Inverse Filtering of Tube Models with Frequency Dependent Tube Terminations

K. Schnell, A. Lacroix
Institute of Applied Physics
Goethe-University, Frankfurt am Main
Lacroix@iap.uni-frankfurt.de

Abstract
The tube model, realized by lattice filters in discrete time, can be used to describe the propagation of plane sound waves through the vocal tract. The tube model which is treated in this contribution contains two prescribed terminations. One for the lip opening and one for the constriction at the glottis. These two terminations are frequency dependent and the lip termination must be adapted to the specific speech sound. To estimate the parameters of this tube model, standard algorithms like the Burg-method and related methods are not applicable. Therefore a procedure is proposed to estimate the parameters of these tube models in an adequate way. The procedure is based on inverse filtering, which is carried out iteratively. The analysis of test signals, which are generated from prescribed systems, shows almost perfect results by the proposed algorithm. The analysis of consonants shows, that the corresponding constrictions in the vocal tract area functions can be observed. Using these estimated constrictions it is possible to synthesize VCV transitions too, which implies the typical formant movements.

1. Introduction
The linear prediction and the Burg-Method [1] delivers a fast algorithm for the parameter estimation of a tube model with specific tube terminations, which are only a simplification of the actual conditions for the vocal tract. For a tube model with two frequency dependent tube terminations the conventional algorithms like Burg are not applicable. One reason is that the standard inverse lattice filter has the usual forward and backward prediction error in the upper and lower path; for arbitrary terminations of the tube this is not valid. In [2] a procedure is proposed to estimate the parameters of a tube model with a time dependent real glottis impedance. The procedure uses the method of steepest descent to minimize a spectral distance and needs long time for computation in contrast to the algorithm described here.

2. Tube model
The tube model, realized by lattice filters in discrete-time describes the propagation of plane sound waves. The model can be described by a concatenation of scattering transfer matrices \( T_i \), which consist of tube elements and two-port adaptors

\[
T_i = k(r_i) \begin{bmatrix} 1 & r_i z^{-1} \\ r_i z^{-1} & 1 \end{bmatrix}, \quad \begin{bmatrix} X^u \\ X^l \end{bmatrix} = T_i \begin{bmatrix} X^u_{i+1} \\ X^l_{i+1} \end{bmatrix}. \tag{1}
\]

\( x^u \) describes the forward propagation in the upper path of the tube model and \( x^l \) the backward propagation in the lower path. The two-port adaptor describes the scattering of the sound waves at the tube junctions. \( k(r) \) is a factor, which determine the kind of the wave quantities and is set equal to one for the inverse filter. The tube elements are realized by a pure delay for lossless wave propagation. The tube termination at the lip opening can be modeled by a pole zero system \( L(z) \) from Laine [3]. \( L(z) \) depends on the area of the lip opening, which depends on the specific speech sound. This tube termination is frequency dependent and an additional positive factor \( \alpha < 1 \) is involved, which can model loss effects of the vocal tract. The resulting reflection coefficients from the analysis of voiced sounds by the Burg-Method show no constriction of the vocal tract areas near the glottis. A fixed termination \( G(z) \) at the opposite tube end is assumed, to model the impedance of the vocal tract at the glottis. \( G(z) \) includes a zero, so that the termination is frequency dependent too.

3. Inverse filtering
The inverse filter is depicted in figure 1. On the left side the termination for the lip opening is located and on the right side the termination at the glottis. \( x^u_0 \) is the input of the inverse

![Figure 1: Inverse tube model.](image-url)
filter and represents the speech signal which is to be analyzed. For the parameter estimation the problem is to determine the reflection coefficients in a way, that the power of the output signal \( x_u^* \) is a minimum. In contrast to the Burg-method and related methods, too, the output power cannot be minimized after each section, but for the entire inverse filter. Therefore each coefficient is estimated by minimizing the power of the output \( x_u^* \), whereas the other coefficients are fixed. Only the sections behind the estimated section \( T_i \) affects the output power. So the two input signals \( x_{\text{eq},i}^* \) and \( x_{\text{eq},-i}^* \) of \( T_i \) should be filtered with respect to the minimal power of the output \( x_u^* \). \( r_i \) is the variable for this minimization. The criterion for the minimum is:

\[
E\left[ x_u^2 \right] \rightarrow \text{min.} \quad \Rightarrow \quad \frac{\partial E\left[ x_u^2 \right]}{\partial r_i} = 0
\]  

which results in the estimated reflection coefficient

\[
\hat{r}_i = - \frac{\left( u_{\text{eq},i}^* \cdot u_{\text{eq},-i}^* + l_{\text{eq},i}^* \cdot l_{\text{eq},-i}^* \right)}{\left( u_{\text{eq},i}^* \cdot u_{\text{eq},-i}^* - 2 \cdot u_{\text{eq},i}^* \cdot l_{\text{eq},-i}^* \cdot l_{\text{eq},i}^* \right)}. \tag{3}
\]

The expected values \( E[x] \) are replaced by time averages \( \bar{x}_i \), \( u_{\text{eq},i}^\beta \) and \( l_{\text{eq},i}^\beta \) are the filtered signals \( x_{\text{eq},i}^* \) and \( x_{\text{eq},-i}^* \) behind \( T_i \):

\[
u_{\text{eq},i}^\beta(n) = f_{\text{eq},i}^\beta(n) \cdot x_{\text{eq},i}^*(n),
\]

\[
l_{\text{eq},i}^\beta(n) = f_{\text{eq},i}^\beta(n) \cdot x_{\text{eq},-i}^*(n - 1).
\]

The finite impulse responses \( f_{\text{eq},i}^\beta(n) \) can be derived from the matrix \( F_i \) which describes the tube sections between the output of the inverse filter and the estimated section

\[
F_i = \begin{bmatrix} F_{i1} & F_{i2} \\ F_{i1}^{-1} & F_{i2}^{-1} \end{bmatrix} = \begin{bmatrix} 1 & \sum_{k=1}^{N_n} r_{sk} \cdot z^{-1} \\ 0 & 0 \end{bmatrix} \prod_{k=1}^{N_n} \left( r_{sk} \cdot z^{-1} \right)
\]

for \( i = 1 \ldots N - 2 \) and

\[
F_i = \begin{bmatrix} F_{i1} & F_{i2} \\ F_{i1}^{-1} & F_{i2}^{-1} \end{bmatrix} = \begin{bmatrix} 1 & g(z) \cdot z^{-1} \\ 0 & 0 \end{bmatrix} \quad \text{for} \quad i = N - 1.
\]

\( \hat{r}_i \) is the optimal coefficient on condition that the other coefficients are determined. The algorithm starts with all parameters \( r_i \) equal to zero which requires no knowledge about the solution. Then the parameters are estimated one after the other by eq. (3) from \( r_i \) till \( r_{N - 3} \) which represents one iteration of the algorithm. Since the solution of one parameter \( r_i \) depends on the other parameters, the next iteration improves the estimate because the \( \hat{r}_i \) of the last iteration are used for the calculation of \( F_i \) in the next iteration. In rare cases it happens that the magnitude of \( \hat{r}_i \) is greater than one, which implies an unstable system. This can be prevented simply by setting the magnitude of \( \hat{r}_i \) to 0.99 in this case. The tube terminations as well as the length of the tube have to be set before the algorithm starts.

3.1. Filtering of periodic signals

The algorithm is specialized for the analysis of periodic signals. Therefore a single period is analyzed with the assumption that the signal values outside of the period are determined by periodic continuation. The values beyond the periods are necessary for the initial values of the delay elements of the inverse filter. With this assumption a time delay transforms a period, represented as a vector \( x \) of length \( N \)

\[
\begin{bmatrix} x_N & x_{N-1} & \ldots & x_1 \end{bmatrix}^T
\]

into

\[
\begin{bmatrix} x_N & x_{N-1} & \ldots & x_1 \end{bmatrix}^T
\]

so that no signal values get lost by processing. By this definition of the time delay the estimation algorithm is independent of the phase of the analyzed signal to a great extent. This is particularly important, if several periods are averaged in the spectral domain to produce a single period.

3.2. Initialization of the inverse filter

Before the first parameter \( r_i \) can be calculated, the signals \( x^*_i \) and \( x^*_0 \) at the beginning of the inverse filter have to be prepared. \( x^*_0 \) is the signal, which is to be analyzed, whereas \( x^*_i \) is calculated by filtering of \( x^*_0 \) according to

\[
x^*_i(z) = \alpha L(z) \cdot x^*_0(z).
\]

This is realized by a repeated filtering of the signal period \( x^*_0 \) about ten times, in order to reach the steady state for the signal \( x^*_i \). In this way the recursive part in the inverse filtering process is performed.

4. Analysis of test signals

Signals from prescribed systems are analyzed to test whether the algorithm is able to reach the optimal solution. To produce a test signal a tube model with frequency dependent terminations is excited by an impulse train. Then one period of the output signal represents the test signal. For the analysis of the test signal the same tube terminations and the same tube length are used, as for the generation, to enable a perfect match. The estimated magnitude responses are shown in figure 2 with the power spectrum of the analyzed test signal for comparison.

![Figure 2: solid line top: power spectrum of the test signal (as polygon); dashed line center: Estimated magnitude response of the tube model after the first iteration; dashed line bottom: Estimated magnitude response of the tube model after 50 iterations.](image-url)
perfect match to the spectrum of the analyzed test signal. Analyses of other test signals show similar results.

5. Analysis of speech signals

5.1. Analysis of speech periods
Especially consonants demonstrate fluctuations from one speech period to another period of voiced speech. Therefore the results of the analysis of adjacent speech periods can differ. To avoid this problem, several periods are averaged in the spectral domain to produce a single period. At first the lengths of the adjacent speech periods are standardized to the mean of the lengths. This is done by repeating or omitting sample values within the second half of the period. Then the magnitude spectrum is calculated from all standardized periods. The average of these spectra is calculated and the signal from the IDFT delivers the single period \( p \).

5.2. Prefiltering
To separate the influence of the excitation and radiation of voiced speech from the vocal tract, the speech signal should be filtered before the actual algorithm starts. The excitation together with the radiation can be modeled by real poles \( B(z) \). These real poles are estimated by a repeated Burg-method of first order. The signal \( p \) is filtered by the real zeros \( B(z) \) and is now used for the input signal \( x_0 \) of the inverse filter. This adaptive prefiltering yields adequate results. In the case of consonants, maybe due to an additional noisy component and/or nonstationary effects, the real poles are corrected sometimes in a way, so that the proportions of the estimated areas are better represented.

5.3. Results of analyzed speech
The sampling rate of all presented analyzed speech signals is 22 kHz, which determines the length of each tube element. Using the termination \( G(z) \) the length of the tube model, which represents the vocal tract, is determined by the number of tube sections. Therefore the number of tube sections is chosen, which is close to the assumed length of the vocal tract with an output power of the inverse filter which is minimum or at least close to it. Furthermore \( L(z) \) is adapted to the lip opening of the corresponding sounds. \( G(z) = -0.37 \cdot (z + 0.135) \) is chosen for all presented results. Figure 3 shows the estimated magnitude response compared to the spectrum of the analyzed vowel /a:/ which is represented by the period \( p \).

In figures 4-8 estimated vocal tract area functions are shown for consonants. On the left side of these figures the lip opening is located and on the right side the glottis. The area functions are presented in a logarithmic scale, since by logarithmic values the proportions of the areas are better visible. Figures 4 and 5 show the results by the analysis of the voiced fricatives /z/ and /Z/ which correspond to the areas of the unvoiced fricatives /s/ and /S/. The notation of the speech sounds is SAMPA.

![Figure 4](image)

**Figure 4:** Estimated vocal tract area function of /z/ in logarithmic scale; mouth to the left and glottis to the right.

![Figure 5](image)

**Figure 5:** Estimated vocal tract area function of /Z/ in logarithmic scale.

Figures 6-8 show the results by the analysis of the voiced plosives. The following sound was the neutral vowel (schwa).

![Figure 6](image)

**Figure 6:** Estimated vocal tract area function of /b/ in logarithmic scale.

![Figure 7](image)

**Figure 7:** Estimated vocal tract area function of /d/ in logarithmic scale.

![Figure 8](image)

**Figure 8:** Estimated vocal tract area function of /l/ in logarithmic scale.

![Figure 3](image)

**Figure 3:** line spectrum; power spectrum of the vowel /a/; solid line: estimated magnitude response after 50 iterations.
6. VCV-Transition

The estimated constrictions of the vocal tract can be used to model the voiced part of a vowel-consonant-vowel-transition. The transition can be created by superimposing the constriction areas on a vowel-vowel-transition in a way, so that the constriction is dominant in the center of the transition. This procedure is described in [4], in which the constrictions are estimated by a different procedure for another tube model [2]. In contrast to [4], here the logarithmic area function are used to mix the vowel areas with the areas of the constriction, which avoid the knowledge about the absolute area values. In figure 9 the transition of the area functions is depicted for the utterance [a:Sa:] in SAMPA. The constriction of the consonant /S/ (fig. 5) is in the center of the picture.

Figure 9: Area functions of the synthesized voiced transition of [a:Sa:].

The excitation of the tube model is an impulse train, which is filtered by the system $1/B(\tau)$ with real poles. Figure 10 shows the spectrogram from the synthesized speech [a:Sa:] by the area transition of the voiced part of figure 9. The frequency range up to 4 kHz is shown where the movement of the formants can be well observed, which correspond to the locus theory. In comparison to the synthesized speech an utterance of natural speech [a:Sa:] is depicted in figure 11, with similar movements of the formants.

Figure 10: Spectrogram of synthesized [a:Sa:].

Figure 11: Spectrogram of natural utterance [a:Sa:].

7. Conclusions

The proposed procedure is able to estimate the reflection coefficients of a tube model with prescribed frequency dependent terminations. The two terminations represent the influence of the lip opening and of the glottis. Analyses of test signals, which are generated by a prescribed system, yield almost perfect results. Especially for the analysis of consonants it is favorable to calculate the average of adjacent periods in the spectral domain to obtain a single period. The estimated vocal tract area functions show constrictions at the corresponding positions of the analyzed consonants. Using these estimated constrictions, it is possible to synthesize VCV transitions, in which typical movements of the formants can be observed.

8. References