Computers and Electrical Engineering 000 (2017) 1–11

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

A training-based speech regeneration approach with cascading mapping models☆

Hamid R. Sharifzadeh[a,*], Amir HajiRassouliha[a], Ian V. McLoughlin[b], Iman T. Ardekani[a], Jacqueline E. Allen[c], Abdolhossein Sarrafzadeh[a]

[a] Signal Processing Lab, Unitec Institute of Technology, Auckland, New Zealand
[b] School of Computing, The University of Kent, Kent, United Kingdom
[c] Department of Otolaryngology, North Shore Hospital, Auckland, New Zealand

ARTICLE INFO

Article history:
Received 15 February 2016
Revised 5 June 2017
Accepted 5 June 2017
Available online xxx

Keywords:
Speech reconstruction
Whispers
Electrolarynx
Laryngectomy
Time alignment

ABSTRACT

Computational speech reconstruction algorithms have the ultimate aim of returning natural sounding speech to aphonic and dysphonic patients as well as those who can only whisper. In particular, individuals who have lost glottis function due to disease or surgery, retain the power of vocal tract modulation to some degree but they are unable to speak anything more than hoarse whispers without prosthetic aid. While whispering can be seen as a natural and secondary aspect of speech communications for most people, it becomes the primary mechanism of communications for those who have impaired voice production mechanisms, such as laryngectomees.

In this paper, by considering the current limitations of speech reconstruction methods, a novel algorithm for converting whispers to normal speech is proposed and the efficiency of the algorithm is explored. The algorithm relies upon cascading mapping models and makes use of artificially generated whispers (called *whisperised* speech) to regenerate natural phonated speech from whispers. Using a training-based approach, the mapping models exploit whisperised speech to overcome frame to frame time alignment problems that are inherent in the speech reconstruction process. This algorithm effectively regenerates missing information in the conventional frameworks of phonated speech reconstruction, and is able to outperform the current state-of-the-art regeneration methods using both subjective and objective criteria.

1. Introduction

The human voice is the most magnificent instrument for communication, capable of expressing deep emotions, conveying oral history through generations, or of starting a war. However, those who suffer from aphonia (no voice) and dysphonia (voice disorders) are unable to make use of this critical form of communication. They are typically unable to project anything more than hoarse whispers [1].

Whispered speech is useful for quiet and private communications in daily life [2–4]. Unimpaired speakers occasionally use whispers to communicate in the public locations such as libraries, cinema theatres, or during lectures and meetings.

☆ Reviews processed and recommended for publication to the Editor-in-Chief by Guest Editor Dr. E. Abdel-Raheem.

* Corresponding author.
E-mail address: hsharifzadeh@unitec.ac.nz (H.R. Sharifzadeh).

http://dx.doi.org/10.1016/j.compeleceng.2017.06.007
0045-7906/

But whispered speech becomes the primary communicative mechanism for many people experiencing voice box difficulties [5]. There is no definitive estimate of the global population suffering some form of voice problem, but information from a number of studies [6,7] suggests that one third of the population have impaired voice production at some point in their lives (temporary) and further that the number of new patients with significant, long lasting voice problems (e.g. laryngectomees) are annually around 35,000 in OECD countries[1].

Patients reduced to whispering have generally lost their pitch generation mechanism [1] through physiological blocking of vocal cord vibrations or, in pathological cases, blocking through disease or exclusion by an operation. Typical prostheses for voice impaired patients (esophageal speech [8], transoesophageal puncture (TEP) [9], and electrolarynx devices [10]) allow patients to regain limited speaking ability but do not generate natural sounding speech; at best their sound is monotonous or robotised [11,12]. Additional drawbacks of traditional prostheses are difficulty of use and risk of infection from surgical insertion [13]. Thus, within a speech processing framework, recent computational reconstruction methods (and particularly whispers to phonated speech) are aiming to regenerate natural sounding speech for aphonic and dysphonic individuals. Furthermore, comparing with traditional prostheses, these methods would be non-invasive and non-surgical.

In recent years, various techniques have been proposed for converting whispers to normal speech [14–17]. The driving idea of all these methods is based on the assumption of whispers are missing some acoustic and spectral features comparing with normal speech; hence, the problem of converting whispers to normal speech is formalised as a reconstruction issue [4,18]. Through this approach, these methods aim to add or enhance missing or modified features and increase the signal similarity of whispers to normal speech. In general, these reconstruction methods can be classified into two major groups of training and non-training based methods. Utilising machine learning algorithms is the basis of training-based methods (whispers are mapped to the corresponding normal speech), while non-training methods rely upon whisper enhancement and pitch regeneration.

These reconstruction methods (either training-based or non-training) suffer from range of disadvantages including problems in converting continuous speech (due to using phoneme switching) [15], being computationally expensive (due to using highly overlapped frames for spectral enhancement, or using jump Markov linear system for pitch and voicing parameters) [4], and more importantly lack of naturalness in regenerated output (due to simplified time alignment and spectral features assumptions) [16]. Particularly, the training-based approaches suffer from intelligibility and over-smoothing problems while lack of naturalness due to time misalignment between normal and whispered speech can be seen as the main disadvantage of the state-of-the-art methods. Furthermore, different nature of acoustic and spectral features has been often ignored in training based whisper-to-speech methods, which leads to matching problems. In this paper, we focus on a training-based approach, and propose a novel reconstruction algorithm to improve the efficiency in phonated speech regeneration. In our algorithm, an intermediary layer called “artificial whisper” or “*whisperised* speech” is introduced to lessen the effect of inconsistent spectral features and time alignment between natural and whispered speech.

This algorithm effectively regenerates missing information in the conventional frameworks of phonated speech reconstruction. Results of objective and subjective evaluations demonstrate that the proposed method successfully improves the reconstructed speech quality. As an expanded version of our previous work [19], this paper presents further discussions on time alignment, provides the results of detailed subjective and objective evaluations, compares the outcome with other computational methods and electrolarynx samples, and yields further improvement by increasing the size of training datasets.

Section 2explains whisperised speech while Section 3 addresses time alignment problem and describes our reconstruction algorithm using cascading mapping models. The algorithm analysis including some examples are demonstrated in Section 4. Performance analysis and the scores of subjective and objective experiments are presented in Section 5 and finally, the paper is concluded in Section 6.

2. Whisperised speech

Whispers and natural speech have different acoustic and spectral characteristics; the most significant physical characteristic of whispers is the absence of vocal cord vibration, resulting in missing pitch and harmonics [20]. Using a source filter model [21], exhalation can be identified as the source of excitation in whispered speech, with the shape of the pharynx adjusted to prevent vocal cord vibration in normal speakers. The open glottis in whispers acts like a distributed excitation source and the turbulent aperiodic noise can be seen as the primary excitation in whispered speech [22]. Whispered vowels and diphthongs also differ from fully phonated ones. Formant frequencies tend to be higher than in normal speech [2], particularly the first formant which shows the greatest difference between two kinds of speech.

Whisperised speech or artificial whisper is a whisper-like speech which is derived from normal speech by taking pitch off (i.e. eliminating periodic glottal excitation or removing long term prediction coefficients in standard source-filter model). The basic structure of analysis and synthesis parts employed in this paper for generating whisperised speech (W') from normal phonated speech (S) is presented in Fig. 1.

Generally, there are two kinds of pitch filters: the pitch filter at the analysis stage (coder), which is a non-recursive pitch prediction filter, and the pitch filter at the synthesis stage (decoder), which is the inverse filter to the pitch prediction, i.e.

[1] OECD refers to the countries which are the members of the Organisation for Economic Co-operation and Development (OECD). The mission of the organisation is to promote policies that will improve the economic and social well-being of people around the world. Currently, 35 countries are the members of this organisation which includes many of the world’s most economically advanced countries and some emerging countries.

Please cite this article as: H.R. Sharifzadeh et al., A training-based speech regeneration approach with cascading mapping models, Computers and Electrical Engineering (2017), http://dx.doi.org/10.1016/j.compeleceng.2017.06.007