High-Level and Low-Level Synthesis

1.1 Differentiating Between Low-Level and High-Level Synthesis

We need to differentiate between low- and high-level synthesis. Linguistics makes a broadly equivalent distinction in terms of human speech production: low-level synthesis corresponds roughly to phonetics and high-level synthesis corresponds roughly to phonology. There is some blurring between these two components, and we shall discuss the importance of this in due course (see Chapter 11 and Chapter 25).

In linguistics, phonology is essentially about planning. It is here that the plan is put together for speaking utterances formulated by other parts of the linguistics, like semantics and syntax. These other areas are not concerned with speaking—their domain is how to arrive at phrases and sentences which appropriately reflect the meaning of what a person has to say. It is felt that it is only when these pre-speech phrases and sentences are arrived at that the business of planning how to speak them begins. One reason why we feel that there is somewhat of a break between sentences and planning how to speak them is that those same sentences might be channelled into a part of linguistics which would plan how to write them. Diagrammed this looks like:

```
graphology writing plan
```

The task of phonology is to formulate a speaking plan, whereas that of graphology is to formulate a writing plan. Notice that in either case we end up simply with a plan—not with writing or speech: that comes later.

1.2 Two Types of Text

It is important to note the parallel between graphology and phonology that we can see in the above diagram. The apparent equality between these two components conceals something very important, which is that the graphology pathway, leading eventually to a rendering of the sentence as writing, does not encode as much of the information available...
at the sentence level as the phonology does. Phonology, for example, encodes prosodic information, and it is this information in particular which graphology barely touches.

Let us put this another way. In the human system graphology and phonology assume that the recipient of their processes will be a human being—it has been this way since human beings have written and spoken. Speech and writing are the result of systematic renderings of sentences, and they are intended to be decoded by human beings. As such the processes of graphology and phonology (and their subsequent low-level rendering stages: ‘graphetics’ and phonetics) make assumptions about the device (the human being) which is to input them for decoding.

With speech synthesis (designed to simulate phonology and phonetic rendering), text-to-speech synthesis (designed to simulate human beings reading aloud text produced by graphology/graphetics) and automatic speech recognition (designed to simulate human perception of speech) such assumptions cannot be made. There is a simple reason for this: we really do not yet have adequate models of all the human processes involved to produce other than imperfect simulations.

Text in particular is very short on what it encodes, and as we have said, the shortcomings lie in that part of a sentence which would be encoded by prosodic processing were the sentence to be spoken by a human being. Text makes one of two assumptions:

- the text is not intended to be spoken, in which case any expressive content has to be text-based—that is, it must be expressed using the available words and their syntactic arrangement;
- the text is to be read out aloud; in which case it is assumed that the reader is able to supply an appropriate expression and prosody and bring this to the speech rendering process.

By and large, practised human readers are quite good at actively adding expressive content and general prosody to a text while they are actually reading it aloud. Occasionally mistakes are made, but these are surprisingly rare, given the look-ahead strategy that readers deploy. Because the process is an active one and depends on the speaker and the immediate environment, it is not surprising that different renderings arise on different occasions, even when the same speaker is involved.

1.3 The Context of High-Level Synthesis

Rendering processes within the overall text-to-speech system are carried out within a particular context—the prosodic context of the utterance. Whether a text-to-speech system is trying to read out text which was never intended for the purpose, or whether the text has been written with human speech rendering in mind, the task for a text-to-speech system is daunting. This is largely down to the fact that we really do not have an adequate model of what it is that human readers bring to the task or how they do it. There is an important point that we are developing throughout this book, and that is that it is not a question of adding prosody or expression, but a question of rendering a spoken version of the text within a prosodic or expressive framework. Let us term these the additive model and the wrapper model. We suggest that high-level synthesis—the development of an utterance plan—is conducted within the wrapper context.
Conceptually these are very different approaches, and we feel that one of the problems encountered so far is that attempts to add prosody have failed because the model is too simplistic and insufficiently reflective of what the human strategy is. This seems to focus on rendering speech within an appropriate prosodic wrapper, and our proposals for modelling the situation assume a hierarchical structure which is dominated by this wrapper (see Chapter 34).

Prosody is a general term, and can be viewed as extending to an abstract characterisation of a vehicle for features such as expressive or, more extremely, emotive content. An abstract characterisation of this kind would enumerate all the possibilities for prosody, and part of the rendering task would be to highlight appropriate possibilities for particular aspects of expression. Notice that it would be absurd to attempt to render ‘prosody’ in this model (since it is simultaneously everything of a prosodic nature), just as it is absurd to try to render syntax in linguistic theory (since it simultaneously characterises all possible sentences in a language). Unfortunately some text-to-speech systems have taken this abstract characterisation of prosody and, calling it ‘neutral’ prosody, have attempted to add it to the segmental characterisation of particular utterances. The results are not satisfactory because human listeners do not know what to make of such a waveform: they know it cannot occur.

Let us summarise what is in effect a matter of principle in the development of a model within linguistic theory:

- Linguistic theory is about the knowledge of language and of a particular language which is shared between speakers and listeners of that language.
- The model is static and does not include—its strictest form—processes involving drawing on this knowledge for characterising particular sentences.

As we move through the grammar toward speech we find the linguistic component referred to as phonology—a characterisation for a particular language of everything necessary to build utterance plans to correspond with the sentences the grammar enumerates (all of them). Again there is no formal means within this type of theory for drawing on that knowledge for planning specific utterances.

Within the phonology there is a characterisation of prosody, a recognisable sub-component—intonational, rhythmic and prominence features of utterance planning. Again prosody enumerates all possibilities and, we say again, with no recipe for drawing on this knowledge. Although prosody is normally referred to as a sub-component of phonology, we prefer to regard phonological processes as taking place within a prosodic context: that is, prosodic processes are logically prior to phonological processes. Hence the wrapper model referred to above.

Pragmatics is a component of the theory which characterises expressive content—along the same theoretical lines as the other components. It is the interaction between pragmatics and prosody which highlights those elements of prosody which associate with particular pragmatic concepts. So, for example, the pragmatic concept of anger (deriving from the bio-psychological concept of anger) is associated with features of prosody which when combined uniquely characterise expressive anger. Prosody is not, therefore, ‘neutral’ expression; it the characterisation of all possible prosodic features in the language.
1.4 Textual Rendering

Text-to-speech systems, by definition, take written text which would have been derived from a writing plan devised by a 'graphology', and use it to generate a speaking plan which is then spoken. Notice from the diagram above, though, that human beings obviously do not do this; writing is an alternative to speaking, it does not precede it. The exception, of course, is when human beings themselves take text and read it out loud. And it is this human behaviour which text-to-speech systems are simulating.

We shall see that an interesting problem arises here. The operation of graphology—the production of a plan for writing out phrases or sentences—constitutes a 'lossy' encoding process: information is lost during the process. What eventually appears on paper does not encode all of a speaker’s intentions. For example, the mood of the writer does not come across, except perhaps in the choice of particular words. The mood of a speaker, however, invariably does come across. Immediately, though, we can observe that mood (along with emotion or intention) could not have been encoded in the phrase or sentence except via actual words—so much of what a third party detects of a person’s mood is conveyed by tone-of-voice. It is not actually expressed in the sentence to begin with.

The human system gets away with this lossy encoding that we come across in written text because human readers in general find no difficulty in restoring what has been removed—or at least some acceptable substitute. For example, compare the following two written sentences:

It was John.
It wasn’t Mary, it was John.

The way in which the words It was John are spoken differs, although the text remains the same. No native speaker of English who can also read fails to make this difference. But a text-to-speech system struggles to add the contrastive emphasis which a listener would expect. This is an easy example—it is not hard to imagine variations in rendering an utterance which are much more subtle than this.

Some researchers have tried to estimate what is needed to perform this task of restoring semantic or pragmatic information at the reading aloud stage. And to a certain extent restoration is possible. But most agree that there are subtleties which currently defeat the most sophisticated algorithms because they rest on unknown factors such as world knowledge—what a speaker knows about the world way beyond the current linguistic context.

The above diagram is therefore too simple. There is a component missing—something which accounts for what a speaker, in planning and rendering an utterance, brings to the process which would not normally be encoded. A more appropriate diagram would look like this:
In linguistics, much of a person’s expression (mood, emotion, attitude, intention) is characterised by a component called **pragmatics**, and it is here that the part of language which has little to do with choice of actual words is formulated. Pragmatics has a direct input into phonology (and phonetics) and influences the way in which an utterance is actually spoken.

Pragmatics (Verschueren 2003) is maturing late. The semantic and syntactic areas matured earlier, as well as phonology (without reference to expression—the phonology of what has been called the *neutral* or expressionless utterance plan). It was therefore not surprising that the earlier speech technology models, adopted in speech synthesis and automatic speech recognition, were not sensitive to expression—they omitted reference to pragmatics or its output. We reiterate many times in this book that a major reason for the persistent lack of convincing naturalness in speech synthesis is that systems are based on a pragmatics-free model of linguistics.

A pragmatics-free model of linguistics fails to accommodate the variability associated with what we might call in very general terms expression or tone-of-voice. The kinds of things reflected here are a speaker’s emotional state, their feelings toward the person they’re speaking to, their general attitude, the environmental constraints which contribute to the choice of style for the speech, etc. There are many facets to this particular problem which currently preoccupy many researchers (Tatham and Morton 2004).

One of the tasks of this book will be to introduce the theoretical concepts necessary to enable an expression information channel to link up with speech planning, and to show how this translates into a better plan for rendering into an actual speech soundwave. Although this book is not about automatic speech recognition, we suggest that consideration of expression, properly modelled in phonology and phonetics, points to a considerable improvement in the performance of automatic speech recognition systems.