A PHYSICAL MODEL FOR VOICE SYNTHESIS

43.70.Bk

Kob, Malte
Dep. of Phoniatry, Pedaudiology and Communication Disorders
University Hospital Aachen
Pauwelsstrasse 30
D-52074 Aachen, Germany
Tel: +49-241-80-88956
Fax: +49-241-80-82513
E-mail: mkob@ukaachen.de

ABSTRACT

The human voice can be described as the joint product of several functional components of the voice organ. A time-domain model is presented that includes a multiple-mass vocal fold model, noise generation, wave propagation through the vocal tract, and radiation at the mouth. A graphic user interface allows the modification of parameters like lung pressure, glottis configuration and vocal tract shape during the calculation. The application of the model to singing voice synthesis (e. g. overtone singing) and generation of pathologic vocal fold movement is demonstrated.

INTRODUCTION

A satisfactory on-line modeling of voice based upon the physiology of an individual speaker is not yet possible, even though the advantages of such an approach are tempting:

- the synthesis of individual voices could be based upon the actual physiology
- speech transmission could use a set of physically-based parameters for increased naturalness of the speech
- a comparison of recorded voices with results from simulations could give diagnostical hints

Limitations of physical models available today are the computational costs for an accurate simulation of the physiology of the voice organs. Especially the three-dimensional description of the flow in the glottis is crucial for an accurate modeling of the vocal fold movement [1]. More restrictions occur in the availability of geometrical and physical properties of the voice organ. This work presents a model implemented in MATLAB for easy implementation and extension of algorithms and visualization of their results.

MODEL COMPONENTS

For singing voice synthesis a combined model, named VOX, was used, consisting of separate models for vocal folds (VF), vocal tract (VT) and noise generation. The degree of interaction between the models and the choice of parameters can be controlled in the graphical user interface (GUI) with buttons and sliders that change parameters such as:

- Reflected flow from glottis back into VT, noise generation, radiation impedance
- Adjustment of lung pressure, vocal fold and vocal tract properties

Future work will provide switches for the reflected flow from the VT into the glottis and the reflected flow from the lungs into the glottis.

The signal flow of the combined model is given in Figure 1. After initialisation of the global parameters the vocal fold signal is calculated sample-wise. The sound pressure wave is fed into the vocal tract model and the resulting pressure wave at the mouth is stored.

All user-defined parameters are accessible during calculation and allow an arbitrary change of the boundary conditions.

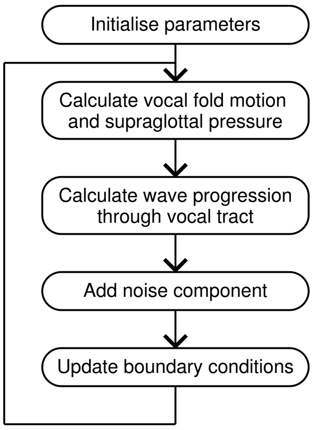
Initialise parameters
Calculate vocal fold motion and supraglottal pressure
Calculate wave progression through vocal tract
Add noise component
Update boundary conditions

Figure 1: Signal flow of the combined model VOX

Vocal Folds

The model presented in this thesis is based on the 16-mass model developed by I. R. Titze [2] but includes some more recently published modifications. The values of parameters and the equations are discussed in [3]. The implemented model offers the following features:

- Modeling of modal, head and falsetto register is possible
- Variation of individual masses or spring stiffness in an arbitrary number of sections allows simulation of e. g. singer's nodules
- All parameters can be changed and pressure or flow values can be monitored during calculation

Figure 2 illustrates the geometry of the vocal fold model.

The arrangement of n masses $m_{m,i}$, with $i = 1..n$ represents the vocalis muscle, $m_{v,i}$ represents the mucosa membrane. Each of the big masses $m_{v,i}$ is connected to the boundary by a spring with stiffness k^b and damping D^b, and to the small mass $m_{m,i}$ by a spring with stiffness k^m and damping D^m.

The most important difference between this model and classical two-mass models is the segmentation of the VF in n segments. In Titze's model, the number of 16 masses is fixed. The sum of the arbitrary number of n segments of width a is the longitudinal length l_g of the VF. There are no springs between the border and the small masses m_m, which is more close to the actual configuration of the vocal folds, because the mucosa membrane is not directly attached to the *cricoid cartilage*.

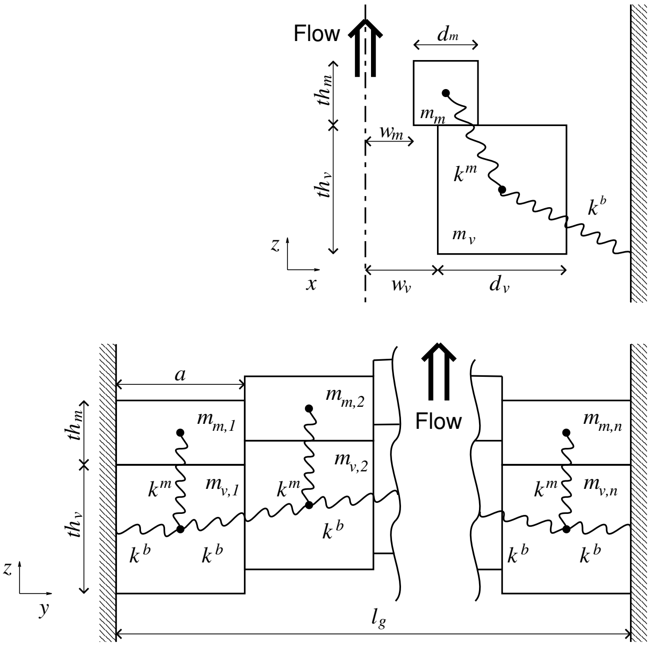

Figure 2: Sectional view (top) and side view (bottom) of the implemented vocal fold model

Vocal Tract

Two models have been implemented: a waveguide approach based on the model of J. L. Kelly and C. C. Lochbaum (KL) [4] and a model using a multiconvolution algorithm (CTIM).

Reflection type line analog

The implemented waveguide model is based upon the general structure of the iterative approach by Barjau [5]. Only frequency-independent losses are taken into account by multiplication of the transmitted pressure wave by a damping factor.

In Figure 3 the signal flow for the waveguide calculation is shown. For each sample, the user-defined parameters are imported from the main program *VOX*. Depending on the spacing of the vocal tract data, the sample rate is calculated. When the noise generation module is active, the sound pressure in appropriate segments is augmented by turbulent noise. Now the vectors are shifted and the loop is repeated until the time limit is reached.

The algorithm can be used in interactive mode for the generation of single pressure samples or in stand-alone mode for the calculation of vocal tract transfer functions (VTTF). A more detailed description and the MATLAB code of the algorithm is given in [3].

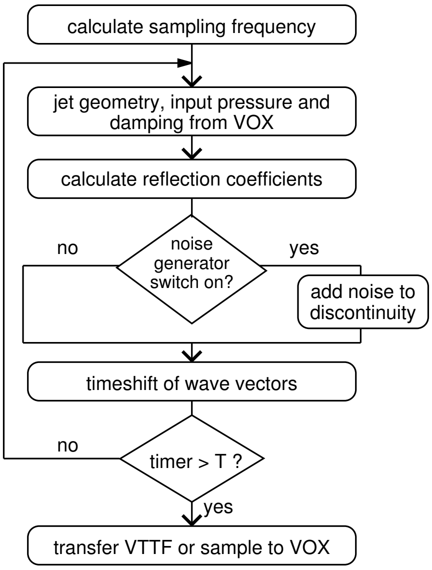
calculate sampling frequency
jet geometry, input pressure and damping from VOX
calculate reflection coefficients
no
noise generator switch on?
yes
add noise to discontinuity
timeshift of wave vectors
no
timer > T ?
yes
transfer VTTF or sample to VOX

Figure 3: Algorithm of the waveguide model

Multiconvolution technique

The algorithm of A. Barjau *et al.* [5] is an improved multiconvolution technique based on the work of J. Martínez *et al.* [6, 7]. Whereas the time-domain algorithm described by Martínez requires a spacing of the discontinuities that is restricted to multiples of phase velocity times time step, the interesting aspect of the CTIM (continuous-time interpolated multiconvolution) algorithm is its independence of the spatial VT discretization and the sample rate. Originally, the algorithm has been invented for the impedance calculation of woodwind instruments with toneholes and moderately changing cross-sections. The calculation scheme differs from the waveguide approach in the following aspect: An update of the tube segment geometry is necessary after changes in the vocal tract geometry only. Therefore, the constant parameters that are used for the convolution routine can be calculated once before the loop is entered. As a consequence, the algorithm is much faster than the waveguide approach. However, for certain geometries the CTIM algorithm does not produce stable results.

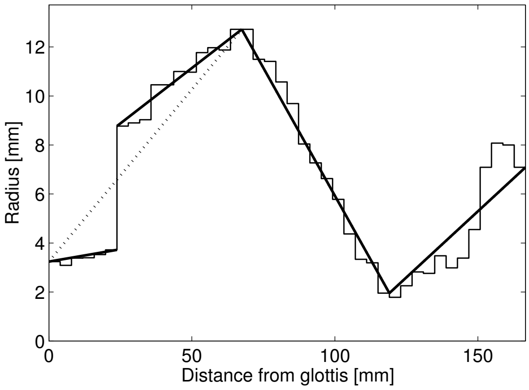

Figure 4: Comparison of an EAF used in the waveguide (thin) and in the CTIM model (thick)

The calculation of the tube segments is done with the help of a graphic user interface and allows the instantaneous interpretation of the results during construction. A pseudo-MATLAB code of the algorithm is given in [3]. In Figure 4 the equivalent area function (EAF) used for each method is depicted for the vowel [i:]. The dashed line indicates a less accurate approximation with 3 instead of 4 segments.

Noise Generation

During voice production noise is generated either when an air stream is directed against an obstacle or when a divergent configuration or a discontinuity of the bounding walls causes a laminar flow to separate from the boundary. In both cases the laminar flow is disturbed and vortices are shed that travel some distance and then break down and dissipate their energy in turbulent air movement.

The noise generation algorithm used here is based on the work of D. J. Sinder [8]. In this implementation each vortex is programmed as a struct in which the parameter values are the struct elements. Equation (1) gives the relation between the expectation value for the shedding interval $E\left\{T_{shed}\right\}$ and vortex diameter D, Strouhal number St and jet velocity U_j.

$$E\left\{T_{shed}\right\} = \frac{D}{StU_j}. \tag{1}$$

The Strouhal number $St = \frac{f_{shed}D}{U_j}$ is a dimensionless measure for the ratio of acceleration due to the unsteadiness of the flow and convective acceleration due to the non-uniformity of the flow [9]. It corresponds to the mean frequency of the generated noise spectrum. From the actual parameters flow, EAF and glottal minimum area the flow velocity is calculated.

The shedding interval T_{shed} is calculated according to equation (1) and, if the time difference between the last vortex generation exceeds T_{shed}, a new vortex with initial velocity 0 $\frac{m}{s}$ is generated. For each vortex, position and diameter are then calculated from the actual flow parameters. The vortex production rate increases when the flow increases or when the diameter of the opening decreases.

$$P_{noise} = \frac{-\pi d\rho_0}{A}[\vec{\Gamma} \times \vec{v} \cdot \vec{U}]. \tag{2}$$

In equation (2) $\Gamma = \frac{1}{2}v^2T$ is the vortex rotation and $\vec{U}$ denotes the normal component of the surrounding flow. According to equation (2) the contribution of the vortex to the turbulent sound is calculated.

During propagation the object properties are updated until the conditions for the life of the vortex are expired.

Two significant differences to Sinder's algorithm have been implemented. At first, the assumption of energy loss of the vortex during propagation has been added. It is assumed that the vortex energy is reduced exponentionally with growing distance at an estimated rate of 40 dB for a travel distance of $4\,D$. As a second difference the source smoothing algorithm has not been implemented.

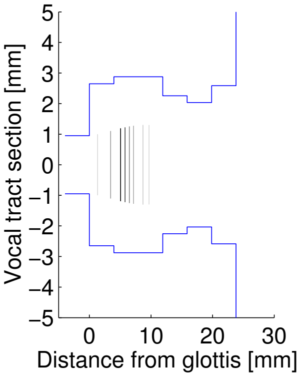

Figure 5: Propagation of vortices in the vocal tract

In Figure 5 the propagation of vortices through the vocal tract is depicted for the configuration of the vowel [æ:]. The glottis is indicated as the negative section in the drawing, whereas the positive section represents the supraglottal space. The noise contribution of each vortex is represented by its gray scale, darker lines indicate a stronger noise production. It was found that the aspiration noise contribution is significantly affected by the vocal tract geometry. This findings are in agreement with measurement results [3].

APPLICATIONS

Two examples for the application of the above model are presented here: the modeling of the VT transfer function of an overtone singer and the modeling of the vocal fold movement of a singer with vocal fold nodules.

More results for other configurations of the voice organ have been obtained such as sustained vowels in different registers [3].

Modeling of Overtone Singing

The synthesis of overtone singing aims at a verification of the modelled process of sound generation. In a first step the vocal tract geometry is modelled using MRI data that were obtained by Adachi and Yamada [10] from a Xöömij overtone singer for four different melody pitches F_6, G_6, A_6, C_7 which equal 1397 Hz, 1568 Hz, 1760 Hz and 2093 Hz.

In Figure 6 the comparison of the original and the modified area functions of the overtone A_6 as well as the corresponding vocal tract transfer functions (VTTF) are shown. The area functions taken from S. Adachi (thin solid, visible at the mouth) have been modified at the mouth opening (thick solid). The VTTF calculation has been carried out using the CTIM algorithm.

An improvement of the second resonance by about 15 dB could be achieved by matching two resonance frequencies [3, 11]. Similar results can be obtained using the KL model, although the amplification is less impressive compared to the CTIM model. An auralization of the sound using the VF model in modal register configuration and the above configuration of the VT is clearly identified as a sound with two pitches.

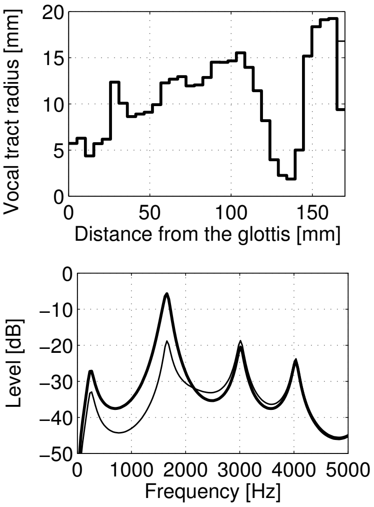

Figure 6: Area functions (top) and VTTF (bottom) of overtone A_6

Modeling of Voice Pathologies

Singer's nodules are characterized by a local increase of tissue, in most cases on opposite locations of the vocal folds. In the set-up for the calculation of the voice with a singer's nodule one mass of one segment out of 11 vocal fold segments is increased. A factor of the value 3 has been chosen, both for vocal fold masses and the mucosa masses.

The observation of the animation of the masses reveals an irregular VF movement that starts up regular with the first (1,0) mode. After a few cycles, the oscillation shifts to the second (longitudinal) (2,0) mode and stays in a periodic but rather complex movement pattern. The fundamental frequency is divided by two compared to the simulation of a healthy voice.

The sound pressure and the spectrum of the resulting signal are depicted in Figure 7. The signal exhibits a significantly lower fundamental frequency that is due to the irregular VF movement pattern in the (2,0) mode.

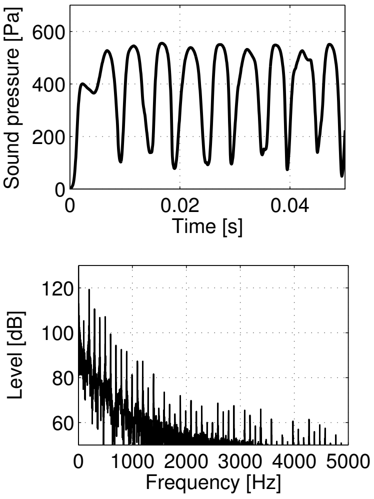

Figure 7: Supraglottal pressure for fold movement with nodule

In contrast to the simulations the *in vivo* observation of the vocal folds in patients suffering from singer's nodules does not necessarily exhibit an irregular VF movement. This might be caused by an automatic compensatory parameter change.

It was found that the irregular movement can be turned into a regular one by increasing the active stress $T_{v,act}$ on the *vocalis* muscle. With increasing $T_{v,act}$, the fundamental frequency rises and, at a certain value, the oscillation falls back to the first (1,0) mode. It was found out by successive increasing of $T_{v,act}$ that the change between the two states of oscillation takes place when the active stress is increased by ca. 5%.[1]

[1] Sound examples and animations of the vocal fold movement for these simulations and pictures of the graphic user interface can be found on the Internet [12].

CONCLUSIONS

This work presents a physical model for voice synthesis that serves both as a research device and as an educational tool for modeling, visualization and auralization of the human voice. The challenge of building such a model is to find a trade-off between a detailed description of the physiology and the physical relations that are involved on one hand and, on the other hand, the implementation of fast and stable algorithms. Due to the complexity of the voice generation process only the most important models for generation of sustained vowels have been implemented so far. The focus was laid on the mathematical description of the main components of voice generation with an emphasis on singing voice synthesis. As a consequence, no attempt has been made to "improve" the generated signals by means of post-processing like filtering to obtain a more realistic sound.

For all models, algorithms described in literature were modified, implemented in MATLAB and checked with respect to their suitability for singing voice synthesis. Modifications have been applied to the models to include latest developments. These models are coupled in several ways. The influence of their coupling has been investigated and it could be shown that the simple oscillator-filter approach is of limited accuracy [3]. The combined model includes generation of noise that is inserted into the vocal tract at approptiate locations.

The application of the model to the synthesis of overtone singing gives insight into the physics of this amazing singing style. The modeling of a pathologic vocal fold configuration demonstrates the relation between changes in the vocal fold parameters and the resulting sound wave.

Future work could aim at the implementation of an asymmetric vocal fold model, a lung model, and a fast but stable algorithm for the wave propagation through the vocal tract.

References

[1] X. Pelorson *et al.* (1995): Description of the flow through in-vitro models of the glottis during phonation. acta acustica **3** 191-202.

[2] I. R. Titze (1973): The Human Vocal Cords: A Mathematical Model Part I. Phonetica **28** 129-170.

[3] M. Kob (2002): Physical modeling of the singing voice. Dissertation, University of Technology Aachen, in press.

[4] J. L. Kelly and C. C. Lochbaum (1962): Speech synthesis. Proceedings of the 4^{th} International Congress on Acoustics, 1-4.

[5] A. Barjau, D. H. Keefe and S. Cardona (1999): Time-domain simulation of acoustical waveguides with arbitrarily spaced discontinuities. J. Acoust. Soc. Am. **105** (3) 1951-1964.

[6] J. Martínez and J. Agulló (1988): Conical bores. Part I: Reflection functions associated with discontinuities. J. Acoust. Soc. Am. **84** (5) 1613-1619.

[7] J. Martínez, J. Agulló and S. Cardona (1988): Conical bores. Part II: Multiconvolution. J. Acoust. Soc. Am. **84** (5) 1620-1627.

[8] D. J. Sinder (1999): Speech Synthesis Using an Aeroacoustic Fricative Model. Ph.D. Thesis, University of New Jersey.

[9] A. Hirschberg (1992): Some fluid dynamic aspects of speech. Bulletin de la Communication Parlée **2** 7-30.

[10] S. Adachi and M. Yamada (1999): An acoustical study of sound production in biphonic singing, Xöömij. J. Acoust. Soc. Am. **105** (5) 2920-2932.

[11] M. Kob (2001): Untersuchung der Eigenschaften des Obertongesangs. Fortschritte der Akustik – DAGA 01, 436-437.

[12] URL: http://www.akustik.rwth-aachen.de/˜malte/vox