A system for the comparison of glottal source models for articulatory speech synthesis

Peter Birkholz and Christiane Neuschaefer-Rube
Clinic of Phoniatrics, Pedaudiology, and Communication Disorders
University Hospital Aachen and RWTH Aachen University
pbirkholz@ukaachen.de, cneuschaefer@ukaachen.de

Introduction

Natural sounding speech synthesis poses one of the most difficult challenges for a biomechanical model of the vocal folds, because it must perform convincingly in multiple tasks. For example, it must oscillate over a frequency range of at least two octaves for a variety of supraglottal loads, simulate different voice qualities, behave realistically during phonation onset and offset, and finally, sound natural. Only few of the existing models were actually evaluated with respect to the quality of the produced voice in this context, or compared with each another under the same conditions (Birkholz, 2011). We present an approach to embed different vocal fold models into the articulatory speech synthesizer VocalTractLab (www.vocaltractlab.de) based on a common control interface. This allows to easily exchange the vocal fold model used for the synthesis of an arbitrary utterance and so to compare the performance of different models under the same conditions in running speech.

Method

Models of the vocal folds usually differ in the number and function of their control parameters. For example, some models are controlled by muscle activation patterns, while other models are controlled by physical parameters, e.g. glottal rest area and vocal fold tension. To make these models exchangeable in our articulatory speech synthesizer, a common model-independent interface for the dynamic control of these parameters is required. Our approach is to assume a fixed set of glottal rest configurations for different phonation types and functions. We distinguish, for example, rest configurations for modal phonation, pressed phonation, breathy phonation, whisper, a glottal stop, and abducted vocal folds for voiceless sounds. Each implemented vocal fold model must provide an individual setting of its parameters to realize the corresponding phonation type or functions. At the model-independent level of control in the synthesizer, laryngeal articulation is specified as a sequence of discrete gestures, each of which represents movement toward one of the glottal rest configurations at an adjustable speed. The change of the parameters is governed by critically damped 6th order linear systems to ensure smooth shape transitions. Fundamental frequency is controlled independently from the other parameters by a sequence of F_0 targets. Therefore, each implemented vocal fold model must provide a mechanism to map a given F_0 on the affected parameters, e.g. a vocal fold tension parameter. For the aerodynamic-acoustic simulation of the vocal system based on a non-linear transmission-line circuit model, the vocal folds are represented as two abutting tube sections with time-varying cross-sectional areas. Hence, each implemented model must define a mapping from the geometry of the glottis to the areas and lengths of these tube sections.

Results

Currently, one geometrical vocal fold model and two variants of the two-mass model are implemented in the system. The models can be exchanged without difficulty for the articulatory synthesis of arbitrary utterances to evaluate the sound of the voices in running speech.

Discussion

We hope that the presented system will facilitate the future development and evaluation of vocal fold models by their integration into a complete articulatory speech synthesizer based on a common control interface.

References

Peter Birkholz (2011). A survey of self-oscillating lumped-element models of the vocal folds. In: B. J. Kröger, P. Birkholz (eds.) *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2011* (TUDPress, Dresden), pp. 47-58