Helsinki University of Technology
Department of Electrical and Communications Engineering

Sami Lemmetty

Review of Speech Synthesis Technology

This Master's Thesis has been submitted for official examination for the degree of Master of Science in Espoo on March 30, 1999.

Supervisor of the Thesis

Professor Matti Karjalainen

HELSINKI UNIVERSTY OF TECHNOLOGY Abstract of the Master's Thesis

Author:	Sami Lemmetty	
Name of the Thesis:	Review of Speech Synthesis Technology	
Date:	March 30, 1999	Number of pages: 104
Department:	Electrical and Communications Engineering	
Professorship:	Acoustics and Audio Signal Processing (S-89)	
Supervisor:	Professor Matti Karjalainen	

Synthetic or artificial speech has been developed steadily during the last decades. Especially, the intelligibility has reached an adequate level for most applications, especially for communication impaired people. The intelligibility of synthetic speech may also be increased considerably with visual information. The objective of this work is to map the current situation of speech synthesis technology. Speech synthesis may be categorized as restricted (messaging) and unrestricted (text-to-speech) synthesis. The first one is suitable for announcing and information systems while the latter is needed for example in applications for the visually impaired. The text-to-speech procedure consists of two main phases, usually called high- and low-level synthesis. In high-level synthesis the input text is converted into such form that the low-level synthesizer can produce the output speech. The three basic methods for low-level synthesis are the formant, concatenative, and articulatory synthesis. The formant synthesis is based on the modeling of the resonances in the vocal tract and is perhaps the most commonly used during last decades. However, the concatenative synthesis which is based on playing prerecorded samples from natural speech is becoming more popular. In theory, the most accurate method is articulatory synthesis which models the human speech production system directly, but it is also the most difficult approach. Since the quality of synthetic speech is improving steadily, the application field is also expanding rapidly. Synthetic speech may be used to read e-mail and mobile messages, in multimedia applications, or in any kind of human-machine interaction. The evaluation of synthetic speech is also an important issue, but difficult because the speech quality is a very multidimensional term. This has led to the large number of different tests and methods to evaluate different features in speech. Today, speech synthesizers of various quality are available as several different products for all common languages, including Finnish.

Keywords: speech synthesis, synthesized speech, text-to-speech, tts, artificial speech, speech synthesizer, audio-visual speech.

TEKNILLINEN KORKEAKOULU Diplomityön tiivistelmä

Tekijä:	Sami Lemmetty	
Työn nimi:	Katsaus puhesynteesiteknologiaan	
Päivämäärä:	30.3.1999	Sivumäärä: 104
Osasto:	Sähkö- ja tietoliikennetekniikan osasto	
Professuuri:	Akustiikka ja äänenkäsittelytekniikka (S-89)	
Työn valvoja:	Professori Matti Karjalainen	

Synteettinen eli keinotekoisesti tuotettu puhe on kehittynyt varsin nopeasti viimeisten vuosikymmenten aikana. Erityisesti puheen ymmärrettävyys on saavuttanut riittävän tason moniin kommunikaatiovaikeuksia omaavien ihmisten tarpeisiin ja sovelluksiin. Synteettisen puheen ymmärrettävyyttä voidaan lisäksi parantaa merkittävästi lisäämällä visuaalista informaatiota (puhuva pää). Tämän työn tarkoitus on kartoittaa puhesynteesiteknologian nykytila. Puhesynteesi voidaan jakaa rajoitetun ja rajoittamattoman sanaston synteesiin. Rajoitetun sanaston synteesi soveltuu hyvin erilaisiin kuulutus- ja informaatiojärjestelmiin, kun taas esimerkiksi näkövammaissovelluksiin tarvitaan useimmiten rajoittamattoman sanaston synteesiä. Rajoittamattoman sanaston synteesi voidaan jakaa korkean- ja matalan tason synteesiin. Korkean tason synteesi huolehtii tekstin esikäsittelystä (numerot, lyhenteen jne.), analyysistä sekä tarvittavan tiedon välittämisestä varsinaisen puhesignaalin tuottavan matalan tason syntetisaattorin ohjaamiseksi. Varsinaisen puhesynteesin tuottamiselle on kolme perusmenetelmää. Yleisin menetelmä on formanttisynteesi, missä mallinnetaan ihmisen ääniväylän resonanssikohtia. Yleistymässä on myös luonnollisesta puheesta poimittujen lyhyiden ääninäytteiden toistamiseen perustuva aikatason synteesi. Kolmas vaihtoehto on mallintaa ihmisen puheentuottojärjestelmää suoraan, mikä on kuitenkin teknisesti ja laskennallisesti varsin raskasta. Puheen luonnollisuuden parantuessa sitä on alettu käyttää yhä useammassa eri sovelluskohteessa, kuten erilaiset lukulaitteet (sähköposti, tekstiviesti jne.), multimedia, tai mikä tahansa ihmisen ja koneen välinen vuorovaikutus. Koska puheen laatu on varsin monitahoinen kysymys, on myös sen laadun arvioiminen varsin hankalaa ja monimutkaista. Tämän vuoksi on olemassa lukuisia eri menetelmiä synteettisen puheen laadun ja erilaisten ominaisuuksien arvioimiseksi. Puhesyntetisaattoreita on tällä hetkellä saatavilla lukuisia erilaisia ja eritasoisia kaikille yleisimmille kielille, myös suomeksi.

Avainsanat: puhesynteesi, audiovisuaalinen puhesynteesi, tts, keinotekoinen puhe, synteettinen puhe.

PREFACE

This thesis has been carried out at Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing. The work is a part of the larger audio-visual speech synthesis project at HUT and it is supported by TEKES, Nokia Research Center, Sonera, Finnish Federation for the Visually Impaired, and Timehouse Oy. This thesis is based on the research and literature made by several speech synthesis researchers and research groups over the world during last decades.

First, I would like to thank my supervisor Professor Matti Karjalainen for his instruction and guidance during this work. Without his encouragement, feedback, and motivation this work could not have been done. I wish also to give my thanks to Dr. Tech. Unto K. Laine for his comments and discussions considering my work, Mr. Matti Airas, Mr. Martti Rahkila, and Mr. Tero Tolonen for their support when the Windows95 has shown some of its concealed and unpredictable features, and the people at the Acoustics Laboratory for giving me the most enjoyable and inspiring working environment. I would also thank all the project participants for their feedback and comments during my work.

Espoo, March 30, 1999

Sami Lemmetty

TABLE OF CONTENTS

1. INTRODUCTION 1

1.1 PROJECT DESCRIPTION 1

1.2 INTRODUCTION TO SPEECH SYNTHESIS 2

2. HISTORY AND DEVELOPMENT OF SPEECH SYNTHESIS 4

2.1 FROM MECHANICAL TO ELECTRICAL SYNTHESIS 4

2.2 DEVELOPMENT OF ELECTRICAL SYNTHESIZERS 6

2.3 HISTORY OF FINNISH SPEECH SYNTHESIS 10

3. PHONETICS AND THEORY OF SPEECH PRODUCTION 11

3.1 REPRESENTATION AND ANALYSIS OF SPEECH SIGNALS 11

3.2 SPEECH PRODUCTION 14

3.3 PHONETICS 17

3.3.1 English Articulatory Phonetics 18

3.3.2 Finnish Articulatory Phonetics 20

4. PROBLEMS IN SPEECH SYNTHESIS 22

4.1 TEXT-TO-PHONETIC CONVERSION 22

4.1.1 Text preprocessing 22

4.1.2 Pronunciation 24

4.1.3 Prosody 25

4.2 PROBLEMS IN LOW LEVEL SYNTHESIS 26

4.3 LANGUAGE SPECIFIC PROBLEMS AND FEATURES 27

5. METHODS, TECHNIQUES, AND ALGORITHMS 28

5.1 ARTICULATORY SYNTHESIS 28

5.2 FORMANT SYNTHESIS 29

5.3 CONCATENATIVE SYNTHESIS 32

5.3.1 PSOLA Methods 34

5.3.2 Microphonemic Method 36

5.4 LINEAR PREDICTION BASED METHODS 37

5.5 SINUSOIDAL MODELS 38

5.6 HIGH-LEVEL SYNTHESIS 40

5.6.1 Text Preprocessing 40

5.6.2 Pronunciation 41

5.6.3 Prosody 42

5.7 OTHER METHODS AND TECHNIQUES 45

6. APPLICATIONS OF SYNTHETIC SPEECH 47
6.1 APPLICATIONS FOR THE BLIND 47
6.2 APPLICATIONS FOR THE DEAFENED AND VOCALLY HANDICAPPED 48
6.3 EDUCATIONAL APPLICATIONS 49
6.4 APPLICATIONS FOR TELECOMMUNICATIONS AND MULTIMEDIA 49
6.5 OTHER APPLICATIONS AND FUTURE DIRECTIONS 50

7. APPLICATION FRAMEWORKS 51
7.1 SPEECH APPLICATION PROGRAMMING INTERFACE 51
7.1.1 Control Tags 52
7.2 INTERNET SPEECH MARKUP LANGUAGES 54
7.3 MPEG-4 TTS 56
7.3.1 MPEG-4 TTS Bitstream 57
7.3.2 Structure of MPEG-4 TTS Decoder 58
7.3.3 Applications of MPEG-4 TTS 59

8. AUDIOVISUAL SPEECH SYNTHESIS 60
8.1 INTRODUCTION AND HISTORY 60
8.2 TECHNIQUES AND MODELS 61

9. PRODUCTS 64
9.1 INFOVOX 64
9.2 DECTALK 65
9.3 BELL LABS TEXT-TO-SPEECH 66
9.4 LAUREATE 68
9.5 SOFTVOICE 68
9.6 CNET PSOLA 69
9.7 ORATOR 69
9.8 EUROVOCS 70
9.9 LERNOUT & HAUSPIES 70
9.10 APPLE PLAIN TALK 71
9.11 ACUVOICE 71
9.12 CYBERTALK 72
9.13 ETI ELOQUENCE 72
9.14 FESTIVAL TTS SYSTEM 73
9.15 MODELTALKER 73
9.16 MBROLA 73
9.17 WHISTLER 74
9.18 NEUROTALKER 74
9.19 LISTEN2 75
9.20 SPRUCE 75
9.21 HADIFIX 76

9.22 SVOX 76
9.23 SYNTE2 AND SYNTE3 77
9.24 TIMEHOUSE MIKROPUHE 78
9.25 SANOSSE 78
9.26 SUMMARY 78

10. SPEECH QUALITY AND EVALUATION 79
10.1 SEGMENTAL EVALUATION METHODS 80
10.1.1 Diagnostic Rhyme Test (DRT) 81
10.1.2 Modified Rhyme Test (MRT) 81
10.1.3 Diagnostic Medial Consonant Test (DMCT) 83
10.1.4 Standard Segmental Test 83
10.1.5 Cluster Identification Test (CLID) 83
10.1.6 Phonetically Balanced Word Lists (PB) 84
10.1.7 Nonsense words and Vowel-Consonant transitions 84
10.2 SENTENCE LEVEL TESTS 84
10.2.1 Harvard Psychoacoustic Sentences 84
10.2.2 Haskins Sentences 85
10.2.3 Semantically Unpredictable Sentences (SUS) 85
10.3 COMPREHENSION TESTS 86
10.4 PROSODY EVALUATION 86
10.5 INTELLIGIBILITY OF PROPER NAMES 87
10.6 OVERALL QUALITY EVALUATION 87
10.6.1 Mean Opinion Score (MOS) 87
10.6.2 Categorical Estimation (CE) 88
10.6.3 Pair Comparison (PC) 88
10.6.4 Magnitude and Ratio Estimation 88
10.7 FIELD TESTS 89
10.8 AUDIOVISUAL ASSESSMENT 89
10.9 SUMMARY 90

11. CONCLUSIONS AND FUTURE STRATEGIES 91

REFERENCES

APPENDIX A: SPEECH SYNTHESIS DEMONSTRATION CD

APPENDIX B: SUMMARY OF SPEECH SYNTHESIS PRODUCTS

LIST OF SYMBOLS

$a(k)$	Filter coefficients
A	Area
ω	Radian frequency
λ	Warping parameter
BW	Bandwidth
c	Sound velocity
$e(n)$	Discrete error signal
F0	Fundamental frequency
F1, F2, F3,...	Formant frequencies
G	Gain
l	Length
OQ	Open Quotient
Q	Q-value (formant frequency divided by its bandwidth)
U_m, U_g	Volume velocity at mouth and glottis
VO	Degree of Voicing
$y(n)$	Discrete speech signal
z^{-k}	Delay of k samples
Z_m, Z_g	Acoustic impedance at mouth and glottis

LIST OF ABBREVIATIONS

ACR	Absolute Category Rating
ANN	Artificial Neural Network
C	Consonant
CLID-test	Cluster Identification Test
DCR	Degradation Category Rating
DFT	Discrete Fourier Transform
DMOS	Degradation Mean Opinion Score
DRT	Diagnostic Rhyme Test
DTMF	Dual Tone Multi-Frequency
FAP	Facial Animation Parameter
HMM	Hidden Markov Model
HTML	Hypertext Markup Language
IDFT	Inverse Discrete Fourier Transform
IPA	International Phonetic Association
LPC	Linear Predictive Coding
MOS	Mean Opinion Score
MPEG	Moving Picture Expert Group
MRT	Modified Rhyme Test
NN	Neural Network
PARCAS	Parallel-Cascade
PCM	Pulse Code Modulation
PSOLA	Pitch Synchronous Overlap and Add
Pt, Pa, Pb	Target phoneme, phoneme after, and phoneme before
SAM-PA	Speech Assessment Methods - Phonetic Alphabet
SAPI	Speech Application Programming Interface
SSML	Speech Synthesis Markup Language
STML	Spoken Text Markup Language
SUS	Semantically Unpredictable Sentences
TTS	Text-to-Speech
V	Vowel
WLP	Warped Linear Prediction

1. INTRODUCTION

1.1 Project Description

This is a pre-study for a larger audiovisual speech synthesis project that is planned to be carried out during 1998-2000 at Helsinki University of Technology. The main objective of this report is to map the situation of today's speech synthesis technology and to focus on potential methods for the future of this project. Usually literature and articles in the area are focused on a single method or single synthesizer or the very limited range of the technology. In this report the whole speech synthesis area with as many methods, techniques, applications, and products as possible is under investigation. Unfortunately, this leads to a situation where in some cases very detailed information may not be given here, but may be found in given references.

The objective of the whole project is to develop high quality audiovisual speech synthesis with a well synchronized talking head, primarily in Finnish. Other aspects, such as naturalness, personality, platform independence, and quality assessment are also under investigation. Most synthesizers today are so called stand-alones and they do not work platform independently and usually do not share common parts, thus we can not just put together the best parts of present systems to make a state-of-the-art synthesizer. Hence, with good modularity characteristics we may achieve a synthesis system which is easier to develop and improve.

The report starts with a brief historical description of different speech synthesis methods and speech synthesizers. The next chapter includes a short theory section of human speech production, articulatory phonetics, and some other related concepts. The speech synthesis procedure involves lots of different kinds of problems described in Chapter 4. Various existing methods and algorithms are discussed in Chapter 5 and the following two chapters are dedicated to applications and some application frameworks. The latest hot topic in the speech synthesis area is to include facial animation into synthesized speech. A short introduction to audiovisual speech synthesis is included in Chapter 8. Although the audiovisual synthesis is not the main purpose of this report, it will be discussed briefly to give a general view of the project. A list of available synthesizers and some ongoing speech synthesis projects is introduced in Chapter 9 and, finally, the last two chapters contain some evaluation methods, evaluations, and future discussion. The end of the thesis contains a collection of some speech synthesis related literature, WEB-sites, and some sound examples stored on an accompanying audio compact disc.

1.2 Introduction to Speech Synthesis

Speech is the primary means of communication between people. Speech synthesis, automatic generation of speech waveforms, has been under development for several decades (Santen et al. 1997, Kleijn et al. 1998). Recent progress in speech synthesis has produced synthesizers with very high intelligibility but the sound quality and naturalness still remain a major problem. However, the quality of present products has reached an adequate level for several applications, such as multimedia and telecommunications. With some audiovisual information or facial animation (talking head) it is possible to increase speech intelligibility considerably (Beskow et al. 1997). Some methods for audiovisual speech have been recently introduced by for example Santen et al. (1997), Breen et al. (1996), Beskow (1996), and Le Goff et al. (1996).

The text-to-speech (TTS) synthesis procedure consists of two main phases. The first one is text analysis, where the input text is transcribed into a phonetic or some other linguistic representation, and the second one is the generation of speech waveforms, where the acoustic output is produced from this phonetic and prosodic information. These two phases are usually called as high- and low-level synthesis. A simplified version of the procedure is presented in Figure 1.1. The input text might be for example data from a word processor, standard ASCII from e-mail, a mobile text-message, or scanned text from a newspaper. The character string is then preprocessed and analyzed into phonetic representation which is usually a string of phonemes with some additional information for correct intonation, duration, and stress. Speech sound is finally generated with the low-level synthesizer by the information from high-level one.

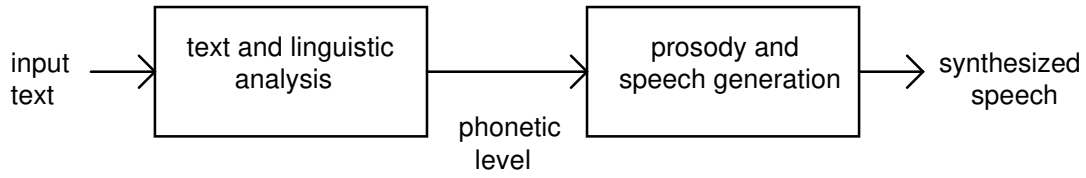

Fig. 1.1. Simple text-to-speech synthesis procedure.

The simplest way to produce synthetic speech is to play long prerecorded samples of natural speech, such as single words or sentences. This concatenation method provides high quality and naturalness, but has a limited vocabulary and usually only one voice. The method is very suitable for some announcing and information systems. However, it is quite clear that we can not create a database of all words and common names in the world. It is maybe even inappropriate to call this speech synthesis because it contains only recordings. Thus, for unrestricted speech synthesis (text-to-speech) we have to use shorter pieces of speech signal, such as syllables, phonemes, diphones or even shorter segments.

Another widely used method to produce synthetic speech is formant synthesis which is based on the source-filter-model of speech production described in Figure 1.2 below. The method is sometimes called terminal analogy because it models only the sound source and the formant frequencies, not any physical characteristics of the vocal tract (Flanagan 1972). The excitation signal could be either voiced with fundamental frequency (F0) or unvoiced noise. A mixed excitation of these two may also be used for voiced consonants and some aspiration sounds. The excitation is then gained and filtered with a vocal tract filter which is constructed of resonators similar to the formants of natural speech.

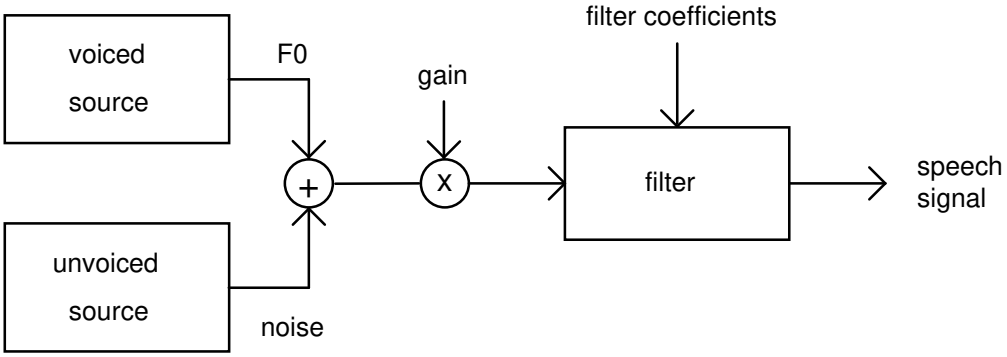

Fig. 1.2. Source-filter model of speech.

In theory, the most accurate method to generate artificial speech is to model the human speech production system directly (O'Saughnessy 1987, Witten 1982, Donovan 1996). This method, called articulatory synthesis, typically involves models of the human articulators and vocal cords. The articulators are usually modeled with a set of area functions of small tube sections. The vocal cord model is used to generate an appropriate excitation signal, which may be for example a two-mass model with two vertically moving masses (Veldhuis et al. 1995). Articulatory synthesis holds a promise of high-quality synthesized speech, but due to its complexity the potential has not been realized yet.

All synthesis methods have some benefits and problems of their own and it is quite difficult to say which method is the best one. With concatenative and formant synthesis, very promising results have been achieved recently, but also articulatory synthesis may arise as a potential method in the future. Different synthesis methods, algorithms, and techniques are discussed more closely in Chapter 5.

2. HISTORY AND DEVELOPMENT OF SPEECH SYNTHESIS

Artificial speech has been a dream of the humankind for centuries. To understand how the present systems work and how they have developed to their present form, a historical review may be useful. In this chapter, the history of synthesized speech from the first mechanical efforts to systems that form the basis for today's high-quality synthesizers is discussed. Some separate milestones in synthesis-related methods and techniques will also be discussed briefly. For more detailed description of speech synthesis development and history see for example Klatt (1987), Schroeder (1993), and Flanagan (1972, 1973) and references in these.

2.1 From Mechanical to Electrical Synthesis

The earliest efforts to produce synthetic speech were made over two hundred years ago (Flanagan 1972, Flanagan et al. 1973, Schroeder 1993). In St. Petersburg 1779 Russian Professor Christian Kratzenstein explained physiological differences between five long vowels (/a/, /e/, /i/, /o/, and /u/) and made apparatus to produce them artificially. He constructed acoustic resonators similar to the human vocal tract and activated the resonators with vibrating reeds like in music instruments. The basic structure of resonators is shown in Figure 2.1. The sound /i/ is produced by blowing into the lower pipe without a reed causing the flute-like sound.

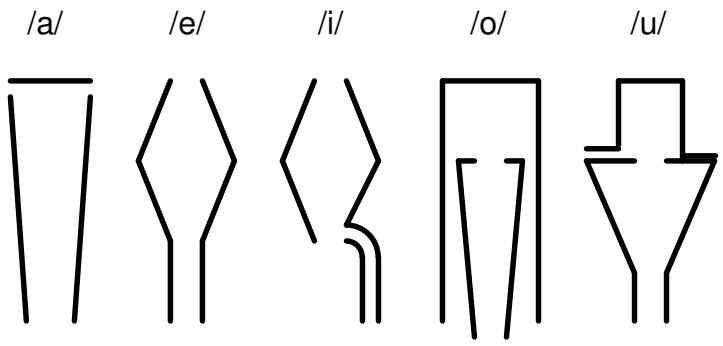

Fig. 2.1. Kratzenstein's resonators (Schroeder 1993).

A few years later, in Vienna 1791, Wolfgang von Kempelen introduced his "Acoustic-Mechanical Speech Machine", which was able to produce single sounds and some sound combinations (Klatt 1987, Schroeder 1993). In fact, Kempelen started his work before Kratzenstein, in 1769, and after over 20 years of research he also published a book in which he described his studies on human speech production and the experiments with his speaking machine. The essential parts of the machine were a pressure chamber for

the lungs, a vibrating reed to act as vocal cords, and a leather tube for the vocal tract action. By manipulating the shape of the leather tube he could produce different vowel sounds. Consonants were simulated by four separate constricted passages and controlled by the fingers. For plosive sounds he also employed a model of a vocal tract that included a hinged tongue and movable lips. His studies led to the theory that the vocal tract, a cavity between the vocal cords and the lips, is the main site of acoustic articulation. Before von Kempelen's demonstrations the larynx was generally considered as a center of speech production. Kempelen received also some negative publicity. While working with his speaking machine he demonstrated a speaking chess-playing machine. Unfortunately, the main mechanism of the machine was concealed, legless chess-player expert. Therefore his real speaking machine was not taken so seriously as it should have (Flanagan et al. 1973, Schroeder 1993).

In about mid 1800's Charles Wheatstone constructed his famous version of von Kempelen's speaking machine which is shown in Figure 2.2. It was a bit more complicated and was capable to produce vowels and most of the consonant sounds. Some sound combinations and even full words were also possible to produce. Vowels were produced with vibrating reed and all passages were closed. Resonances were effected by deforming the leather resonator like in von Kempelen's machine. Consonants, including nasals, were produced with turbulent flow trough a suitable passage with reed-off .

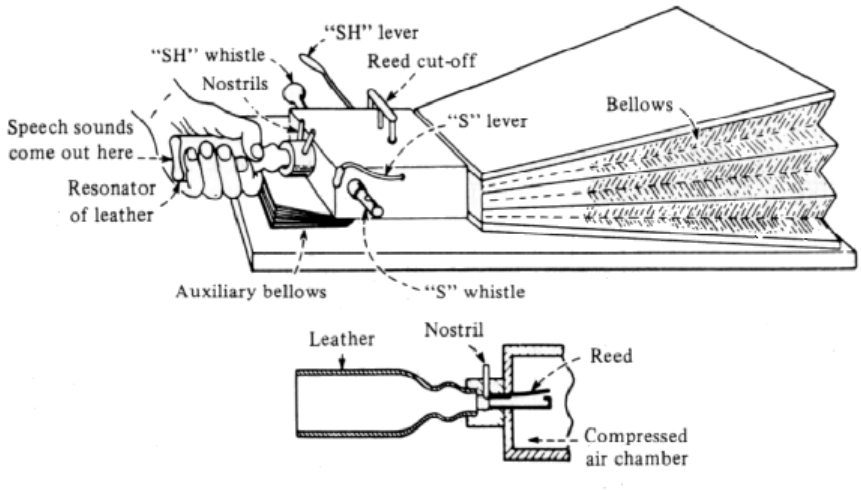

Fig. 2.2. Wheatstone's reconstruction of von Kempelen's speaking machine (Flanagan 1972).

The connection between a specific vowel sound and the geometry of the vocal tract was found by Willis in 1838 (Schroeder 1993). He synthesized different vowels with tube resonators like organ pipes. He also discovered that the vowel quality depended only on the length of the tube and not on its diameter.

In late 1800's Alexander Graham Bell with his father, inspired by Wheatstone's speaking machine, constructed same kind of speaking machine. Bell made also some questionable experiments with his terrier. He put his dog between his legs and made it growl, then he modified vocal tract by hands to produce speech-like sounds (Flanagan 1972, Shroeder 1993).

The research and experiments with mechanical and semi-electrical analogs of vocal system were made until 1960's, but with no remarkable success. The mechanical and semi-electrical experiments made by famous scientists, such as Herman von Helmholz and Charles Wheatstone are well described in Flanagan (1972), Flanagan et al. (1973), and Shroeder (1993).

2.2 Development of Electrical Synthesizers

The first full electrical synthesis device was introduced by Stewart in 1922 (Klatt 1987). The synthesizer had a buzzer as excitation and two resonant circuits to model the acoustic resonances of the vocal tract. The machine was able to generate single static vowel sounds with two lowest formants, but not any consonants or connected utterances. Same kind of synthesizer was made by Wagner (Flanagan 1972). The device consisted of four electrical resonators connected in parallel and it was excited by a buzz-like source. The outputs of the four resonators were combined in the proper amplitudes to produce vowel spectra. In 1932 Japanese researchers Obata and Teshima discovered the third formant in vowels (Schroeder 1993). The three first formants are generally considered to be enough for intelligible synthetic speech.

First device to be considered as a speech synthesizer was VODER (Voice Operating Demonstrator) introduced by Homer Dudley in New York World's Fair 1939 (Flanagan 1972, 1973, Klatt 1987). VODER was inspired by VOCODER (Voice Coder) developed at Bell Laboratories in the mid-thirties. The original VOCODER was a device for analyzing speech into slowly varying acoustic parameters that could then drive a synthesizer to reconstruct the approximation of the original speech signal. The VODER consisted of wrist bar for selecting a voicing or noise source and a foot pedal to control the fundamental frequency. The source signal was routed through ten bandpass filters whose output levels were controlled by fingers. It took considerable skill to play a sentence on the device. The speech quality and intelligibility were far from good but the potential for producing artificial speech were well demonstrated. The speech quality of VODER is demonstrated in accompanying CD (track 01).

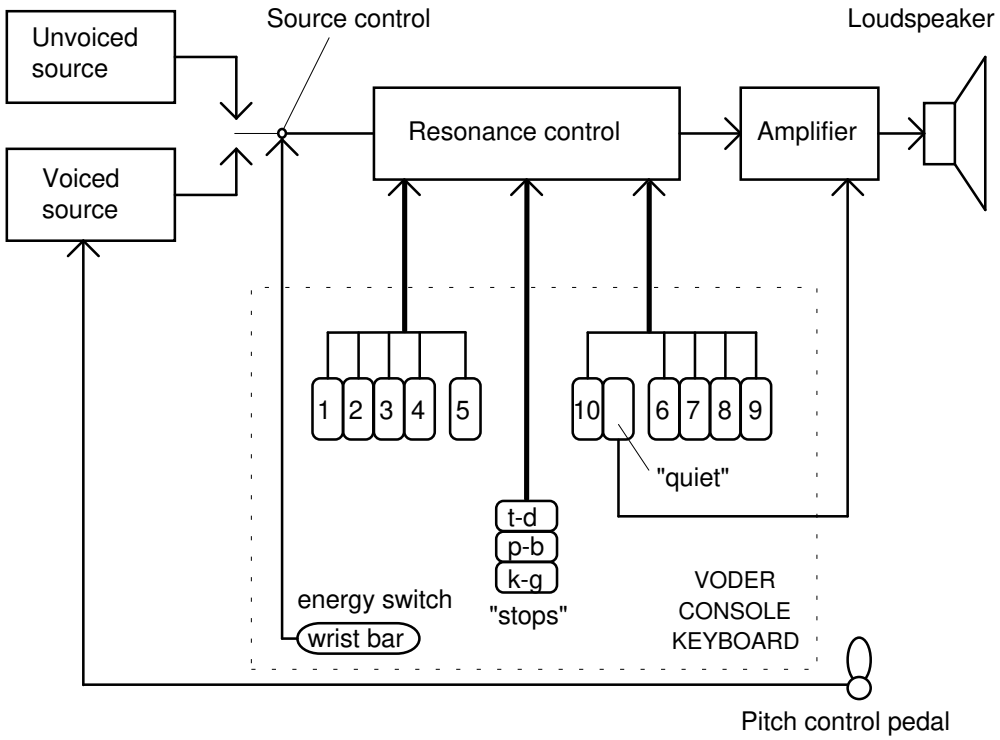

Fig. 2.3. The VODER speech synthesizer (Klatt 1987).

After demonstration of VODER the scientific world became more and more interested in speech synthesis. It was finally shown that intelligible speech can be produced artificially. Actually, the basic structure and idea of VODER is very similar to present systems which are based on source-filter-model of speech.

About a decade later, in 1951, Franklin Cooper and his associates developed a Pattern Playback synthesizer at the Haskins Laboratories (Klatt 1987, Flanagan et al. 1973). It reconverted recorded spectrogram patterns into sounds, either in original or modified form. The spectrogram patterns were recorded optically on the transparent belt (track 02).

The first formant synthesizer, PAT (Parametric Artificial Talker), was introduced by Walter Lawrence in 1953 (Klatt 1987). PAT consisted of three electronic formant resonators connected in parallel. The input signal was either a buzz or noise. A moving glass slide was used to convert painted patterns into six time functions to control the three formant frequencies, voicing amplitude, fundamental frequency, and noise amplitude (track 03). At about the same time Gunnar Fant introduced the first cascade formant synthesizer OVE I (Orator Verbis Electris) which consisted of formant resonators connected in cascade (track 04). Ten years later, in 1962, Fant and Martony introduced an improved OVE II synthesizer, which consisted of separate parts to model the transfer function of the vocal tract for vowels, nasals, and obstruent consonants. Possible excitations were voicing, aspiration noise, and frication noise. The OVE

projects were followed by OVE III and GLOVE at the Kungliga Tekniska Högskolan (KTH), Sweden, and the present commercial Infovox system is originally descended from these (Carlson et al. 1981, Barber et al. 1989, Karlsson et al. 1993).

PAT and OVE synthesizers engaged a conversation how the transfer function of the acoustic tube should be modeled, in parallel or in cascade. John Holmes introduced his parallel formant synthesizer in 1972 after studying these synthesizers for few years. He tuned by hand the synthesized sentence "I enjoy the simple life" (track 07) so good that the average listener could not tell the difference between the synthesized and the natural one (Klatt 1987). About a year later he introduced parallel formant synthesizer developed with JSRU (Joint Speech Research Unit) (Holmes et al. 1990).

First articulatory synthesizer was introduced in 1958 by George Rosen at the Massachusetts Institute of Technology, M.I.T. (Klatt 1987). The DAVO (Dynamic Analog of the VOcal tract) was controlled by tape recording of control signals created by hand (track 11). In mid 1960s, first experiments with Linear Predictive Coding (LPC) were made (Schroeder 1993). Linear prediction was first used in low-cost systems, such as TI Speak'n'Spell in 1980, and its quality was quite poor compared to present systems (track 13). However, with some modifications to basic model, which are described later in Chapter 5, the method has been found very useful and it is used in many present systems.

The first full text-to-speech system for English was developed in the Electrotehnical Laboratory, Japan 1968 by Noriko Umeda and his companions (Klatt 1987). It was based on an articulatory model and included a syntactic analysis module with sophisticated heuristics. The speech was quite intelligible but monotonous and far away from the quality of present systems (track 24).

In 1979 Allen, Hunnicutt, and Klatt demonstrated the MITalk laboratory text-to-speech system developed at M.I.T. (track 30). The system was used later also in Telesensory Systems Inc. (TSI) commercial TTS system with some modifications (Klatt 1987, Allen et al. 1987). Two years later Dennis Klatt introduced his famous Klattalk system (track 33), which used a new sophisticated voicing source described more detailed in (Klatt 1987). The technology used in MITalk and Klattalk systems form the basis for many synthesis systems today, such as DECtalk (tracks 35-36) and Prose-2000 (track 32). For more detailed information of MITalk and Klattalk systems, see for example Allen et al. (1987), Klatt (1982), or Bernstein et al. (1980).

The first reading aid with optical scanner was introduced by Kurzweil in 1976. The Kurzweil Reading Machines for the Blind were capable to read quite well the multifont written text (track 27). However, the system was far too expensive for average

customers (the price was still over $ 30 000 about ten years ago), but were used in libraries and service centers for visually impaired people (Klatt 1987).

In late 1970's and early 1980's, considerably amount of commercial text-to-speech and speech synthesis products were introduced (Klatt 1987). The first integrated circuit for speech synthesis was probably the Votrax chip which consisted of cascade formant synthesizer and simple low-pass smoothing circuits. In 1978 Richard Gagnon introduced an inexpensive Votrax-based Type-n-Talk system (track 28). Two years later, in 1980, Texas Instruments introduced linear prediction coding (LPC) based Speak-n-Spell synthesizer based on low-cost linear prediction synthesis chip (TMS-5100). It was used for an electronic reading aid for children and received quite considerable attention. In 1982 Street Electronics introduced Echo low-cost diphone synthesizer (track 29) which was based on a newer version of the same chip as in Speak-n-Spell (TMS-5220). At the same time Speech Plus Inc. introduced the Prose-2000 text-to-speech system (track 32). A year later, first commercial versions of famous DECtalk (tracks 35-36) and Infovox SA-101 (track 31) synthesizer were introduced (Klatt 1987). Some milestones of speech synthesis development are shown in Figure 2.4.

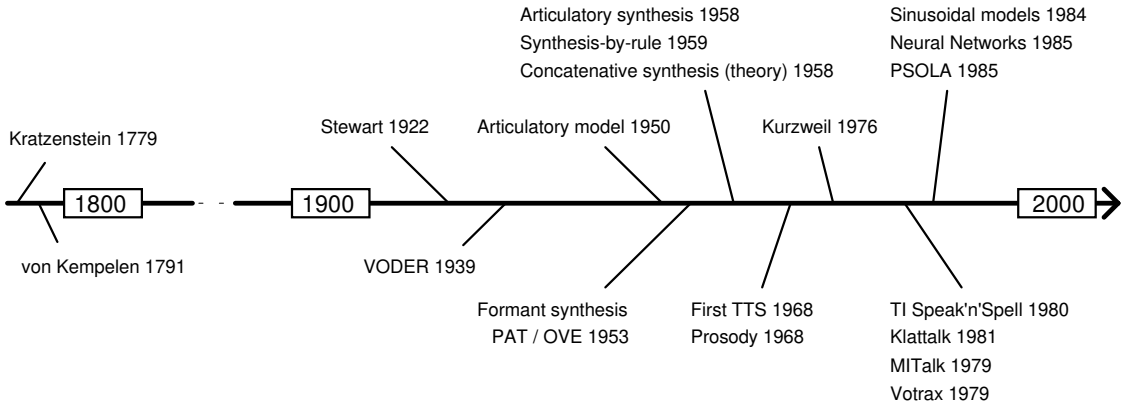

Fig. 2.4. Some milestones in speech synthesis.

Modern speech synthesis technologies involve quite complicated and sophisticated methods and algorithms. One of the methods applied recently in speech synthesis is hidden Markov models (HMM). HMMs have been applied to speech recognition from late 1970's. For speech synthesis systems it has been used for about two decades. A hidden Markov model is a collection of states connected by transitions with two sets of probabilities in each: a transition probability which provides the probability for taking this transition, and an output probability density function (pdf) which defines the conditional probability of emitting each output symbol from a finite alphabet, given that that the transition is taken (Lee 1989).

Neural networks have been applied in speech synthesis for about ten years and the latest results have been quite promising. However, the potential of using neural networks have not been sufficiently explored. Like hidden Markov models, neural networks are also used successfully with speech recognition (Schroeder 1993).

2.3 History of Finnish Speech Synthesis

Although Finnish text corresponds well to its pronunciation and the text preprocessing scheme is quite simple, researchers had paid quite little attention to Finnish TTS before early 1970's. On the other hand, compared to English, the potential number of users and markets are quite small and developing process is time consuming and expensive. However, this potential is increasing with the new multimedia and telecommunication applications.

The first proper speech synthesizer for Finnish, SYNTE2, was introduced in 1977 after five years research in Tampere University of Technology (Karjalainen et al. 1980, Laine 1989). SYNTE2 was also among the first microprocessor based synthesis systems and the first portable TTS system in the world. About five years later an improved SYNTE3 synthesizer was introduced and it was a market leader in Finland for many years. In 1980's, several other commercial systems for Finnish were introduced. For example, Amertronics, Brother Caiku, Eke, Humanica, Seppo, and Task, which all were based on the Votrax speech synthesis chip (Salmensaari 1989).

From present systems, two concatenation-based synthesizers, Mikropuhe and Sanosse, are probably the best known products for Finnish. Mikropuhe has been developed by Timehouse Corporation during last ten years. The first version produced 8-bit sound only from the PC's internal speaker. The latest version is much more sophisticated and described more closely in Chapter 9. Sanosse synthesizer has been developed during last few years for educational purposes for University of Turku and the system is also adopted by Sonera (former Telecom Finland) for their telecommunication applications. Also some multilingual systems including Finnish have been developed during last decades. The best known such system is probably the Infovox synthesizer developed in Sweden. These three systems are perhaps the most dominant products in Finland today (Hakulinen 1998).

3. PHONETICS AND THEORY OF SPEECH PRODUCTION

Speech processing and language technology contains lots of special concepts and terminology. To understand how different speech synthesis and analysis methods work we must have some knowledge of speech production, articulatory phonetics, and some other related terminology. The basic theory of these topics will be discussed briefly in this chapter. For more detailed information, see for example Fant (1970), Flanagan (1972), Witten (1982), O'Saughnessy (1987), or Kleijn et al (1998).

3.1 Representation and Analysis of Speech Signals

Continuous speech is a set of complicated audio signals which makes producing them artificially difficult. Speech signals are usually considered as voiced or unvoiced, but in some cases they are something between these two. Voiced sounds consist of fundamental frequency (F0) and its harmonic components produced by vocal cords (vocal folds). The vocal tract modifies this excitation signal causing formant (pole) and sometimes antiformant (zero) frequencies (Witten 1982). Each formant frequency has also an amplitude and bandwidth and it may be sometimes difficult to define some of these parameters correctly. The fundamental frequency and formant frequencies are probably the most important concepts in speech synthesis and also in speech processing in general.

With purely unvoiced sounds, there is no fundamental frequency in excitation signal and therefore no harmonic structure either and the excitation can be considered as white noise. The airflow is forced through a vocal tract constriction which can occur in several places between glottis and mouth. Some sounds are produced with complete stoppage of airflow followed by a sudden release, producing an impulsive turbulent excitation often followed by a more protracted turbulent excitation (Kleijn et al. 1998). Unvoiced sounds are also usually more silent and less steady than voiced ones. The differences between these are easy to see from Figure 3.2 where the second and last sounds are voiced and the others unvoiced. Whispering is the special case of speech. When whispering a voiced sound there is no fundamental frequency in the excitation and the first formant frequencies produced by vocal tract are perceived.

Speech signals of the three vowels (/a/ /i/ /u/) are presented in time- and frequency domain in Figure 3.1. The fundamental frequency is about 100 Hz in all cases and the formant frequencies F1, F2, and F3 with vowel /a/ are approximately 600 Hz, 1000 Hz, and 2500 Hz respectively. With vowel /i/ the first three formants are 200 Hz, 2300 Hz,

and 3000 Hz, and with /u/ 300 Hz, 600 Hz, and 2300 Hz. The harmonic structure of the excitation is also easy to perceive from frequency domain presentation.

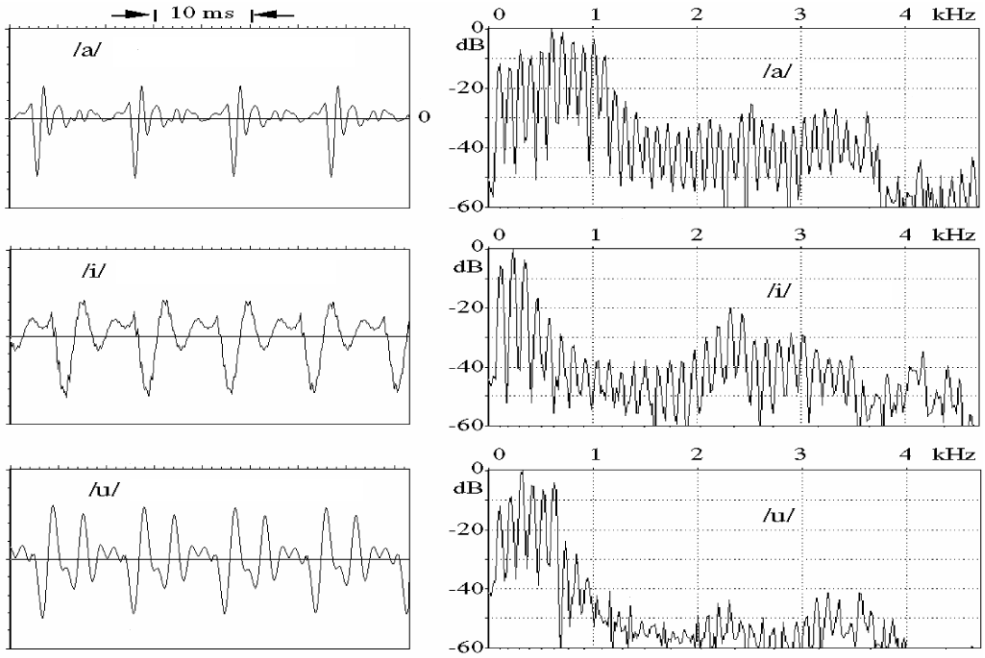

Fig. 3.1. The time- and frequency-domain presentation of vowels /a/, /i/, and /u/.

It can be seen that the first three formants are inside the normal telephone channel (from 300 Hz to 3400 Hz) so the needed bandwidth for intelligible speech is not very wide. For higher quality, up to 10 kHz bandwidth may be used which leads to 20 kHz sampling frequency. Unless, the fundamental frequency is outside the telephone channel, the human hearing system is capable to reconstruct it from its harmonic components.

Another commonly used method to describe a speech signal is the *spectrogram* which is a time-frequency-amplitude presentation of a signal. The spectrogram and the time-domain waveform of Finnish word *kaksi* (two) are presented in Figure 3.2. Higher amplitudes are presented with darker gray-levels so the formant frequencies and trajectories are easy to perceive. Also spectral differences between vowels and consonants are easy to comprehend. Therefore, spectrogram is perhaps the most useful presentation for speech research. From Figure 3.2 it is easy to see that vowels have more energy and it is focused at lower frequencies. Unvoiced consonants have considerably less energy and it is usually focused at higher frequencies. With voiced consonants the situation is something between of these two. In Figure 3.2 the frequency axis is in kilohertz, but it is also quite common to use an auditory spectrogram where the frequency axis is replaced with Bark- or Mel-scale which is normalized for hearing properties.

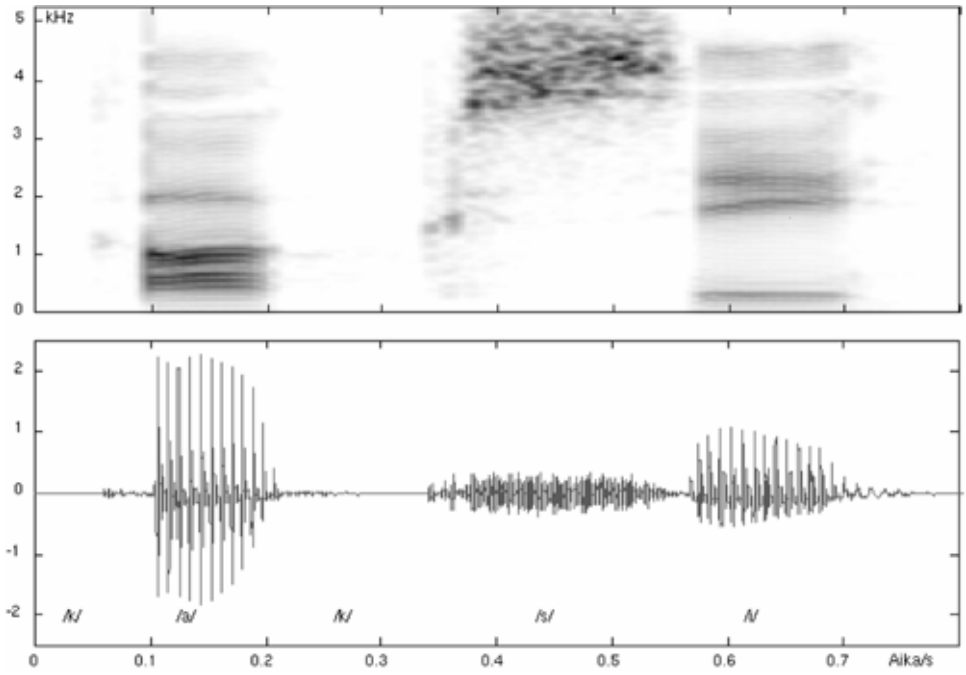

Fig. 3.2. Spectrogram and time-domain presentation of Finnish word *kaksi* (two).

For determining the fundamental frequency or pitch of speech, for example a method called cepstral analysis may be used (Cawley 1996, Kleijn et al. 1998). Cepstrum is obtained by first windowing and making Discrete Fourier Transform (DFT) for the signal and then logaritmizing power spectrum and finally transforming it back to the time-domain by Inverse Discrete Fourier Transform (IDFT). The procedure is shown in Figure 3.3.

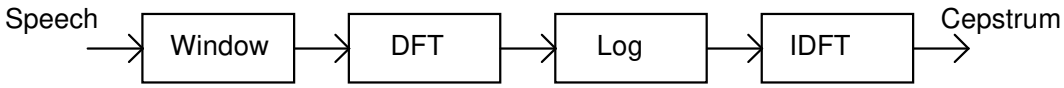

Fig. 3.3. Cepstral analysis.

Cepstral analysis provides a method for separating the vocal tract information from excitation. Thus the reverse transformation can be carried out to provide smoother power spectrum known as homomorphic filtering.

Fundamental frequency or intonation contour over the sentence is important for correct prosody and natural sounding speech. The different contours are usually analyzed from natural speech in specific situations and with specific speaker characteristics and then applied to rules to generate the synthetic speech. The fundamental frequency contour can be viewed as the composite set of hierarchical patterns shown in Figure 3.4. The overall contour is generated by the superposition of these patterns (Sagisaga 1990). Methods for controlling the fundamental frequency contours are described later in Chapter 5.

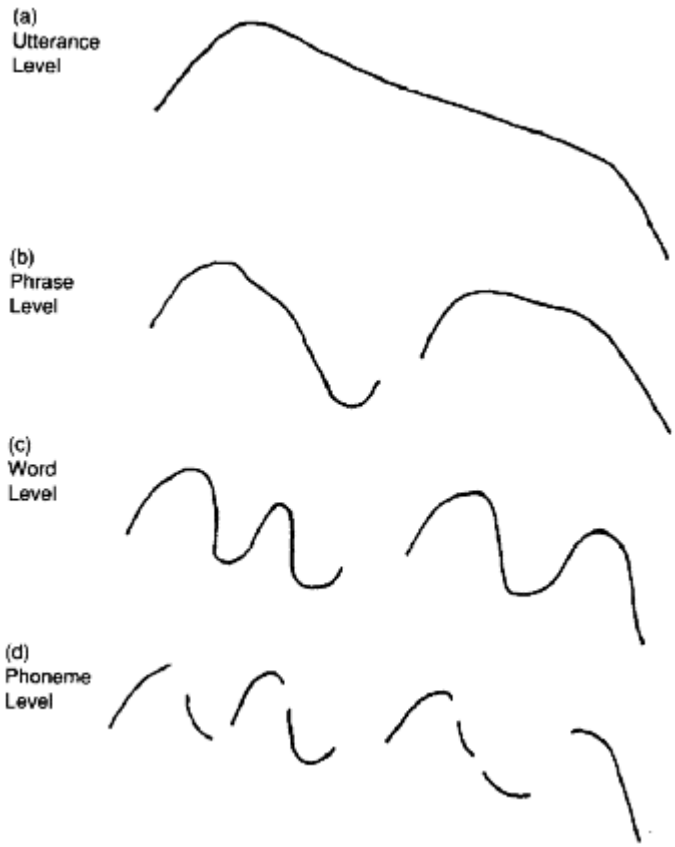

Fig. 3.4. Hierarchical levels of fundamental frequency (Sagisaga 1990).

3.2 Speech Production

Human speech is produced by vocal organs presented in Figure 3.5. The main energy source is the lungs with the diaphragm. When speaking, the air flow is forced through the glottis between the vocal cords and the larynx to the three main cavities of the vocal tract, the pharynx and the oral and nasal cavities. From the oral and nasal cavities the air flow exits through the nose and mouth, respectively. The V-shaped opening between the vocal cords, called the glottis, is the most important sound source in the vocal system. The vocal cords may act in several different ways during speech. The most important function is to modulate the air flow by rapidly opening and closing, causing buzzing sound from which vowels and voiced consonants are produced. The fundamental frequency of vibration depends on the mass and tension and is about 110 Hz, 200 Hz, and 300 Hz with men, women, and children, respectively. With stop consonants the vocal cords may act suddenly from a completely closed position in which they cut the air flow completely, to totally open position producing a light cough or a glottal stop. On the other hand, with unvoiced consonants, such as /s/ or /f/, they may be completely open. An intermediate position may also occur with for example phonemes like /h/.

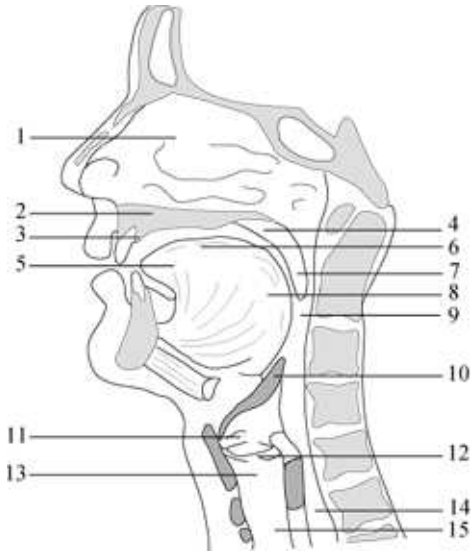

Fig. 3.5. The human vocal organs. (1) Nasal cavity, (2) Hard palate, (3) Alveoral ridge, (4) Soft palate (Velum), (5) Tip of the tongue (Apex), (6) Dorsum, (7) Uvula, (8) Radix, (9) Pharynx, (10) Epiglottis, (11) False vocal cords, (12) Vocal cords, (13) Larynx, (14) Esophagus, and (15) Trachea.

The pharynx connects the larynx to the oral cavity. It has almost fixed dimensions, but its length may be changed slightly by raising or lowering the larynx at one end and the soft palate at the other end. The soft palate also isolates or connects the route from the nasal cavity to the pharynx. At the bottom of the pharynx are the epiglottis and false vocal cords to prevent food reaching the larynx and to isolate the esophagus acoustically from the vocal tract. The epiglottis, the false vocal cords and the vocal cords are closed during swallowing and open during normal breathing.

The oral cavity is one of the most important parts of the vocal tract. Its size, shape and acoustics can be varied by the movements of the palate, the tongue, the lips, the cheeks and the teeth. Especially the tongue is very flexible, the tip and the edges can be moved independently and the entire tongue can move forward, backward, up and down. The lips control the size and shape of the mouth opening through which speech sound is radiated. Unlike the oral cavity, the nasal cavity has fixed dimensions and shape. Its length is about 12 cm and volume 60 cm^3. The air stream to the nasal cavity is controlled by the soft palate.

From technical point of view, the vocal system may be considered as a single acoustic tube between the glottis and mouth. Glottal excited vocal tract may be then approximated as a straight pipe closed at the vocal cords where the acoustical

impedance $Z_g = \infty$ and open at the mouth ($Z_m = 0$). In this case the volume-velocity transfer function of vocal tract is (Flanagan 1972, O'Saughnessy 1987)

$$V(\omega) = \frac{Z_m}{Z_g} = \frac{U_m}{U_g} = \frac{1}{\cos(\frac{\omega l}{c})} \quad , \tag{3.1}$$

where l is the length of the tube, ω is radian frequency and c is sound velocity. The denominator is zero at frequencies $F_i = \omega_i/2\pi$ (i=1,2,3,...), where

$$\frac{\omega_i l}{c} = (2i-1)\frac{\pi}{2}, \text{ and} \qquad F_i = \frac{(2i-1)}{4l}, \tag{3.2}$$

If l=17 cm, $V(\omega)$ is infinite at frequencies F_i = 500, 1500, 2500,... Hz which means resonances every 1 kHz starting at 500 Hz. If the length l is other than 17 cm, the frequencies F_i will be scaled by factor $17/l$ so the vocal tract may be approximated with two or three sections of tube where the areas of adjacent sections are quite different and resonances can be associated within individual cavities. Vowels can be approximated with a two-tube model presented on the left in Figure 3.6. For example, with vowel /a/ the narrower tube represents the pharynx opening into wider tube representing the oral cavity. If assumed that both tubes have an equal length of 8.5 cm, formants occur at twice the frequencies noted earlier for a single tube. Due to acoustic coupling, formants do not approach each other by less than 200 Hz so formants F1 and F2 for /a/ are not both at 1000 Hz, but rather 900 Hz and 1100 Hz, respectively (O'Saughnessy 1987).

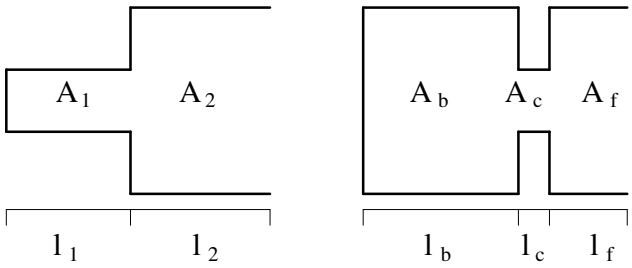

Fig. 3.6. Examples of two- and three-tube models for the vocal tract.

Consonants can be approximated similarly with a three-tube model shown on the right in Figure 3.5., where the narrow middle tube models the vocal tract constriction. The back and middle tubes are half-wavelength resonators and the front tube is a quarter-wavelength resonator with resonances

$$\frac{ci}{2l_b}, \frac{ci}{2l_c}, \frac{c(2i-1)}{4l_f}, \qquad \text{for i = 1, 2, 3,...} \tag{3.3}$$

where l_b, l_c, and l_f are the length of the back, center, and front tube, respectively. With the typical constriction length of 3 cm the resonances occur at multiples of 5333 Hz and can be ignored in applications that use less than 5 kHz bandwidth (O'Saughnessy 1987).

The excitation signal may be modeled with a two-mass model of the vocal cords which consists of two masses coupled with a spring and connected to the larynx by strings and dampers (Fant 1970, Veldhuis et al. 1995).

Several other methods and systems have been developed to model the human speech production system to produce synthetic speech. These methods are related with articulatory synthesis described in Chapter 5. The speech production system, models, and theory are described more closely in Fant (1970), Flanagan (1972), Witten (1982), and O'Saughnessy (1987).

3.3 Phonetics

In most languages the written text does not correspond to its pronunciation so that in order to describe correct pronunciation some kind of symbolic presentation is needed. Every language has a different phonetic alphabet and a different set of possible phonemes and their combinations. The number of phonetic symbols is between 20 and 60 in each language (O'Saughnessy 1987). A set of phonemes can be defined as the minimum number of symbols needed to describe every possible word in a language. In English there are about 40 phonemes (Breen et al. 1996, Donovan 1996). Due to complexity and different kind of definitions, the number of phonemes in English and most of the other languages can not be defined exactly.

Phonemes are abstract units and their pronunciation depends on contextual effects, speaker's characteristics, and emotions. During continuous speech, the articulatory movements depend on the preceding and the following phonemes. The articulators are in different position depending on the preceding one and they are preparing to the following phoneme in advance. This causes some variations on how the individual phoneme is pronounced. These variations are called allophones which are the subset of phonemes and the effect is known as coarticulation. For example, a word *lice* contains a light /l/ and *small* contains a dark /l/. These l's are the same phoneme but different allophones and have different vocal tract configurations. Another reason why the phonetic representation is not perfect, is that the speech signal is always continuous and phonetic notation is always discrete (Witten 1982). Different emotions and speaker characteristics are also impossible to describe with phonemes so the unit called phone is usually defined as an acoustic realization of a phoneme (Donovan 1996).

The phonetic alphabet is usually divided in two main categories, vowels and consonants. Vowels are always voiced sounds and they are produced with the vocal cords in vibration, while consonants may be either voiced or unvoiced. Vowels have considerably higher amplitude than consonants and they are also more stable and easier to analyze and describe acoustically. Because consonants involve very rapid changes

they are more difficult to synthesize properly. The articulatory phonetics in English and Finnish are described more closely in the end of this chapter.

Some efforts to construct language-independent phonemic alphabets were made during last decades. One of the best known is perhaps IPA (International Phonetic Alphabeth) which consists of a huge set of symbols for phonemes, suprasegmentals, tones/word accent contours, and diacritics. For example, there are over twenty symbols for only fricative consonants (IPA 1998). Complexity and the use of Greek symbols makes IPA alphabet quite unsuitable for computers which usually requires standard ASCII as input. Another such kind of phonetic set is SAMPA (Speech Assessment Methods - Phonetic Alphabet) which is designed to map IPA symbols to 7-bit printable ASCII characters. In SAMPA system, the alphabets for each language are designed individually. Originally it covered European Communities languages, but the objective is to make it possible to produce a machine-readable phonetic transcription for every known human language. Alphabet known as Worldbet is another ASCII presentation for IPA symbols which is very similar to SAMPA (Altosaar et al. 1996). American linguists have developed the Arpabet phoneme alphabet to represent American English phonemes using normal ASCII characters. For example a phonetic representation in DECtalk system is based on IPA and Arpabet with some modifications and additional characters (Hallahan 1996). Few examples of different phonetic notations are given in Table 3.1.

Table 3.1. Examples of different phonetic notations.

IPA	IPA-ASCII	SAMPA	DECtalk	Example
i	i	i:	iy	beet
I	I	I	ih	bit
ɛ	E	e	ey	bet
æ	&	{	ae	at
ə	@	@	ax	about
ʌ	V	V	ah	but

Several other phonetic representations and alphabets are used in present systems. For example MITalk uses a set of almost 60 two-character symbols for describing phonetic segments in it (Allen et al. 1987) and it is quite common that synthesis systems use the alphabet of their own. There is still no single generally accepted phonetic alphabet.

3.3.1 English Articulatory Phonetics

Unlike in Finnish articulatory phonetics, discussed in the next chapter, the number of phonetic symbols used in English varies by different kind of definitions. Usually there are about ten to fifteen vowels and about twenty to twenty-five consonants.

English vowels may be classified by the manner or place of articulation (front-back) and by the shape of the mouth (open - close). Main vowels in English and their classification are described in Figure 3.7 below. Sometimes also some diphthongs like /ou/ in *tone* or /ei/ in *take* are described separately. Other versions of definitions of English vowels may be found for example in Rossing (1990) and O'Saughnessy (1987).

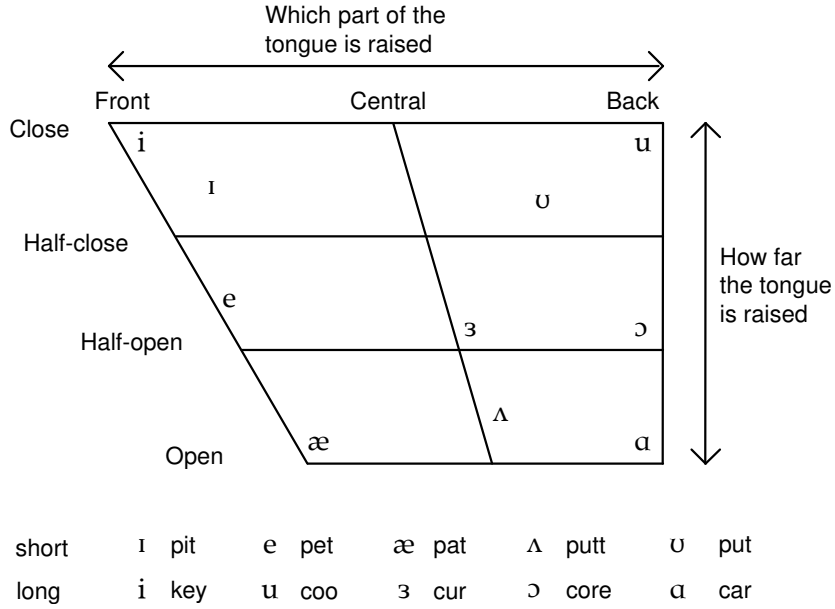

Fig. 3.7. The classification of the main vowels in English (Cawley 1996).

English consonants may be classified by the manner of articulation as plosives, fricatives, nasals, liquids, and semivowels (Cawley 1990, O'Saughnessy 1987). Plosives are known also as stop consonants. Liquids and semivowels are also defined in some publications as approximants and laterals. Further classification may be made by the place of articulation as labials (lips), dentals (teeth), alveolars (gums), palatals (palate), velars (soft palate), glottal (glottis), and labiodentals (lips and teeth). Classification of English consonants is summarized in Figure 3.8.

place / manner	labial	labio-dental	dental	alveolar	palate-alveoral	palatal	velar	glottal
plosive	p b			t d			k g	
fricative		f v	q ð	s z	ʒ ʃ			h
nasal	m			n			ŋ	
liquid				r l				
semivowel	w					j		

Fig. 3.8. Classification of English consonants (Cawley 1996).

Some sounds such as /dʒ/ (as in *gin* or *judge*) and /tʃ/ (as in *chin* or *church*) are sometimes included as separate consonants usually named as affricates (Rossing 1990). Finally, consonants may be classified as voiced and unvoiced. Voiced consonants are /b/, /d/, /g/, /v/, /z/, /ʒ/, /ð/, /l/, /r/, and /j/, others are unvoiced.

3.3.2 Finnish Articulatory Phonetics

There are eight vowels in Finnish. These vowels can be divided into different categories depending how they are formulated: Front/back position of tongue, wideness/roundness of the constriction position, place of the tongue (high or low), and how open or close the mouth is during articulation. Finnish vowels and their categorization are summarized in Figure 3.9.

Vowels		front		back	
		wide	round	wide	round
Close	high	i	y		u
Close-mid Open-mid	mid	e	ö		o
Open	low	ä		a	

Fig. 3.9. Classification of Finnish vowels.

Finnish consonants can be divided into the following categories depending on the place and the manner of articulation:

1. Plosives or stop consonants: /k, p, t, g, b, d/. The vocal tract is closed causing stop or attentuaned sound. When the tract reopens, it causes noise-like, impulse-like or burst sound.
2. Fricatives: /f, h, s/. The vocal tract is constricted in some place so the turbulent air flow causes noise which is modified by the vocal tract resonances. Finnish fricatives are unvoiced.
3. Nasals: /n, m, ŋ/. The vocal tract is closed but the velum opens a route to the nasal cavity. The generated voiced sound is affected by both vocal and nasal tract.
4. Tremulants: /r/. Top of the tongue is vibrating quickly (20 - 25 Hz) against the alveoral ridge causing voiced sound with an effect like amplitude modulation.
5. Laterals: /l/. The top of the tongue closes the vocal tract leaving a sideroute for the air flow.
6. Semivowels: /j, v/. Semivowels are almost like vowels, but they are more unstable and not as context-free as normal vowels.

The consonant categories are summarized in Figure 3.10. For example, for phoneme /p/, the categorization will be unvoiced bilabial-plosive.

Consonants	labial		dental alveoral			palatal	velar	laryng.
	bi-lab.	labio-dent.	pro	medio	post			
plosive (tenuis)	p		t				k	
plosive (media)	b			d			g	
fricative (sibilants)			s					
fricative (spirants)		f						h
nasal	m		n				ŋ	
tremulant			r					
lateral			l					
semivowel		v				j		

Fig. 3.10. Classification of Finnish consonants.

When synthesizing consonants, better results may be achieved by synthesizing these six consonant groups with separate methods because of different acoustic characteristics. Especially the tremulant /r/ needs a special attention.

4. PROBLEMS IN SPEECH SYNTHESIS

The problem area in speech synthesis is very wide. There are several problems in text pre-processing, such as numerals, abbreviations, and acronyms. Correct prosody and pronunciation analysis from written text is also a major problem today. Written text contains no explicit emotions and pronunciation of proper and foreign names is sometimes very anomalous. At the low-level synthesis, the discontinuities and contextual effects in wave concatenation methods are the most problematic. Speech synthesis has been found also more difficult with female and child voices. Female voice has a pitch almost twice as high as with male voice and with children it may be even three times as high. The higher fundamental frequency makes it more difficult to estimate the formant frequency locations (Klatt 1987, Klatt et al. 1990). The evaluation and assessment of synthesized speech is neither a simple task. Speech quality is a multidimensional term and the evaluation method must be chosen carefully to achieve desired results. This chapter describes the major problems in text-to-speech research.

4.1 Text-to-Phonetic Conversion

The first task faced by any TTS system is the conversion of input text into linguistic representation, usually called text-to-phonetic or grapheme-to-phoneme conversion. The difficulty of conversion is highly language depended and includes many problems. In some languages, such as Finnish, the conversion is quite simple because written text almost corresponds to its pronunciation. For English and most of the other languages the conversion is much more complicated. A very large set of different rules and their exceptions is needed to produce correct pronunciation and prosody for synthesized speech. Some languages have also special features which are discussed more closely at the end of this chapter. Conversion can be divided in three main phases, text preprocessing, creation of linguistic data for correct pronunciation, and the analysis of prosodic features for correct intonation, stress, and duration.

4.1.1 Text preprocessing

Text preprocessing is usually a very complex task and includes several language dependent problems (Sproat 1996). Digits and numerals must be expanded into full words. For example in English, numeral 243 would be expanded as *two hundred and forty-three* and 1750 as *seventeen-fifty* (if year) or *one-thousand seven-hundred and fifty* (if measure). Related cases include the distinction between *the 747 pilot* and *747 people*. Fractions and dates are also problematic. 5/16 can be expanded as *five-sixteenths* (if fraction) or *May sixteenth* (if date). Expansion ordinal numbers have been found also

problematic. The first three ordinals must be expanded differently than the others, 1st as *first*, 2nd as *second*, and 3rd as *third*. Same kind of contextual problems are faced with roman numerals. Chapter III should be expanded as *Chapter three* and Henry III as *Henry the third* and *I* may be either a pronoun or number. Roman numerals may be also confused with some common abbreviations, such as MCM. Numbers may also have some special forms of expression, such as 22 as *double two* in telephone numbers and 1-0 as *one love* in sports.

Abbreviations may be expanded into full words, pronounced as written, or pronounced letter by letter (Macon 1996). There are also some contextual problems. For example kg can be either *kilogram* or *kilograms* depending on preceding number, St. can be *saint* or *street*, Dr. *doctor* or *drive* and ft. *Fort*, *foot* or *feet*. In some cases, the adjacent information may be enough to find out the correct conversion, but to avoid misconversions the best solution in some cases may be the use of letter-to-letter conversion. Innumerable abbreviations for company names and other related things exists and they may be pronounced in many ways. For example, N.A.T.O. or RAM are usually pronounced as written and SAS or ADP letter-by-letter. Some abbreviations such as MPEG as *empeg* are pronounced irregularly.

Special characters and symbols, such as '$', '%', '&', '/', '-', '+', cause also special kind of problems. In some situations the word order must be changed. For example, *$71.50* must be expanded as *seventy-one dollars and fifty cents* and *$100 million* as *one hundred million dollars*, not as *one hundred dollars million*. The expression '1-2' may be expanded as *one minus two* or *one two*, and character '&' as *et* or *and*. Also special characters and character strings in for example web-sites or e-mail messages must be expanded with special rules. For example, character '@' is usually converted as *at* and e-mail messages may contain character strings, such as some header information, which may be omitted. Some languages also include special non ASCII characters, such as accent markers or special symbols.

Written text may also be constructed in several ways, like in several columns and pages as in a normal newspaper article. This may cause insuperable problems especially with optical reading machines.

In Finnish, the text preprocessing scheme is in general easier but contains also some specific difficulties. Especially with numerals and ordinals expansion may be even more difficult than in other languages due to several cases constructed by several different suffixes. The two first ordinals must be expanded differently in some cases and with larger numbers the expansion may become rather complex. With digits, roman numerals, dates, and abbreviations same kind of difficulties are faced as in other languages. For example, for Roman numerals I and III, there is at least three possible

conversion. Some examples of the most difficult abbreviations are given in Table 4.1. In most cases, the correct conversion may be concluded from the type of compounding characters or from other compounding information. But to avoid misconversions, some abbreviations must be spelled letter-by-letter.

Table 4.1. Some examples of the text parsing difficulties for Finnish in some contexts.

Text	Different possibilities in different contexts
s	sekuntia, sivua, syntynyt
kg	(1) kilogramma, (2) kilogrammaa
III	kolmos (olut), (Kaarle) kolmas, (luku) kolme
mm	millimetriä, muun muassa
min	minimi, minuuttia
huom	huomenna, huomio
kk	kuukausi, kuukautta, keittokomero
os.	osoite, omaa sukua, osasto

In Finnish the conversion of common numbers is probably more complicated than in English. The suffixes, such as *s* in ordinals are included after every single number. For example ordinal *1023.* is pronounced as "tuhanne*s* kahdeskymmene*s* kolma*s*". In some cases, the conversion of a number may be concluded from the suffix of the following word, but sometimes the situation may be very ambiguous which can be seen from the following examples:

- 100 (sadan) markan alennus. - 100 (sata) markan rahaa.
- 15 (viisitoista) koiran omistajaa. - 15 (viidentoista) koiran ryhmä.
- halasin 5 (viittä) henkilöä. - kohtasin 5 (viisi) henkilöä.

It is easy to see from previous examples that, for correct conversion in every possible situation, a very complicated set of rules is needed.

4.1.2 Pronunciation

The second task is to find correct pronunciation for different contexts in the text. Some words, called *homographs*, cause maybe the most difficult problems in TTS systems. Homographs are spelled the same way but they differ in meaning and usually in pronunciation (e.g. fair, lives). The word *lives* is for example pronounced differently in sentences "Three *lives* were lost" and "One *lives* to eat". Some words, e.g. *lead*, has different pronunciations when used as a verb or noun, and between two noun senses (He followed her *lead* / He covered the hull with *lead*). With these kind of words some semantical information is necessary to achieve correct pronunciation.

The pronunciation of a certain word may also be different due to contextual effects. This is easy to see when comparing phrases *the end* and *the beginning*. The pronunciation of *the* depends on the initial phoneme in the following word. Compound words are also problematic. For example the characters 'th' in mother and hothouse is pronounced differently. Some sounds may also be either voiced or unvoiced in different context. For example, phoneme /s/ in word *dogs* is voiced, but unvoiced in word *cats* (Allen et al. 1987).

Finding correct pronunciation for proper names, especially when they are borrowed from other languages, is usually one of the most difficult tasks for any TTS system. Some common names, such as Nice and Begin, are ambiguous in capitalized context, including sentence initial position, titles and single text. For example, the sentence *Nice is a nice place* is very problematic because the word *Nice* may be pronounced as /niis/ or /nais/. Some names and places have also special pronunciation, such as Leicester and Arkansas. For correct pronunciation, these kind of words may be included in a specific exception dictionary. Unfortunately, it is clear that there is no way to build a database of all proper names in the world.

In Finnish, considerably less rules are needed because in most cases words are pronounced as written. However, few exceptions exists, such as /ŋ/ in words *kenkä* and *kengät*. Finnish alphabet contains also some foreign origin letters which can be converted in text preprocessing, such as taxi - taksi (x - ks) and pizza (zz - ts). The letter pairs v and w, c and s, or å and o are also usually pronounced the same way (Karjalainen 1978).

4.1.3 Prosody

Finding correct intonation, stress, and duration from written text is probably the most challenging problem for years to come. These features together are called prosodic or suprasegmental features and may be considered as the melody, rhythm, and emphasis of the speech at the perceptual level. The intonation means how the pitch pattern or fundamental frequency changes during speech. The prosody of continuous speech depends on many separate aspects, such as the meaning of the sentence and the speaker characteristics and emotions. The prosodic dependencies are shown in Figure 4.1. Unfortunately, written text usually contains very little information of these features and some of them change dynamically during speech. However, with some specific control characters this information may be given to a speech synthesizer.

Timing at sentence level or grouping of words into phrases correctly is difficult because prosodic phrasing is not always marked in text by punctuation, and phrasal accentuation is almost never marked (Santen et al. 1997). If there is no breath pauses in speech or if they are in wrong places, the speech may sound very unnatural or even the meaning of

the sentence may be misunderstood. For example, the input string "John says Peter is a liar" can be spoken as two different ways giving two different meanings as "John says: Peter is a liar" or "John, says Peter, is a liar". In the first sentence Peter is a liar, and in the second one the liar is John.

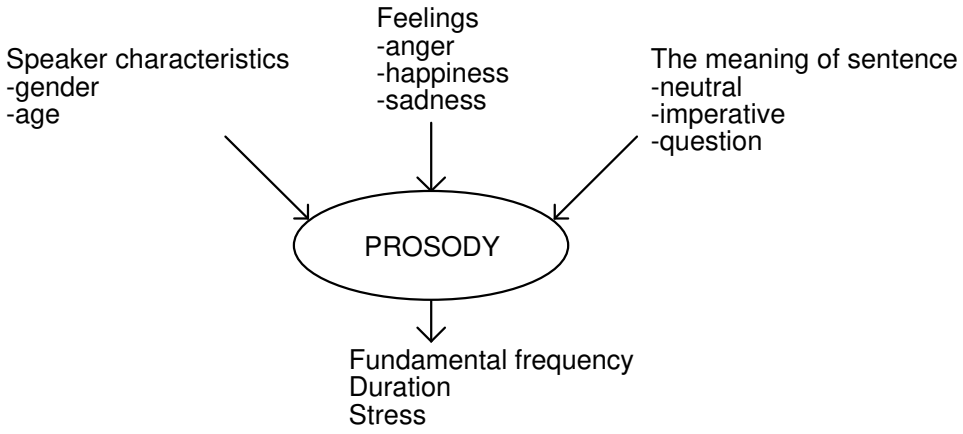

Fig. 4.1. Prosodic dependencies.

4.2 Problems in Low Level Synthesis

There are many methods to produce speech sounds after text and prosodic analysis. All these methods have some benefits and problems of their own.

In articulatory synthesis (see 5.1), the collection of data and implementation of rules to drive that data correctly is very complex. It is almost impossible to model masses, tongue movements, or other characteristics of the vocal system perfectly. Due to this complexity, the computational load may increase considerably.

In formant synthesis (see 5.2), the set of rules controlling the formant frequencies and amplitudes and the characteristics of the excitation source is large. Also some lack of naturalness, especially with nasalized sounds, is considered a major problem with formant synthesis.

In concatenative synthesis (see 5.3), the collecting of speech samples and labeling them is very time-consuming and may yield quite large waveform databases. However, the amount of data may be reduced with some compression method. Concatenation points between samples may cause distortion to the speech. With some longer units, such as words or syllables, the coarticulation effect is a problem and some problems with memory and system requirements may arise.

4.3 Language Specific Problems and Features

For certain languages synthetic speech is easier to produce than in others. Also, the amount of potential users and markets are very different with different countries and languages which also affects how much resources are available for developing speech synthesis. Most of languages have also some special features which can make the development process either much easier or considerably harder.

Some languages, such as Finnish, Italian, and Spanish, have very regular pronunciation. Sometimes there is almost one-to-one correspondence with letter to sound. The other end is for example French with very irregular pronunciation. Many languages, such as French, German, Danish and Portuguese also contain lots of special stress markers and other non ASCII characters (Oliveira et al. 1992). In German, the sentential structure differs largely from other languages. For text analysis, the use of capitalized letters with nouns may cause some problems because capitalized words are usually analyzed differently than others.

In Japanese, almost every spoken syllable is in CV form which makes the synthesis a bit easier than with other languages. On the other hand, conversion from Kanji to Kana symbols must be performed when using a TTS system (Hirokawa 1989). In Chinese and many other Asian languages which are based on non ASCII alphabet, words are not delimited with whitespace (space, tab etc.) and word boundaries must therefore be reconstructed for such languages separately (Santen et al. 1997). However, these languages usually contain a designated symbol as sentence delimiter which makes the end-of-the-sentence detection easier, unlike in English where the period may be the sentence delimiter or used to mark abbreviation (Kleijn et al. 1998). In some tone languages, such as Chinese, the intonation may be even used to change the meaning of the word (Breen 1992).

5. METHODS, TECHNIQUES, AND ALGORITHMS

Synthesized speech can be produced by several different methods. All of these have some benefits and deficiencies that are discussed in this and previous chapters. The methods are usually classified into three groups:

- Articulatory synthesis, which attempts to model the human speech production system directly.
- Formant synthesis, which models the pole frequencies of speech signal or transfer function of vocal tract based on source-filter-model.
- Concatenative synthesis, which uses different length prerecorded samples derived from natural speech.

The formant and concatenative methods are the most commonly used in present synthesis systems. The formant synthesis was dominant for long time, but today the concatenative method is becoming more and more popular. The articulatory method is still too complicated for high quality implementations, but may arise as a potential method in the future.

5.1 Articulatory Synthesis

Articulatory synthesis tries to model the human vocal organs as perfectly as possible, so it is potentially the most satisfying method to produce high-quality synthetic speech. On the other hand, it is also one of the most difficult methods to implement and the computational load is also considerably higher than with other common methods (Kröger 1992, Rahim et al. 1993). Thus, it has received less attention than other synthesis methods and has not yet achieved the same level of success.

Articulatory synthesis typically involves models of the human articulators and vocal cords. The articulators are usually modeled with a set of area functions between glottis and mouth. The first articulatory model was based on a table of vocal tract area functions from larynx to lips for each phonetic segment (Klatt 1987). For rule-based synthesis the articulatory control parameters may be for example lip aperture, lip protrusion, tongue tip height, tongue tip position, tongue height, tongue position and velic aperture. Phonatory or excitation parameters may be glottal aperture, cord tension, and lung pressure (Kröger 1992).

When speaking, the vocal tract muscles cause articulators to move and change shape of the vocal tract which causes different sounds. The data for articulatory model is usually derived from X-ray analysis of natural speech. However, this data is usually only 2-D

when the real vocal tract is naturally 3-D, so the rule-based articulatory synthesis is very difficult to optimize due to the unavailability of sufficient data of the motions of the articulators during speech. Other deficiency with articulatory synthesis is that X-ray data do not characterize the masses or degrees of freedom of the articulators (Klatt 1987). Also, the movements of tongue are so complicated that it is almost impossible to model them precisely.

Advantages of articulatory synthesis are that the vocal tract models allow accurate modeling of transients due to abrupt area changes, whereas formant synthesis models only spectral behavior (O'Saughnessy 1987). The articulatory synthesis is quite rarely used in present systems, but since the analysis methods are developing fast and the computational resources are increasing rapidly, it might be a potential synthesis method in the future.

5.2 Formant Synthesis

Probably the most widely used synthesis method during last decades has been formant synthesis which is based on the source-filter-model of speech described in Chapter 2. There are two basic structures in general, parallel and cascade, but for better performance some kind of combination of these is usually used. Formant synthesis also provides infinite number of sounds which makes it more flexible than for example concatenation methods.

At least three formants are generally required to produce intelligible speech and up to five formants to produce high quality speech. Each formant is usually modeled with a two-pole resonator which enables both the formant frequency (pole-pair frequency) and its bandwidth to be specified (Donovan 1996).

Rule-based formant synthesis is based on a set of rules used to determine the parameters necessary to synthesize a desired utterance using a formant synthesizer (Allen et al. 1987). The input parameters may be for example the following, where the open quotient means the ratio of the open-glottis time to the total period duration (Holmes et al. 1990):

- Voicing fundamental frequency (F0)
- Voiced excitation open quotient (OQ)
- Degree of voicing in excitation (VO)
- Formant frequencies and amplitudes (F1...F3 and A1...A3)
- Frequency of an additional low-frequency resonator (FN)
- Intensity of low- and high-frequency region (ALF, AHF)

A cascade formant synthesizer (Figure 5.1) consists of band-pass resonators connected in series and the output of each formant resonator is applied to the input of the following one. The cascade structure needs only formant frequencies as control information. The main advantage of the cascade structure is that the relative formant amplitudes for vowels do not need individual controls (Allen et al. 1987).

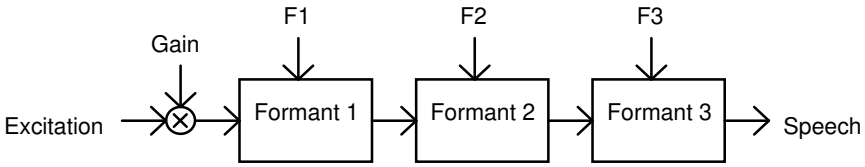

Fig. 5.1. Basic structure of cascade formant synthesizer.

The cascade structure has been found better for non-nasal voiced sounds and because it needs less control information than parallel structure, it is then simpler to implement. However, with cascade model the generation of fricatives and plosive bursts is a problem.

A parallel formant synthesizer (Figure 5.2) consists of resonators connected in parallel. Sometimes extra resonators for nasals are used. The excitation signal is applied to all formants simultaneously and their outputs are summed. Adjacent outputs of formant resonators must be summed in opposite phase to avoid unwanted zeros or antiresonances in the frequency response (O'Saughnessy 1987). The parallel structure enables controlling of bandwidth and gain for each formant individually and thus needs also more control information.

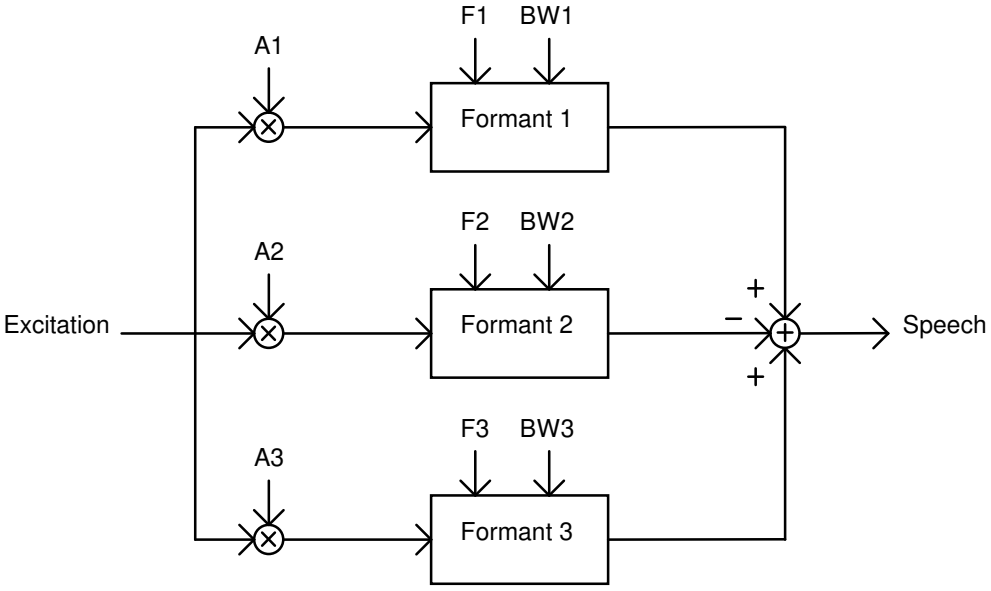

Fig. 5.2. Basic structure of a parallel formant synthesizer.

The parallel structure has been found to be better for nasals, fricatives, and stop-consonants, but some vowels can not be modeled with parallel formant synthesizer as well as with the cascade one.

There has been widespread controversy over the quality and suitably characteristics of these two structures. It is easy to see that good results with only one basic method is difficult to achieve so some efforts have been made to improve and combine these basic models. In 1980 Dennis Klatt (Klatt 1980) proposed a more complex formant synthesizer which incorporated both the cascade and parallel synthesizers with additional resonances and anti-resonances for nasalized sounds, sixth formant for high frequency noise, a bypass path to give a flat transfer function, and a radiation characteristics. The system used quite complex excitation model which was controlled by 39 parameters updated every 5 ms. The quality of Klatt Formant Synthesizer was very promising and the model has been incorporated into several present TTS systems, such as MITalk, DECtalk, Prose-2000, and Klattalk (Donovan 1996). Parallel and cascade structures can also be combined by several other ways. One solution is to use so called PARCAS (Parallel-Cascade) model introduced and patented by Laine (1982) for SYNTE3 speech synthesizer for Finnish. In the model, presented in Figure 5.3, the transfer function of the uniform vocal tract is modeled with two partial transfer functions, each including every second formant of the transfer function. Coefficients k1, k2, and k3 are constant and chosen to balance the formant amplitudes in the neutral vowel to keep the gains of parallel branches constant for all sounds (Laine 1982).

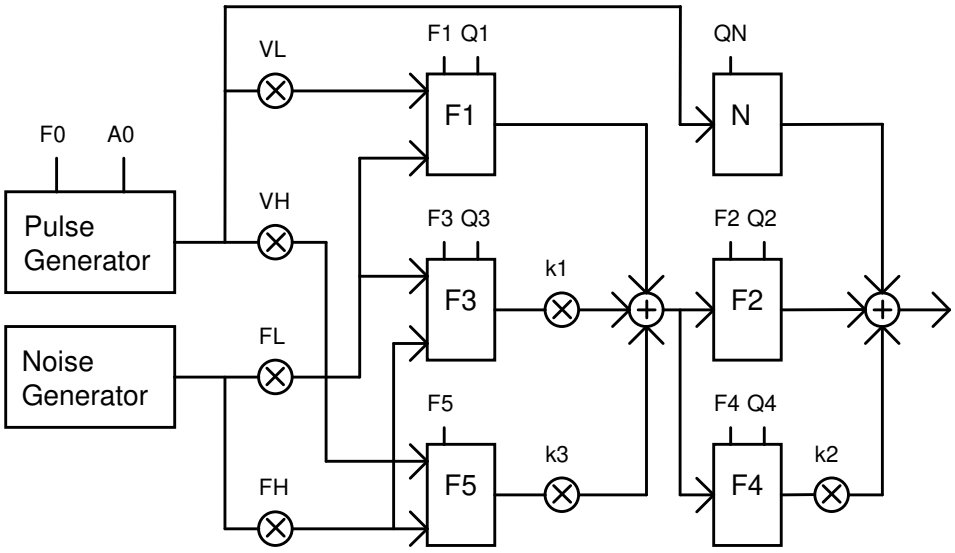

Fig. 5.3. PARCAS model (Laine 1989).

The PARCAS model uses a total of 16 control parameters:

- F0 and A0 - fundamental frequency and amplitude of voiced component
- Fn and Qn - formant frequencies and Q-values (formant frequency / bandwidth)
- VL and VH - voiced component amplitude, low and high
- FL and FH - unvoiced component amplitude, low and high
- QN - Q-value of the nasal formant at 250 Hz

The used excitation signal in formant synthesis consists of some kind of voiced source or white noise. The first voiced source signals used were simple sawtooth type. In 1981 Dennis Klatt introduced more sophisticated voicing source for his Klattalk system (Klatt 1987). The correct and carefully selected excitation is important especially when good controlling of speech characteristics is wanted.

The formant filters represent only the resonances of the vocal tract, so additional provision is needed for the effects of the shape of the glottal waveform and the radiation characteristics of the mouth. Usually the glottal waveform is approximated simply with -12dB/octave filter and radiation characteristics with simple +6dB/octave filter.

5.3 Concatenative Synthesis

Connecting prerecorded natural utterances is probably the easiest way to produce intelligible and natural sounding synthetic speech. However, concatenative synthesizers are usually limited to one speaker and one voice and usually require more memory capacity than other methods.

One of the most important aspects in concatenative synthesis is to find correct unit length. The selection is usually a trade-off between longer and shorter units. With longer units high naturalness, less concatenation points and good control of coarticulation are achieved, but the amount of required units and memory is increased. With shorter units, less memory is needed, but the sample collecting and labeling procedures become more difficult and complex. In present systems units used are usually words, syllables, demisyllables, phonemes, diphones, and sometimes even triphones.

Word is perhaps the most natural unit for written text and some messaging systems with very limited vocabulary. Concatenation of words is relative easy to perform and coarticulation effects within a word are captured in the stored units. However, there is a great difference with words spoken in isolation and in continuos sentence which makes the continuous speech to sound very unnatural (Allen et al. 1987). Because there are hundreds of thousands of different words and proper names in each language, word is not a suitable unit for any kind of unrestricted TTS system.

The number of different syllables in each language is considerably smaller than the number of words, but the size of unit database is usually still too large for TTS systems. For example, there are about 10,000 syllables in English. Unlike with words, the coarticulation effect is not included in stored units, so using syllables as a basic unit is not very reasonable. There is also no way to control prosodic contours over the sentence. At the moment, no word or syllable based full TTS system exists. The current synthesis systems are mostly based on using phonemes, diphones, demisyllables or some kind of combinations of these.

Demisyllables represents the initial and final parts of syllables. One advantage of demisyllables is that only about 1,000 of them is needed to construct the 10,000 syllables of English (Donovan 1996). Using demisyllables, instead of for example phonemes and diphones, requires considerably less concatenation points. Demisyllables also take account of most transitions and then also a large number of coarticulation effects and also covers a large number of allophonic variations due to separation of initial and final consonant clusters. However, the memory requirements are still quite high, but tolerable. Compared to phonemes and diphones, the exact number of demisyllables in a language can not be defined. With purely demisyllable based system, all possible words can not be synthesized properly. This problem is faced at least with some proper names (Hess 1992). However, demisyllables and syllables may be successfully used in a system which uses variable length units and affixes, such as the HADIFIX system described in Chapter 9 (Dettweiler et al. 1985).

Phonemes are probably the most commonly used units in speech synthesis because they are the normal linguistic presentation of speech. The inventory of basic units is usually between 40 and 50, which is clearly the smallest compared to other units (Allen et al. 1987). Using phonemes gives maximum flexibility with the rule-based systems. However, some phones that do not have a steady-state target position, such as plosives, are difficult to synthesize. The articulation must also be formulated as rules. Phonemes are sometimes used as an input for speech synthesizer to drive for example diphone-based synthesizer.

Diphones (or dyads) are defined to extend the central point of the steady state part of the phone to the central point of the following one, so they contain the transitions between adjacent phones. That means that the concatenation point will be in the most steady state region of the signal, which reduces the distortion from concatenation points. Another advantage with diphones is that the coarticulation effect needs no more to be formulated as rules. In principle, the number of diphones is the square of the number of phonemes (plus allophones), but not all combinations of phonemes are needed. For example, in Finnish the combinations, such as /hs/, /sj/, /mt/, /nk/, and /ŋp/ within a word are not possible. The number of units is usually from 1500 to 2000, which increases the

memory requirements and makes the data collection more difficult compared to phonemes. However, the number of data is still tolerable and with other advantages, diphone is a very suitable unit for sample-based text-to-speech synthesis. The number of diphones may be reduced by inverting symmetric transitions, like for example /as/ from /sa/.

Longer segmental units, such as triphones or tetraphones, are quite rarely used. Triphones are like diphones, but contains one phoneme between steady-state points (half phoneme - phoneme - half phoneme). In other words, a triphone is a phoneme with a specific left and right context. For English, more than 10,000 units are required (Huang et al. 1997).

Building the unit inventory consists of three main phases (Hon et al. 1998). First, the natural speech must be recorded so that all used units (phonemes) within all possible contexts (allophones) are included. After this, the units must be labeled or segmented from spoken speech data, and finally, the most appropriate units must be chosen. Gathering the samples from natural speech is usually very time-consuming. However, some is this work may be done automatically by choosing the input text for analysis phase properly. The implementation of rules to select correct samples for concatenation must also be done very carefully.

There are several problems in concatenative synthesis compared to other methods.

- Distortion from discontinuities in concatenation points, which can be reduced using diphones or some special methods for smoothing signal.
- Memory requirements are usually very high, especially when long concatenation units are used, such as syllables or words.
- Data collecting and labeling of speech samples is usually time-consuming. In theory, all possible allophones should be included in the material, but trade-offs between the quality and the number of samples must be made.

Some of the problems may be solved with methods described below and the use of concatenative method is increasing due to better computer capabilities (Donovan 1996).

5.3.1 PSOLA Methods

The PSOLA (Pitch Synchronous Overlap Add) method was originally developed at France Telecom (CNET). It is actually not a synthesis method itself but allows prerecorded speech samples smoothly concatenated and provides good controlling for pitch and duration, so it is used in some commercial synthesis systems, such as ProVerbe and HADIFIX (Donovan 1996).

There are several versions of the PSOLA algorithm and all of them work in essence the same way. Time-domain version, TD-PSOLA, is the most commonly used due to its computational efficiency (Kortekaas et al. 1997). The basic algorithm consist of three steps (Charpentier et al. 1989, Valbret et. al 1991). The analysis step where the original speech signal is first divided into separate but often overlapping short-term analysis signals (ST), the modification of each analysis signal to synthesis signal, and the synthesis step where these segments are recombined by means of overlap-adding. Short term signals $x_m(n)$ are obtained from digital speech waveform $x(n)$ by multiplying the signal by a sequence of pitch-synchronous analysis window $h_m(n)$:

$$x_m(n) = h_m(t_m - n)x(n), \tag{5.1}$$

where m is an index for the short-time signal. The windows, which are usually Hanning type, are centered around the successive instants t_m, called pitch-marks. These marks are set at a pitch-synchronous rate on the voiced parts of the signal and at a constant rate on the unvoiced parts. The used window length is proportional to local pitch period and the window factor is usually from 2 to 4 (Charpentier 1989, Kleijn et al. 1998). The pitch markers are determined either by manually inspection of speech signal or automatically by some pitch estimation methods (Kortekaas et al. 1997). The segment recombination in synthesis step is performed after defining a new pitch-mark sequence.

Manipulation of fundamental frequency is achieved by changing the time intervals between pitch markers (see Figure 5.4). The modification of duration is achieved by either repeating or omitting speech segments. In principle, modification of fundamental frequency also implies a modification of duration (Kortekaas et al. 1997).

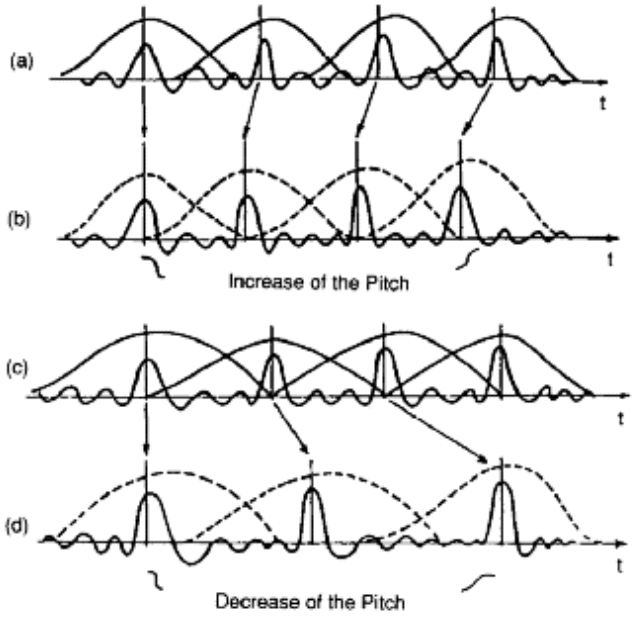

Fig. 5.4. Pitch modification of a voiced speech segment.

Another variations of PSOLA, Frequency Domain PSOLA (FD-PSOLA) and the Linear-Predictive PSOLA (LP-PSOLA), are theoretically more appropriate approaches for pitch-scale modifications because they provide independent control over the spectral envelope of the synthesis signal (Moulines et al. 1995). FD-PSOLA is used only for pitch-scale modifications and LP-PSOLA is used with residual excited vocoders.

Some drawbacks with PSOLA method exists. The pitch can be determined only for voiced sounds and applied to unvoiced signal parts it might generate a tonal noise (Moulines et al. 1990).

5.3.2 Microphonemic Method

The basic idea of the microphonemic method is to use variable length units derived from natural speech (Lukaszewicz et al. 1987). These units may be words, syllables, phonemes (and allophones), pitch periods, transients, or noise segments (Lehtinen et al. 1989). From these segments a dictionary of prototypes is collected.

Prototypes are concatenated in time axis with PSOLA-like method. If the formant distances between consecutive sound segments is less than two critical bandwidths (Barks), the concatenation is made by simple linear amplitude-based interpolation between the prototypes. If the difference is more than two Barks, an extra intermediate prototype must be used because the simple amplitude-based interpolation is not sufficient for perceptually acceptable formant movements (Lukaszewicz et al. 1987). The overlap-add processes of prototypes are shown in Figure 5.5.

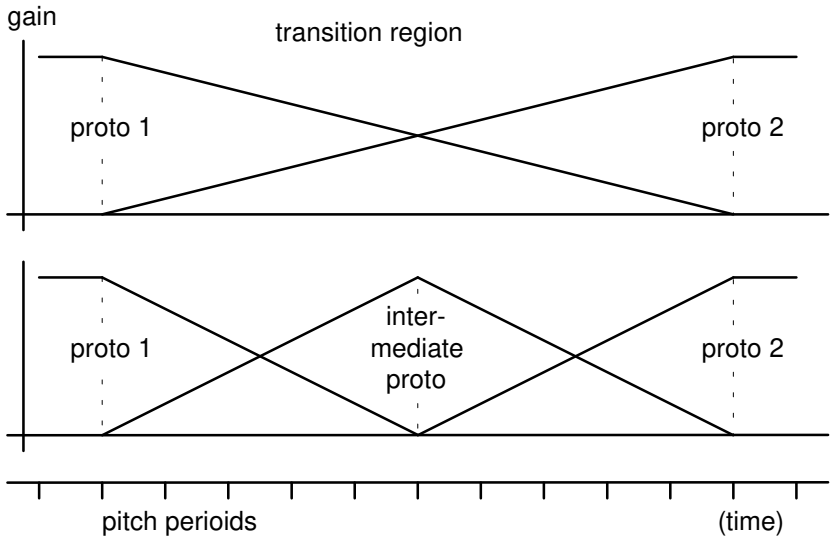

Fig. 5.5. Linear amplitude based interpolation (Lukaszewicz et al. 1997).

Some consonants need special attention. For example, stop consonants can be stored as direct waveform segments as several variants in the different vowel context (Lehtinen et

al. 1989). With fricatives, prototypes of about 50 ms of total length and 10 ms units from them are randomly selected for concatenation with the interpolation method above. Most voiced consonants act like vowels, but the context dependence variability is higher (Lukaszewicz et al. 1987).

The benefits of the microphonemic method is that the computational load and storage requirements are rather low compared to other sample based methods (Lehtinen 1990). The biggest problem, as in other sample based methods, is how to extract the optimal collection of prototypes from natural speech and the developing of rules for concatenating them.

5.4 Linear Prediction based Methods

Linear predictive methods are originally designed for speech coding systems, but may be also used in speech synthesis. In fact, the first speech synthesizers were developed from speech coders (see 2.1). Like formant synthesis, the basic LPC is based on the source-filter-model of speech described in Chapter 1. The digital filter coefficients are estimated automatically from a frame of natural speech.

The basis of linear prediction is that the current speech sample *y(n)* can be approximated or predicted from a finite number of previous *p* samples *y(n-1)* to *y(n-k)* by a linear combination with small error term *e(n)* called residual signal. Thus,

$$y(n) = e(n) + \sum_{k=1}^{p} a(k)\, y(n-k)\,, \tag{5.2}$$

and

$$e(n) = y(n) - \sum_{k=1}^{p} a(k)\, y(n-k) = y(n) - \tilde{y}(n)\ , \tag{5.3}$$

where $\tilde{y}(n)$ is a predicted value, *p* is the linear predictor order, and *a(k)* are the linear prediction coefficients which are found by minimizing the sum of the squared errors over a frame. Two methods, the covariance method and the autocorrelation method, are commonly used to calculate these coefficients. Only with the autocorrelation method the filter is guaranteed to be stable (Witten 1982, Kleijn et al. 1998).

In synthesis phase the used excitation is approximated by a train of impulses for voiced sounds and by random noise for unvoiced. The excitation signal is then gained and filtered with a digital filter for which the coefficients are *a(k)*. The filter order is typically between 10 and 12 at 8 kHz sampling rate, but for higher quality at 22 kHz sampling rate, the order needed is between 20 and 24 (Kleijn et al. 1998, Karjalainen et al. 1998). The coefficients are usually updated every 5-10 ms.

The main deficiency of the ordinary LP method is that it represents an all-pole model, which means that phonemes that contain antiformants such as nasals and nasalized vowels are poorly modeled. The quality is also poor with short plosives because the time-scale events may be shorter than the frame size used for analysis. With these deficiencies the speech synthesis quality with standard LPC method is generally considered poor, but with some modifications and extensions for the basic model the quality may be increased.

Warped Linear Prediction (WLP) takes advantages of human hearing properties and the needed order of filter is then reduced significally, from orders 20-24 to 10-14 with 22 kHz sampling rate (Laine et al. 1994, Karjalainen et al. 1998). The basic idea is that the unit delays in digital filter are replaced by following all-pass sections

$$\tilde{z}^{-1} = D_1(z) = \frac{z^{-1} - \lambda}{1 - \lambda z^{-1}} \ , \tag{5.4}$$

where λ is a warping parameter between -1 and 1 and $D_1(z)$ is a warped delay element and with Bark scale it is $\lambda = 0.63$ with sampling rate of 22 kHz. WLP provides better frequency resolution at low frequencies and worse at high frequencies. However, this is very similar to human hearing properties (Karjalainen et al. 1998).

Several other variations of linear prediction have been developed to increase the quality of the basic method (Childers et al. 1994, Donovan 1996). With these methods the used excitation signal is different from ordinary LP method and the source and filter are no longer separated. These kind of variations are for example multipulse linear prediction (MLPC) where the complex excitation is constructed from a set of several pulses, residual excited linear prediction (RELP) where the error signal or residual is used as an excitation signal and the speech signal can be reconstructed exactly, and code excited linear prediction (CELP) where a finite number of excitations used are stored in a finite codebook (Campos et al. 1996).

5.5 Sinusoidal Models

Sinusoidal models are based on a well known assumption that the speech signal can be represented as a sum of sine waves with time-varying amplitudes and frequencies (McAulay et al. 1986, Macon 1996, Kleijn et al. 1998). In the basic model, the speech signal $s(n)$ is modeled as the sum of a small number L of sinusoids

$$s(n) = \sum_{l=1}^{L} A_l \cos(\omega_l n + \phi_l) \ , \tag{5.5}$$

where $A_l(n)$ and $\phi_l(n)$ represent the amplitude and phase of each sinusoidal component associated with the frequency track ω_l. To find these parameters $A_l(n)$ and $\phi_l(n)$, the DFT

of windowed signal frames is calculated, and the peaks of the spectral magnitude are selected from each frame (see Figure 5.6). The basic model is also known as the McAulay/Quatieri Model. The basic model has also some modifications such as ABS/OLA (Analysis by Synthesis / Overlap Add) and Hybrid / Sinusoidal Noise models (Macon 1996).

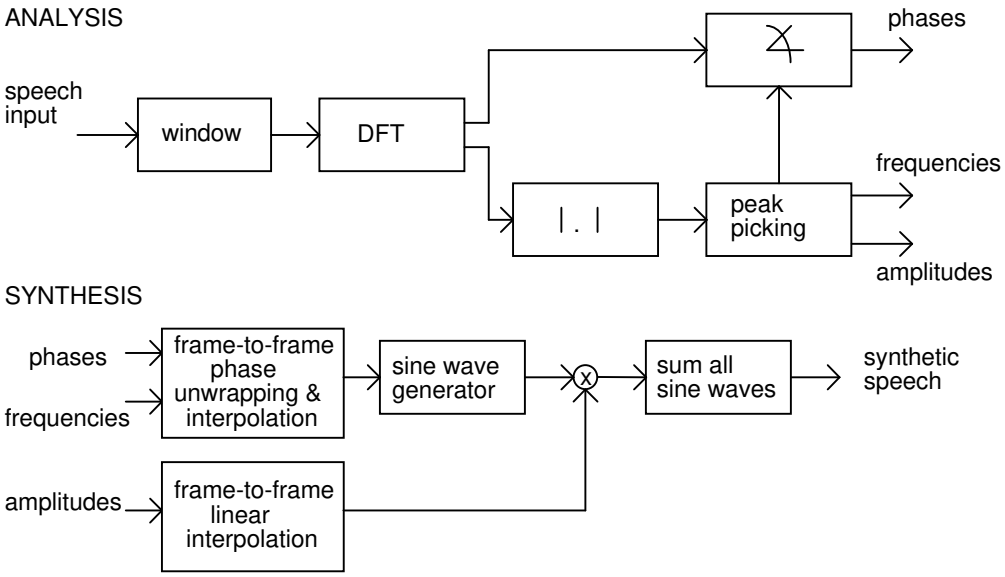

Fig. 5.6. Sinusoidal analysis / synthesis system (Macon 1996).

While the sinusoidal models are perhaps very suitable for representing periodic signals, such as vowels and voiced consonants, the representation of unvoiced speech becomes problematic (Macon 1996).

Sinusoidal models are also used successfully in singing voice synthesis (Macon 1996, Macon et al. 1997). The synthesis of singing differs from speech synthesis in many ways. In singing, the intelligibility of the phonemic message is often secondary to the intonation and musical qualities. Vowels are usually sustained longer in singing than in normal speech, and naturally, easy and independent controlling of pitch and loudness is also required. The best known singing synthesis system is perhaps the LYRICOS which is developed at Georgia Institute of Technology. The system uses sinusoidal-modeled segments from an inventory of singing voice data collected from human vocalist maintaining the characteristics and perceived identity. The system uses a standard MIDI-interface where the user specifies a musical score, phonetically-spelled lyrics, and control parameters such as vibrato and vocal effort (Macon et al. 1997).

5.6 High-Level Synthesis

With high-level synthesis the input text or information is transcribed in such format that low-level voice synthesizer is capable to produce the acoustic output. A proper implementation of this is the fundamental challenge in all present systems and will probably be for years to come. The procedure consists of three main phases.

- Text preprocessing where numerals, special characters, abbreviations, and acronyms are expanded into full words.
- Pronunciation analysis where the pronunciation of certain words, including homographs and proper names, are determined.
- Prosodic analysis where the prosodic features of speech are determined.

After high-level synthesizer, the information is delivered to drive some low-level system. The type of used data depends on the driven system. For example, for formant synthesizer, at least fundamental frequency, formant frequencies, duration, and amplitude of each sound segment is needed.

5.6.1 Text Preprocessing

The first task of all TTS systems is to convert input data to proper form for a synthesizer. In this stage, all non-alphabetical characters, numbers, abbreviations, and acronyms must be converted into a full spelled-out format. Text preprocessing is usually made with simple one-to-one lookup tables, but in some cases additional information of neighboring words or characters is needed. This may lead to a large database and complicated set of rules and may cause some problems with real-time systems. Input text may also contain some control characters which must be delivered through the text parser without modifications. The conversion must neither affect abbreviations which are a part of another. For example, if the character *M* is in some context converted as *mega*, the abbreviation *MTV* should not be converted as *megaTV*. However, character strings or abbreviations which are not in a lookup table and consist only of consonants can be always converted letter-by-letter because those kind of words do no exist in any language.

Numbers are perhaps the most difficult to convert correctly into spelled-out format. Numbers are used in several relations, such as digits, dates, roman numerals, measures, and mathematical expressions. Numbers between 1100 and 1999 are usually converted as years like 1910 as *nineteen-ten*. Expressions in form 12/12/99 or 11/11/1999 may be converted as dates, if the numbers are within acceptable values. However, the expression 2/5 is more difficult because it may be either *two divided by five* or *the second of may*. In some cases, the correct conversion is possible to conclude from

compounding information (measures etc.) or from the length of the number (dates, phone numbers etc.). However, there will be always some ambiguous situations.

In some cases with measures, usually currencies, the order of some character and value is changed. For example $3.02 is converted as *three dollars and two cents*. In these situations, the numerical expressions which are already in spelled-out format must be recognized to avoid the misconversion like *$100 million* to *one hundred dollars million*.

Some abbreviations and acronyms are ambiguous in different context like described in Chapter 4. For common abbreviation like st., the first thing to do is to check if it is followed by a capitalized word (potential name), when it will be expanded as *saint*. Otherwise, if it is preceded a capitalized word, an alphanumeric (5th), or a number, it will be expanded as *street* (Kleijn et al. 1998).

The parser may be implemented by straight programming using for example C, LISP or PERL or a parsing database with separate interface may be used. The latter method provides more flexibility for corrections afterwards but may have some limitations with abbreviations which have several different versions of correct conversion. A line in a converting database may look like for example following:

<"rules", "abbrev", "preceding info", "following info", "converted abbreviation">

where the "rules" may contain information of in which cases the current abbreviation is converted, e.g., if it is accepted in capitalized form or accepted with period or colon. Preceding and following information may contain also the accepted forms of ambient text, such as numbers, spaces, and character characteristics (vowel/consonant, capitalized etc.).

Sometimes different special modes, especially with numbers, are used to make this stage more accurate, for example, math mode for mathematical expressions and date mode for dates and so on. Another situation where the specific rules are needed is for example the E-mail messages where the header information needs special attention.

5.6.2 Pronunciation

Analysis for correct pronunciation from written text has also been one of the most challenging tasks in speech synthesis field. Especially, with some telephony applications where almost all words are common names or street addresses. One method is to store as much names as possible into a specific pronunciation table. Due to the amount of excisting names, this is quite unreasonable. So rule-based system with an exception dictionary for words that fail with those letter-to-phoneme rules may be a much more reasonable approach (Belhoula et al. 1993). This approach is also suitable for normal

pronunciation analysis. With morphemic analysis, a certain word can be divided in several independed parts which are considered as the minimal meaningful subpart of words as prefix, root, and affix. About 12 000 morphemes are needed for covering 95 percent of English (Allen et al.1987). However, the morphemic analysis may fail with word pairs, such as heal/health or sign/signal (Klatt 1987).

Another perhaps relatively good approach to the pronunciation problem is a method called *pronunciation by analogy* where a novel word is recognized as parts of the known words and the part pronunciations are built up to produce the pronunciation of a new word, for example pronunciation of word *grip* may be constructed from *grin* and *rip* (Gaved 1993). In some situations, such as speech markup languages described later in Chapter 7, information of correct pronunciation may be given separately.

5.6.3 Prosody

Prosodic or suprasegmental features consist of pitch, duration, and stress over the time. With good controlling of these gender, age, emotions, and other features in speech can be well modeled. However, almost everything seems to have effect on prosodic features of natural speech which makes accurate modeling very difficult. Prosodic features can be divided into several levels such as syllable, word, or phrase level. For example, at word level vowels are more intense than consonants. At phrase level correct prosody is more difficult to produce than at the word level.

The pitch pattern or fundamental frequency over a sentence (intonation) in natural speech is a combination of many factors. The pitch contour depends on the meaning of the sentence. For example, in normal speech the pitch slightly decreases toward the end of the sentence and when the sentence is in a question form, the pitch pattern will raise to the end of sentence. In the end of sentence there may also be a continuation rise which indicates that there is more speech to come. A raise or fall in fundamental frequency can also indicate a stressed syllable (Klatt 1987, Donovan 1996). Finally, the pitch contour is also affected by gender, physical and emotional state, and attitude of the speaker.

The duration or time characteristics can also be investigated at several levels from phoneme (segmental) durations to sentence level timing, speaking rate, and rhythm. The segmental duration is determined by a set of rules to determine correct timing. Usually some inherent duration for phoneme is modified by rules between maximum and minimum durations. For example, consonants in non-word-initial position are shortened, emphasized words are significantly lengthened, or a stressed vowel or sonorant preceded by a voiceless plosive is lengthened (Klatt 1987, Allen et al. 1987). In general, the phoneme duration differs due to neighboring phonemes. At sentence level, the speech rate, rhythm, and correct placing of pauses for correct phrase boundaries are

important. For example, a missing phrase boundary just makes speech sound rushed which is not as bad as an extra boundary which can be confusing (Donovan 1996). With some methods to control duration or fundamental frequency, such as the PSOLA method, the manipulation of one feature affects to another (Kortekaas et al. 1997).

The intensity pattern is perceived as a loudness of speech over the time. At syllable level vowels are usually more intense than consonants and at a phrase level syllables at the end of an utterance can become weaker in intensity. The intensity pattern in speech is highly related with fundamental frequency. The intensity of a voiced sound goes up in proportion to fundamental frequency (Klatt 1987).

The speaker's feelings and emotional state affect speech in many ways and the proper implementation of these features in synthesized speech may increase the quality considerably. With text-to-speech systems this is rather difficult because written text usually contains no information of these features. However, this kind of information may be provided to a synthesizer with some specific control characters or character strings. These methods are described later in Chapter 7. The users of speech synthesizers may also need to express their feelings in "real-time". For example, deafened people can not express their feelings when communicating with speech synthesizer through a telephone line. Emotions may also be controlled by specific software to control synthesizer parameters. Such system is for example HAMLET (Helpful Automatic Machine for Language and Emotional Talk) which drives the commercial DECtalk synthesizer (Abadjieva et al. 1993, Murray et al. 1996).

This section shortly introduces how some basic emotional states affect voice characteristics. The voice parameters affected by emotions are usually categorized in three main types (Abadjieva et al. 1993, Murray et al. 1993):

- *Voice quality* which contains largely constant voice characteristics over the spoken utterance, such as loudness and breathiness. For example, angry voice is breathy, loud, and has a tense articulation with abrupt changes while sad voice is very quiet with a decreased articulation precision.
- *Pitch contour* and its dynamic changes carry important emotional information, both in the general form for the whole sentence and in small fluctuations at word and phonemic levels. The most important pitch features are the general level, the dynamic range, changes in overall shape, content words, stressed phonemes, emphatic stress, and clause boundaries.
- *Time characteristics* contain the general rhythm, speech rate, the lengthening and shortening of the stressed syllables, the length of content words, and the duration and placing of pauses.

The number of possible emotions is very large, but there are five discrete emotional states which are commonly referred as the primary or basic emotions and the others are altered or mixed forms of these (Abadjieva et al. 1993). These are anger, happiness, sadness, fear, and disgust. The secondary emotional states are for example whispering, shouting, grief, and tiredness.

Anger in speech causes increased intensity with dynamic changes (Scherer 1996). The voice is very breathy and has tense articulation with abrupt changes. The average pitch pattern is higher and there is a strong downward inflection at the end of the sentence. The pitch range and its variations are also wider than in normal speech and the average speech rate is also a little bit faster.

Happiness or *joy* causes slightly increased intensity and articulation for content words. The voice is breathy and light without tension. Happiness also leads to increase in pitch and pitch range. The peak values of pitch and the speech rate are the highest of basic emotions.

Fear or *anxiety* makes the intensity of speech lower with no dynamic changes. Articulation is precise and the voice is irregular and energy at lower frequencies is reduced. The average pitch and pitch range are slightly higher than in neutral speech. The speech rate is slightly faster than in normal speech and contains pauses between words forming almost one third of the total speaking time (Murray et al. 1993, Abadjieva et al. 1993).

Sadness or *sorrowness* in speech decreases the speech intensity and its dynamic changes. The average pitch is at the same level as in neutral speech, but there are almost no dynamic changes. The articulation precision and the speech rate are also decreased. High ratio of pauses to phonation time also occurs (Cowie et al. 1996). Grief is an extreme form of sadness where the average pitch is lowered and the pitch range is very narrow. Speech rate is very slow and pauses form almost a half of the total speaking time (Murray et al. 1993).

Disgust or *contempt* in speech also decreases the speech intensity and its dynamic range. The average pitch level and the speech rate are also lower compared to normal speech and the number of pauses is high. Articulation precision and phonation time are increased and the stressed syllables in stressed content words are lengthened (Abadjieva et al. 1993).

Whispering and *shouting* are also common versions of expression. Whispering is produced by speaking with high breathiness without fundamental frequency, but the emotions can still be conveyed (Murray et al. 1993). Shouted speech causes an increased

pitch range, intensity and greater variability in it. Tiredness causes a loss of elasticity of articulatory muscles leading to lower voice and narrow pitch range.

5.7 Other Methods and Techniques

Several other methods and experiments to improve the quality of synthetic speech have been made. Variations and combinations of previously described methods have been studied widely, but there is still no single method to be considered distinctly the best. Synthesized speech can also be manipulated afterwards with normal speech processing algorithms. For example, adding some echo may produce more pleasant speech. However, this approach may easily increase the computational load of the system.

Some experiments to show the use of a combination of the basic synthesis methods have been made, because different methods show different success in generating individual phonemes. Time domain synthesis can produce high-quality and natural sounding speech segments, but in some segment combinations the synthesized speech is discontinuous at the segment boundaries and if a wide-range variation of fundamental frequency is required, overall complexity will increase. On the other hand, formant synthesis yields more homogeneous speech allowing a good control of fundamental frequency, but the voice-timbre sounds more synthetic. This approach leads to the hybrid system which combines the time- and frequency-domain methods. The basic idea of a hybrid system is shown in Figure 5.7 (Fries 1993).

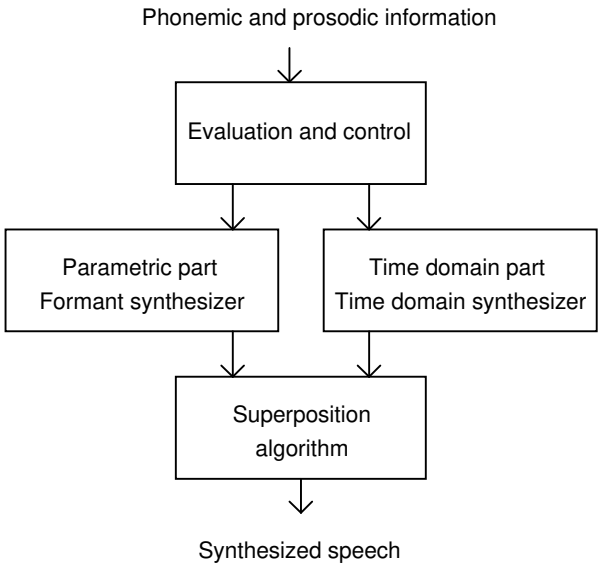

Fig. 5.7. Basic idea of the hybrid synthesis system

Several methods and techniques for determining the control parameters for a synthesizer may be used. Recently, the artificial intelligence based methods, such as Artificial Neural Networks (ANN), have been used to control synthesis parameters, such as

duration, gain, and fundamental frequency (Scordilis et al. 1989, Karjalainen et al. 1991, 1998). Neural networks have been applied in speech synthesis for about ten years and they use a set of processing elements or nodes analogous to neurons in the brain. These processing elements are interconnected in a network that can identify patterns in data as it is exposed to the data. An example of using neural networks with WLP-based speech synthesis is given in Figure 5.8.

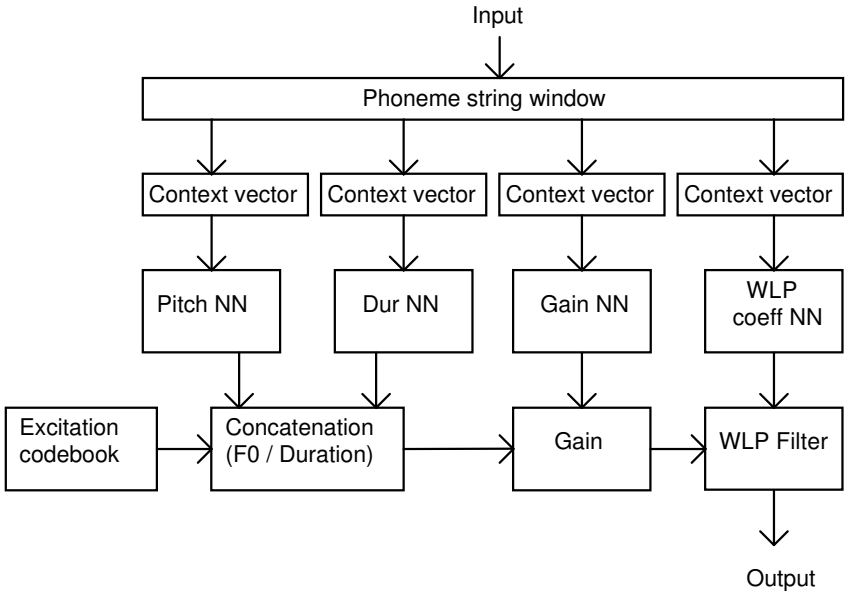

Fig. 5.8. An example of using neural networks in WLP synthesis (Karjalainen et al. 1998).

For more detailed discussion of using neural networks in speech synthesis, see for example (Rahim et al. 1993), (Cawley et al. 1993a, 1993b), (Cawley 1996), or (Karjalainen et al. 1991, 1998) and references in them.

Another common method in speech synthesis, and especially in speech recognition and analysis of prosodic parameters from speech, is for example the use of hidden Markov models (HMMs). The method is based on a statistical approach to simulate real life stochastic processes (Rentzepopoulos et al. 1992). A hidden Markov model is a collection of states connected by transitions. Each transition carries two sets of probabilities: a transition probability, which provides the probability for taking the transition, and an output probability density function, which defines the conditional probability of emitting each output symbol from a finite alphabet, given that that the transition is taken (Lee 1989).

6. APPLICATIONS OF SYNTHETIC SPEECH

Synthetic speech may be used in several applications. Communication aids have developed from low quality talking calculators to modern 3D applications, such as talking heads. The implementation method depends mostly on used application. In some cases, such as announcement or warning systems, unrestricted vocabulary is not necessary and the best result is usually achieved with some simple messaging system. With suitable implementation some funds may also be saved. On the other hand, some applications, such as reading machines for the blind or electronic-mail readers, require unlimited vocabulary and a TTS system is needed.

The application field of synthetic speech is expanding fast whilst the quality of TTS systems is also increasing steadily. Speech synthesis systems are also becoming more affordable for common customers, which makes these systems more suitable for everyday use. For example, better availability of TTS systems may increase employing possibilities for people with communication difficulties.

6.1 Applications for the Blind

Probably the most important and useful application field in speech synthesis is the reading and communication aids for the blind. Before synthesized speech, specific audio books were used where the content of the book was read into audio tape. It is clear that making such spoken copy of any large book takes several months and is very expensive. It is also easier to get information from computer with speech instead of using special bliss symbol keyboard, which is an interface for reading the Braille characters.

The first commercial TTS application was probably the Kurzweil reading machine for the blind introduced by Raymond Kurzweil in the late 1970's. It consisted of an optical scanner and text recognition software and was capable to produce quite intelligible speech from written multifont text (Klatt 1987). The prices of the first reading machines were far too high for average user and these machines were used mostly in libraries or related places. Today, the quality of reading machines has reached acceptable level and prices have become affordable for single individual, so a speech synthesizer will be very helpful and common device among visually impaired people in the future. Current systems are mostly software based, so with scanner and OCR system, it is easy to construct a reading machine for any computer environment with tolerable expenses. Regardless of how fast the development of reading and communication aids is, there is always some improvements to do.

The most crucial factor with reading machines is speech intelligibility which should be maintained with speaking rates ranging from less than half to at least three times normal rate (Portele et al. 1996). Naturalness is also an important feature and makes the synthetic speech more acceptable. Although the naturalness is one of the most important features, it may sometimes be desirable that the listener is able to identify that speech is coming from machine (Hess 1992), so the synthetic speech should sound natural but somehow "neutral".

When the output from a speech synthesizer is listened for the first time, it may sound intelligible and pleasant. However, during longer listening period, single clicks or other weak points in the system may arise very annoying. This is called an annoying effect and it is difficult to perceive with any short-term evaluation method, so for these kind of cases, the feedback from long-term users is sometimes very essential.

Speech synthesis is currently used to read www-pages or other forms of media with normal personal computer. Information services may also be implemented through a normal telephone interface with keypad-control similar to text-tv. With modern computers it is also possible to add new features into reading aids. It is possible to implement software to read standard check forms or find the information how the newspaper article is constructed. However, sometimes it may be impossible to find correct construction of the newspaper article if it is for example divided in several pages or has an anomalous structure.

A blind person can not also see the length of an input text when starting to listen it with a speech synthesizer, so an important feature is to give in advance some information of the text to be read. For example, the synthesizer may check the document and calculate the estimated duration of reading and speak it to the listener. Also the information of bold or underlined text may be given by for example with slight change of intonation or loudness.

6.2 Applications for the Deafened and Vocally Handicapped

People who are born-deaf can not learn to speak properly and people with hearing difficulties have usually speaking difficulties. Synthesized speech gives the deafened and vocally handicapped an opportunity to communicate with people who do not understand the sign language. With a talking head it is possible to improve the quality of the communication situation even more because the visual information is the most important with the deaf and dumb. A speech synthesis system may also be used with communication over the telephone line (Klatt 1987).

Adjustable voice characteristics are very important in order to achieve individual sounding voice. Users of talking aids may also be very frustrated by an inability to convey emotions, such as happiness, sadness, urgency, or friendliness by voice. Some tools, such as HAMLET (Helpful Automatic Machine for Language and Emotional Talk) have been developed to help users to express their feelings (Murray et al. 1991, Abedjieva et al. 1993). The HAMLET system is designed to operate on a PC with high quality speech synthesizer, such as DECtalk.

With keyboard it is usually much slower to communicate than with normal speech. One way to speed up this is to use the predictive input system that always displays the most frequent word for any typed word fragment, and the user can then hit a special key to accept the prediction. Even individual pre-composed phrases, such as greetings or salutes, may be used.

6.3 Educational Applications

Synthesized speech can be used also in many educational situations. A computer with speech synthesizer can teach 24 hours a day and 365 days a year. It can be programmed for special tasks like spelling and pronunciation teaching for different languages. It can also be used with interactive educational applications.

Especially with people who are impaired to read (dyslexics), speech synthesis may be very helpful because especially some children may feel themselves very embarrassing when they have to be helped by a teacher (Klatt 1987). It is also almost impossible to learn write and read without spoken help. With proper computer software, unsupervised training for these problems is easy and inexpensive to arrange.

A speech synthesizer connected with word processor is also a helpful aid to proof reading. Many users find it easier to detect grammatical and stylistic problems when listening than reading. Normal misspellings are also easier to detect.

6.4 Applications for Telecommunications and Multimedia

The newest applications in speech synthesis are in the area of multimedia. Synthesized speech has been used for decades in all kind of telephone enquiry systems, but the quality has been far from good for common customers. Today, the quality has reached the level that normal customers are adopting it for everyday use.

Electronic mail has become very usual in last few years. However, it is sometimes impossible to read those E-mail messages when being for example abroad. There may be no proper computer available or some security problems exists. With synthetic

speech e-mail messages may be listened to via normal telephone line. Synthesized speech may also be used to speak out short text messages (sms) in mobile phones.

For totally interactive multimedia applications an automatic speech recognition system is also needed. The automatic recognition of fluent speech is still far away, but the quality of current systems is at least so good that it can be used to give some control commands, such as yes/no, on/off, or ok/cancel.

6.5 Other Applications and Future Directions

In principle, speech synthesis may be used in all kind of human-machine interactions. For example, in warning and alarm systems synthesized speech may be used to give more accurate information of the current situation. Using speech instead of warning lights or buzzers gives an opportunity to reach the warning signal for example from a different room. Speech synthesizer may also be used to receive some desktop messages from a computer, such as printer activity or received e-mail.

In the future, if speech recognition techniques reach adequate level, synthesized speech may also be used in language interpreters or several other communication systems, such as videophones, videoconferencing, or talking mobile phones. If it is possible to recognize speech, transcribe it into ASCII string, and then resynthesize it back to speech, a large amount of transmission capacity may be saved. With talking mobile phones it is possible to increase the usability considerably for example with visually impaired users or in situations where it is difficult or even dangerous to try to reach the visual information. It is obvious that it is less dangerous to listen than to read the output from mobile phone for example when driving a car.

During last few decades the communication aids have been developed from talking calculators to modern three-dimensional audiovisual applications. The application field for speech synthesis is becoming wider all the time which brings also more funds into research and development areas. Speech synthesis has also several application frameworks which are described in the following chapter.

7. APPLICATION FRAMEWORKS

Several methods and interfaces for making the implementation of synthesized speech in desired applications easier have been developed during this decade. It is quite clear that it is impossible to create a standard for speech synthesis methods because most systems act as stand alone device which means they are incompatible with each other and do not share common parts. However, it is possible to standardize the interface of data flow between the application and the synthesizer.

Usually, the interface contains a set of control characters or variables for controlling the synthesizer output and features. The output is usually controlled by normal play, stop, pause, and resume type commands and the controllable features are usually pitch baseline and range, speech rate, volume, and in some cases even different voices, ages, and genders are available. In most frameworks it is also possible to control other external applications, such as a talking head or video.

In this chapter, three approaches to standardize the communication between a speech synthesizer and applications are introduced. Most of the present synthesis systems support so called Speech Application Programming Interface (SAPI) which makes easier the implementation of speech in any kind of application. For Internet purposes several kind of speech synthesis markup languages have been developed to make it possible to listen to synthesized speech without having to transfer the actual speech signal through network. Finally, one of the most interesting approaches is probably the TTS subpart of MPEG-4 multimedia standard which will be introduced in the near future.

7.1 Speech Application Programming Interface

SAPI is an interface between applications and speech technology engines, both text-to-speech and speech recognition (Amundsen 1996). The interface allows multiple applications to share the available speech resources on a computer without having to program the speech engine itself. Speech synthesis and recognition applications usually require plenty of computational resources and with SAPI approach lots of these resources may be saved. The user of an application can also choose the synthesizer used as long as it supports SAPI. Currently SAPIs are available for several environments, such as MS-SAPI for Microsoft Windows operating systems and Sun Microsystems Java SAPI (JSAPI) for JAVA based applications. In this chapter, only the speech synthesis part is discussed.

SAPI text-to-speech part consists of three interfaces. The *voice text* interface which provides methods to start, pause, resume, fast forward, rewind, and stop the TTS engine during speech. The *attribute interface* allows access to control the basic behavior of the TTS engine, such as the audio device to be used, the playback speed (in words per minute), and turning the speech on and off. With some TTS systems the attribute interface may also be used to select the speaking mode from predefined list of voices, such as female, male, child, or alien. Finally, the *dialog interface* can be used to set and retrieve information regarding the TTS engine to for example identify the TTS engine and alter the pronunciation lexicon.

7.1.1 Control Tags

The SAPI model defines 15 different control tags that can be used to control voice characteristics, phrase modification, and low-level synthesis. The voice character tags can be used to set high-level general characteristics of the voice, such as gender, age, or feelings of the speaker. The tag may also be used to tell the TTS engine the context of the message, such as plain text, e-mail, or address and phone numbers. The phrase modification tags may be used to adjust the pronunciation at word-by-word or phrase-by-phrase level. User can control for example the word emphasis, pauses, pitch, speed, and volume. The low-level tags deal with attributes of the TTS engine itself. User can for example add comments to the text, control the pronunciation of a word, turn prosody rules on and off, or reset the TTS engine to default settings. Only the reset tag of the low-level tags is commonly used (Amundsen 1996).

The control tags are separated with the backslash symbol from text to be spoken (\Tag="Parameter" or "value"\). The control tags are not case sensitive, but white-space sensitive. For example, \spd=200\ is the same as \SPD=200\, but \Spd=200\ is not the same as \ Spd=200 \. If the TTS engine encounters an unknown control tag, it just ignores it. The following control tags and their examples are based on MS-SAPI (Amundsen 1996).

The voice character control tags:

Chr Used to set the character of the voice. More than one charasteristic can be applied at the same time. The default value is normal and the other values may be for example angry, excited, happy, scared, quiet, loud, and shout.

\Chr="Angry","Loud"\ Give me that! \Chr="Normal"\ Thanks.
\Chr="Excited"\ I am very excited. \Chr="Normal"\

Ctx Used to set the context of spoken text. The context parameter may be for example address, C, document, E-mail, numbers/dates, or spreadsheet. The default value is unknown. In the following example the TTS engine converts

the "W. 7th St." to "West seventh street", but fails to do so when the \ctx="unknown"\ tag is used. The e-mail address is converted as "sami dot lemmetty at hut dot fi".

\Ctx="Address"\ 1204 W. 7th St., Oak Ridge, TN.
\Ctx="E-mail"\ sami.lemmetty@hut.fi.
\Ctx="unknown"\ 129 W. 1st Avenue.

Vce Used to set additional characteristics of the voice. Several character types can be set in a single call. Character types may be for example language, accent, dialect, gender, speaker, age, and style.

\Vce=Language="English", Accent="French"\ This is English with a French accent. \Vce=Gender="Male"\ I can change my gender easily from male to \Vce=Gender="Female"\ female.

The phrase modification control tags:

Emp Used to add emphasis to a single word followed by the tag. In the following sentences the words "told" and "important" are emphasized.

I \Emp\ told you never go running in the street.
You must listen to me when I tell you something \Emp\ important.

Pau Used to place a silent pause into the output stream. The duration of pause is given in milliseconds.

Pause of one \Pau=1000\ second.

Pit Used to alter the base pitch of the output. The pitch is measured in Hertz between 50 Hz and 400 Hz. The pitch setting does not automatically revert the default value after a message has been spoken so it must be done manually.

\Pit=200\ You must listen to me \Pit=100\.

Spd Used to set the base speed of the output. The speed is measured in words per minute between 50 and 250.

\Spd=50\ This is slow, but \Spd=200\ this is very fast.

Vol Used to set the base volume of the output. The value can range from 0 (quiet) to 65535 (loud).

\Vol=15000\ Hello. \Vol=60000\ Hello!! \Vol=30000\

The low-level TTS control tags:

Com Used to add comments to the text passed to the TTS engine. These comments will be ignored by the TTS engine.

\Com="This is a comment"\

Eng Used to call an engine-specific command.

Mrk Used to mark specific bookmarks. Can be used for signalling such things as page turns or slide changes once the place in the text is reached.

Prn Used to embed custom pronunciations of words using the International Phonetic Alphabet (IPA).

Pro Used to turn on and off the TTS prosody rules. Value 1 turns the settings off and value 0 turns them on.

Prt Used to tell the TTS engine what part of speech the current word is. The categories may be for example abbreviation, noun, adjective, ordinal number, preposition, or verb. The following example defines word "is" as a verb, and word "beautiful" as an adjective.

This flower \Prt="V"\ is \Prt="Adj"\ beautiful.

Rst Used to reset the control values to those that existed at the start of the current session.

7.2 Internet Speech Markup Languages

Most synthesizers accept only plain text as input. However, it is difficult to analyze the text and find correct pronunciation and prosody from written text. In some cases there is also need to include the speaker features or emotional information in the output speech. With some additional information in input data it is possible to control these features of speech easily. For example, with some information about if the input sentence is in a question, imperative, or neutral form, the controlling of prosody may become significantly easier. Some commercial systems allow the user to place same kind of annotations in the text to produce more natural sounding speech. These are for example DECtalk and the Bell Labs system described more closely in Chapter 9.

In normal HTML (Hyper-Text Markup Language), certain markup tags like ... are used to delimit paragraphs and help the web-browser to construct the correct output. These and same kind of additional tags may be used to help a speech synthesizer produce correct output with different kind of pronunciations, voices and other features.

For example, to describe happiness, we may use tags <happy>...</happy> or to describe a question <quest>...</quest>. Speaker's features and used language may be controlled by same way with tags <gender=female> or <lang=fin>. Some words and common names have anomalous pronunciation which may be corrected with same kind of tags. Local stress markers may also be used to stress a certain word in a sentence.

The first attempt to develop a TTS markup language was called SSML (Speech Synthesis Markup Language), developed at the Centre for Speech Technology Research (CSTR) in the University of Edinburgh, England, in 1995 (Taylor et al. 1997). It included control tags for phase boundaries, language, and made possible to define a pronunciation of a specific word and include emphasis tags in the sentence. In the following example, *pro* defines the pronunciation of the word and *format* defines the used lexicon standard. With tag <phrase> it is even possible to change the meaning of the whole sentence.

```
<ssml>
<define word= "edinburgh" pro="EH1 D AH0 N B ER2 OW0" format= "cmudict.1.0>
<phrase> I saw the man in the park <phrase> with the telescope </phrase>
<phrase> I saw the man <phrase> in the park with the telescope </phrase>
<phrase> The train is now standing on platform <emph> A </emph>
<language="italian">
<phrase> continua in italiano </phrase>
</ssml>
```

Currently the development of the language is continuing with Bell Laboratories (Sproat et al. 1997). The latest version is called STML (Spoken Text Markup Language). SUN Microsystems is also participating in the development process to merge their JSML (Java Speech Markup Language) to achieve one widespread system in the near future. Currently, the controllable features are much wider than in SSML.

The structure of STML is easiest to apprehend from the example below. The used language and the default speaker of that language are set simply with tags <language id> and <speaker id>. The tag <genre type> allows to set the type of text like plain prose, poetry, or lists. The tag <div type> specifies a particular text-genre-specific division with list items. With tag <emph> the emphasis level of the following word is specified. The tag <phonetic> specifies that the enclosed region is a phonetic transcription in one of a predefined set of schemes. The tag <define> is used to specify the lexical pronunciation of a certain word. The tag <intonat> specifies the midline and amplitude of pitch range with absolute scale in hertz or relative multiplier compared to normal pitch for the speaker. The tag <bound> is used to define an intonational boundary between 0 (weakest) and 5 (strongest). The <literal mode> is used for spelling mode and the <omitted> tag specifies the region that is emitted from output speech.

In the following example some of the essential features of STML are presented.

```
<!doctype stml system>
<stml>
<language id=english>
<speaker id=male1>
<genre type=plain>
In this example, we see some <emph> particular </emph> STML tags, including:
<genre type=list>
language specification <div type=item>
speaker specification <div type=item>
text type (genre) specifications <div type=item>
<phonetic scheme=native> f&amp;n"etik </phonetic> specifications
phrase boundary <bound type=minor> specifications
</genre>

<define word="edinburgh" pro="e 1 d i n b 2 r @@" scheme="cstr">
The Edinburgh and Bell labs systems now pronounce word Edinburgh correctly.

Some text in <literal mode=spell> literal mode </literal>
<omitted verbose=yes> you hear nothing </omitted>
<rate speed=250 scheme=wpm> this is faster </rate>
...
</genre>
</speaker>
</language>
</stml>
```

Markup languages provide some advantages compared to for example SAPI which provides tags only for speaker directives, not for any text description. In theory, anything specifiable in the text which can give an instruction or description to a TTS system could be in a synthesis markup language. Unfortunately, several systems are under development and the making of an international standard is a considerable problem.

7.3 MPEG-4 TTS

The MPEG-4 Text-to-Speech (M-TTS) is a subpart of the standard which is currently under development in ISO MPEG-4 committee (ISO 1997). It specifies the interface between the bitstream and listener. Naturally, due to various existing speech synthesis techniques the exact synthesis method is not under standardization. The main objective is to make it possible to include narration in any multimedia content without having to record natural speech. Also controlling of facial animation (FA) and moving picture (MP) is supported. Because MPEG-4 TTS system is still under development, it is only discussed briefly below. Further and more up-to-date information is available in MPEG-homepage (MPEG 1998).

7.3.1 MPEG-4 TTS Bitstream

The M-TTS bitstream consists of two parts, the sequence part and the sentence part. Both parts begin with start code and ID code. The sequence part contains the information of what features are included in the bit stream. It consists of enable flags for gender, age, speech rate, prosody, video, lip-shape, and trick mode. Used language is also specified in this part with 18 bits. The sentence part contains all the information which is enabled in the sequence part and the text to be synthesized with phonetic symbols used. Also the length of silence sections are defined in this section. Used variables are described in Table 7.1., where for example the notation '8 x L' in TTS_Text means that the TTS_Text is indexed with the Length_of_Text.

Table 7.1. The M-TTS sentence part description.

String	Description	Bits
Silence	Set to 1 when the current position is silence.	1
Silence_Duration	Silence segment in milliseconds (0 prohibited).	12
Gender	Speakers gender. 1 if male and 0 if female.	1
Age	Speaker age, 8 levels, below 6 to over 60.	3
Speech_Rate	Synthetic speech rate in 16 levels.	4
Length_of_Text	Length of TTS_Text data in bytes (L).	12
TTS_Text	Character string containing the input string.	8 x L
Dur_Enable	Set to 1 when duration data exists.	1
F0_Contour_Enable	Set to 1 when pitch contour information exists.	1
Energy_Contour_Enable	Set to 1 when energy contour information exists.	1
Number_of_Phonemes	Number of phonemes needed for synthesis of input text. (NP)	10
Phonemes_Symbols_Lenght	The length of Phoneme_Symbols data in bytes. (P)	13
Phoneme_Symbols	The indexing number for the current phoneme.	8 x P
Dur_each_Phoneme	The duration of each phoneme in milliseconds.	12 x NP
F0_Contour_each_Phoneme	The pitch for the current phoneme in Hz. (Half of the real pitch in Hz for three points: 0, 50, 100 %)	8 x NP x 3
Energy_Contour_each_Phoneme	The energy level of current phoneme in integer dB for three points (0, 50, 100% positions of the phoneme).	8 x NP x 3
Sentence_Duration	The duration of sentence in milliseconds.	16
Position_in_Sentence	The position of the current stop in a sentence. (Elapsed time in milliseconds)	16
Offset	The duration of short pause before the start of speech in msec.	10
Number_of_Lip_Shape	The number of lip-shape patterns to be processed. (N)	10
Lip_Shape_in_Sentence	The position of each lip shape from the beginning of the sentence in milliseconds. (L)	16 x N
Lip_Shape	The indexing number for the current lip shape for MP.	8 x L

The parameters are described more closely in ISO (1996). For example, prosody_enable bit in sequence part enables duration, f0contour, and energy contour in sentence part making prosodic features available.

7.3.2 Structure of MPEG-4 TTS Decoder

The structure of an M-TTS decoder is presented in Figure 7.1. Only the interfaces are the subjects of standardization process. There are five interfaces:

1. Between demux and the syntactic decoder
2. Between the syntactic decoder and the speech synthesizer
3. From the speech synthesizer to the compositor
4. From the compositor to speech synthesizer
5. Between the speech synthesizer and the phoneme-to-FAP (Facial Animation Parameter) converter

When decoder receives the M-TTS bitstream it is first demultiplexed (1) and sent to the syntactic decoder which specifies the bitstream sent to speech synthesizer (2) including some of the following: The input type of the M-TTS data, control commands stream, input text to be synthesized, and some additional information, such as prosodic parameters, lip-shape patterns, and information for the trick mode operation.

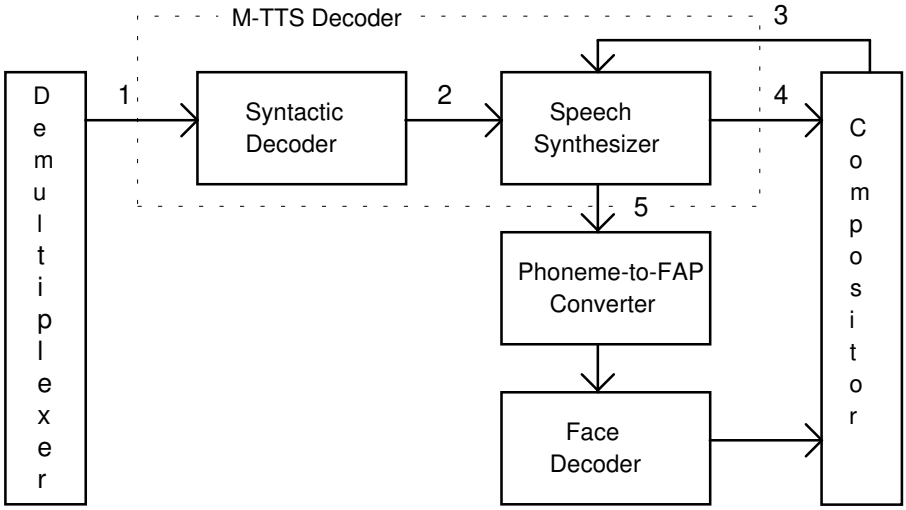

Fig. 7.1. MPEG-4 Audio TTS decoder architecture.

The interface from compositor to speech synthesizer (3) is defined to allow the local control of synthesized speech by user. The user interface can support several features, such as trick mode (play, stop, forward, backward etc.) and prosody (speech rate, pitch, gender, age etc.). Trick mode is synchronized with moving picture.

The interface from speech synthesizer to compositor (4) defines the data structure of encoded digital speech. It is identical to the interface for digitized natural speech to the compositor. Finally, the interface between speech synthesizer and phoneme-to-FAP

converter (5) defines the data structure between these modules. The phoneme-to-FAP converter is driven and synchronized with speech synthesizer by phoneme information. The data structure consists of phoneme symbol and duration with average fundamental frequency.

7.3.3 Applications of MPEG-4 TTS

M-TTS presents two application scenarios for the M-TTS Decoder, MPEG-4 Story Teller on Demand (STOD) and MPEG-4 Audio Text-to-Speech with Moving Picture. These scenarios are only informative and they are not under standardization process. Naturally, MPEG-4 TTS may be used in several other audio-visual related applications, such as dubbing-tools for animated pictures or Internet voice.

Story Teller on Demand is an application where user can select huge databases or story libraries stored on hard disk, CD-ROM or other media. The system reads the story via M-TTS decoder with the MPEG-4 facial animation or with appropriately selected images. The user can stop and resume speaking at any moment he wants with for example mouse or keyboard. The gender, age, and the speech rate of the story teller are also easily adjustable. With the STOD system, the narration with several features can be easily composed without recording the natural speech and so the required disk space is considerably reduced.

Audio Text-to-Speech with Moving picture is an application where the synchronized playback of the M-TTS decoder and encoded moving picture is the main objective. The decoder can provide several granularities of synchronization for different situations. Aligning only the composition time of each sentence, coarse granularity of synchronization and trick mode functionality can be easily achieved. For finer synchronization granularity the lip shape information may be utilized. The finest granularity can be achieved by using the prosody and video-related information. With this synchronization capability, the M-TTS decoder may be used for moving picture dubbing by utilizing the lip shape pattern information.

In the future M-TTS or other similar approaches may be used in several multimedia and telecommunication applications. However, it may take some time before we have full synthetic newsreaders and narrators. Some of the present synthesizers are using a same kind of controlling approach in their system, but there is still no efforts for widespread standard.

8. AUDIOVISUAL SPEECH SYNTHESIS

8.1 Introduction and History

Speech communication relies not only on audition, but also on visual information. Facial movements, such as smiling, grinning, eye blinking, head nodding, and eyebrow rising give an important additional information of the speaker's emotional state. The emotional state may be even concluded from facial expression without any sound (Beskow 1996). Fluent speech is also emphasized and punctuated by facial expressions (Waters et al. 1993). With visual information added to synthesized speech it is also possible to increase the intelligibility significantly, especially when the auditory speech is degraded by for example noise, bandwidth filtering, or hearing impairment (Cohen et al. 1993, Beskow 1996, Le Goff et al. 1996). The visual information is especially helpful with front phonemes whose articulation we can see, such as labiodentals and bilabials (Beskow et al. 1997). For example, intelligibility between /b/ and /d/ increases significantly with visual information (Santen et al. 1997). Synthetic face also increases the intelligibility with natural speech. However, the facial gestures and speech must be coherent. Without coherence the intelligibility of speech may be even decreased. For example, an interesting phenomenon with separate audio and video is so called McGurk effect. If an audio syllable /ba/ is dubbed onto a visual /ga/, it is perceived as /da/ (Cohen et al. 1993, Cole et al. 1995).

Human facial expression has been under investigation for more than one hundred years. The first computer-based modeling and animations were made over 25 years ago. In 1972 Parke introduced the first three-dimensional face model and in 1974 he developed the first version of his famous parameteric three-dimensional model (Santen et al. 1997). Since the computer capabilities have increased rapidly during last decades, the development of facial animation has been also very fast, and will remain fast in the future when the users are becoming more comfortable with the dialogue situations with machines.

Facial animation has been applied to synthetic speech for about ten years. Most of the present audiovisual speech synthesizers are based on a parametric face model presented by Parke in 1982. The model consisted of a mesh of about 800 polygons that approximated the surface of a human face including the eyes, the eyebrows, the lips, and the teeth. The polygon surface was controlled by using 50 parameters (Beskow 1996). However, present systems contain a number of modifications to Parke model to improve it and to make it more suitable for synthesized speech. These are usually a set of rules

for generating facial control parameter trajectories from phonetic text, and a simple tongue model, which were not included in the original Parke model.

Audiovisual speech synthesis may be used in several applications. Additional visual information is very helpful for hearing impaired people. It can be used as a tool for interactive training of speechreading. Also a face with semi-transparent skin and a well modeled tongue can be used to visualize tongue positions in speech training for deaf children (Beskow 1996). It may be used in information systems in public and noisy environments, such as airports, train stations and shopping centers. If it is possible to make the talking head look like some certain individual, it may be utilized in videoconferencing or used as a synthetic newsreader. Multimedia is also an important application field of talking heads. A full synthetic story teller requires considerably less storage capacity compared to for example movie clips.

8.2 Techniques and Models

Perhaps the easiest approach for audiovisual speech is to use pre-stored images to represent all the possible shapes under interest and combine these with for example some morphing method similar to concatenative speech synthesis. This method may be quite successful for some limited applications, but is very inflexible, since there is no way to control different facial features independently of each other (Beskow 1996). Because of this, the talking head is usually implemented with some kind of parametric model. There are two usually used basic methods:

- Two or three-dimensional parametric model which can be viewed as a geometric description of the facial surface that can be deformed using a limited set of control parameters and rendered using standard computer graphics techniques. Method is similar to formant synthesis.
- Muscle based controlling, where the face surface is modeled with facial muscle activation parameters. The method is perhaps theoretically the most elegant because it model face movements directly as the articulatory synthesis models the vocal system. However, there are several difficulties involved in modeling all the muscles needed to simulate for example articulate lip motion.

Due to difficulties with muscle based implementation, most researchers have found the parametric model more feasible (Beskow 1996, Le Goff et al. 1996) and most of the present systems are using a parametric model descended from Parke model. Naturally, the mouth and the lips are the most important in facial models, but with eyes, eyebrows, jaw, and tongue it is possible to make the audiovisual speech more natural and intelligible (Santen et al. 1997).

In visual part, the equivalence of phonemes is called as visemes. One example of how the set of visemes can be formed from phonemes of standard English is represented in Table 8.1. The phonetic SAM-PA representation which has been used is described earlier in Chapter 4. Certainly, due to articulation effect, this set of visemes is not enough to represent accurate mouth shapes (Breen et al. 1996).

Table 8.1. Set of visemes formed by phonemes of standard British English.

Viseme group	Phonemes (SAM-PA)
Consonant 1	p, b, m
Consonant 2	f, v
Consonant 3	D, T
Consonant 4	s, z
Consonant 5	S, Z
Consonant 6	t, d, n, l, r
"Both"	w, U, u, O
Vowel 1	Q, V, A
Vowel 2	3, i, j
Vowel 3	@, E, I, {

Like in concatenative speech synthesis, diphone-like units may be used to avoid discontinuities and to include coarticulation effect in used units. In visual part, these units are called as di-visemes. A di-viseme records the change in articulation produced when moving from one viseme to another. The number of video recordings with di-visemes is 128 and may be reduced to less than 50 if the further approximation is made that the coarticulation due to vowel production greatly outweighs the effects produced by consonants (Breen et al. 1996). Longer units may also be used, such as tri-visemes which contains the immediate right and left context effect on a center viseme. However, the number of needed video recordings is approximately 800, which is clearly unrealistic.

Audiovisual speech suffers mostly of the same problems as normal speech synthesis. For example, phoneme /t/ in *tea* differs in lip shape to the same phoneme in *two*, and Finnish phoneme /k/ in *kissa* and *koira* is visually very different compared to acoustical difference. These differences in facial movements due to context are the visual correlate of the speech effect known as coarticulation (Breen et al. 1996).

The speech synthesizer and the facial animation are usually two different systems and they must be synchronized somehow. If the synchronization process is not done properly, the quality and intelligibility of audiovisual speech may even decrease significantly. Usually, the speech synthesizer is used to provide the information for controlling the visual part. The computational requirements for the visual part are usually considerably higher than for the audio, so some kind of feedback from the visual

part may be needed to avoid lag between audio and video. The lag may be avoided by buffering the audio images and adjusting the frame rate if necessary. The structure of an audiovisual speech synthesizer is presented in Figure 8.1.

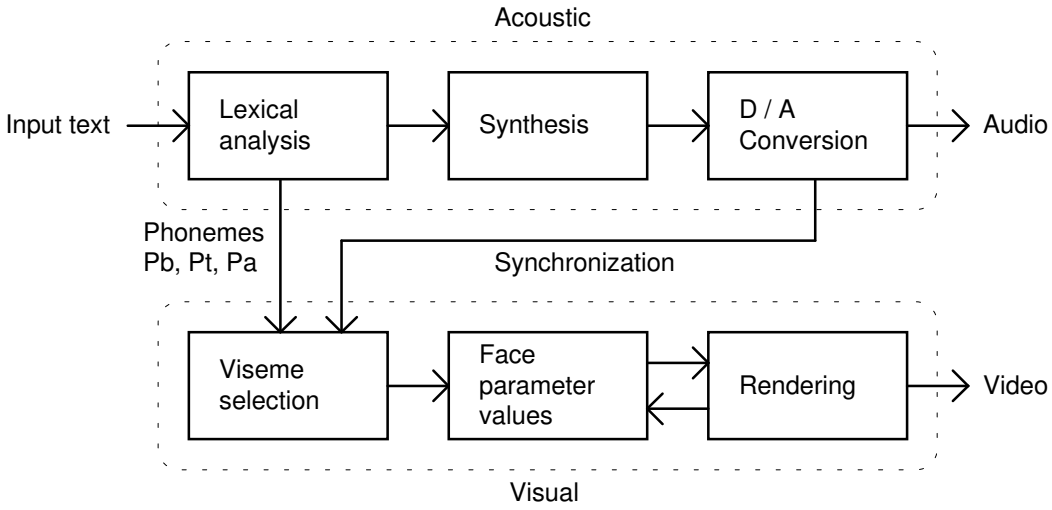

Fig. 8.1. Structure of the audiovisual synthesizer.

The sequence to control the visual synthesizer is processed as three phoneme frames: the target phoneme (Pt), the phoneme before (Pb) and the one after the target (Pa). The transformation from phonetic representation to the face parameter values is based on corresponding visemes.

9. PRODUCTS

This chapter introduces some of the commercial products, developing tools, and ongoing speech synthesis projects available today. It is clear that it is not possible to present all systems and products out there, but at least the most known products are presented. Some of the text in this chapter is based on information collected from Internet, fortunately, mostly from the manufacturers and developers official homepages. However, some criticism should be bear in mind when reading the "this is the best synthesis system ever" descriptions from these WWW-sites.

First commercial speech synthesis systems were mostly hardware based and the developing process was very time-consuming and expensive. Since computers have become more and more powerful, most synthesizers today are software based systems. Software based systems are easy to configure and update, and usually they are also much less expensive than the hardware systems. However, a stand alone hardware device may still be the best solution when a portable system is needed.

The speech synthesis process can be divided in high-level and low-level synthesis. A low-level synthesizer is the actual device which generates the output sound from information provided by high-level device in some format, for example in phonetic representation. A high-level synthesizer is responsible for generating the input data to the low-level device including correct text-preprocessing, pronunciation, and prosodic information. Most synthesizers contain both, high and low level system, but due to specific problems with methods, they are sometimes developed separately.

9.1 Infovox

Telia Promotor AB Infovox speech synthesizer family is perhaps one of the best known multilingual text-to-speech products available today. The first commercial version, Infovox SA-101, was developed in Sweden at the Royal Institute of Technology in 1982. The system is originally descended from OVE cascade formant synthesizer (Ljungqvist et al. 1994). Several versions of current system are available for both software and hardware platforms.

The latest full commercial version, Infovox 230, is available for American and British English, Danish, Finnish, French, German, Icelandic, Italian, Norwegian, Spanish, Swedish, and Dutch (Telia 1997). The system is based on formant synthesis and the speech is intelligible but seems to have a bit of Swedish accent. The system has five different built-in voices, including male, female, and child. The user can also create and store individual voices. Aspiration and intonation features are also adjustable. Individual

articulation lexicons can be constructed for each language. For words which do not follow the pronunciation rules, such as foreign names, the system has a specific pronunciation lexicon where the user can store them. The speech rate can be varied up to 400 words per minute. The text may be synthesized also word by word or letter by letter. Also DTMF tones can be generated for telephony applications. The system is available as a half length PC board, RS 232 connected stand-alone desktop unit, OEM board, or software for Macintosh and Windows environments (3.1, 95, NT) and requires only 486DX33MHz with 8 Mb of memory.

Telia has also recently introduced English, German, and Dutch versions of new and improved Infovox 330 software for Windows 95/NT environments. Other languages are under development and will be released soon. Unlike earlier systems, Infovox 330 is based on diphone concatenation of pre-recorded samples of speech. The new system is also more complicated and requires more computational load than earlier versions.

9.2 DECTalk

Digital Equipment Corporation (DEC) has also long traditions with speech synthesizers. The DECtalk system is originally descended from MITalk and Klattalk described earlier in Chapter 2. The present system is available for American English, German and Spanish and offers nine different voice personalities, four male, four female and one child. The present system has probably one of the best designed text preprocessing and pronunciation controls. The system is capable to say most proper names, e-mail and URL addresses and supports a customized pronunciation dictionary. It has also punctuation control for pauses, pitch, and stress and the voice control commands may be inserted in a text file for use by DECtalk software applications. The speaking rate is adjustable between 75 to 650 words per minute (Hallahan 1996). Also the generation of single tones and DTMF signals for telephony applications is supported.

DECtalk software is currently available for Windows 95/NT environments and for Alpha systems running Windows NT or DIGITAL UNIX. A software version for Windows requires at least Intel 486-based computer with 50 MHz processor and 8 Mb of memory. The software provides also an application programming interface (API) that is fully integrated with computer's audio subsystem. Three audio formats are supported, 16- and 8-bit PCM at 11 025 Hz sample rate for standard audio applications and 8-bit μ-law encoded at 8 000 Hz for telephony applications (Hallahan 1996).

The software version has also three special modes, speech-to-wave mode, the log-file mode, and the text-to-memory mode. The speech-to-wave mode, where the output speech is stored into wav-file, is essential for slower Intel machines which are not able to perform real-time speech synthesis. The log-file mode writes the phonemic output in

to file and the text-to-memory mode is used to store synthesized speech data into buffers from where the applications can use them (Hallahan 1996).

A hardware version of DECtalk is available as two different products, DECtalk PC2 and DECtalk Express. DECtalk PC2 is an internal ISA/EISA bus card for IBM compatible personal computers and uses a 10 kHz sample rate. DECtalk Express is an external version of the same device with standard serial interface. The device is very small (92 x 194 x 33 mm, 425 g) and so suitable for portable use. DECtalk speech synthesis is also used in well known Creative Labs Sound Blaster audio cards know as TextAssist. These have also a Voice editing tool for new voices.

The present DECtalk system is based on digital formant synthesis. The synthesizer input is derived from phonemic symbols instead of using stored formant patterns as in a conventional formant synthesizer (Hallahan 1996). The system uses 50 different phonemic symbols including consonants, vowels, diphthongs, allophones, and a silence. Symbols are based on the Arpabet phoneme alphabet which is developed to represent American English phonemes with normal ASCII characters. Also IPA symbols for American English are supported.

Digital is also developing a talking head called DECface (Waters et al. 1993). The system is a simple 2D representation of frontal view. The model consists of about 200 polygons mostly presenting mouth and teeth. The jaw nodes are moved vertically as a function of displacement of the corners of the mouth and the lower teeth are displacement along with the lower jaw. For better realism, the eyelid movements are also animated.

9.3 Bell Labs Text-to-Speech

AT&T Bell Laboratories (Lucent Technologies) has also very long traditions with speech synthesis since the demonstration of VODER in 1939. The first full TTS system was demonstrated in Boston 1972 and released in 1973. It was based on articulatory model developed by Cecil Coker (Klatt 1987). The development process of the present concatenative synthesis system was started by Joseph Olive in mid 1970's (Bell Labs 1997). Present system is based on concatenation of diphones, context-sensitive allophonic units or even of triphones.

The current system is available for English, French, Spanish, Italian, German, Russian, Romanian, Chinese, and Japanese (Möbius et al. 1996). Other languages are under development. The development is focused primarily for American English language with several voices, but the system is multilingual in the sense that the software is identical for all languages, except English. Some language specific information is naturally needed, which is stored externally in separate tables and parameter files.

The system has also good text-analysis capabilities, as well as good word and proper name pronunciation, prosodic phrasing, accenting, segmental duration, and intonation. Bell Laboratories have particular activity for developing statistical methods for handling these problematic aspects. The latest commercial version for American English is available as several products, for example TrueTalk provided by Entropic Research and WATSON FlexTalk by AT&T.

The architecture of the current system is entirely modular (Möbius et al. 1996). It is designed as pipeline presented in Figure 9.1 where each of 13 modules handle one particular step for the process. So members of a research group can work on different modules separately and an improved version of a given module can be integrated anytime as long as the communication between the modules and the structure of the information to be passed along is properly defined. Another advantage of this structure is that it is possible to interrupt and initiate processing anywhere in the pipeline and assess TTS information in that particular point, or to insert tools or programs to modify TTS parameters.

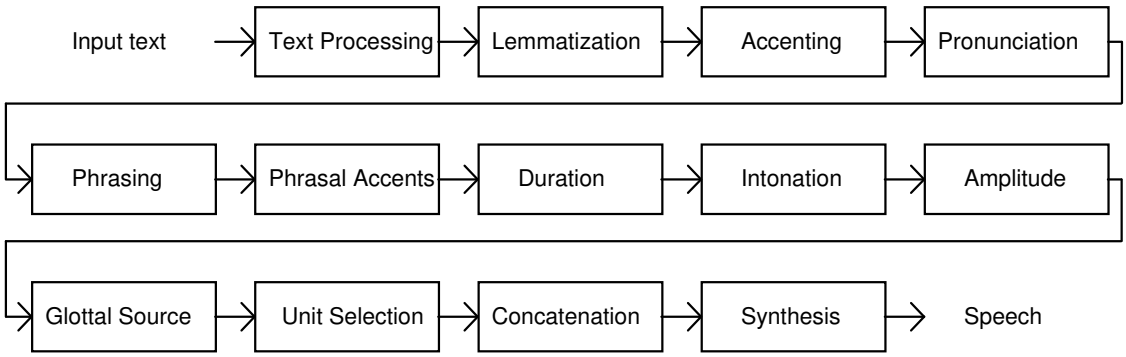

Fig. 9.1. Modules of the English TTS system.

The text processing module handles the end-of-the-sentence detection, text normalization (expansion of numbers, abbreviations etc.), and makes some grammatical analysis. The Accenting module handles the assignment of levels of prominence to various words in the sentence. The pronunciation module handles the pronunciation of words and names and the disambiguation of homographs. The phrasing module contains the breaking of long stretches of text into one or more intonational units. The duration determines the appropriate segmental durations for phonemes in the input on the basis of linguistic information. Intonation module computes the fundamental frequency contour. The glottal source determines the parameters of the glottal source (glottal open quotient, spectral tilt, and aspiration noise) for each sentence. The unit selection module handles the selection of appropriate concatenative units given the phoneme string to be synthesized. Finally, the selected units are concatenated and synthesized (Santen et al. 1997).

Bell Laboratories are also developing an Internet Speech Markup Language with CSTR. The main objective is to combine the present Internet Markup Languages into a single standard.

9.4 Laureate

Laureate is a speech synthesis system developed during this decade at BT Laboratories (British Telecom). To achieve good platform independence Laureate is written in standard ANSI C and it has a modular architecture shown in Figure 9.2 below (Gaved 1993, Morton 1987). The Laureate system is optimized for telephony applications so that lots of attention have been paid for text normalization and pronunciation fields. The system supports also multi-channel capabilities and other features needed in telecommunication applications.

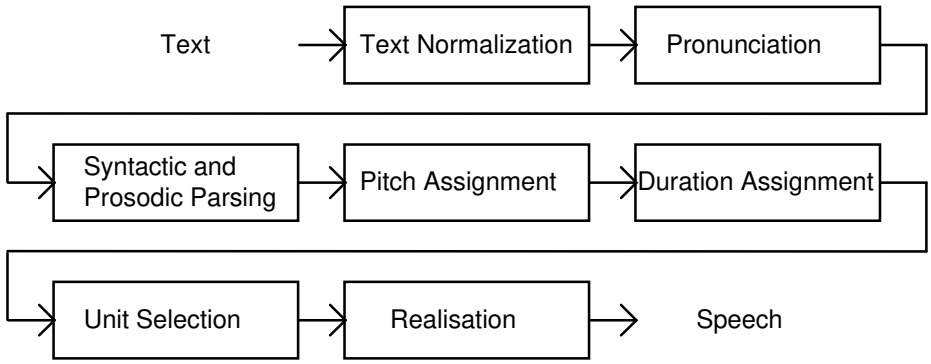

Fig. 9.2. Overview of Laureate.

The current version of Laureate is available only for British and American English with several different accents. Prototype versions for French and Spanish also exist and several other European languages are under development. A talking head for the system has been also recently introduced (Breen et al. 1996). More information, including several pre-generated sound examples and interactive demo, is available at the Laureate home page (BT Laboratories 1998).

9.5 SoftVoice

SoftVoice, Inc. has over 25 years of experience in speech synthesis. It is known for SAM (Software Automatic Mouth) synthesizer for Commodore C64 (SAM-synthesizer) and Amiga (Narrator), Apple (original MacinTalk), and Atari computers in the early 1980's which were probably the first commercial software based systems for personal home computers.

The latest version of SVTTS is the fifth generation multilingual TTS system for Windows is available for English and Spanish with 20 preset voices including males,

females, children, robots, and aliens. Languages and parameters may be changed dynamically during speech. More languages are under development and the user may also create an unlimited number of own voices. The input text may contain over 30 different control commands for speech features. Speech rate is adjustable between 20 and 800 words per minute and the fundamental frequency or pitch between 10 and 2000 Hz. Pitch modulation effects, such as vibrato, perturbation, and excursion, are also included. Vocal quality may be set as normal, breathy, or whispering and the singing is also supported. The output speech may be also listened in word-by-word or letter-by-letter modes. The system can also return mouth shape data for animation and has capable to send synchronization data for the other user's applications. The basic architecture of the present system is based on formant synthesis.

The speech quality of SoftVoice is not probably the best of the available products, but with the large number of control characters and different voices makes it very useful for several kinds of multimedia applications.

9.6 CNET PSOLA

One of the most promising methods for concatenation synthesis was introduced in mid 1980's by France Telecom CNET (Centre National d'Etudes Télécommunications). The synthesizer is a diphone based synthesizer which uses the famous PSOLA algorithm discussed earlier in chapter 5.

The latest commercial product is available from Elan Informatique as ProVerbe TTS system. The concatenation unit used is diphone sampled at 8 kHz rate. The ProVerbe Speech Unit is a serial (RS232 or RS458) connected external device (150x187x37 mm) optimized for telecommunication applications like e-mail reading via telephone. The system is available for American and British English, French, German, and Spanish. The pitch and speaking rate are adjustable and the system contains a complete telephone interface allowing connection directly to the public network. ProVerbe has also an ISA connected internal device which is capable also multichannel operation. Internal device is available also for Russian language and has same features as serial unit.

9.7 ORATOR

ORATOR is a TTS system developed by Bell Communications Research (Bellcore). The synthesis is based on demisyllable concatenation (Santen 1997, Macchi et al. 1993, Spiegel 1993). The latest ORATOR version provides probably one of the most natural sounding speech available today. Special attention on text processing and pronunciation of proper names for American English is given and the system is thus suitable for telephone applications. The current version of ORATOR is available only for American English and supports several platforms, such as Windows NT, Sun, and DECstations.

9.8 Eurovocs

Eurovocs is a text-to-speech synthesizer developed by Technologie & Revalidatie (T&R) in Belgium. It is a small (200 x 110 x 50 mm, 600g) external device with built-in speaker and it can be connected to any system or computer which is capable to send ASCII via standard serial interface RS232. No additional software on computer is needed. Eurovocs system uses the text-to-speech technology of Lernout and Hauspie speech products described in the following chapter, and it is available for Dutch, French, German, Italian, and American English. One Eurovocs device can be programmed with two languages. The system supports also personal dictionaires. Recently introduced improved version contains also Spanish and some improvements in speech quality and device dimensions have been made.

9.9 Lernout & Hauspies

Lernout & Hauspies (L&H) has several TTS products with different features depending on the markets they are used. Different products are available optimized for application fields, such as computers and multimedia (TTS2000/M), telecommunications (TTS2000/T), automotive electronics (TTS3000/A), consumer electronics (TTS3000/C). All versions are available for American English and first two also for German, Dutch, Spanish, Italian, and Korean (Lernout & Hauspie 1997). Several other languages, such as Japanese, Arabic, and Chinese are under development. Products have a customizable vocabulary tool that permits the user to add special pronunciations of words which do not succeed with normal pronunciation rules. With a special transplanted prosody tool it is possible to copy duration and intonation values from recorded speech for commonly used sentences which may be used for example in information and announcement systems.

Recently, a new version for PC multimedia (TTS3000/M) has been introduced for Windows 95/NT with Software Developer's kit (API) and a special E-mail preprocessing software. The E-mail processing software is capable to interpret the initials and names in addresses and handle the header information. The new version contains also Japanese and supports run-time switching between languages. System supports wav-formats with 8 kHz and 11 kHz. The architecture is based on concatenation of rather long speech segments, such as diphones, triphones, and tetraphones.

9.10 Apple Plain Talk

Apple has developed three different speech synthesis systems for their MacIntosh Personal Computers. Systems have different level of quality for different requirements. The PlainTalk products are available for MacIntosh computers only and they are downloadable free from Apple homepage.

MacinTalk2 is the wavetable synthesizer with ten built-in voices. It uses only 150 kilobytes of memory, but has also the lowest quality of PlainTalk family, but runs in almost every Macintosh system.

MacinTalk3 is a formant synthesizer with 19 different voices and with considerably better speech quality compared to MacinTalk2. MacinTalk3 supports also singing voices and some special effects. The system requires at least Macintosh with a 68030 processor and about 300 kb of memory. MacinTalk3 has the largest set of different sounds.

MacinTalkPro is the highest quality product of the family based on concatenative synthesis. The system requirements are also considerably higher than in other versions, but it has also three adjustable quality levels for slower machines. Pro version requires at least 68040 PowerPC processor with operating system 7.0 and uses about 1.5 Mb of memory. The pronunciations are derived from a dictionary of about 65,000 words and 5,000 common names.

9.11 AcuVoice

AcuVoice is a software based concatenative TTS system (AcuVoice 1997). It uses syllable as a basic unit to avoid modeling co-articulation effects between phonemes. Currently the system has only American English male voice, but female voice is promised to release soon. The database consists of over 60 000 speech fragments and requires about 150 Mb of hard disk space. The memory requirement is about 2.7 Mb. The system supports personal dictionaries and allows also the user to make changes to the original database. A dictionary of about 60 000 proper names is also included and names not in the dictionary are produced by letter-to-sound rules which models how humans pronounce the names which are unfamiliar to them. Additions and changes to the dictionary are also possible. The maximum speech rate is system speed dependent and is at least over 20 words per minute. The output of the synthesizer may also be stored in 8- or 16-bit PCM file.

AcuVoice is available as two different products, AV1700 for standard use and AV2001 multichannel developer's kit which is also MS-SAPI compliant. The products are available for Windows 95/NT environments with 16-bit sound card, and for Solaris x86 and SPARC UNIX workstations.

9.12 CyberTalk

CyberTalk is a software based text-to-speech synthesis system for English developed by Panasonic Technologies, Inc. (PTI), USA (Panasonic 1998). The system is a hybrid formant/concatenation system which uses rule-based formant synthesis for vowels and sonorants, and prerecorded noise segments for stops and fricatives. Numbers and some alphanumerical strings are produced separately with concatenation synthesis. The CyberTalk software is available for MS Windows with male and female voices. The sound engine requires 800 kb of memory and the speech data from 360 kb to 3.5 Mb depending on voice configuration. The system has over 100,000 words built-in lexicon and separate customizable user lexicon.

9.13 ETI Eloquence

ETI Eloquence is a software based TTS system developed by Eloquent Technology, Inc., USA, and is currently available for British and American English, Mexican and Castillian Spanish, French, German, and Italian. Other languages, such as Chinese are also under development. For each language the system offers seven built-in voices including male, female, and child. All voices are also easily customizable by user. The system is currently available for Windows95/NT requiring at least 468 processor at 66 MHz and 8 Mb of memory, and for IBM RS/6000 workstations running AIX.

Adjustable features are gender, head size, pitch baseline, pitch fluctuation, roughness, breathiness, speech, and volume. The head size is related to the vocal tract size, low pitch fluctuation produces a monotone sounding voice and a high breathiness value makes the speech sound like a whisper.

The architecture consists of three main modules, the text module, the speech module, and the synthesizer. The text module has components for text normalization and parsing. The speech module uses the information from text module to determine parameter values and durations for the synthesizer. Speech is synthesized with Klatt-style synthesizer with few modifications (Herz 1997).

One special feature in the system is different text processing modes, such as math mode which converts the number 1997 as *one-thousand-ninety-seven* instead of *nineteen-ninety-seven* and several spelling modes, such as radio mode which converts the input string *abc* as *alpha, bravo, charlie*. The system also supports customized dictionaries where the user can add special words, abbreviations and roots for overriding the default pronunciation. The system can handle common difficulties with compound words, such as the *th* between words hothouse and mother and with common abbreviations, such as St. (saint or street).

The system contains also several control symbols for emphasizing a particular word, expressing boredom or excitement, slow down or speed up, switch voices and even languages during the sentence. Virtually any intonation pattern may be generated.

9.14 Festival TTS System

The Festival TTS system was developed in CSTR at the University of Edinburgh by Alan Black and Paul Taylor and in co-operation with CHATR, Japan. The current system is available for American and British English, Spanish, and Welsh. The system is written in C++ and supports residual excited LPC and PSOLA methods and MBROLA database. With LPC method, the residuals and LPC coefficients are used as control parameters. With PSOLA or MBROLA the input may be for example standard PCM files (Black et al. 1997). As a University program the system is available free for educational, research, and individual use. The system is developed for three different aspects. For those who want simply use the system from arbitrary text-to-speech, for people who are developing language systems and wish to include synthesis output, such as different voices, specific phrasing, dialog types and so on, and for those who are developing and testing new synthesis methods.

The developers of Festival are also developing speech synthesis markup languages with Bell Labs and participated development of CHATR generic speech synthesis system at ATR Interpreting Telecommunications Laboratories, Japan. The system is almost identical to Festival, but the main interests are in speech translation systems (Black et al. 1994).

9.15 ModelTalker

ASEL ModelTalker TTS system is under development at University of Delaware, USA. It is available for English with seven different emotional voices, neutral, happy, sad, frustrated, assertive, surprise, and contradiction. English female and child voices are also under development. The system is based on concatenation of diphones and the architecture consists of seven largely independent modules, text analysis, text-to-phoneme rules, part of speech rules, prosodic analysis, discourse analysis, segmental duration calculation, and intonational contour calculation.

9.16 MBROLA

The MBROLA project was initiated by the TCTS Laboratory in the Faculté Polytechnique de Mons, Belgium. The main goal of the project is to develop multilingual speech synthesis for non-commercial purposes and increase the academic research, especially in prosody generation. It is a method like PSOLA, but named

MBROLA, because of PSOLA is a trademark of CNET. The MBROLA-material is available free for non-commercial and non-military purposes (Dutoit et al. 1993, 1996).

The MBROLA v2.05 synthesizer is based on diphone concatenation. It takes a list of phonemes with some prosodic information (duration and pitch) as input and produces speech samples of 16 bits at the sampling frequency of the diphone database currently used, usually 16 kHz. It is therefore not a TTS system since it does not accept raw text as input, but it may be naturally used as a low level synthesizer in a TTS system. The diphone databases are currently available for American/British/Breton English, Brazilian Portuguese, Dutch, French, German, Romanian, and Spanish with male and/or female voice. Several other languages, such as Estonian, are also under development.

The input data required by MBROLA contains a phoneme name, a duration in milliseconds, and a series of pitch pattern points composed of two integers each. The position of the pitch pattern point within the phoneme in percent of its total duration, and the pitch value in Hz at this position. For example, the input "_ 51 25 114" tells the synthesizer to produce a silence of 51 ms, and to put a pitch pattern point of 114 Hz at 25% of 51ms.

9.17 Whistler

Microsoft Whistler (Whisper Highly Intelligent Stochastic TaLkER) is a trainable speech synthesis system which is under development at Microsoft Research, Richmond, USA. The system is designed to produce synthetic speech that sounds natural and resembles the acoustic and prosodic characteristics of the original speaker and the results have been quite promising (Huang et al. 1996, Huang et al. 1997, Acero 1998). The speech engine is based on concatenative synthesis and the training procedure on Hidden Markov Models (HMM). The speech synthesis unit inventory for each individual voice is constructed automatically from unlabeled speech database using the Whisper speech recognition system (Hon et al. 1998). The use of speech recognition for labeling the speech segments is perhaps the most interesting approach for this, usually time-consuming task in concatenative synthesis. The text analysis component is derived from Lernout & Hauspie's TTS engine and, naturally, the speech engine supports MS Speech API and requires less than 3 Mb of memory.

9.18 NeuroTalker

The INM (International Neural Machines, Canada) NeuroTalker is a TTS system with OCR (Optical Character Recognition) for American English with plans to release the major EU languages soon (INM 1997). The system allows the user to add specialized pronounced words and pronunciation rules to the speech database. The system recognizes most of the commonly used fonts, even when mixed or bolded. It is also

capable to separate text from graphics and make corrections to text which can not be sometimes easily corrected through an embedded speller, such as numbers or technical terms. The system requires at least Intel 486DX with 8 Mb of memory and support most scanners available. The NeuroTalker is available as two products, the standard edition with normal recognition and synthesis software, and an audiovisual edition for the visually impaired.

9.19 Listen2

Listen2 is a text-to-speech system from JTS Microconsulting Ltd., Canada, which uses the ProVoice speech synthesizer. The current system is available as an English and international version. The English version contains male and female voices and the international version also German, Spanish, French, and Italian male voices. The languages may be switched during speech and in English version the gender and pitch may be changed dynamically. The speech output may also be stored in a separate wav-file. The system requires at least a 486-processor with 8 Mb of memory and a 16-bit sound card. The system has special e-mail software which can be set to announce for incoming mail with subject and sender information. The speech quality of Listen2 is far away from the best systems, but it is also very inexpensive.

9.20 SPRUCE

SPRUCE (SPeech Response from UnConstrained English) is a high-level TTS system, currently under development at Universities of Bristol and Essex. The system is capable of creating parameter files suitable for driving most of the low-level synthesizers, including both formant and concatenation systems. A parallel formant synthesizer is usually used, because it gives more flexibility than other systems (Tatham et al. 1992a). In general, the system is capable to drive any low-level synthesizer based on diphones, phonemes, syllables, demi-syllables, or words (Lewis et al 1993). The system is successfully used to drive for example the DECtalk, Infovox, and CNET PSOLA synthesizers (Tatham et al. 1992b, 1995, 1996).

SPRUCE architecture consists of two main modules which are written in standard C. The first on is a module for phonological tasks which alter the basic pronunciation of an individual word according to its context, and the second is a module for prosodic task which alters the fundamental frequency and duration throughout the sentence (Lewis et al. 1997).

The system is based on inventory of syllables obtained from recorded natural speech to build the correct output file. The set of syllables is about 10 000 (Tatham et al. 1996). The top level of the system is dictionary based where the pronunciation of certain words are stored for several situations. For example, in weather forecast the set of used words

is quite limited and consists of lots of special concepts, and with announcement systems the vocabulary may be even completely fixed. The word lexicon consists of 100 000 words which requires about 5 Mb disk space (Lewis et al. 1997).

9.21 HADIFIX

HADIFIX (HAlbsilben, DIphone, SufFIXe) is a TTS system for German developed at University of Bonn, Germany. The system is available for both male and female voices and supports control parameters, such as duration, pitch, word prominence and rhythm. Inserting of pauses and accent markers into the input text and synthesis of singing voice are also supported.

The system is based on concatenation of demisyllables, diphones, and suffixes (Portele et al. 1991, 1992). First, the input text is converted into phonemes with stress and phrasing information and then synthesized using different units. For example, the word *Strolch* is formed by concatenating *Stro* and *olch*.

The concatenation of two segments is made by three methods. Diphone concatenation is suitable when there is some kind of stable part between segments. Hard concatenation is the simplest case of putting samples together with for example glottal stops. This also happens at each syllable boundary in demisyllable systems. Soft concatenation takes place at the segment boundaries where the transitions must be smoothed by overlapping (Portele et al. 1994).

The inventory structure consists of 1080 units (750 for initial demisyllables, 150 for diphones, and 180 for suffixes) which is sufficient to synthesize nearly all German words including uncommon sound combinations originating from foreign languages (Portele et al. 1992).

9.22 SVOX

SVOX is a German text-to-speech synthesis system which has been developed at TIK/ETHZ (Swiss Federal Institute of Technology, Zurich). The SVOX system consists of two main modules. The transcription module includes the text analysis and the phonological generation which are speaker and voice independent. Phonological representation is generated from each input sentence and it includes the respective phoneme string, the accent level per syllable, and the phrase boundaries (position, type, and strength). The second one, phono-acoustical module, includes all the speaker-dependent components that are required to generate an appropriate speech signal from the phonological representation (Pfister 1995).

9.23 SYNTE2 and SYNTE3

SYNTE2 was the first full text-to-speech system for Finnish and it was introduced in 1977 after five years of research in Tampere University of Technology (Karjalainen et al. 1980, Laine 1989). The system is a portable microprocessor based stand-alone device with analog formant synthesizer. The basic synthesis device consists of a Motorola 68000 microprocessor with 2048 bytes of ROM and 256 bytes of RAM, a set of special D/A-converters to generate analog control signals, and an analog signal processing part for sound generation, which is a combination of cascade and serial type formant synthesizers. SYNTE2 takes an ASCII string as input and some special characters may be used to control features, such as speech rate, intonation, and phoneme variation (Karjalainen et al. 1980). The information hierarchy of SYNTE2 is presented in Figure 9.3. More detailed discussion of SYNTE2 see (Karjalainen 1978), (Karjalainen et al. 1980), or (Laine 1989).

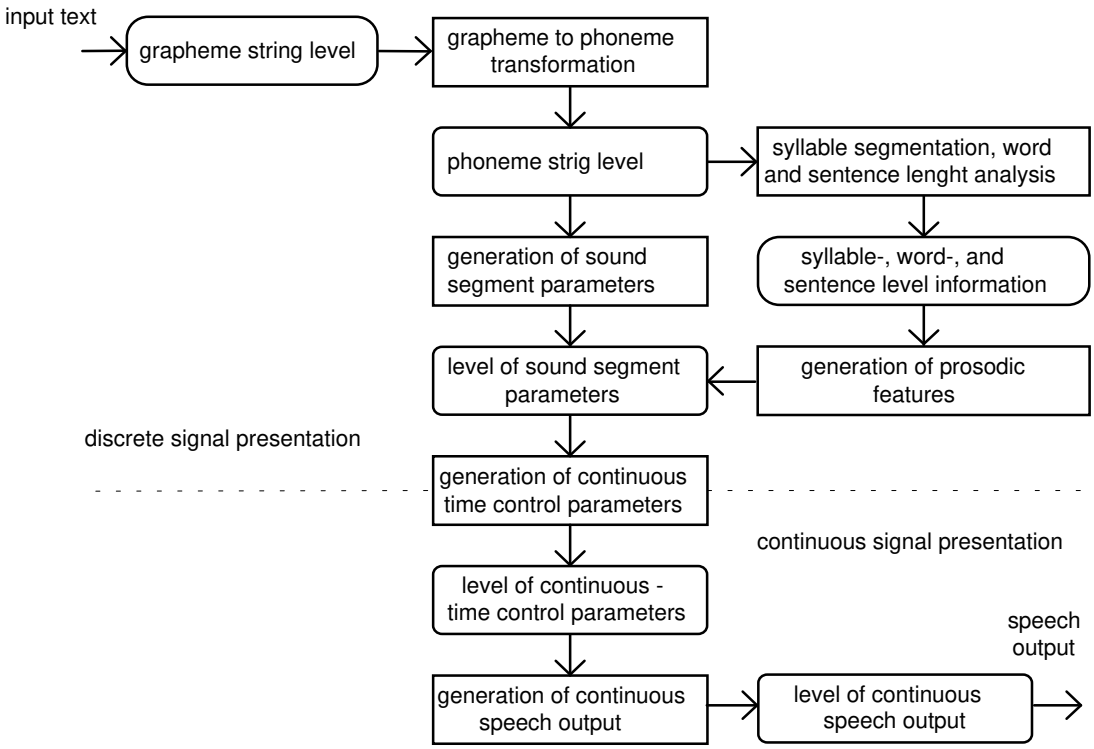

Fig. 9.3. SYNTE2 as an information hierarchy.

An improved version, SYNTE3, was introduced about five years later. The synthesis was based on a new parallel-cascade (PARCAS) model which is described earlier in Chapter 5 and more closely by Laine (1989). The speech quality was slightly improved and the commercial version system is still use.

9.24 Timehouse Mikropuhe

Mikropuhe is a speech synthesizer for Finnish developed by Timehouse Inc. It is currently available for Windows 95/NT and Macintosh computers with one Finnish male voice. Also robotic and Donald Duck voices are available. Several other voices including female voice are under development. The synthesis is based on the microphonemic method concatenating about 10 ms long samples uttered from natural speech. The system uses 22050 Hz sampling frequency with 16 bits, but it works also with 8 bit sound cards.

The controllable features are the speech rate, the pitch, the pitch randomness, the peacefulness, and the duration of pauses between words. The speech rate can be adjusted between about 280 to 3200 characters per minute. The pitch can be set between 25 Hz and 300 Hz and the randomness up to 48 %. The duration of pauses between words can be set up to one second. The latest version of Mikropuhe (4.11) is available only for PC environments and it contains also singing support. All features can be also controlled by control characters within a text. The system also supports a personal abbreviation list with versatile controls and the output of the synthesizer can be stored into a separate wav-file.

9.25 Sanosse

Sanosse synthesis has been developed originally for educational purposes for the University of Turku. The system is based on concatenative synthesis and it is available for Windows 3.1/95/NT environments. The adjustable features are the speech rate, word emphasis, and the pauses between words. The input text can also be synthesized letter-by-letter, word-by-word, or even syllable-by-syllable. The feature can also be controlled with control characters within a text. Sanosse synthesis is currently use in aLexis software which is developed for computer based training for reading difficulties (Hakulinen 1998). The original Sanosse system is also adopted by Sonera for their telephony applications.

9.26 Summary

The product range of text-to-speech synthesizers is very wide and it is quite unreasonable to present all possible products or systems available out there. Hopefully, most of the famous and commonly used products are introduced in this chapter. Most products described here are also demonstrated on the accompanying audio CD described in Appendix A. Some of the currently available speech synthesis products are summarized in Appendix B.

10. SPEECH QUALITY AND EVALUATION

Synthetic speech can be compared and evaluated with respect to intelligibility, naturalness, and suitability for used application (Klatt 1987, Mariniak 1993). In some applications, for example reading machines for the blind, the speech intelligibility with high speech rate is usually more important feature than the naturalness. On the other hand, prosodic features and naturalness are essential when we are dealing with multimedia applications or electronic mail readers. The evaluation can also be made at several levels, such as phoneme, word or sentence level, depending what kind of information is needed.

Speech quality is a multi-dimensional term and its evaluation contains several problems (Jekosh 1993, Mariniak 1993). The evaluation methods are usually designed to test speech quality in general, but most of them are suitable also for synthetic speech. It is very difficult, almost impossible, to say which test method provides the correct data. In a text-to-speech system not only the acoustic characteristics are important, but also text pre-processing and linguistic realization determine the final speech quality. Separate methods usually test different properties, so for good results more than one method should be used. And finally, how to assess the test methods themselves.

The evaluation procedure is usually done by subjective listening tests with response set of syllables, words, sentences, or with other questions. The test material is usually focused on consonants, because they are more problematic to synthesize than vowels. Especially nasalized consonants (/m/ /n/ /ng/) are usually considered the most problematic (Carlson et al. 1990). When using low bandwidth, such as telephone transmission, consonants with high frequency components (/f/ /th/ /s/) may sound very annoying. Some consonants (/d/ /g/ /k/) and consonant combinations (/dr/ /gl/ /gr/ /pr/ /spl/) are highly intelligible with natural speech, but very problematic with synthesized one. Especially final /k/ is found difficult to perceive. The other problematic combinations are for example /lb/, /rp/, /rt/, /rch/, and /rm/ (Goldstein 1995).

Some objective methods, such as Articulation Index (AI) or Speech Transmission Index (STI), have been developed to evaluate speech quality (Pols et al. 1992). These methods may be used when the synthesized speech is used through some transmission channel, but they are not suitable for evaluating speech synthesis in general. This is because there is no unique or best reference and with a TTS system, not only the acoustic characteristics are important, but also the implementation of a high-level part determines the final quality (Pols et al. 1992). However, some efforts have been made to evaluate

objectively for example the quality of automatic segmentation methods in concatenative synthesis (Boeffard et al. 1993).

When repeating the test procedure to the same listening group, the test results may increase significally by the learning effect which means that the listeners get familiar with the synthetic speech they hear and they understand it better after every listening session (Neovius et al. 1993). Concentration problems, on the other hand, may decrease the results especially in segmental methods. Therefore, the decision of using naive or pro listeners in listening tests is important.

Several individual test methods for synthetic speech have been developed during last decades. Some researchers even complain that there are too many existing methods which make the comparisons and standardization procedure more difficult. On the other hand, there is still no test method to give undoubtedly the correct results. The most commonly used methods are introduced in this chapter. Also some computer softwares have been developed for making the test procedure easier to perform. One of these is for example the SAM SOAP (A Speech Output Assessment Package) which is implemented in PC-environment and contains several different test methods (Howard-Jones et al. 1991).

10.1 Segmental Evaluation Methods

With segmental evaluation methods only a single segment or phoneme intelligibility is tested. The very commonly used method to test the intelligibility of synthetic speech is the use of so called rhyme tests and nonsense words. The rhyme tests have several advantages (Jekosh 1993). The number of stimuli is reduced and the test procedure is not time consuming. Also naive listeners can participate without having to be trained and reliable results can be obtained with relatively small subject groups, which is usually from 10 to 20. The learning effects can also be discarded or measured. With these features the rhyme tests are easy and economic to perform. The obtained measure of intelligibility is simply the number of correctly identified words compared to all words and diagnostic information can be given by confusion matrices. Confusion matrices give information how different phonemes are misidentified and help to localize the problem points for development. However, rhyme tests have also some disadvantages. With monosyllabic words only single consonants are tested, the vocabulary is also fixed and public so the system designers may tune their systems for the test, and the listeners might remember the correct answers when participating in the test more than once. For avoiding these problems Jekosh (1992) has presented CLID-test described later in this chapter. Rhyme tests are available for many languages and they are designed for each language individually. The most famous segmental tests are the Diagnostic and Modified Rhyme Tests described below. Some developers or

vendors, such as Bellcore and AT&T have also developed word lists for diagnostic evaluation of their own (Delogu et al 1995).

10.1.1 Diagnostic Rhyme Test (DRT)

The Diagnostic Rhyme Test, introduced by Fairbanks in 1958, uses a set of isolated words to test for consonant intelligibility in initial position (Goldstein 1995, Logan et al. 1989). The test consists of 96 word pairs which differ by a single acoustic feature in the initial consonant. Word pairs are chosen to evaluate the six phonetic characteristics listed in Table 10.1. The listener hears one word at the time and marks to the answering sheet which one of the two words he thinks is correct. Finally, the results are summarized by averaging the error rates from answer sheets. Usually, only total error rate percentage is given, but also single consonants and how they are confused with each other can be investigated with confusion matrices.

Table 10.1. The DRT characteristics.

Characteristics	Description	Examples
Voicing	voiced - unvoiced	veal - feel, dense - tense
Nasality	nasal - oral	reed - deed
Sustension	sustained - interrupted	vee - bee, sheat - cheat
Sibilation	sibilated - unsibilated	sing - thing
Graveness	grave - acute	weed - reed
Compactness	compact - diffuse	key - tea, show - sow

DRT is a quite widely used method and it provides lots of valuable diagnostic information how properly the initial consonant is recognized and it is very useful as a developing tool. However, it does not test any vowels or prosodic features, so it is not suitable for any kind of overall quality evaluation. Other deficiency is that the test material is quite limited and the test items do not occur with equal probability, so it does not test all possible confusions between consonants. Thus, confusions presented as matrices are hard to evaluate (Carlson et al. 1990).

10.1.2 Modified Rhyme Test (MRT)

The Modified Rhyme Test, which is a sort of extension to the DRT, tests for both initial and final consonant apprehension (Logan et al. 1989, Goldstein 1995). The test consists of 50 sets of 6 one-syllable words which makes a total set of 300 words. The set of 6 words is played one at the time and the listener marks which word he thinks he hears on a multiple choice answer sheet. The first half of the words are used for the evaluation of the initial consonants and the second one for the final ones. Table 10.2 summarizes the test format (Shiga et al. 1994).

Table 10.2. Examples of the response sets in MRT.

	A	**B**	**C**	**D**	**E**	**F**
1	bad	back	ban	bass	bat	bath
2	beam	bead	beach	beat	beak	bean
3	bus	but	bug	buff	bun	buck
...						
26	led	shed	red	bed	fed	wed
27	sold	told	hold	fold	gold	cold
28	dig	wig	big	rig	pig	fig
...						

Results are summarized as in DRT, but both final and initial error rates are given individually (Pisoni et al. 1980). Also same kind of problems are faced with MRT as with DRT.

Logan et al. (1989) have presented this test for nine synthesizers and natural speech. They also performed an open response version of the test and found out that the intelligibility decreased significally when the multiple choice answer sheet is excluded. The results are summarized in Figure 10.1 where the three error rates for each synthesizer are shown for the initial consonants, the final consonants, and the average of these respectively. The test and the results are also summarized in Santen et al. (1997).

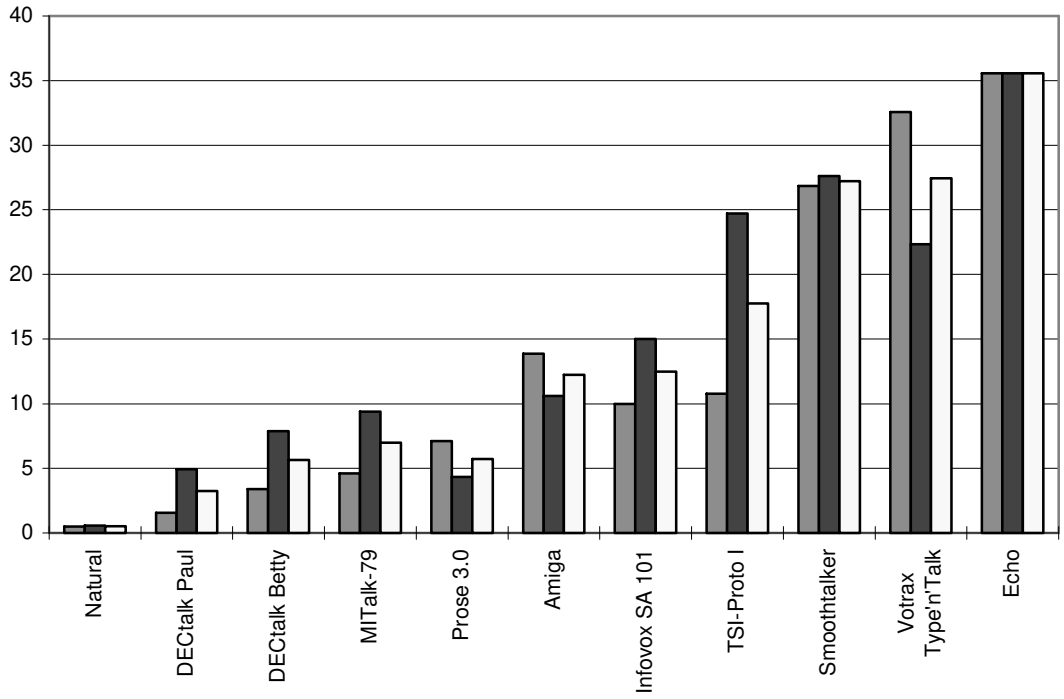

Fig. 10.1. The example of DRT (Logan et al. 1989).

10.1.3 Diagnostic Medial Consonant Test (DMCT)

Diagnostic Medial Consonant Test is same kind of test like rhyme tests described before. The material consists of 96 bisyllable word pairs like "stopper - stocker" which were selected to differ only with their intervocalic consonant. As in DRT, these differences are categorized into six distinctive features and score in each of these categories provides information on diagnosing system deficiencies. The listeners task is to choose correct word from two possible alternatives in the answer sheet. These scores are averaged together to provide an overall measure of system segmental intelligibility.

10.1.4 Standard Segmental Test

The SAM Standard Segmental Test (Jekosh 1993, Pols et al. 1992) uses lists of CV, VC, and VCV nonsense words. All consonants that can occur at the respective positions and three vowels /a/, /i/, and /u/ are the basic items of the test material. For each stimulus, the missing consonant must be filled to the response sheet, so the vowels are not tested at all. The test material is available and used for at least English, German, Swedish, and Dutch. Examples may be found for example in (Goldstein 1995).

10.1.5 Cluster Identification Test (CLID)

The Cluster Identification Test was developed under the ESPRIT project SAM (Jekosh 1992, 1993). The test is based on statistical approach. The test vocabulary is not predefined and it is generated for each test sequence separately. The test procedure consists of three main phases: word generator, phoneme-to-grapheme converter and an automatic scoring module. Word generator generates the test material in phonetic representation. The user can determine the number of words to be generated, the syllable structure (e.g., CCVC, VC,...), and the frequency of occurrence of cluster, initial, medial, and final cluster separately. Syllable structures can also be generated in accordance of their statistical distribution. For example, the structure CCVC occurs more often than CCCVCCC. Used words are usually nonsense. Since most of the synthesizers do not accept phoneme strings, the string has to be converted into graphemic representation. Finally, the error rates are automatically fetched from computer. Initial, medial, and final clusters are scored individually. Also confusion matrices for investigating mix-ups between certain phonemes are easy to generate from the data. In CLID test the open response answering sheet is used and the listener can use either a phonemic or a graphemic transcription. Used sound pressure level (SPL) can be also chosen individually (Kraft et al. 1995).

10.1.6 Phonetically Balanced Word Lists (PB)

In the Phonetically Balanced Word Lists the monosyllabic test words are chosen so that they approximate the relative frequency of phoneme occurrence in each language (Logan et al. 1989, Goldstein 1995). The first this kind of word list was developed in Harvard University during the Second World War. The relative difficulty of the stimulus items was constrained so that items that were always missed or always correct were removed, leaving only those items that provided useful information. The open response set was used. Several other balanced word lists have been developed (Goldstein 1995). For example, the Phonetically Balanced-50 word discrimination test (PB-50) consists of 50 monosyllabic words which approximates the relative frequency of occurrence in English. The PD-100 test is developed to compare for phonetic discrimination and for overall recognition accuracy. The test material includes examples of all possible consonants both in initial and final position and all vowels are in medial position.

10.1.7 Nonsense words and Vowel-Consonant transitions

The use of nonsense words (logotoms), mostly transitions between vowels (V) and consonant (C) is one of the most commonly used evaluation method for synthetic speech. This method provides high error rates and excellent diagnostic material especially when open response set is used. Usually a list of VC, CV, VCV or CVC words is used, but longer words, such as CVVC, VCCV, or CCCVCCC, are sometimes needed. Especially when testing diphone-based systems, longer units must be used to test all CV-, VC-, VV-, and CC-diphone-units. Test words are usually symmetric, like /aka/, /iki/, /uku/ or /kak/, /kik/, /kuk/. Common examples of these methods can be found for example in Carlson et al. (1990) and Dutoit et al. (1994).

10.2 Sentence Level Tests

Several sets of sentences have been developed to evaluate the comprehension of synthetic speech. Sentences are usually chosen to model the occurrence frequency of words in each particular language. Unlike in segmental tests, some items may be missed and the given answer may still be correct, especially if meaningful sentences are used (Pisoni et al. 1980, Allen et al. 1987).

10.2.1 Harvard Psychoacoustic Sentences

Harvard Psychoacoustic Sentences is a closed set of 100 sentences developed to test the word intelligibility in sentence context. The sentences are chosen so that the various segmental phonemes of English are represented in accordance with their frequency of occurrence. The test is easy to perform, no training of the subjects is needed and the scoring is simple. However, when using fixed set of sentences, the learning effect is

very problematic (Pisoni et al. 1980, Kleijn et al. 1998). The first five sentences of the test material are (Allen et al. 1987):

- The birch canoe slid on the smooth planks
- Glue the sheet to the dark blue background
- It's easy to tell the depth of a well
- These days a chicken leg is a rare dish
- Rice is often served in round bowls

Nevertheless the number of sentences is large, the subject may also be familiar with the test material without listening to it. For example, the first one of these sentences is used in many demonstrations or sound examples.

10.2.2 Haskins Sentences

Haskins sentences are also developed to test the speech comprehension in sentence or word level. Unlike in Harvard sentences, the test material is anomalous which means that the missed items can not be concluded from context as easily as with use of meaningful sentences (Pisoni et al. 1980). As in Harvard sentences, a fixed set of sentences is used and due to learning effect the test subjects can be used only once for reliable results. The first five sentences of the test material are (Allen et al. 1987):

- The wrong shot led the farm
- The black top ran the spring
- The great car met the milk
- The old corn cost the blood
- The short arm sent the cow

It is easy to see that these sentences are more difficult to perceive than Harvard sentences and they are not faced in real life situations.

10.2.3 Semantically Unpredictable Sentences (SUS)

The SUS-test is also an intelligibility test on sentence level (Goldstein 1995, Pols et al. 1992). The words to be tested are selected randomly from a pre-defined list of possible candidates. These are mostly mono-syllabic words with some expectations. The test contains five grammatical structures described with examples in Table 10.3 below. As in Haskins sentences, the missed item can not be concluded from textual context.

Table 10.3. Grammatical structures in SUS-test (Jekosh 1993).

	Structure	Example
1	Subject - verb - adverbial	The table walked through the blue truth.
2	Subject - verb - direct object	The strong way drank the day.
3	Adverbial - verb - direct object	Never draw the house and the fact.
4	Q-word - transitive verb - subject - direct object	How does the day love the bright word.
5	Subject - verb - complex direct object	The plane closed the fish that lived.

In the actual test, fifty sentences, ten of each grammatical structure, are generated and played in random order to test subjects. If the test procedure is run more than once, a learning effect may be observed. But because the sentence set is not fixed, the SUS-test is not as sensitive to for example the learning effect as previously described test sentences.

10.3 Comprehension tests

Most of the test methods above are used to test how the single phoneme or word is recognized. In comprehension tests a subject hears a few sentences or paragraphs and answers to the questions about the content of the text, so some of the items may be missed (Allen et al. 1987). It is not important to recognize one single phoneme, if the meaning of the sentence is understood, so the 100% segmental intelligibility is not crucial for text comprehension and sometimes even long sections may be missed (Bernstein et al. 1980). No significant differences were obtained in understanding between natural and synthetic voice (Goldstein 1995). Only with prosody and naturalness the differences are perceptible which may also influence to the concentration of test subjects.

10.4 Prosody evaluation

Evaluation of the prosodic features in synthesized speech is probably one of the most challenging tasks in speech synthesis quality evaluation. Prosody is also one of the least developed parts of existing TTS systems and needs considerable attention for the research in the future. For more discussion about prosody, see Chapter 5.

Prosodic features may be tested with test sentences which are synthesized with different emotions and speaker features. The listeners task is to evaluate for example with five level scale how well the certain characteristic in speech is produced. Evaluation may be made also by other kind of questions, such as "Does the sentence sound like a question, statement or imperative".

10.5 Intelligibility of Proper Names

With some proper names, such as Leicester, Edinburgh, or Begin, the correct pronunciation is usually almost impossible to find from written text. Places like Nice and Begin are also ambiguous when they are in the initial position of the sentence. For applications, such as automatic telephone directory inquiry service, the correct pronunciation of common names is very important. Unfortunately, almost infinite number of first- and surnames exist with many different versions of pronunciation. Without any special rules for names, the mispronunciation percent may be even 40 % (Belhoula 1993). With morphological analysis or pronunciation-by-analogy like methods described in chapter 5 it is possible to increase the speech intelligibility with common names considerably. With a large exception library it is possible to achieve even 90 % intelligibility.

10.6 Overall Quality Evaluation

Methods presented in this chapter are mostly developed for evaluating single features of speech quality. Several methods have been developed to evaluate speech quality in general and these methods are also suitable to measure overall quality or acceptability of synthetic speech (Klaus et al. 1993).

10.6.1 Mean Opinion Score (MOS)

Mean Opinion Score is probably the most widely used and simplest method to evaluate speech quality in general. It is also suitable for overall evaluation of synthetic speech. MOS is a five level scale from bad (1) to excellent (5) and it is also know as ACR (Absolute Category Rating). The listener's task is simply to evaluate the tested speech with scale described in Table 10.4 below. In the same table a kind of opposite version of MOS scale, so called DMOS (Degradation MOS) or DCR (Degradation Category Rating), is presented. DMOS is an impairment grading scale to measure how the different disturbances in speech signal are perceived.

Table 10.4. Scales used in MOS and DMOS.

	MOS (ACR)	DMOS (DCR)
5	Excellent	Inaudible
4	Good	Audible, but not annoying
3	Fair	Slightly annoying
2	Poor	Annoying
1	Bad	Very annoying

However, the use of simple five level scale is easy and provides some instant explicit information, the method gives any segmental or selected information on which parts of the synthesis system should be improved (Goldstein 1995).

10.6.2 Categorical Estimation (CE)

In categorical estimation methods the speech is evaluated by several attributes or aspects independently (Kraft et al. 1995). Possible attributes may be like in Table 10.5 which are from Categorical Rating Test (CRT) performed by Kraft et al (1995) for five German synthesizers.

Table 10.5. Examples of possible attributes for Categorical Estimation.

Attribute	Ratings
pronunciation	not annoying ... very annoying
speed	much too slow ... much too fast
distinctness	very clear ... very unclear
naturalness	very natural ... very unnatural
stress	not annoying ... very annoying
intelligibility	very easy ... very hard
comprehensibility	very easy ... very hard
pleasantness	very pleasant ... very unpleasant

The method indicates well some individual strong and weak points in system and is easy to perform so it is useful for overall assessment of synthetic speech.

10.6.3 Pair Comparison (PC)

Pair comparison methods are usually used to test system overall acceptance (Kraft et al. 1995). An average listener of a speech synthesizer will listen to artificial speech for perhaps hours per day so the small and negligible errors may become very annoying because of their frequent occurrences. Some of this effect may be apparent if few sentences are frequently repeated in the test procedure (Kraft et al. 1995).

Stimuli from each synthesizer are compared in pairs with all n(n-1) combinations, and if more than one test sentence (m) is used each version of a sentence is compared to all the other version of the same sentence. This leads total number of n(n-1)m comparison pairs. The category "equal" is not allowed (Goldstein 1995).

10.6.4 Magnitude and Ratio Estimation

Magnitude and ratio estimation methods are used to make direct numerical estimate to the perceived sensory magnitudes produced by different stimuli, such as loudness and brightness. Nonsensory variables, such as emotional experience may also be used

(Pavlovic et al. 1990). Unlike in pair comparison or categorical estimation, which use the interval scale, magnitude estimation method uses absolute ratio scale. In ratio estimation, a modulus or a standard stimulus is used with tested signal and in magnitude estimation, no modulus is given.

10.7 Field Tests

The most optimal way to test the suitably for individual application is to perform the test in a real environment. In that case the quality of the whole system, not only the speech quality, is usually tested. For example, when testing the reading machines with the optical scanner the overall quality is affected also by the quality of scanner and the text recognition software, or when using speech synthesis in telephony applications, the quality of telephone transmission line is very effective to the overall results. In some situations, it is not possible to perform the test in a real environment, because the environment is not known beforehand. Conditions may be very different for example over the telephone line, in the airplane cockpit or in the classroom.

10.8 Audiovisual Assessment

As mentioned before, the visual information may increase the speech intelligibility significally (Beskow et al. 1997), especially with front vowels and labial consonants. Audiovisual speech is important especially in noisy environments. The intelligibility of audiovisual speech can be evaluated the same way as normal speech. It is also feasible to compare the results to other combinations of natural and synthetic face and speech. It is easy to see from Figure 10.2 that the intelligibility increases with facial information. The test results are based on test made by Beskow et al. (1997). The audio signal was degraded by adding white noise and the signal-to-noise ratio was 3 dB.

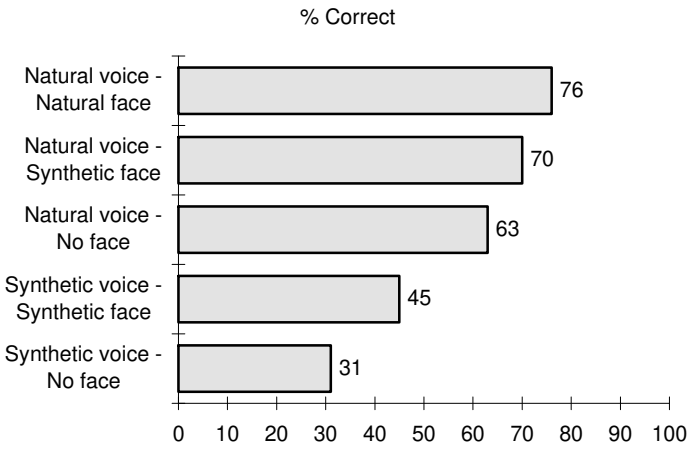

Fig. 10.2. Results from intelligibility tests (Beskow et al. 1997).

It is obvious that the highest improvement is achieved with bilabials and labiodentals. On the other hand, with palatal and velar consonants, there is no improvements to intelligibility due to back articulatory movements (Beskow et al. 1997, Le Goff et al. 1996).

10.9 Summary

Like presented in this chapter, synthesized speech can be evaluated by many methods and at several levels. All methods give some kind of information on speech quality, but it is easy to see that there is no test to give the one and only correct data. Perhaps the most suitable way to test a speech synthesizer is to select several methods to assess each feature separately. For example using segmental, sentence level, prosody, and overall tests together provides lots of useful information, but is on the other hand very time-consuming.

The test methods must be chosen carefully because there is no sense to have the same results from two tests. It is also important to consider in advance what kind of data is needed and why. It may be even reasonable to test the method itself with a very small listening group to make sure the method is reasonable and will provide desirable results.

The assessment methods need to be developed as well as speech synthesizers. Feedback from real users is essential and necessary to develop speech synthesis and the assessment methods.

11. CONCLUSIONS AND FUTURE STRATEGIES

Speech synthesis has been developed steadily over the last decades and it has been incorporated into several new applications. For most applications, the intelligibility and comprehensibility of synthetic speech have reached the acceptable level. However, in prosodic, text preprocessing, and pronunciation fields there is still much work and improvements to be done to achieve more natural sounding speech. Natural speech has so many dynamic changes that perfect naturalness may be impossible to achieve. However, since the markets of speech synthesis related applications are increasing steadily, the interest for giving more efforts and funds into this research area is also increasing. Present speech synthesis systems are so complicated that one researcher can not handle the entire system. With good modularity it is possible to divide the system into several individual modules whose developing process can be done separately if the communication between the modules is made carefully.

The three basic methods used in speech synthesis have been introduced in Chapter 5. The most commonly used techniques in present systems are based on formant and concatenative synthesis. The latter one is becoming more and more popular since the methods to minimize the problems with the discontinuity effects in concatenation points are becoming more effective. The concatenative method provides more natural and individual sounding speech, but the quality with some consonants may vary considerably and the controlling of pitch and duration may be in some cases difficult, especially with longer units. However, with for example diphone methods, such as PSOLA may be used. Some other efforts for controlling of pitch and duration have been made by for example Galanes et al. (1995). They proposed an interpolation/decimation method for resampling the speech signals. With concatenation methods the collecting and labeling of speech samples have usually been difficult and very time-consuming. Currently most of this work can be done automatically by using for example speech-recognition.

With formant synthesis the quality of synthetic speech is more constant, but the speech sounds slightly more unnatural and individual sounding speech is more difficult to achieve. Formant synthesis is also more flexible and allows a good control of fundamental frequency. The third basic method, the articulatory synthesis, is perhaps the most feasible in theory especially for stop consonants because it models the human articulation system directly. On the one hand, the articulatory based methods are usually rather complex and the computational load is high, so the potential has not been realized yet. On the other hand, computational capabilities are increasing rapidly and the analysis

methods of speech production are developing fast, so the method may be useful in the future.

Naturally, some combinations and modifications of these basic methods have been used with variable success. An interesting approach is to use a hybrid system where the formant and concatenative methods have been applied in parallel to phonemes where they are the most suitable (Fries 1993). In general, combining the best parts of the basic methods is a good idea, but in practice, controlling of synthesizer may become difficult.

Also some speech coding methods have been applied to speech synthesis, such as Linear Predictive Coding and Sinusoidal Modeling. Actually, the first speech synthesizer, VODER, was developed from the speech coding system VOCODER (Klatt 1987, Schroeder 1993). Linear Prediction has been used for several decades, but with the basic method the quality has been quite poor. However, with some modifications, such as Warped Linear Prediction (WLP), considerable achievements have been reported (Karjalainen et al. 1998). Warped filtering takes advantage of hearing properties, so it is perhaps useful in all source-filter based synthesis methods. Sinusoidal models have also been applied to speech synthesis for about a decade. Like PSOLA methods, the sinusoidal modeling is best suited for periodic signals, but the representation of unvoiced speech is difficult. However, the sinusoidal methods have been found useful with singing voice synthesis (Macon 1996).

Several normal speech processing techniques may be used also with synthesized speech. For example, adding some reverberation it may be possible to increase the pleasantness of synthetic speech afterwards. Other effects, such as digital filtering, chorus, etc., can be also be used to generate different voices. However, using these kind of methods may increase the computational load. Most information of the speech signal is focused at the frequency range less than 10 kHz. However, using higher sample rate than necessary, the speech may sound slightly more pleasant.

Some other techniques have been applied to speech synthesis, such as Artificial Neural Networks and Hidden Markov Models. These methods have been found promising for controlling the synthesizer parameters, such as gain, duration, and fundamental frequency.

As mentioned earlier, the high-level synthesis is perhaps the least developed part of present synthesizers and needs special attention in the future. Especially controlling prosodic features has been found very difficult and the synthesized speech still sounds usually synthetic or monotonic. The methods for correct pronunciation have been developed steadily during last decades and the present systems are quite good, but improvements with especially proper names are needed. Text preprocessing with numbers and some context-dependent abbreviations is still very problematic. However,

the development of semantic parsing or text understanding techniques may provide a major improvement in high-level speech synthesis.

As long as speech synthesis needs to be developed, the evaluation and assessment play one of the most important roles. Different levels of testing and the most common test methods have been discussed in the previous chapter. Before performing a listening test, the method used should be tested with smaller listener group to find out possible problems and the subjects should be chosen carefully. It is also impossible to say which test method provides the valid data and it is perhaps reasonable to use more than one test.

It is quite clear that there is still very long way to go before text-to-speech synthesis, especially high-level synthesis, is fully acceptable. However, the development is going forward steadily and in the long run the technology seems to make progress faster than we can imagine. Thus, when developing a speech synthesis system, we may use almost all resources available, because in few years todays high resources are available in every personal computer. Regardless how fast the development process will be, speech synthesis, whenever used in low-cost calculators or state-of-the-art multimedia solutions, has probably the most promising future. If speech recognition systems someday achieve a generally acceptable level, we may develop for example a communication system where the system may first analyze the speakers' voice and its characteristics, transmit only the character string with some control symbols, and finally synthesize the speech with individual sounding voice at the other end. Even interpretation from a language to another may became feasible. However, it is obvious that we must wait for several years, maybe decades, until such systems are possible and commonly available.

REFERENCES AND LITERATURE

Abadjieva E., Murray I., Arnott J. (1993). Applying Analysis of Human Emotion Speech to Enhance Synthetic Speech. *Proceedings of Eurospeech 93* (2): 909-912.

Acero A. (1998). Source-Filter Models for Time-Scale Pitch-Scale Modification of Speech. *Proceedings of ICASSP98*.

AcuVoice, Inc. Homepage (1998). <http://www.acuvoice.com>.

Allen J., Hunnicutt S., Klatt D. (1987). *From Text to Speech: The MITalk System*. Cambridge University Press, Inc.

Altosaar T., Karjalainen M., Vainio M. (1996). A Multilingual Phonetic Representation and Analysis for Different Speech Databases. *Proceedings of ICSLP 96* (3).

Amundsen M. (1996). *MAPI, SAPI, and TAPI Developers Guide*. Sams Publishing. <http://book.ygm.itu.edu.tr/Book/mapi_sapi/index.htm>

Apple Speech Technologies Home Page (1998). <http://www.apple.com/macos/speech/>.

Barber S., Carlson R., Cosi P., Di Benedetto M., Granström B., Vagges K. (1989). A Rule Based Italian Text-to-Speech System. *Proceedings of Eurospeech 89* (1): 517-520.

Belhoula K. (1993). Rule-Based Grapheme-to-Phoneme Conversion of Names. *Proceedings of Eurospeech 93* (2): 881-884.

Bell Laboratories TTS Homepage (1998). <http://www.bell-labs.com/project/tts/>.

Bellcore ORATOR Homepage (1998). <http://www.bellcore.com/ORATOR>.

Benoit C. (1995). Speech Synthesis: Present and Future. *European Studies in Phonetic & Speech Communication*. Netherlands. pp. 119-123.

Bernstein J., Pisoni D. (1980). Unlimited Text-to-Speech System: Description and Evaluation of a Microprocessor Based Device. *Proceedings of ICASSP 80* (3): 574-579.

Beskow J. (1996). Talking Heads - Communication, Articulation and animation. *Proceedings of Fonetik-96*: 53-56.

Beskow J., Dahlquist M., Granström B., Lundeberg M., Spens K-E.,. Öhman T. (1997). The Teleface Project - Disability, Feasibility, and Intelligibility. *Proceedings of Fonetik97*, Swedish Fonetics Conf., Umea, Sweden. <http://www.speech.kth.se/~magnusl/teleface_f97.html>

Beskow K., Elenius K., McGlashan S. (1997). The OLGA Project: An Animated Talking Agent in a Dialogue System. *Proceedings of Eurospeech 97*. <http://www.speech.kth.se/multimodal/papers/>

Black A., Taylor P. (1994). CHATR: A Generic Speech Synthesis System. *COLING94*, Japan.

Black A., Taylor P. (1997). *Festival Speech Synthesis System: System Documentation (1.1.1)*. Human Communication Research Centre Technical Report HCRC/TR-83.

Boeffard O., Cherbonnel B., Emerard F., White S. (1993). Automatic Segmentation and Quality Evaluation of Speech Unit Inventories for Concatenation-Based, Multilingual PSOLA Text-to-Speech Systems. *Proceedings of Eurospeech 93* (1): 1449-1452.

Breen A. (1992). Speech Synthesis Models: A Review. *Electronics & Communication Engineering Journal*, vol. 4: 19-31.

Breen A., Bowers E., Welsh W. (1996). An Investigation into the Generation of Mouth Shapes for a Talking Head. *Proceedings of ICSLP 96* (4).

BT Laboratories Laureate home page (1998). <http://www.labs.bt.com/innovate/speech/laureate>

Campos G., Gouvea E. (1996). Speech Synthesis Using the CELP Algorithm. *Proceedings of ICSLP 96* (3).

Carlson R., Fant G., Gobl C., Granström B., Karlsson I., Lin Q. (1989). Voice Source Rules for Text-to-Speech Synthesis. *Proceedings of ICASSP 89* (1): 223-226.

Carlson R., Granström B., Nord L. (1990). Evaluation and Development of the KTH Text-to-Speech System on the Segmental Level. *Proceedings of ICASSP 90* (1): 317-320.

Cawley G., Noakes B. (1993a). Allophone Synthesis Using a Neural Network. *Proceedings of the First World Congress on Neural Networks (WCNN-93)* (2): 122-125. <http://www.sys.uea.ac.uk/~gcc>.

Cawley G., Noakes B. (1993b). LSP Speech Synthesis Using Backpropagation Networks. *Proceedings fo the IEE International Conference on Artificial Neural Networks (ANN-93)*: 291-293. <http://www.sys.uea.ac.uk/~gcc>.

Cawley G. (1996). *The Application of Neural Networks to Phonetic Modelling*. PhD. Thesis, University of Essex, England. <http://www.sys.uea.ac.uk/~gcc/thesis.html>

Charpentier F., Moulines E. (1989). Pitch-Synchronous Waveform Prosessing Techniques for Text-to-Speech Synthesis Using Diphones. *Proceedings of Eurospeech 89* (2): 13-19.

Charpentier F., Stella M. (1986). Diphone Synthesis Using an Overlap-Add Technique for Speech Waveforms Concatenation. *Proceedings of ICASSP 86* (3): 2015-2018.

Childers D., Hu H. (1994). Speech Synthesis by Glottal Excited Linear Prediction. *Journal of the Acoustical Society of America, JASA* vol. 96 (4): 2026-2036.

Cohen M., Massaro D. (1993). Modelling Coarticulation in Synthetic Visual Speech. *Proceedings of Computer Animation 93*, Suisse.

Cole R., Mariani J., Uszkoreit H., Zaenen A., Zue V. (Editors) (1995). *Survey of the State of the Art in Human Language Technology*. <http://www.cse.ogi.edu/CSLU/HLTsurvey/>

Cowie R., Douglas-Cowie E. (1996). Automatic Statistical Analysis of the Signal and Prosodic Signs of Emotion in Speech. *Proceedings of ICSLP 96* (3).

Delogu C., Paolini A., Ridolfi P., Vagges K. (1995). Intelligibility of Speech Produced by Texto-to-Speech Systems in Good and Telephonic Condtions. *Acta Acoustica* 3 (1995): 89-96.

Dettweiler H., Hess W. (1985). Concatenation Rules for Demisyllable Speech Synthesis. *Proceedings of ICASSP 85* (2): 752-755.

Donovan R. (1996). *Trainable Speech Synthesis*. PhD. Thesis. Cambridge University Engineering Department, England. <ftp://svr-ftp.eng.cam.ac.uk/pub/reports/donovan_thesis.ps.Z>.

Dutoit T. (1994). High Quality Text-to-Speech Synthesis: A Comparison of Four Candidate Algorithms. *Proceedings of ICASSP 94* (1): 565-568.

Dutoit T., Leich H. (1992). Improving the TD-PSOLA Text-to-Speech Synthesizer with a Specially Designed MBE Re-Synthesis of the Segments Database. *Proceedings of EUSIPCO-92* (1): 343-346.

Dutoit T., Leich H. (1993). MBR-PSOLA: Text-to-Speech Synthesis Based on an MBE Re-Synthesis of the Segments Database. *Speech Communication,* vol. 13: 435-440.

Dutoit T., Pagel V., Pierret N., Bataille F., Vrecken O. (1996). The MBROLA Project: Towards a Set of High Quality Speech Synthesizers Free of Use for Non Commercial Purposes. *Proceedings of ICSLP 96* (3).

Dynastat, Inc. Homepage (1997). <http://www.realtime.net/dynastat/>.

ELAN Informatique Homepage (1998). <http://www.elan.fr/speech/>.

ETI Eloquence Home Page (1998). <http://www.eloq.com/eti0elo.html>.

Eurovocs Homepage (1998). <http://www.elis.rug.ac.be/t%26i/eurovocs.htm>.

Falaschi A., Giustiniani M., Verola M. (1989). A Hidden Markov Model Approach to Speech Synthesis. *Proceedings of Eurospeech 89* (2): 187-190.

Fant G. (1970). *Acoustic Theory of Speech Production*. Mouton, The Hague.

Flanagan J. (1972). *Speech Analysis, Synthesis, and Perception*. Springer-Verlag, Berlin-Heidelberg-New York.

Flanagan J., Rabiner L. (Editors) (1973). *Speech Synthesis*. Dowden, Hutchinson & Ross, Inc., Pennsylvania.

Fries G. (1993). Phoneme-Depended Speech Synthesis in the Time and Frequency Domains. *Proceedings of Eurospeech 93* (2): 921-924.

Fujisaki H., Ljungqvist M., Murata H. (1993). Analysis and Modeling of Word Accent and Sentence Intonation in Swedish. *Proceedings of ICASSP 93* (2): 211-214.

Galanes F., Savoji M., Pardo J. (1995). Speech Synthesis System Based on a Variable Decimation/Interpolation Factor. *Proceedings of ICASSP 95*: 636-639.

Gaved M. (1993). Pronunciation and Text Normalisation in Applied Text-to-Speech Systems. *Proceedings of Eurospeech 93* (2): 897-900.

George E. (1998). Practical High-Quality Speech and Voice Synthesis Using Fixed Frame Rate ABS/OLA Sinusoidal Modeling. *Proceedings of ICASSP98*.

Goldstein M. (1995). Classification of Methods Used for Assessment of Text-to-Speech Systems According to the Demands Placed on the Listener. *Speech Communication* vol. 16: 225-244.

Gonzalo E., Olaszy G., Németh G. (1993). Improvements of the Spanish Version of the MULTIVOX Text-to-Speech System. *Proceedings of Eurospeech 93* (2): 869-872.

HADIFIX Speech Synthesis Homepage (1997). University of Bonn. <http://www.ikp.uni-bonn.de/~tpo/Hadifix.en.html>

Hakulinen J. (1998). *Suomenkieliset puhesynteesiohjelmistot (The Software Based Speech Synthesizers for Finnish)*. Report Draft, University of Tampere, Department of Computing Science, Speech Interfaces, 26.8.1998. <http://www.cs.uta.fi/research/hci/SUI/reports/ra0298jh.html>.

Hallahan W. (1996). DECtalk Software: Text-to-Speech Technology and Implementation. *Digital Technical Journal*. <http://www.digital.com/DTJ01>.

Hertz S. (1997). *The ETI-Eloquence Text-to-Speech System*. White Paper, Eloquent Technology Inc. <http://www.eloq.com/White1297-1.htm>.

Hess W. (1992). Speech Synthesis - A Solved Problem? *Proceedings of EUSIPCO 92* (1): 37-46.

Heuft B., Portele T., Rauth M. (1996). Emotions in Time Domain Synthesis. *Proceedings of ICSLP 96* (3).

Hirakawa T. (1989). Speech Synthesis Using a Waveform Dictionary. *Proceedings of Eurospeech 89* (1): 140-143.

Holmes W., Holmes J., Judd M. (1990). Extension of the Bandwith of the JSRU Parallel-Formant Synthesizer for High Quality Synthesis of Male and Female Speech. *Proceedings of ICASSP 90* (1): 313-316.

Hon H., Acero A., Huang X., Liu J., Plumpe M. (1998). Automatic Generation of Synthesis Units for Trainable Text-to-Speech Systems. *Proceedings of ICASSP 98 (CD-ROM).*

Howard-Jones P., SAM Partnership. 'SOAP' - A Speech Output Assessment Package for Controlled Multilingual Evaluation of Synthetic Speech. *Proceedings of Eurospeech 91* (1): 281-283.

Huang X., Acero A., Adcock J., Hon H., Goldsmith J., Liu J., Plumpe M. (1996). Whistler: A Trainable Text-to-Speech System. *Proceedings of ICSLP96* (4).

Huang X., Acero A., Hon H., Ju Y., Liu J., Mederith S., Plumpe M. (1997). Recent Improvements on Microsoft's Trainable Text-to-Speech System - Whistler. *Proceedings of ICASSP97* (2): 959-934.

Hunt A., Black A. (1996). Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database. *Proceedings of ICASSP 96*: 373-376.

INM Homepage (1997). <http://www.ineural.com/products.html>.

IPA (1998). International Phonetic Association Homepage. <http://www.arts.gla.ac.uk/IPA/ipa.html>.

ISO/IEC CD 14496-3TTS (1997). Information Technology - Coding of Audiovisual Objects - Part 3: Audio - Subpart 6: Text-to-Speech.

Jekosch U. (1992). The Cluster-Identification Test. *Proceedings of ICSLP 92* (1): 205-208.

Jekosch U. (1993). Speech Quality Assessment and Evaluation. *Proceedings of Eurospeech 93* (2): 1387-1394.

Karjalainen M. (1978). *An Approach to Hierarchical Information Process With an Application to Speech Synthesis by Rule*. Doctorial Thesis. Tampere University of Technology.

Karjalainen M., Altosaar T. (1991). Phoneme Duration Rules for Speech Synthesis by Neural Networks. *Proceedings of Eurospeech 91* (2): 633-636.

Karjalainen M., Altosaar T., Vainio M. (1998). Speech Synthesis Using Warped Linear Prediction and Neural Networks. *Proceedings of ICASSP 98*.

Karjalainen M., Laine U., Toivonen R. (1980). Aids for the Handicapped Based on "SYNTE 2" Speech Synthesizer. *Proceedings of ICASSP 80* (3): 851-854.

Karlsson I., Neovius L. (1993). Speech Synthesis Experiments with the GLOVE Synthesizer. *Proceedings of Eurospeech 93* (2): 925-928.

Klatt D. (1980). Software for a Cascade/Parallel Formant Synthesizer. *Journal of the Acoustical Society of America, JASA*, Vol. 67: 971-995.

Klatt D. (1982). The Klattalk Text-to-Speech Conversion System. *Proceedings of ICASSP 82* (3): 1589-1592.

Klatt D. (1987) Review of Text-to-Speech Conversion for English. *Journal of the Acoustical Society of America, JASA* vol. 82 (3), pp.737-793.

Klatt D., Klatt L. (1990). Analysis, Synthesis, and Perception of Voice Quality Variations Among Female and Male Listeners. *Journal of the Acoustical Society of America, JASA* vol. 87 (2): 820-857.

Klaus H., Klix H., Sotscheck J., Fellbaum K. (1993). An Evaluation System for Ascertaining the Quality of Synthetic Speech Based on Subjective Category Rating Tests. *Proceedings of Eurospeech 93* (3): 1679-1682.

Kleijn K., Paliwal K. (Editors) (1998). *Speech Coding and Synthesis*. Elsevier Science B.V., The Netherlands.

Kortekaas R., Kohlrausch A. (1997). Psychoacoustical Evaluation of the Pitch-Synchronous Overlap-and-Add Speech-Waveform Manipulation Technique Using Single-Formant Stimuli. *Journal of the Acoustical Society of America, JASA*, Vol. 101 (4): 2202-2213.

Kraft V., Portele T. (1995). Quality Evaluation of Five German Speech Synthesis Systems. *Acta Acustica* 3 (1995): 351-365.

Kröger B. (1992). Minimal Rules for Articulatory Speech Synthesis. *Proceedings of EUSIPCO92* (1): 331-334.

Laine U. (1982). PARCAS, a New Terminal Analog Model for Speech Synthesis. *Proceedings of ICASSP 82* (2).

Laine U. (1989). *Studies on Modelling of Vocal Tract Acoustics with Applications to Speech Synthesis*. Thesis for the degree of Doctor of Technology. Helsinki University of Technology.

Laine U., Karjalainen M., Altosaar T. (1994). Warped Linear Prediction (WLP) in Speech Synthesis and Audio Processing. *Proceedings of ICASSP94* (3): 349-352.

Le Goff B., Benoit C. (1996). A Text-to-Audiovisual-Speech Synthesizer for French. *Proceedings of ICSLP96*.

Lee K. (1989). Hidden Markov Models: Past, Present, and Future. *Proceedings of Eurospeech 89* (1): 148-155.

Lehtinen L. (1990). *Puhesynteesi aika-alueessa (Speech Synthesis in Time-Domain).* Lic. Thesis, University of Helsinki.

Lehtinen L., Karjalainen M. (1989). Individual Sounding Speech Synthesis by Rule Using the Microphonemic Method. *Proceedings of Eurospeech 89* (2): 180-183.

Lernout & Hauspies (L&H) Speech Technologies Homepage (1998). <http://www.lhs.com/speechtech/>.

Lewis E., Tatham M. (1993). A Generic Front End for Text-to-Speech Synthesis Systems. *Proceedings of Eurospeech 93* (2): 913-916.

Lewis E., Tatham M. (1997). *SPRUCE - High Specification Text-to-Speech Synthesis.* <http://www.cs.bris.ac.uk/~eric/research/spruce97.html>.

Lindström A., Ljungqvist M., Gustafson K. (1993). A Modular Architecture Supporting Multiple Hypotheses for Conversion of Text to Phonetic and Linguistic Entities. *Proceedings of Eurospeech 93* (2): 1463-1466.

Listen2 Homepage (1997). <http://www.islandnet.com/jts/listen2.htm>.

Ljungqvist M., Lindström A., Gustafson K. (1994). A New System for Text-to-Speech and Its Application to Swedish. *ICSLP 94* (4): 1779-1782.

Logan J., Greene B., Pisoni D. (1989). Segmental Intelligibility of Synthetic Speech Produced by Rule. *Journal of the Acoustical Society of America, JASA* vol. 86 (2): 566-581.

Lukaszewicz K., Karjalainen M. (1987). Microphonemic Method of Speech Synthesis. *Proceedings of ICASSP87* (3): 1426-1429.

Macchi M., Altom M., Kahn D., Singhal S., Spiegel M. (1993). Intelligibility as a Function of Speech Coding Method for Template-Based Speech Synthesis. *Proceedings of Eurospeech 93* (2): 893-896.

Macon M. (1996). *Speech Synthesis Based on Sinusoidal Modeling*. Doctorial Thesis, Georgia Institute of Technology.

Macon M., Clements C. (1996). Speech Concatenation and Synthesis Using an Overlap-Add Sinusoidal Model. *Proceedings of ICASSP 96*: 361-364.

Macon M., Jensen-Link L., Oliverio J., Clements M., George E. (1997). A Singing Voice Synthesis System Based on Sinusoidal Modeling. *Proceedings of ICASSP97*.

Mariniak A. (1993). A Global Framework for the Assessment of Synthetic Speech Without Subjects. *Proceedings of Eurospeech 93* (3): 1683-1686.

MBROLA Project Homepage (1998). <http://tcts.fpms.ac.be/synthesis/mbrola.html>.

McAulay R., Quatieri T. (1986). Speech Analysis-Synthesis Based on Sinusoidal Representation. *Proceedings of ASSP-34* (4): 744-754.

Meyer P., Rühl H., Krüger R., Kugler M., Vogten L., Dirkensen A., Belhoula K. PHRITTS - A Text-to-Speech Synthesizer for the German Language. *Proceedings of Eurospeech 93* (2): 877-880.

ModelTalker Homepage (1997). University of Delaware (ASEL). <http://www.asel.udel.edu/speech/Dsynterf.html>.

Morton K. (1987). The British Telecom Research Text-to-Speech Synthesis System - 1984-1986. *Speech Production and Synthesis*. Unpublished PhD Thesis. University of Essex. pp. 142-172. <http://wrangler.essex.ac.uk/speech/archive/bt>.

Morton K. (1991). Expectations for Assessment Techniques Applied to Speech Synthesis. *Proceedings of the Institute of Acoustics* vol. 13 Part 2. <http:wrangler.essex.ac.uk /speech/archive/assess/assess.html>.

Moulines E., Emerard F., Larreur D., Le Saint Milon J., Le Faucheur L., Marty F., Charpentier F., Sorin C. (1990). A Real-Time French Text-to-Speech System Generating High-Quality Synthetic Speech. *Proceedings of ICASSP 90* (1): 309-312.

Moulines E., Laroche J. (1995). Non-Parametric Techniques for Pitch-Scale Modification of Speech. *Speech Communication* 16 (1995): 175-205.

MPEG Homepage (1998). <http://drogo.cselt.stet.it/mpeg/>

Murray I., Arnott J., Alm N., Newell A. (1991). A Communication System for the Disabled with Emotional Synthetic Speech Produced by Rule. *Proceedings of Eurospeech 91* (1): 311-314.

Murray I., Arnott L. (1993). Toward the Simulation of Emotions in Synthetic Speech: A Review of the Literature on Human Vocal Emotion. *Journal of the Acoustical Society of America, JASA* vol. 93 (2): 1097-1108.

Murray I., Arnott L. (1996) Synthesizing Emotions in Speech: Is It Time to Get Excited? *Proceedings of ICSLP 96* (3).

Möbius B., Schroeter J., Santen J., Sproat R., Olive J. (1996). Recent Advances in Multilingual Text-to-Speech Synthesis. *Fortschritte der Akustik, DAGA-96*.

Möbius B., Sproat R., Santen J., Olive J. (1997). The Bell Labs German Text-to-Speech System: An Overview. *Proceedings of the European Conference on Speech Communication and Technology* vol. 5: 2443-2446.

Neovius L., Raghavendra P. (1993). Comprehension of KTH Text-to-Speech with "Listening Speed" Program. *Proceedings of Eurospeech 93* (3): 1687-1690.

Ohala J. (1996). Ethological Theory and the Voice Expression of Emotion in the Voice. *Proceedings of ICSLP 96* (3).

Olaszy G. (1989). MULTIVOX - A Flexible Text-to-Speech System for Hungarian, Finnish, German, Esperanto, Italian and Other Laguages for IBM-PC. *Proceedings of Eurospeech 89* (2): 525-528.

Olaszy G., Németh G. (1997). Prosody Generation for German CTS/TTS Systems (From Theoretical Intonation Patterns to Practical Realisation). *Speech Communication*, vol. 21 (1997): 37-60.

Oliveira L., Viana M., Trancoso I. (1992). A Rule Based Text-to-Speech System for Portuguese. *Proceedings of ICASSP 92* (2): 73-76.

O'Saughnessy D. (1987). *Speech Communication - Human and Machine*, Addison-Wesley.

Panasonic CyberTalk Homepage (1998). <http://www.research.panasonic.com/pti/stl_web_demo/demo.html>.

Pavlovic C., Rossi M., Espesser R. (1990). Use of the Magnitude Estimation Technique for Assessing the Performance of Text-to-Speech Synthesis System. *Journal of the Acoustical Society of America, JASA* vol. 87 (1): 373-382.

Pfister B. (1995). *The SVOX Text-to-Speech System*. Computer Engineering and Networks Laboratory, Speech Processing Group, Swiss Federal Institute of Technology, Zurich. <http://www.tik.ee.ethz.ch/~spr/publications/Pfister:95d.ps>.

Pisoni D, Hunnicutt S. (1980). Perceptual Evaluation of MITalk: The MIT Unrestricted Text-to-Speech System. *Proceedings of ICASSP 80* (3): 572-575.

Pols L. (1994). Voice Quality of Synthetic Speech: Representation and Evaluation. *Proceedings of ICSLP 94* (3): 1443-1446.

Pols L., SAM-partners (1992). Multilingual Synthesis Evaluation Methods. *Proceedings of ICSLP 92* (1): 181-184.

Portele T., Höfer F., Hess W. (1994). *A Mixed Inventory Structure for German Concatenative Syntesis.* University of Bonn. <ftp://asl1.ikp.uni-bonn.de/pub/vm41/tpnpal94.ps.gz>.

Portele T., Krämer J. (1996). Adapting a TTS System to a Reading Machine for the Blind. *Proceedings of ICSLP 96* (1).

Portele T., Steffan B., Preuss R., Hess W. (1991). German Text-to-Speech Synthesis by Concatenation of Non-Parametric Units. *Proceedings of Eurospeech 91* (1): 317-320.

Portele T., Steffan B., Preuss R., Sendlmeier W., Hess W. (1992). HADIFIX - A Speech Synthesis System for German. *Proceedings of ICSLP 92* (2): 1227-1230.

Rabiner L., Shafer R. (1978). *Digital Processing of Speech Signals*, Prentice-Hall.

Rahim M., Goodyear C., Kleijn B., Schroeter J., Sondhi M. (1993). On the Use of Neural Networks in Articulatory Speech Synthesis. *Journal of the Acoustical Society of America, JASA* vol. 93 (2): 1109-1121.

Renzepopoulos P., Kokkinakis G. (1992). Multilingual Phoneme to Grapheme Conversion System Based on HMM. *Proceedings of ICSLP 92* (2): 1191-1194.

Rossing T. (1990). *The Science of Sound*. Addison-Wesley.

Rutledge J., Cummings K., Lambert D., Clements M. (1995). Synthesized Styled Speech Using the KLATT Synthesizer. *Proceedings of ICASSP 95*: 648-651.

Sagisaga Y. (1990). Speech Synthesis from Text. *IEEE Communications Magazine*, vol. 28 1, pp. 35-41, 55.

Salmensaari O. (1989). *Puhesyntetisaattoritesti (Speech Synthesizer Test)*. HUT Acoustics Laboratory. Unpublished report. Espoo 1.12.1989.

Santen J. (1993). Timing in Text-to-Speech Systems. *Proceedings of Eurospeech 93* (2): 1397-1404.

Santen J., Sproat R., Olive J., Hirschberg J. (editors) (1997). *Progress in Speech Synthesis*, Springer-Verlag New York Inc. (Includes CD-ROM).

Scherer K. (1996). Adding the Affective Dimension: A New Look in Speech Analysis and Synthesis. *Proceedings of ICSLP 96* (3).

Schroeder M. (1993). A Brief History of Synthetic Speech. *Speech Communication* vol. 13, pp. 231-237.

Scordilis M., Gowdy J. (1989). Neural Network Based Generation of Fundamental Frequency Contours. *Proceedings of ICASSP 89* (1): 219-222.

Shiga Y., Hara Y., Nitta T. (1994). A Novel Segment-Concatenation Algorith for a Cepstrum-Based Synthesizer. *Proceedings of ICSLP 94* (4): 1783-1786.

SoftVoice, Inc. Homepage (1997). <http://www.text2speech.com/>.

Spiegel M. (1993). Using the ORATOR Synthesizer for a Public Reverse-Directory Service: Design, Lessons, and Recommendations. *Proceedings of Eurospeech 93* (3): 1897-1900.

Sproat R. (1996). Multilingual Text Analysis for Text-to-Speech Synthesis. *Proceedings of ICSLP 96* (3).

Sproat R., Taylor P., Tanenblatt M., Isard A. (1997). A Markup Language for Text-to-Speech Synthesis. *Proceedings of Eurospeech 97*.

SVOX Text-to-Speech Synthesis Homepage (1997). <http://www.tik.ee.ethz.ch/cgi-bin/w3svox>.

Tatham M., Lewis E. (1992a). Prosodic Assignment in SPRUCE Text-to-Speech Synthesis. *Proceedings of Institute of Acoustics*, vol. 14, Part 6 (1992): 447-454.

Tatham M., Lewis E. (1992b). Prosodics in a Syllable-Based Text-to-Speech Synthesis System. *Proceedings of ICSLP92* (2): 1179-1182.

Tatham M., Lewis E. (1995). Naturalness in a High-Level Synthetic Speech System. *Proceedings of Eurospeech 95* (3): 1815-1818.

Tatham M., Lewis E. (1996). Improving Text-to-Speech Synthesis. *Proceedings of ICSLP 96* (3).

Tatham M., Morton K., (1972). /p/ and /pp/ in Finnish: Duration of the Voiceless Phase in Intervocalic Context. *Occasional Papers* No 13/1972, Language Centre, University of Essex. <http://wrangler.essex.ac.uk/speech/archive/>.

Taylor P., Isard A. (1997). SSML: A Speech Synthesis Markup Language. *Speech Communication* vol. 21: 123-133.

Telia Promotor Home Page (1998). <http://www.infovox.se>.

Valbret H., Moulines E., Tubach J. (1991). Voice Transformation Using PSOLA Techique. *Proceedings of Eurospeech 91* (1): 345-348.

Veldhuis R., Bogaert I., Lous N. (1995). Two-Mass Models for Speech Synthesis. *Proceedings of Eurospeech 95* (3): 1853-1856.

Waters K., Levergood T. (1993). DECface: An Automatic Lip-Synchronization Algorithm for Synthetic Faces. *DEC Technical Report Series*, Cambridge Research Laboratory, CRL 93/4. <http://www.crl.research.digital.com/projects/facial/facialdoc.html>.

Witten I. (1982). *Principles of Computer Speech*, Academic Press Inc.