

# SALTcymru\_

## Document 2

An overview of the Festival speech synthesis system, and its possible enhancement for use in Wales in a pre-competitive research stage

Prepared by the  
Language Technologies Unit (Canolfan Bedwyr), Bangor University

April 2008



Llywodraeth Cynulliad Cymru  
Welsh Assembly Government

**This document was prepared as part of the SALT Cymru project, funded by the Welsh Assembly Government under the Knowledge Exploitation Fund's Knowledge Exchange Programme, reference HE 06 KEP 1002**

### **Introduction and relevance to the field**

In the SALT Cymru survey, it is noted that speech synthesis (text-to-speech) is one of the key areas both for users and developers. In particular, *speech enabled communication aids for disabled users and those with specific needs* was the SALT category developed by the greatest proportion of SALT developers participating in the survey. Further, 18% of developers said they would be interested in developing speech synthesis further in future.

Festival provides a useful system for the development and deployment of speech synthesis. It is available as open-source software with very few restrictions, and its comparatively liberal license means that it can be used commercially without the need to expose industrially valuable source code. Further, Festival is a modular system, and basic modules exist within its framework to enable most of the required functions of a speech synthesis system. This means that development time can be targeted to improve and enhance the features required by the developers, without necessarily needing to develop a complete system.

### **Ease of use**

Festival, in its native form, runs from the command line, thus knowledge of its commands is required in order to be able to use it. An ordinary user should not be expected to have this level of expertise with the system in order to be able to use it, and thus a simpler interface is sought.

Such an interface exists, and has been developed by Bangor University's Language Technologies Unit. It works with Microsoft Windows through the operating system's Application Programming Interface, and integrates Festival with Windows so that it can be used with any speech-enabled application that conforms to Windows' standards. This allows a variety of screen readers and similar assistive technologies to work with any Festival voice. The interface can be downloaded freely under the same open-source license as Festival itself.

### **Ease of development**

As a modular system, Festival can be used with little or no extra development. Festival voices exist for English, Welsh and many other languages, and can be deployed by simply downloading the required voice modules and executing the relevant voice commands. The voices can also be operated in a more user-friendly manner through the Windows interface, as explained in the previous paragraph.

Festival contains standard modules that provide the basic functionalities of a speech synthesis system, including speech output, phrasing and intonation. It could, therefore, be deployed with little or no additional development beyond its integration with the developers' speech input and output systems.

The main challenge to Festival developers is improvement of the voice quality from that which is readily available. Most of the voices currently available for Festival (including all those available in Welsh) are diphone voices. These are suitable for screen readers and other applications where the user is likely to listen to the voice for extended periods of

time and hence familiarise themselves with the voice's qualities. Diphone voices are, however, less suitable for telephony applications, and other scenarios where the user will only have limited exposure to the voice, and where utterances are less likely to be repeated. For these situations, unit selection voices are the preferred synthesis technique.

A framework to develop unit selection (and other) voices within Festival already exists. Called Festvox, it uses a series of scripts and pre-written code to reduce the length of time, and the amount of programming expertise, required for voice development. Literature has been written on developing unit selection voices within this framework for languages that did not previously possess such a resource<sup>1</sup>, and as such, the process for developing a new unit selection voice is reasonably well-trodden and well-described. It does, however, require a developer with a level of expertise in speech synthesis (or time to be set aside to learn the fundamentals of the field) and would need, at the very least, six months of one person's development time, plus additional time and a speaker to record the required unit selection speech database.

In developing voices for Festival, it is worth bearing in mind that it was principally designed as a *development* framework rather than a system for deploying voices. It is true that the computational load of a Festival voice is not excessive, and Festival voices can be run on a standard PC with relatively modest processing power. However, if a voice is to be run on a low-power embedded or hand-held system, alternatives to Festival should be sought. In particular, Flite, developed by many of the Festival team, offers a speech synthesis system with a smaller footprint, suitable for a wider range of devices. A mechanism exists to transfer voices from Festival to Flite for their deployment.

### **Potential worth to Welsh SMEs**

Festival offers a ready answer to the problem of speech synthesis for those requiring to develop it. Acceptable solutions are already available for English-language and Welsh-language synthesis using the package. These solutions are not universally applicable, but do however give results that are suitable for the majority of PC users that are likely to rely on speech synthesis for their day-to-day computer use.

It is evident that those SMEs willing to invest time and effort in Festival development should reap rewards from doing so. To a large extent, off-the-shelf solutions are available for most of the likely deployments within Wales. Development of higher-quality voices for Festival, while challenging, should not be an insurmountable problem, especially if time is invested within companies in building knowledge of speech synthesis techniques. It is felt that a significant competitive edge would be gained by a company able to develop, for example, a high-quality telephony system using a newly developed Festival unit selection voice.

Festival is also well-integrated with Windows, meaning that most of the research time for new Festival developments can be taken up in the core work of voice building, rather than being occupied in integration work which is not central to the development process. This has the effect of reducing a company's time to market on Festival development, and increasing profitability.

---

<sup>1</sup> For the Amharic language, see, e.g. [www.cs.cmu.edu/~awb/papers/ssw5/amharic.pdf](http://www.cs.cmu.edu/~awb/papers/ssw5/amharic.pdf)