Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 34 (2015) 43–60

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Text-to-speech synthesis system with Arabic diacritic recognition system☆

Ilyes Rebai *, Yassine BenAyed

MIRACL: Multimedia InfoRmation System and Advanced Computing Laboratory, University of Computer Science and Multimedia, Tunis Street km. 10, Technopole, Sfax, Tunisia

Received 26 January 2014; received in revised form 5 April 2015; accepted 17 April 2015
Available online 27 April 2015

Abstract

Text-to-speech synthesis system has been widely studied for many languages. However, speech synthesis for Arabic language has not sufficient progresses and it is still in its first stage. Statistical parametric synthesis based on hidden Markov models was the most commonly applied approach for Arabic language. Recently, synthesized speech quality based on deep neural networks was found as intelligible as human voice. This paper describes a Text-To-Speech (TTS) synthesis system for modern standard Arabic language based on statistical parametric approach and Mel-cepstral coefficients. Deep neural networks achieved state-of-the-art performance in a wide range of tasks, including speech synthesis. Our TTS system includes a diacritization system which is very important for Arabic TTS application. Our diacritization system is also based on deep neural networks. In addition to the use deep techniques, different methods were also proposed to model the acoustic parameters in order to address the problem of acoustic models accuracy. They are based on linguistic and acoustic characteristics (e.g. letter position based diacritization system, unit types based synthesis system, diacritic marks based synthesis system) and based on deep learning techniques (stacked generalization techniques). Experimental results show that our diacritization system can generate a diacritized text with high accuracy. As regards the speech synthesis system, the experimental results and subjective evaluation show that our proposed method for synthesis system can generate intelligible and natural speech.

Keywords: Text-to-speech synthesis; Statistical parametric; Deep neural networks; Natural language processing; Diacritization system

1. Introduction

Text-To-Speech (TTS) system is one of the most important technologies due to the expanding field of applications, such as: multimedia, telecommunication and aids for handicaps. People that speak Arabic as their native language are more than 442 million around the world. Five million Arabic people are blind around the world (Zaki et al., 2010).

☆ This paper has been recommended for acceptance by Mark J.F. Gales.

* Corresponding author. Tel.: +216 21447305.
E-mail address: rebai_ilyes@hotmail.fr (I. Rebai).

http://dx.doi.org/10.1016/j.csl.2015.04.002
0885-2308/© 2015 Elsevier Ltd. All rights reserved.

Hence, Arabic TTS with intelligible and natural speech quality is required. However, the field of speech synthesis for Arabic language has not sufficient progresses and it is still in its first stage. This could be explained by the fact that:

- One of the problems facing computer processing of the Arabic text is the absence of the diacritic marks in the modern text (Elshafei et al., 2006). These marks are used to identify the right pronunciation of the text.
- It is difficult to obtain an Arabic speech database for speech synthesis task. The solution consists of developing a specific speech database (Chouireb and Guerti, 2008; Hamad and Hussain, 2011).
- Linguistic researches for Arabic language are limited.

While native Arabic readers can determine the appropriate vocalization of the text with minimal difficulty, computer processing of Arabic text is often obstructed by the lack of diacritic signs. For instance, a given Arabic TTS would not generate speech from undiacritized text because there are different pronunciations of the same undiacritized word. To make the problem easier for English readers, if the words “read, red, ride, rude” are written without vowels, they will be the same word “rd” and it will be impossible to determine the correct meaning and pronunciation of this word. Therefore, the Arabic diacritization system is very important for an Arabic TTS application. This system recognizes the missing diacritics of the input text. The problem of diacritic sign restoration can be solved by three approaches: standard Arabic dictionaries, rule-based approach and machine learning approach.

Speech synthesis system is the process of generation of speech, as output, from text, as input. The two popular approaches are *concatenative speech synthesis* (known as corpus-based approach) (Hunt and Black, 1996) and *statistical parametric speech synthesis* (called also knowledge-based approach) (Black et al., 2007). In the first approach, desired speech is produced by selecting and concatenating required segments from pre-recorded speech by human. Many systems use a corpus of fixed length units, typically phonemes or diphones. Other concatenative systems use more varied, non-uniform units speech segments database. For instance, in (Hamad and Hussain, 2011), the authors developed an Arabic TTS based on allophone and diphone concatenation method. This variety of speech segments allows the generation of more natural speech. The highest speech quality is generated based on unit selection (Hunt and Black, 1996; Clark et al., 2007). The basic concept of this method consists of concatenating speech segments without modification. It uses a large database, including units in different phonetic and prosodic contexts. However, large speech database requires a huge memory storage. Furthermore, to have different voice styles and emotions, another speech database is required for such style or emotion which increases the required storage capacity (Zen et al., 2009). These issues make corpus-based synthesis systems not suitable for devices with limited resources.

In direct contrast to the concatenation of pre-recorded speech units approach, Statistical Parametric speech Synthesis (SPS) approach consists of converting a set of parametric representations to speech waveform. During the recent years, SPS approach has been growing fast in popularity and the generated speech quality has been found to be as intelligible as human voice (Zen et al., 2009, 2013; Tokuda et al., 2013; Ekpenyong et al., 2014). In such SPS system, pre-recorded speech database is replaced by a set of generative models (e.g. neural networks, hidden Markov models). These techniques are used to model the acoustic parameters (spectral and excitation parameters) extracted from a speech database. Subsequently, the target speech waveform is reproduced from the appropriate speech parameters through a source-filter model. The main advantages of the SPS approach over the concatenative approach are the small memory footprint and the flexibility of voice characteristics’ modification (e.g. style, emotions) (Nose et al., 2005, 2007; Barra-Chicote et al., 2010).

Statistical parametric synthesis systems are composed of two parts: training part (generative model is used to create models that map the linguistic features into acoustic parameters) and synthesis part (reconstruct the speech waveform from the predicted parametric representations using a vocoder, e.g. MLSA-based vocoder: Mel Log Spectrum Approximation (Imai et al., 1983)). Since the last decade, Hidden Markov Models (HMMs) have been widely used in speech synthesis for many languages (Qian et al., 2006; Abdel-Hamid et al., 2006; Fares et al., 2008; Bahaadini et al., 2011; Phan et al., 2013). These models fall within the category of shallow architectures which are based on a single hidden layer of non linear transformation. In the last few years, deep learning has emerged as a new area of machine learning research (Deng and Yu, 2014). Deep learning techniques are impacting a wide range of signal and image processing applications. For instance, deep learning using neural network achieved high performance in many tasks, including speech processing (speech recognition and synthesis) (Martin et al., 2013) and computer vision (Ciresan et al., 2010). Opposed to shallow techniques, deep architectures are based on many layers of non-linear transformations. With the fast development of hardware and software, it became possible to use neural networks with