis and perception of voice quality variations among female and male talkers", J. Acoust. Soc. Am. **87**, 820–857.

Koizumi, T. & Taniguchi, S. (**1988**). "Perceptual significance of glottal source – vocal tract interaction", The Second Symposium on Advanced Man-Machine Interface through Spoken Language, 8-1–8-8.

Krishnamurthy, A.K. & Childers, D.G. (**1986**). "Two-channel speech analysis", IEEE Trans. Acoust., Speech, Signal Processing **34**, 730–743.

Kroon, P. & Deprettere, E.F. (**1988**). "A class of analysis-by-synthesis predictive coders for high quality coding at rates between 4.8 and 16 kbit/s", IEEE J. Selected Areas in Communications **6**, 353–363.

Kuwabara, H. & Ohgushi, K. (**1987**). "Contributions of vocal tract resonant frequencies and bandwidths to the personal perception of speech", Acustica **63**, 120–128.

LaRiviere, C. (**1975**). "Contributions of fundamental frequency and formant frequencies to speaker identification", Phonetica **31**, 185–197.

Laver, J. (**1980**). *The phonetic description of voice quality* (Cambridge University Press, Cambridge).

Lent, K. (**1989**). "An efficient method for pitch shifting digitally sampled sounds", Computer Music J. **13**, 65–71.

LeRoux, J. & Gueguen, C. (**1977**). "A fixed point computation of partial correlation coefficients", IEEE Trans. Acoust., Speech, Signal Processing **25**, 258–259.

Levitt, H. & Rabiner, L. (**1967**). "Use of sequential strategy in intelligibility testing", J. Acoust. Soc. Am. **42**, 609–612.

Levitt, H. (**1971**). "Transformed up-down methods in psychoacoustics", J. Acoust. Soc. Am. **49**, 467–477.

Lin, Q. (**1990**). *Speech production theory and articulatory speech synthesis* (Doctoral thesis, Dept. of Speech Communication & Music Acoustics, Royal Institute of Technology, Stockholm).

Ma, C., Kamp, Y. & Willems, L.F. (**1992**). "A singular value decomposition approach to glottal closure determination from the speech signal", To appear in J. Acoust. Soc. Am.

Mackie, K., Dermody, P. & Katsch, R. (**1987**). "Assessment of evaluation measures for processed speech", Speech Communication **6**, 309–316.

Markel, J.D. & Gray, Jr., A.H. (**1974**). "A linear prediction vocoder simulation based upon the autocorrelation method", IEEE Trans. Acoust.Speech Signal Process. **2**, 124–134.

Markel, J.D. & Gray, Jr., A.H. (**1976**). *Linear Prediction of Speech* (Springer-Verlag, New York).

Mathes, R.C. & Miller, R.L. (**1947**). "Phase effects in monaural phase perception", J. Acoust. Soc. Am. **19**, 780–797.

Matsumoto, H., Hiki, S., Sone, T. & Nimura, T. (**1973**). "Multidimensional representation of personal quality of vowels and its acoustical correlates", IEEE Trans. Audio Electroacoust. **AU-21**, 428–436.

Miller, J.E. & Mathews, M.V. (**1963**). "Investigation of the glottal waveshape by automatic inverse filtering", J. Acoust. Soc. Am. **35**, 1876 (A).

Miller, J.E. (**1964**). "Decapitation and recapitation, a study of voice quality", J. Acoust. Soc. Am. **36**, 2002 (A).

Monsen, R.B & Engebretson, A.M. (**1977**). "Study of variations in the male and female glottal wave", J. Acoust. Soc. Am. **62**, 981–993.

Mosteller, F. (**1951**). "Remarks on the method of paired comparisons: III. A test of significance for paired comparisons when equal standard deviations and equal correlations are assumed", Psychometrica **16**, 207–218.

Moulines, E. & Di Francesco, R. (**1990**). "Detection of the glottal closure by jumps in the statistical properties of the speech signal", Speech Communication **9**, 401-418.

Nakatani, L.H. & Dukes, K.D. (**1973**). "A sensitive test of speech communication quality", J. Acoust. Soc. Am. **53**, 1083–1092.

Nolan, F. (**1983**). *The phonetic basis of speaker recognition,* (Cambridge University Press, Cambridge).

Nord, L., Ananthapadmanabha, T.V. & Fant, G. (**1984**). "Signal analysis and perceptual tests of vowel responses with an interactive source filter model", Speech Trans. Lab. Q. Prog. Stat. Rep. **2–3**, Royal Institute of Technology, Stockholm, 25-52.

Pabon, J.P.H. (**1991**). "Objective acoustic voice-quality parameters in the computer phonetogram", J. Voice **3**, 203.

Patterson, R.D. (**1987**). "A pulse ribbon model of monaural phase perception", J. Acoust. Soc. Am. **82**, 1561–1586.

Peterson, D.B. & Barney, H.L. (**1952**). "Control methods used in a study of the vowels", J. Acoust. Soc. Am. **24**, 175–184.

Philips Export BV, (**1983**). "MEA8000 voice synthesizer: principles and interfacing", Philips Technical Publication **101**.

Phillips, J.P.N. (**1964**). "On the presentation of stimulus objects in the method of paired comparisons", Am. J. Psychology **77**, 660-664.

Pinto, N.B., Childers, D.G. & Lalwani, A.L. (**1989**). "Formant speech synthesis: improving production quality", IEEE Trans. Acoust., Speech, Signal Processing **37**, 1870–1887.

Pisoni, D.B. (**1982**). "Perception of speech: the human listener as a cognitive interface", Speech Technology **1**, 10–23.

Pisoni, D.B. & Koen, E. (**1982**). "Some comparisons of intelligibility of synthetic and natural speech at different speech-to-noise ratios", J. Acoust. Soc. Am. **71**, S94(A).

Pisoni, D.B., Nusbaum, H.C. & Greene, B.G. (**1985**). "Perception of synthetic speech generated by rule", Proc. of the IEEE **73**, 1665-1676.

Plomp, R. & Steeneken, H.J.M. (**1969**). "Effect of phase on the timbre of complex tones", J. Acoust. Soc. Am. **46**, 409–421.

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

Pols, L.C.W. (**1988**). "Improving synthetic speech quality by systematic evaluation", Proc. Second Symposium on Advanced Man-Machine Interface through Spoken Language, Hawaii, 17/1–17/9.

Pols, L.C.W. (**1992**). "Quality assessment of text-to-speech synthesis

by rule", in *Advances in Speech Signal Processing*, edited by S. Furui, and M.M. Sondhi (Marcel Dekker, INC., New York), 387–416.

Pratt, R.L. (**1987**). "Quantifying the performance of text-to-speech synthesizers", Speech Technology **3**, 54–64.

Price, P.J. (**1989**). "Male and female voice source characteristics: inverse filtering results", Speech Communication **8**, 261–277.

Rabiner, L.R. & Schafer, R.W. (**1978**). *Digital Processing of Speech Signals* (Prentice-Hall, Inc., Englewood Cliffs, New Jersey).

Repp, B.H. & Crowder, R.G. (**1990**). "Stimulus order effects in vowel discrimination", J. Acoust. Soc. Am. **88**, 2080–2090.

Rietveld, A.C.M. & Gussenhoven, C. (**1985**). "On the relation between pitch excursion size and prominence", J. Phonetics **13**, 299–308.

Rothauser, E.H., Urbanek, G.E. & Pachl, W.P. (**1968**). "Isopreference method for speech evaluation", J. Acoust. Soc. Am. **44**, 408–418.

Rosenberg, A.E. (**1971**). "Effect of glottal pulse shape on the quality of natural vowels", J. Acoust. Soc. Am. **49**, 583–590.

Sambur, M.R., Rosenberg, A.E., Rabiner, L.R. & McGonegal, C.A. (**1978**). "On reducing the buzz in LPC synthesis", J. Acoust. Soc. Am. **63**, 918–924.

Scheffé, H. (**1952**). "An analysis of variance for paired comparisons", J. Am. Statist. Ass. **47**, 381–400.

Schoentgen, J. (**1990**). "Non-linear signal representation and its applications to the modelling of the glottal waveform", Speech Communication **9**, 189–201.

Schroeder, M.R. & Strube, H.W. (**1986**). "Flat-spectrum speech", J. Acoust. Soc. Am. **79**, 1580–1583.

Siegel, S. (**1956**). *Nonparametric statistics for the behavioral sciences* (McGraw-Hill, New York).

Sondhi, M.M. (**1975**). "Measurement of the glottal waveform", J. Acoust. Soc. Am. **57**, 228–232.

Sondhi, M.M. (**1990**). "Models of speech production for speech analysis and synthesis", J. Acoust. Soc. Am. **87**, S14.

Spiegel, M.F., Altom, M.J. & Macchi, M.J. (**1990**). "Comprehensive assessment of the telephone intelligibility of synthesized and natural speech", Speech Communication **9**, 279–291.

Steeneken, H.J.M. (**1992**). *On measuring and predicting speech intelligibility* (Doctoral thesis, University of Amsterdam).

Strube, H.W. (**1974**). "Determination of the instant of glottal closure from the speech wave", J. Acoust. Soc. Am. **56**, 1625–1629.

Studebaker, G.A. (**1985**). "A "rationalized" arcsine transform", J. Speech Hearing Res. **28**, 455–462.

Tenpaku, S. & Hirahara, T. (**1990**). "A glottal waveform model for high quality speech synthesis", J. Acoust. Soc. Am. **88**, S152.

Titze, I.R. (**1989**). "A four-parameter model of the glottis and vocal fold contact area", Speech Communication **8**, 191–201.

Torgerson, W.S. (**1958**). *Theory and methods of scaling* (Wiley & Sons, New York).

Traunmüller, H. (**1981**). "Perceptual dimension of openness in vowels", J. Acoust. Soc. Am. **69**, 1465–1475.

Un, C.K. & Magill, D.T. (**1975**). "The residual-excited linear prediction vocoder with transmission rate below 9.6kbit/s", IEEE Trans. Comm. **COM-23**, 1466–1474.

Van Bezooijen, R. & Pols, L.C.W. (**1990**). "Evaluating text-to-speech systems: some methodological aspects", Speech Communication **9**, 263–270.

Van Dijk-Kappers, A.M.L. & Marcus, S.M. (**1989**). "Temporal decomposition of speech", Speech Communication **8**, 125–135.

Van Hemert, J.P. (**1984**). "Multipulse Excitation (MPE): the possibilities and restrictions of a new speech synthesizer", Ann. Prog. Rep. **19**, Institute for Perception Research, Eindhoven, 20–24.

Vogten, L.L.M. (**1980**). "Evaluation of LPC formant-coded speech with a speech interference test", Ann. Prog. Rep. **15**, Institute for Perception Research, Eindhoven, 33–41.

Vogten, L.L.M. (**1983**). *Analyse, zuinige codering en resynthese van spraakgeluid* (Doctoral thesis, Eindhoven University of Technology).

Voiers, W.D. (**1969**). "The effects of masking voices on the apprehen-

sibility of six consonant attributes", Rep. No. AFCRL-69-0157 prepared under contract AF19(628)-5883 for the Air Force Cambridge Research Laboratories, Bedford Mass.

Voiers, W.D. (**1977**). "Diagnostic evaluation of speech intelligibility", in *Speech intelligibility and speaker recognition*, edited by M.E. Hawley (Dowden Hutchinson Ross, Stroudburg), 374–387.

Von Helmholtz, H.L.F. (**1877**). *On the sensations of Tone* (English translation of 4th edition by A.J.Ellis, Dover, New York, 1954).

Waterham, R.P. (**1989**). *The "Pocketstem": an easy-to-use speech communication aid for the vocally handicapped* (Doctoral thesis, Eindhoven University of Technology).

Willems, L.F. (**1986**). "Robust formant analysis", Ann. Prog. Rep. **21**, Institute for Perception Research, Eindhoven, 34–40.

Wong, D.J., Markel, J.D. & Gray, Jr., A.H. (**1979**). "Least squares glottal inverse filtering from an acoustic speech wave", IEEE Trans. Acoust., Speech, Signal Processing **27**, 350–355.

Wood, L.C. & Pearce, D.J.B. (**1989**). "Excitation synchronous formant analysis", IEE Proc. **136**, 110–118.

Summary

ONE of the applications of speech-coding algorithms is the generation of the output speech of text-to-speech systems. Linear Predictive Coding (LPC) is still one of the most powerful algorithms used for this purpose. Despite its many advantages such as the capability to resynthesize highly intelligible speech, the possibility to manipulate perceived aspects of speech, and the power to provide accurate estimates of speech parameters, LPC also has its shortcomings. LPC speech lacks naturalness and speaker characteristics are degraded. The research reported on in this dissertation was aiming for two things. One was the assessment of some limitations of LPC as a scheme for speech analysis, manipulation and synthesis, the other was the exploration of ways to remove some of the drawbacks of LPC.

Intelligibility is an important attribute of speech quality. We developed a Monosyllabic Adaptive Speech Interference Test (MASIT) which was used to evaluate the intelligibility of different speech-coding schemes. It was shown that, in the case of synthetic speech, differences in intelligibility are not always magnified by adding interfering speech. This means that MASIT provides information which cannot be obtained by traditional articulation tests. In particular, MASIT can be applied to assess the performance of speech-coding schemes in noisy environments.

Intelligibility of synthetic speech is just one attribute of speech quality. Other factors, such as naturalness, also play an important role. As LPC lacks naturalness, we tried to determine which requirements are needed for the generation of natural-sounding speech. We implemented a pitch-synchronous analysis-resynthesis system with which we systematically manipulated the amplitude and phase spectra of the LPC

residue. These stimuli were judged by subjects in a paired-comparison experiment. For female voices, both amplitude and phase information were neccesary to synthesize more natural-sounding speech, whereas for male voices amplitude information alone was sufficient to make the synthetic speech almost indistinguishable from natural speech.

Besides intelligibility and naturalness, preservation of speaker characteristics may also be an important feature of a high quality speech-coding scheme. We found that the proper reproduction of the prosodic structure of natural speech by means of LPC resynthesis is not sufficient to preserve reliably the identity of different speakers. We decided to concentrate our efforts on the improvement of the LPC speech-coding scheme. We implemented a glottal-excited (GE) LPC speech synthesizer which incorporates a more detailed model of the human voice source. We developed algorithms to estimate the parameters of the GE-LPC synthesizer from the speech signal. In this way we synthesized a set of hybrid speech stimuli, i.e. speech stimuli for which the voice-source function of one speaker is modulated by the vocal-tract transfer function of another speaker. The hybrid stimuli were presented to listeners in a perception experiment. The listeners had to identify which speaker they thought had produced the stimulus. The experimental results clearly show that listeners use both voice-source and vocal-tract information to perform the identification task. After the experiments were finished, subjects were asked which criteria they had used to identify the speakers. Their answers were remarkably similar, and could be related to physical aspects of the stimuli. Besides the expected importance of the vocal tract filter, the spectral balance between high and low-frequency components of the voice-source spectrum and the flutter of F_0 proved to be important perceptual cues for speaker identity.

Samenvatting

EEN van de toepassingen van spraakcoderingsalgorithmen is de generatie van spraak in een tekst-naar-spraak systeem. "Linear Predictive Coding (LPC)" is één van de meest krachtige algorithmen die voor dit doel gebruikt wordt. Ondanks de vele voordelen die de LPC techniek biedt, zoals de mogelijkheid om goed verstaanbare spraak te resynthetiseren, de mogelijkheid om perceptief belangrijke eigenschappen van spraak te manipuleren en het vermogen om spraakparameters uit het spraaksignaal te bepalen, kent de LPC-techniek ook zijn beperkingen. LPC-spraak schiet tekort met betrekking tot natuurlijkheid en sprekereigenschappen worden aangetast. Het onderzoek dat in dit proefschrift beschreven wordt, kent een tweetal doelstellingen. Op de eerste plaats wilden we een inzicht krijgen in de beperkingen van de LPC-techniek met betrekking tot de analyse, manipulatie en synthese van spraak. Op de tweede plaats wilden we onderzoeken hoe sommige van de nadelen van LPC opgeheven zouden kunnen worden.

Verstaanbaarheid is een belangrijk attribuut van spraakkwaliteit. Er is een Monosyllabische Spraak Interferentie Test (MASIT) ontwikkeld, die gebruikt is om de verstaanbaarheid van verschillende spraakcoderingstechnieken te evalueren. Het bleek dat in het geval van synthetische spraak, het toevoegen van interferentiespraak niet altijd leidt tot het vergroten van bestaande verschillen in verstaanbaarheid. Dit betekent dat MASIT informatie oplevert die niet met traditionele articulatietesten verkregen kan worden. MASIT is met name geschikt om te bepalen hoe bruikbaar spraakcoderingsalgorithmen zijn in lawaaiige omgevingen.

Verstaanbaarheid is slechts één attribuut van spraakkwaliteit. Andere factoren, zoals natuurlijkheid, spelen ook een belangrijke rol.

Omdat LPC-spraak tekort schiet met betrekking tot natuurlijkheid, hebben we geprobeerd vast te stellen welke condities nodig zijn om natuurlijkklinkende spraak te genereren. Er is een pitch-synchroon analyse-resynthese systeem geïmplementeerd, waarmee we systematisch de amplitude- en fase-spectra van het LPC residu hebben gemanipuleerd. Deze stimuli werden aangeboden aan luisteraars in een paarsgewijze-vergelijkingen experiment. In het geval van vrouwenspraak waren zowel amplitude- als fase-informatie noodzakelijk om natuurlijkklinkende spraak te synthetiseren, terwijl in het geval van mannenspraak alleen amplitude-informatie toereikend was om de synthetische spraak vrijwel ononderscheidbaar te maken van natuurlijke spraak.

Naast verstaanbaarheid en natuurlijkheid, kan het behoud van sprekeridentiteit ook een belangrijke eigenschap zijn van een geavanceerd spraakcoderingsalgorithme. Uit informele experimenten met LPC-resynthese bleek dat een correcte reproduktie van de prosodische structuur van natuurlijke spraak niet voldoende is om het behoud van de identiteit van verschillende sprekers te waarborgen. We hebben daarom besloten om onze inspanningen te concentreren op het verbeteren van de LPC synthesizer. De glottal-excited (GE) LPC spraaksynthesizer is geïmplementeerd die een meer gedetailleerd model bevat van de menselijke stembron. Er zijn algorithmen ontwikkeld om de parameters van de GE-LPC synthesizer te bepalen uit het spraaksignaal. Op deze wijze is een set hybride spraakstimuli gesynthetiseerd, d.w.z. spraakstimuli waarvoor de stembronfunctie van de ene spreker gemoduleerd wordt met de spraakkanaaloverdrachtsfunctie van een andere spreker. Deze hybride stimuli zijn in een perceptie-experiment aangeboden aan luisteraars. De luisteraars moesten aangeven welke spreker de stimulus geproduceerd had. De experimentele resultaten laten duidelijk zien dat de luisteraars bij het uitvoeren van de identificatietaak informatie gebruiken van zowel de stembron als het spraakkanaal. Na afloop van de experimenten werd aan de luisteraars gevraagd welke criteria zij gehanteerd hadden om de sprekers te identificeren. De antwoorden van de luisteraars vertoonden veel overeenkomsten en konden gerelateerd worden aan fysische eigenschappen van de stimuli. Naast de verwachte belangrijke rol van het spraakkanaal, zijn de spectrale balans tussen hoge en lage frequen-

tiecomponenten in het spectrum van de stembron en de "F_0-flutter" voor de waargenomen sprekeridentiteit.

Curriculum Vitae

18 mei 1960	Geboren te Vlaardingen.
aug 1972 – juni 1978	Eckartcollege Eindhoven. Atheneum B.
sept 1978 – feb 1986	Technische Universiteit Eindhoven. Technische Natuurkunde (met lof). Afstudeerrichting: Psychofysica. Afstudeeronderwerp: "The Strike Note of Bells".
mei 1986 – mei 1987	Onderzoeksmedewerker in dienst van het SPIN/ASSP programma[1], gedetacheerd op het Instituut voor Perceptie Onderzoek (IPO), Eindhoven. Projecttitel: "Start-evaluatie van beschikbare methoden voor analyse, manipulatie en resynthese van spraak".
mei 1987 – nov 1989	Onderzoeksmedewerker in dienst van het SPIN/ASSP programma, gedetacheerd op het Instituut voor Perceptie Onderzoek (IPO). Projecttitel: "Fysische correlaten van waargenomen sprekeridentiteit en sprekerkenmerken".
dec 1989 – heden	Wetenschappelijk medewerker in dienst van het Natuurkundig Laboratorium van Philips, gedetacheerd op het Instituut voor Perceptie Onderzoek (IPO).

[1] "StimuleringsProjektteam Informaticaonderzoek Nederland (SPIN)"-programma "Analyse en Synthese van SPraak (ASSP)".

Stellingen

behorende bij het proefschrift
On the quality of synthetic speech
van J.H. Eggen

I

De spraakinterferentietest van Nakatani & Dukes (1973) is gebaseerd op het idee dat kleine verschillen in verstaanbaarheid vergroot kunnen worden door het toevoegen van interferentiespraak. Dit idee gaat niet altijd op voor synthetische spraak.

Nakatani, L.H. and Dukes, K.D. (1973): *A sensitive test of speech communication quality*, J. Acoust. Soc. Am. **53**, 1083–1092.

II

De conclusie van Miller (1964) en Matsumoto et al. (1973), dat de in het spraaksignaal aanwezige akoestische informatie van het spraakkanaal meer zou bijdragen aan de identiteit van een spreker dan de in het spraaksignaal aanwezige akoestische informatie van de stembron, is niet algemeen geldig.

Miller, J.E. (1964): *Decapitation and recapitation, a study of voice quality*, J. Acoust. Soc. Am. **36**, 2002 (A).

Matsumoto, H., Hiki, S., Sone, T. and Nimura, T. (1973): *Multidimensional representation of personal quality of vowels and its acoustical correlates*, IEEE Trans. Audio Electroacoust. **AU–21**, 428–436.

III

Het door Taylor (1988) ontwikkelde "layered protocols" model voor mens–computer communicatie, kan met succes toegepast worden in het onderzoek naar betere gebruikersinterfaces voor consumentenelectronica.

Taylor, M.M. (1988): *Layered protocols for computer–human dialogue. I: Principles*, Int. J. Man-Machine Studies **28**, 175–218.

IV

De "misjudged octave" hypothese, door Jones (1930) opgesteld om de waargenomen toonhoogte van klokken te verklaren, is te eenvoudig. Andere deeltonen, zoals de duodeciem en het dubbeloctaaf, kunnen eveneens de waargenomen toonhoogte van een klok beïnvloeden.

Jones, A.T.J. (1930): *The strike note of bells*, J. Acoust. Soc. Am. **1**, 373–381.

Eggen, J.H. and Houtsma A.J.M. (1986): *The pitch perception of bell sounds*, IPO annual progress report **21**, 15–23.

V

Gilman (1991) trekt een parallel tussen de randvoorwaarden die nodig zijn voor een onderzoeksteam om te komen tot nieuwe onderzoeksresultaten en de basiselementen die nodig zijn voor een jazzgroep om nieuwe muziek te creëren. In deze analogie onderbelicht Gilman de invloed die de interactie tussen het onderzoekteam, c.q. de jazzgroep, enerzijds en de buitenwereld, c.q. het publiek, anderzijds heeft op het uiteindelijke resultaat.

Gilman, J.J. (1991): *Research management today*, Physics Today, March 1991, 42–48.

VI

Het succes van de "Musical Instrument Digital Interface" standaard (MIDI) voor electronische muziekapparatuur toont aan dat een goed gedefinieerde en flexibele specificatie aanleiding kan geven tot een groot aantal nieuwe toepassingen die nooit door de ontwerpers van de specificatie voorzien waren.

Aikin, J. (1986): *MIDI, Musical Instrument Digital Interface*, Keyboard, January 1986, 28–31.

VII

De huidige generatie consumentenelectronica ondersteunt onvoldoende de intenties van de gebruiker.

VIII

Recente ontwikkelingen op het gebied van de mens-machine interactie zullen, na verdere uitwerking, een grote groep niet-muzikaal geschoolde mensen in staat stellen deel te nemen aan het maken van muziek op een niveau dat tot nu toe ongekend is.

IX

In de filmindustrie is het van groot belang het juiste geluid te combineren met het juiste beeld. Dat deze combinatie niet altijd overeen hoeft te stemmen met de fysische werkelijkheid, bewijst het ruimteschip dat we op het bioscoopscherm met donderend geraas door de lege ruimte zien schieten.