

A Brief Outline of Aculab TTS: Multilingual TTS for Computer Telephony

Alex Monaghan

Aculab plc, Milton Keynes, UK
Alex.Monaghan@Aculab.com

Abstract

The requirements of the computer telephony (CT) industry place conflicting demands on text-to-speech (TTS) systems. Multilingual functionality and high quality output at telephone bandwidth requires detailed linguistic and acoustic analysis. At the same time, the need for robustness together with a high channel count and small memory footprint means that systems must be extremely efficient and databases must be kept small. We present a system which provides TTS for six languages, with 100 channels of highly natural output on a single DSP card.

1. Introduction

The Aculab multilingual TTS system has been developed over the past five years to provide unparalleled accuracy, quality and efficiency for CT applications. The current version supports six languages (UK and US English, French, German, Dutch, and Latin American Spanish), with multiple voices configurable for each language. Further languages and voices are under development. Telephone bandwidth versions (8kHz sampling rate) of all voices are available.

Aculab TTS is a concatenative synthesis system which uses a deliberately small database of natural speech to perform mixed-unit waveform concatenation. Any necessary modification of pitch and duration is performed by time-domain techniques.

Aculab's approach to multilingual TTS is based on three principles:

- The use of a multilingual system architecture, and multilingual modules wherever possible, so that language-specific elements are kept to a minimum;
- The application of detailed linguistic analysis and sophisticated symbolic processing, allowing us to produce accurate pronunciations and natural-sounding prosody;
- The careful design of speech databases, so that a small database contains a large number of units requiring minimal modification to produce high-quality speech.

Aculab TTS is specifically designed to run on Aculab's award-winning CT hardware, providing over 100 channels of telephone bandwidth speech on a single DSP card. The only restriction on the number of channels is the number of cards which can be accessed by the application. At the time of writing, Aculab TTS is available free of charge to users of Aculab's DSP hardware. For more information about these products, see Aculab's website at <http://www.aculab.com>.

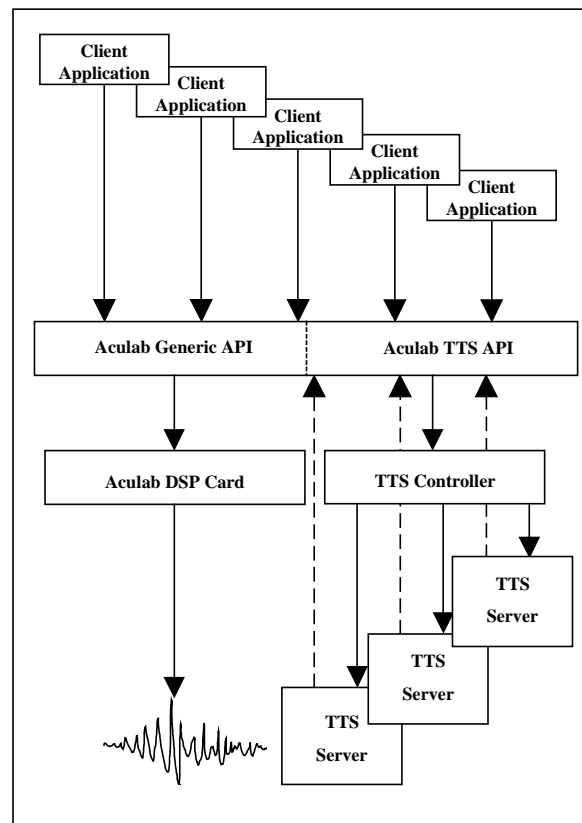


Figure 1. Typical application architecture using Aculab TTS with multiple multilingual servers.

2. Architecture

Aculab TTS is essentially a client-server system. In order to allow for applications with multiple servers, the system includes a controller process which manages any number of servers. The client process sends requests to the controller or server, according to an Applications Programming Interface (API) which provides compatibility with other Aculab products and is compliant with various appropriate standards. Figure 1 illustrates a typical system architecture for an application using Aculab TTS. The API is documented at http://www.aculab.com/support_main/downloads_main.htm.

2.1. Commonality

All the languages which Aculab TTS supports are available on the same TTS server. They share a common system architecture, and they make use of the same modules in most cases. For example, there is a single program which performs

letter-to-sound conversion for unknown words in all languages: this engine uses a separate rule file for each language, so adding a new language simply involves writing a new set of context-sensitive rewrite rules. The same is true of the postlexical rules, the syntactic analysis, and lower-level modules such as unit selection and modification.

The common system architecture is similar to classic rule-based TTS systems such as MITalk and INFOVOX. Any particular language is synthesised by slotting the appropriate modules into this architecture, and perhaps adding or skipping some modules, as illustrated in Figure 2. The same common architecture also supplies most of the functions which are required by the modules: functions such as "get next word" or "compare syntactic categories" are built in, so that new modules can be written in an efficient and consistent manner. This approach also ensures that all data types are the same across languages, allowing us to produce language-independent modules and development tools.

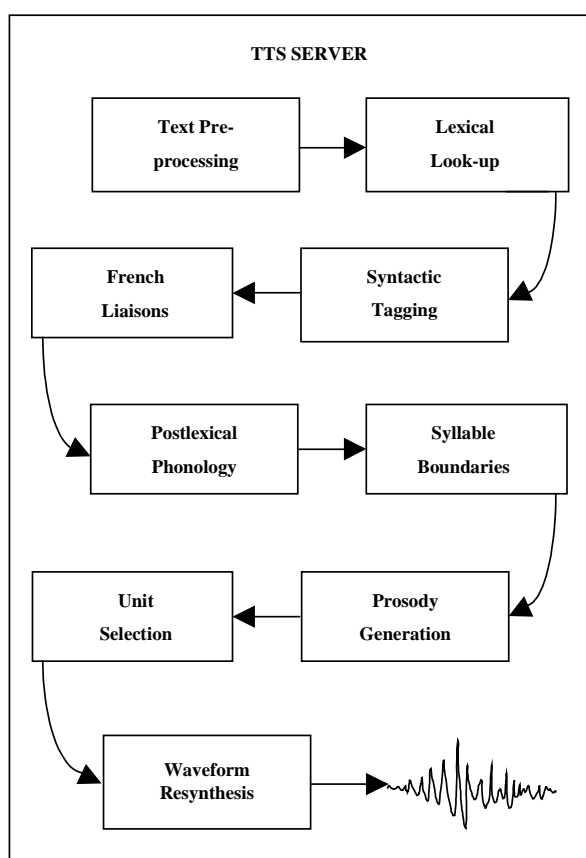


Figure 2. Sequence of processing modules used for French in a typical multilingual Aculab TTS server.

2.2. Modularity

The Aculab TTS system is truly modular. Any set of modules may be executed in any order. This is probably more flexible than we need, but it means that modules may be included or omitted and that modules from different languages may be mixed for particular purposes (if, for example, we wish to synthesise a German text with French prosody).

Special modules may be added for particular languages, such as liaison processing for French, simply by inserting them at the appropriate point in the architecture. Similarly,

modules may apply to one, some, or all languages as discussed in the section on prosody below.

2.3. Multilinguality

As far as possible, Aculab TTS processes all languages in the same way. We aim to exploit the linguistic commonalities between groups of languages, without compromising the naturalness or accuracy of our speech output.

2.4. Tools

In addition to the API for rapid application development, Aculab TTS provides three high-level tools for developers and content authors of CT applications: a lexicon manager, an email pre-processor, and a mark-up language.

The lexicon manager TTSLexMan allows users to build and maintain their own lexical resources, which can be used to override the standard lexica (to accommodate local or regional pronunciations) or to supplement the standard lexica with additional items such as customer names, products and brands.

The email pre-processor handles email headers intelligently, extracting the desired information and discarding other items which would not be suitable for synthesis. It also handles embedded messages, indented text, attached files, emoticons and many other phenomena peculiar to email.

The mark-up language is based on XML and W3C standards, and provides control of many aspects of the synthetic output: pronunciation, prosody, speech rate, etc. At the time of writing, this tool is not yet fully functional but it is scheduled for release early in 2002.

3. Summary

Aculab TTS is a multilingual TTS system designed for CT applications. It currently provides high-quality concatenative TTS for six languages. The quality is comparable with the best commercial TTS systems.

Three major advantages of Aculab TTS are the number of channels of speech available (100 on a single DSP card), the small memory footprint, and the cost (currently free if used with appropriate Aculab DSP hardware).

Future versions of Aculab TTS will include new languages (Italian, Brazilian Portuguese and others), new voices for existing languages, improved waveform modification and concatenation techniques, and compliance with emerging standards for speech. For more details, see Monaghan & Sannier (this vol.) and Monaghan et al. (2001).

You can submit text to Aculab TTS by emailing ttsdemo@aculab.com and our online TTS demonstration will generate a .wav file in real time and send you the URL.

4. References

- [1] Monaghan, A.I.C. & Sannier, F., "A Metrical Model of Prosody for French TTS", in Proceedings of the 4th International Workshop on Speech Synthesis, Pitlochry, 2001.
- [2] Monaghan, A., Kassaei, M., Luckin, M., Amador-Hernandez, M., Lowry, A., Faulkner, D., and Sannier, F., "Multilingual TTS for Computer Telephony: The Aculab Approach", in Proceeding of Eurospeech 2001, Aalborg, Denmark, pp.513-516.