Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 38 (2016) 46–67

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Concept-to-Speech generation with knowledge sharing for acoustic modelling and utterance filtering[☆]

Xin Wang, Zhen-Hua Ling[*], Li-Rong Dai

National Engineering Laboratory for Speech and Language Information Processing, University of Science and Technology of China, Hefei, Anhui 230027, PR China

Received 28 February 2015; received in revised form 19 October 2015; accepted 16 December 2015
Available online 6 January 2016

Abstract

A Concept-to-Speech (CTS) system converts the conceptual representation of a sentence-to-be-spoken into speech. While some CTS systems consist of independently built text generation and Text-to-Speech (TTS) modules, the majority of the existing CTS systems enhance the connection between these two modules with a prosodic prediction module that utilizes linguistic knowledge from the text generator to predict prosodic features for TTS generation. However, knowledge embodied within the individual modules has the potential to be shared in more ways. This paper describes knowledge sharing for acoustic modelling and utterance filtering in a Mandarin CTS system. First, syntactic information generated by the text generator is propagated to a hidden Markov model (HMM) based acoustic model within the TTS module and replaces the symbolic prosodic phrasing features therein. Our experimental results show that this approach alleviates the local hard-decision problem in automatic prosodic phrasing for Mandarin CTS systems and achieves a comparable performance to the traditional approach without explicit prosodic phrasing. Second, the acoustic features of multiple synthetic utterances expressing the same input concept are utilized to evaluate the utterance candidates. With this 'post-processing' mechanism, our CTS system is able to filter out inferior synthetic utterances and find an acceptable candidate to express the input concept.

Keywords: Concept-to-Speech; Speech synthesis; Hidden Markov model; Natural language generation

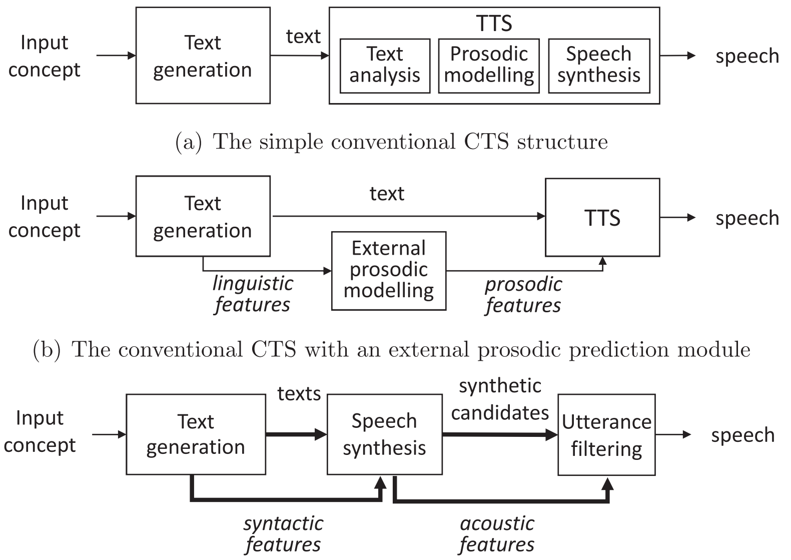
1. Introduction

The transfer of human language proficiency to machines has been explored for decades. One branch of exploration, the Concept-to-Speech (CTS) approach, endeavors to enable machines to produce speech based on abstract representations of the sentence to be spoken. Because humans articulate their ideas by translating them into syntactic, phonological and phonetic codes that guide the vocal tract articulator to produce acoustic waveforms, a similar process is to some extent adopted by CTS systems wherein the input concept undergoes syntactic and phonological processing and eventually drives the 'articulator' to synthesize speech. Another type of human language proficiency, reading text

[☆] This paper has been recommended for acceptance by S. King.

[*] Corresponding author at: University of Science and Technology of China, No. 96, JinZhai Road, Hefei, Anhui, China. Tel.: +86 551 63603140.
E-mail addresses: wangx064@mail.ustc.edu.cn (X. Wang), zhling@ustc.edu.cn (Z.-H. Ling), lrdai@ustc.edu.cn (L.-R. Dai).

http://dx.doi.org/10.1016/j.csl.2015.12.003



(a) The simple conventional CTS structure

(b) The conventional CTS with an external prosodic prediction module

(c) The proposed CTS method which shares syntactic features for acoustic modelling and acoustic features for utterance filtering. The bold arrows represent the information flow for multiple utterances.

Fig. 1. Structures of the conventional vs. proposed CTS method.

out loud, has also been transferred to machines. This technique is known as Text-to-Speech (TTS) synthesis. One obvious difference between CTS and TTS is that the CTS approach must convert conceptual representations into sentences before articulating them while TTS directly accepts a concrete sentence as input. Because conceptual representation is typically domain dependent, CTS is incorporated into a spoken dialogue system wherein conceptual representations can be defined and provided by front-end modules. For example, well-known CTS systems have been deployed in an inquiry system for a water-supply network (Young and Fallside, 1979) and a multimedia medical briefing system (McKeown and Pan, 1997).

Although CTS may appear to be more complex than TTS due to the additional task of sentence composition, CTS can be practically implemented by concatenating a text generator with a TTS module, as shown in Fig. 1(a). This solution has been adopted by several dialogue systems that generate spoken responses (Litman et al., 1998; Rudnicky et al., 1999; Walker, 2000). However, this solution abandons the valuable linguistic knowledge that is available in the text generator, which cannot be fully retrieved by the error-prone text-analyzer in the TTS module.

Finding improvements to knowledge sharing amongst the components of a CTS system is a vital issue in CTS research (Young and Fallside, 1979). The consensus among researchers seems to be that linguistic information such as discourse, semantic and syntactic knowledge permits finer control of prosody modelling in CTS systems. Existing CTS systems based on this idea typically incorporate external prosodic prediction modules to convert linguistic information into prosodic symbols. This idea is illustrated in Fig. 1(b). For example, the system in Pan (2002) utilizes various types of linguistic information and an instance-based learning algorithm to predict ToBI symbols (Silverman et al., 1992) for CTS generation in the medical briefing domain. Other CTS systems apply similar ideas, and their performance confirms that predicting prosodic symbols based on the knowledge from text generation is effective for CTS tasks (Danlos et al., 1986; Theune et al., 1997; Dorffner et al., 1990; Nakatani and Chu-Carroll, 2000; Fawcett, 1990; Teich et al., 1997; Takada et al., 2007; Schnell and Hoffmann, 2004; Hitzeman et al., 1998; Pan, 2002).

Although the above strategy is informative, we wonder whether it represents the best way to address the CTS task, particularly for Mandarin. Initially, the pipeline structure in Fig. 1(b) only shares the linguistic information from text generation with the prosodic model or a prosodic phrasing model for Mandarin. We think that propagating the linguistic information as well as the prosodic features to the speech synthesizer is a worthy pursuit (Badino et al., 2012). However, because automatic prosodic phrasing for Mandarin is challenging, we doubt that the improvements produced by linguistic information in both automatic prosodic phrasing and acoustic modelling can prevent the unacceptable synthetic