http://e-flt.nus.edu.sg/

Electronic Journal of Foreign Language Teaching
2006, Vol. 3, No. 1, pp. 131-135

Review of "VoiceText"

Software Version	English, January 2005
VoiceText Type	Engine
Sampling Rate	16kHz
Minimum system requirements	CPU 400MHz Disk space 650MB Memory 256MB OS WindowsXP/2000/2003Server,Linux
Original developer	Voiceware <http://www.voiceware.co.kr>
Distributor	Pentax <http://voice.pentax.co.jp/pentaxtts/>

Reviewed by Hideto D. Harashima

1 Introduction

Modern technology has developed to the point where it now allows language teachers to create audio materials of their own and present them online to learners for "anywhere-anytime learning." It suggests a great opportunity for all language teachers to tailor listening materials for the specific levels and the needs of their students. However, creating original audio material is not an easy job for a language teacher, especially when he or she is a non-native speaker of the target language. (Some of the concrete difficulties and the advantages of using TTS technology, in general, are discussed, with details, in Harashima's forthcoming paper "Online Listening Materials Made Easy by Text-To-Speech Technology".) What is desired is a system which renders standard voices of the target language which are easy to edit and re-edit, and is within the reach of an ordinary teacher's in terms of the costs. TTS is a technology that meets these demands.

2 What is TTS?

TTS means Text-To-Speech conversion technology or engine. It is computer software that converts a written text into a voice file. Speech synthesis technology has been around for many years. According to Jacobson (n.d.), the first text-to-speech system was completed at Bell Laboratories as early as 1968. Kilickaya (2006) also states that the first TTS was implemented in the Speak and Spell handheld electronic learning aid by Texas Instruments in 1978. Until recent years, machine-made voices had been of such poor quality that they were often ridiculed as "robot voices". However, corpus-based speech synthesis technology has developed so rapidly in recent years that it can now be considered a valid replacement for human voices for various purposes, including language learning.

Studies of speech synthesis in the past had focused on developing good *vocoders*, or voice generating machines. Recent corpus-based and context-sensitive TTS adopts a different approach to speech synthesis. It collects samples of real human voices on the phoneme and the word level. Then it assembles these elements so as to read aloud a given text before it adjusts the sound junctures. Finally, appropriate intonations and accents are added to the reading in accordance with the context. It is a highly sophisticated process.

There are a number of TTS engines available on the market, including PowerTTS, Speechfy, Ultra Hal Reader, Natural Voices, ReadPlease, Festival, RealSpeak, Cepstral, NeoSpeech, TextAloud, Microsoft Reader, eLite, WavePad and VoiceText. Many of these products offer free online demonstrations or sample downloads, but the available functions are usually limited. Among these TTSs, VoiceText is considered to be a cutting-edge product with excellent quality and utility. It will be reviewed in detail below.

3 VoiceText

VoiceText is a TTS system originally developed by a Korean company Voiceware, which is now a 100% subsidiary of Pentax in Japan. VoiceText is available in two types: server/client type and engine type. The server/client type does voice conversion over the network system, whereas the engine type does it inside a (standalone) computer. This review is on the engine type.

The VoiceText engine is not usually sold to the general public; it is mainly marketed to corporate customers or to software developers to be customized for different software packages. Demonstrations are available on the Pentax webpage at http://voice.pentax.co.jp/pentaxtts/ttsdemoplay.asp.

VoiceText offers voices in four different languages: English, Japanese, Chinese and Korean. This review covers only the English voices, i.e. those of *Kate* and *Paul*.

3.1 Method of speech synthesis

There are two major currents in modern speech synthesis technology: rule-based and corpus-based speech synthesis. The rule-based method synthesizes speech based on phonemes, while the corpus-based method does so based on larger units such as syllabi, words or phrases, using statistical prosody analysis models such as HMM (the Hidden Markov Model) and the regression tree model[1]. VoiceText adopts the latter. The corpus-based speech synthesis is considered to be more able to cope with prosodic variations of human speech. Hence it is more natural-sounding, including the insertion of speech that sounds as if it has been affected by emotions. VoiceText is categorized as one of the corpus-based TTSs, but it partially adopts phoneme-based synthesis for uncommon words. It also incorporates context analysis as pre-treatment to speech synthesis, resulting in higher naturalness of pitch contours, which effectively makes VoiceText one of the most sophisticated TTSs currently available.

3.2 Basic operations

The operations on VoiceText are very easy and straightforward. As soon as the software is booted up, a simple editing screen appears (see Fig. 1).

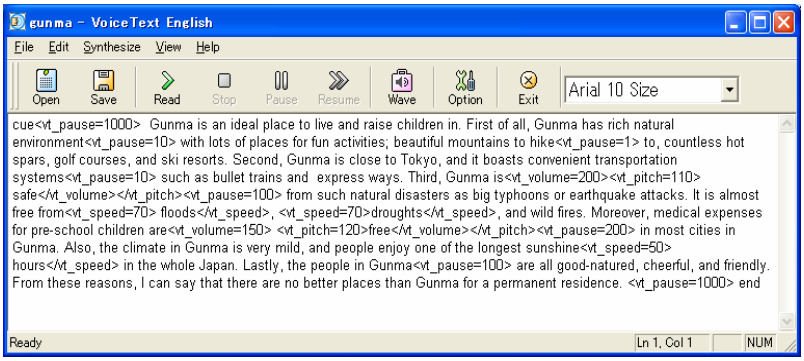

Fig. 1: VoiceText Editing Screen

Into this screen a user can type any given text or copy-and-paste a text from a webpage. Then he or she clicks the “Read” button, and the system immediately starts reading the text aloud. (Needless to say, the speakers on the computer must be turned on beforehand.) At this point, if the user does not like the way the text is being read, he or she can add some prosodic amendments, which are described below. He or she can then click the “Save” button to save the text and the “Wave” button to save the speech as a WAV file. Later, the user might want to convert the file into a more compact file format such as MP3, using other software, such as Audacity.

3.4 Features

VoiceText differentiates itself from other TTSs by two distinct features: the prosodic feature tags and the user dictionary with a sound editor. The prosodic feature tags are little tags which a user implants at both ends of a particular part of a text (look closely at Fig. 1.). There are tags for controlling speech speed, pause, pitch and volume. By utilizing these control tags, a user can modify the speech to make it sound closer to what he or she desires, adding some elements of emotion and emphasis.

The other feature, the user dictionary, allows a user to build up his or her own personal dictionary within VoiceText. The accompanying sound editor is especially useful when the user desires to have VoiceText pronounce new or foreign words and acronyms properly. A user first stores some vocabulary items in the dictionary. Then he or she assigns proper pronunciations to them using either IPA phonetic symbols or the Carnegie sound-spelling system. These newly created pronunciations in the custom dictionary override the originally assigned pronunciations in the main text. Figure 2 shows how IPA symbols are chosen to represent word pronunciation.

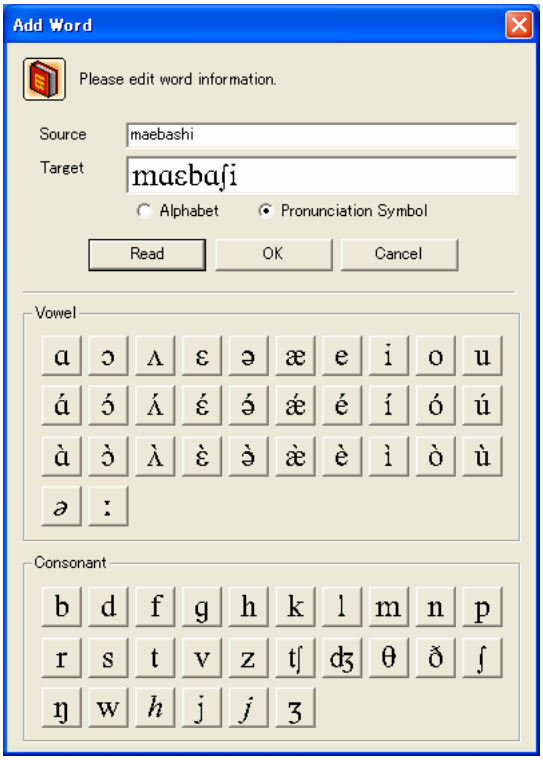

Fig. 2: User Dictionary Sound Editor Screen with IPA Palette

4 Evaluation

4.1 Voice quality

The voice quality of VoiceText is excellent, for it uses the real voice samples of (probably professional) radio actor *Paul* and actress *Kate*, who both speak standard American English, at the sampling rate of 16KHz. The voices are crystal clear and without hum noise.

4.2 Intonation

Individual word intonations sound quite natural, standard and flawless. Phrasal intonations are also at a satisfactory level, owing to VoiceText's contextually sensitive speech synthesis feature. However, there are some cases in which it adopts slightly different intonation patterns from what we normally expect, but these are within the acceptable range. It is also inevitable for VoiceText to tend to sound somewhat flat, and with monotonous intonations. We cannot expect VoiceText, or any other TTS for that matter, to exhibit emotive intonations for phrases such as "Oh, my God! What a fool I am!"

4.3 Juncture

Junctures between phonemes are, in fact, the Achilles' heel of many corpus-based TTSs. Engineers sweat day and night about how they can conjoin two adjacent phonemes without pitch wobbles coming through. This is not a problem with common words themselves. It surfaces with uncommon or foreign words, and mostly at the junctures around short words. For example, for a phrase such as "at an organization," VoiceText slightly blurs at the end of *at*.

4.4 Pronunciation of rare words

As might be imagined, VoiceText is not very good at pronouncing rare and foreign words. The name of my hometown in Japan, Maebashi, can never be pronounced perfectly by VoiceText. Nonetheless, we can put this word into the aforementioned custom dictionary and assign quasi-perfect pronunciation to it. Otherwise, we can find an alternative by applying a little hack, i.e. common sound spelling, to the text such as *mah-eigh-bah-she*.

4.5 Overall evaluation

The overall quality of VoiceText is superb such that people who hear the voices for the first time are startled by their natural sounds. In fact, in a survey which I conducted with my students who repeatedly heard those "voices" online during e-learning course navigations and instructions, 62.7% of them (N=59) answered that they did not notice that the voices were not the products of natural human speech. Furthermore, the interface of VoiceText is very simple and easy to operate, which is also an advantage.

5 Conclusion

I have used a number of TTSs, but nothing has come close to the satisfaction level which I have experienced when using VoiceText. It is certainly one of the best TTSs currently available, and I can recommend it to everybody who reads this review. I believe the demand for a high quality TTS among foreign language educators will grow with time. The articulate and composed pronunciation of standard American speech that VoiceText offers is perfect for short voice navigation/instruction/guidance, as well as for listening tests. The listening component of college entrance examinations bear increasingly significance these days in Japan and elsewhere. If VoiceText can replace human radio actors/actresses for these tests, it will be a solution to the

problems faced by test administrators, because it is easier to use, faster, more secure in terms of possible information leaks, and more economical in the long run.

Notes

[1] The discussion of these highly technical and mathematical analyses is beyond the scope of this review. Readers who are interested in how these models are utilized for natural speech processing are encouraged to refer to such works as Blunsom (2004), and Iwahashi and Sagisaka (2000).

References

Blunsom, P. (2004). *Hidden Markov Models*. Retrieved May 19, 2006, from http://www.cs.mu.oz.au/460/2004/materials/hmm-tutorial.pdf

Harashima, H.D. (Forthcoming). Online listening materials made easy by text-to-speech technology. Place: Publisher?

Iwahashi, N., & Sagisaka, Y. (2000). Statistical modelling of speech segment duration by Constrained Tree Regression [Electronic version]. *IEICE Trans. Inf. & Syst., Vol. E83-D, No. 7*, July 2000, 1550-1559.

Jacobson, K. (n.d.). Approaches to speech synthesis. Retrieved May 19, 2006, from http://umsis.miami.edu/~kjacobso/speechsynth/speechsynth.htm

Kilickaya, Ferit. (2006). 'Text-to-speech technology': What does it offer to foreign language learners? *CALL-EJ Online*, *7*(2). Retrieved January 31, 2006, from http://www.tell.is.ritsumei.ac.jp/callejonline/journal/7-2/Kilickaya.html