Source/Filter – Model

Flohberger Markus

maxiko@panacea.at

TU-Graz

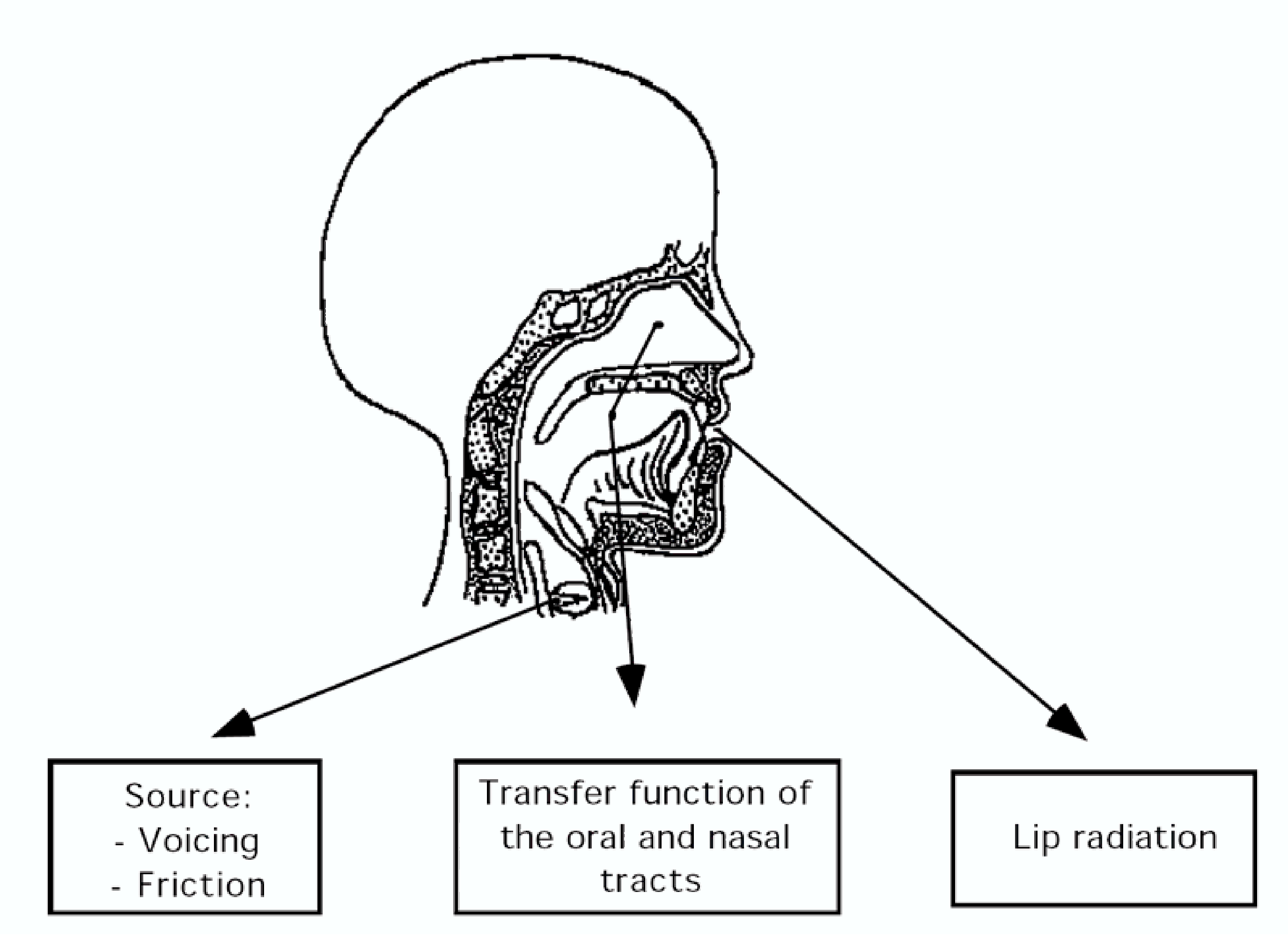
Overview (1/2)

- Introduction to Source/Filter–Model
- Acoustic Tube Models
  - Lossless Uniform Tube
  - Nonuniform Tube: Considering Losses
  - Discrete–Time Model: Concatenated Tubes
- Linear Prediction
  - Assuming Stationarity
  - Block Oriented Adapdation
  - Efficient Computation

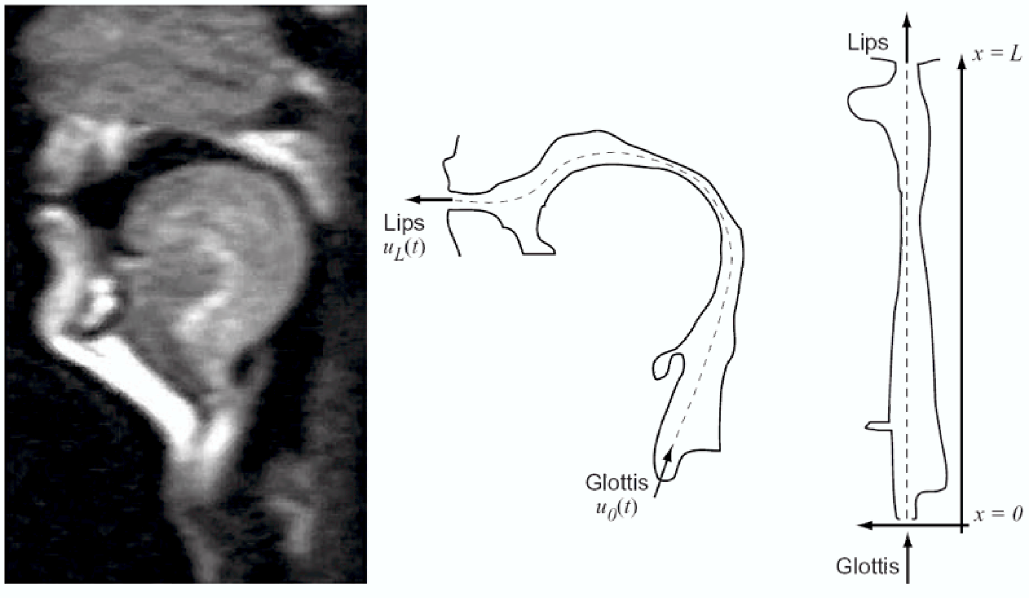
Overview (2/2)

- Formant Synthesizer
  - Source Modelling
  - Vocal Tract Modelling
  - Example: Klatt Synthesizer

Introduction (1/1)



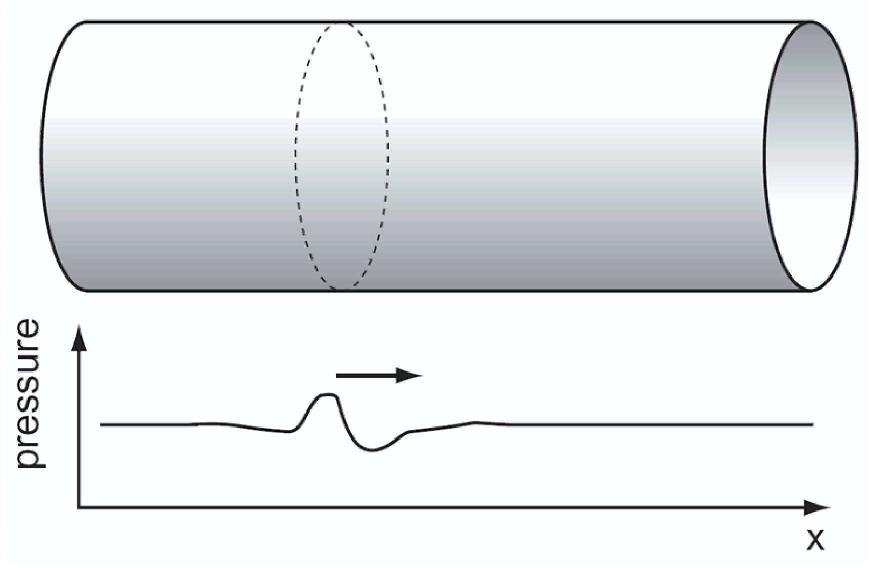
Acoustic Tube Models (1/7)



- Uniform Lossless Tube
- Nonuniform Tube: Considering Losses
- Discrete–Time Concatenated Tube Model

Lossless Uniform Tube Model (2/7)

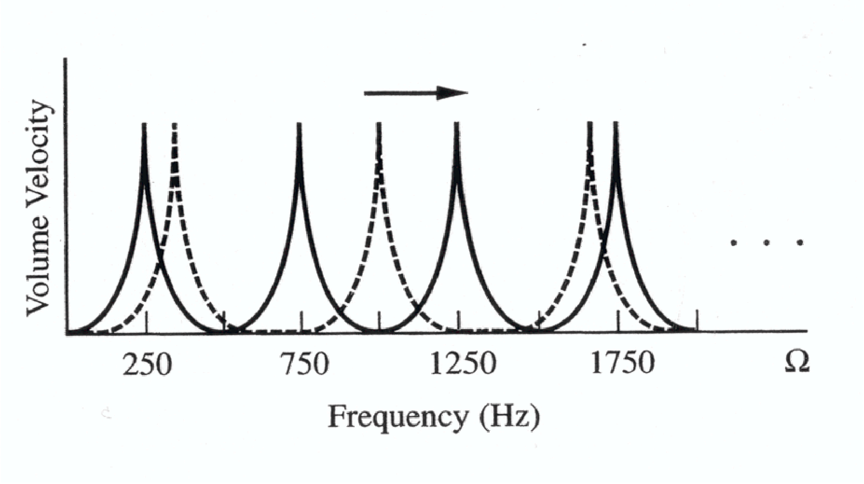
- Constant cross–section
- Moving piston corresponds an ideal particle velocity source
- The open end represents the opened lips



Lossless Uniform Tube Model (3/7)

- The transfer function is described by its poles, which correspond to the peaks in the spectrum

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

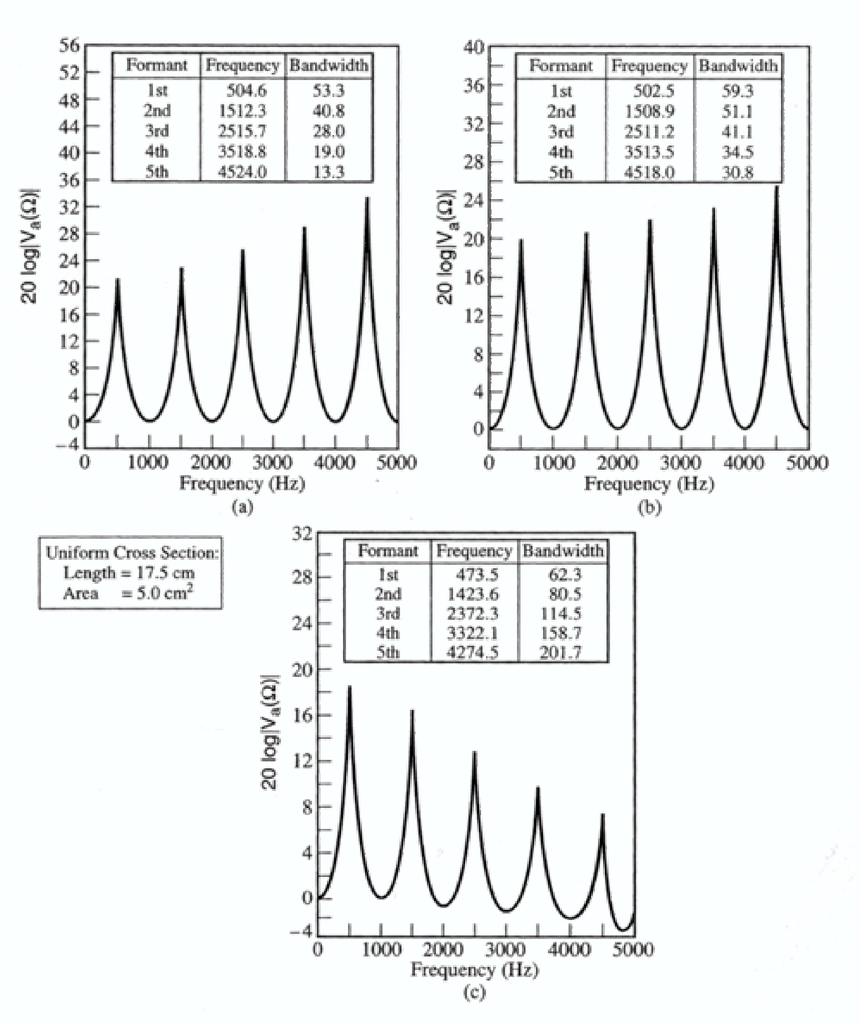


Complete Model (4/7)

- Wall Vibration
  - The walls move under the pressure induced by sound propagation
- Viscosity and Thermal Loss
  - Friction of air particles along wall
  - Heat loss through the vibrating wall
- Boundary Effects (Losses at in–/output)
  - Sound radiation is modelled by an acoustic impedance
  - Acoustic impedance for losses at glottis

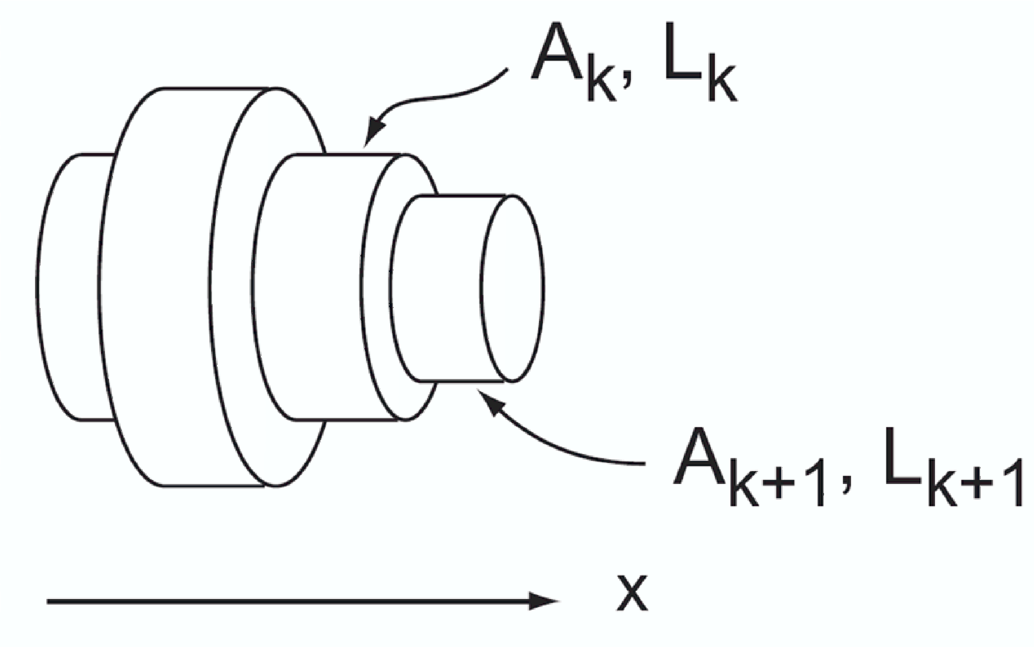
Complete Model (5/7)

- The losses are modelled by partial differential equations coupled to the wave equation
- Can only be solved by numerical simulation
- Shift of resonant frequencies
- Alteration of band-widths



Concatenated Tube Model (6/7)

- Concatenation of short lossless tubes
- Energy loss only at boundaries (glottis, lips)
- Model is linear



Concatenated Tube Model (7/7)

- Boundary conditions (junctions, glottis r_G and lips r_L) $\rightarrow$ Pressure and volume velocity continuous in time and space
- At discontinuities occurs propagation and reflection $\rightarrow$ Reflection coefficients are a function of the cross–section areas
- All–pole transfer function $V(z)$ is a function of the reflection coefficients
  $\rightarrow$ Estimate area functions, thus obtain $V(z)$
- Model not consistent with underlying physics, however formant bandwidths can be controlled with boundaries at glottis and lips

Linear Prediction (1/9)

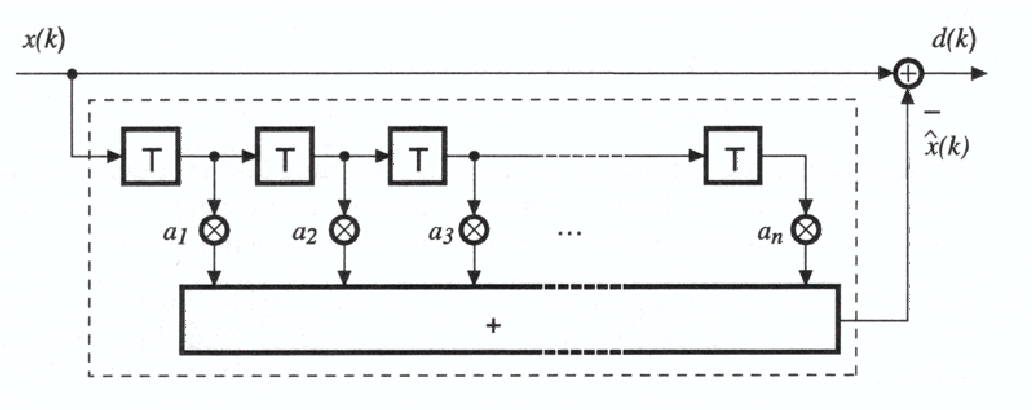
- Signal sample can be described as a linear combination of the preceding samples
- Model coefficients are calculated by minimizing the mean square error between the predicted and the original signal
- The system is an all–pole linear filter that simulates the source spectrum and the vocal tract transfer function

Formulation (2/9)

Prediction Error:

$$d[n] = x[n] - \hat{x}[n] = v[n] - \sum_{k=1}^{m} (c_k + a_k) \cdot x[n-k]$$

if the predictor coefficients $a_k = -c_k$ then $d[n] = v[n]$
$\rightarrow$ equals an inverse filtering



Stationarity (3/9)

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

- c_k and $h[n]$ are time–invariant
- v[n] is white noise sequence

$$\frac{\partial E\left\{d[n]^2\right\}}{\partial a_\lambda} = E\left\{2 \cdot d[n] \cdot \frac{\partial d[n]}{\partial a_\lambda}\right\} = -2 \cdot E\left\{d[n] \cdot x[n-\lambda]\right\} \overset{!}{=} 0$$

$$= -2 \cdot E\left\{\left[x[n] - \sum_{k=1}^{p} a_k \cdot x[n-k]\right] \cdot x[n-\lambda]\right\}$$

Normal Equations (4/9)

$$\frac{\partial E\left\{d[n]^2\right\}}{\partial a_\lambda} = -2 \cdot \varphi_{(x,x)}[\lambda] + 2 \cdot \sum_{k=1}^{m} a_k \cdot \varphi_{(x,x)}[\lambda - k] \stackrel{!}{=} 0$$

$$\begin{bmatrix} \varphi_{(x,x)}[1] \\ \varphi_{(x,x)}[2] \\ \vdots \\ \varphi_{(x,x)}[p] \end{bmatrix} = \begin{bmatrix} \varphi_{(x,x)}[0] & \varphi_{(x,x)}[-1] & \dots & \varphi_{(x,x)}[1-p] \\ \varphi_{(x,x)}[1] & \varphi_{(x,x)}[0] & \dots & \varphi_{(x,x)}[2-p] \\ \vdots & \vdots & \dots & \vdots \\ \varphi_{(x,x)}[p-1] & \varphi_{(x,x)}[p-2] & \dots & \varphi_{(x,x)}[0] \end{bmatrix} \cdot \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix}$$

$$\boldsymbol{\varphi}_{(x,x)} = \boldsymbol{R}_{(x,x)} \cdot \boldsymbol{a} \Rightarrow \boldsymbol{a} = \boldsymbol{R}_{(x,x)}^{-1} \cdot \boldsymbol{\varphi}_{(x,x)}$$

Block Oriented (5/9)

- Vocal tract and source are time–varying
- Only slow changes assumed
  - Characteristics are fixed for a short–time interval
  - Approximation with window
- Length of window (time/frequency resolution)
- Type of window (e.g. Hamming)
  - Trade–Off between Side/Mainlobe characteristics
- Block length *10...30ms*, sampling frequency *8kHz* $\rightarrow$ *N=80...240* samples

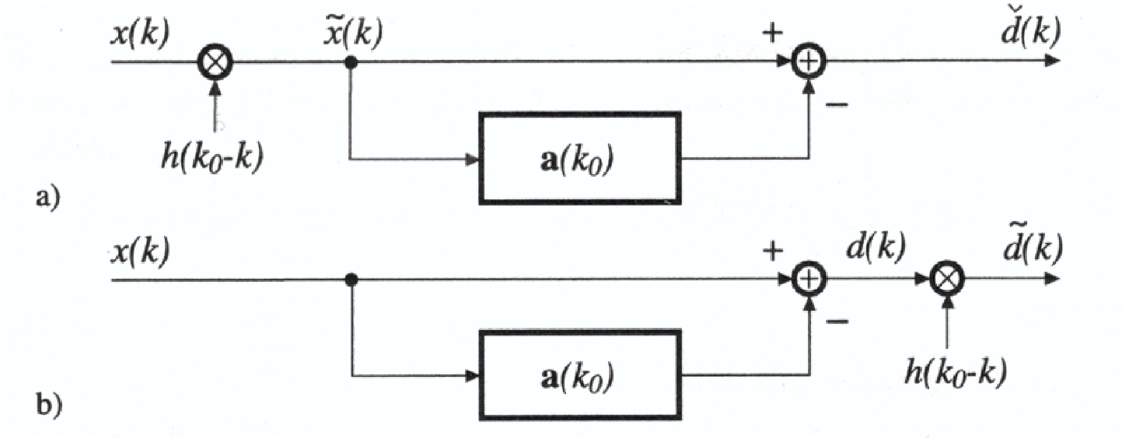
Methods (6/9)

- Autocorrelation Method: Window $h[n]$ applied to $x[n]$

$$r = R_{(\tilde{x},\tilde{x})} a$$

- Covariance Method: Window $h[n]$ applied to $d[n]$

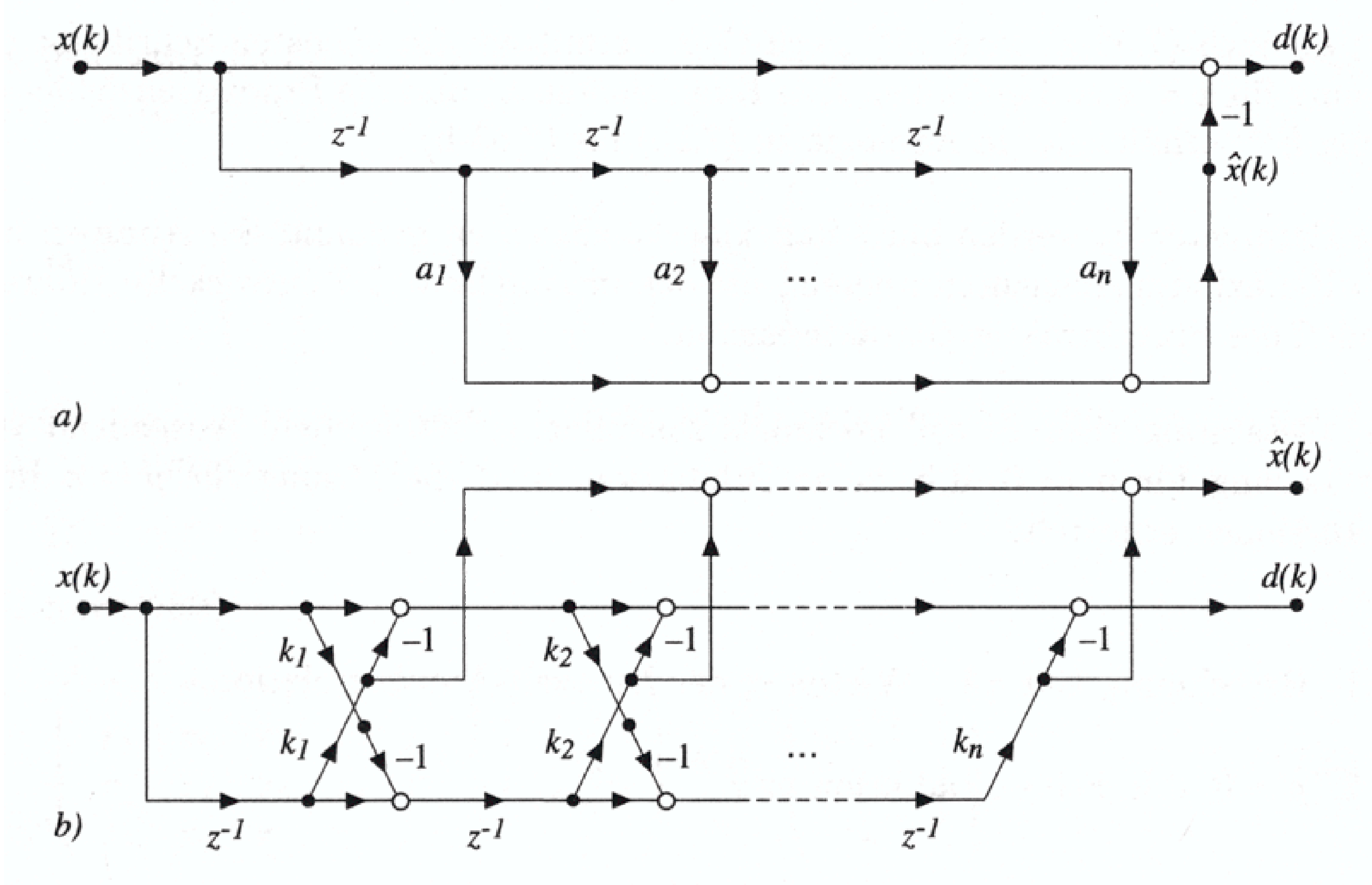
$$\hat{r}_0 = \hat{R}_{(x,x)} a$$



Levinson–Durbin–Algortihm (7/9)

- Effective algorithm for solving normal equation system of the autocorrelation method, delivers predictor coefficients a_k and reflection coefficients k_p
  - Gauss algorithm: p^3 multiplications and additions
  - Recursive algorithm: p^2 multiplications and additions
- The predictor can be implemented in direct form or lattice structure

Levinson–Durbin–Algortihm (8/9)



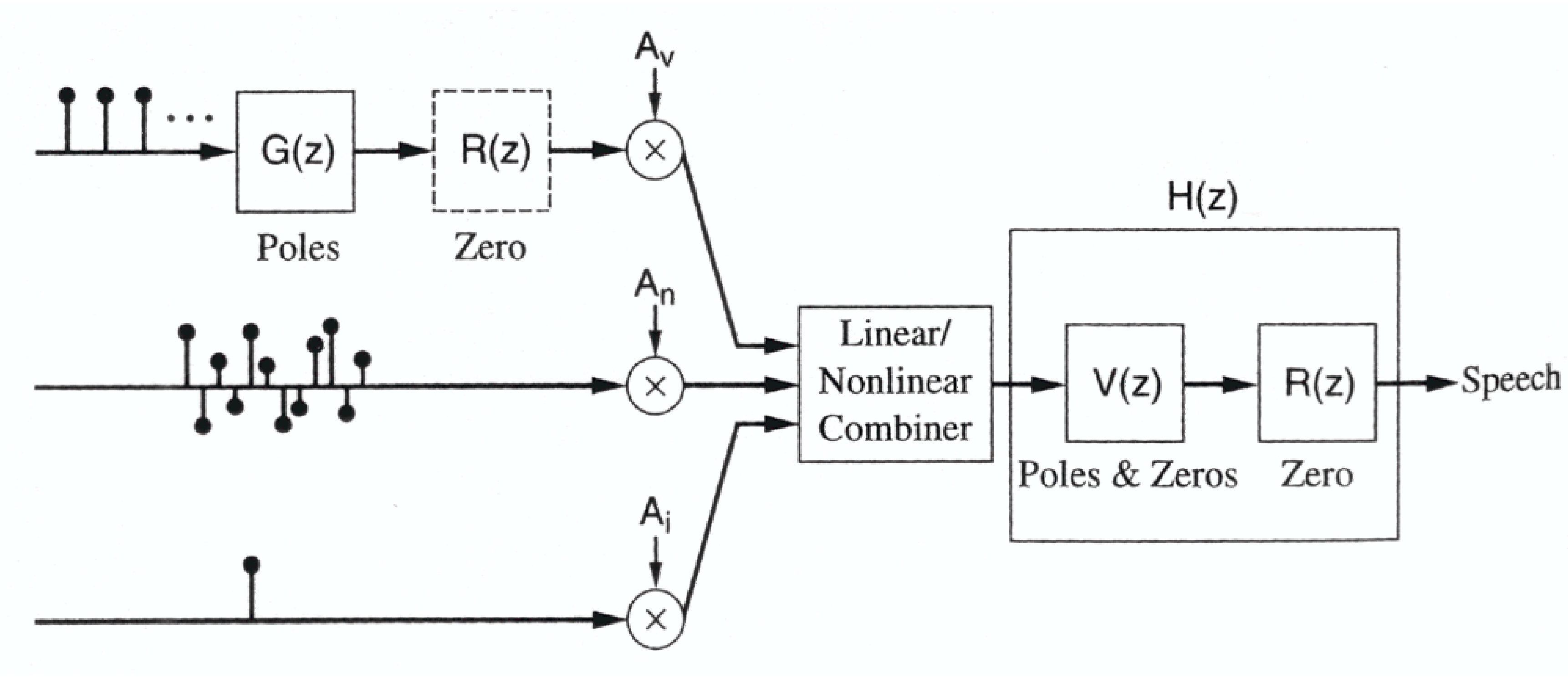
Levinson–Durbin–Algortihm (9/9)

- Computation
  1. Out of speech measurement the autocorrelation coefficients can be computed. After solving the recursion, the cross–sectional areas can be calculated.
  2. Out of the cross–sectional areas the reflection and predictor coefficients and hence the transfer function of the vocal tract can be computed.
- Comparison $x[n]$ and $d[n]$
  - Whitening in the spectrum of $d[n]$
  - Reduction in dynamics in $d[n]$

Formant Synthesizer (1/16)

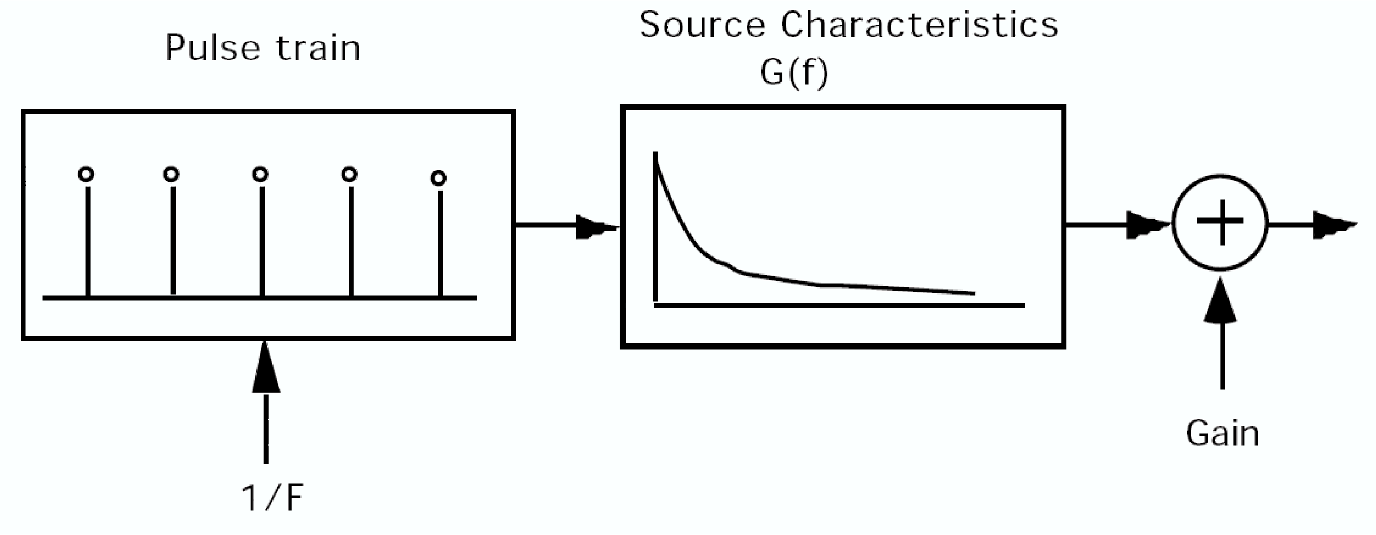
- Based on the source/filter–theory of speech production
- Vocal tract transfer function can be modelled by simulating formant frequencies and formant amplitudes
- Artificial reconstruction of formant characteristics by exciting resonators by a source
  - Voicing Source $\rightarrow$ Simulates vocal fold vibration
  - Noise Generator $\rightarrow$ Simulates constriction in vocal tract
- Pros: Good restitution of the speech signal
  Cons: Automatic techniques are unsatisfactory

Speech Production Model (2/16)



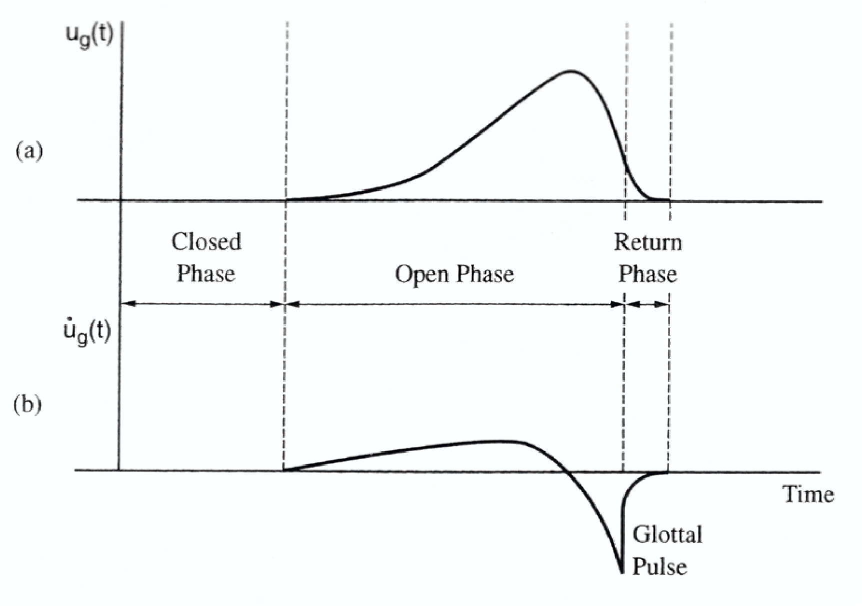
Voicing Source (3/16)

- Vocal folds are vibrating periodically (voiced sounds)
- Modelled as an impulse gernerator and a linear filter with frequency response $G(f)$



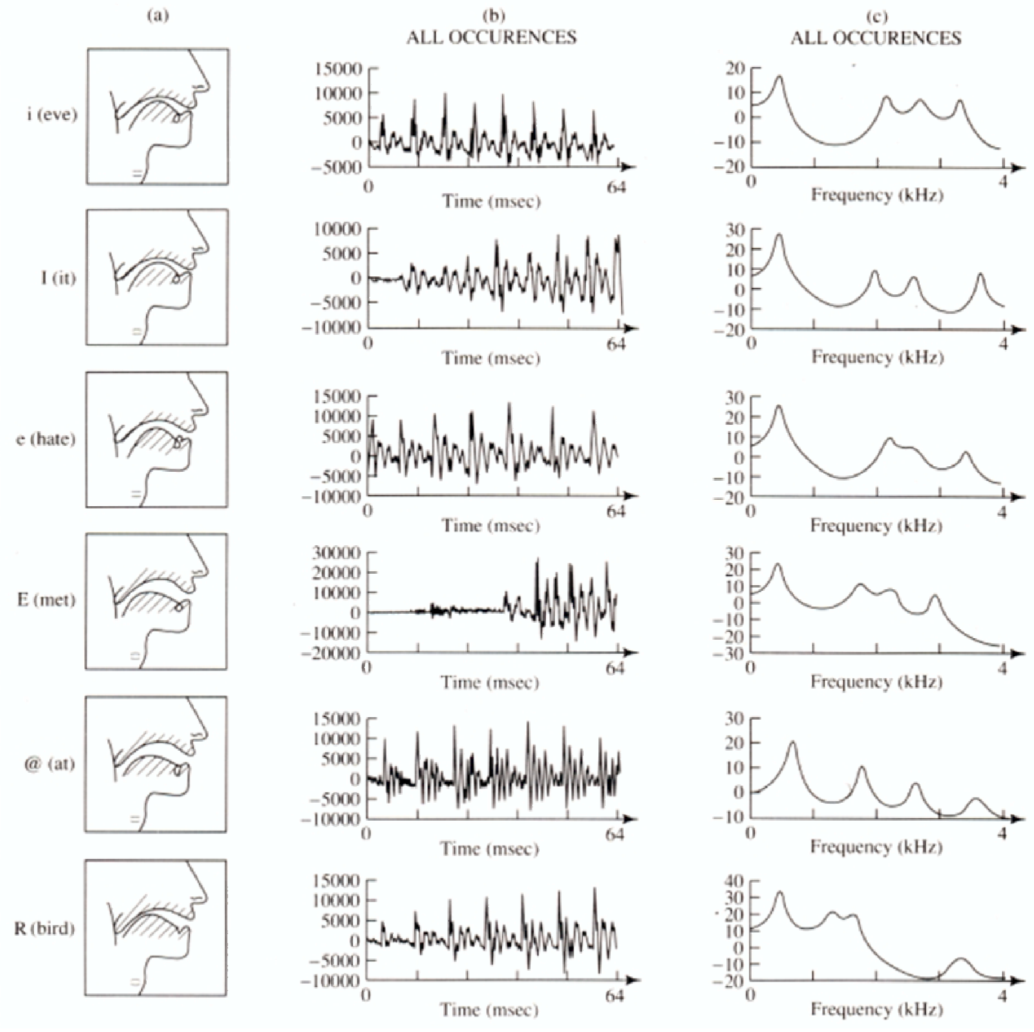
Models for $G(f)$ (4/16)

- Low–pass filter with variable slope
- Mathematical function e.g. by Fant and Liljencrants



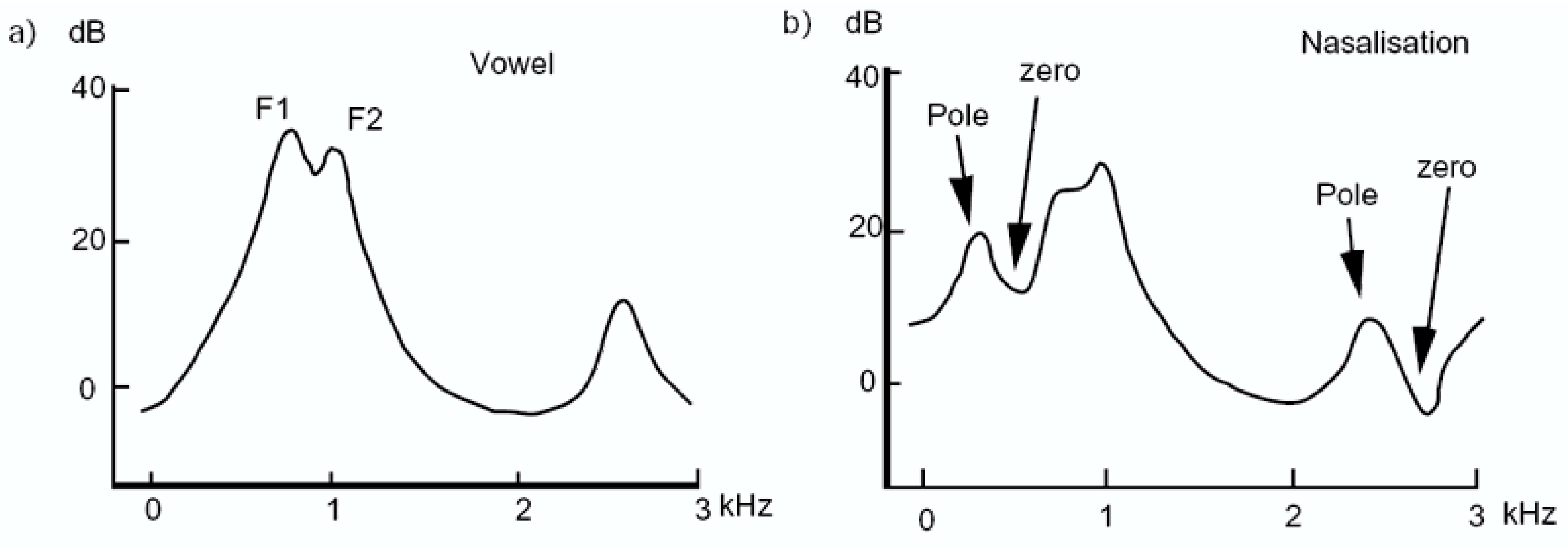
- Mechanical simulation of the vocal fold vibration (Animation!)

Vocal Tract Modelling (5/16)



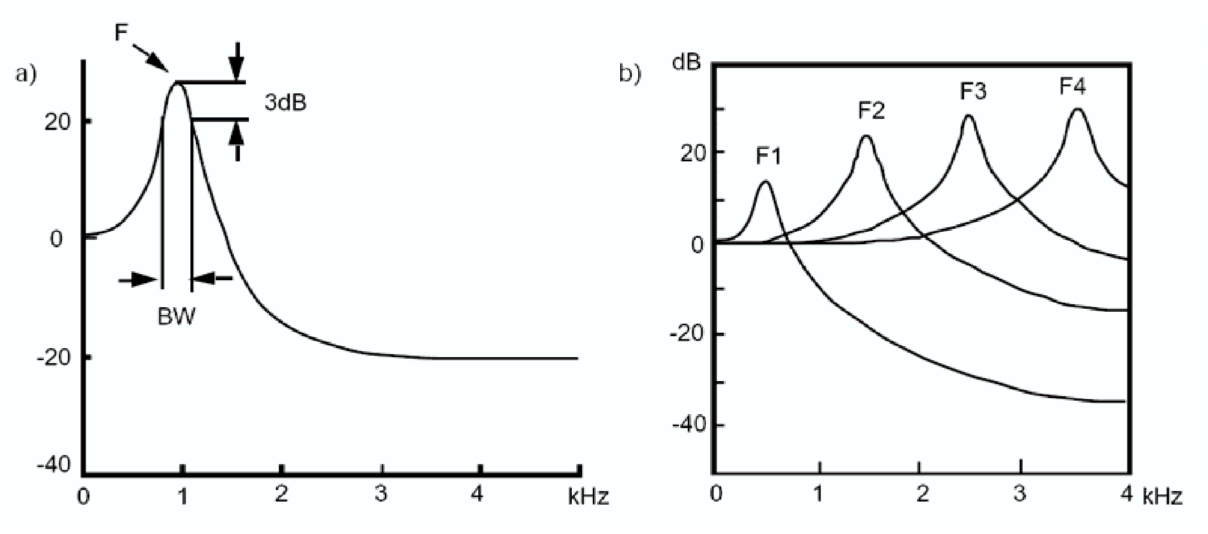
Vocal Tract Modelling (6/16)

- Vowels: Set of poles
- Nasalized Vowel: Set of zeros an poles



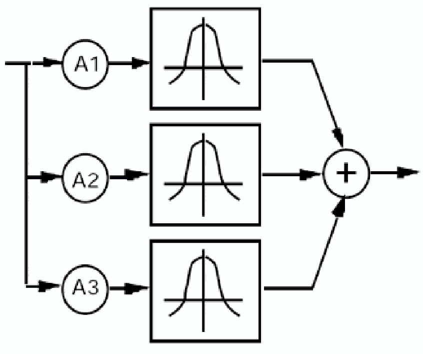
Vocal Tract Modelling (7/16)

- Formants modelled by bandpass–filter
- Anti–Formants modelled by bandstop–filter
- Parameters: Frequency and Bandwidth
- Spectrum modelled by superposition

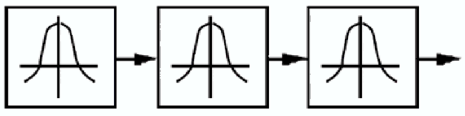


Vocal Tract Modelling (8/16)

- Parallel Configuration
  - Control of each formant amplitude
  - Convenient for consonant production
- Cascade Configuration
  - Direct replica of formant energy distribution
  - Convenient for producing vowels



a) Parallel configuration

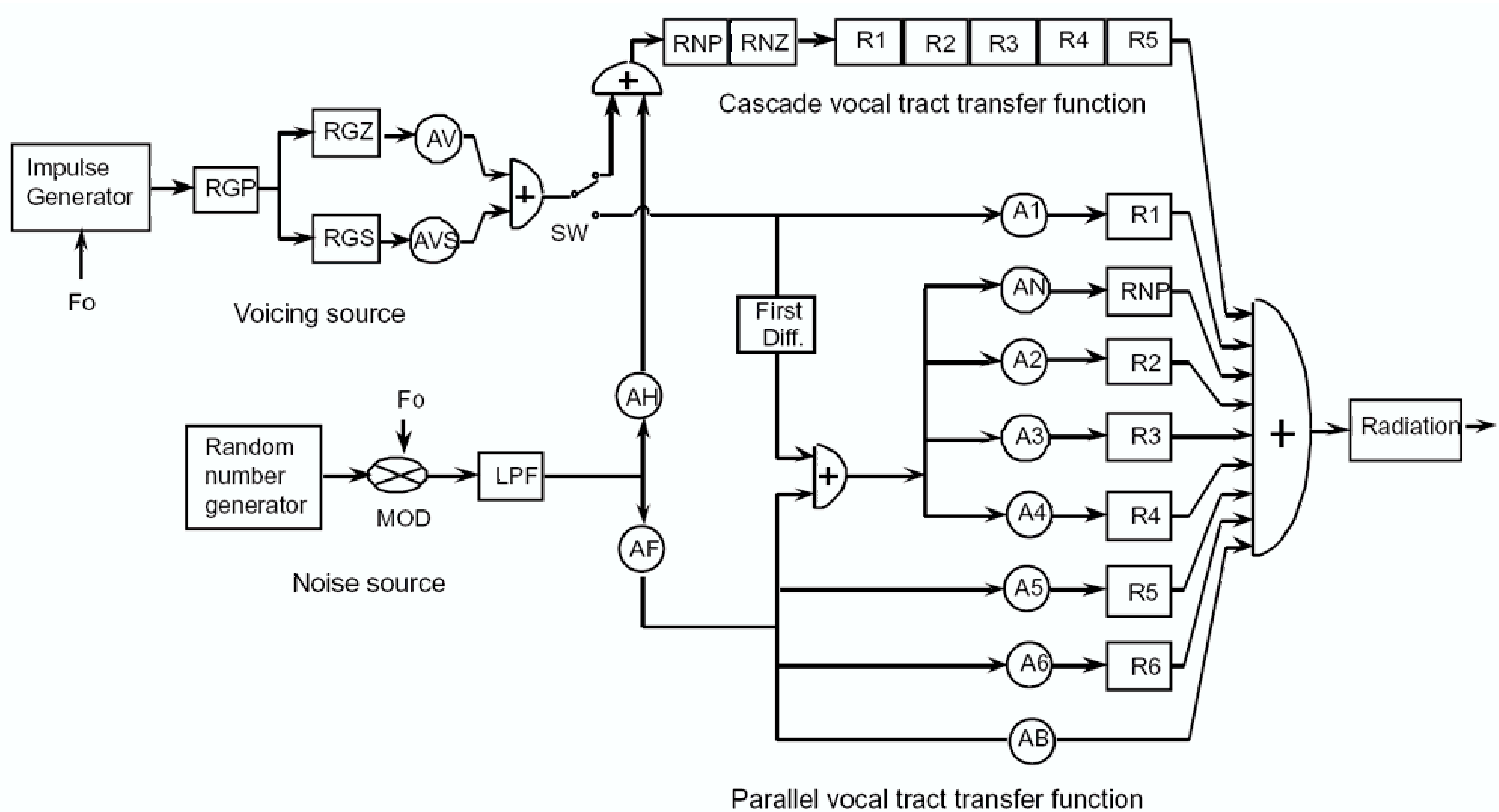

b) Cascade configuration

Klatt Formant Synthesizer (9/16)

Computer simulation of an electrical structure consisting of resonators in cascade/parallel

- Controlled by 40 parameters
- Male/Female Voice
- Two voicing sources
  - Vowels
  - Voiced fricatives
- Friction source

Klatt Formant Synthesizer (10/16)



Parameters (11/16)

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"

Parameters (12/16)

Symbol	C/V	Min.	Max.	Name
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)

Parameters (13/16)

Symbol	C/V	Min.	Max.	Name
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)

Parameters (14/16)

Symbol	C/V	Min.	Max.	Name
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)

Klatt Examples (15/16)

1. **Vowel**

```
TIME = 000; F1=450; F2=1450; F3=2450; F0=100; AV=72
TIME + 400; AV=0
```

2. **Syllable** (‘bah’)

```
TIME = 000; F1=400; F2=1000; F3=2000; F0=120; AV=72
TIME +  20; F1=650; F2=1200; F3=2500; AV=72
TIME +  20; F1=750; F2=1150; F3=2500; AV=72
TIME = 400; F1=750; F2=1000; F3=2300; F0=90; AV=72
TIME + 30; AV=0
```

Other parameters have been set to default values

Other Examples (16/16)

3. **Multivox**, TU Budapest
4. **Multipulse Linear Prediction**, Bishnu Atal, 1982
5. **DECtalk** male voice to make it sound female
6. **Female voice**, Dennis Klatt, 1986

Source: *Klatt's 'History of speech synthesis'*

Summary (1/2)

- Acoustic Tube Models
  - Simpliest Model: Lossless uniform tube
  - Complete model: Considering losses and nonuniformity of the area function
  - Tube Model: Transfer Function
- Linear Prediction
  - Optimum predictor coefficients assuming stationarity
  - Time–varying methods
  - Levinson–Durbin–Algorithm

Summary (2/2)

- Formant Synthesizer
  - Source/Tract Modelling
  - Klatt Synthesizer
  - Audio Examples

Literature(1/1)

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.