International Journal of Multimedia Data Engineering and Management
Volume 7 • Issue 2 • April-June 2016

Unit-Selection Speech Synthesis Method Using Words as Search Units

Hiroyuki Segi, Department of Computer and Information Science, Seikei University, Tokyo, Japan

ABSTRACT

Unit-selection speech-synthesis systems have been proposed. In most of the unit-selection speech-synthesis systems, search units are rather short such as syllables, phonemes and diphones. However, when applied to large speech databases, shorter units produce more voice-waveform candidates and a larger speech database cannot be used without narrow pruning for practical use. Narrow pruning impairs the quality of the synthesized speech. Here the author examined the possibility of using words as search units. Subjective evaluations indicated that 70% of the speech synthesized by the proposed method sounded more natural than that synthesized by a conventional method. The five-point mean opinion score of the synthesized speech was 3.5, and 21% was judged to sound as natural as human speech. These results demonstrate the effectiveness of unit-selection speech synthesis using words as search units.

KEYWORDS

Broadcast Program, Mean Opinion Score, Search Unit, Speech Database, Speech Synthesis, Unit Selection, Word Unit

1. INTRODUCTION

There is a strong need for higher quality Text-To-Speech (TTS) conversion in broadcasting services. The development of a TTS system that can generate synthesized speech that sounds similar to a human voice could improve access to text information in both data broadcasts (Sakai, 2007) and broadband contents (Baba, 2012) for visually impaired and mobile receivers. Moreover, a high-quality TTS system could facilitate the development of automatic spoken broadcasts, such as weather reports (Segi, 2013) and even automatic television broadcasts, by combining speech with computer-graphic animations generated from a script (Hayashi, 2013; Doke, 2012).

Several types of TTS system have been reported to date. One group utilizes the compilation of recorded speech sounds, which is employed in airport and train announcements (Demeur, 1987). Although the speech synthesized by this method has not yet been evaluated, it is widely considered

DOI: 10.4018/IJMDEM.2016040104

to achieve human voice quality based on its use in broadcast systems. However, the content of the speech synthesized by this method is limited to combinations of recorded phrases connected by silent sections. Thus, it cannot be utilized for the speech synthesis of arbitrary input sentences. Moreover, this method does not take coarticulation into account, suggesting that the naturalness of the synthesized speech is degraded without sufficent silence sections. Indeed, 91% of synthesized speech with coarticulation was evaluated as more natural than synthesized speech without coarticulation in a previous study (Segi, 2010).

A second group of TTS systems employs Hidden Markov Models (HMMs) (Zen, 2009; Toda, 2007). This method analyzes the speech data, extracts prosody and voice-quality components from the speech data respectively, and allows them to be controlled independently (Kawahara, 1999). For example, HMM TTS systems can extract the feature parameters of phoneme "a" from a speech database, and use them to synthesize speech. This method has several advantages as it is easy to use for voice conversion, has good performance with small speech databases, and does not require a high-performance Central Processing Unit (CPU) or large memory. However, the naturalness of the speech synthesized using this method is not so high (Zen, 2008; Takaki, 2011; Nose, 2013).

A third group of TTS systems uses the Pitch Synchronous OverLap and Add (PSOLA) method (Moulines, 1990). This technique converts the pitch of short-period waveform to the target pitch, and connects the waveform samples with overlap. It is necessary to determine the boundaries of the phonemes and fundamental periods before speech can be synthesized using this method. Although these boundaries can be set automatically, they must be modified manually to improve the quality of the synthesized speech. Therefore, it is difficult to increase speech-database size in order to spend much time and money on manual modification of large speech databases. This restriction means that large pitch conversion is necessary and results in less natural sounding synthesized speech.

The fourth group of TTS systems employs the unit-selection method (Hunt, 1996; Conkie, 2011; Toda, 2002). Although similar to PSOLA, this method can use a larger speech database. Significant pitch conversion is not always necessary with this approach, and the synthesized speech is considered to sound more natural than that produced by the PSOLA method.

If synthesized speech sounds unnatural, users cannot tolerate it for long periods of time. Therefore, it is desired to achieve natural synthesized speech. Among TTS systems described above, the unit-selection speech synthesis method tends to perform best in this respect (Kawai, 2004, 2006). In these papers, speech samples synthesized by 10 commercially available systems and XIMERA, which is a proposed TTS system using the unit-selection method, was evaluated. All the speech synthesis method of 10 commercially available systems is not clear but they include HMM and PSOLA TTS system. The results showed the superiority of XIMERA over commercially available ones. Therefore, the unit-selection method can synthesize more natural speech than any other methods. To achieve more natural synthesized speech in the unit-selection speech synthesis method, a larger speech database is needed (Segi, 2004). Here, we propose a unit-selection speech-synthesis method that uses words as search units, in order to increase the size of the speech database.

The current paper is organized as follows. Section 2 describes conventional work of unit-selection speech-synthesis methods. Section 3 describes the proposed unit-selection speech-synthesis method using words as search units. Section 4 describes an experiment to compare words with phonemes as search units of speech synthesis by performing subjective evaluations of naturalness. The results of our study are summarized in Section 5.

2. PREVIOUS WORK ON UNIT-SELECTION SPEECH SYNTHESIS METHODS

To achieve more natural synthesized speech, a larger speech database is needed. In most of the unit-selection speech synthesis systems, search units are rather short such as half-phones (Conkie, 2011; Toda, 2002), phonemes (Hunt, 1996), and triphones (Yan, 2010). A shorter unit, however, produces a larger number of candidates of voice waveform and a larger speech database cannot be used without

Requirements to a Search Engine for Semantic Multimedia Content

Lydia Weiland, Felix Hanser and Ansgar Scherp (2014). *International Journal of Multimedia Data Engineering and Management (pp. 53-65).*

www.igi-global.com/article/requirements-to-a-search-engine-for-semantic-multimedia-content/120126?camid=4v1a

Default Reasoning for Forensic Visual Surveillance based on Subjective Logic and Its Comparison with L-Fuzzy Set Based Approaches

Seunghan Han and Walter Stechele (2011). *International Journal of Multimedia Data Engineering and Management (pp. 38-86).*

www.igi-global.com/article/default-reasoning-forensic-visual-surveillance/52774?camid=4v1a