Building Synthetic Voice: A Journey from the Known Voices to a Kiwi Voice

Hira Sathu **Ranjana Shukla** **Jun Li**

UNITEC Institute of
Technology,
Auckland, NZ
hsathu@unitec.ac.nz

This paper relates to a work in progress and is an offshoot of an earlier paper. In the earlier paper the authors had undertaken the addition of audio and still images with a view to augment existing text chat features for an effective online learning process. In this improved version of the application, Microsoft Agent technology in conjunction with Microsoft Text-to-Speech (TTS) engines have been used to provide synthesised voice output, which is based on an American accent at the moment but can be changed to a British accent at short notice. The next step involves the design for a Kiwi accented voice the idea for which took shape during the authors' visit to TalkLink, situated within the Unitec campus. Here various hardware and software applications are used by TalkLink to provide synthetic voices to aid / teach their clients. These voices are either American or British accented synthetic/digitized voices in the absence of a Kiwi accented speech engine. Therefore, the authors visualised the need for a Kiwi voice engine for the in-house developed application, as well as a need to explore the use of the Kiwi Speech engine for porting on to other applications/ platforms by users such as TalkLink and the Royal New Zealand Foundation for the Blind. Since the development of a speech engine from scratch might take couple of years, the authors started exploring the possible use of available tools and resources for this purpose. During this search authors learnt that Carnegie Mellon (CM) University at Pittsburgh, USA, has a speech research centre. The researchers at CM have been working on various aspects of speech synthesis over a long time. They have developed a number of tools for developing synthetic voices. These tools are available as open source along with good documentation. This paper relates to the first of the three phases for the development of the Kiwi Voice Speech engine. The authors investigate appropriate tools for this development, from a range of tools that include Flite, festVox and Festival, developed by the research team at CM. The authors are currently analysing the tools based on the three basic parts of a TTS engine. Briefly the first part relates to text analysis for finding the basic utterance. The second part relates to linguistic analysis for pronunciation and the third part relates to appropriate waveform generation. The authors' intenton to adopt the Festival approach is expected to reduce the cost and time for developing a Kiwi accented TTS.

Keywords

Speech Synthesis, TTS, Speech Engines, Text analysis, Normalization, Lexicons, Waveform Synthesis

1. INTRODUCTION

Incorporation of Voice as an additional dimension in the interactive e-learning has been well established. The authors had undertaken the addition of audio and still images with a view to augment existing text chat features for an effective online learning process. The speech engines used for this purpose were the commonly available Microsoft (MS) text-to-speech (TTS) engines (Mary and Mike). These had been configured to provide speech output capabilities through Microsoft Agents. The authors chanced to visit TalkLink an on Campus organization that supports community programmes by way of aiding clients that require speech support. Here various hardware and software applications are used by TalkLink to provide synthetic voices to aid / teach their clients. Although they would have preferred a Kiwi accented voice, they had to be content with the American or British synthetic voices in the absence of a Kiwi accented speech engine. The other organization that used speech synthesis was Royal New Zealand Foundation for the Blind (RNZFB), this organization too uses speech synthesis that outputs non Kiwi accented voice. In the case of RNZFB, the speech is output in a male voice. The authors envisaged undertaking the development of a speech synthesis engine that would present an option to choose a Kiwi voice with an additional option of its being male or female. Therefore, the authors visualised the need for a Kiwi voice engine for the in-house developed Humanized Virtual Presentation (HVP) application (Shukla, Sathu and Zhang, 2001). The need to explore use of the Kiwi Speech engine for porting on to other applications / platforms by users such as TalkLink and the RNZFB

came as an afterthought. The project for development of a speech engine was divided into three phases. The first phase was to be the development of the concept design; the second phase was to be the development of the prototype Kiwi Voice Speech engine and the final phase was anticipated to be the refinement of the Kiwi Voice Speech engine for commercial applications. This paper pertains to Phase 1, the concept development for the Kiwi Voice speech engine. The paper begins by explaining the HVP application and the background study for the project in Section 2 that lead the study team into the area of text to speech conversion. Section3 develops the contextual background of the various commonly available speech engines. The anatomy of the speech engine (synthesis) is discussed in Section 4 to provide the necessary background information for selection of a suitable approach which is covered in Section 5.

2. HVP APPLICATION AND BACKGROUND

The HVP application that has been developed provides a web-based, voice-enabled classroom facility for the students to make online presentations. It is a client-server application that makes use of both TTS conversion technology and the MS online chat service. Work has been done over the last few months to redesign the application to incorporate a couple of new features. New voice animation feature has been added. The Microsoft Agent interactive and animated characters have been included in the application to enrich the human-computer interaction interface. The gender attribute has also been added to the synthesized voice to enhance the user experience. The application can now automatically set the gender attribute of voice to match the speaker's gender information stored in the server, and thus providing a gender identity support to the application. The team felt that the additional feature of a Kiwi accented voice would further enhance the user experience.

Further research was undertaken by way of approaching organizations that may have leads to our requirement of a Kiwi Voice TTS. TalkLink Communication Centres that provide communication solutions to persons with disabilities were approached. The authors visited the centre at Carrington Road. The technical solutions used by them cover a range of devices that produce speech output. Most devices such as KiwiCom, Macaw and BIGmack output digitized speech. Some devices like AlphaTalker, DeltaTalker and Liberator output synthesized and / or digitized speech (TalkLink, 2004). However, all hardware/ software based solutions output British or the American accented voice. The authors visualized that TalkLink would not be able to meet contingencies where a client would have preferred a Kiwi accented voice.

Further consideration and probing areas where a Kiwi accented speech engine would find a need, revealed Royal New Zealand Foundation for the Blind (RNZFB) as a possible candidate. In fact RNZFB had recently announced offering of a male voice access to the New Zealand Herald for the visually impaired. ScanSoft Inc. has deployed a service that converts text from the paper's web site via XML feeds into natural sounding speech using Speechify software through a program developed by the RNZFB (ScanSoft, 2004). RNZFB on contact intimated that work was underway to introduce an option for its clients to listen to the news in a female voice. RNZFB was open to the suggestion of an option of a Kiwi voice engine integration into their solution as and when available. The next stage involved exploring popular speech engines.

3. POPULAR SPEECH ENGINES

The speech engines that the team was aware of were the Microsoft TTS since integrated into the HVP application developed by the team for the earlier project. Speechify from ScanSoft Inc. was learnt of only in connection with RNZFB. In addition the team attempted to know what help could be had from other educational institution work. In this regard the CM research work was found of greatest interest. These and the other common / popular speech engines are discussed briefly below.

- **Microsoft**. As the worldwide leader in software, services and Internet technologies, MS is also a leading player in the speech solutions market. A key part of Microsoft's strategy is to work with partners to develop the broadest selection of languages available with the wide-ranging country type of TTS products. For example, Microsoft in addition to their in house TTS development have licensed TTS engines from Lernout & Hauspie (L & H). This solution converts any computer-readable text into in-

telligible, human-sounding synthetic speech. The algorithms for parameterized segment concatenation, on which the technology is based, store basic human voice segments such as diphones, triphones and tetraphones (Aculab.com, 2004) and use them to convert the text into speech.

■ **Speechify**. This synthetic voice TTS from ScanSoft is built with recorded voices that enable new applications to blend their TTS and pre-recorded prompts smoothly. Speechify TTS voices can also be customized to a specific voice thereby lowering the time and the burden of expensive recording sessions. However, obtaining of free source code was anticipated as the bottleneck in view of Scan Soft being a commercial organization.

■ **Flite.** This TTS engine is targeted for small devices like the PDAs and Telephones that have limited memory and CPU resources. It consists of a small and an efficient run-time speech synthesis engine. It is not considered suitable as a take off point for the development of Kiwi Voice TTS engine even though it is offered as a free source by their authors at CM. Flite consists of two voices that at best can be considered as examples.

■ **Festvox.** This like Flite is also from CM, created for developing synthetic voices that are more systemic and better documented. It comes with free documentation and specific scripts to build new voices in the supported languages like US and UK English. It also provides example speech databases to help build new voices with links, demos and a repository of new voices (festvox.org, 2004). This is again considered as a limited source for the Kiwi voice objective when compared to the Festival Speech synthesis system discussed below.

■ **Festival.** It is a large sized TTS engine that is typically resource hungry. The availability of faster CPUs and larger memory size in current systems has overcome this resource constraint. It is a multilingual synthesizer with the default language set up at start. Currently festival uses a diphone synthesiser with the duration and intonation generated from statistically trained models. The architecture of Festival has been designed to support large number of tools and to be flexible that permit new voices being developed easily (CMU, 2002). The authors at University of Edinburgh provide the system code in C++, scripts for building, example databases and the instruction without any restrictions. These form the main reason for adopting Festival Speech Synthesiser as the preferred choice for a start point towards the journey to a Kiwi Voice. The next step for the team was to develop an understanding of the anatomy of TTS synthesis engines, which is discussed briefly in the next section.

4. ANATOMY OF SPEECH SYNTHESIS

This section covers the basic constituent parts required for developing a speech engine for a specific language. Based on the literature available about TTS engines, authors have considered three vital components in text to speech conversion: Analysing the given text, creating a lexicon and actual wave synthesis for speech generation.

4.1 Text Analysis.

This entails the text spoken being analysed independent of its meaning. This helps in simplifying the design since it does not take into account the context. The text is first converted in to a string of words for which the pronunciation is already available in the pronunciation dictionary (basic pronunciation dictionaries have list of words and their pronunciations). Since it is impossible to include all the possible words, it is expected that during this analysis some part of the text will be identified as non-standard words, such as numbers, symbols, acronyms, names etc, For these words some rules are designed to generate the pronunciation. To explain this point Jurafsky & Martin (Jurafsky & Martin, 2000, 123) sight an example of a small article for text to speech conversion. Out of 561 words, in the article 3% words did not have the pronunciation. Text analysis next involves dividing a given sentence into shorter length text strings (according to how much can be spoken in a single breath). Reorganizing the sentence in this way also helps in the later stage of waveform generation. It makes the system more efficient during waveform generation since the speech generation is done in smaller parts. Larger sentences would take longer time for the speech output.

Last part of text analysis is normalization (Black & Lenzo, 2003). During this process the text is analysed based on the context. For example when numbers are used in a text, they are spoken differently based on the context in which these are used. Con-

sider the following sentence: "On Sept 6, 1998, authors' institute bought 1998 computers." The 6 is pronounced as sixth, year is pronounced as nineteen ninety eight and number of computers are pronounced as one thousand nine hundred and ninety eight. Once analysis of the given text is done, a string of voice codes (phones) for speech waveform generation is prepared. This is done using the available lexicon and other methodologies.

4.2 Lexicons

A lexicon consists of an explicit list of every word of the language including abbreviations (Jurafsky & Martin, 2000, 66) and proper nouns such as (Jane, China). From lexicon pronunciation are generated using pronunciation dictionary and letter to sound rules. There are number pronunciation dictionaries (list of words and their pronunciation) available for American and British English (CMUdict is for American English and CELEX for British English). Letter to sound rules are used only as a back up for the words that are not listed explicitly. Festival offers option to add new entries in the current lexicon. Festival's letter to sound rule software allows rules to be developed or automatically generated. The choice of hand coding or automatic coding depends on the language and the individual software developer. Hand coding of rules may take more time. For English language, automatic letter to sound rule technique is used in Festival.

4.3 Waveform Synthesis

There are three basic techniques for waveform synthesis:

- Articulatory Synthesis- In this method human vocal tract, tongue, lips etc are modelled. This method produces very natural sounding speech, but requires too much computational power.

- Formant synthesis- In this method the speech signal is broken into overlapping parts, voicing, aspiration, nasality etc. The output of individual generator is used to compose the speech from subparts. With careful handling very good quality of speech could be obtained. Automatic prediction of these parameters is harder, thus getting good quality human voice is difficult.

- Concatenative synthesis- Here the speech waveform is created by joining parts of natural speech from recorded human voice. This type of speech synthesizer requires more memory and disk space. There is a direct relationship between voice quality and database size. Concatenative synthesis gives the most natural sounding speech and it is less complex than the other two.

5. APPROACH FOR DEVELOPING A KIWI SPEECH ENGINE

The above discussion helps in identifying a suitable approach for the design of a Kiwi voice TTS. For developing the pilot Kiwi speech engine it was decided that it would be worthwhile to concentrate on waveform synthesis and utilize the existing text analysis and lexicon that is already available for English. It is appreciated that basic language being English, text analysis and lexicon should be the same. For speech synthesis concatenative approach is being considered here. For this approach we would require recording and using a diphone database of Kiwi accented English. This diphone database would consist of all the possible phone-phone transitions. In a language usually number of diphones are typically between 1000 to 2000 (Jurafsky & Martin, 2000, 274).

5.1 Defining the Diphone List

The clear articulation of diphones is essential. One of the techniques used for this purpose uses target words embedded in the sentences and it is seen that diphones are pronounced with desired consistency. For obtaining best results, the vocal variation should be minimal.

5.2 Recording Environment

For recording we require a uniform set of pronunciation as far as possible. Since the recording may be done in multiple sittings, the speaker's voice and the recording environment must be consistent for all these sittings. To achieve this it is even recommended that recording be done at the same time of the day. Quality of the recording instrument, its setting, microphone distance etc are quite crucial too. Head mounted microphone gives best results in this regard.

6. CONCLUSION

The above discussion reveals that there is a need for a Kiwi TTS engine. The analysis of various TTS engines and the evaluation of the TTS technology

has lead the team to select the concatenative approach for Kiwi voice prototype development phase. This phase is anticipated to involve the study and use of Festival TTS tools. Selection and recording of the Kiwi voice to form a representation of a typical Kiwi voice database is expected to be an exciting and challenging part of the paper. To make a Kiwi voice we need to create a pronunciation dictionary and consider generation of waveforms. Patience is to be our watchword since "the process of building a voice is not necessarily going to work the first time" (Black & Lenzo, 2003). It is expected that once the prototype Kiwi Voice engine is refined it would find commercial applications such as speech enabled ATM for the visually impaired and other applications that prefer a Kiwi Voice over the common British and the US English version.

REFERENCES

Aculab.com (2004). "Prosody with Text-To - Speech (TTS)". Accessed May 5, 2004, http://www.aculab.com/products/pdf_files/L&HTTS5c.pdf

Black, A. W. and Lenzo, K. A. (2003). "Building Synthetic Voices", Language Technologies Institute, Carnegie Mellon University and Cepstral, LLC, http://festvox.org/festvox/festvox_toc.html

CMU Speech Software (2002). Accessed April 16, 2002, http://festvox.org/

Festvox.org (2004). Accessed on March 10, 2004, http://festvox.org/

Jurafsky & Martin. (2000), Speech and Language Processing, Prentice-Hall, pp.66, 123, 274.

ScanSoft Inc. (2004). Accessed March 9, 2004. . http://www.scansoft.com/news/pressrelease/2003/20031201_rnzfb.asp

Shukla, R., Sathu, H., Tang, Z. (2001) "Humanized Virtual Presentation and Chat using TTS". Journal of Contemporary Business Issues, volume 9, No2, pp.61-66

TalkLink Communication Centres, (2002). Accessed April 16, 2002,http://www.talklink.org.nz/howwehelp/hightech.html