Neurocomputing 150 (2015) 367–376

Contents lists available at ScienceDirect

Neurocomputing

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

A study of glottal excitation synthesizers for different voice qualities

Jesús B. Alonso [a,*], Miguel A. Ferrer [a], Patricia Henríquez [a], Karmele López-de-Ipina [b], Josue Cabrera [a], Carlos M. Travieso [a]

[a] Instituto para el Desarrollo Tecnológico y la Innovación en Comunicaciones (IDeTIC), Universidad de Las Palmas de Gran Canaria, Despacho D-102, Pabellón B, Ed. de Eletrónica y Comunicaciones, Campus de Tafira, 35017 Las Palmas, Spain

[b] System Engineering and Automation Department, University of the Basque Country, Polytechnic School, Donostia 20008, Spain

ARTICLE INFO

Article history:
Received 15 December 2013
Received in revised form
24 March 2014
Accepted 2 May 2014
Available online 23 October 2014

Keywords:
Speech synthesis and generation
Speech perception and psychoacoustics
Speech analysis

ABSTRACT

The aim of this paper is to analyze the improvements that are observed in the glottal excitation synthesizers when the possible manifestations of non-linear behavior are characterized in glottal excitation. This paper proposes a new model based on the modification of a classic glottal excitation synthesizer and to study the improvements regarding different glottal excitation synthesizers. The proposed model tries to improve the naturalness of the synthesized voice using the synthesis of the sub-harmonics. The proposed model is included in a generic synthesizer of sustained vowels in order to get an assessment about the quality of the synthesis of different qualities of voice, where speakers with pathologies in the phonatory system are used to simulate the behavior of low quality voices. The different models are adjusted using genetic algorithms.

The assessment of the different glottal excitation synthesizers is obtained using an objective measure of similarity between the original signals and the synthesized signals based on temporal and spectral measurements. In addition, the quality of the proposed glottal excitation model is evaluated with a study of subjective perception.

1. Introduction

The quality of synthesized voice depends especially on two factors [1]: intelligibility and naturalness. Intelligibility depends on the ability of the synthesizer to reproduce formants (frequencies, band-widths and transitions) that are specific characteristics of the vocal tract, while naturalness depends mainly on the glottal excitation waveform (glottal air flow velocity) of voiced sounds.

Many features of phonatory system are manifested in the glottal excitation. For this reason, it is essential to implement properly glottal excitation waveform. Some of these features include the presence of sub-harmonics [2–5]. In this context, sub-harmonics are considered as secondary vibrational modes, in which there is a non-integer relation with the main vibrational mode. This is due to the properties of vocal folds, such as viscosity of the mucus, asymmetry in the folds or desynchronization in the movement [6,7]. Many of these characteristics are manifested in voice signal where non-linear behavior is observed. The non-linear behavior is more obvious in low quality voices [6–8]. Classical glottal excitation models do not take into account these factors, and thus they fail when synthesizing voices of different qualities, especially with poor quality.

The dynamics of vocal folds has been extensively studied for decades, obtaining several models that simulate its behavior and the generation of the glottal excitation waveforms [9,10].

There are two kinds of glottal source models: glottal models and vocal folds models. On the one hand, glottal models use glottal acoustic parameters that describe the dynamic movement of the glottal wave or its derivative. On the other hand, vocal folds models are focused on the vibration of vocal folds and on the production of the glottal wave.

1.1. Glottal models

Glottal models are known for their flexibility and precision. The Liljencrants Fant model, also known as the LF model [11], and the Veldhuis R++ model [12] stand out because they are more complex and are capable of shaping the closure of the glottis and the asymmetrical forms of glottal excitation. The model most commonly used is the LF model, which generates the derivative glottal pulse. This model reproduces a high quality synthetic speech, suggesting that the LF model provides sufficient degrees of freedom to model different type of voices such as the pathological voice [13] or the singing voice [5]. Rosenberg proposed a glottal excitation model [30] which allows the adjustment of several parameters such as amplitude, width and skewness of the pulse. Veldhuis proposed the Rosenberg++ model as an extension of Rosenberg's parametric

* Corresponding author.
E-mail address: jalonso@dsc.ulpgc.es (J.B. Alonso).

http://dx.doi.org/10.1016/j.neucom.2014.05.081
0925-2312/

model, focusing on the derivative of glottal excitation. These improvements were introduced to increase the flexibility of the original model, incorporating control over the phase and the asymmetry glottal excitation. In the model proposed by [14], the glottal excitation is represented by two polynomial segments, which allows that the degree of detail of the model is easily modified. The model [15] is a relatively simple model that describes the derivative of the glottal excitation and it is used due to two reasons: it is generated with a simple second order polynomial model and it allows encapsulating the effects of radiation on the lips. Other model presented in [16] suggests synthesizing glottal excitation through a shaping function. This function transforms a trigonometric function in the desired function. It uses a nonlinear function without memory that transforms a cycle of a harmonic in the desired waveform.

1.2. Vocal folds models

Vocal folds models can be grouped into four main categories:

- Mechanical models, where the glottal source is represented by mechanical oscillators. The one mass model [17], the two-mass model [18] and the standard multiple masses model [19] are all included in this group.
- Continuum model [10], where the vocal folds are represented as a continuous and deformable means.
- Ribbon model [10], where the vibration of the vocal folds is depicted through the movement of a tape.
- Body-cover models [20], where the vocal fold is divided into two masses. The body primarily consists of muscle mass and the tissues that cover it.

Mechanical models in general and the two-mass model in particular are the best due to its high degree of efficiency without requiring complicated calculations. In the one mass model, the vibration of the vocal folds is simulated from a system composed of a single mass, a spring and a piston. Modified models that focus on the interaction source–tract have been also developed [21]. The two-mass model provides a realistic simulation of the glottis properties. The vocal folds are divided into an upper mass and a lower mass motivated for the anatomical and functional division between the mucosa and the basis of the vocal folds. Each part consists of an oscillator composed of a mass, a spring and a piston. The springs represent the elastic properties of the vocal folds, while the piston represents dissipative forces, such as viscosity and friction. In addition, there is an interaction between the two masses, which is represented by a rigid coupling. The springs have a non-linear feature which simulates the stiffness in the vocal folds. During the glottal closure, there is a contact force resulting in the deformation. The two-mass model provides a realistic simulation of the glottal properties, allowing the reproduction of a natural voice with a reasonable computational load. The two-mass model cannot be used for the various modes of longitudinal vibration observed in the human phonation. Titze attempted to improve this model proposing one model with 16 masses, two rows of eight masses. The masses of the upper row primarily represent the mucosa, while the masses of the lower row are used for the vocal ligament and the vocalise muscle (the basis of the vocal folds). In addition, longitudinal tensions are represented. The constant of the springs from both rows increases in a non-linear direction with respect to the elongation of the vocal folds. The multiple masses model is a complex model with a high computational load. It is possible to adjust parameters that have a direct physiological correlation, thereby increasing the naturalness of the sound. Other authors [22,23] have proposed a two-mass model where each mass represented one of the vocal folds, differentiating between the left fold and the right fold. This model characterizes the elasticity and viscosity of the vocal folds. The masses move to the right and left with different speeds. If the conditions for the glottal excitement are sufficient, both folds collide and ricochet. However, the simplicity in the model does not provide an explanation for a very important phenomenon, namely the effect of the vibration of the mucosa (mucosal wave effect). This effect occurs in the vocal folds because of the distribution of the mass in the tissues of the folds [24]. This can be described as a second mode of vibration, superimposed on the first mode. An explanation for this phenomenon can be obtained by using the two-mass model, but with two masses for each fold. In addition, there are three springs per fold representing the elastic behavior of both masses and their coupling. Viscosity losses are represented with two parameters for each fold. The presence of the mucosa vibration can be estimated using this model and inverse filtering techniques [22]. The two-mass model proposed by Ishizaka and Flanagan has led to additional models. For example, Drioli and Avanzini [25] proposed a nonlinear model of glottal excitation using a reduced number of parameters, based on a glottal excitation waveform estimated, starting with the vocal tract filter parameters and using inverse filtering techniques.

In summary, the different voice synthesizer models that have been developed do not take into account the manifestations of non-linear behavior (sub-harmonics) that are present in the voice signal or do not implement vocal fold models using multiple masses which generate non-explicit sub-harmonics. However, multiple studies [2–8] indicate the presence of sub-harmonics in voice signals, mainly on speeches with low voice quality. This paper recognizes the need to investigate the presence of sub-harmonics in the glottal excitation for certain types of voices. To give an example of applications, parametric coders can be developed with a high degree of naturalness of synthesized voices, especially taking into account the importance of speech processing with pathological voices in a world with an increasingly aging population [31]. This kind of tools could be applied for example in the rehabilitation of pathological voices.

In this paper, we propose that the performance of a glottal excitation synthesizer improve regarding the naturalness when a fundamental harmonic and a sub-harmonic are used in the generation of glottal excitation. The utility of this approach is demonstrated by means of a comparative study between different models of glottal excitation. In Section 2 we introduce a new model of glottal excitation that generates a fundamental harmonic and sub-harmonics. Section 3 describes the methodology and Section 4 shows the data used in the study. Results and conclusions are presented in Sections 5 and 6.

2. Proposed system

We propose to add a small improvement to a classic glottal excitation model which allows the possibility of synthesizing sub-harmonics. We have chosen to use the LF model for its robustness and simplicity. This model has five parameters t_p, t_e, t_c, t_a and t_0 which are described in [11]. Fig. 1 shows the proposed glottal excitation model:

The proposed glottal excitation model generates a signal $P(n)$ that includes the sum of two signals:$P_1(n)$ and $P_2(n)$. The first signal $P_1(n)$ is a complete cycle of glottal flow signal generated with the LF1 model ($S_1(n)$), with the same frequency than the fundamental harmonic (frequency of vibration of the vocal folds) and with amplitude A_1. LF1 model has five parameters: t_{p1}, t_{e1}, t_{c1}, $t_{a1}$and t_0(the same as the LF model). The second signal $P_2(n)$ is a double cycle of glottal flow signal with the same frequency than a pseudo-subharmonic of the fundamental harmonic, which it is called sub harmonic in this paper and it is generated with LF2 model ($S_2(n)$), and where each cycle has amplitude A_2 and A_3 respectively. The frequency of the sub-harmonic is related to the