Parameter Estimation of Vocal Tract Models in SPEAK

Karl Schnell, Frank Ranostaj, Arild Lacroix

Institut für Angewandte Physik, J.W.Goethe-Universität Frankfurt am Main, D-60325 Germany
Robert-Mayer-Str. 2-4, Email: {Schnell, Ranostaj, Lacroix}@iap.uni-frankfurt.de

Introduction

In the field of speech processing interactive and multimedia presentations support a better understanding of models and algorithms. SPEAK is a multimedia system for the demonstration of speech acoustics and investigations of vocal-tract models. The program has a modular structure which allows to construct these complex models from simple elements. Every element is interactively adjustable and the results are displayed immediately and presented audibly. For example in a vocal-tract model, consisting of a signal source and a acoustic tubes, the alteration of the cross-sectional areas affects the synthesized output signal simultaneously.

The inverse process is also possible in SPEAK [1] where vocal tract configurations are estimated from speech signals. In this contribution the capability is extended by implementation of a proper estimation algorithm.

SPEAK is implemented in Java, which supports the graphical and audio requirements of the application. Furthermore, due to the underlying concepts of Java, it is possible to start this application directly from our homepage [2].

Structure of SPEAK

For the investigation of the above mentioned elements alone or in combination, a modular structure for the software is chosen. With the aid of SPEAK it is possible to combine several elements acoustically correct and allows the construction of different detailed models of the vocal-tract. SPEAK is focused on source-filter-models and consequently these elements appear in two categories: signal-sources and filters. Several signal sources are available including one with a signal shape similar to the glottal pulse. Other useful excitations are white noise, an impulse train, or arbitrary recorded signals, for example residual signals from speech. For the immediate analysis of speech utterances a microphone input as source is also provided.

The filter elements are general pole-zero-systems, FIR-systems, and time discrete models of acoustic tubes. Each filter is displayed by a graphical user interface, and can be adjusted therein. It is possible to display several of it's filter characteristics, namely the impulse response, pole-zero plot, phase and magnitude of the frequency response. Furthermore it is possible to present the magnitude of the system function above the z-plane in a rotating 3-D graphic. All views are updated immediately when the filter parameters are modified, fig. 1a.

Two different kinds of links for these elements are provided, one which acts as an abstract signal link, while the other connects acoustic tubes providing an acoustical correct adaptation. With this link it is also possible to model

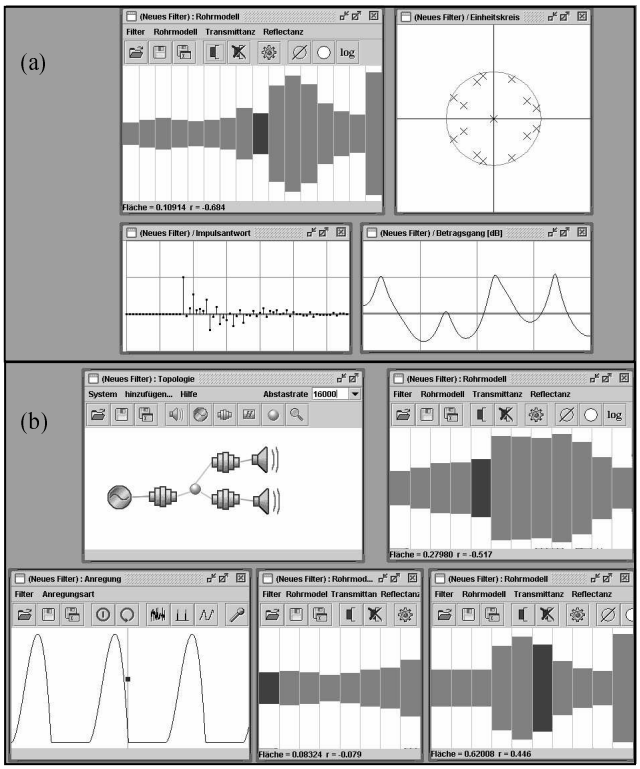

Figure 1: Screenshots of SPEAK: a) Tube model with views of system properties; b) branched tube configuration.

branching of tubes, like it is necessary to describe the junction of nasal and vocal tract, fig. 1b.

Speech Analysis by Tube Models

The tube model is realized by lattice filters in discrete time and can be described by a concatenation of scattering transfer matrices $\boldsymbol{T}_i$; each matrix consists of the i'th tube element and a two-port adaptor for an area discontinuity

$$\boldsymbol{T}_i = \begin{pmatrix} 1 & r_i \cdot \alpha \cdot z^{-1} \\ r_i & \alpha \cdot z^{-1} \end{pmatrix}, \qquad \begin{pmatrix} X_i^u \\ X_i^l \end{pmatrix} = \boldsymbol{T}_i \begin{pmatrix} X_{i-1}^u \\ X_{i-1}^l \end{pmatrix}. \qquad (1)$$

X_i^u describes the forward propagation and X_i^l the backward propagation of sound waves. r_i is the reflection coefficient and α is the attenuation factor modeling the distributed losses within the vocal tract. Although the losses are in fact frequency dependent the distributed losses are assumed to be frequency independent because of the computational costs of the estimation process. The tube termination at the lip opening is modeled by a real factor λ or by the pole-zero system $L(z)$ from Laine [3] with an additional positive factor $\beta < 1$. The termination λ is frequency independent in contrast to the termination $L(z)$ which additionally depends on the area of the lip opening.

Inverse filtering

For the analysis of speech, estimation algorithms are

implemented in SPEAK which estimate the areas of the vocal tract models from the speech signal. Since SPEAK is able to process the speech signals in real time, estimation algorithms of low computational costs are required. Prior to the analysis, the speech signals are filtered by a repeated adaptive preemphasis to eliminate the influence of the excitation and radiation. Then the parameters of the tube model are estimated by inverse filtering. One implemented estimation algorithm is the Burg method [4] which yields only optimal results, if the tube termination and the sound propagation within the vocal tract are assumed to be lossless. Therefore an iterative inverse filtering procedure [5] is implemented in SPEAK, too, suitable for lossy tube models which deviate from the standard LPC-model. In contrast to the Burg method the power of the output of the entire inverse filter is minimized. The minimization of the expectation of $(x_N^u)^2$ for each tube section

$$\mathrm{E}\left[x_N^{u\,2}\right] \rightarrow \min. \quad \Rightarrow \quad \frac{\partial \mathrm{E}\left[x_N^{u\,2}\right]}{\partial r_i} = 0 \qquad (2)$$

yields the *i*'th estimated reflection coefficient

$$\hat{r}_i = -\frac{\overline{u_i^{11} \cdot u_i^{12}} + \overline{u_i^{11} \cdot l_i^{11}} + \overline{l_i^{12} \cdot u_i^{12}} + \overline{l_i^{12} \cdot l_i^{11}}}{\overline{u_i^{12} \cdot u_i^{12}} + 2 \cdot \overline{u_i^{12} \cdot l_i^{11}} + \overline{l_i^{11} \cdot l_i^{11}}} \qquad (3)$$

with time averages $\bar{x}$. $u_i^{\lambda\beta}(n)$ and $l_i^{\lambda\beta}(n)$ are the filtered signals x_{i-1}^u and x_{i-1}^l behind $\boldsymbol{T}_i$:

$$u_i^{\lambda\beta}(n) = f_i^{\lambda\beta}(n) * x_{i-1}^u(n),$$

$$l_i^{\lambda\beta}(n) = \alpha \cdot f_i^{\lambda\beta}(n) * x_{i-1}^l(n-1). \qquad (4)$$

$f_i^{\lambda\beta}(n)$ are finite impulse responses and can be derived from the matrix $\boldsymbol{F}_i$ describing the tube sections between the estimated section $\boldsymbol{T}_i$ and the output of the inverse filter x_N^u

$$\text{with } \boldsymbol{F}_i = \begin{pmatrix} F_i^{11} & F_i^{12} \\ F_i^{21} & F_i^{22} \end{pmatrix} = \prod_{k=0}^{N-i-1} \boldsymbol{T}_{N-k} \quad \text{for } i = 1 \ldots N-1 \qquad (5)$$

and $\boldsymbol{F}_N = \boldsymbol{I}$ for $i = N$. Before the first parameter r_1 can be calculated, the signals x_0^u and x_0^l at the beginning of the inverse filter are initialized with $x_0^u = x$ and $x_0^l = \beta \cdot h_L(i) * x_0^u$ or $x_0^l = \lambda \cdot x_0^u$ in the case of a real termination; $h_L(i)$ is the impulse response of the lip termination $L(z)$ and x is the speech signal filtered by the adaptive preemphasis. $\hat{r}_i$ is the optimal coefficient on condition that the other coefficients are determined. Therefore the estimation of the reflection coefficients by equation (3) is carried out iteratively. One iteration calculate $\hat{r}_i$ for $i = 1 \ldots N$ and the resulting coefficients are bounded to $|r_i| < 0.99$. This iterative estimation procedure is a modification of the algorithm described in [5]. Additionally to the existing algorithm distributed losses within the vocal tract, represented by the attenuation factor α, are introduced as can be seen from eq. (1). Furthermore equations (4) and (5) are affected since no explicit tube termination for the glottis is prescribed. The algorithm is implemented in Java by an efficient routine due to the requirements of real time processing which is achieved by avoiding memory allocations.

Speech Analysis

Figure 2 shows the analysis of the vowel /i:/ (sampling rate 22 kHz) by different tube models and estimation algorithms. The influence of the losses can be investigated. It can be seen that the introduction of losses increases the ratios of the estimated cross sectional areas in certain regions. This is consistent with the fact that in the lossless case the reflection coefficients have to model the decrease of the sound wave amplitudes caused by the losses.

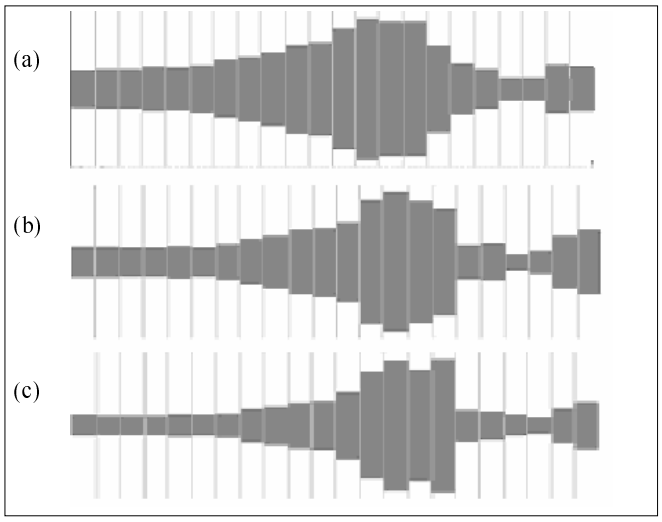

Figure 2: Estimated cross sections of the vocal-tract for the vowel /i:/. The lips are located at the right side. (a) Estimated by the Burg method with a lossless tube model; (b) iterative inverse filtering with the Laine termination (lip opening = 4,5 cm²) and no distributed losses; (c) same as (b) but with distributed losses.

Conclusion

The capability of SPEAK is extended by a real time analysis module which is able to include physical effects of sound propagation in tubes. The effect of losses either caused by terminations or by distributed losses during wave propagation is clearly demonstrated. It can be observed that in some cases more realistic cross-sectional areas are achieved. For that purpose an iterative inverse filtering is provided which is suitable for lossy tube models.

References

[1] Ranostaj F., Lacroix A.: Ein Experimentalsystem zur Sprechakustik und Sprachproduktion, Proc. ESSV-2003, Karlsruhe Germany, pp. 280-285, 2003.

[2] www.rz.uni-frankfurt.de/~lacroix/index.html

[3] Laine U. K.: Modeling of lip radiation impedance in the z-domain, Proc. ICASSP'82, Paris, pp. 1992-1995, 1982.

[4] Burg J.: A new Analysis Technique for Time Series Data, NATO Advanced Study Inst. on Signal Processing, Enschede, 1968.

[5] Schnell K., Lacroix A.: Inverse Filtering of Tube Models with Frequency Dependent Tube Terminations, Proc. EUROSPEECH-2001, Aalborg, pp. 2467-2470, 2001.