Source/Filter–Model

– Acoustic Tube Models
– Linear Prediction
– Formant Synthesizer

Markus Flohberger
maxiko@sbox.tugraz.at

Graz, 19.11.2003

1 Introduction

Speech synthesis methods that simulate speech production mechanisms are mainly based on the source/filter–theory. The model consists of three parts, the source, the filter and lip radiation, as seen in figure 1. The vocal tract can be modelled as an acoustic tube with varying cross–sectional area. A linear filter simulates the vocal tract that varies over time, which in turn is driven by an adequate source. To this category belong two main techniques: The Formant Synthesizer and Diphone Concatenation.

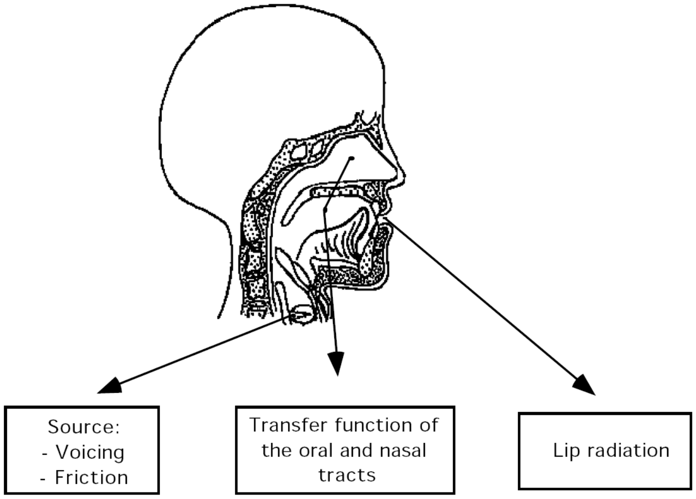


Figure 1: Source/Filter–Model

2 Acoustic Tube Models

The wave equation is a linear partial differential equation whose solution describes pressure and velocity of air particles in a sound wave as a function of time and space, and provides a means to approximately describe the acoustics of speech production. Sound in the vocal tract can be described by a simple *lossless uniform tube model.* This model can be extended to a *nonuniform tube model with energy loss*, including vocal tract wall vibration, viscosity, thermal conduction of air particles and radiation loss of the lips. The losses are described by partial differential equations which are coupled to the wave equation. These equation set can be solved by numerical simulation. To get a closed–form transfer function of air particle velocity from the glottis to the lips an alternative approach can be followed. This model approximates the vocal tract by a *concatenation of acoustic tubes*, each of small length and uniform cross–section. This model assumes that the vocal tract is linear and

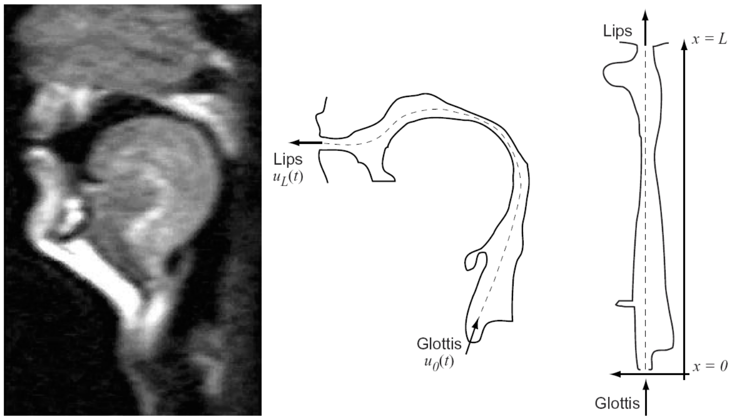


Figure 2: Vocal tract to acoustic tube model

time–invariant. More *accurate models* consider a *time–varying vocal tract* and *nonlinear coupling between vocal folds and vocal tract.*

2.1 Uniform Tube Model: Lossless Case

- Constant (time– and space–invariant) cross–section $A(x,t) = A$
- Moving piston provides an ideal particle velocity source
- The open end represents opened lips

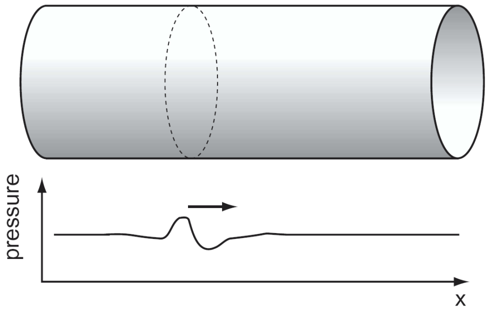


Figure 3: Lossless uniform tube model

Volume Velocity $u(x,t)$**:** Rate of flow of air particles perpendicularly through a specified area. $u(x,t)$ is used instead of the air particle velocity $v(x,t)$ and is related by $u(x,t) = A \cdot v(x,t)$.

Wave Equation:

$$-\frac{\partial p}{\partial x} = \frac{\rho}{A}\frac{\partial u}{\partial t} \qquad -\frac{\partial u}{\partial x} = \frac{A}{\rho \cdot c^2}\frac{\partial p}{\partial t}$$

$\frac{\rho}{A}$ is the *acoustic inductance* and $\frac{A}{\rho \cdot c^2}$ the *acoustic capacitance*. These quantities describe the inertia (mass or density) and the springiness (elasticity) of the air medium.

General Solutions:

$$u(x,t) = u^+\left(t - \frac{x}{c}\right) - u^-\left(t + \frac{x}{c}\right)$$

$$p(x,t) = \frac{\rho \cdot c}{A}\left[u^+\left(t - \frac{x}{c}\right) - u^-\left(t + \frac{x}{c}\right)\right]$$

Velocity and pressure is comprised by a *forward– and backward–travelling wave*. The steady–state solutions can be represented in the frequency domain, because ultimately we are interested in the frequency response of the tube.

Pressure *(p)*	↔	Voltage *(v)*
Volume Velocity u	↔	Current *(i)*
Acoustic Inductance $\frac{\rho}{A}$	↔	Electric Inductance L
Acoustic Capacitance $\frac{A}{\rho \cdot c^2}$	↔	Electric Capacitance C

Table 1: Analogies between Acoustic Tube and Electrical Transmission Line

We assume a sinusoidal volume velocity source on the one end and on the other end the pressure equals zero, which represents opened lips. With these boundary conditions we get the particular solution, which is referred to as *standing wave*. The forward– and backward–travelling waves have added and made the wave shape stationary in time. Pressure and velocity are 90° out of phase.

Acoustic Impedance:

$$Z_A(\Omega) = \frac{p(x,t)}{u(x,t)}$$

Transfer Function:

$$V_a(s) = \frac{1}{\prod_{k=1}^{\infty}(s - s_k)(s - s_k^*)}$$

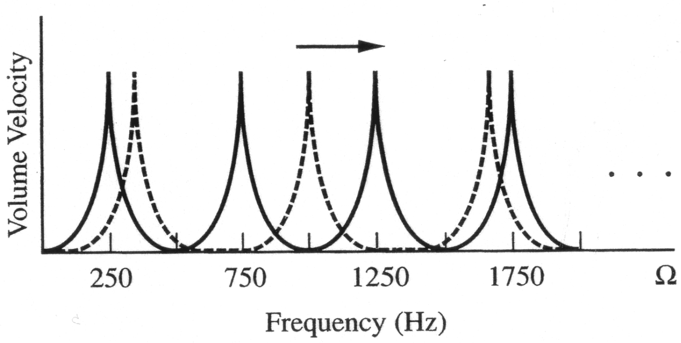


Figure 4: Dependence of the formants from the tube–length

2.2 Nonuniform Tube Model: Considering Losses

Energy loss can be described by partial differential equations coupled to the wave equation. The closed–form solution of these equations is quite difficult to obtain and so numerical simulation is preferred, which requires discretization in time and space.

2.2.1 Wall Vibration

The cross–section of the vocal tract is nonuniform and time varying. The walls of the vocal tract are pliant and so can move under the pressure induced by the sound propagation.
It is assumed that small, differential pieces are locally reacting. Each of these pieces can be modelled by a mass m_w, spring constant k_w and a damping constant b_w per unit surface area and is called *lumped parameter model*, as can be seen in figure 5. The change of the cross–section due to pressure changes is small compared to the average cross–section. As a result of this loss the bandwidth gets nonzero and the resonant frequencies increase.

2.2.2 Viscosity and Thermal Loss

The wave equation has to be extended by a

- *Resistive Term:* Represents the energy loss due to friction of the air particles along the wall.
- *Conductive Term:* Represents the heat loss through the vibrating walls.

Due to these losses the resonant frequencies decrease and the resonant bandwidth broadens.

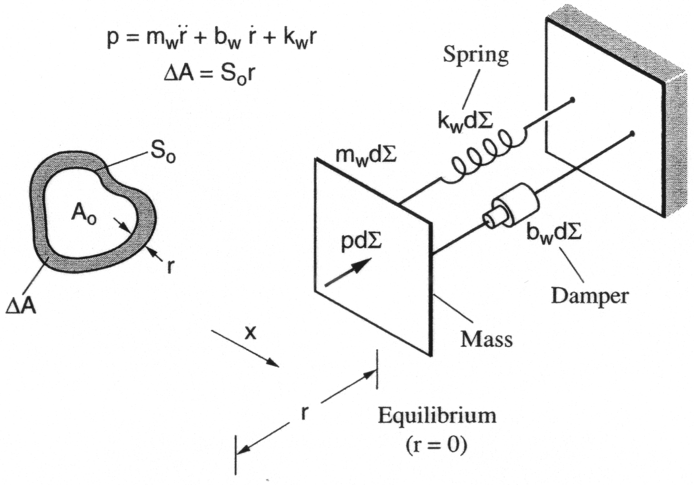


Figure 5: Lumped parameter model of vibrating wall

2.2.3 Boundary Effects

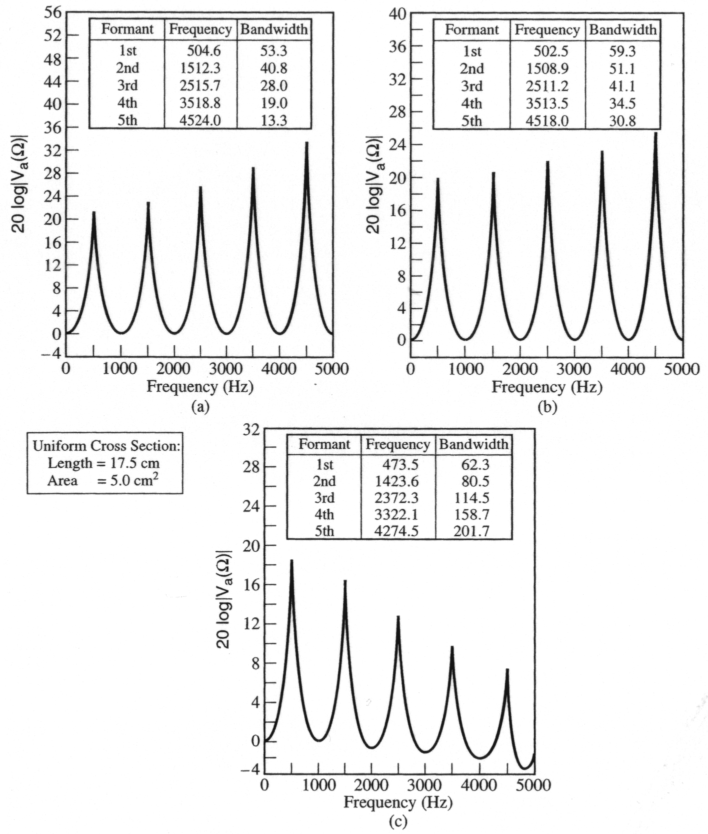
Until now it was assumed, that the pressure at the lips is zero, the volume velocity source is ideal and there is no energy loss at the output or input of the uniform tube.

- The description of sound radiation at the lips is quite complicated, but it can be simplified by an acoustic impedance. The consequences of radiation are broader resonant bandwidths and a decrease of the resonant frequencies.
- There can be found an acoustic impedance for the glottis by simplifying a nonlinear, time varying two–mass vocal fold model. This impedance broadens the resonant bandwidth.

2.2.4 Complete Model

A complete model involves *boundary condition effects*, energy loss due to *vibrating walls*, *viscosity* and *thermal conduction*. The effects can be summarised and are depicted in figure 6:

1. There is a shift of the resonant frequencies due to the various sources of energy loss.
2. Bandwidths of the lower resonances are controlled by vibrating walls and glottal impedance loss.
3. Bandwidths of the higher resonances are controlled by radiation, viscous and thermal loss.



(a)

Formant	Frequency	Bandwidth
1st	504.6	53.3
2nd	1512.3	40.8
3rd	2515.7	28.0
4th	3518.8	19.0
5th	4524.0	13.3

(b)

Formant	Frequency	Bandwidth
1st	502.5	59.3
2nd	1508.9	51.1
3rd	2511.2	41.1
4th	3513.5	34.5
5th	4518.0	30.8

(c)

Formant	Frequency	Bandwidth
1st	473.5	62.3
2nd	1423.6	80.5
3rd	2372.3	114.5
4th	3322.1	158.7
5th	4274.5	201.7

Figure 6: (a) vibrating walls, (b) vibrating walls, viscosity and thermal loss (c) vibrating walls, viscosity, thermal loss and radiation

The complete model leads to cumbersome time and spacial numerical simulation. With additional approximations a transfer function through concatenated tubes can be found. This approach leads to a discrete–time model which can be computed efficiently.

2.3 Discrete–Time Model Based on Tube Concatenation

Sound Propagation in the Concatenated Tube Model

- Concatenation of short lossless uniform tubes, as seen in figure 7.
- Energy loss appears only at the two boundaries (glottis, lips). Typical it is assumed, that the glottal impedance is infinite and so the only loss ocurrs at the lips.
- This model is linear and easy to work with.

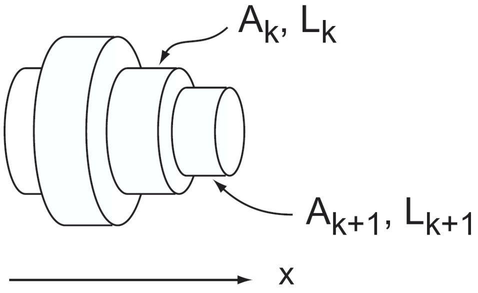


Figure 7: Concatenation of short tubes

The concatenated tube model provides a convenient transition from a continuous–time to a discrete–time all–pole model, and the resulting all–pole model leads naturally into the *linear prediction speech analysis*. Each tube has a cross–section area A_k, a length l_k and a time of propagation down the length of the tube τ_k. We can assume planar wave propagation (no loss), so that the pressure and velocity relations for each tube satisfy the wave equation. The solution for pressure and velocity for the k–th tube can be written as $(0 \leq x \leq l_k)$:

$$u_k(x,t) = u_k^+\left(t - \frac{x}{c}\right) - u_k^-\left(t + \frac{x}{c}\right)$$
$$p_k(x,t) = \frac{\rho \cdot c}{A_k}\left[u_k^+\left(t - \frac{x}{c}\right) - u_k^-\left(t + \frac{x}{c}\right)\right]$$

To give a specific solution requires boundary conditions between two adjacent tubes. Pressure and volume velocity have to be *continuous in time and space* everywhere in the system.

$$u_k(l_k,t) = u_{k+1}(0,t)$$

$$p_k(l_k,t) = p_{k+1}(0,t)$$

At discontinuities in the area function $A(x,t)$ there occur *propagation and reflection* of the travelling wave. This can be described by the reflection coefficient r_k:

$$r_k = \frac{A_{k+1} - A_k}{A_{k+1} + A_k}$$

$$u_{k+1}^+(t) = (1 + r_k) \cdot u_k^+(t - \tau_k) + r_k \cdot u_{k+1}^-(t)$$
$$u_k^-(t + \tau_k) = -r_k \cdot u_k^+(t - \tau_k) + (1 - r_k) \cdot u_{k+1}^-(t)$$

Boundary conditions at the glottis and the lips have also to be considered:

- The forward–travelling wave is reflected back at the lips boundary with the reflection coefficient for the lips r_L. There is no backward–travelling wave from free space.

- The boundary condition at the glottis can also be modelled by a reflection coefficient r_g.

The concatenated tube model is not necessarily consistent with the underlying physics of sound propagation. However it is possible to control the formant bandwidths with these two boundaries (glottis, lips).

Discrete Time Realization We assume that all tubes are of equal length $\Delta x = \frac{l}{N}$ and the time to propagate through one tube is $\tau = \frac{\Delta x}{c}$. This is equivalent to a discretization of the continuous–space tube by Δx in space, corresponding to τ in time. This leads to a periodicity of the transfer function representation.

$$V(z) = \frac{A \cdot z^{-\frac{N}{2}}}{1 - \sum_{k=1}^{N} a_k \cdot z^{-k}}$$

- The poles of the transfer function correspond to the formants of the vocal tract.
- The resulting system is stable. (all poles inside unit cycle)
- The all–pole transfer function is only a function of the reflection coefficients, and these are only a function of the cross–section areas. Therefore, if we could estimate the area functions, we could then obtain the all-pole discrete–time transfer function.

Complete Discrete Time Model The discrete–time speech production model can be described as:

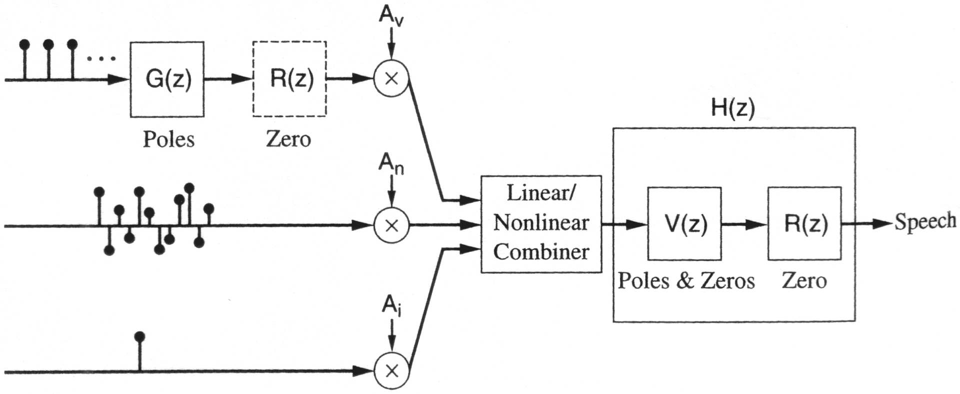


Figure 8: Complete discrete time model

$$\begin{aligned}\text{Periodic voiced speech: } X(z) &= A_v \cdot G(z) \cdot H(z) = A_v \cdot G(z) \cdot V(z) \cdot R(z) \\ \text{Fricatives (noise): } X(z) &= A_n \cdot U(z) \cdot H(z) = A_n \cdot U(z) \cdot V(z) \cdot R(z) \\ \text{Plosives: } X(z) &= A_i \cdot H(z) = A_i \cdot V(z) \cdot R(z)\end{aligned}$$

- $H(z) = V(z) \cdot R(z)$ is the discrete–time transfer function from volume velocity input to pressure output.

- $R(z) = Z_r(z)$ denotes the discrete–time radiation impedance. $R(z) \approx 1 - \alpha \cdot z^{-1}$ is approximately a differentiation of volume velocity to obtain pressure. $R(z)$ has a zero at α inside the unit circle.

- $V(z)$ is the stable discrete–time all–pole vocal tract transfer function from the volume velocity at the glottis to volume velocity at the lips. Function of speech sound, speaker and speaking style. Poles inside the unit circle.

- A_v is controlling the loudness of the sound and is determined by the subglottal pressure.

- **Periodic (voiced) speech:** $G(z)$ is the z–transform of the glottal flow input $g[n]$ over one cycle. Differs with speech sound, speaker and speaking style.

  $$G(z) = \frac{1}{(1 - \beta \cdot z)^2}$$

  For real $\beta < 1$ represents two identical poles outside the unit circle. This model considers only the loss by radiation at the lips. The source for periodic (voiced) speech is the impulse sequence $\sum_{k=-\infty}^{\infty} \delta[n - k \cdot P]$.

- **Fricatives:** $H(z)$ denotes the z–transform of a noise sequence $u[n]$.

- **Plosives:** The input for a plosive is simply modelled by a impulse.

In the noise and impulse source state, oral tract constrictions may give zeros as well as poles. Zeros in the transfer function also occur for nasal consonants and nasalized vowels. In these cases $V(z)$ has poles inside the unit circle, but may have zeros inside and outside the unit circle.

$$X(z) = A \cdot \frac{(1 - \alpha \cdot z^{-1}) \prod_{k=1}^{M_i} (1 - a_k \cdot z^{-1}) \prod_{k=1}^{M_o} (1 - b_k \cdot z)}{(1 - \beta \cdot z)^2 \prod_{k=1}^{C_i} (1 - c_k \cdot z^{-1})(1 - c_k^* \cdot z^{-1})}$$

$(1 - a_k \cdot z^{-1})(1 - b_k \cdot z)$ are zeros inside and outside the unit circle, due to oral and/or nasal tract configurations.

Vocal Fold/Vocal Tract Interaction Up to now it was assumed that:

- there is no glottal loss (glottal impedance is infinite)
- glottal air flow source isn't coupled with the vocal tract

This allowed us to model the output speech waveform as an ideal volume velocity from the glottal source convolved with a linear vocal tract impulse response. In reality there is a nonlinear coupling between the glottal airflow velocity and the pressure in the vocal tract. The treatment of this interaction is quite complicated and not further examined here.

3 Linear Prediction

A signal sample can be described as a linear combination of the preceding samples. The algorithm calculates model coefficients by minimizing the mean square error between the predicted signal and the original signal. These coefficients are recalculated every 5 to $20ms$. The system is an all–pole linear filter that simulates the source spectrum and the vocal tract transfer function. Speech sounds with periodic or impulsive sources can be classified as 'deterministic' sounds, and with noise sources ($\sigma_u^2 = 1$) as 'stochastic' sounds.

3.1 Model

The relationship between the input sequence $v[n]$ and the output sequence $x[n]$ can be represented by the difference equation:

$$x[n] = \sum_{k=0}^{m} b_k \cdot v[n-k] - \sum_{k=1}^{m} c_k \cdot x[n-k]$$

The transfer function is given by:

$$H(z) = \frac{B(z)}{C(z)} = \frac{\sum_{k=0}^{m} b_{m-k} \cdot z^k}{\sum_{k=0}^{m} c_{m-k} \cdot z^k}$$

This is the general case of a pole–zero model, however for speech–synthesis the all-pole model is used, with the difference equation:

$$x[n] = b_0 \cdot v[n] - \sum_{k=1}^{m} c_k \cdot x[n-k]$$

and the transfer function:

$$H(z) = \frac{b_0 \cdot z^m}{\sum_{k=0}^{m} c_{m-k} \cdot z^k}$$

In this model the nasal–tract and the filter of the glottis and the lips is neglected.

3.1.1 Why All-Pole Model?

A stable pole–zero system function can by decomposed into a minimum phase system $H_{min}(z)$ and an all–pass system $H_{AP}(z)$.

$$H(z) = H_{min}(z) \cdot H_{AP}(z)$$

For speech–synthesis it is sufficient to consider the minimum phase system, because the human ear is insensitive to the phase variations of the all–pass system. This implies two important consequences:

1. For the minimum phase exists a stable inverse system $H_{min}(z)^{-1} = \frac{1}{H_{min}(z)}$. Due to filtering the speech output signal $x[n]$ with the inverse vocal tract system the input signal $v[n]$ can be extracted.

2. A minimum phase pole–zero system can be represented by an all–pole system with an infinite amount of poles, which can be approximated by a filter of m–th degree.

This is the justification for the use of an all–pole filter in speech–synthesis.

3.2 Formulation

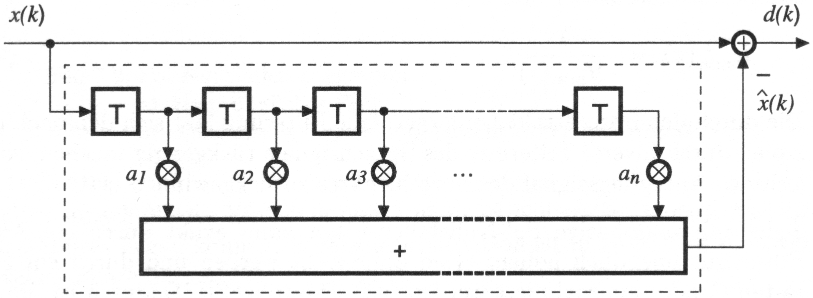


Figure 9: Forming of the prediction error signal $d[n]$

The basic idea behind linear prediction analysis is that each speech sample is approximated as a linear combination of past speech samples. The filter coefficients are determined by linear prediction, which implies inverse filtering of $x[n]$, so that the input signal $v[n]$ is extracted (usefull for coding and synthesis). The true coefficients aren't known, but an estimation can be stated, which is called a linear predictor of order p:

$$\widehat{x}[n] = \sum_{k=1}^{p} a_k \cdot x[n-k]$$

with an predictor error ($p = m$ and $b_0 = 1$):

$$\begin{aligned} d[n] &= x[n] - \widehat{x}[n] \\ &= v[n] - \sum_{k=1}^{m} c_k \cdot x[n-k] - \sum_{k=1}^{m} a_k \cdot x[n-k] \\ &= v[n] - \sum_{k=1}^{m} (c_k + a_k) \cdot x[n-k] \end{aligned}$$

If $a_k = -c_k$ then $d[n] = v[n]$, which equals an inverse filtering.

3.3 Optimum Predictor Coefficients assuming stationarity

The minimal mean–squared prediction error is defined as:

$$E\left\{d\,[n]^2\right\} \overset{!}{=} min$$

It is assumed, that the coefficients c_k and the impulse response $h[n]$ are time–invariant and that the order $p = m$ is known. The input sequence $v[n]$ is a real, stationary, uncorrelated and average–free sequence (White Noise).

Autocorrelation Function of $v[n]$

$$\varphi_{vv}[\lambda] = E\left\{v[n] \cdot v[n \pm \lambda]\right\} = \begin{cases} \sigma_v^2 & \text{for } \lambda = 0 \\ 0 & \text{for } \lambda = \pm 1, \pm 2, \ldots \end{cases}$$

Autocorrelation Function of $x[n]$

$$\varphi_{xx}[\lambda] = E\left\{x[n] \cdot x[n \pm \lambda]\right\} = \sigma_v^2 \cdot \sum_{k=0}^{\infty} h[k] \cdot h[k \pm \lambda]$$

Crosscorrelation Function of $x[n]$ and $v[n]$

$$\varphi_{vx}[\lambda] = E\left\{v[n] \cdot x[n \pm \lambda]\right\} = h[\lambda] \cdot \sigma_v^2$$

To get the minimal mean–squared prediction error we derive the equation with respect to a_λ.

$$\begin{aligned} \frac{\partial E\left\{d[n]^2\right\}}{\partial a_\lambda} &= E\left\{2 \cdot d[n] \cdot \frac{\partial d[n]}{\partial a_\lambda}\right\} = E\left\{-2 \cdot d[n] \cdot x[n-\lambda]\right\} \overset{!}{=} 0 \\ &= -2 \cdot E\left\{d[n] \cdot x[n-\lambda]\right\} = -2 \cdot E\left\{\left[x[n] - \sum_{k=1}^{p} a_k \cdot x[n-k]\right] \cdot x[n-\lambda]\right\} \\ &= -2 \cdot \varphi_{xx}[\lambda] + 2 \cdot \sum_{k=1}^{p} a_k \cdot \varphi_{xx}[\lambda - k] \overset{!}{=} 0 \end{aligned}$$

With $\lambda = 1, 2, \ldots p$ follow the so called *normal equations.*

$$\begin{bmatrix} \varphi_{xx}[1] \\ \varphi_{xx}[2] \\ \vdots \\ \varphi_{xx}[p] \end{bmatrix} = \begin{bmatrix} \varphi_{xx}[0] & \varphi_{xx}[-1] & \ldots & \varphi_{xx}[1-p] \\ \varphi_{xx}[1] & \varphi_{xx}[0] & \ldots & \varphi_{xx}[2-p] \\ \vdots & \vdots & \ldots & \vdots \\ \varphi_{xx}[p-1] & \varphi_{xx}[p-2] & \ldots & \varphi_{xx}[0] \end{bmatrix} \cdot \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix}$$

$$\varphi_{xx} = \mathbf{R}_{xx}\mathbf{a}$$

The solution for the optimal coefficients:

$$\mathbf{a}_{opt} = \mathbf{R}_{xx}^{-1}\varphi_{xx}$$

φ_{xx}	...	correlation vector
$\mathbf{R}_{xx}$	...	correlation matrix (symmetric, real → positive–definite Toeplitz matrix)

For optimal setting of the predictor coefficients the resulting error power can be solved in closed form:

$$\sigma_d^2 = E\left\{(x[n] - \widehat{x}[n])^2\right\} = \ldots = \sigma_x^2 - \sum_{\lambda=1}^{p} a_\lambda \cdot \varphi_{xx}[\lambda]$$

The ratio $G_p = \frac{\sigma_x^2}{\sigma_d^2}$ is called prediction gain. It is a quantity for bit–rate reduction due to predictive coding. G_p increases above a prediction order of $p = 3 - 4$ only slowly.

3.4 Block Oriented Adaptation of the Linear Predictor

The vocal tract and its source are time–varying. It is assumed, that the characteristics are changing slowly, so that the vocal tract and thus its transfer function remain fixed for a short–time interval. A sliding window is used to make the short–time approximation valid. The choice of the length of the window is a tradeoff between time– and frequency resolution. Also the specific shape of the window contributes to the time– and frequency resolution properties e.g.: A rectangular window has a narrower mainlobe, but higher sidelobes than a Hamming window. The optimization takes place for blocks with N samples. The block length is typically about $10 \ldots 30ms$ and with a sampling frequency of $8kHz$ one block contains $N = 80 \ldots 240$ samples. There are two possible procedures, the window can by applied on the input sequence $x[n]$ or on the prediction error sequence $d[n]$, as depicted in figure 10.

Autocorrelation Method The window $h[n]$ is applied to the input signal $x[n]$ which contains N samples until n_0.

$$\tilde{x}[n] = x[n] \cdot h[n_0 - n]$$

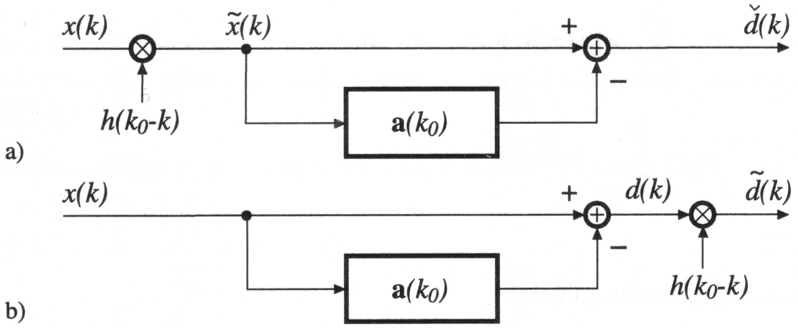


Figure 10: (a) autocorrelation method (b) covariance method

Covariance Method

$$\tilde{d}[n] = d[n] \cdot h[n_0 - n]$$

3.4.1 Stationary Solution (Autocorrelation Method)

The sequence $\tilde{x}[n]$ has finite energy and is only nonzero in the interval $n_1 = n_0 - N - 1 \leq n \leq n_0$. The predictor error signal $\tilde{d}[n]$ is due to consideration of the finite prediction order p limited to the interval $n_1 = n_0 - N - 1 \leq n \leq n_0 + p = n_2$ and also possesses finite energy. For the optimization of the energy of the sequence $\check{d}[n]$, following term can be minimized:

$$\hat{E}\left\{\check{d}^2[n]\right\} = \sum_{n=n1}^{n_2} \check{d}^2[n] \overset{!}{=} min$$

The same approach as for stationarity can be followed, but the short–time autocorrelation function has to be used.

$$\hat{\varphi}_{\tilde{x}\tilde{x}}[\lambda] = \hat{E}\left\{\tilde{x}[n] \cdot \tilde{x}[n+\lambda]\right\} = \sum_{n_1}^{n_2} \tilde{x}[n] \cdot \tilde{x}[n+\lambda] = \sum_{n_1+\lambda}^{n_0} x[n] \cdot x[n+\lambda]$$

With the simplification in notation $r_\lambda \doteq \hat{\varphi}_{\tilde{x}\tilde{x}}$ it follows:

$$\begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_p \end{bmatrix} = \begin{bmatrix} r_0 & r_1 & \dots & r_{p-1} \\ r_1 & r_2 & \dots & r_{p-2} \\ \vdots & \vdots & \dots & \vdots \\ r_{p-1} & r_{p-2} & \dots & r_0 \end{bmatrix} \cdot \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix}$$

$$\mathbf{r} = \mathbf{R}_{\tilde{x}\tilde{x}}\mathbf{a}$$

$\mathbf{r}$ ... short–time correlation vector

$\mathbf{R}_{\tilde{x}\tilde{x}}$ ... short–time correlation matrix (symmetric, Toeplitz matrix)

The autocorrelation method is a suboptimal method, but leads to the normal equation system, which can be solved efficiently by the *Levinson–Durbin algorithm.*

3.4.2 Instationary Solution (Covariance Method)

The energy of the predictor error signal $\tilde{d}[n]$ is minimized over an interval of length N.

$$\hat{E}\left\{\tilde{d}^2[n]\right\} = \sum_{n=n1}^{n_0} \tilde{d}^2[n] \overset{!}{=} min$$

With the relationship

$$d[n] = x[n] - \sum_{k=1}^{p} a_k \cdot x[n-k]$$

the partial derivative to a_λ yields

$$\begin{aligned}
\frac{\partial \hat{E}\left\{\tilde{d}[n]^2\right\}}{\partial a_\lambda} &= \hat{E}\left\{2 \cdot \tilde{d}[n] \cdot \frac{\partial \tilde{d}[n]}{\partial a_\lambda}\right\} \hat{E}\left\{-2 \cdot \tilde{d}[n] \cdot x[n-\lambda]\right\} \overset{!}{=} 0 \\
&= -2 \cdot E\left\{d[n] \cdot x[n-\lambda]\right\} = -2 \cdot \sum_{n=n_1}^{n_0} \left[x[n] - \sum_{k=1}^{p} a_k \cdot x[n-k]\right] \cdot x[n-\lambda]) \\
\rightarrow \quad & \sum_{n=n_1}^{n_0} x[n] \cdot x[n-\lambda] = \sum_{k=1}^{p} a_k \sum_{n=n_1}^{n_0} x[n-k] \cdot x[n-\lambda]
\end{aligned}$$

Considering the factor $1/N$ in this equation system the short–time autocorrelation function measuring values $\hat{\varphi}_{xx}[n,-k,-\lambda]$ can be substituted. Introducing the simplification:

$$\hat{r}_{i,\lambda} \doteq N \cdot \hat{\varphi}_{xx}[n,-k,-\lambda] = \sum_{n=n_1}^{n_0} x[n-k] \cdot x[n-\lambda]$$

Thus follows the equation system:

$$\begin{bmatrix} \hat{r}_{0,1} \\ \hat{r}_{0,2} \\ \vdots \\ \hat{r}_{0,p} \end{bmatrix} = \begin{bmatrix} \hat{r}_{1,1} & \hat{r}_{1,2} & \dots & \hat{r}_{1,p} \\ \hat{r}_{1,2} & \hat{r}_{2,2} & \dots & \hat{r}_{2,p} \\ \vdots & \vdots & \dots & \vdots \\ \hat{r}_{1,p} & \hat{r}_{2,p} & \dots & \hat{r}_{p,p} \end{bmatrix} \cdot \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix}$$

$$\mathbf{\hat{r}_0} = \mathbf{\hat{R}}_{xx}\mathbf{a}$$

$\mathbf{\hat{R}}_{xx}$ has no Toeplitz structure, so more complicated algorithms are needed to solve this equation system.

3.4.3 Levinson–Durbin–Algorithm

This is an effective algorithm for solving the equation system of the autocorrelation method.

$$\mathbf{r} = \mathbf{R}_{\tilde{x}\tilde{x}}\mathbf{a}$$

Solving the equation system by building the inverse matrix, would require p^3 multiplications and additions, on the contrary the recursive solution only needs p^2 multiplications and additions. This algorithm finds from the known solution $\mathbf{a}^{p-1}$ the solution $\mathbf{a}^p$ with little effort. Modifying the notation $\alpha_k = -a_k$. Starting from $p = 0$ (no prediction), the solution for $p = m$ is calculated in m recursions.

Algorithm

1. Calculation of $m + 1$ values of the short–time autocorrelation

2. $p = 0$, no prediction
$$d[n] = x[n], \quad \hat{E}^{(0)} = r_0, \quad \alpha_0^{(0)} \doteq 1$$

3. for $p \geq 1$

   (a) $q = \sum_{k=0}^{p-1} \alpha_k^{(p-1)} r_{p-k}, \qquad k_p = -\dfrac{q}{\hat{E}^{(p-1)}}$

   (b) $\alpha_p^{p-1} = 0$

   (c) $\alpha_k^{(p)} = \alpha_k^{(p-1)} + k_p \cdot \alpha_{p-k}^{(p-1)} \qquad 0 \leq k \leq p$

   (d) $\hat{E}^{(p)} = \hat{E}^{(p-1)} \cdot (1 - k_p^2)$

   (e) $p = p + 1$

4. Repeat step 3 while $p \leq m$

5. $a_k = -\alpha_k^{(n)}$

The algorithm delivers the predictor coefficients a_k and the reflection coefficients k_p. The predictor can be implemented in direct form or lattice structure, as can be seen in figure 11.

Two approaches:

- The cross–sectional area ratios aren't known. Out of a speech measurement the reflection coefficients k_p are known. The reflection coefficients k_p coincide with the reflection coefficients of the tube model, so the cross–sectional area ratios can be computed by:
$$\frac{A_{i+1}}{A_i} = \frac{1 + k_i}{1 - k_i}$$

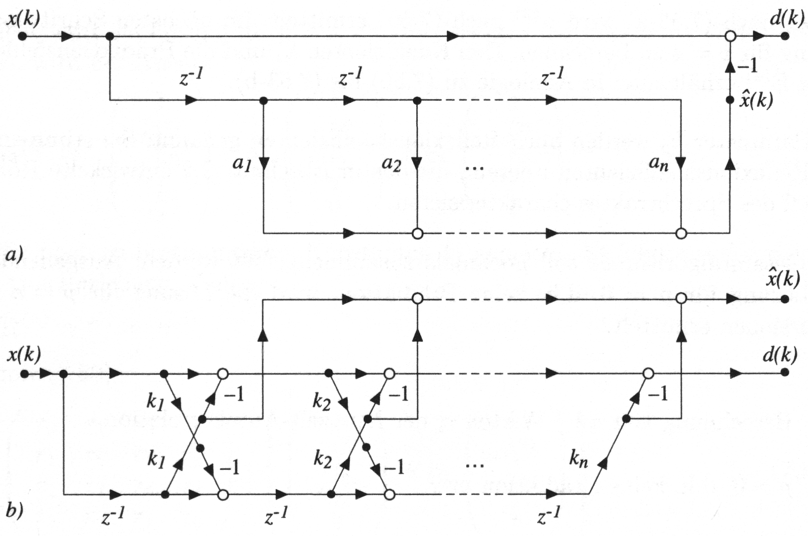


Figure 11: (a) direct form (b) lattice structure

- The cross–sectional areas of the tubes are known. Out of these the reflection coefficients and the predictor coefficients and hence the transfer function of the vocal tract can be computed.

The effect of the predictive filtering can be seen, if the waveform and spectrum of $x[n]$ and $g[n]$ are compared. There is a reduction of dynamics in the time signal $g[n]$, which corresponds to the glottis signal and a whitening in the spectrum of $g[n]$ due to the inverse filtering, as depicted in figure 12.

3.5 Sequential Adaptation of the Linear Predictor

In block–oriented adaptation the predictor coefficients are calculated after N samples. In sequential adaptation the coefficients are altered after every sample.

4 Formant Synthesizer

This synthesizer type is based on the source/filter–theory of speech production. The basic assumption is that the transfer function of the vocal tract can be modelled by simulating formant frequencies and formant amplitudes. The synthesis consists of the artificial reconstruction of the formant characteristics. This is done by exciting a set of resonators by a voicing source, which simulates the vocal fold vibration or a noise generator, that simulates a constriction somewhere in the vocal tract.

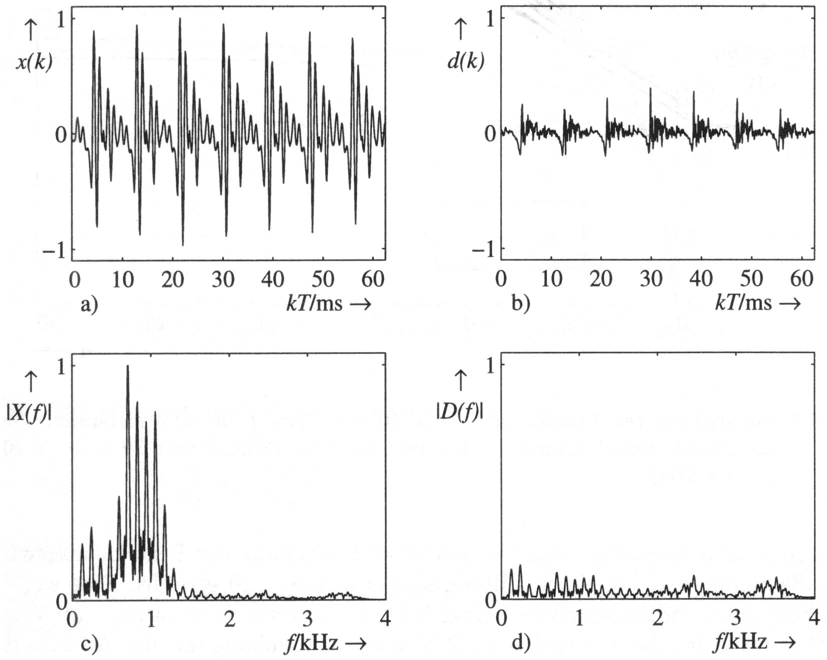


Figure 12: (a) speech signal $x[n]$ (b) prediction error signal $d[n]$ (c) spectrum of $x[n]$ (d) spectrum of $d[n]$

4.1 Source Modelling

Speech sounds can be divided into those produced by a periodic vibration of the vocal folds (voiced sounds), and those generated without vocal fold vibrations, but with plosive or friction noise (voiceless sounds).

Two Sources

- Voicing Source, producing a quasi–periodic wave
- Friction Source, producing noise

Voicing Source The model is composed of an impulse generator that simulates the acoustic energy and a linear filter whose frequency response $G(f)$ approximates the glottal wave form. Different timing of the glottal opening results in different spectral characteristics e.g.: the slope of the spectrum. A gain control allows the adjustment of the voicing amplitude, as can be seen in figure 13.

Models for characterising $G(f)$

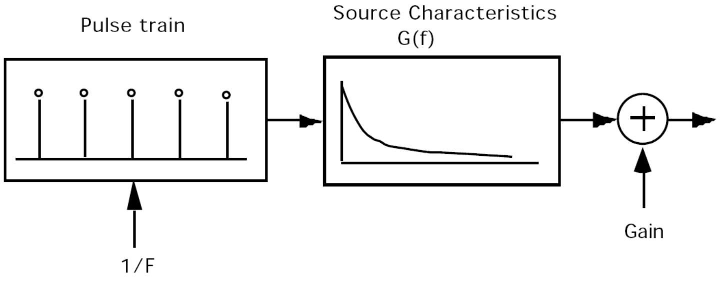


Figure 13: Voicing Source

- Low–pass filter with variable slope (simple)
- Mathematical function e.g. by Fant and Liljencrants (figure 14)
  The discrete time radiation impedance $R(z) = 1 - \alpha \cdot z^{-1}$, can approximately be thought as a differentiator of volume velocity to obtain pressure at the lips.

  $$x(t) \approx A \cdot \frac{d}{dt}[u_g(t) * v(t)] = A \cdot \left[\frac{d}{dt}u_g(t)\right] * v(t)$$

  So the source to the vocal tract can thought to be the derivative of the glottal airflow. The glottal waveform can be composed by two time–reversed exponentially decaying sequences and is described by the so called LF–Parameters.
- Mechanical simulation of the vocal fold vibration

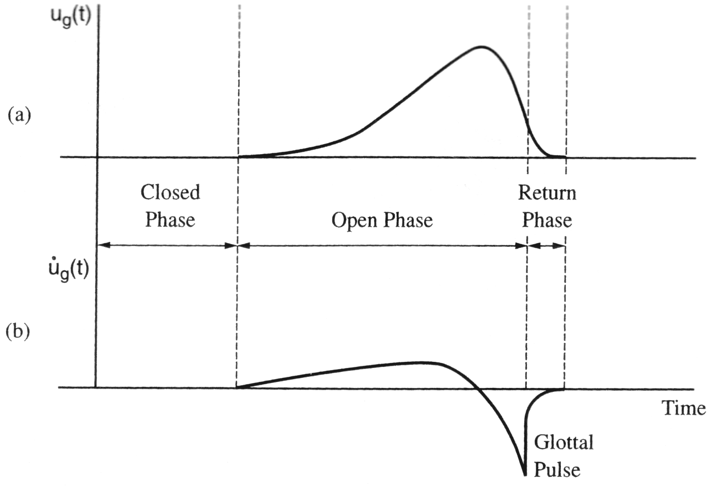


Figure 14: (a) glottal airflow (b) derivative of the glottal airflow

Friction Source To produce voiceless sounds, the vocal folds are non–vibrating and held open. The air becomes turbulent across a constriction formed in the vocal tract. The usual model of a friction source consists of a pseudo random white noise generator and a gain parameter. This model is not very accurate, because the actual friction source is not located in the larynx, but at the location of the main constriction.

4.2 Vocal Tract Modelling

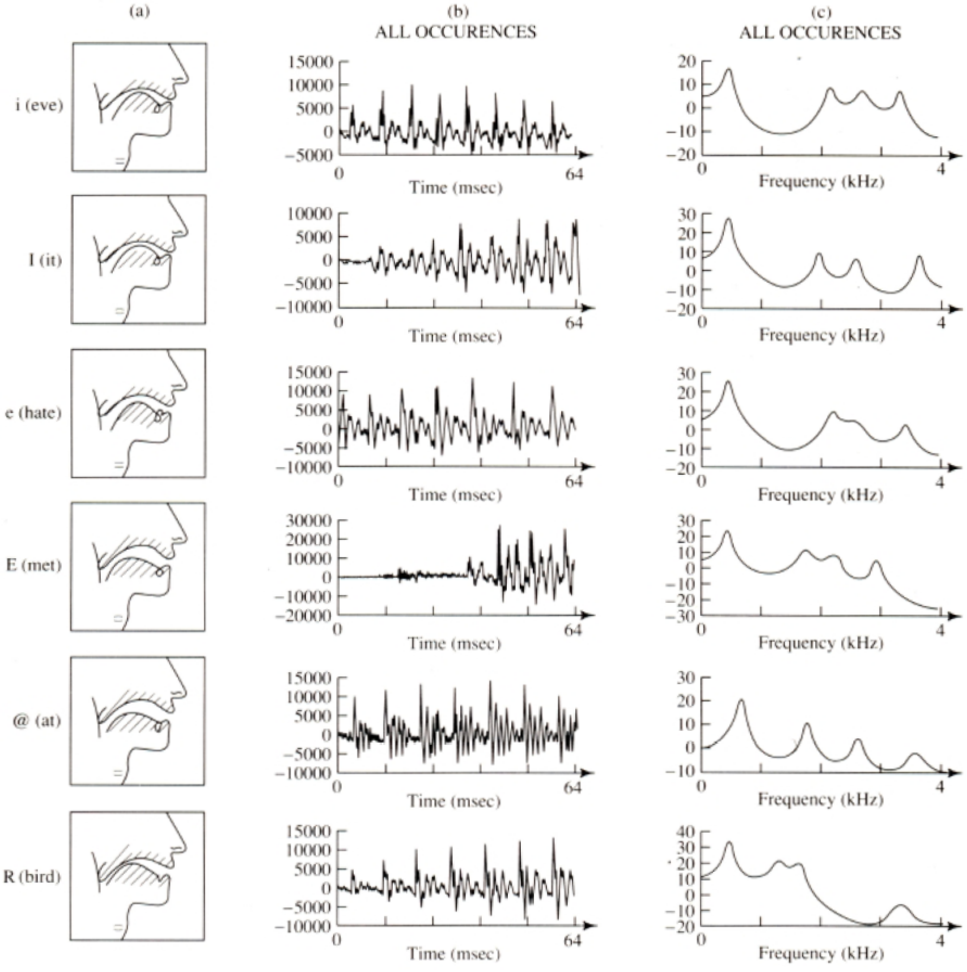


Figure 15: Formant characteristics of vowels

The source/filter modelling of vocal tract behaviour is intended to simulate the formant characteristics of the various speech sounds, as can be seen in figure 15. It is assumed that there is no source/tract interaction. For vowels the sound wave propagates through the vocal tract, where some frequencies are augmented and others are depressed. The resonance frequencies for a given sound are called the formants. These resonances may be assumed to behave as a all–pole filter, and thus can be modelled with a set of poles. The poles are identified as peaks in the frequency spectrum. The nasal cavities can be coupled with the vocal tract, and anti–resonants (anti–formants) appear in the spectrum, as a result of the damping properties of the nasal cavities. They can be modelled as zeros in the transfer function. Such spectras can be seen in figure 16.

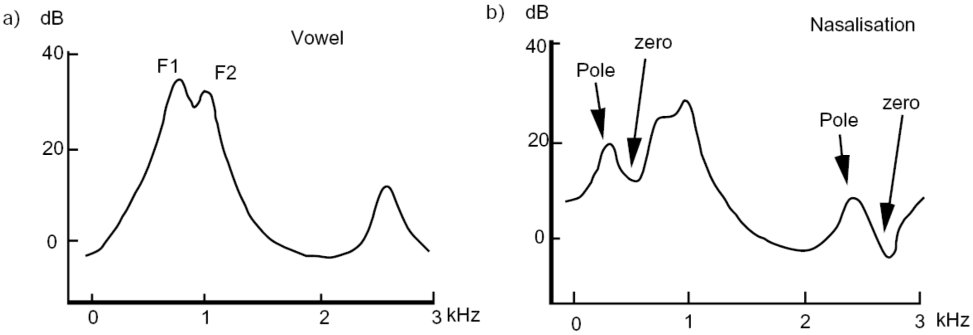


Figure 16: (a) vowel (b) nasalized vowel

A good approximation is to form each formant with an electrical resonator, such as a band–pass filter (pole filter). The band–pass filter can be described by two parameters, the resonance frequency F and the bandwith BW. A spectrum including the formants can be modelled by a superposition of resonators at different frequencies, as in figure 17.

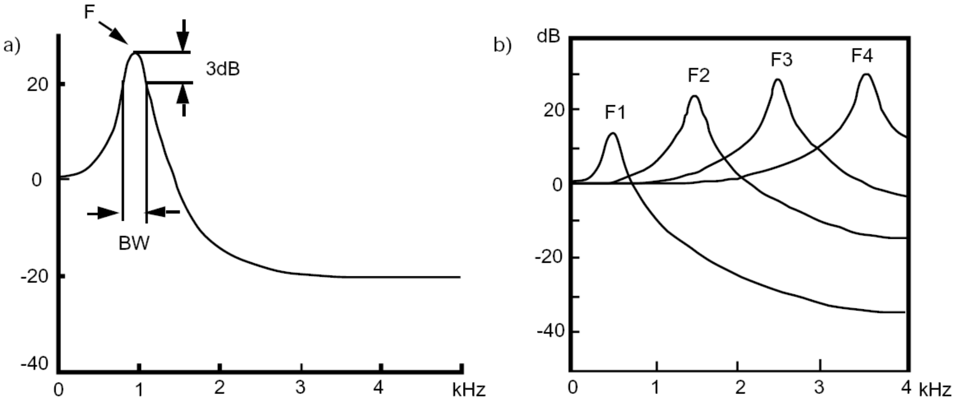
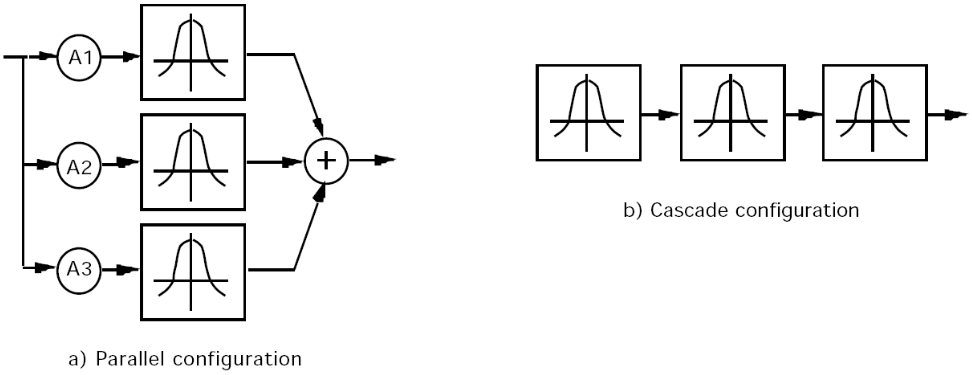


Figure 17: (a) Spectrum of a single formant resonator (b) Superposition of resonators

These resonators can be combined in parallel or cascade configuration. The parallel approach allows the control of each formant amplitude, but is not an accurate imitation of the vocal tract behaviour in speech. It is better adapted at producing consonants. The cascade approach direct produces a replica of the total formant energy distribution, and is good for synthesising vowel sounds. Both configurations are depicted in figure 18.



a) Parallel configuration

b) Cascade configuration

Figure 18: Resonator configurations

4.3 Lip Radiation

At the mouth opening the sound wave is constrained in its propagation. The modification of the propagation medium is identified by the radiation impedance. Higher frequencies are better transmitted through the opening than the low frequency counterparts, so that this can be modelled by a high–pass filter.

4.4 Klatt Formant Synthesizer

This synthesizer was proposed by Dennis Klatt in 1979. This is a direct application of the source/filter–theory, implemented by a computer simulation of an electrical structure, consisting of resonators combined in cascade or parallel. It is a cascade/parallel synthesizer which allows simulation of male and female voices. 40 parameters determine the output wave. 34 of these can be varied dynamically and the remaining are constant. A block diagram of the synthesizer can be seen in figure 19.

The synthesizer implements two voicing sources, one destined for vowels and the other for voiced fricatives. The low–pass filters (RGP, RGS, RGZ) modify the spectral shape of the source. The friction source is implemented by a random noise generator. The vocal tract in its cascade configuration is realised with five resonators ($R_1..R_5$), whose central frequence and bandwith can be adjusted. The supplemental resonator (RNP) and anti–resonator (RNZ) are used for nasalized sounds. The parallel configuration consists of seven formant

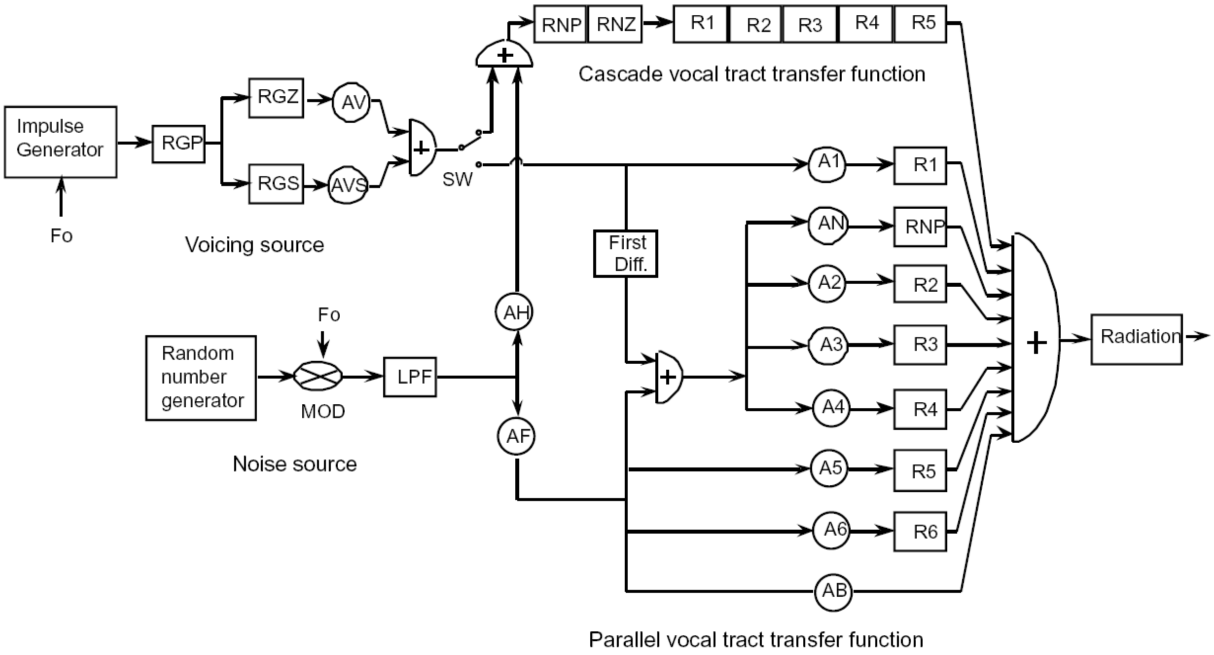


Figure 19: Block diagram of the Klatt 79 synthesizer

resonators ($R_1..R_6, RNP$), each having an individual gain control ($A_1..A_6, AN$). The whole set of the control parameters can be seen in table 3.

Symbol	C/V	Min.	Max.	Name
DU	C	30	5000	Duration of the utterance (ms)
NWS	C	1	20	Update interval for parameter reset (ms)
SR	C	5000	20000	Output sampling rate (Hz)
NF	C	1	6	Number of formants in cascade branch
SW	C	0	1	0=Cascade, 1=Parallel tract excitation by AV
G0	C	0	80	Overall gain scale factor (dB)
F0	V	0	500	Fundamental frequency (Hz)
AV	V	0	80	Amplitude of voicing (dB)
AVS	V	0	80	Amplitude of quasi-sinusoidal voicing (dB)
FGP	V	0	600	Frequency of glottal resonator ”RGP”
BGP	V	50	2000	Bandwidth of glottal resonator ”RGP”
FGZ	V	0	5000	Frequency of glottal anti-resonator ”RGZ”
BGZ	V	100	9000	Bandwidth of glottal anti-resonator ”RGZ”
BGS	V	100	1000	Bandwidth of glottal resonator ”RGS”
AH	V	0	80	Amplitude of aspiration (dB)
AF	V	0	80	Amplitude of frication (dB)
F1	V	180	1300	Frequency of 1st formant (Hz)
B1	V	30	1000	Bandwidth of 1st formant (Hz)
F2	V	550	3000	Frequency of 2nd formant (Hz)
B2	V	40	1000	Bandwidth of 2nd formant (Hz)
F3	V	1200	4800	Frequency of 3rd formant (Hz)
B3	V	60	1000	Bandwidth of 3rd formant (Hz)
F4	V	2400	4990	Frequency of 4th formant (Hz)
B4	V	100	1000	Bandwidth of 4th formant (Hz)
F5	V	3000	6000	Frequency of 5th formant (Hz)
B5	V	100	1500	Bandwidth of 5th formant (Hz)
F6	V	4000	6500	Frequency of 6th formant (Hz)
B6	V	100	4000	Bandwidth of 6th formant (Hz)
FNP	V	180	700	Frequency of nasal pole (Hz)
BNP	V	40	1000	Bandwidth of nasal pole (Hz)
FNZ	V	180	800	Frequency of nasal zero (Hz)
BNZ	V	40	1000	Bandwidth of nasal zero (Hz)
AN	V	0	80	Amplitude of nasal formant (dB)
A1	V	0	80	Amplitude of 1st formant (dB)
A2	V	0	80	Amplitude of 2nd formant (dB)
A3	V	0	80	Amplitude of 3rd formant (dB)
A4	V	0	80	Amplitude of 4th formant (dB)
A5	V	0	80	Amplitude of 5th formant (dB)
A6	V	0	80	Amplitude of 6th formant (dB)
AB	V	0	80	Amplitude of bypass path (dB)

Table 3: Control parameters of the Klatt Synthesizer

References

[1] Thomas F. Quatieri. *Speech Signal Processing*. Prentice Hall PTR, Upper Saddle River, first edition, 2002.

[2] T. Styger and E. Keller. *Fundamentals of Speech Synthesis and Speech Recognition*, chapter 6, Formant Synthesis, pages 109–128. Wiley, 1994.

[3] W. Hess, U. Heute, and P. Vary. *Digitale Sprachsignalverarbeitung*. Teubner, Stuttgart, first edition, 1998.