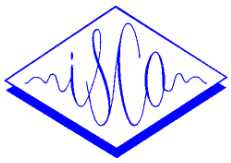

Second ESCA/IEEE Workshop
on Speech Synthesis
Mohonk, New Paltz, NY, USA
September 12-15, 1994

THE BT LAUREATE TEXT-TO-SPEECH SYSTEM

A P Breen

Abstract

Over the last four years BT laboratories have been developing a sophisticated text to speech system, called Laureate. The approach adopted in the Laureate system has been to maximise the usability of the system for differing applications, while retaining an effective research and development tool.
During the design phase of the system, a number of specific requirements were identified, prominent amongst these being the desire to ensure good speech quality and naturalness, but not at the cost of intelligibility.

Introduction

For centuries scientists have been struggling to generate synthetic speech. The very early attempts were based on mechanical analogues of the human vocal tract, but as science and technology progressed these crude mechanical analogues were replaced by sophisticated electronic models of the vocal apparatus. Such models were capable of producing highly natural sounding synthetic speech provided explicit parametric information on the speech signal was available. Unfortunately, these advances in the synthesis of speech were not matched by a corresponding increase in our understanding of how to automatically generate the appropriate parametric information required by these synthesis models. As a consequence, the quality of synthetic speech from text sounded robotic and unnatural. With the advent of the computer revolution, the possibilities for modelling, encoding and storing elements of the speech signal have led to a resurgence of interest in text to speech generation.

Today, the quality of synthetic speech from text has improved to such an extent that real world applications are now viable. As a consequence of this, text to speech synthesis has moved from being a predominantly academic discipline to one of increasing importance to the telecommunications and computer industries. This shift in emphasis has led to a need by industry to develop methods of research and development which enable novel ideas to be investigated, tested and developed while still ensuring that products can be produced rapidly and cost effectively. The solution to this problem as adopted by the synthesis and analysis group at BT is described in this paper.

The Laureate text to speech system

Stated simply there are two problems facing text to speech developers. They are technical constraints and theoretical unknowns. With the recent increase in computing power and the availability of cheap storage, many of the technical constraints have been solved or at least significantly eased. However, there has not been an equal increase in our understanding of the production and perception of speech. A number of the latest developments in text to speech have tended towards encoding ignorance rather than attempting to model the underlying mechanics. This has led to a divergence in system design, those systems which attempt to understand and model the speech signal and those which attempt to encode significant aspects of the speech signal. Neither approach has, as yet, been shown to be clearly superior to the other.

Text to speech systems typically follow one theory or practice whether it be modelling or encoding. In other words the system as a whole is an implementation of a single approach. The Laureate architecture does not make this assumption. It has been designed specifically with the aim of allowing a number of different approaches to co-exist within the same computational framework.

The basic design premise of Laureate v2.0 is shown graphically in figure 1. The different sized balloons represent different shades and types of linguistic theory, which in isolation are incompatible with each other. The purpose of the Laureate core and component interface is to provide a consistent and sufficient minimum set of standard linguistic representations which are common across the different models and production methods, so allowing different theories to coexist. There are a number of potential problems in mixing different theories:

1. Each theory will require its own internal data structures.
2. Different theories may have different and even conflicting linguistic definitions.
3. Different theories will require different amounts of given knowledge.

Figure 2. shows diagramatically, how the Laureate achitecture attempts to address these problems. The Laureate system consists of two main parts, a core component and a number of satellite components. The core represents the backbone of the system. All information within the system passes from one component to another via the linguistic object contained within this core. The linguistic object is a dynamic database of information, which has a simple understanding of the relative hierarchical structure of language. Satellite components can only access the object via a set of formalised linguistic questions or statements. This means that each satellite component is completely isolated from the rest of the system. It may, therefore, have its own internal representational linguistic structure which does not necessarily conform to that of any other component within the system. The formalised linguistic interface ensures that data returned to the linguistic object can be used by other components of the system in a consistent manner. Due to the modular nature of the system components may be placed at different points along the backbone. Thus, components which require large amounts of linguistic knowledge can be placed after components which add such knowledge to the linguistic object.

In addition to the above, the linguistic object design makes it possible to develop different languages within the same framework. Satellite components may change, but the overall architecture will remain the same, considerably reducing development time. To highlight this concept further, consider the following. The current Laureate TTS system is based on a unit inventory approach, however it differs from many other systems in that it sees units as aspects of an underlying language process. In other words the units are referenced using phonological knowledge rather than phonetic and acoustic labels. Prosody is not currently viewed as an appropriate problem of encoding, hence the Laureate system has a set of prosody models which are imposed on the basic elemental sounds selected from the unit inventory. However, should an appropriate method of encoding prosody be developed, the Laureate architecture is capable of supporting this alternative approach without the need to redesign or modify large aspects of the system. This is possible because all language knowledge within the system is stored in the linguistic object.

However "meaning" is distributed throughout the system, i.e. the meaning of an element of the linguistic object is not an attribute of the object but of the satellite components which access it. This makes the Laureate code both flexible and robust to theoretical changes.

So far this section has considered the speech theoretic aspects of the design. However, there are a number of practical benefits of the Laureate v2.0 architecture. These are listed over the page:

1. Divergent theories may be developed and tested under one unifying framework. The modular nature of the design makes it possible to develop and test a specific theory of speech much more quickly than would otherwise be possible. In other words, as the linguistic object is not committed to one particular theory of language it is possible for a new component to use all the information it can from other components without the overhead of having to conform to a given view of language.

2. The concept of a linguistic object means that few assumptions have been made about the type of linguistic information needed, making the system very flexible and so extending the life time of the code. As stated above this is particularly significant when foreign languages are considered. Only those components which are specific to the language being implemented need to be changed.

3. It is possible to include extra components or remove components very quickly reducing development and downstreaming time. As research progresses more advanced features will need to be tested within a TTS framework. This architecture enables researchers to concentrate on developing these new features without the overhead of having to design and integrate such modifications into code which was never designed to support them.

4. As each component can be developed in isolation the cost of upgrading the system once it has been downstreamed is greatly reduced. The design enables part release of code, in other words the cost of downstreaming a new or improved component is predominantly the cost of the component testing.

5. The code has been written in ANSI standard C, which is available on most hardware platforms. Languages other than 'C' may be included in the system provided they are capable of linking with 'C'.

6. Each component in the system may be configured to run as a separately scheduled task.

7. The code can be easily tailored to specific application needs by the addition of extra application specific components.

8. The code has been designed to support multi-channel operation.

The need for assessment

The ability to independently assess the quality of a text to speech system is vital to its successful development. Researchers in this area are aware of an effect known as the "golden ear", this is where the researchers, due to the effect of repeated listening, lose their ability to objectively evaluate the quality of the synthetic speech being produced. Unfortunately assessment remains a complex and expensive enterprise, due to the lack of satisfactory metrics and automated objective methods.

Because of the above, BT has taken care has to develop a set of subjective tests which can be repeatedly used to assess three main aspects of development [1] listed below:

- Whether the system is improving compared to earlier versions
- How similar the synthetic speech is to natural speech
- How well the system is performing in comparison to its competitors.

The metric used in all these tests was the "listening effort" required by subjects to understand a set of short, meaningful independent sentences. This type of test assesses the general effects of the systems on the listener's ease of understanding, rather than any particular domain such as pronunciation etc.The non-diagnostic tests described above are supplemented with a series of informal diagnostic assessments which are not conducted with the same experimental rigour.

Acknowledgements

The Laureate text to speech system has been designed and developed through the combined efforts of a number of individuals, prominent amount these are: Dr. A. Lowry, Ms. M. Gaved, Mr. P. Jackson, Mr. R. Ashworth, Mr. D. Koopman, Mr. P. Deans, Mr. S. Minnis and Mr. M. Edgington.

References

1. Evans, K., "An on-going series of subjective experiments to assess speech output from text-to-speech systems", International Telegraph and Telephone Consultative Committee (CCITT), ITU, TSS, study period 1993-1996, COM 12-18, Question 5.

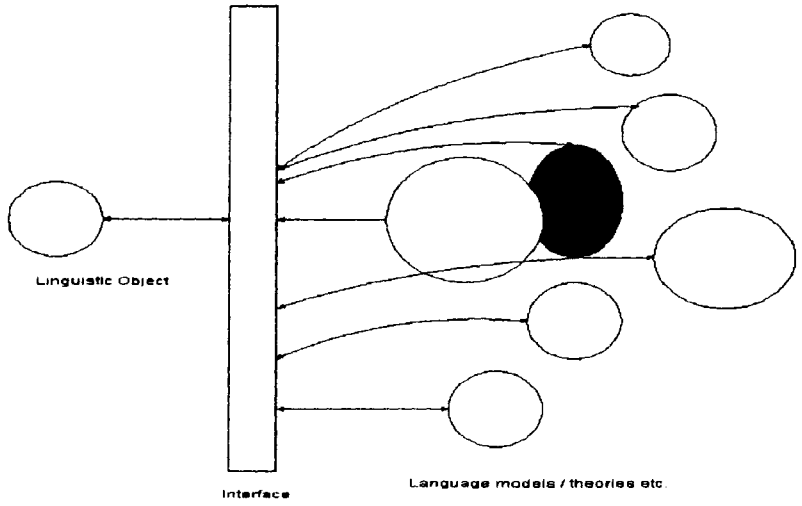

Figure 1.

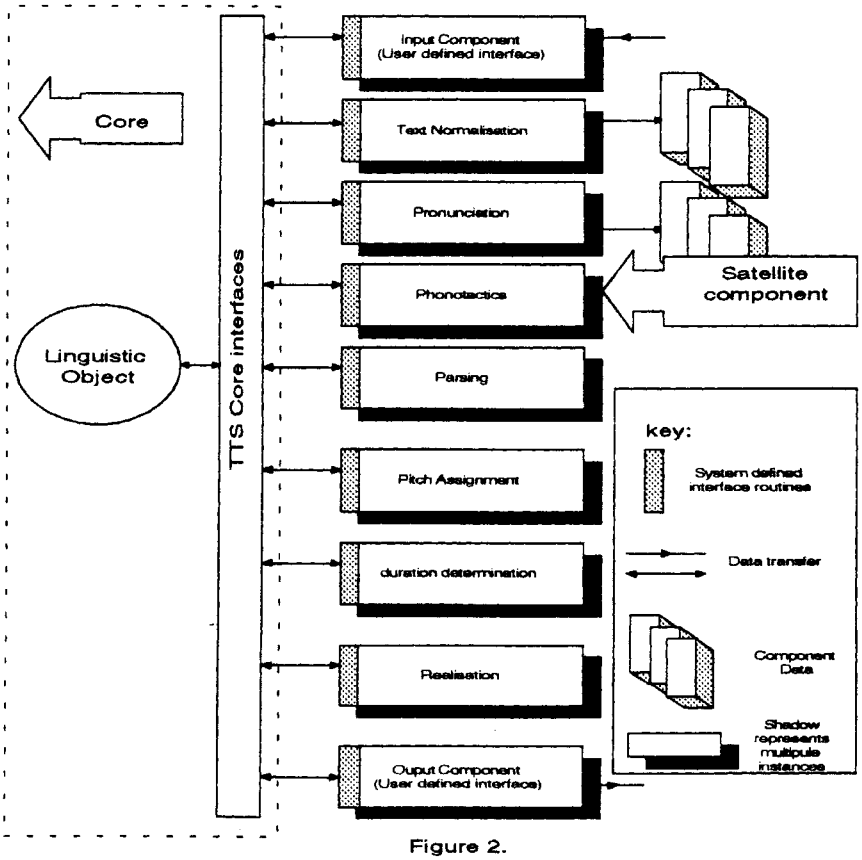

Figure 2.