Speech Communication 82 (2016) 85–96

Contents lists available at ScienceDirect

Speech Communication

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

Extension of the single-matrix formulation of the vocal tract: Consideration of bilateral channels and connection of self-oscillating models of the vocal folds with a glottal chink

Benjamin Elie*, Yves Laprie

LORIA, INRIA / CNRS / Université de Lorraine, Vandoeuvre-les-Nancy, France

ARTICLE INFO

Article history:
Received 14 September 2015
Revised 1 June 2016
Accepted 6 June 2016
Available online 17 June 2016

Keywords:
Speech synthesis
Vocal folds
Glottal chink
Lateral consonants

ABSTRACT

The paper presents extensions of the single-matrix formulation (Mokhtari et al., 2008, Speech Comm. 50(3) 179 – 190) that enable self-oscillation models of vocal folds, including glottal chink, to be connected to the vocal tract. They also integrate the case of a local division of the main air path into two lateral channels, as it may occur during the production of lateral consonants. Provided extensions are detailed by a reformulation of the acoustic conditions at the glottis, and at the upstream and downstream connections of bilateral channels. The simulation framework is validated through numerical simulations. The introduction of an antiresonance in the transfer function due to the presence of asymmetric bilateral channels is confirmed by the simulations. The frequency of the antiresonance agrees with the theoretical predictions. Simulations of static vowels reveal that the behavior of the vocal folds is qualitatively similar whether they are connected to the single-matrix formulation or to the classic reflection-type line analog model. Finally, the acoustic effect of the glottal chink on the production of vowels is highlighted by the simulations: the shortening of the vibrating part of the vocal folds lowers the amplitude of the glottal flow, and therefore lowers the global acoustic level radiated at the lips. It also introduces an offset in the glottal flow waveform.

1. Introduction

Time-domain continuous speech synthesizers are commonly based on simplified physical models to compute the acoustic propagation along the vocal tract (Birkholz, 2014; Kelly and Lochbaum, 1962; Maeda, 1982; Mokhtari et al., 2008; Story, 2013) and/or the self-sustaining oscillations of the vocal folds (Erath et al., 2011; Ishizaka and Flanagan, 1972; Moisik and Esling, 2014; Pelorson et al., 1994). In comparison with finite-element-based methods (Fleischer et al., 2015), which require a huge amount of time, their low computation time makes them interesting for continuous speech synthesis.

Simplified acoustic models use the plane wave assumption to compute the acoustic propagation along a set of acoustic tubes. The dimensions of the elementary tubes (or *tubelets*) approximate the geometry of the vocal tract. In regards to the typical dimensions of the human vocal tract, these models are valid up to frequencies around 5 kHz (Story, 2004).

Articulatory synthesis bridges the gap between the articulatory and acoustic domains of speech. This is thus an invaluable tool to apprehend the acoustic impact of the speech articulator's gestures and their temporal coordination, that of the anatomic characteristics of the human vocal tract, and of the interactions between the vocal folds and the vocal tract. In order to enable speech production to be studied via articulatory synthesis, several aspects should be covered by the numerical simulations of the speech aerodynamic/acoustic phenomena. First, the complexity of the vocal tract should be accurately modeled so that the various cavities (nasal tract, paranasal sinuses, sublingual cavities, etc.) can be taken into account during the simulation. Then, the simulation framework should be able to deal with time-varying geometries of the vocal tract in order to simulate word-level or phrase-level utterances. This constrains the time trajectory of each articulator to be accurately modeled. Finally, the acoustic coupling between the glottal source, i.e. the vocal folds, and the vocal tract needs to be realistically modeled.

So far, there is no known time-domain continuous speech synthesizer that can deal with all these constraints. The scientific literature about speech synthesis based on simplified physical models identifies two main techniques: the *reflection-type line ana-*

* Corresponding author.
E-mail address: benjamin.elie@inria.fr (B. Elie).

http://dx.doi.org/10.1016/j.specom.2016.06.002
0167-6393/

log method (Kelly and Lochbaum, 1962), which is called RTLA in this paper, and the *transmission line circuit analog* (Maeda, 1982) method, called TLCA.

The reader may find a detailed review of existing techniques for speech synthesis in Kröger and Birkholz (2009). Basically, RTLA has the advantage of accurately accounting for the frequency dependence of acoustic losses, but suffers from the constraints on the tubelet dimensions in regards to the chosen simulation frequency. As a consequence, the total length of the vocal tract cannot be modified during the simulation. This is an important issue for continuous speech synthesis since the length of the vocal tract varies during natural speech production. Its use is usually limited to studies about the self-sustained motion of the vocal folds (Bailly et al., 2008; Lous et al., 1998; Story and Titze, 1995) coupled with simplified acoustic resonators. Using RTLA to simulate running speech constrains the vocal tract to unrealistic simplified geometries (Story, 2013). On the other hand, many continuous speech synthesizers use TLCA (Birkholz and Jackèl, 2004; Ishizaka and Flanagan, 1972; Laprie et al., 2014; Maeda, 1982). It is based on the electric-acoustic analogy: the vocal tract acoustics is seen as a lumped electric circuit. The main advantage of TLCA is its flexibility of use with time-varying geometries of the vocal tract, including length variation and uneven spatial sampling of the vocal tract. However, this analogy does not allow the frequency dependence of the acoustic losses and the acoustic radiation to be accurately taken into account. Recently, the *Single-Matrix Formulation* (SMF) (Mokhtari et al., 2008) has been a major contribution to TLCA models: it is now possible to compute the acoustic propagation along a vocal tract modeled as a waveguide network, where each waveguide represents a side cavity. Consequently, using SMF is a useful tool to study the acoustic effects of the numerous side cavities in the context of continuous speech synthesis.

Another important challenge to tackle when dealing with articulatory synthesis is the glottal source model. Indeed, in order to thoroughly study the phonatory mechanisms, one should include a glottis model that is able to realistically account for the coupling between the vocal folds and the vocal tract. Many efforts to simulate the production of the glottal source have been made, and self-oscillating models of vocal folds, based on lumped mass-spring systems, have been of particular interest. The reader may find detailed reviews of these models in Birkholz et al. (2011); Erath et al. (2013). Curiously, most of these studies have neglected the possibility of including a posterior glottal opening to simulate air leakage. This glottal opening, also called *glottal chink*, allows a DC component to appear in the glottal flow waveform, as commonly observed *in vivo* (Klatt and Klatt, 1990). A partial closure of the glottis during the oscillating cycle of the vocal folds may be useful to simulate voiced fricatives (Elie and Laprie, 2016) and breathiness, which is an important acoustic cue, especially for gender identification (Klatt and Klatt, 1990).

First attempts at modeling a glottal chink have used parametric models (Cranen and Schroeter, 1995; 1996). In these papers, two models of glottal leakage are proposed: firstly, the glottal chink is caused by a partial abduction of the vocal folds, i.e. only a portion of the vocal folds vibrates, the other part is abducted and forms a triangular glottal chink, and secondly, the glottal chink is formed in the inter-arytenoid portion of the glottis. In the second case, the vocal folds vibrate along their whole length. Later on, (Wilhelms-Tricarico, 1994) proposed a modification of the classic two-mass model by Ishizaka and Flanagan (1972) to include the glottal chink by connecting an electric branch in parallel to the vocal fold model. The glottal system is then connected to the first resonance of the vocal tract. Recently, (Zañartu et al., 2014) have studied the effect of the glottal chink on self-oscillating movements of the vocal folds. Although this study is a significant advance in glottis modeling, its connection with RTLA models of acoustic propagation does not make it suitable for continuous speech synthesis with realistic time-varying geometries of the vocal tract.

Starting from the single-matrix formulation presented in Mokhtari et al. (2008), this paper details the theory and the methodology for extending it by overcoming its limitations. The limitations of SMF are the following: it does not offer the possibility to connect a self-oscillating model of the vocal folds, and the configuration of anastomozing waveguides, i.e. the local division of the main oral tract into two lateral channels, as it may occur during the production of lateral consonants, is not discussed. The aim is then to propose a complete simulation framework for speech synthesis that can account for the complexity of the vocal tract geometry and its numerous cavities taken simultaneously, that can deal with a time-varying realistic model of the vocal tract, including length variation, and that realistically models the acoustic coupling between the glottal source and the vocal tract, including a glottal leakage.

The paper is organized as follows. The transmission line circuit analog and the original single matrix formulation are detailed in Section 2. It also includes the required acoustic conditions at the glottis for integrating self-oscillating models of the vocal folds. The main aspects of the simulation framework, called *Extended Single-Matrix Formulation* (ESMF), are detailed in Section 3. They consist in the mathematical formulations for introducing the case of anastomozing waveguides into the single-matrix formulation, as well as the mathematical formulations to connect self-oscillating models of the vocal folds and a glottal chink to the single-matrix formulation. Finally, Section 4 presents numerical simulations that illustrate the accuracy of the extended single-matrix formulation to deal with the new features.

2. Theoretical background

This section summarizes the single-matrix formulation of the vocal tract by Mokhtari et al. (2008), which is itself derived from the *transmission line circuit analog* model by Maeda (1982). The present paper provides modifications in the formulation, taking into account the internal resistance of a noise source pressure to simulate frication noise. This reformulation is motivated by the fact that many quantities introduced in this section are used to demonstrate the contributions detailed in the next section. Yet, for the sake of brevity, not all computation details have been provided, and one may refer to the original papers (Maeda, 1982; Mokhtari et al., 2008) for more details.

2.1. Transmission line circuit analog model

The vocal tract is modeled as a concatenation of cylindrical tubes (or *tubelets*) along which plane waves propagate. The length and the cross-sectional areas of the tubelets are such that they approximate the vocal tract geometry. TLCA represents each tubelet as lumped circuit elements. Fig. 1 shows the lumped circuit elements of a single tube section and Table 1 details the acoustic-electric analogy.

The terms W_R, W_C, and W_L are constant terms denoting respectively the resistance, the stiffness, and the mass of the vocal tract walls per area unit. Chosen values for this study are those provided in Birkholz and Jackèl (2004), namely $W_R = 8000$ kg.m^{-2}.s^{-1}, $W_C = 8.45 \times 10^6$ kg.m^{-2}.s^{-2}, and $W_L = 21$ kg.m^{-2}. By convention, indices follow the air flow direction. For instance, considering the vocal tract, index 1 denotes the glottis connection, and index N denotes the lip termination.

Note that unlike in Maeda (1982); Mokhtari et al. (2008), lumped circuit elements include a frication noise source. It is made up of a pressure source P_{n_i}, with an internal resistance R_{n_i} (Birkholz, 2014; Maeda, 1996; Sondhi and Schroeter, 1987),