Pole-Zero Modeling of Vocal Tract for Fricative Sounds

Minsheng Liu, Arild Lacroix
Institut fuer Angewandte Physik, University of Frankfurt
Robert-Mayer-Str.2-4; 60325 Frankfurt am Main, Germany
e-mail: Liu@iap.uni-frankfurt.de,
Lacroix@iap.uni-frankfurt.de

ABSTRACT

This paper presents a pole-zero model based on a multi-tube acoustic model for fricative sounds. This model consists of the front and back cavity formed by oral tract and pharynx, in which the excitation source is located at the point of constriction. The transfer function of this model including poles and zeros is derived and its properties are investigated. Small losses such as viscous friction which is an important for the fricative sound in the vocal tract are considered and the results show, if the vocal tract is lossless, the numerator part of the pole-zero model is symmetric. The transfer function with small losses overcomes the limitation of the symmetry. This method is applied by employing the inverse filtering and an adaptive algorithm to analyse fricative sounds.

1 INTRODUCTION

For fricative sounds the excitation source is inside the vocal tract, where the acoustic waves propagate in two directions. Thus the vocal tract is separated by the source of excitation at the constriction into two cavities. The back cavity traps energy and introduces antiresonances, resulting in zeros in the transfer function. In this case the LPC method widely used in the processing of speech signals is not properly adapted. The method of acoustic modeling has its advantages [1] and in [3] a simple model for fricative sounds is discussed. The source location for fricative sounds has been investigated [2]. In [4] an improved vocal tract model is used to represent nasal sounds. Till now transfer functions characterized by anti-resonances as well as resonances for fricative sounds have not been derived and computed on the basis of the acoustic tube model. In this contribution the acoustic tube model is improved by using a three-port adaptor at the position of the excitation, therefore a proper transfer function including zeros in the frequency response is derived and zeros can be calculated from the numerator polynomial of the transfer function.

2 POLE-ZERO MODELING FOR FRICATIVE SOUNDS

The unvoiced fricatives [f], [s] and [$\int$] are produced by a steady air flow which becomes turbulent in the region of a constriction within the vocal tract. The location of the constriction determines which fricative sound is produced. For the fricative [f] the constriction is near the lips, for [s] it is near the teeth; and for [$\int$] it is near the back of the oral tract. Sounds are radiated from the lips. In this case the vocal tract can be characterized by the acoustic tube model depicted in fig. 1. The excitation source is located at the junction of the section between two tubes.

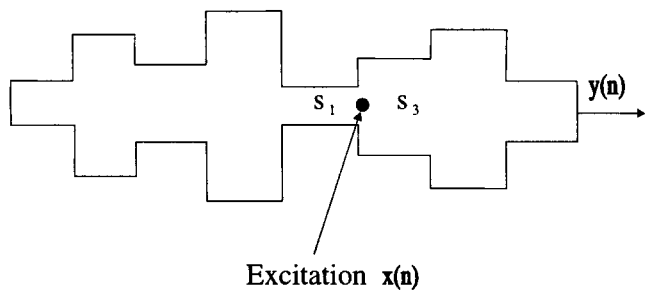

Fig. 1 Acoustic tube model for production of fricative.

We can see that the acoustic wave at the point of excitation propagates forward into the front cavity and backward into the back cavity. Together with the excitation source, there are three wave-directions at the location of the excitation source.Thus we use a three-port-adaptor to describe the acoustic characteristics at the excitation source, as shown in fig. 2. The relationship between the traveling waves in the remaining adjacent tubes can be represented by two-port-adaptors[5].

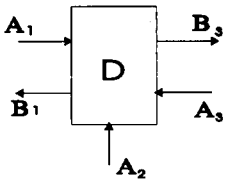

Fig. 2 Three-port adaptor at excitation source

The entire vocal tract can be realized using two- and three-port-adaptor in discrete-time domain as shown in fig. 3.

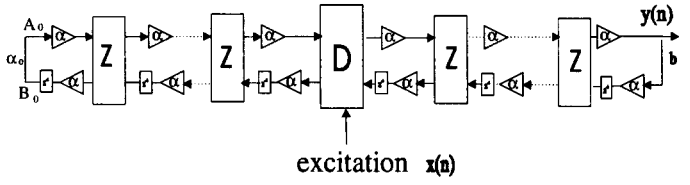

Fig. 3 Digital realization of the model in Fig. 1 in discrete-time domain

the constants α are lossy coefficients; Z and D express the two- and three-port-adaptor; the constant α_0 describes the opening of the glottis, its value lies in [-1,1].

Derivation of the Transfer Function

The three-port adaptor can be presented by a scattering matrix from the continuity equations for pressure and flow [6]. The linear relation for the flow is given by:

$$\begin{pmatrix} B_1 \\ B_2 \\ B_3 \end{pmatrix} = \mathbf{S} \begin{pmatrix} A_1 \\ A_2 \\ A_3 \end{pmatrix} \tag{1}$$

with

$$\mathbf{S} = \begin{pmatrix} \alpha_1 - 1 & \alpha_1 & \alpha_1 \\ \alpha_2 & \alpha_2 - 1 & \alpha_2 \\ \alpha_3 & \alpha_3 & \alpha_3 - 1 \end{pmatrix} \tag{2}$$

and

$$\alpha_i = \frac{S_i}{S_1 + S_2 + S_3} \tag{3}$$

S_1 and S_3 are the cross sectional areas of the front and the back cavity and S_2 describes the coupling of the excitation source to the vocal tract.

The acoustic wave at the end of the back cavity is partly reflected; there is the relation $A_0 = \alpha_0 B_0$ and the equation(1) can be transformed into the following equation:

$$\begin{pmatrix} B_2 \\ A_2 \end{pmatrix} = \mathbf{T_D} \begin{pmatrix} A_3 \\ B_3 \end{pmatrix} \tag{4}$$

$\mathbf{T_D}$ is a 2 x 2 matrix and is the function of the scattering transfer matrix of the back cavity, its elements are fractions.

The transfer function is finally derived and can be expressed as:

$$H(z) = \frac{1}{(0,1)\mathbf{T_D}\mathbf{T_F}\begin{pmatrix} -1 \\ 1 \end{pmatrix}} \tag{5}$$

$\mathbf{T_F}$ is the scattering transfer matrix of the front cavity.

If we have chosen the back cavity consisting of two sectional tubes, the first tube is a single section tube, the number of sections of the second tube is n; then the transfer function can be expressed:

$$H(z) = \frac{\alpha_0 z^{n+1} + \alpha_0\alpha_1^2 r_0 z^n + \alpha_2^2 r_0 z + \alpha_1^2\alpha_2^2}{(0,1)\mathbf{T_{DD}}\mathbf{T_F}\begin{pmatrix} -1 \\ 1 \end{pmatrix}} \tag{6}$$

r_0 is the reflection coefficient between the two tubes in the back cavity; α_1 and α_2 are the lossy coefficients of the back cavity. $\mathbf{T_{DD}}$ is a 2x2 matrix and its elements are polynomials.

Properties of the Transfer Function

We can see that the numerator of the transfer function is a polynomial containing the zeros in the transfer function. From this polynomial we can calculate the zero locations after estimating the model parameters.

It can be seen that the back cavity determines the zero locations of the transfer function, while the front cavity has no effect on it.

It can be seen that if the vocal tract is lossless($\alpha_1 = 1, \alpha_2 = 1$) and the glottis is closed or open($\alpha_0 = 1, \alpha_0 = -1$), the numerator polynomial of the transfer function is symmetric or antisymmetric,

$$z^{n+1} + r_0 z^n \pm r_0 z \pm = 0 \tag{7}$$

To break the symmetric or antisymmtric relation there are two ways, either a loss is to take into account in vocal tract or the factor α_0 of the glottis changes in (-1, 1).

If the area at the junction is $S_1 = 0$, the back cavity is eliminated, and the matrix $\mathbf{T_D}$ becomes a standard scattering transfer matrix of a two-port adaptor and the pole-zero model is transformed into the well-known all-pole model.

QUANTITATIVE CHARACTERIZATION OF FUNCTIONAL VOICE DISORDERS USING MOTION ANALYSIS OF HIGHSPEED VIDEO AND MODELING

Thomas Wittenberg *Patrick Mergell* *Monika Tigges* *Ulrich Eysholdt*

Department of Phoniatrics & Pedaudiology
ENT-Clinic at the University Erlangen-Nürnberg
Bohlenplatz 21, 91054 Erlangen, Germany
Wittenberg@phoni.med.uni.erlangen.de

ABSTRACT

A semiautomatic motion analysis software is used to extract elongation-time diagrams (trajectories) of vocal fold vibrations from digital highspeed video sequences. By combining digital image processing with biomechanical modeling we extract characteristic parameters such as phonation onset time and pitch. A modified two-mass model of the vocal folds is employed in order to fit the main features of simulated time series to those of the extracted trajectories. Due to the variation of the model parameters, general conclusions can be made about laryngeal dysfunctions such as functional dysphonia. We show the first results of semi-automatic motion analysis in combination with model simulations as a step towards a computer aided diagnosis of voice disorders.

	Hyper-functional	Hypo-functional
mean pitch	high	normal
phonation onset	pathological hard	normal, soft
sound	hoarse, creaky, pressed, soundless	soft soundless, breathy
amplitudes	restrained with intensity	widened with intensity
closure	relatively long	relatively short
irregularity	period	amplitudes

Table 1. Characteristics of hyper- and hypofunctional Dysphonias [8]

1. MOTIVATION

One important section of phoniatrics deals with the diagnosis and classification of functional disphonia. Morphological or organic voice disorders, such as cysts, polyps, carcinoma or granuloma can be detected and classified with the naked eye and the supplement of an endoscope or a laryngeal mirror. In contrast, functional dysphonia can only be diagnosed with the support of a highly refined imaging device, since no morphological change is observed. The dysfunction is hidden in the aperiodic, asymmetric and transitory oscillation patterns. For the recording of such a short-time scale motion without the violation of Shannons sampling theorem, a highspeed video camera has to be used [1, 3]. Fig.(1) shows exemplarily one period of vocal fold vibrations of a hyperfunctional dysphonia recorded with digital highspeed video camera. In our department of phoniatrics, a digital highspeed video camera has been applied in clinical routine examinations of vocal cord disorders in addition to the conventional videostroboscopy system for the past three years [2, 9].

One main goal of our research is to get a deeper insight into mechanisms hidden in functional dysphonia. Tab.(1) shows a comparison of the general features of of the two major categories of functional voice disorders, the hyper- and hypofunctional dysphonias [8]. In this context the prefixes *hyper* and *hypo* refer to the global tonus of all muscles involved in phonation, whereas *hyper/hypo* means, the muscle activity is too high, too low respectively. *Hypo*functional dysphonias are usually found in male patients whereas hyperfunctional dysphonias are specific for female patients. Even for the trained phoniatric it is very difficult to distinguish stridently between the hyper- and hypofunctions of the laryngeal muscles, since one extreme muscle tonus is usually compensated by the contrary tonus in an other muscle.

In the past, for the diagnosis of functional dysphonias, the phonation onset as one characteristic feature has been described qualitatively and subjectively with the terms *hard,normal* and *soft*. Typical for a soft phonation onset is an incomplete prephonatic closure after the adduction. Moreover, it is characterised by a relatively long onset time. In contrast, normal and hard, phonation onsets, are related to short onset times and a complete closure prior to the first oscillation maximum. The prephonatic closures are usually longer than in the case of soft onsets [7]. Fig.(2) shows three different kymograms of phonation onsets, which have been subjectively classified as hard, normal and soft, respectively.

It is our aim is to create a quantitative classification base to complement the subjective visual and auditive diagnosis of the phoniatric.

2. MOTION ANALYSIS

A semi-automatic motion analysis algorithm has been designed and implemented to extract the vocal cord trajectories from the highspeed image sequences [9].

This work has been funded by the 'Deutsche Forschungsgemeinschaft' (DFG)

Figure 1. Example a of digital highspeed video of a Male subject, 32 years, with hyperfunctional dyphonia. One oscillation period of a vocal cord vibration. The recording speed was 1922 frames/s, with a resolution of 128x64 pixels x 8 bit grayscale.

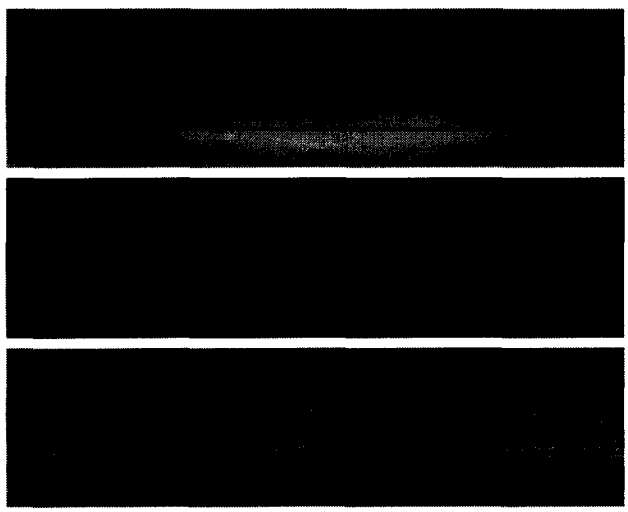

Figure 2. Different phonation onsets : hard phonation onset (top), normal phonation onset (center), soft phonation onset (bottom). From each image frame of the highspeed sequence (128x64 pixels x 8 bit grayscale, 1922 frames/s), one single image line (128 pixel) centered in anterior-posterior direction, and perpendicular to the principal glottal axis has been extracted. All image lines are arranged in chronolgical order, the time running from left to right. A time span of 0.25 seconds is shown.

In Fig.(1), one cycle of a vocal cord vibration is depicted. In the center of each single frame the two vocal folds can been seen. The dark area between the vocal folds is called glottis.
Using apriori knowledge about the anatomic design of the larynx and its physiological behaviour, the motion analysis problem can be broken down into two separate tasks:

- The detection and segmentation of the glottal area in each single frame of the sequence, and
- The calculation of selected trajectory motion points on the edge between the segmented area and the bordering vocal cords.

The first task is solved by applying a region growing algorithm to each single frame and thus separating the glottal area from the rest of the image. The seed for the region growing process can be calculated from the coordinates of the grayscale-minimum of each frame under the hypothesis, that the glottal area is always corresponding to the darkest region of the image and therefore contains the minimum grayscale value. To verify the spatial location of a potential seed, information about the location and expansion of the glottal area from the previous frames and a dynamic mean grayscale value are used. If the coordinates of the potential seed are outside the previous glottal area plus a tolerance region, or the value of the minimum grayscale is above the mean grayscale value of the past $n = 20$ frames, this seed will be rejected and a glottal closure is assumed.
The second task deals with the detection and calculation of significant tracking points on the edge of the vocal folds, which can be used to represent their motion during phonation and speech. In the past it has been useful to analyze three pairs of points on the vocal folds which correspond to the dorsal, central and ventral section of the larynx [6]. In our algorithm, these points can be detected from the glottal area calculated before. The first step consists of calculating the principle axis of the glottal area. This axis defines the temporal angular orientation and spatial expansion of the glottis. The principal axis is then divided by three orthogonal lines into four segments of equal length. Starting from the intersection points, these lines are traced until the border of the glottal area is reached. These endpoints mark the edge between the glottal area and the vocal folds as the temporal location of the individual tracking points. If these points are persued over all successive frames of an image sequence, an oscillating time series is generated. Analogous to the electro-glottogram (EGG) of the larynx, we call these trajectories **H**ighspeed-**G**lotto**G**rams (HGG).

Fig.(3) shows the resulting trajectories from the motion analysis of three different phonation onset modes of the kymograms in Fig.(2).

3. AUTOMATIC PARAMETER EXTRACTION: PHONATION ONSET TIME AND PITCH

In this section we present the automatic extraction of the phonation onset time τ from the HGG-data as one important parameter for the classification of functional dysphonias.
From the motion trajectories, as depicted in Fig.(3), the fundamental frequency f_0 can be obtained from its corresponding power spectrum. By using a bandpass filter ($f_0 \pm 10\%\ f_0$) the signal noise and motion analysis artifacts in the curve can be suppressed. From the filtered trajectory, the series of amplitude maxima is extracted using a simple peak-picking-algorithm and ignoring all peaks which are not spaced by the cycle duration. The resulting sequence of maxima describes the envelope curve of the phonation onset. The phonation onset time can be defined as the duration of amplitude growth from 32.2% to 67.8% of the saturation amplitude. These two thresholds have been chosen to be able to relate the experimentally obtained phonation onset time with a set of model parameters constituting a certain laryngeal configuration [4].