Integrating Festival and Windows

Rhys James Jones, Ambrose Choy, Briony Williams

e-Welsh Unit, Canolfan Bedwyr
University of Wales, Bangor, Wales, LL57 2EN, UK
r.j.jones@bangor.ac.uk, a.choy@bangor.ac.uk, b.williams@bangor.ac.uk

10.21437/Interspeech.2006-39

Abstract

Festival is a popular open-source development and execution environment for speech synthesis. It has been well-integrated within many environments, particularly Unix ones, but so far has not been easy to integrate natively into Windows. We present two solutions to this: an MSAPI interface, which allows Festival voices to work with a range of speech-enabled Windows applications, and SpeechServer, a client-server architecture which allows Festival to operate within a Flash (or other) application within a web browser. While the motivation for this work was to enable new Welsh diphone Festival voices to be used within screenreaders and other Windows programs, the MSAPI interface is now modularised, allowing it to work with any Festival voice.

Index terms: speech synthesis, Festival, Windows, MSAPI, integration

Introduction

1.1. "Classical" Festival

The system used for this work is Festival [1]. It offers a powerful and flexible platform for the development and deployment of speech synthesis systems. It is a multi-lingual system, and the accompanying Festvox [2] framework simplifies the development of new voices.

Festival has proved a popular choice for those developing new speech synthesizers, due to its modular nature and the fact that it is free software. Festival and the accompanying Edinburgh Speech Tools library are distributed under an X11-type licence allowing unrestricted commercial and non-commercial use.

To date, however, Festival has been relatively inflexible in the way it can be integrated into other speech-enabled applications. While it can be compiled on a variety of Unix platforms, on Cygwin (a Linux-like environment for Windows), and natively on Windows, it is usually run either from the command line or via Festival's server functionality. An ordinary end user would not be expected to be familiar with Festival's command-line interface. Any applications that wish to use Festival as a server or speech engine must use its native API. While this provides flexibility to programmers when writing new applications, it does not offer a solution to, for example, the many Windows applications that use the operating system's standard speech application interface.

1.2. Motivation for this work

This work derives from the Welsh and Irish Speech Processing Resources (WISPR) project [3], one of whose aims was to further develop text-to-speech (TTS) systems for Welsh. To this end, a basic first-generation diphone voice, that had already been partially developed for Welsh outside the Festival framework [4] was integrated by Festival's developers into Festival itself. This paper describes the attempts made by the authors to further integrate it into Windows itself.

The standard speech synthesis and recognition interface in Windows is Microsoft's Speech Application Programming Interface [5], or MSAPI. MSAPI-enabled applications can make use of any MSAPI-enabled voice that has been installed in Windows.

Within Wales, there is great demand for a Welsh-speaking screen reader or web browser for blind people and educational applications. Previously, potential users used English voices, which would pronounce words with an English phoneset and using an English lexicon or letter-to-sound rules. While such a voice can be deciphered with practice, there are clear advantages, particularly in pedagogical areas, of having a native Welsh voice using a full Welsh phoneset and lexicon.

Additionally, the Welsh Language Act provides an onus on public bodies to make available in Welsh any services that are already available in English. As more and more such bodies and governmental organizations provide speech-driven services on their websites, the demand for Welsh speech technology, particularly speech synthesis, also increases.

Since the majority of applications that are required to operate with a Welsh voice run within Windows, it becomes apparent that a system is required within Windows which offers Welsh speech synthesis to existing applications. A system which is compliant with MSAPI will offer the greatest benefit to users, as it will enable any standard speech-enabled Windows application to run with a Welsh voice.

2. Integrating speech synthesis and Windows

2.1. Investigating alternative run-time systems

It was briefly investigated whether existing speech synthesis frameworks, working within Windows, would offer an easy route to integrating a Festival voice within the operating system:

- **Flite** [6] is a popular light-weight speech synthesis system, designed as a 'lite' version of Festival. It includes an MSAPI interface. While there is a mechanism to port existing Festival voices to Flite, this is by no means a

straightforward process. Futhermore, Flite does not offer any letter-to-sound (LTS) rule functionality. The ability to use LTS rules is more critical for Welsh than for many languages, as Welsh exhibits the phenomenon of initial consonant mutation, where the first grapheme of a word may change according to well-defined phonological rules. A Welsh lexicon must therefore contain every possible mutated form of a word in addition to its baseform, increasing its size and complexity. Therefore, in the existing Welsh TTS system, LTS rules are extensively used, and only a small lexicon is present containing exceptions to the standard LTS and syllable stress rules. It was not felt practical to port the existing TTS system to a framework which does not allow LTS rules.

- **FreeTTS** [7], a Java-based speech synthesis system based on Festival, also offers a way to port existing Festival voices. Again, though, this is not by any means a simple process, and LTS rule functionality is not present.

As no alternative framework appears to be practical, it was decided to attempt to integrate the existing Festival voice with Windows' MSAPI.

2.2. First attempt at integration

The first attempt at integration used Cygwin, a Linux-like environment for Windows. This was chosen for ease and speed of development, as it offered an environment that was very similar to a Unix one, on which Festival has been well-developed.

Festival's client/server functions were used in this first development. The MSAPI code used to integrate Festival's server and Windows was based on that which was already part of Flite, but with changes made to include calls to Festival's API, and Festival's specific commands.

The end result of the first development was an installable executable for Windows which, when run, allowed the Festival voice to be included as one of the options for the Windows speech engine. It could be selected within Windows Control Panel (as can be seen in Figure 1), or any speech-enabled application. It offered users the chance to change speaking rate and volume, and also implemented a 'shut-up' command, allowing speech to be muted mid-utterance.

A limited alpha release of Festival's MSAPI installer was made available to specific users within Wales. As it presented the first opportunity to run a native Welsh TTS voice within Windows, it was welcomed. However, feedback indicated that users found it a cumbersome system to use, as it relied on having the Festival server running in the background as an executable before the voice could be initialized, and for the duration of its use.

It was also not supported on versions of Windows other than Windows XP. Due to the requirement for Festival's server to keep a network socket connection open, the operation of the voice on machines without an active network connection was also not guaranteed.

Further, the initial deployment was also limited to one voice, included within the installation itself. This was because the voice initialization commands for Festival were contained within the code itself, which had been compiled before being included within the installation, and could not be changed by the end-user.

An additional problem was that the initial Welsh voice could not cope with any characters outside the standard 7-bit ASCII range. Any attempt to input accented characters, whether by the user or an application, caused the MSAPI engine to crash.

It was thus clear that refinement of this initial development was necessary.

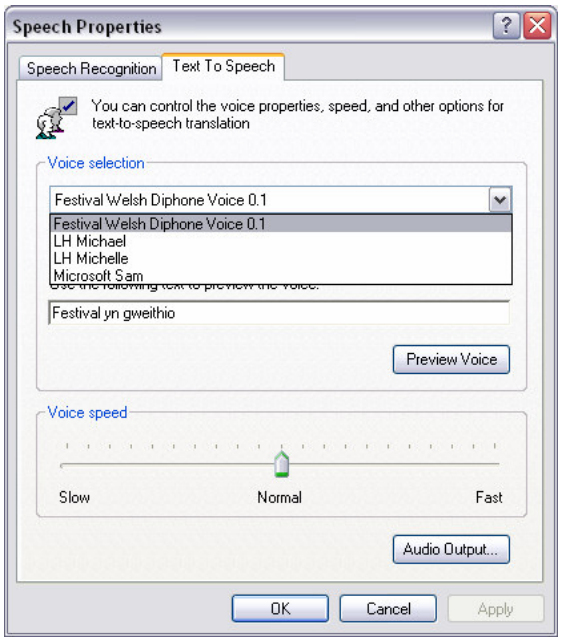


Figure 1. *A Welsh Festival voice appearing in Windows Control Panel, via the first version MSAPI interface.*

2.3. Refining integration

2.3.1. Architecture

A decision was made to move away from a Cygwin build of Festival. This enabled the static link libraries (libfestival.lib, libestools.lib, libestbase.lib, libeststring.lib) produced by the Windows native build of Festival to be included in this MSAPI dll. Thus the TCP/IP communication was replaced by standard Festival API calls.

This architecture allows more flexibility in the configuration of machine and in the version of the operating system required. Tests showed that the new architecture worked with all versions of Windows from Windows 98 onwards.

This new installation is, however, still limited to one voice, which can not be changed without recourse to the MSAPI code itself.

2.3.2. UTF-8 support

UTF-8 is a transformation format of ISO 10646. It is a concise Unicode-based system. It allows characters that can be expressed in 7 bits to be encoded in one 8-bit byte, and so-called 'extended characters' (typically accented ones) to be expressed in two or more 8-bit bytes [8]. The majority of

characters in Western alphabets can thus be expressed in one byte.

UTF-8 is growing in popularity, and all current versions of major operating systems/desktops, and most of their major applications, are UTF-8 compliant. At its lowest level, however, Festival still assumes one-byte characters, and the Festival rules for Welsh do not accept UTF-8 encoded characters.

Accented characters are used by many Festival voices in different languages, but the accented characters they require can usually be found within an 8-bit character codepage which is usually the default in Windows. For some Welsh accented characters, such as ŵ and ŷ, this is not the case, and thus a more comprehensive solution is required, involving a Unicode variant such as UTF-8.

While it is possible to input accents to the existing voice, they need to be input with a special character following the affected letter, e.g., a+ is input for â. The user is not expected to be familiar with this necessity. Therefore a need exists to enable accented characters to be input and converted to the internal format.

This is accomplished by adding a hook to the main Festival code near the start of the flow of control, within the Text module. This allows a pre-tokenization function to be replaced with one which accomplishes transliteration of any input UTF-8 characters into the format required by the Welsh voice. The use of a Festival hook to implement this means that it can be easily adapted for other languages which require Unicode/UTF-8 support.

The disadvantage of such an approach is that a version of the Festival code is required which is different from the ones officially released. In practice this does not pose a problem: the voice is usually deployed within an installer, for which Festival has been precompiled. Further, the transliteration code has been included in the current Festival development branch, so will soon be part of the standard Festival distribution.

A patch containing the changes made to the Festival code is freely available via http://www.e-gymraeg.org/wispr/ .

2.3.3. Changing voices

The MSAPI deployment thus far has been for a unique Welsh voice. The voice initialization commands for Festival have been hard-coded into the system itself, and can not be changed without altering the source code. This is an inflexible arrangement, especially in view of Festival's modular nature which allows multiple voices to be used, and switched between.

Therefore, the MSAPI code was changed so that a configuration file was read by it whenever speech was required. The configuration file includes details of which voice is to be initialized, and its sampling rate. This allows voices to be changed without needing to alter the code. As the configuration file is read each time the voice is initialized, the text file can be edited in the middle of a session, and the voice to be used can be changed 'on-the-fly'.

2.4. Usage

The MSAPI interface can be used with any compliant Windows application. This allows Welsh-speaking computer users to use a 'talking' web browser for the first time, via Firefox and its Foxyvoice plugin [9]. Standard screen-readers and educational word processors can also be used. An example of EdGair, a Welsh educational word processor which works within the MSAPI interface, can be seen in Figure 2.

The applications of the interface are, however, broader than just one language. The facility for changing voices allows developers of other Festival voices to integrate them with the MSAPI framework, assuming that these voices do not need different UTF-8 conversion rules from Welsh. Thus, with a minimum of work, a large number of pre-existing voices can also be used within Windows.

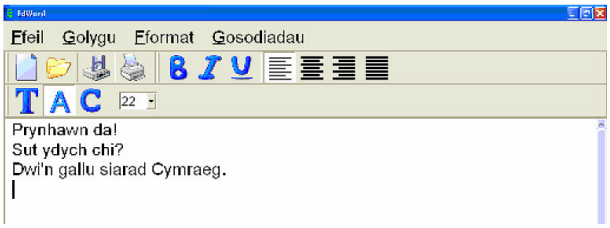


Figure 2. *EdGair, an MSAPI-compliant educational word processor in Welsh.*

3. SpeechServer: Festival through a web browser

The previous section has demonstrated how Festival can be deployed within Windows and the MSAPI framework. Often this is not a practical solution. It might not be practical, for instance, to download a large executable just to allow Festival to read text from a single web page.

SpeechServer was developed to integrate Festival easily with web-based applications. It essentially acts as a bridging interface. Through it, web clients communicate with the Festival speech engine running on a remote server. SpeechServer allows this to happen in a controlled, managed and stable manner.

3.1. Implementation

The SpeechServer client is presented to the user as a Macromedia Flash application (or object). Flash is chosen as it allows multimedia interaction, and communication with a remote host via the internet standard TCP/IP protocol, whilst still remaining a small executable.

Most importantly, the Flash object was chosen because it is capable of streaming sound files within the web page without launching any external applications. As a result, it provides a link, transparent to the user, between the web browser and the remote server on which Festival is running.

On the remote server, SpeechServer itself constantly monitors and manages multiple incoming requests from web clients, by launching a new thread within the application and sending the appropriate request to the Festival server. The function of the Festival server is thus solely to generate the synthesized voice.

Once the Festival engine generates the voice in a wav file, SpeechServer immediately converts it into a more manageable format by compressing it into an MP3 file. The information of this smaller sound file will then be sent back to the Web client and the Flash object streams the MP3 file through the web page.

The operation and interaction of the SpeechServer client and server is shown in Figure 3.

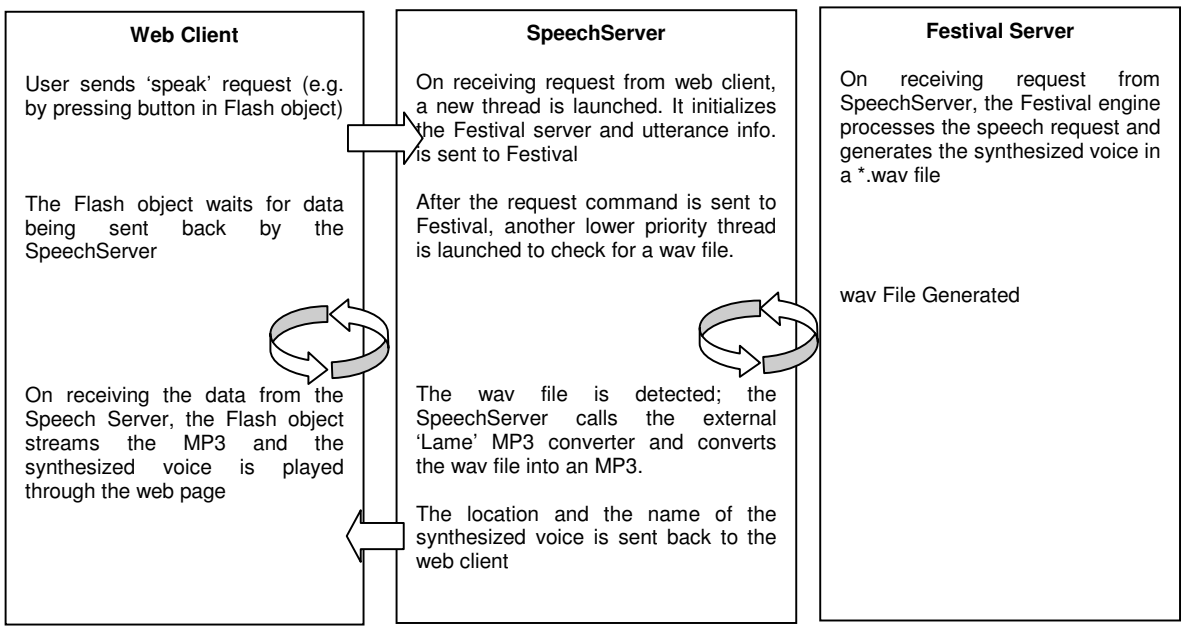


Figure 3. *Event diagram of the web client, SpeechServer and Festival server processes.*

3.2. Usage

A demonstration application of SpeechServer is provided via http://welshtime.notlong.com/, where the small Flash application has been incorporated into a Welsh and Irish speaking clock. The application demonstrates the ability of SpeechServer not only to send speech commands to Festival, but also to allow Festival's full functionality to be accessed through a web browser. In that example, a button allows the voice to be switched from Welsh to Irish.

SpeechServer thus presents a way by which Festival can 'run' in a web browser, without a large computational overhead or the necessity to install a new application on the user's computer.

Conclusions

This paper has presented two methods by which Festival can be presented to a user and integrated within a Windows framework, without any need for special computing knowledge.

While the original motivation for the work was to allow easy integration of a Welsh voice with Windows, the end results are not dependent on any particular voice or indeed language. It is hoped that they will be of use to voice developers and users throughout the world.

It is believed that both Festival's MSAPI interface and SpeechServer extend the usefulness of this popular speech synthesis framework. It is hoped in particular that the release of the MSAPI interface to the speech community, under a free license, will encourage yet more TTS voices to be taken out of a laboratory setting and presented, in a user-friendly way, to those who could benefit from them.

Acknowledgements

This work was accomplished under the WISPR project, funded under the European Union's INTERREG IIIA strand, with additional financial support from the Welsh Language Board.

References

[1] P. A. Taylor, A. W. Black, and R. J. Caley. 'The architecture of the Festival speech synthesis system', in *Proc. 3rd International Workshop on Speech Synthesis,* Sydney, Australia, 1998.

[2] http://www.festvox.org/

[3] B. Williams, D. Prys, and A. Ní Chasaide. 'Creating an Ongoing Research Capability in Speech Technology for Two Minority Languages: Experiences from the WISPR Project.', in *Proc. Interspeech,* 2005, vol. 1, pp. 189-192.

[4] B. Williams. Text-to-speech synthesis for Welsh and Welsh English. In *Proc. Eurospeech, 1995,* vol. 2, pp. 1113-1116.

[5] http://www.microsoft.com/speech/download/sdk51/

[6] A. W. Black and K. A. Lenzo. 'Flite: a Small Fast Run-Time Synthesis Engine,' in *Proc. 4th ISCA Workshop on Speech Synthesis*, 2001.

[7] http://freetts.sourceforge.net/

[8] J. Spolsky. 'The Absolute Minimum Every Software Developer Absolutely, Positively Must Know About Unicode and Character Sets (No Excuses!)' Online article. http://www.joelonsoftware.com/articles/Unicode.htm

[9] http://foxyvoice.kenche.info/