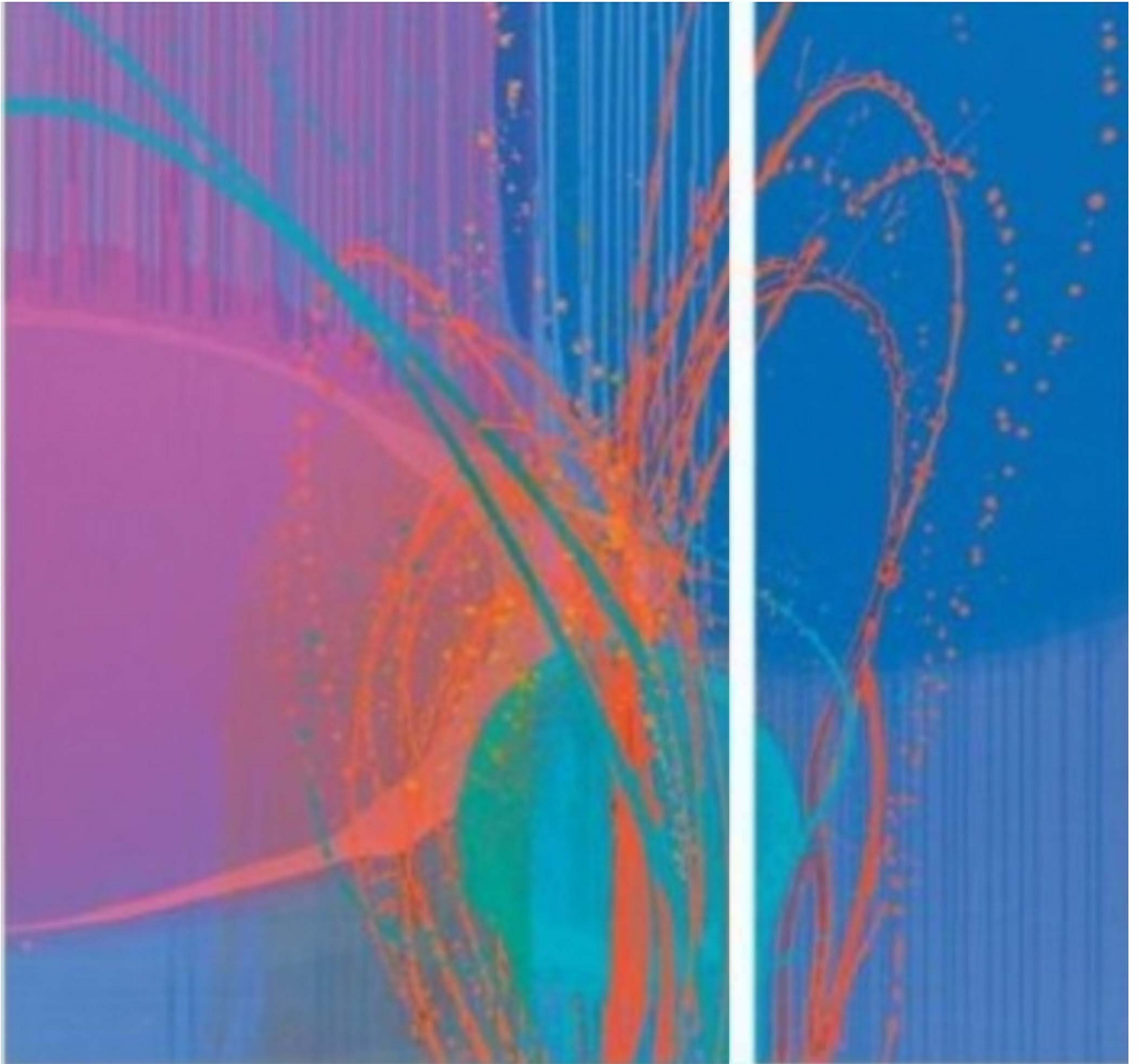




Text-to-Speech Synthesis

Paul Taylor

CAMBRIDGE

Text-to-Speech Synthesis

Text-to-Speech Synthesis provides a complete, end-to-end account of the process of generating speech by computer. Giving an in-depth explanation of all aspects of current speech synthesis technology, it assumes no specialised prior knowledge.

Introductory chapters on linguistics, phonetics, signal processing and speech signals lay the foundation, with subsequent material explaining how this knowledge is put to use in building practical systems that generate speech. The very latest techniques such as unit selection, hidden-Markov-model synthesis and statistical text analysis are covered, and explanations of the more-traditional techniques such as format synthesis and synthesis by rule are also provided.

By weaving together the various strands of this multidisciplinary field, this book is designed for graduate students in electrical engineering, computer science and linguistics. It is also an ideal reference for existing practitioners in the fields of human communication interaction and telephony.

PAUL TAYLOR is a Visiting Lecturer in Engineering at the University of Cambridge and Chief Executive Officer of Phonetic Arts. He received his Ph.D. in Linguistics from Edinburgh University, where he went on to serve as a lecturer and Director in the Centre for Speech Technology Research. He was also a co-founder and Chief Technical Officer at Rhetorical Systems.

Text-to-Speech Synthesis

PAUL TAYLOR
University of Cambridge



CAMBRIDGE
UNIVERSITY PRESS

Cambridge University Press
978-0-521-89927-7 - Text-to-Speech Synthesis
Paul Taylor
Frontmatter
[More information](#)

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo, Delhi
Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521899277

© Cambridge University Press 2009

This publication is in copyright. Subject to statutory exception
and to the provisions of relevant collective licensing agreements,
no reproduction of any part may take place without
the written permission of Cambridge University Press.

First published 2009

Printed in the United Kingdom at the University Press, Cambridge

A catalogue record for this publication is available from the British Library

ISBN 978-0-521-89927-7 hardback

Cambridge University Press has no responsibility for the persistence or
accuracy of URLs for external or third-party internet websites referred to
in this publication, and does not guarantee that any content on such
websites is, or will remain, accurate or appropriate.

Foreward

Speech processing technology has been a mainstream area of research for more than 50 years. The ultimate goal of speech research is to build systems that mimic (or potentially surpass) human capabilities in understanding, generating and coding speech for a range of human-to-human and human-to-machine interactions.

In the area of speech coding a great deal of success has been achieved in creating systems that significantly reduce the overall bit rate of the speech signal (from order of 100 kilobits per second, to rates on the order of 8 kilobits per second or less), while maintaining speech intelligibility and quality at levels appropriate for the intended applications. The heart of the modern cellular industry is the 8 kilobit per second speech coder, embedded in VLSI logic on the more than 2 billion cellphones in use worldwide at the end of 2007.

In the area of speech recognition and understanding by machines, steady progress has enabled systems to become part of everyday life in the form of call centers for the airlines, financial, medical and banking industries, help desks for large businesses, form and report generation for the legal and medical communities, and dictation machines that enable individuals to enter text into machines without having to explicitly type the text. Such speech recognition systems were made available to the general public as long as 15 years ago (in 1992 AT&T introduced the Voice Recognition Call Processing system which automated operator-assisted calls, handling more than 1.2 billion requests each year with error rates below 0.5% penetrated almost every major industry since that time. Simple speech understanding systems have also been introduced into the marketplace and have had varying degrees of success for help desks (e.g., the How May I Help You system introduced by AT&T for customer care applications) and for stock trading applications (IBM system), among others.

It has been the area of speech generation that has been the hardest speech technology area to obtain any viable degree of success. For more than 50 years researchers have struggled with the problem of trying to mimic the physical processes of speech generation via articulatory models of the human vocal tract, or via terminal analog synthesis models of the time-varying spectral and temporal properties of speech. In spite of the best efforts of some outstanding speech researchers, the quality of synthetic speech generated by machine was unnatural most of the time and has been unacceptable for human use in most real world applications. In the late 1970s the idea of generating speech by concatenating basic speech units (in most cases diphone units which represented pieces of pairs of phonemes) was investigated and shown to be practical once researchers learned how to reliably excise diphones from human speech. After more than a decade of investigation as to how to optimally concatenate diphones, the resulting synthetic speech was often highly intelligible (a big improvement over earlier systems) but regrettably remained highly unnatural. Hence concatenative speech synthesis systems remained lab curiosities but were not employed in real world applications such as reading email, user interactions in dialogue systems, etc. The really big breakthrough in speech synthesis came in the late 1980s when Yoshinori Sagisaka at ATR in Japan made the leap from single diphone tokens as the basic unit set for speech synthesis, to multiple di-

phone tokens, extracted from carefully designed and read speech databases. Sagisaka realized that in the limiting case, where you had thousands of tokens of each possible diphone of the English language, you could literally concatenate the “correct” sequence of diphones and produce natural sounding human speech. The new problem that arose was deciding exactly which of the thousands of diphones should be used at each diphone position in the speech being generated. History has shown that, like most large scale computing problems, there are solutions that make the search for the optimum sequence of diphones (from a large, virtually infinite database) possible in reasonable time and memory. The rest is now history as a new generation of speech researchers investigated virtually every aspect of the so-called “unit selection” method of concatenative speech synthesis, showing that high quality (both intelligibility and naturalness) synthetic speech could be obtained from such systems for virtually any task application.

Once the problem of generating natural sounding speech from a sequence of diphones was solved (in the sense that a practical demonstration of the feasibility of such high quality synthesis was made with unit selection synthesis systems), the remaining long-standing problem was the conversion from ordinary printed text to the proper sequence of diphones, along with associated prosodic information about sound duration, loudness, emphasis, pitch, pauses, and other so-called suprasegmental aspects of speech. The problem of converting from text to a complete linguistic description of associated sound was one that has been studied almost as long as synthesis itself; much progress had been made in almost every aspect of the linguistic description of speech as in the acoustic generation of high quality sounds.

It is the success of unit selection speech synthesis systems that has motivated the research of Paul Taylor, the author of this book on text-to-speech synthesis systems. Paul Taylor has been in the thick of the research in speech synthesis systems for more than 15 years, having worked at ATR in Japan on the CHATR synthesizer (the system that actually demonstrated near perfect speech quality on some subset of the sentences that were input), at the Centre for Speech Technology research at the University of Edinburgh on the Festival synthesis system, and as Chief Technical Officer of Rhetorical Systems, also in Edinburgh.

Based on decades of research and the extraordinary progress over the past decade, Taylor has put together a book which attempts to tie it all together and to document and explain the processes involved in a complete text-to-speech synthesis system. The first nine chapters of the book address the problem of converting printed text to a sequence of sound units (which characterize the acoustic properties of the resulting synthetic sentence), and an accompanying description of the associated prosody which is most appropriate for the sentence being spoken. The remaining eight chapters (not including the conclusion) provide a review of the associated signal processing techniques for representing speech units and for seamlessly tying them together to form intelligible and natural speech sounds. This is followed by a discussion of the three generations of synthesis methods, namely articulatory and terminal analog synthesis methods, simple concatenative methods using a single representation for each diphone unit, and the unit selection method based on multiple representations for each diphone unit. There is a single chapter devoted to a promising new synthesis approach, namely a statistical method based on the popular Hidden Markov Model

(HMM) formulation used in speech recognition systems.

According to the author, “Speech synthesis has progressed remarkably in recent years, and it is no longer the case that state-of-the-art systems sound overtly mechanical and robotic”. Although this statement is true, there remains a great deal more to be accomplished before speech synthesis systems are indistinguishable from a human speaker. Perhaps the most glaring need is expressive synthesis that not only imparts the message corresponding to the printed text input, but imparts the emotion associated with the way a human might speak the same sentence in the context of a dialogue with another human being. We are still a long way from such emotional or expressive speech synthesis systems.

This book is a wonderful addition to the literature in speech processing and will be a must-read for anyone wanting to understand the blossoming field of text-to-speech synthesis.

Lawrence Rabiner

August 2007

1 INTRODUCTION

This is a book about getting computers to read out loud. It is therefore about three things: the process of reading, the process of speaking, and the issues involved in getting computers (as opposed to humans) to do this. This field of study is known both as **speech synthesis**, that is the “synthetic” (computer) generation of speech, and **text-to-speech** or **TTS**; the process of converting written text into speech. As such it compliments other language technologies such as **speech recognition**, which aims to convert speech into text, **machine translation** which converts writing or speech in one language into writing or speech in another.

I am assuming that most readers have heard some synthetic speech in their life. We experience this in a number of situations; some telephone information systems have automated speech response, speech synthesis is often used as an aid to the disabled, and Professor Stephen Hawking has probably contributed more than anyone else to the direct exposure of (one particular type of) synthetic speech. The *idea* of artificially generated speech has of course been around for a long time - hardly any science fiction film is complete without a talking computer of some sort. In fact science fiction has had an interesting effect on the field and our impressions of it. Sometimes (less technically aware) people believe that perfect speech synthesis exists because they “heard it on Star Trek”¹. Often makers of science fiction films fake the synthesis by using an actor, although usually some processing is added to the voice to make it sound “computerised”. Some actually use real speech synthesis systems, but interestingly these are usually not state of the art systems, as these sound too natural, and may mislead the viewer². One of the films that actually makes a genuine attempt to predict how synthetic voices will sound is the computer HAL in the film 2001: A Space Odyssey [265]. The fact that this computer spoke with a calm and near humanlike voice gave rise to the sense of genuine intelligence in the machine. While many parts of this film were wide of the mark (especially, the ability of HAL to understand, rather than just recognise human speech), the makers of the film just about got it right in predicting how good computer voices would be in the year in question.

¹ Younger readers please substitute the in-vogue science-fiction series of the day

² In much the same way, when someone types the wrong password on a computer, the screen starts flashing and saying “access denied”. Some even go so far as to have a siren sounding. Those of us who use computers know this never happens, but in a sense we go along with the exaggeration as it adds to the drama.

Speech synthesis has progressed remarkably in recent years, and it is no longer the case that state-of-the-art systems sound overtly mechanical and robotic. That said, it is normally fairly easy to tell that it is a computer talking rather than a human, and so substantial progress is still to be made. When assessing a computer's ability to speak, one fluctuates between two judgments. On the one hand, it is tempting to paraphrase Dr Johnson's famous remark [61] "Sir, a talking computer is like a dog's walking on his hind legs. It is not done well; but you are surprised to find it done at all." And indeed, even as an experienced text-to-speech researcher who has listened to more synthetic speech than could be healthy in one life, I find that sometimes I am genuinely surprised and thrilled in a naive way that here we have a talking computer: "like wow! it talks!". On the other hand it is also possible to have the impression that computers are quite dreadful at the job of speaking; they make frequent mistakes, drone on, and just sound plain *wrong* in many cases. These impressions are all part of the mysteries and complexities of speech.

1.1 WHAT ARE TEXT-TO-SPEECH SYSTEMS FOR?

Text-to-speech systems have an enormous range of applications. Their first real use was in reading systems for the blind, where a system would read some text from a book and convert it into speech. These early systems of course sounded very mechanical, but their adoption by blind people was hardly surprising as the other options of reading braille or having a real person do the reading were often not possible. Today, quite sophisticated systems exist that facilitate human computer interaction for the blind, in which the TTS can help the user navigate around a windows system.

The mainstream adoption of TTS has been severely limited by its quality. Apart from users who have little choice (as in the case with blind people), people's reaction to old style TTS is not particularly positive. While people may be somewhat impressed and quite happy to listen to a few sentences, in general the novelty of this soon wears off. In recent years, the considerable advances in quality have changed the situation such that TTS systems are more common in a number of applications. Probably the main use of TTS today is in call-centre automation, where a user calls to pay an electricity bill or book some travel and conducts the entire transaction through an automatic dialogue system. Beyond this, TTS systems have been used for reading news stories, weather reports, travel directions and a wide variety of other applications.

While this book concentrates on the practical, engineering aspects of text-to-speech, it is worth commenting that research in this field has contributed an enormous amount to our general understanding of language. Often this has been in the form of "negative" evidence, meaning that when a theory thought to be true was implemented in a TTS system it was shown to be false; in fact as we shall see, many linguistic theories have fallen when rigorously tested in speech systems. More positively, TTS systems have made a good testing grounds for many models and theories, and TTS systems are certainly interesting in their own terms, without reference to application or use.

1.2 WHAT SHOULD THE GOALS OF TEXT-TO-SPEECH SYSTEM DEVELOPMENT BE?

One can legitimately ask, regardless of what application we want a talking computer for, is it really necessary that the quality needs to be high and that the voice needs to sound like a human? Wouldn't a mechanical sounding voice suffice? Experience has shown that people are in fact very sensitive, not just to the words that are spoken, but to the way they are spoken. After only a short while, most people find highly mechanical voices irritating and discomforting to listen to. Furthermore tests have shown that user satisfaction increases dramatically the more "natural" sounding the voice is. Experience (and particularly commercial experience) shows that users clearly want natural sounding (that is human-like) systems.

Hence our goals in building a computer system capable of speaking are to first build a system that clearly gets across the message, and secondly does this using a human-like voice. Within the research community, these goals are referred to as **intelligibility** and **naturalness**.

A further goal is that the system should be able to take any written input; that is, if we build an English text-to-speech system, it should be capable of reading any English sentence given to it. With this in mind, it is worth making a few distinctions about computer speech in general. It is of course possible to simply record some speech, store it on a computer and play it back. We do this all the time; our answer machine replays a message we have recorded, the radio plays interviews that were previously recorded and so on. This is of course simply a process of playing back what was originally recorded. The idea behind text-to-speech is to "play back" messages that weren't originally recorded. One step away from simple playback is to record a number of common words or phrases and recombine them, and this technique is frequently used in telephone dialogue services. Sometimes the result is acceptable, sometimes not, as often the artificially joined speech sounded stilted and jumpy. This allows a certain degree of flexibility, but falls short of open ended flexibility. Text-to-speech on the other hand, has the goal of being able to speak anything, regardless of whether the desired message was originally spoken or not.

As we shall see in Chapter 13, there are a number of techniques for actually generating the speech. These generally fall into two camps, which we can call bottom-up and concatenative. In the bottom-up approach, we generate a speech signal "from scratch", using our knowledge of how the speech production system works. We artificially create a basic signal and then modify it, much the same way that the larynx produces a basic signal which is then modified by the mouth in real human speech. In the concatenative approach, there is no bottom-up signal creation per se; rather we record some real speech, cut this up into small pieces, and then recombine these to form "new" speech. Sometimes one hears the comment that concatenative techniques aren't "real" speech synthesis in that we aren't generating the signal from scratch. This point may or may not be relevant, but it turns out that at present concatenative techniques far out perform other techniques, and for this reason concatenative techniques currently dominate.

1.3 THE ENGINEERING APPROACH

In this book, we take what is known as an **engineering approach** to the text-to-speech problem. The term “engineering” is often used to mean that systems are simply bolted together, with no underlying theory or methodology. Engineering is of course much more than this, and it should be clear that great feats of engineering, such as the Brooklyn bridge were not simply the result of some engineers waking up one morning and banging some bolts together. So by “engineering”, we mean that we are tackling this problem in the best traditions of other engineering; these include, working with the materials available and building a practical system that doesn’t for instance take days to produce a single sentence. Furthermore, we don’t use the term engineering to mean that this field is only relevant or accessible to those with (traditional) engineering backgrounds or education. As we explain below, TTS is a field relevant to people from many different backgrounds.

One point of note is that we can contrast the engineering approach with the scientific approach. Our task is to build the best possible text-to-speech system, and in doing so, we will use any model, mathematics, data, theory or tool that serves our purpose. Our main job is to build an *artefact* and we will use any means possible to do so. All artefact creation can be called engineering, but *good* engineering involves more: often we wish to make good use of our resources (we don’t want to use a hammer to crack a nut); we also in general want to base our system on solid principles. This is for several reasons. First, using solid (say mathematical) principles assures us we are on well tested ground; we can trust these principles and don’t have to experimentally verify every step we do. Second, we are of course not building the last ever text-to-speech system; our system is one step in a continual development; by basing our system on solid principles we hope to help others to improve and build on our work. Finally, using solid principles has the advantage of helping us diagnose the system, for instance to help us find why some components do perhaps better than expected, and allow the principles of which these components are based to be used for other problems.

Speech synthesis has also been approached from a more scientific aspect. Researchers who pursue this approach are not interested in building systems for their own sake, but rather as models which will shine light on human speech and language abilities. As such, the goals are different, and for example, it is important in this approach to use techniques which are at least plausible possibilities for how humans would handle this task. A good example of the difference is in the concatenative waveform techniques which we will use predominantly; recording large numbers of audio waveforms, chopping them up and gluing them back together can produce very high quality speech. It is of course absurd to think that this is how humans do it. We bring this point up because speech synthesis is often used (or was certainly used in the past) as a testing ground for many theories of speech and language. As a leading proponent of the scientific viewpoint states, so long as the two approaches are not confused, no harm should arise (Huckvale [225]).

1.4 OVERVIEW OF THE BOOK

I must confess to generally hating sections entitled “how to read this book” and so on. I feel that if I bought it, I should be able to read it any way I damn well please! Nevertheless, I feel some guidelines may be useful.

This book is what one might call an *extended text book*. A normal text book has the job of explaining a field or subject to outsiders and this book certainly has that goal. I qualify this by using the term “extended” for two reasons. Firstly the book contains some original work, and is not simply a summary, collection or retelling of existing ideas. Secondly, the book aims to take the reader right up to the current state of the art. In reality this can never be fully achieved, but the book is genuinely intended to be an “all you ever need to know”. More modestly, it can be thought of as “all that I know and can explain”. In other words: this is it: I certainly couldn’t write a second book which dealt with more advanced topics.

Despite these original sections, the book is certainly not a monograph. This point is worth reinforcing: because of my personal involvement in many aspects of TTS research over the last 15 years, and specifically because of my involvement in the development of many well-known TTS systems, including Chatr [53], Festival [55], and rVoice, many friends and colleagues have asked me whether this is a book about those systems or the techniques behind those systems. Let me clearly state that this is not the case; *Text-to-Speech Synthesis* is not a system book that describes one particular system; rather I aim for a general account that describes current techniques without reference to any particular single system or theory.

1.4.1 Viewpoints within the book

That said, this book aims to provide a single, coherent picture of text-to-speech, rather than simply a list of available techniques. And while not being a book centred on any one system, it is certainly heavily influenced by the general philosophy that I have been using (and evolving) over the past years, and I think it is proper at this stage to say something about what this philosophy is and how it may differ from other views. In the broadest sense, I adopt what is probably the current mainstream view in TTS, that this is an engineering problem, that should be approached with the aim of producing the best possible system, rather than with the aim of investigating any particular linguistic or other theory.

Within the engineering view, I again have taken a more specialist view in posing the text-to-speech problem as one where we have a single integrated text analysis component followed by a single integrated speech synthesis component. I have called this the **common form model** (this and other models are explained in Chapter 3.) While the common form model differs significantly from the usual “pipelined” models, most work that has been carried out in one framework can be used in the other without too much difficulty.

In addition to this, there are many parts which can be considered as original (at least to my knowledge) and in this sense, the book diverges from being a pure text book at these points.

Specifically, these parts are

1. the common form model itself
2. the formulation of text analysis as a decoding problem,
3. the idea that text analysis should be seen as a semiotic classification and verbalisation problem,
4. the model of reading aloud,
5. the general unit selection framework.
6. the view that prosody is composed of the functionally separate systems of affective, augmentative and suprasegmental prosody.

With regard to this last topic, I should point out that my views on prosody diverge considerably from the mainstream. My view is that mainstream linguistics, and as a consequence much of speech technology has simply got this area of language badly wrong. There is a vast, confusing and usually contradictory literature on prosody, and it has bothered me for years why several contradictory competing theories (of say intonation) exist, why no-one has been able to make use of prosody in speech recognition and understanding systems, and why all prosodic models that I have tested fall far short of the results their creators say we should expect. This has led me to propose a completely new model of prosody, which is explained in Chapters 3 and 6.

1.4.2 Readers' backgrounds

This book is intended for both an academic and commercial audience. Text-to-speech or speech synthesis does not fall neatly into any one traditional academic discipline, and so the level and amount of background knowledge will vary greatly depending on a particular reader's background. Most TTS researchers I know come from an electrical engineering, computer science or linguistics background. I have aimed the book at being directly accessible to readers with these backgrounds, but the book should in general be accessible to those from other fields.

I assume that all readers are computer literate and have some experience in programming. To this extent, concepts such as algorithm, variable, loop and so on are assumed. Some areas of TTS are mathematical, and here I have assumed that the entry level is that of an advanced high school or first year university course in maths. While some of the mathematical concepts are quite advanced, these are explained in full starting with the entry level knowledge. For those readers with little mathematical knowledge (or inclination!) don't worry; many areas of TTS do not require much maths. Even for those areas which do, I believe a significant understanding can still be achieved by reading about the general principles, studying the graphs, and above all, trying the algorithms in practice. Digital filters can seem like a complicated and abstract subject to many; but I have seen few people fail to grasp its basics when give the opportunity to play around with filters in a GUI package.

My commercial experience made it clear that it was difficult to find software developers with any knowledge of TTS. It was seen as too specialist a topic and even for those who were interested

in the field, there was no satisfactory introduction. I hope this book will help solve this problem, and I have aimed it at being accessible to software engineers (regardless of academic background) who wish to learn more about this area. While this book does not give a step by step recipe for building a TTS system, it does go significantly beyond the theory, and tries to impart a feel for the subject and tries to pass on some of the “folklore” that is necessary for successful development. I believe the book covers enough ground such that a good software engineer should not have too much difficulty with implementation.

1.4.3 Background and specialist sections

The book contains a significant amount of background material. This is included for two reasons. Firstly, as just explained, I wanted to make the book seem complete to any reader outside the field of speech technology. I believe it is to the readers’ benefit to have introductory sections on phonology or signal processing in a single book, than have to resort to the alternative of pointing the reader to other works.

There is a second reason however, which is that I believe that the traditional approach to explaining TTS is too disjointed. Of course TTS draws upon many other disciplines, but the differences between these to me are often overstated. Too often, it is believed that only “an engineer” (that is someone who has a degree in engineering) can understand the signal processing, only “a linguist” (again, degree in linguistics) can understand the phonetics and so on. I believe that this view is very unhelpful; it is ridiculous to believe that someone with the ability to master signal processing isn’t able to understand phonetics and vice versa. I have attempted to bridge these gaps by providing a significant amount of background material, but in doing so have tried to make this firstly genuinely accessible and secondly focused on the area of text to speech. I have therefore covered topics found in introductory texts in engineering and linguistics, but tried to do so in a novel way that makes the subject matter more accessible to readers with different backgrounds. It is difficult to judge potential readers’ exposure to the fundamentals of probability as this is now taught quite extensively. For this reason, I have assumed a knowledge of this in the body of the book, and have included an reference section on this topic in the appendix.

The book is written in English and mostly uses English examples. I decided to write the book and focus on one language rather than make constant references to the specific techniques or variations that would be required for every language of interest to us. Many newcomers (and indeed many in the field who don’t subscribe to the data-driven view) believe that the differences between languages are quite substantial, and that what works for English, is unlikely to work for French, Finnish, or Chinese. While languages obviously do differ, in today’s modern synthesisers these differences can nearly all be modelled by training and using appropriate data; the same core engine suffices in all cases. Hence concentrating on English does not mean we are building a system that will only work on one language.

2

COMMUNICATION AND LANGUAGE

Before delving into the details of how to perform text-to-speech conversion, we will first examine some of the fundamentals of communication in general. This chapter looks at the various ways in which people communicate and how communication varies depending on the situation and the means which are used. From this we can develop a general model which will then help us specify the text-to-speech problem more exactly in the following chapter.

2.1 TYPES OF COMMUNICATION

We experience the world through our senses and we can think of this as a process of gaining **information**. We share this ability with most other animals; if an animal hears running water it can infer that there is a stream nearby; if it sees a ripe fruit it can infer there is food available. This ability to extract information from the world via the senses is a great advantage in the survival of any species. Animals can however cause information to be created: many animals make noises, such as barks or roars, or gestures such as flapping or head nodding, which are intended to be interpreted by other animals. We call the process of *deliberate creation* of information with the *intention that it be interpreted* as **communication**.

The prerequisites for communication are an ability to create information in one being, an ability to transmit this information, and an ability to perceive the created information by another being. All three of these pre-requisites strongly influence the nature of communication; for example, animals that live in darkness or are blind would be unlikely to use a visual system. But despite these restrictions, it is clear that there are still many possible ways to make use of the possibilities of creation, medium and perception to communicate. We will now examine the three fundamental communication techniques that form the basis for human communication.

2.1.1 Affective communication

The most basic and common type of communication is **affective** communication, where we express a primary emotional state with external means. A good example of this is the expression of pain, where we might let out a yell or cry upon hurting ourselves. A defining characteristic of



Figure 2.1 (British) road sign, indicating a slippery road and high chance of skidding.

this type of communication is that the intensity of the external form is clearly a function of the intensity of feeling; the more intense the pain the louder the yell. Other primary mental states such as happiness, anger and sadness can be expressed in this way. This type of communication is one that is common to most higher animals. While the *ability* to express these affective states is common among animals, the precise means by which these are expressed is not universal or always obvious. A high pitched yell, squeal or cry often means pain, but it by no means obvious that a dog's wagging tail or a cat's purring are expressions of happiness.

2.1.2 Iconic communication

Though fundamental and powerful, affective communication is severely limited in the range of things it can be used to express. Happiness can readily be conveyed, but other simple mental states such as hunger or tiredness are significantly more difficult to convey. To express more complex messages, we can make use of a second communication mechanism known as **iconic** communication. An iconic system is one where the created **form** of the communication somehow resembles what the intended **meaning**. We use the term "form" here in a technical sense that allows us to discuss the common properties of communication systems: in acoustic communication, "form" can be thought of as a type of sound; in visual communication form might be types of hand signals, facial expressions and so on. For example, it is common to communicate tiredness iconically by the "sleeping gesture", where someone closes her eyes, puts her hands together and places her head sideways on her hands. The person isn't really asleep - she using a gesture that (crudely) mimics sleep to indicate tiredness. Another good example of iconic communication is the road sign shown in Figure 2.1.1. In this case, the form is the diagram, and the meaning is *slippery road*, and the fact that the form visually resembles what can happen when a road is slippery means that this communication is iconic. Note that just with the sleep example, the form isn't a particularly accurate picture of a car, road, skid or so on; the idea is to communicate the essence

of the meaning and little else.

Anecdotally, we sometimes think of pre-human communication also working like this, and in such a system, the presence of a saber tooth tiger might be communicated by imitating its growl, or by acting like a tiger and so on. Such systems have a certain advantage of transparency, in that if the intended recipient does not know what you mean, a certain amount of mimicry or acting may get the point across. In essence, this is just like the road sign example, in that the form and meaning have a certain direct connection. When travelling in a country where we don't speak the language, we often resort to miming some action with the hope that the meaning will be conveyed.

Iconic systems though have several drawbacks. One is that while it may be easy to imitate a tiger, it is less easy to imitate more abstract notions such as "nervous" or "idea". More importantly though, iconic systems can suffer from a lack of precision: when a first caveman imitates the tiger, a second caveman may not get this reference exactly - he might be sitting there thinking, "well, it could be a tiger, or perhaps a lion. Or maybe a large dog". And by the time the first caveman has mimed his way through this and the action of "creeping up behind", both have probably departed to the great cave in the sky. While useful and important, iconic communication clearly has its limits.

2.1.3 Symbolic communication

In contrast to iconic and affective communication, we also have **symbolic** communication in which we give up the idea that the form must indicate the meaning. Rather we use a series of correspondences between form and meaning, in which the relationship is not direct. In symbolic systems, a tiger might be indicated by waving the left hand, a lion by waving the right. There is no reason why these forms should indicate what they do, it is merely a matter of convention. The advantage is that it is easier to devise a system where the forms are clear and distinct from one another - less confusion will arise. The disadvantage is that the fact that left-arm-wave means tiger, and right-arm-wave means lion has to be *learned*, if you don't know, seeing the movement in isolation won't give clue to the meaning. Despite the disadvantage in needing to learn, using conventions rather than icons can be hugely advantageous in that the conventions can be relatively brief; one noise or gesture may represent the tiger and this need not be acted out carefully each time. This brevity and clarity of form leads to a swiftness and precision seldom possible with iconic communication.

Once a convention-based communication system is used, it soon becomes clear that it is the notion of **contrast** in form that is the key to success. To put it another way, once the form no longer needs to resemble the meaning, the communication system gains benefit from making the forms as distinct from one another as possible. To show this point, consider the following experiment.

We grouped eight subjects into pairs, and in isolation, asked each pair to design a communication system based on colour cards. The premise was that the pair were in a noisy pub, one individual was at the bar while the other was sitting down at a table. We said there were four basic concepts to communicate: "I would like a drink of water", "I would like some food", "I would like a beer" and "I would like to listen to some music". Each pair was given a set of 100 different



Figure 2.2 With three symbols in isolation only three meanings are possible. When we combine the symbols in sequences the number of meanings exponentially increases, while the ability to uniquely perceive the symbols is unaffected.

coloured cards arranged across the spectrum, and asked to pick one card for each concept and memorise which colour was associated with which concept. In a simple test, it was clear that the pairs could remember the colours and their meanings, and could effectively communicate these simple concepts with this system. The real purpose of the experiment was different though. While three of the groups chose a blue colour for the concept “I want a drink of water”, for the other categories there was no commonality in the choice. So “I want a beer” had two pairs chose green, one yellow and one black. Interestingly most gave some justification as to why each colour was chosen (green means “yes, lets have another”) and so on; but it is clear that this justification was fairly superficial in that each group choose quite different colours. This demonstrates one of the main principles of symbolic communication system, namely that the relationship between the form of the symbol (in this case the colour) and its meaning is in general **arbitrary**. These form/meaning pairs must be learned, but as we shall see, in many cases the “cost” of learning the pairing to an individual is less than the “cost” of potential confusion and imprecision that can arise from having the form resemble the meaning. This simple study demonstrates another vital point. In each case, the pairs chose quite different colours; so group A chose a blue, a red, a yellow and a green, while group B chose a red, a yellow, a light blue and a black. In no case did a group choose say a green, a light blue, a medium blue and a dark blue for instance. In other words, the groups chose a set of colours that had clear contrast with each other. In a way this makes obvious sense; picking colours that are significantly different from one another lessens the possibility that they will be confused, but it does demonstrate a crucial point in symbolic systems, namely that the systems are based on the notion of contrast between forms rather than the absolute “value” of each form.

2.1.4 Combinations of symbols

This simple communication system works well because we only have four concepts we wish to convey. What happens when we need to communicate more concepts? We can see that the simple

colour contrast will eventually reach its limits. If say 20 concepts are required, it soon becomes a difficult task to associate each with just one colour. Setting aside the issue of whether the individuals can remember the colours (let us assume they have a crib sheet to remind them), a major problem arises because as more concepts get added, a new colour is associated with each one, and as the range of colours is fixed, each new added concept increases the likelihood that it will get confused with another concept which has a similar colour. While it may be possible for an individual to distinguish 20 colours, asking them to do so for 1000 colours seems unlikely to meet with success. We therefore need a separate solution. The obvious thing to do is use **combinations** of colour cards. For instance, we could associate the colour black with the concept *not*, pink with *large*, and then use these and one of the original colours to indicate a much larger range of concepts (so holding a black card, pink card and blue card would say “I don’t want a large drink of water”). Note that this is different from holding a single card that is a mixed combination of black pink and blue (which might be a murky brown). The reason why the combination system works is that each colour card on its own is clear, distinct and unconfusable, but by combining them, we can generate a large number of concepts from a small number of form/meaning pairs. The power of combination is considerable; in the limit, the number of complex forms exponentially multiplies, so that if we have N simple forms and M instances of them, we have a total of N^M complex forms; so 10 forms combined in 10 different ways would produce over 10 billion different complex forms.

We call communication systems that use a fixed number of simple forms which can be combined into a much larger number of forms a **symbolic language**. Here we use the term “language” in both a very precise and general sense. Precise, in that it we use it only to mean this principle of combination, but general in the sense that all human languages make use of this principle in this sense, as are other systems, such as written numbers (we can create any number from a combination of the 10 digits), mathematics and even email addresses. Human language often goes under the description of **natural language** and we shall use that term here to distinguish it from the principle of languages in general. One of the main distinctive properties of natural language is the sheer scale of things that it can express. While we can talk of other animals having sophisticated communication systems (bee dancing etc), none has anything like the scale of natural language, where a person might know say 10,000 words, and be able to combine these into trillions and trillions of sentences.

2.1.5 Meaning, form and signal

In our discussion so far, we have been somewhat vague and a little informal about what we really mean by “meaning” and “form”. The study of meaning (often called semantics), is notoriously difficult, and has been the subject of debate since the earliest philosophy. It is an inherently tricky subject, because firstly it requires an understanding of the world (what the world contains, how the things in the world relate to one another and so on), and secondly, from a cognitive perspective, it is exceedingly difficult to peer inside someone’s brain and find out exactly how they represent the concept of “warm” or “drink” and so on. As we shall see, developing a precise model of

semantics does not seem to be necessary for text-to-speech, and so we shall largely bypass this topic in a formal sense.

Our notion of “form” also needs some elaboration. A proper understanding of this part of communication is vital for our purposes, but thankfully we can rely on an established framework to help us. The main issue when discussing form is to distinguish the general nature of the form from any specific instantiation it may have. Going back to our simple colour example; it is important to realise that when one individual holds up a red card, it isn’t the red card itself that is the important point, rather that red was used rather than the other choices of green, blue and yellow. So for example, a different red card could be used successfully, and probably any red card would do, as it is not the particular red card itself which is the form, but rather that the card has the property “red” and it is this property and its contrasting values that is the basis of the communication system. In some accounts, the term **message** is used instead of form.

We use the term **signal** to refer to the physical manifestation of the form. Here signal is again used in a specific technical sense: it can be thought of as the thing which is stored or the thing which is transmitted during communication. The idea of “red” (and the idea that it contrasts with other colours) is the form; the physical light waves that travel are the signal. The above example also demonstrates that the same form (red) can have a number of quite different signals, each dependent on the actual mode of transmission.

It is important to see that the relationship between meaning and form, and form and signal are quite separate. For instance, in our example we have associated the colours red, yellow, blue and green each with a concept. Imagine now the sender holds up an orange card. Most likely the receiver will see this signal and correctly interpret the physical light-waves into a form “orange”. After this however she is stuck because she can’t determine what concept is matched with this orange form. The signal/form relationship has worked well, but it has not been possible to determine the correct concept. This situation is similar to that when we read a new word - while we are quite capable of using our eyes to read the physical pattern from the page and create an idea of the letters in our head, if we don’t know the form/meaning mapping of the word, we can not understand it.

Finally, we use the term **channel** or **medium** to refer to the means by which the message is converted and transmitted as a signal. In this book, we deal mainly with the **spoken channel** and the **written channel**.

2.2 HUMAN COMMUNICATION

The above examination of communication in the abstract has of course been leading us to an explanation of how human communication operates. Taking spoken communication first, let us make a primary distinction between the two main aspects of this, which we shall call the **verbal** component and the **prosodic** component. We use the term “verbal” to mean the part of human communication to with **words**. In some books, the term “verbal” and “language” are synonymous, but because “language” has so many other, potentially confusable meanings, we will stick with the

terms “verbal” or “verbal component” throughout. The verbal component is a symbolic system in that we have a discrete set of words which can be arranged in sequences to form a very large number of sentences.

In contrast, the prosodic component (often just called **prosody**) is different in that it is not primarily symbolic. We use prosody for many purposes: to express emotion or surprise, to emphasise a word, to indicate where a sentence ends and so on. It is fundamentally different in that it is not purely discrete, does not generally have units which combine in sequences and so has a much more limited range of meanings that can be expressed. That said, prosody should not be taken as a poor relative of the verbal component; if we say “its wonderful” with words, but use a tone of voice that indicates disgust; the listener will guess we are being sarcastic. If the prosodic and verbal component disagree, the prosodic information wins. To a certain extent, these systems can be studied separately; in fact nearly all work in the field of linguistics is solely interested in the verbal component. The two do interact though; first in the obvious sense that they share the same signal (speech), but also in more complex ways in that prosody can be used to emphasise one particular word in a sentence.

2.2.1 Verbal communication

The symbolic model of the verbal component states that the principle unit of form is the **phoneme**. Phonemes operate in a contrastive system of sounds, in which a fixed number of phonemes can combine to a much larger number of **words** (a vocabulary of more than 10,000 words is common, but many millions of words are possible). Next, these words are taken as the basic units, and they further combine into **sentences**. Verbal language is therefore seen as a combination of two systems; one that makes words from phonemes, and another that makes sentences from words. As the word/sentence system has quite a large set of basic units to start with, the number of combined units is vast. The two systems have different properties with regard to meaning. In the first system, the basic building blocks, the phonemes, have no direct (arbitrary, iconic or other) meaning, they are simply building blocks. These are combined to form words, and these do have meaning. The relation between a word’s meaning and the phonemes which comprise it is completely arbitrary and so the meaning and pronunciation of each word has to be learned. The second system operates differently, in that the basic building blocks (the words) do have inherent meaning, and the meaning of the sentence is related to the meaning of the components. Exactly how the sentence meaning relates to the word meaning is still an issue of considerable debate; one the one hand there is considerable evidence that it is a function, in the sense that if we have JOHN KICKED MARY¹ we can substitute different people for JOHN and MARY (e.g. ANN KICKED GEORGE), or different verbs (e.g. JOHN KISSED MARY) and predict the sentence meaning from the parts. On the other hand, there is also considerable evidence that sentence meaning is not a simple function, in that JOHN KICKED THE BUCKET means John has died, and has nothing to do

¹ From here we will use a small caps typeface when we refer to form/message such as JOHN and a courier typeface for the written signal, e.g. j ohn.

with kicking or buckets.

From this basic model, we can now explore some of the fundamental features of language. In doing so, we follow Hockett [214], who stated the principle defining properties of verbal language. Among these are:

Arbitrary

Phonemes do not have inherent meaning but words do; a word is in effect a form/meaning pair in which the correspondence between the meaning and the form (ie. the sequence of phonemes) is arbitrary. Because of this, there is no way to guess the meaning of a word if it is not learned, so as a reader, you won't know the meaning of the word PANDICULATION unless you are informed of its meaning. From this, it follows that words of similar form can have complete different meanings, so BIT and PIT mean quite different things despite only having a slight sound difference, whereas DOG and WOLF are similar in meaning, but share no sounds.

Duality

The duality principle states that verbal language is not one system but two; the first system uses a small inventory of forms, phonemes, which sequentially combine to form a much larger inventory of words. The phonemes don't carry meaning but the words do, and this association is arbitrary. The second system combines the large number of words into an effectively limitless number of sentences.

Productive

The **productive** property means that there is an effectively unlimited number of things that one can say in a verbal language. It is not the case that we have a fixed number of messages and in communication simply choose one from this selection; rather many of the sentences we produce are unique, that is, the particular combination of words we use have never been used before. Of course many of the sentences we say are conventionalised ("How do you do?" etc), but most are unique. It might seem questionable if this is really the case as most people don't feel they are being particularly "original" when they speak. However, a simple examination of the combinatorics of combining even a modest vocabulary of 10,000 words shows that a vast number of possibilities exist (certainly larger than the number of atoms in the universe).

Discrete

Both the basic types of linguistic unit, the phoneme and the word, are **discrete**. Phonemes are distinct from one another and form a finite and fixed set. The set of words is less fixed (we can create new words), but nevertheless all words are distinct from one another and are countable. What is remarkable about the discreteness property of language is that the primary signal of language, speech, is not discrete and hence there is a mismatch between form and signal. Speech signals are continuous functions of pressure over time (see Chapter 10) and so speaking is a process of

converting a discrete phoneme representation into a continuous acoustic one. This process is reversed during listening, where we recover a discrete representation from the continuous signal. As we shall see, this is a remarkably complex process, and is one of the main reasons why automatic speech synthesis and speech recognition are difficult.

2.2.2 Linguistic levels

Another common way to look at verbal language is as a series of so-called linguistic **levels**. So far we have talked about the duality of form such that we have a phonemes which combine to form words which combine to form sentences. In our model, sentences, words and phonemes are considered the primary units of form, and suffice to describe the basic properties of verbal language. It is however extremely useful to make use of a number of secondary units, which we shall now introduce. **Morphemes** are units used to describe word formation or the internal structure of words. This describes for instance the connection between the three English words BIG, BIGGER, BIGGEST, and the study of this is called **morphology**. Next we have **syntax**, often defined as the study of the patterns of words within a sentence. Syntax describes the process of how nouns, verbs, prepositions and so on interact, and why some sequences of words occur but not others. The fundamental units of syntax are the **syntactic phrase** and sentence. With these levels, we can for a given sentence describe a **syntactic hierarchy** where a sentence comprises a number of phrases, each of these comprises a number of words, which then comprise morphemes, which then comprise phonemes.

We can also consider the **phonological hierarchy**, which is a different, parallel hierarchy which focuses on the sound patterns in a sentence. In this, we have the additional units of **syllables**, which are a structural sound unit used to group phonemes together; and **phonological phrases**, another structural sound unit that groups words together within the sentence. Representations using these units can also be expressed in a hierarchy of sentences, prosodic phrases, words, syllables and phonemes.

The difference between the two hierarchies is that the phonological one concentrates on the sound patterns alone, and is not concerned with any aspect of meaning, while the syntactic one ignores the sound patterns, and concentrates on the grammatical and semantic relationships between the units.

Next we consider levels between the form and the signal, which in spoken communication is an acoustic waveform. Here we have **speech acoustics** which involves working with the acoustic speech sound signal. Next we have **phonetics**, which is often described as the study of the physical speech sounds in a language. Phonetics is often studied independently of other linguistic areas, as it is seen as part of the system which links form and signal rather than as part of the meaning/form verbal linguistic system. **Phonology** is the study of sound patterns as they relate to linguistic use; in contrast to phonetics, this is considered part of the core linguistic system as phonology describes the form in the meaning/form system. That said, phonology and phonetics are closely related and phonology is sometimes described as the organisational system of phonetic units.

Semantics is sometimes given a broad definition as the study of “meaning” in the general sense, but is sometimes given a much narrower definition, and is concerned only with certain aspects of meaning, such as analytic truth, reference, and logic. Finally **pragmatics** concerns a number of related areas including the aspects of meaning not covered by semantics, the interaction of language, a speaker’s knowledge and the environment and the use of language in conversations and discourses. Sometimes pragmatics is described as being “parallel” to semantics/syntax/phonology etc as it is the study of how language is used.

It is important to realise that these terms are used both to refer to the linguistic level and the area of study. In traditional linguistics these terms are often taken as absolute levels that exist in language, such that one sometimes sees “turf wars” between phoneticians and phonologists about who best should describe a particular phenomenon. From our engineering perspective, we will find it more convenient to adopt these terms, but not worry too much about the division between each; rather it is best to view these levels as salient points in a continuum stretching from pragmatics to acoustics.

2.2.3 Affective Prosody

We now turn to a different component of human communication. Broadly speaking, we can state that **prosody** has two main communicative purposes, termed **affective** and **augmentative**. The affective use of prosody can be used to convey a variety of meanings: perhaps the most straightforward case is when it is used to indicate primary emotions such as anger or pain. Prosody is particularly good at this; when we hurt ourselves and want to tell someone we usually let out a yell rather than say “look, I’ve cut off my fingers” in a calm measured tone. The use of prosody to convey primary emotion is also largely universal; when listening to someone who speaks a different language we find they use more or less the same form to indicate the same meanings. In fact this extends to animals too; we are often in little doubt as to what an animal snarl or growl is trying to convey. Beyond primary emotions we can consider a whole range of meanings that are expressible via prosody. These include secondary emotions such as angst or distrust, speech acts such as question, statement and imperative; modes of expression such as sarcasm or hesitancy and so on.

Many of these phenomena are language specific (in contrast to primary emotion), and have a conventional or arbitrary character such that a speaker of that language has to learn the appropriate prosodic pattern in the same way that they learn the form/meaning correspondences for words in verbal language. In fact we can see these effects as lying on a scale with universal systems lying at one end and conventional arbitrary systems lying at the other. In many languages, a rising pitch pattern at the end of a sentence indicates a question; but this is not universally so. Ladd [269] shows that the pattern is reversed in Hungarian such that certain questions are indicated by a falling pitch.

One of the central debates in the field of prosodic linguistics is just how far this scale extends. Crudely speaking, the argument concerns whether prosody is *mainly* a universal, non-arbitrary

system, with a few strange effects such as Hungarian question intonation; or whether it is in fact much more like verbal language in that it is *mainly* arbitrary, and has rules of syntax and so on.

2.2.4 Augmentative Prosody

The second aspect of prosody is called **augmentative prosody**. The basic purpose of this is to *augment* the verbal component. Taking a simple example first, we see that many sentences are highly confusable with respect to syntactic structure. In sentences such as

(1) BILL DOESN'T DRINK BECAUSE HE'S UNHAPPY

we see that there are two completely different interpretations which depend on which syntactic structure is used. The speaker can choose to insert a pause, and if she does so, she will bias the listener's choice as to which interpretation is intended, and hence facilitate understanding. In addition to helping with phrasing structure, this component of prosody is also used to draw attention to words by emphasising them.

This use of prosody is called augmentative because it augments the verbal component and therefore helps the listener to find the intended form and meaning from the signal. In contrast to affective prosody, the use of augmentative prosody here does not add any new meaning or extra aspect to the existing meaning, it is simply there to help find the verbal content. One of the most interesting aspects of augmentative prosody is that speakers show considerable variation in how and when they use it. While one speaker may pause at a certain point, another speaker will speak straight on, and we even find that the same speaker will quite happily use a pause on one occasion and not on another. We can explain this variation by considering the range of conversational situations where augmentative prosody might be used.

In general, when a speaker generates a message that is potentially ambiguous, it is often entirely obvious from the form of words used, the background knowledge and the situation what interpretation is intended, and in such cases no or little augmentative prosody is used. In other cases however, we may find that the speaker suspects the listener may have difficulty in interpretation, we find that here the speaker does indeed make use of prosody. In other words, prosody is usually only used when it is needed. Hence if we consider a sentence in isolation, it is nearly impossible to predict what prosody a speaker will use. A consequence of this is that it is clearly possible to use more or use less prosody for any given sentence. We can think of this as a scale of prosodic intensity, and this applies to both affective and augmentative prosody. In many cases, speakers utter sentences at the lower end of this scale, in which case we say they are speaking with **null prosody** or **neutral prosody** to indicate that there is no true prosody being used.

2.3 COMMUNICATION PROCESSES

In this section, we will build a general model of communication that will serve as a framework in which to analyse the various communicative processes involved in spoken and written language.

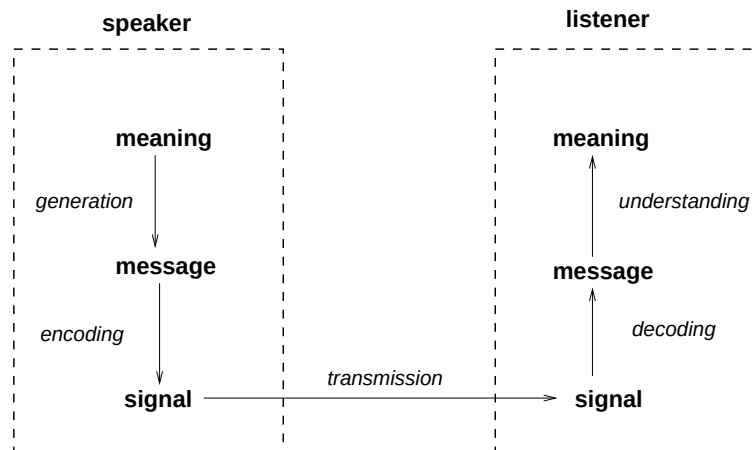


Figure 2.3 Processes involved in communication between two speakers.

For illustration, let us consider a dialogue between two participants, A and B. In a normal conversation, A and B take it in turns to speak, and these are known as **dialogue turns**. Suppose, A decides to say something to B. To do this, A firstly **generates** a linguistic form, which we shall call the **message**. He then **encodes** this as a **speech signal**, which is a form which can be **transmitted** to B (lets assume they are in the same room). On receiving the signal, B first **decodes** the signal and creates her own message as linguistic form. Next she attempts to **understand** this message and so find the meaning.

In line with our model of language explained before, we have three types of representation: the meaning, the form and the signal. In communicating, A and B use 4 separate systems to move from one to the other; these are:

generation : meaning to form

encoding : form to signal

decoding : signal to form

understanding : form to meaning

This model works equally well for written language; A converts his thought into a message, then encodes this as writing and sends this to B (in this case the “sending” can be quite general, for example an author writing a novel and a reader subsequently picking it up). Once received, B attempts to decode the writing into a message, and then finally understand this message. A diagram of this is shown in Figure 2.3.

2.3.1 Communication factors

We start our discussion by simply asking, why do the participants say the things they do? While “free-will” plays a part in this, it is a mistake to think that “free-will” is the sole governing factor

in how conversations happen. In fact there are a number of guiding principles involved, which we shall now explore.

The first concerns the understood conventions of communication. In human conversation, participants don't just say everything on their mind; to do so would swamp the conversation. Rather, they only say what they need to in order to get their point across. We can state this as a case of **effectiveness** and **efficiency**; we wish to effectively communicate our message so that it is understood, but also to do so efficiently so that we don't waste energy. In fact it is clear that conversation works by adopting a set of understood conventions that govern or guide the conversation. The most famous exposition of these principles was by Grice, who defined four main principles, of conversation which are now commonly known as **Grice's maxims** [184]. These principles state for instance that it is expected that one participant should not be overly long or overly short in what is said; one should say just the right amount to make the thought clear. Grice's maxims are not meant as prescriptive guides as to how conversation should progress, or even statements about how they do progress, but rather a set of tacit conventions, which we normally adhere to, and which we notice when they are broken.

The second factor concerns the **common ground** that the participants share. This is a general term which covers the background knowledge, situation and mutual understand that the conversational participants share. When in the same room, they can point to objects or refer to objects, actions and so on, in a different way to how such things can be referred to if they are more distant. If one says "have a seat" and there is a chair in the room, this means something straightforward. The situation can significantly alter the intended meaning of an utterance. Take for instance the sentence "my car has broken down". Now this can be taken as a simple statement of fact, but if someone says this upon entering a mechanic's workshop, the implication is that the person's car has broken down and he'd like the mechanic to fix it. If by contrast, this is the first thing someone says when arriving for dinner, the intention might be to apologise for being late because an unforeseen travel delay occurred. To take another example, if two experts in the same research field are talking, they can talk assuming that the other has a detailed knowledge of the subject matter. This is different to when an expert talks to a lay person.

Finally, the knowledge of the **channel** helps form the message. Participants don't generate messages and then decided whether to speak them or write them down; they know what medium is going to be used before they start. This allows the participants to best exploit the strengths of each form of communication; in spoken language extensive use may be made of prosody; in written language the author may take the liberty of explaining something in a complicated fashion knowing that the reader can go back and read this several times.

2.3.2 Generation

In our model, the process of converting a meaning into a message is called **message generation** (elsewhere the term **language generation** is often used, but this is too potentially confusing for our purposes). To demonstrate how generation works, let us use the above mentioned communi-

cation factors to consider situation where a participant in a conversation wants to know the score in a particular football game. We should realise that the participant has considerable **choice** in converting this thought into a message. For example, the participant can make a statement “I want to know the score” or a question “what is the score”; the participant has choices over mood, and may choose to say “can you tell me the score?”, may choose between active and passive “who has scored?” vs “have any goals been scored?”. The participant has choices over vocabulary, may well qualify the message, may decide on a casual, formal, polite, demanding tone and so on. In addition the participant can choose to emphasise words, pick different intonational tunes, perhaps make a pointing gesture and so on. The important point is that the participant have a wide range of options in turning the thought into the message. The result is the message, and a perfectly transparent communication system would send this to the receiver with all this detail and with no chance of error.

It is sometimes assumed that messages themselves convey or represent the communicated meaning. While in certain cases this may be possible, we have just demonstrated that in most conversations, to extract (understand) the correct meaning one must know the situation and shared knowledge.

2.3.3 Encoding

Encoding is the process of creating a signal from a message. When dealing with speech, we talk of **speech encoding** and when dealing with writing we talk of **writing encoding**. Speech encoding by computer is more commonly known as **speech synthesis**, which is of course the topic of this book.

The most significant aspect of speech encoding is that the nature of the two representations, the message and the speech signal, are dramatically different. A word (e.g. HELLO) may be composed of four phonemes, /h eh l ow/, but the speech signal is a continuously varying acoustic waveform, with no discrete components or even four easily distinguishable parts. If one considers the message, we can store each phoneme in about 4-5 bits, so we would need say 2 bytes to store /h eh l ow/. By contrast, a reasonably high fidelity speech waveform² would need about 20,000 bytes, assuming that the word took 0.6 seconds to speak. Where has all this extra content come from, what is it for and how should we view it?

Of course, the speech signal is significantly more **redundant** than the message: by various compression techniques it is possible to reduce its size by a factor of 10 without noticeably affecting it. Even so, this still leaves about three orders of magnitude difference. The explanation for this is that the speech waveform contains a substantial amount of information that is not directly related to the message. First we have the fact that not all speakers sound the same. This may be a fairly obvious thing to say, but consider that all English speakers can say exactly the same message, but they all do so in subtly different ways. Some of these effects are not controllable by

² Say 16 bits dynamic range and 16,000Hz sampling rate. (See Chapter 10 for an explanation)

the speaker and determined by purely physiological means. Some are due to prosody, where for instance, the speaker can have different moods, can affect the signal by smiling when talking, can indicate different emotional states and so on. This fact is hugely significant for speech synthesis, as it demonstrates that encoding the message is really only part of the problem; the encoding system must do several other tasks too if the speech signal is to sound genuinely human-like to the receiver.

2.3.4 Decoding

Decoding is the opposite of encoding; its job is to take a signal and create a message. **Speech decoding** is the task of creating a message from the signal; the technology that does this is most commonly known as **speech recognition**. The challenges for speech decoding (speech recognition) systems are just the converse of speech encoding; from a continuous signal, which contains enormous amounts of information, the system must extract the small amount of content that is the message. Speech recognition is difficult because of the problem of **ambiguity**. For each part of a message, a word or a phoneme, if one studies the speech patterns produced from that one sees that there is an enormous amount of variability. This is partly caused by the factors we described above, namely speaker physiology, emotional state and so on, and partly from the encoding process itself, where a phonemes characteristics are heavily influenced by the other phonemes around it. This variability wouldn't be a problem *per se*, but it turns out that the patterns for each phoneme overlap with one another, such that given a section of speech, it is very hard to determine which phoneme it belonged to. These aspects make speech recognition a difficult problem.

Text decoding, also called **text analysis**, is a similar task where we have to decode the written signal to create a message. If the writing is on paper, we face many of the same problems as with speech recognition; namely ambiguity that arises from variability, and this is particularly true if we are dealing with hand writing. Computerised writing is significantly easier as we are dealing with a much cleaner representation in which the representation is discrete and each character is uniquely identified.

For many types of message, the text decoding problem is fairly trivial. For instance, for simple text such as `the man walked down the street`, it should be possible for a simple system to take this and generate the correct message all the time. Significant problems do arise from a number of sources. First, ambiguity exists even in computerised written signals. One type is **homograph** ambiguity, where two different words have the same sequence of characters (e.g. `polish` can mean "from Poland" or "to clean", `console` can mean "to comfort" or "a part of a machine").

One final point should be made when comparing text and speech decoding. In speech recognition, the problem is understood as being difficult, and so error rates of say 10% are often seen as good system performance on difficult tasks. As we have seen text decoding is fundamentally easier, but along with this so are expectations. So while for some tasks the vast majority of words may be decoded successfully, the expectation is also high and so error rates of 0.1% may be seen

as too high.

2.3.5 Understanding

Understanding is the process of creating a thought from the message. Understanding can be thought of as the reverse of generation and also bears similarity to decoding in the sense that we are trying to recover a thought from a representation (the message), which also includes information generated from other sources such as the background, situation and so on.

In understanding, we also have an ambiguity problem, but this is quite different from the ambiguity problem in decoding. For example, it is perfectly possible when hearing a speech signal to decode it into the correct message, but then fail to understand it because of ambiguity. For instance, if someone says “did you see the game last night”, it is quite possible to correctly identify all the words, but fail to determine which game is being referred to.

2.4 DISCUSSION

This chapter has introduced a model of communication and language that will serve as the basis for the next chapter, which looks at text-to-speech conversion in detail, and the subsequent chapters, which explore the various problems that occur and the techniques that have been developed to solve them. We should stress that what we have presented here is very much a *model*; that is a useful engineering framework in which to formulate, test and design text-to-speech systems.

In building this model, we have drawn from a number of fields, including semiotics, information theory, psychology, linguistics and language evolution. These fields have overlapping spheres of subject matter, but unfortunately are often studied in isolation such that none really gives as complete and detailed a model of communication as we might wish for. Each of these separate disciplines is a vast field in itself, so we have only touched on each area in the briefest possible way.

Considering semiotics first, we should look at the works of Ferdinand de Saussure, who is widely acknowledged to have invented the field of semiotics and is considered by many as the first modern linguist. Saussure was the first to emphasise the significance of the arbitrary nature of form and meaning in communication and proposed that the study of the relationship between the two should be the central topic in linguistics [391], [392]. The philosopher Charles Sanders Pierce also pioneered early work in semiotics and semantics, and gave us the framework of iconic and symbolic semiotics which is used here [350]. The second main area we draw from is communication theory or information theory, and from this we have the principles of encoding, decoding, channels, messages and signals. The pioneering work in this field was produced by Claude Shannon [400], who was originally interested in the problem of accurate transmission of information along telephone wires (he worked for the telephone company AT&T). It soon became clear that the implications of this work were much broader (Weaver [401]) and it has been widely

adopted as an approach to the analysis of signals arising from linguistic sources, such that this **information theoretic** approach is now dominant in speech recognition and many other fields. Saussure did not in general deal much with signals, and Shannon and Weaver did not deal much with meaning, but their work clearly intersects regarding the issue of form and message (the term form is used in linguistics and semiotics and message in information theory, but they both refer to the same thing.)

The field of linguistics is often described as the science of language, but in the main this field focuses very much on just the meaning/form verbal part of language, as opposed to the entire process which includes participants, conversations, signals and so on. Linguistics generally approaches language without reference to use; the idea is that a language is an entity and it can be studied independently of any purpose to which it is put³. This approach is part tradition and part pragmatic; as linguistics can trace its lineage back to the philosophical musings of ancient Greek philosophers. In addition there is the practical justification that by isolating effects it makes them easier to study. In modern times, the field of linguistics has been dominated by Noam Chomsky[89], [90], [88], who in addition to many other innovations has championed the idea that the primary function of language is not communication at all, but rather forms the basis of the human mind; it evolved for this reason and its communicative purpose was a useful by product. Despite this mismatch between linguistics and our more practical needs, this field does provide the basics of a significant chunk of the overall picture of human communication, including the models of phonetics, phonology, words, syntax and so on that are central to our problem. Many linguistics text books tend to be heavily skewed towards one of the many incompatible theories within the discipline, so care needs to be taken when reading any book on the subject. A good introductory book which explains the “hard issues”, which is accessible to those with a scientific background and which doesn’t follow any particular theory is Lyons [290].

The study of human communication in terms of participants, conversations, background knowledge and so on is often performed outside the field of mainstream linguistics by psychologists and philosophers. Good books on human conversation and language use include Clark [95] and Clark and Clark [96], Searle [395], Grice [184] and Power [361].

2.5 SUMMARY

Communication

- A semiotic system has three levels of representation; meaning, form, and signal.
- in affective communication, the meaning, form and signal are continuous and simply related.
- in iconic communication, the form resembles the meaning in some way. Iconic systems can be either discrete or continuous

³ In fact in surveying more than twenty introductory text books in linguistics found in the Cambridge University book shop, I did not find even one that mentioned that language is used for communication!

- in symbolic communication, the form does not resemble the meaning and the correspondences between these have to be learned.
- Symbolic systems are discrete, and this and the arbitrary meaning/form relationship mean that communication of complex meanings can occur more effectively than with the other types.
- Human language operates as a discrete symbolic system.

Components of language

- Human communication can be divided into two main components:
 - The verbal component describes the system of phonemes, words and sentences. It is a discrete symbolic system in which a finite number of units can be arranged to generate an enormous number of messages.
 - The prosodic component is by contrast mostly continuous, and in general expresses more basic meanings in a more direct
- Affective prosody is when a speaker uses prosody to express emotion, speech act or other information which is conveyed sentence by sentence
- Augmentative prosody is used to disambiguate and reinforce the verbal component.
- Both types of prosody are optional, in that speakers can choose to use no prosody if they wish and just communicate via the verbal component. Writing in fact uses just this component.

Communication

- We can break communication into four processes in spoken communication:
 1. generation is the conversion of meaning into form by the speaker
 2. encoding is the conversion of form into signal by the speaker
 3. decoding is the conversion of signal into form by the listener
 4. understanding is the conversion of form into meaning by the listener
- The equivalent processes exist for written communication also, where we have a writer and reader instead of a speaker and listener.

3

THE TEXT-TO-SPEECH PROBLEM

We now turn to an examination of just what is involved in performing text-to-speech (TTS) synthesis. In the previous chapter, we described some of the basic properties of language, the nature of signal, form and meaning, and the four main processes of generation, encoding, decoding and understanding. We will now use this framework to explain how text-to-speech can be performed.

In TTS, the input is writing and the output speech. While it is somewhat unconventional to regard it as such, here we consider writing as a signal, in just the same way as speech. Normal reading then is a process of decoding the signal into the message, and then understanding the message to create meaning. We stress this, because too often no distinction is made at all between signal and form in writing; we here about “the words on the page”. More often an informal distinction is made in that it is admitted that real writing requires some “tidying up” to find the linguistic form, for instance by “text normalisation” which removes capital letters, spells out numbers or separates punctuation. Here we take a more structured view in that we see linguistic form as clean, abstract and unambiguous, and written form as a noisy signal that has been encoded from this form.

The process of reading aloud then is one of taking a signal of one type, writing, and converting it into a signal in another type, speech. The questions we will explore in this chapter are:

- How do we decode the written signal?
- Does the message that we extract from the written signal contain enough information to generate the speech we require?
- When someone is reading aloud, what are the relationships between the author, reader and listener?
- What do listeners want from a synthetic voice?

3.1 SPEECH AND WRITING

It is clear that natural (human) language developed first using sound as its medium. While humans clearly also communicate with facial and hand gestures and so on, these systems are not able to express the same wide range of expressions as natural language. Quite *why* human language

evolved using sound as its medium is not clear, perhaps it was down to an ability to make distinct patterns more easily with the mouth, than say by waving hands. It is important to realise that this choice of sound may well have been a complete accident of evolution, today we communicate in many different ways with vision either by reading or by interpreting sign language. For most of human history, speech was the only (sophisticated) means of communication, but a second key invention, perhaps 10,000 years ago, was that of writing.

To begin with, we need to look at the nature of speech and writing in more depth. We will do this from a number of angles, looking first at the physical differences between these signals, before moving on to look at how these differences influence communication.

3.1.1 Physical nature

The speech signal is a continuous, acoustic waveform. It is created by the operation of the vocal organs in response to motor control commands from the brain. We only have a partial understanding of how this process works, but it is clear that the coordinated movement of the vocal organs to produce speech is extremely complicated. One of the main reasons why the relationship between the speech signal and the message it encodes is that speech is continuous and the form is discrete. In addition to this, speech is highly variable with respect to the discrete message as many other factors, including prosody, speaker physiology, a speaker's regional accent and so on all affect the signal. A single message can generate many potential signals. Finally, speech is not permanent; unless a recording device is present, once something is spoken, it is lost for ever and if no listener is present or a listener does not hear the speech then the message that the speech encodes will never be decoded.

Writing is fundamentally different from speech in a number of ways. Firstly writing is primarily seen as **visual**: traditionally, writing always took the form of visible marks on a page or other material, and today we are quite used to seeing writing on computer and other screens. This is not however a complete defining property; we consider braille a form of writing and its key benefit is of course that one does not need to be able to see to read it. It has been argued [199] that the main defining feature of writing is that it is **permanent** in a way that speech is not. Whether the writing is created with a pen on ink or impressions on a page, the key point is that once written it does not disappear. We can of course destroy what we have done, but this is not the point; the idea is that with writing we can record information, either so that we can access it at a later date, or send it over some distance (as a letter for instance).

What then does writing record? Writing can be used to record natural language, but in saying this we should make it clear that writing attempts to record the linguistic form of a message, rather than the actual signal. That is, if we write down what someone says, we do not attempt to record all the information in the speech, such that a reader could exactly mimic the original; rather we decode the speech into a linguistic message, and then encode that. Of course quite often the message we wish to write has never been spoken, and so when writing we generate a linguistic form from our thoughts, and then encode this as writing without ever considering speech.

We explained in Chapter 2 that spoken language had a dual structure of sentences made of words, and words comprised of phonemes. We have a similar dual structure in written communication, where sentences are made of words (as before), but where words are made of **graphemes**. Graphemes are in many ways analogous to phonemes, but differ in that they vary much more from language to language. In alphabetic languages like English and French, graphemes can be thought of as letters or characters (we will define these terms later); in syllabic writing like Japanese hiragana, each grapheme represents a syllable or mora, and in languages like Chinese each grapheme represents a full word or part of a word.

It should be clear that even apart from the visual and permanency aspects writing signals and speech signals are very different in nature. The most important difference is that writing is discrete and “close to the message”, while speaking is continuous, more complex, and does not encode the message in any trivial way. The complexities of the speech encoding and decoding arise because we are using an apparatus (the vocal organs) for a purpose quite different from their original one. In fact, one can state that it is a marvel that speech can be used at all for communication messages. In contrast, writing is a cultural invention, and was designed explicitly and solely for the purposes of communicating form. If we consider it a “good” invention, it should not be surprising that the encoding mechanism between linguistic form and writing is much simpler. This is not to say that the reading and writing process are trivial; it takes considerable skill for the brain to interpret the shapes of letters and decode messages from them; but this skill is one shared with other visual processing tasks, and not something specific to writing. The difficulty of decoding a written signal depends very much on the specific message. Often the relationship between the writing on the page and the words is obvious and transparent; but often it is complex enough for the decoding and encoding processes to be non-trivial, such that we have to resort to sophisticated and principled techniques to perform these operations.

3.1.2 Spoken form and written form

We have just explained the differences between spoken and written signals. Turning now to form, we can ask how (if at all) do the spoken and written form differ? One way of looking at this is to consider the combined generation/encoding process of converting meaning into a signal. In general we can convert the same meaning into either a spoken signal or a written one, so we see that speech and writing (in general) share the same meaning representation but different signal representations. The question now is, how much of the overall language system do the two share? It is difficult to answer this conclusively, as to do so (in humans at least) would require a much more thorough understanding of the processes than we currently have. It is however possible to consider a model where the meaning to form systems for both spoken and written language share a common system from meaning to the second level of form. In other words, the meaning-to-word system is the same for both. The divergence occurs when we consider the primary level of form; in spoken language words are comprised of phonemes, in written language, they are comprised of graphemes. From this point until the signal, the two systems are separate. Depending on the

language, the primary forms may be broadly similar (for instance in Spanish) such that one can deduce the pronunciation from the spelling and vice versa, or quite different, (for example in Chinese).

While a few problems certainly occur, this model seems to fit the facts very well. Whether we write or say the sentence THE MAN WALKED DOWN THE STREET, there are six words and they occur in the same order in the spoken and written versions. This fact may be so obvious as to seem superfluous, but it is in fact vital to our discussion. One would find it difficult for instance to teach someone to read French text and speak English words: even if we learned that the pronunciation for *chien* was /d ao g/, we would find it difficult to understand why the written French sentence has different numbers of words in different orders from the spoken English version.

The main point in this realisation of the commonality between the first level of form in written and spoken language is that it shows that if we can decode the form from the written signal, then that we have virtually all the information we require to generate a spoken signal. Importantly, it is *not* necessary to go all the way and uncover the meaning from the written signal; we only have to perform the job of text decoding and not text understanding. There are two qualifications we need to add to this. Firstly, some understanding may be required, or may make the job easier when we are resolving ambiguity in the written signal. For example, when encountering the homograph *polish* it may help in our disambiguation to know that the topic of the sentence is about people from Poland. Secondly, we may need to generate prosody when speaking, and this may not be obtainable from the writing. This issue is dealt with below (Section 3.5.5).

This conclusion is so important for our system that it is worth restating again; by and large, the identity and order of the words to be spoken is all we require to synthesise speech; no higher order analysis or understanding is necessary.

3.1.3 Use

Speech and writing are used for different purposes. When do we use one instead of the other? In certain cases we have no real choice; in face to face communication it is normal to speak: while we could write something down and pass it to our friend to read, this is in general cumbersome. When writing a book, or an article we normally use writing; that is the standard way in which material is stored, copied and distributed. Before the computer era it was fairly clear which communication mode was used for what purpose.

Modern technology certainly gives us more choices; with telephones we can communicate over distance, whereas in the past this was only possible by letter. We can also now quickly send written messages electronically, by email, phone messaging, computer messaging and by other means. Computer or “instant” messaging is particularly interesting in that it assumes many of the characteristics of speech, but exists as a written signal. When phone messaging (“texting” or “sms”) first arrived, many thought the idea absurd; why would you ever want to send a message in writing with a phone when you could simply call the person directly? In fact, many “functional” uses have been found; a common one is for someone to send an address as a text message as this

is easier for the receiver than having to listen and write the address down in case it is forgotten. Uses go beyond this though, and today we see “texting” as particularly popular among teenagers. Anecdotaly, we believe that by using text, teenagers can think about, plan and edit their messages so as to seem more confident, funny and so on; this saves the embarrassment that may ensue from having to actually *talk to someone*.

One major difference is that speech is in general **spontaneous**. During a normal conversation, we usually speak with little planning or forethought; we can’t simply suspend the conversation until we have thought of the best possible way to say something. This accounts for the **disfluencies** in speech, such as hesitations, pauses and restarts. Writing is *usually* free of these effects, but it is important to realise that this is not because of any property of writing itself, but more because the writer can spend time over difficult sections, or can go back and correct what has been written.

Another significant difference is that in spoken communication the speaker can often see or interact with the listeners. In face to face communication the speaker can see the listener and see how they react to what is being spoken; in addition to the listener responding directly (“Yes, I agree”), they may nod their head, give an acknowledgment or show confusion and so on. In response, the speaker can alter or modify their behaviour accordingly. This also happens over the phone, where verbal cues (“uh-huh”, “yeah...”) are given by the listener to help show the speaker their level of comprehension and agreement. This is less effective than with the visual cues; anyone who has taken part in a multi-person call can attest to the difficulty involved in sensing everyone’s understanding and participation in the conversation.

If we compare this to a typical situation where written communication is involved, we see some important differences. For example, imagine someone is reading a newspaper on a train. Here the author of a particular column will not personally know the reader, in fact the author will probably know only a handful of people who read the column. The author and reader are separated in time and space and the communication is one way only (the reader may comment on the article to someone else, but this will not reach the author). Because of these factors, the situation with respect to common ground is very different. First, the author is writing for many readers, and all will have somewhat different degrees and types of background knowledge. Secondly, as no feedback is possible, it is necessary for the author to “get it right first time”, there is no possibility of starting an article and then checking to see if the reader is following. These situational distinctions account for many of the differences in content between spoken and written language; for example, because the level of background information can not be known, writing is often more explicit in defining references, times and places and so on. Informally, we can describe the writing as being “more exact” or “less ambiguous”. Of course, these terms only make sense when considering the speech or writing in isolation; as explained above, most ambiguity is instantly resolved when the situation is taken into account.

3.1.4 Prosodic and verbal content

One of the most notable features of writing is that it nearly exclusively encodes the verbal component of the message alone: the prosody is ignored. This feature seems to be true of all the writing systems known around the world. Because of this it is sometimes said that written language is **impoverished** with respect to spoken language as it can only express part of the message. It is of course possible to represent effects such as emotion in written language, (one can simply write “I am sad”) but this is done by adding information in the verbal component and is different from the systematic use of prosody to indicate this emotion. Quite why prosody is not encoded in writing is not well understood. Possible explanations include the fact that it is often continuous, which creates difficulties if we are using a discrete coding system. Prosody is most powerful for emotive language; and while this is an important part of literature, the ability to express emotion directly is not needed so much for legal documents, newspapers, technical manuals and so on.

We do however have a limited prosodic component in written language. Punctuation can be used to show structure; and underlined, italicised or bold text is used to show emphasis. These are clearly encoding augmentative prosody in that the punctuation and italics are there to help the reader uncover the correct verbal form from the writing. Affective prosody on the other hand is nearly entirely absent; there is no conventional way to shown anger, pain or sarcasm. The only aspect that is present is the ability to express sentence types such as question, statement and imperatives with punctuation at the end of a sentence. It is interesting to note that in most early writing systems (e.g. Ancient Greek) no punctuation was used at all, but as writing evolved punctuation appeared and more recently italics and other font effects have become commonplace due to the wide availability of computer word processing. In the last few years, a system of “emotions” has become common in email and other electronic communication, in which punctuation is used to indicate happy faces :-) and other effects, which attempt to encode affective prosody. Quite why this trend is present is again not clear; while we can ascribe the rise of smiley faces to writing being used for more informal purposes, this doesn’t explain the trend from classical Greek to the English of 200 years ago with its commas, semi-colons and full stops.

3.1.5 Component balance

From the outset, a person knows whether a message should be spoken or written: it is not the case that the person first generates the message and then decides what channel it should be sent in. This principle gives rise to another set of differences between spoken and written signals. Because he knows that a certain message is to be written, the person will, from the outset, structure the message so as to take account of the component balance available in that signal. So if for instance, if a writer wishes to convey something happy, he is quite likely to express this happiness in the verbal component, as the ability to encode affective prosody is not available. So he might say “I’m thrilled that you got that place” instead of perhaps “well done that you got that place”, which could give the impression of being somewhat unmoved. With appropriate expression, the “well done...”

version may sound very enthusiastic if spoken. The different use of components in written and spoken language is called **component balance**.

Simply put, a speaker has a open choice of how much prosody or how much verbal content to put in a message; they will balance these against each other, to create a final signal that conveys the message they wish in a way that is balanced between efficiency and effectiveness. A writer on the other hand has only a very limited ability to use prosody; and so has to rely on the verbal component more. Because of this, he is likely to “beef-up” the verbal component, so that it is less ambiguous and more explicit than the spoken equivalent. In addition, the writer will “verbalise” any emotion if that is required (that is, explicitly explain the emotion by using words).

It is this awareness of the prosodic deficiencies of written language and the ability to get around them that enables good writers to make themselves understood. It should be noted that the majority of people don’t really write anything more than the odd note or shopping list in their day to day lives. While we all learn to read and write at school, it is *reading* that is the essential skill to cope with modern life; only a few of us regularly write. This point is worth making because writing well, that is so that one communicates effectively, is actually quite difficult. One of the main aspects of this is that the ability to encode emotions, make writing unambiguous, and have a good picture of a potential reader in mind while writing, are all skills that require some mastering. We have seen this with the advent of email - many new users of emails send messages which are far too brief (e.g. “yes” in reply to something that was asked days ago) or that come across as being quite rude - the writer has written the words that he would normally use if he were speaking face to face, but without the prosody, the emotion (which could be humour or politeness) is lost and the words themselves seem very abrupt.

3.1.6 Non-linguistic content

In our discussion so far we have equated speech as being the signal used to communicate linguistic messages acoustically and writing as the signal used to communicate linguistic visually or over time. This in fact simplifies the picture too much because we can use our speaking/listening and reading/writing apparatus for types of communication which don’t involve natural language. Just as we can use our vocal organs for whistling, burping, coughing, laughing and of course eating, breathing and kissing, we find a similar story in writing, in that we can use writing for many other purposes than just to encode natural language.

This is a significant concern for us, in that we have to accurately analyse and decode text, and to do so we need to identify any non-linguistic content that is present. When considering “non-linguistic” writing, it helps again to look at the evolution of writing. The key feature of writing is that it is more permanent than speech it can be used to record information. Even the earliest forms of Egyptian and Sumerian are known to have recorded numerical information relating to quantities of grain, livestock and other “accounting” type information. Furthermore, there are several cultures who developed methods for recording and communicating accounting information without developing a system for encoding their language: the most famous of these is the Inca

quipu system, in which numbers could be stored and arithmetic performed. This ability for writing systems to encode non-linguistic information is therefore an ancient invention, and it continues today in printed mathematics, balance sheets, computer programs and so on.[64], [19], [6]

Consider for a moment numbers and mathematics. Without getting to far into the debate about what numbers and mathematics actually are, it is worth discussing a few basic points about these areas. Firstly, in most primitive cultures, there is only a very limited number system present in the language; one might find a word for one, two and three, then a word for a few and a word for lots.

More sophisticated number communication systems (as opposed to numbers themselves) are a cultural invention, such that in Western society, we think of numbers quite differently, and have developed a complex system for talking and describing them. Most scientifically and mathematically minded people believe that numbers and mathematics exist beyond our experience and that advancing mathematics is a process of discovery rather than invention. In other words, if we met an alien civilisation, we would find that they had calculus too, and it would work in exactly the same way as ours. What is a cultural invention is our system of numerical and mathematical notation. This is to some extent arbitrary; while we can all easily understand the mathematics of classical Greece, this is usually only after their notational systems have been translated into the modern system: Pythagoras certainly never wrote $x^2 + y^2 = z^2$. Today, we have a single commonly used mathematical framework, such that everyone knows that one number raised above another (e.g. x^2) indicates the power. The important point for our discussion is to realise that when we write $x^2 + y^2$ this is *not* encoding natural language; it is encoding mathematics. Now, it so happens that we can *translate* $x^2 + y^2$ into a language such as English and we would read this as something like “x to the power of two plus y to the power of two”. It is vital to realise that $x^2 + y^2$ is not an abbreviated, or shorthand form of the English sentence that describes it. Furthermore, we can argue that while mathematics can be described as a (non-natural) language, in that it has symbols, a grammar, a semantics and so on; it is primarily expressed as a meaning and a *written* form only; the spoken form of mathematical expressions is derivative. We can see this from the ease at which we can write mathematics and the relative difficulties we can get in when trying to speak more complex expressions; we can also see this from the simple fact that every mathematically literate person in the western world knows what $x^2 + y^2$ means, can write this, but when speaking has to translate into words in their own language first (e.g. EKS SQUARED PLUS Y SQUARED in English IX QUADRAT PLUS YPSILON QUADRAT in German).

3.1.7 Semiotic systems

A **semiotic system** is a means of relating meaning to form. Natural language is a semiotic system, and probably has by far the most complicated form to meaning relationship, and as we have just seen mathematics is another. Most semiotic systems are much simpler than natural language as they are cultural inventions rather than biological processes (recall that while mathematics itself may be a property of the universe, the mechanism we use for its form and encoding are cultural

inventions). In addition to these examples, we also have computer languages, email addresses, dates, times, telephone numbers, postal addresses and so on. Just as with mathematics, a telephone number is not part of natural language, it is a complete meaning/form/signal system in its own right.

If it were just a case of each semiotic system having its own writing system, we could design a separate system to process each. The problem arises because frequently we mix these systems in the same signal and using the same characters to do so. For example, in the above text, we saw the sentence

(2) Pythagoras certainly never wrote $x^2 + y^2 = z^2$.

in which we clearly mixed natural language text with mathematical text. There are many other cases too; while the mathematics example uses different fonts and so on, many systems share the same characters as natural language writing, so that we get:

- (3) a. I was born on 10/12/67
 b. It weighed 23Kg
 c. I'll send it to paul@yahoo.com

and so on. The point is that these numbers, dates and email addresses are encoded using the same signal as natural language writing. To be able to process this type of writing, we have to identify which parts of the text relate to which semiotic systems.

3.1.8 Writing Systems

To the best of our knowledge¹, the first writing system was that of Sumer which developed into Cuneiform. Perhaps the best known early writing system is the hieroglyphic writing of the Ancient Egyptians, although it is not entirely clear whether these two systems are related.

Initially these systems were **logographic** in nature, meaning that one symbol usually represented one word or one morpheme. Furthermore, the scripts were **pictographic** meaning that the symbols were often iconic, ie pictorial representations of the entity the word stood for. As these systems developed, the iconic nature became less obvious, developing into a more abstract logographic script. Additionally **alphabetic** elements were developed in the language that allowed writers to “spell out” words. These developments should come as no surprise; in Section 2.1 we saw that it is quite natural for a communication system to start highly iconic and then progress to a state when the relation between meaning and form is more arbitrary as this general allows for experienced users to communicate more quickly and more accurately.

Most writing systems can be described as belonging to one of the following groups, but in some cases, e.g. Japanese, the complete writing system uses more than one of these types. Logographic systems, as we have just seen use one symbols to represent words or morphemes. **Syllabic** systems use one symbol per syllable, such as in Japanese hiragana. **Alphabetic** systems

¹ If only someone had taken the trouble to write down this key discovery!

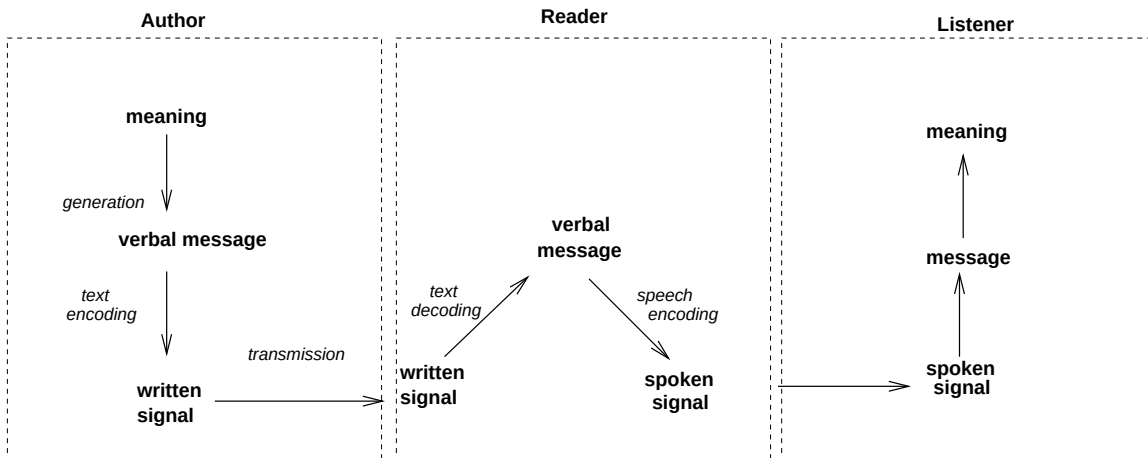


Figure 3.1 Basic model of reading aloud, which shows in the information flow between the author, reader and listener.

use a small list of symbols which roughly correspond to the phonemes of a language. The latin alphabet is of course the most widely used, but its relation to phonemes varies greatly from language to language. **Abjad** systems are ones primarily based on consonants, vowels are usually not written. It is a basic fact of human language that most information that discriminates words is carried by the consonants, and so ignoring these in written language is possible (this is why the abbreviation for DOCTOR is **dr** and not **oo**). The semitic family of languages, including Hebrew and Arabic are abjads.

Writing systems are an important consideration in TTS as in general we have to be able to design a system for any language, and this means being able to handle all writing systems. Problems that have to be solved include, identifying the language, identifying the writing system (for example Japanese uses a logographic and a syllabic system), identifying the computer encoding used, the writing direction and many other issues specific to each language.

3.2 READING ALOUD

We now turn to the problem of **reading aloud**, that is, how to read text and then speak what we have read.

3.2.1 Reading silently and reading aloud

The process of reading, as we normally understand it, is one of taking some text, decoding it into a message, and then understanding it. For clarity, we call this process **reading silently**. By and

large, the process of decoding the text is relatively straightforward for us, and in say a newspaper, we can often successfully decode all the text without error. Understanding is more difficult, in that we may not completely understand an article, and we can perhaps think of various degrees of understanding that we might achieve with respect to a text.

The other participant in this process is the author, and exactly the same factors govern the way she generates a message as with our conversation example of section 2; she assumes some background knowledge on behalf of the reader, makes an assumption about the reader's situation, and is guided by the conversational principles outlined before.

In addition to all these though, she generally assumes that the reader will be reading in the way just described; that is, decoding the writing into the message and then understanding this. This last point is highly significant because this carries with it an assumption that the spoken part of the communication process is not relevant to the task of normal, silent, reading. Examples of the effect of this include a situation where a journalist may include the name of a foreign person or place, but not have to explain how this should be pronounced. The word structure of the sentences can be different, because unlike speaking, there is no need to limit the length of phrase or sentences to chunks that are possible to say with one breath. Far more significantly though, there is little or no specification of prosodic information; it is assumed that the verbal content is enough to create the correct understanding, and emotional or other thoughts are communicated via the verbal component. These problems cause no problem for normal reading, as this speech isn't involved in this process.

Now consider the act of **reading aloud**, that is, reading the text and simultaneously speaking it. This is considerably more complicated in that it involves three agents; an author, a reader and a listener, and two types of signal, text and speech. A diagram of this process is shown in Figure 3.1. This type of reading of course *does* involve speech; and herein lies the problem, because in the majority of cases, the text that has been written has only been written to be read silently; it was never intended to be read aloud. The problem is that as we explained in Section 3.1.5, the component balance is now wrong; a message that was intended to be communicated by writing is now being communicated by speech. While sometimes this is unproblematic, often difficulties can arise from this mismatch, and we can class these difficulties into those to do with prosody and those to do with the verbal content and style. We now investigate these in detail.

3.2.2 Prosody in reading aloud

We have seen already that the written signal contains little or no prosodic information, and when this is decoded, it is decoded into a message that has verbal content only; the prosodic content is generally missing. This is not a problem for reading silently, the reader can understand the thought without needing the prosodic information, and the assumption is that the author has made strong use of the verbal component to compensate for the lack of prosody. If the message is to be spoken aloud we face the issue of what and how much prosody to use. For messages which have little emotion and/or are highly propositional in content, there should not be much of a problem. We

decide to use neutral prosody, and the suprasegmental effects that need to be synthesised can be found from the verbal part of the message. For more emotive messages, or in cases where we think we need significant use of augmentative prosody, we have a fairly serious problem. This is because we have no means of knowing what prosody the speech should be encoded with; the message that we found from the text has no explicit clues as to what the prosody should be.

The only solution is for the reader to generate their own prosody. This is very different from what they are doing with the verbal content; there they are simply decoding from one medium and encoding in another, no imagination, creativity or understanding is required. With prosody generation however, the message does not contain the necessary content, and so the speaker must generate this themselves. We don't really know exactly how people perform this task, but it is plausible that some understanding of the message is undertaken, and from this an attempt is made to generate prosody that the speaker believes is appropriate for the verbal content. Another way of looking at this, is that they in some sense attempt to reproduce the prosody that the author would use if they had decided to speak the message rather than write it.

It is crucial to realise the difference between the process of finding the message by decoding the writing, and the process of prosody generation. In the first, there is a clear right and wrong; the source generated this partial representation, it was clear and specific, and it is our job to find it. Any difficulties that may arise are solely because some distinctions in the structured representation are lost when the message is encoded in the signal (they are ambiguous). In the case of prosody generation, there can be no right or wrong; this information never existed in the first place and so we cannot recover it.

3.2.3 Verbal content and style in reading aloud

Another issue concerns the situation where the reading process is taking place. In our canonical example, the author is writing with the intent that the reader silently reads the text. If the text is then read aloud, we face a potential problem in that the communication convention assumptions used to generate the message have been violated, and we are now converting the message into a signal that it wasn't intended for. A simple example will demonstrate this. Many written documents are quite long; simplistically we can say that this is necessary because the author has a lot to say and it may not be easy to get this across except in a long document (say a book). The author of course assumes that the reader will read this at a time that is convenient and at his own rate. Furthermore, while the author may not explicitly desire this, it is understood that the reader can skip ahead, glance through sections, or reread sections again.

If we now consider the situation of a human or computer reading this document out loud, we see that many of these assumptions no longer apply; the speech progresses linearly from the start of the book, the rate of reading is more or less constant, and that the reader doesn't directly decide when the reading should start and stop. Potentially, this can cause severe problems and we may find that although the system is reading and speaking the text perfectly adequately, the user is not satisfied with the experience because of these violations of communicative norms.

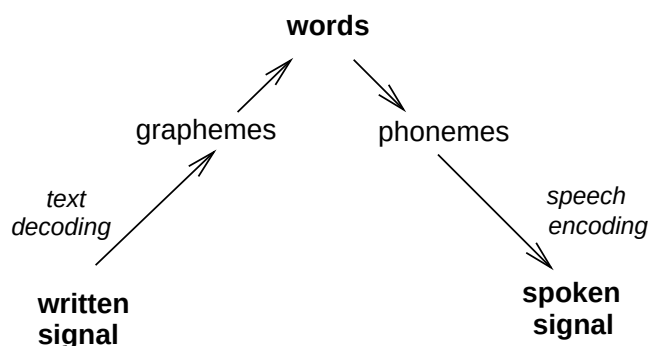


Figure 3.2 The common-form model showing the two processes of text decoding, which finds the words, and speech encoding from those words. For illustration we have shown the different primary levels of form for speech and writing as graphemes and phonemes to show that these are not directly connected.

3.3 TEXT-TO-SPEECH SYSTEM ORGANISATION

The above discussion shed light on some of the issues of spoken and written communication, and the particular process of reading aloud. From our understanding of these processes, we can now define a model of how text-to-speech conversion by computer can be carried out. The next section describes the model we adopt throughout the book, and the sections after that describe alternative models and how these compare.

3.3.1 The Common Form model

In the **common form** model, there are essentially two components; a text analysis system which decodes the text signal and uncovers the form, and a speech synthesis system which encodes this form as speech. The first system is one of resolving ambiguity from a noisy signal so as to find a clean, unambiguous message; the second system is one where we take this message and encode it as a different, noisy, ambiguous and redundant signal. The core idea is that writing and speech share a “common form” and that that, and that alone is the single and correct intermediate representation between the two types of signals. A diagram of this is shown in Figure 3.2.

For purposes of basic explanation, we should take the word “form” here to mean “the words”. Hence, in analysis phase, we are trying to find the unambiguous, underlying words from the text; and in the synthesis phase we are attempting to create a speech signal from these words. In the basic common form model, we always read the words as they are encoded in the text; every word is read in the same order we encounter it. No understanding or other higher level analysis is performed, and as such everything is read with a neutral prosody.

The key features of this model are;

- There are two fundamental processes; text analysis and speech synthesis.

- The task of analysis is to find the form, ie. the words, from the text.
- The task of synthesis is to generate the signal from this form.
- No understanding is required, an accurate decoding of the text into a linguistic form is sufficient.
- Prosodic information can not in general be decoded from the text, and so everything is spoken with neutral prosody.

3.3.2 Other models

A number of different models are found in TTS systems and we will briefly examine these now. These models are not mutually exclusive and many real life systems are a combination of two or more models.

Signal-to-signal model In this model, the process is seen as one of converting the written signal into a spoken one directly. In such models, the process is not seen as one of uncovering a linguistic message from a written signal, and then synthesising from this, but as a process where we try and directly convert the text into speech. In particular, the system is not divided into explicit analysis and synthesis stages.

Pipelined models Quite often the signal-to-signal model is implemented as a pipelined model where the process is seen as one of passing representations from one module to the next. Each module performs one specific task such as part-of-speech tagging, or pause insertion and so on. No explicit distinction is made between analysis and synthesis tasks. These systems are often highly modular, such that each module's job is defined as reading one type of information and producing another. Often the modules are not explicitly linked so that different theories and techniques can co-exist in the same overall system.

Text as language models In this model, the process is seen as basically one of synthesis alone. The text itself is taken as the linguistic message, and synthesis is performed from this. As the text is rarely clean or unambiguous enough for this to happen directly, a **text normalisation** process is normally added, as a sort of pre-processor to the synthesis process itself. The idea here is that the text requires "tidying-up" before it can be used as the input to the synthesiser.

Grapheme and phoneme form models This approach is in many ways similar to the common form model in that first a grapheme form of the text input is found, and this is then converted to a phoneme form for synthesis. Words are not central to the representation as is the case in the common form model. This approach is particularly attractive in languages where the grapheme-phoneme correspondence is relatively direct; in such languages finding the graphemes often means the phonemes and hence pronunciation can accurately be found. For other languages such as English, this is more difficult, and for languages such as Chinese this approach is probably impossible. If it can be performed, a significant advantage of the grapheme/phoneme form model is that, an exhaustive knowledge of the words in a language

is not necessary; little or no use is needed for a lexicon. This fact makes this approach still attractive for small footprint systems.

Full linguistic analysis models The common form model is based on the idea that all we need to uncover from the text is primarily the identity of the words; we don't make heavy use of any other linguistic form or meaning representations. Some systems go much further in terms of linguistic analysis and perform morphological analysis, part-of-speech tagging and syntactic parsing. To some extent, all these are useful for finding the words; the issue is really whether these should be performed as separate stand-alone tasks whose output feeds in to the word finding system, or whether we can find the words with a single, integrated approach. This issue is addressed in detail in Chapter 5. In addition to word identity detection, parsing and other types of linguistic analysis are often seen as being useful for helping with prosody.

Complete prosody generation The common form model allows us to describe the three types of prosody independently. Suprasegmental "prosody" is modelled by the verbal component, and for utterances which have little affective prosody. This means that only augmentative prosody has to be explicitly generated. This contrasts with many systems which state that the F0, phrasing, stress and so on in an utterance are all directly determined by prosody, and so to generate speech we have to generate all these quantities with an explicit prosodic model. If this is the case, then the prosodic part of the system plays a much bigger role.

Prosody from the text Following from the assumptions of the complete prosody model, we find that if every utterance requires a detailed prosodic specification, then this must somehow be generated. A common assumption is that the text does in fact contain enough information to determine prosody, and so many TTS systems have modules which try and predict prosodic representations directly from the text, often with the assumption that this is an analysis process with a right and wrong answer.

3.3.3 Comparison

While some of the above systems may seem quite different from the common form model that we adopt, it should be realised that in many cases the differences can be somewhat superficial and that "under-the-hood" many of the same underlying operations are being performed. It formulates text analysis as a decoding problem, and this allows us to bring many of the powerful statistical text processing techniques to bear on the problem. This is a more powerful approach than the alternatives of either treating text as the synthesis input, which in some way must be normalised; or the approach of performing a full traditional linguistic analysis, which can often be errorful while the reason why we have chosen the common form model is that it simplifies the overall design of the system. The text decoding approach allows us to focus on a single well defined problem such that we are often able to perform this in a single, optimised system. Secondly, by avoiding the "complete prosody" assumption, we bypass many of the difficulties (and we believe impossibilities) that arise from trying to determine the prosodic content from the text.

The common form approach is what one can also term an *integrated* approach. In many

TTS systems, one often finds a huge list of modules arranged in a pipeline; a typical system of this kind might include modules for preprocessing, tokenisation, text normalisation, part-of-speech tagging, parsing, morphological analysis, lexicon, post-lexical rules, intonation, phrasing, duration, F0 generation, unit selection and signal processing. The problem with this approach is that if each module operates independently, it may only be marginally contributing to the overall process. Each module will have to make “hard decisions” to generate its output, and this can lead to a propagation of errors through the system. Furthermore, from an analysis of errors and performance, it may in fact be clear that many of these modules (e.g. a parser) would only be of marginal benefit even if their output was always correct. An integrated approach on the other hand aims to avoid these problems by optimising the system performance on a single metric. In the case of text decoding, the idea is to go from raw text to a word list in a single step. While some of the tasks performed by modules in a pipelined system may still be performed, they will be done implicitly, with few hard decisions until the final output is required.

3.4 SYSTEMS

3.4.1 A Simple text-to-speech system

To demonstrate the common form model, let us now sketch how a text-to-speech system actually works. The input text arrives as a sequence of ascii characters, which can be of any length. To make the processing more manageable, we break input text into separate sentences using a sentence splitting algorithm. The input may only have one sentence, but we don’t know this so we always attempt to find sentence boundaries. For each sentence, we then further divide the input into a sequence of tokens, based on the presence of whitespace, punctuation and so on. Often tokens are just the written encodings of individual words, but can also be encodings of numbers, dates and other types. Next we find the semiotic class of each token. For non-natural language tokens, we use a separate system for each type to decode the text into the underlying form, and then using rules, we translate this form into a natural language form with words. For natural language tokens, we attempt to resolve any ambiguity and so find the words. We attempt a basic prosodic analysis of the text. Although much of the information we might ideally like is missing from the text, we do the best we can and use algorithms to determine the phrasing, prominence patterns and intonational tune of the utterance. This concludes the text and prosodic analysis phase.

The first stage in the synthesis phase is to take the words we have just found and encode them as phonemes. We do this, because this provides a more compact representation for further synthesis processes to work on. The words, phonemes and phrasing form an input specification to the unit selection module. Actual synthesis is performed by accessing a database of pre-recorded speech so as to find units contained there that match the input specification as closely as possible. The pre-recorded speech can take the form of a database of waveform fragments and when a particular sequence of these are chosen, signal processing is used to stitch them together to form a

single continuous output speech waveform.

This is essentially how (one type) of modern TTS works. One may well ask why it takes an entire book to explain this then, but as we shall see, each stage of this process can be quite complicated, and so we give extensive background and justification for the approaches taken. Additionally, while it is certainly possible to produce a system that speaks *something* with the above recipe, it is considerably more difficult to create a system that consistently produces high quality speech no matter what the input is.

3.4.2 Concept to speech

It is not strictly speaking necessary to start with text as input in creating synthetic speech. As we have clearly stated, our TTS model is really two processes, an analysis one followed by a synthesis one, and these are quite different in nature. As the analysis system will never be perfect, it is quite reasonable to ask whether there are situations where we can do away with this component and generate speech “directly”. Often this way of doing things is called **concept-to-speech** in contrast to text-to-speech. We shall stick with the generally used term “Concept-to-speech”, but we should point out that this can however mean a number of different things which we will now explain.

Clearly, in our model, if we do away with text analysis, then the input to the system must be the input to the synthesiser itself, that is, the word-based form or message. We could call this “message-to-speech” or “form-to-speech” synthesis. The idea is that a message-to-speech system would be a component in a larger computer natural language system, and that the message would be generated by a natural language generation system. As this system knows whether it said PROJECT-NOUN or PROJECT-VERB, it can pass this directly into the synthesiser and so bypass any errors the text analysis system may make here. While this would certainly bypass any inadequacies of text analysis, it does suffer from a significant problem in that we don’t have a standard definition of what form actually is. Even with systems that conform to the common-form model, each may have a slightly different idea of what this form should take; one system may just use words, another may use syntax as well. We therefore face a situation of each system potentially exposing a quite different interface to an external system developer. A standard interface could be designed, but standardisations of systems often progress along a lowest common denominator strategy and may end up being somewhat ineffective.

An alternative strategy is to propose a “meaning-to-speech” system, in which semantic representations are the input. Here the message generation and speech encoding parts would be combined into a single system. While this bypasses the nature of how the message should look, it opens up an even thornier problem in that we have even less agreement about how semantic representations should look. Such a system would have a significant advantage in that it could solve the prosody generation problem. As we are synthesising directly from meaning, it would in principle be possible to generate prosody in the signal directly, and so bypass problems of prosodic impoverishment in the text, or problems with how to represent prosodic form.

A final solution is that we use a TTS system as before, but augment the input so as to reduce

ambiguity and explicitly show the system where to generate prosodic effects. This can be done by a number of means, including the use of XML or other markup (explained in Chapter 17). While this will inevitably lack the power and fine control of the two above methods, it has the advantage in that the input is in a more standard form, such that a system developer should be able to get this working and switch from one system to another. It also has the advantage in that one can easily vary how close the system is to “raw” text or “clean” text.

The message-to-form and augmented-text approaches are well covered in this book in the sense that they can be adapted from our main common-form TTS system. Meaning-to-speech is considerably more difficult and has only been researched in a very tentative manner to date. Mainly for this reason, we do not cover this approach in any detail here.

3.4.3 Canned Speech and Limited Domain Synthesis

Today, most applications which use speech output do not use text-to-speech. Rather they make use of a set of recordings which are simply played back when required. These are called **canned speech** or **pre-recorded prompts** depending on the application. A typical way these are deployed is that a designer creates an exhaustive list of the utterances that are needed for an application. Quite often these utterances are specific to that application and a new set will be created and recorded for a different application. Once the utterances have been designed, a speaker is asked to read these, they are recorded and then stored as waveform files. The application then plays back the required file at the required time. A typical system of this kind might be deployed in a train station announcement system or an automated credit card booking system, where the users interact with the system by speech recognition and automated dialogue control.

Such systems initially compare unfavorably with TTS in that they require a new set of utterances for each application, compared to a TTS system which would just be deployed once and can be used for any application. Furthermore, the canned speech approach can only say a very fixed number of things, which can limit the scope of the application (for instance it would be very difficult to speak a user’s name). Finally, if the application has to be updated in some way and new utterances added, this requires additional recordings which may incur considerable difficulty if say the original speaker is unobtainable. Despite these apparent disadvantages, canned speech is nearly always deployed in commercial systems in place of TTS. Part of the reason behind this is technical, part cultural. Technically canned speech is perfectly natural and as users show extreme sensitivity to the naturalness of all speech output this factor can outweigh all others. In recent years, TTS systems have improved considerably in terms of naturalness, and so it is more common to find TTS systems in these applications. There are other non-technical reasons; canned speech is seen as a simple, low-tech solution whereas TTS is seen as complex and hi-tech. The upshot is that most system designers feel that they know where they stand with canned speech, whereas TTS requires some leap of faith. There may be purely business reasons also; while the canned speech approach incurs up front cost in an application, it is a one off cost, and does not increase with the size of deployment. TTS systems by contrast can often be sold like normal soft-

ware where extra licences (and hence cost) is required when the size of the deployment increases.

The main technical drawback of canned speech is that it can only say a fixed number of things. This can be a severe drawback in even simple applications where for example a telephone number is to be read in an answer machine application. A common solution is to attempt to splice together recordings of individual words or phrases so as to create new utterances. The result of such operations varies greatly from acceptable (but clearly spliced) to comically awful. That said, even the resultant poor naturalness of this is often chosen over TTS in commercial situations.

Faced with the choice between fully natural but inflexible canned speech, and somewhat unnatural but fully flexible TTS, some researchers have proposed **limited domain synthesis** systems which aim to combine the benefits of both. There are as many different approaches to this. Some systems attempt to mix canned speech and TTS. Black and Lenzo [52] proposed a system for cleverly joining words and carrier phrases, for use in applications such as a talking clock. The **phrase splicing** approach of Donovan et al [139] used recorded carrier phrases and cleverly spliced with unit selection synthesis from recordings of the same speaker. A somewhat different approach is to use what is basically the same system as for normal unit selection synthesis but to load the database of recordings with words and phrases from the required domain [440], [436], [394], [9].

3.5 KEY PROBLEMS IN TEXT-TO-SPEECH

In this section we identify the main challenges in high quality text-to-speech. First we describe four areas which we believe are the key problems that have been recognised as the main goals of TTS system building (but note, the problems are not always described in this way). Then we go on to describe two additional problems, which are starting to attract more attention and which we believe will be central to TTS research in the future.

3.5.1 Text classification with respect to semiotic systems

As we explained, it is a mistake to think that text is simply or always an encoding of natural language. Rather, we should see text as a common physical signal that can be used to encode many different semiotic systems, of which natural language is just one (rather special) case.

There are two main ways to deal with this problem. The first is the **text-normalisation** approach, which sees the text as the input to the synthesiser and tries to rewrite any “non-standard” text as proper “linguistic” text. The second is to classify each section of text according to one of the known semiotic classes. From there, a parser specific to each classes is used to analyse that section of text and uncover the underlying form. For natural language the text analysis job is now done; but for the other systems an additional stage is needed, where the underlying form is translated into words.

Let us consider the semiotic class approach. Assume for a moment that we can divide an input sentence into a sequence of text tokens, such that the input sentence

(4) Tuesday September 27, 10:33 am ET

would be tokenised as

(5) Tuesday
September
27,
10:33
am
ET

Semiotic classification is therefore a question of assigning the correct class to each of these tokens. This can be done based on the patterns within the tokens themselves (e.g. three numbers divided by slashes (e.g. 10/12/67) is indicative of a date) and optionally the tokens surrounding the one in question (so that if we find “1967” preceded by “in” there is a good chance that this is a year).

Once the class has been identified, it is normally quite straightforward for classes other than natural language to analyse this token to find its underlying form. In dealing with times for instance, the text token might say 10:45am and from this we would find a form something like (hours=10, minutes=45, time_of_day=morning). Finally, we can **translate** this into words. In one sense this is often straightforward as there are often simple well-accepted rules which describe this. In our time example, the word version could be read as TEN FORTY FIVE A M, and it is fairly easy to see how the form-to-words rules would be written for this example. Two issues complicate the matter however. Firstly, often there are many different legitimate ways to translate the same form; with our time example we could have

- (6) a. A QUARTER TO ELEVEN
- B. A QUARTER OF ELEVEN
- C. A QUARTER TO ELEVEN IN THE MORNING
- D. TEN FORTY FIVE IN THE MORNING
- E. TEN FORTY FIVE
- F. TEN FORTY FIVE A M

and so on. All are legitimate, but as the system can read only one, we have to be sensitive to which one this should be. Sometimes preferences depend on the user; in particular a British English speaker may prefer one, an American speaker another. In addition, the context and genre of the text may have an effect; in an airline reservation system it would be useful to make absolutely sure the user knows the time is in the morning, in a business appointment system this may be obvious.

The important point to realise is that there is often no single correct answer to the translation issue, and so systems need to be sensitive to the user expectations.

A further difficulty with translation is that for some content there is no commonly agreed rendering in words. While times, dates and currencies normally present no problem, classes such as email addresses, computer programs and other technical information has no agreed reading. For example, taking the email address:

(7) pat40@cam.ac.uk

We can easily parse this into

<i>user</i>	=	pat40
<i>top level domain</i>	=	uk
<i>secondary domain</i>	=	ac
<i>local domain</i>	=	cam

In effect as all email programs do just this. The problem is how should we read this? Should **pat40** be read P A T FOUR ZERO, P A T FORTY, or PAT FORTY?

In summary then, there are three tasks involved here.

1. Identify which semiotic class a token belongs to. This is often the most difficult task.
2. Analyse each token to find its underlying form. Apart from the case of natural language, this is often quite easy as the semiotic systems are cultural inventions and have been designed to be easy to understand.
3. Translate into natural language. It is often quite easy to generate at least *one* legitimate translation but sometimes difficult to generate the particular version that a user expects.

3.5.2 Decoding natural language text

While most semiotic systems have a fairly simple and direct signal-to-form mapping, this in general is not so for natural language. Once the text has been properly classified in terms of its semiotic class, the general level of ambiguity with respect to form is greatly reduced; it is only in the area of natural language where significant ambiguity still remains.

Ambiguity occurs in many different guises. The most basic is **homograph** ambiguity, where two words share the same form. In cases such as

(8) lets now project the image

and

(9) the project was running behind schedule

we see that two separate words PROJECT-VERB and PROJECT-NOUN both share the same letter sequence, but have different meanings and different pronunciations. In addition to homograph ambiguity, **syntactic ambiguity** is also reasonably common. In the example (from [363])

(10) Police help dog bite victim

we (at least) two different possible syntactic patterns that could give rise to these tokens:

- (11) a. (POLICE HELP DOG) BITE VICTIM
b. POLICE HELP (DOG BITE VICTIM)

The question when building a TTS system is just how much of this ambiguity we actually need to resolve. A comprehensive approach would attempt to resolve all the ambiguity in the text and create a rich, structure for the form. In recognition that the more we attempt the more likely we are to make an error, and also in recognition of the basic engineering principle of efficiency, we could also take a minimalist approach, which would adopt the principle that we should only resolve those ambiguities that affect pronunciation. The word SET is used in many different ways in English, but it is always spelt the same and always pronounced the same, and so we can ignore the multitude of meanings and uses. If we know that some text could be two or more underlying forms, but realise that all these forms are spoken the same way, then we shouldn't bother resolving the ambiguity.

3.5.3 Naturalness

We take the view that a major goal of TTS research is to make the system as **natural** sounding as possible. By natural we mean that the system should sound just like a human; and so eventually we would like our TTS system to be able to create the full range of human speech.

Sometimes the question is raised as to whether we really want a TTS system to sound like a human at all. The concern is raised that if the system sounds too much like a human listeners will get confused and mistake it for a person. We think that in reality this is highly unlikely. All the time we hear "perfect" recordings of people's voices, view photographs, watch films and so on, and are rarely confused. In fact, most of us are used to hearing recorded voices on the telephone in the form of people's answering machines, introductory messages or recorded prompts in IVR systems. Rarely do these confuse us. So no matter how good a system is, it will rarely be mistaken for a real person, and we believe this concern can be ignored.

We still have to ask however whether people really *want* natural sounding voices: wouldn't they be happy with a "robotic" sounding one instead? (which would after all be much easier for us to produce). This is really an empirical question that needs to be resolved from user trials. From what we know however, the answer is over whelming in that people give much higher satisfaction and acceptance ratings to natural sounding systems. In fact we can go so far as to say that most listeners are extremely intolerant of unnaturalness to the extent that they will refuse to use non-natural sounding systems regardless of what other benefits are provided. Quite why this should be so is somewhat of a mystery; after all in many situations we are quite happy with visual caricatures of people as evidenced by cartoons and animation. Somehow we don't consider a drawing of Homer Simpson a pathetic rendition of reality, but we are insistent on him having a natural voice, such that in the Simpsons (and all other cartoons) a real actor is used for his voice.

What then makes one voice sound more natural than another? Again, this is a question that we are not in a position to answer fully, but we can provide a sketch of the factors involved. Firstly, any system that produces obviously non-human artefacts in the speech will readily be judged as unnatural. It is all too easy to generate speech with pops, clicks, buzzes and an endless variety of other mechanical sounds. Even speech that is absent from any of these “error” type of sounds can readily sound unnatural. While speech exhibits considerable variability in the realisation of a particular phoneme, this is controlled by a number of factors including phonemic context, position in sentence, supra-segmental influence and so on. Getting this variation just right is a key task. Beyond this, we can consider a number of issues which are related to individuality. In normal speech, any given instance of speech is obviously speech *from someone*; there is no such thing as speaker-neutral speech. So in a sense, we are completely conditioned to the fact that speech comes from people, all of whom sound slightly different. So when listening to a synthesiser, one naturally thinks that it is from someone, and to sound natural, this someone must have a believable sounding voice. This aspect of naturalness has virtually nothing to do with the message being spoken, it is entirely a property of the system that encoded the speech. Hence we must make the synthesiser sound like someone, either a copy of a real person’s voice, or the creation of a new voice that could pass for a real person unproblematically.

3.5.4 Intelligibility: encoding the message in signal

The final central task is **intelligibility**. For our purposes, we define intelligibility as the ability of a listener to decode the message from the speech. As such, this does not include any measure of comprehension directly, but the idea is that if the listener can decode the message properly they should be able to understand it with the same ability as they could the same message spoken by a human. Because of that we can sometimes measure intelligibility with comprehension tests; this is usually done with content that may be hard to decode, but is often not difficult to understand (e.g. a simple sentence address or telephone number).

Of all the TTS problems, intelligibility is probably the easiest to solve. In fact, it is possible to argue that this problem was solved some time ago, as was possible to generate fairly intelligible sounding speech in TTS systems from the late 1970s onwards. In fact, it can be shown in tests that the intelligibility of a modern TTS system is often only marginally better than much older systems such as MITalk [10], once any text errors have been dealt with². From this it is wrong to conclude that little progress has been made on this issue. Crudely speaking, the story is that while the intelligibility of older systems was reasonably good, research focused on improving naturalness which was uniformly rated as being very poor with these systems. Hence the idea in the research community (though rarely stated as such) was to improve naturalness without making intelligibility any worse.

² Today’s systems are massively more accurate with respect to text analysis, and so any text processing errors in systems such as MITalk have to be eliminated if a comparison of synthesis encoding with respect to intelligibility is to be performed.

The relationship is by no means deterministic or simple, but it is possible to see that in many synthesis paradigms there is an inverse correlation between naturalness and intelligibility. To achieve high intelligibility, formant systems and early diphone systems used very “safe” speech, which was well articulated and close to the average realisation for the message in question. In a sense, this speech was too safe, and containing none of the fluency, slurring, rhythm and heavy coarticulation found in real speech. As these effects are added, there is a chance that they will be added inappropriately, and in so doing will actually cause a loss of intelligibility. The trick therefore is to add in extra naturalness by synthesising variation more appropriately, but to do so in a way that does not cause unnatural variation to occur.

3.5.5 Auxiliary generation for prosody

We now consider the first of the more advanced problems; specifically how to create speech that encodes the full range of prosodic effects and not just the neutral prosody of most of today’s systems. Recall that this is difficult because while the text encodes the verbal message that the author generated, it does not do the equivalent with prosody. The author of course (subconsciously) knows this. As such, it is not the case that he generates a message with verbal and prosodic content as he would do if speaking, rather he generates a message that only contains verbal content; the prosody was neither generated or encoded, and so it we can never recover this from the writing. How then are we to generate prosodic content if we can’t uncover it from the text?

Firstly, let us separate the issue of generating a signal which contains a particular prosodic effect (for example surprise, or accentuation of a particular word) from the issue of deciding how to do this automatically from the text. The issue of the realisation of a particular prosodic form is by no means easy, but it is certainly more tractable than the second issue. In essence it is no different from other parts of the synthesis problem in that we can collect data with this effect, study how supra-segmental features vary with respect to this effect and so on.

It is really the second issue; that of how to determine automatically which prosodic effects to use from the text that is the more difficult problem. We term this process **auxiliary generation** to show that it is a generation (and not text decoding) issue, and that this is done as a separate, auxiliary task to the process of verbal decoding and encoding. How then are we to approach this problem of auxiliary generation? First let us state that just because there is no definite right or wrong, this doesn’t imply an “anything goes” policy. One would not think highly of a system that read a message “we are regret to inform you of the sad demise of Mr...” with a humorous or flippant paralinguistic tone. Furthermore, we probably wouldn’t think too highly of a system that placed emphasis as follows “THE share price is AT and all TIME high because of uncertainties caused BY the forecast of PROFESSOR Merton...” and so on. Now there is nearly always some context which we can construct that makes the emphasis of any particular word seem reasonable. However, using the same knowledge and context that we used to decode the writing, we can see that some emphasis patterns seem more appropriate than others. This leads us to the proposal that the approach to prosody generation should be one of **assumed intent**, which means that we

have to make a *guess* as to what the author *might* have specified in the structured representation if he had intended to speak rather than write the message. This is by no means a certain or easy task but it does give us a reasonable basis on which to solve this problem. We should make two final points about this. Firstly, as we are making an assumption as to what the author might have done, there is no clear right or wrong about this task. Secondly, we have to from the outset build the notion of generation choice into our model of auxiliary completion: just as the author can choose to represent a thought in a number of different ways, we have to assume that he could have exercised choice over the prosodic and paralinguistic components. Therefore, as well as being uncertain because we are making an assumption, it is also perfectly valid to assume that there are a number of equally valid choices.

3.5.6 Adapting the system to the situation

The issue here is that unlike the “normal” communicative situations of reading silently or having a conversation; in the case of reading aloud we have three agents; the author, the listener and the reader (i.e. the TTS system). Most text was never written with the intention that it be read aloud, and because of this, problems can occur in that faithful readings of the text can lead to situations where the reader is speaking something that the listener can not understand, has no knowledge of or is not interested in hearing. When people take on the role of the reader, they often stray from the literal text, using explanatory asides, paraphrasing, synonyms and other devices to make the author’s message understood. Just how faithful the reader is to the text depends very much on the situation; we sometimes describe a good reader as someone who has a good awareness of both the listener and the intent of the author and as such can steer a good compromise between faithfulness and acceptability.

Very few TTS systems themselves make any serious attempt at solving this issue; it is normally seen as a problem for the application to resolve such issues. Even so, it is possible for the TTS to facilitate better control; this can be done by providing natural and easy to control ways to change the speed of the speaking, to have a phrase repeated, a phrase skipped and so on.

3.6 SUMMARY

Speech and writing

- We can consider both speech and writing/text as signals which encode a linguistic, or other, message.
- The primary functional difference is that writing is an invention used to record messages; as such it is significantly “closer” to the message than is the case with speech.
- Most writing is written with the intention that the reader reads it silently. This can present difficulties if this writing is then read aloud.

- Not all writing encodes linguistic messages, it is used to encode messages in many other semiotic systems also.

Reading aloud

- To a large extent, we can perform the task of reading aloud, by decoding a written signal into a message, and then re-encode this into a speech signal.
- In many cases, no prosody is required, but in others a more sophisticated approach is to attempt to generate prosody that is appropriate for the message. This can not however be decoded from the text as that information was never encoded there in the first place.
- These observations lead us to build our system within the framework of the common form model. This has two fundamental processes.
 1. Text analysis; a decoding process that finds the message from the text
 2. Speech synthesis; an encoding process that creates a signal from the message.

Text to Speech Key Challenges

- We can identify four main challenges for any builder of a TTS system:
 1. Semiotic classification of text
 2. Decoding natural language text
 3. Creating natural, human sounding speech
 4. Creating intelligible speech
- We can also identify two current and future main challenges
 1. Generating affective and augmentative prosody
 2. Speaking in a way that takes the listener's situation and needs into account.