Digital Audio and Speech Processing (IN4182)

Speech Production

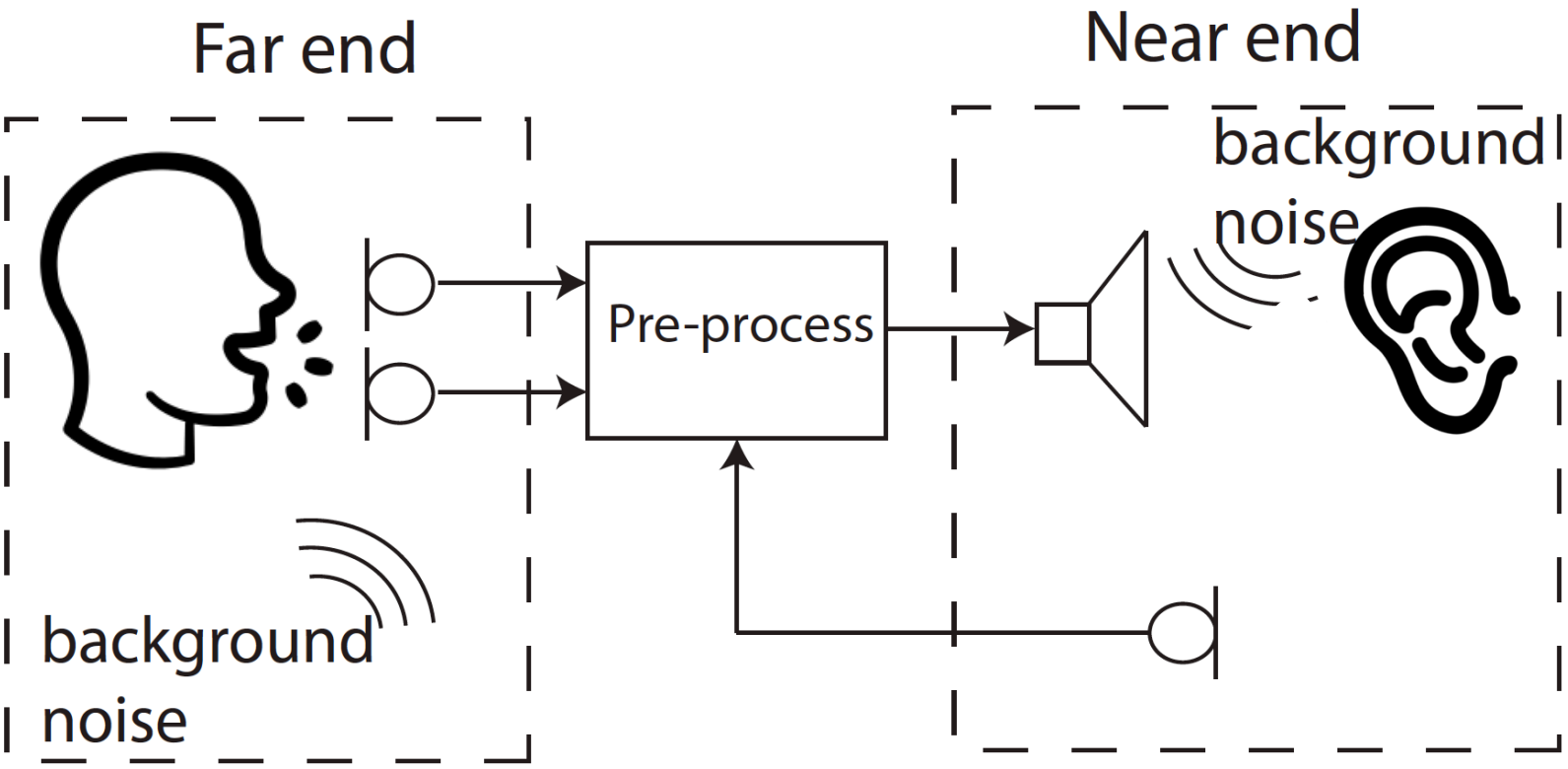
Richard C. Hendriks

04/05/20

TU Delft

Delft University of Technology

Prior Knowledge: Human Listener and Human Talker



TU Delft

Speech Production

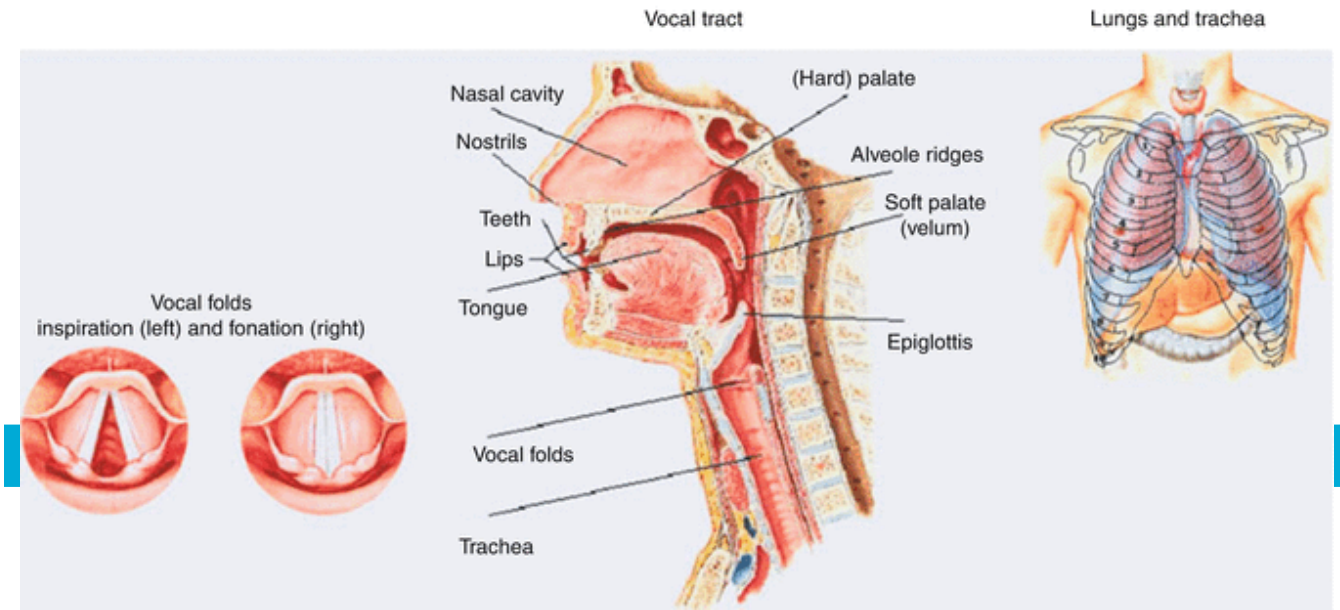
Outline:

- Speech Production – anatomy and physiology.
- The "looks and feel" of digital speech signals
  - Time domain.
  - Frequency domain.
  - Spectrograms.
- Speech production models
  - Acoustic modeling.
  - Discrete-time modeling.

Speech Production - Anatomy

Overview of speech production system:

- Lungs
- Larynx (organ of voice production).
- Vocal Tract
  - throat (pharyngeal cavity).
  - oral+nasal cavity.



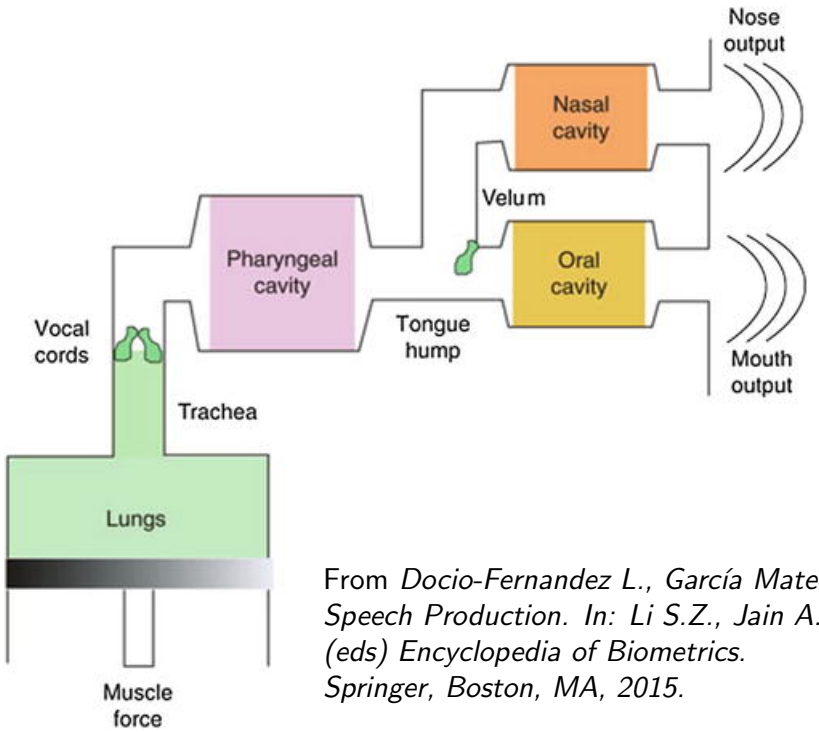
From *Docio-Fernandez L., García Mateo C. Speech Production. In: Li S.Z., Jain A.K. (eds) Encyclopedia of Biometrics. Springer, Boston, MA, 2015.*

TU Delft

Speech Production - Anatomy

Acoustic filter model:

- Lungs+vocal folds: Excitation.
- Cavities: Main acoustic filter.
- Velum: "switch" for nasal sounds.

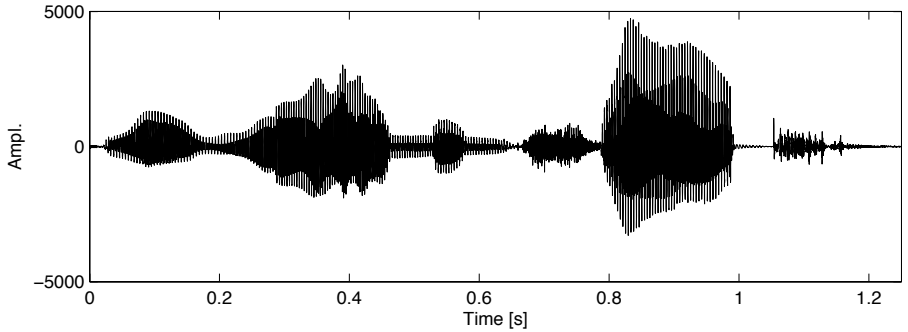


From *Docio-Fernandez L., García Mateo C. Speech Production. In: Li S.Z., Jain A.K. (eds) Encyclopedia of Biometrics. Springer, Boston, MA, 2015.*

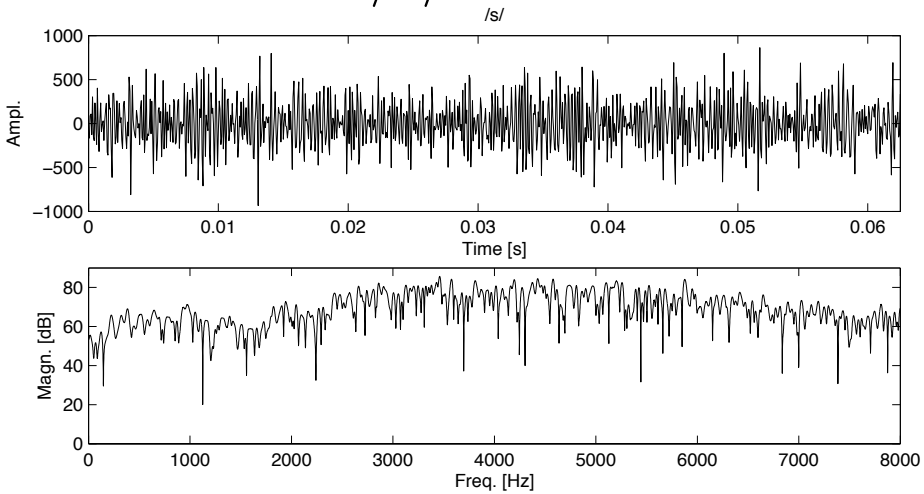
TU Delft

Speech Signals - A First Encounter

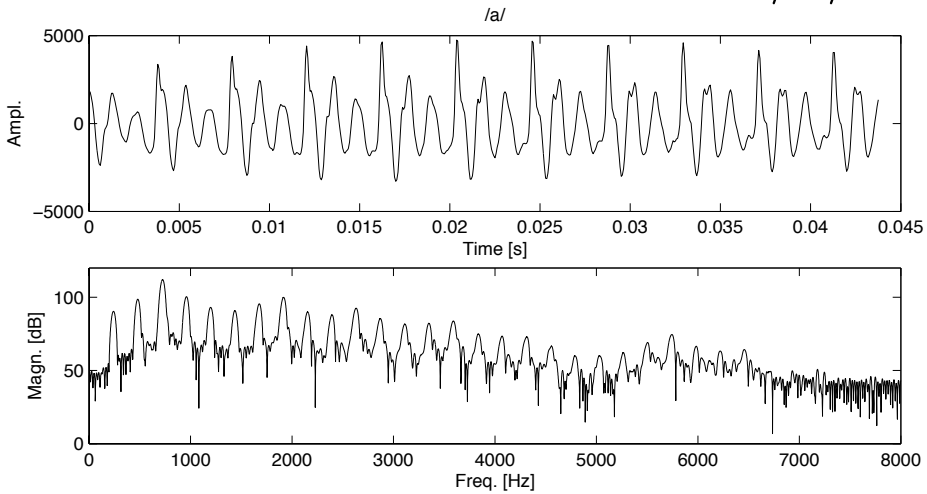
Characteristics of speech change across time due to changing production system:



uvoiced: /s/



voiced: /a/



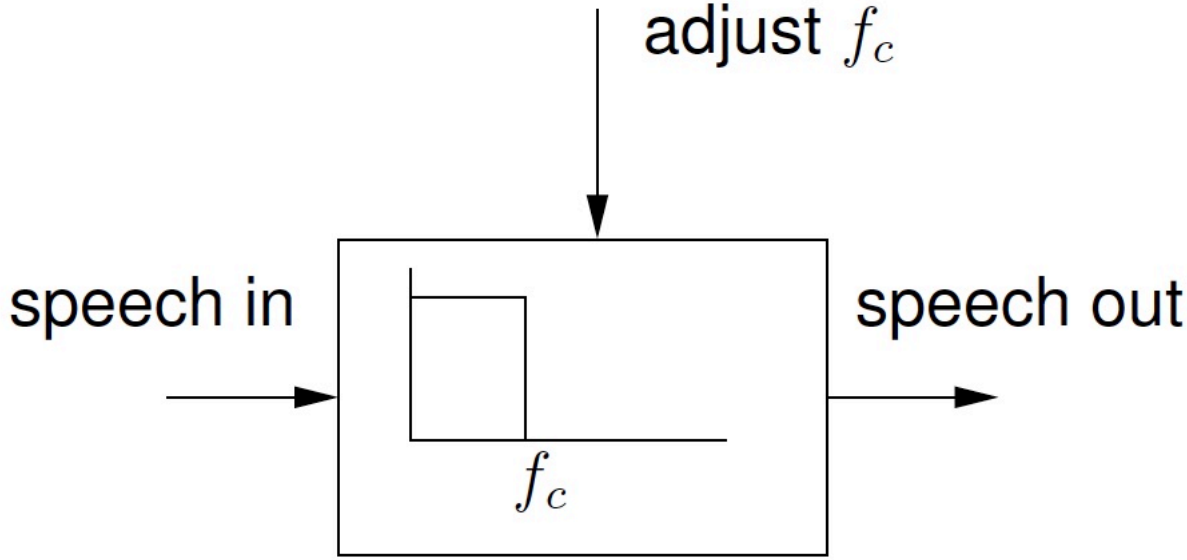
TU Delft

Speech Signals - A First Encounter

Speech bandwidth – Where is the information?

Cut-off frequency f_c:

- $f_c = 500$ Hz.
- $f_c = 1000$ Hz.
- $f_c = 1500$ Hz.
- $f_c = 2000$ Hz.
- $f_c = 3000$ Hz.
- $f_c = 4000$ Hz.
- $f_c = 8000$ Hz.



$s = x * (-1)$? s

Can you hear the difference? x

TU Delft

Speech Production - Excitation

Excitation signal: The air stream signal that enters the paryngeal cavity (throat), i.e., after vocal folds.

Types of excitation:

- *Voiced*: Air pushed through glottis which oscillate, generating quasi-periodic puffs of air (e.g. vowels /a/, /i/, etc.).
- *Unvoiced*: Air forced through constriction somewhere along vocal tract (e.g. /s/, /f/).
- *Mixed*: Quasi-periodic excitation but with constriction along vocal tract (e.g. /z/).
- *Plosive*: Complete closure of vocal tract, build-up of air pressure + release (e.g. /p/, /t/).

Speech Production – Excitation Signal

Voicing:

(a) Vocal chords closed, complete constriction of windpipe.

(b) Air pushes from lungs, building up pressure below vocal chords.

(c) Vocal chords (elastic muscle tissue) cannot withstand air pressure and starts to open.

(d) Air flows through glottis.

(e)-(f) Air pressure below glottis is low, vocal chords begin to close.

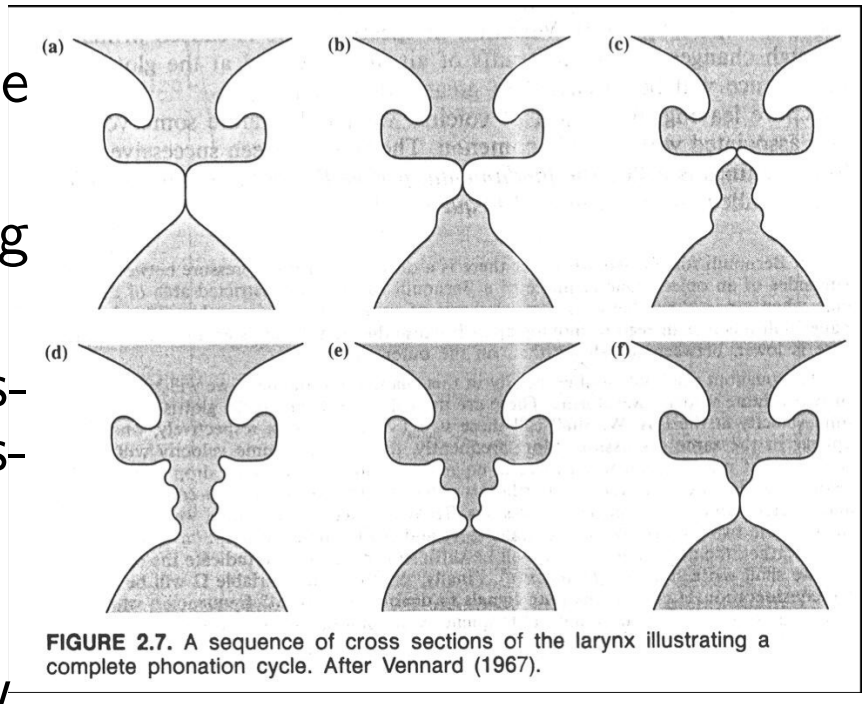


FIGURE 2.7. A sequence of cross sections of the larynx illustrating a complete phonation cycle. After Vennard (1967).

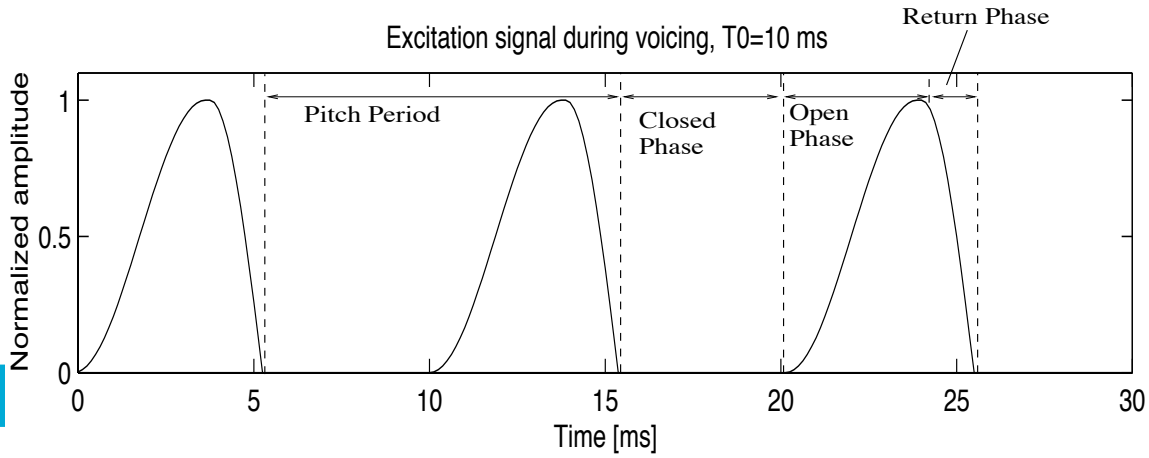
After *Deller, Hansen, and Proakis: "Discrete-Time Processing of Speech Signals", p. 111, IEEE Press, 2000.*

TU Delft

Speech Production - Excitation Signal

Voicing:

- *Fundamental period* T_0: Time interval between successive vocal fold openings.
- *Fundamental frequency* F_0: Rate of vocal chord vibration ($F_0 = 1/T_0$).
- *Pitch*: Perceived fundamental frequency (whether or not present in waveform).



TU Delft

Speech Production - Excitation Signal

Voicing:
Fundamental period/frequency dependent on size and tension of speaker's vocal folds.

- Men: F0 between 50–250 Hz.
- Women: F0 between 120–500 Hz.
- Children: sometimes even higher

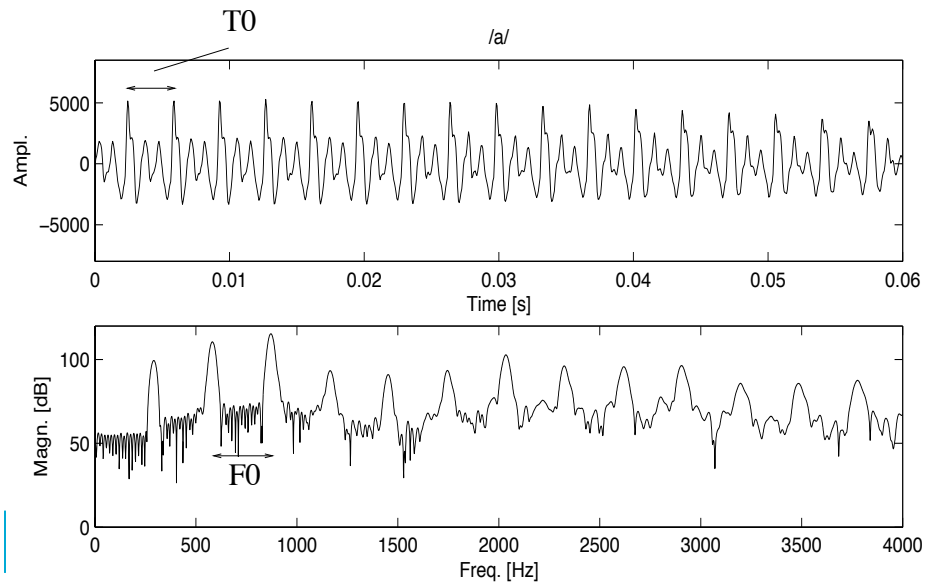
F0 related to stress, emotion, intonation.

Speech Production - Excitation Signal

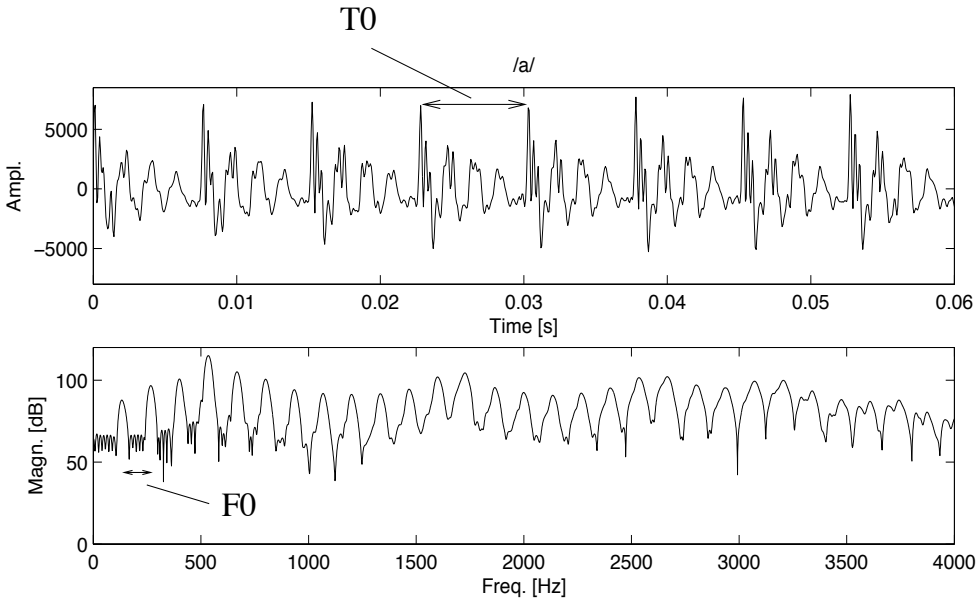
Voicing:

The fundamental period/frequency is evident in the time domain as well as the frequency domain representations of speech.

Female speaker



Male speaker

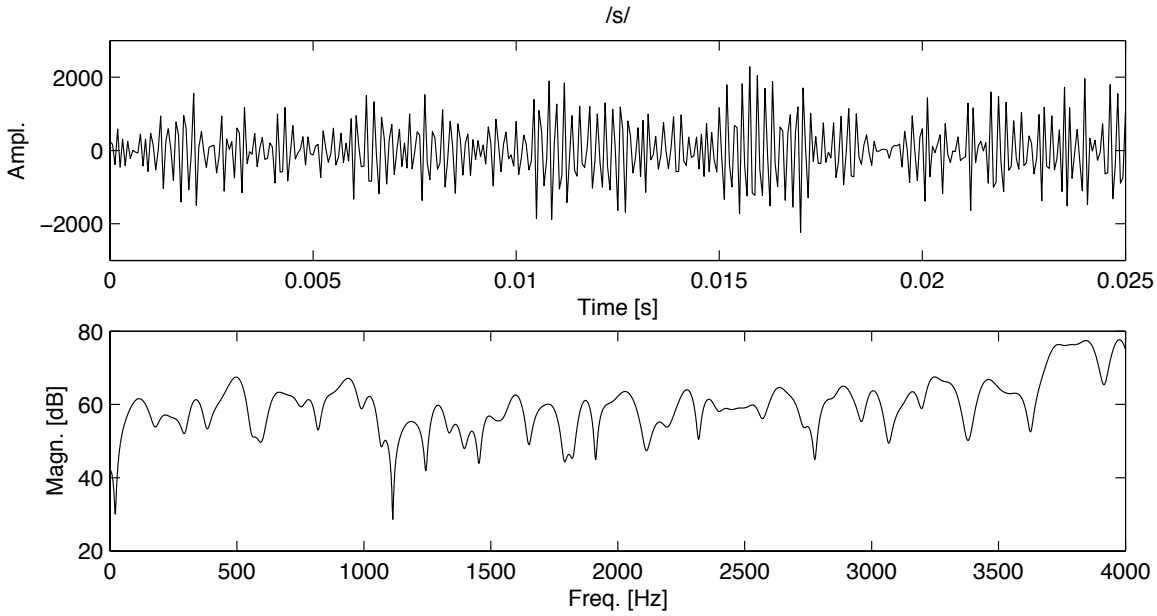


TU Delft

Speech Production - Excitation Signal

Unvoiced regions:

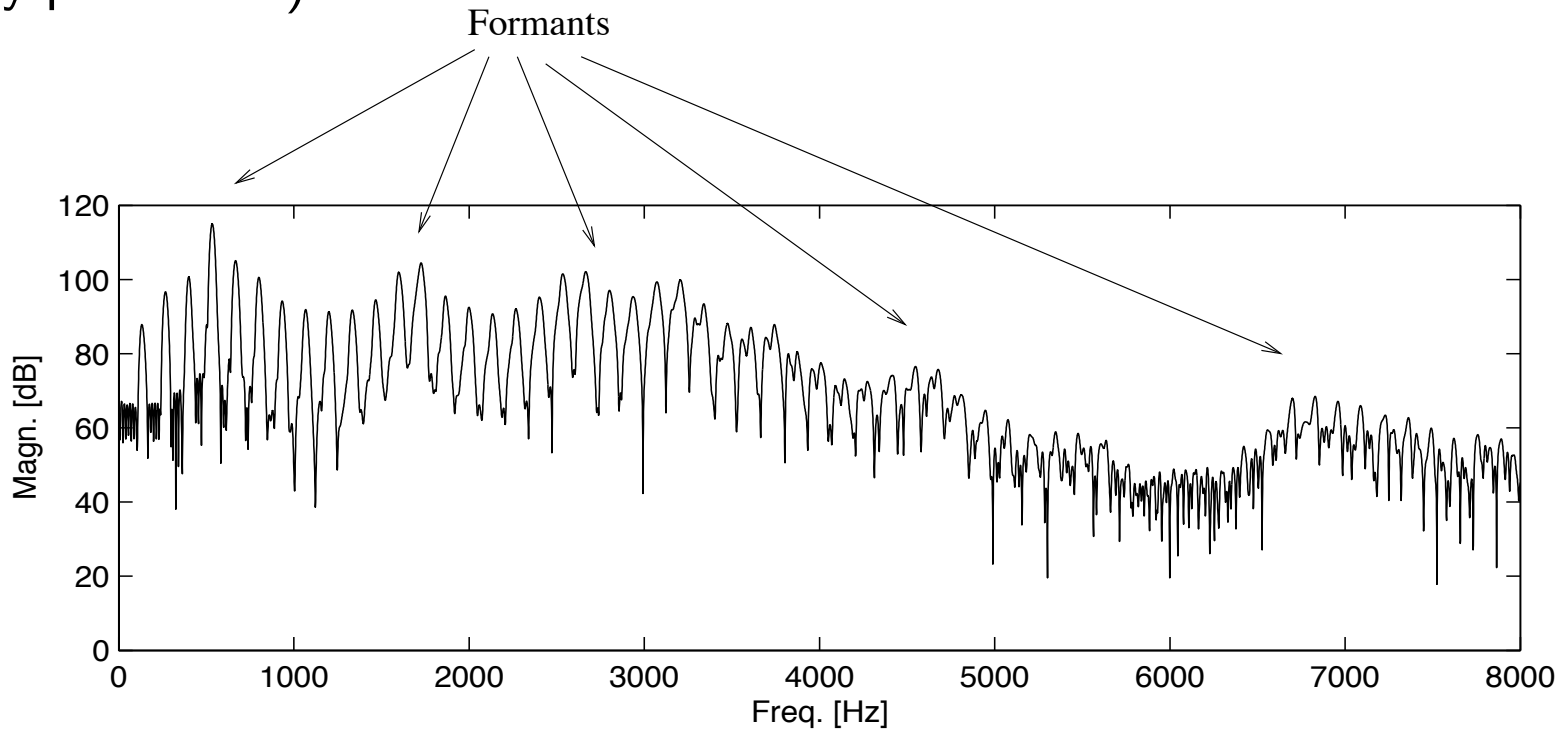
In unvoiced regions, the excitation signal is noise-like (i.e., without the periodicity that characterizes voiced signals.)



TU Delft

Speech Production - The Vocal Tract

- Configuration of vocal tract "shapes" excitation to generate specific speech sound, i.e., overall spectral characteristic determined by vocal tract.
- Resonance frequencies of vocal tract system give rise to peaks in overall spectrum $\sim$ *formants* (3-5 formants within Nyquist band).



Speech Production - The Vocal Tract

- Vocal tract system changes over time $\Rightarrow$ spectral/temporal characteristics of the speech waveform are *time-varying* $\Rightarrow$ only short segments of speech waveform can be assumed to have similar acoustic properties (*"non-stationarity" vs "short-term stationarity"*).
- Speech signals typically assumed stationary over 20-30 ms time frames.
- Typical maximum speech bandwidth 7-8 kHz.

Speech Production - The Vocal Tract

Observation:

- Articulators have mass (tissue/muscles, etc.) and can therefore not move instantly.

Consequences:

- Speech is not string of discrete sound events, but rather output of continuous and relatively slowly varying system.
- Coarticulation: Neigboring speech sounds colour each other.

Categorization of Speech

Phonemes:

- The basic theoretical unit for describing how speech conveys linguistic meaning.
- Each phoneme represents a speech sound class that differentiates words of a language. For example, "cat", "bat", "hat" consists of three speech sounds, the first of which gives word distinctive meaning, i.e., /c/, /b/, and /h/ are different phonemes.
- ≈ 45 phonemes (American English).

Categorization of Speech

Phoneme classes (American English):

- vowels.
- nasals, e.g. /m/ in "mama", /n/ in "no", /G/ in "sing".
- fricatives, e.g., /s/ in "sand", /f/ in "fan".
- plosives, e.g., /t/ in "top", /b/ in "bob".
- dipthongs, e.g., /w/ in "we", /W/ in "out".

Characterizing vowels:

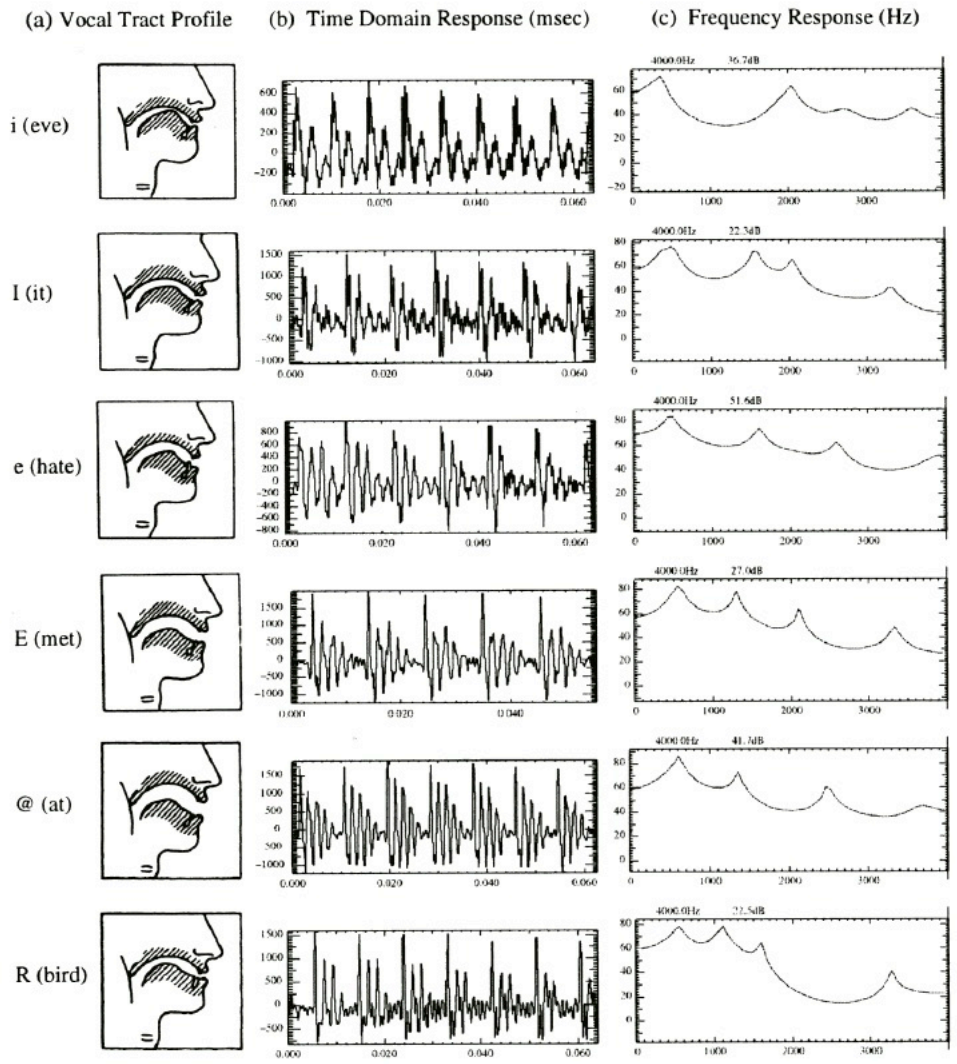


FIGURE 2.10. A collection of features for vowels in American English. Column (a) represents schematic vocal-tract profiles, (b) typical acoustic waveforms, and (c) the corresponding vocal-tract magnitude spectrum for each vowel.

From *Deller, Hansen, and Proakis: "Discrete-Time Processing of Speech Signals", IEEE Press, 2000.*

TU Delft

Characterizing vowels:

Formant locations for vowels:

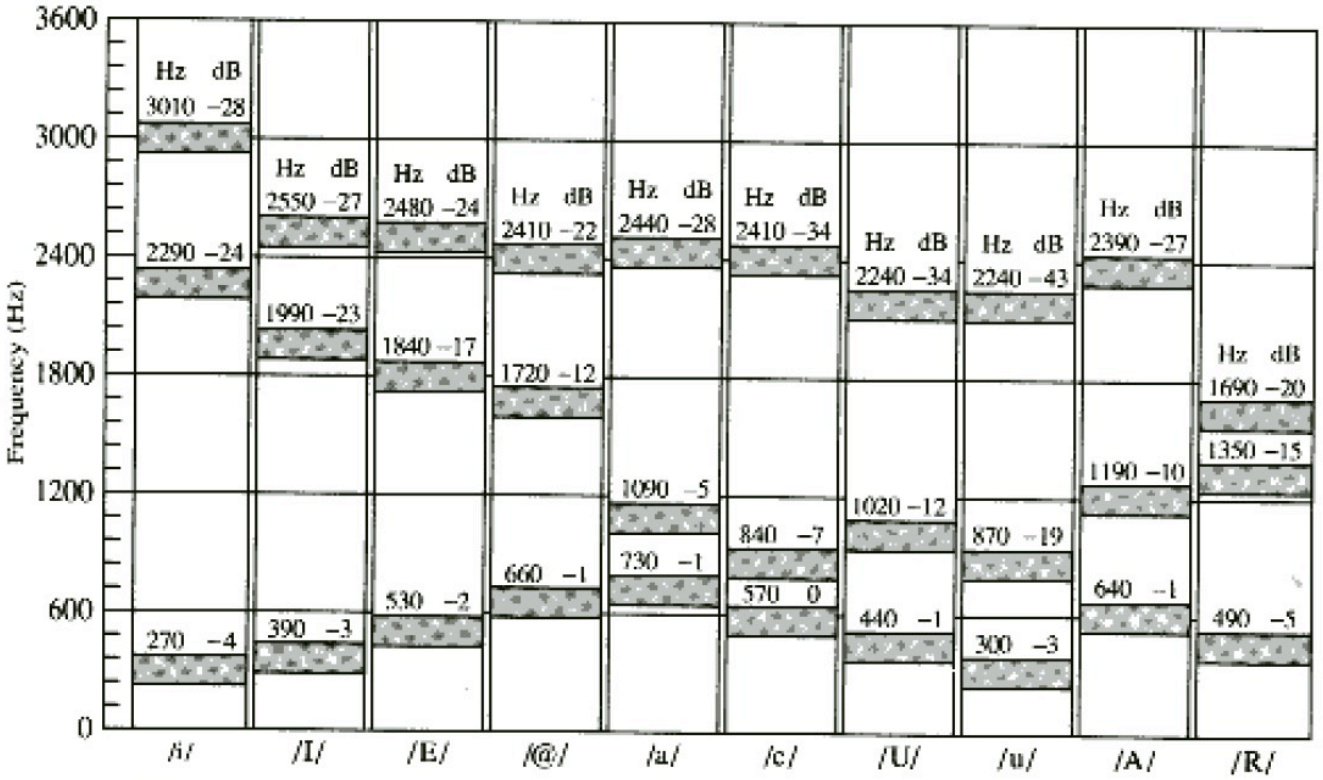


FIGURE 2.11. Average formant locations for vowels in American English (Peterson and Barney, 1952).

From *Deller, Hansen, and Proakis: "Discrete-Time Processing of Speech Signals", IEEE Press, 2000.*

TU Delft

Characterizing vowels:

Formant locations for vowels:

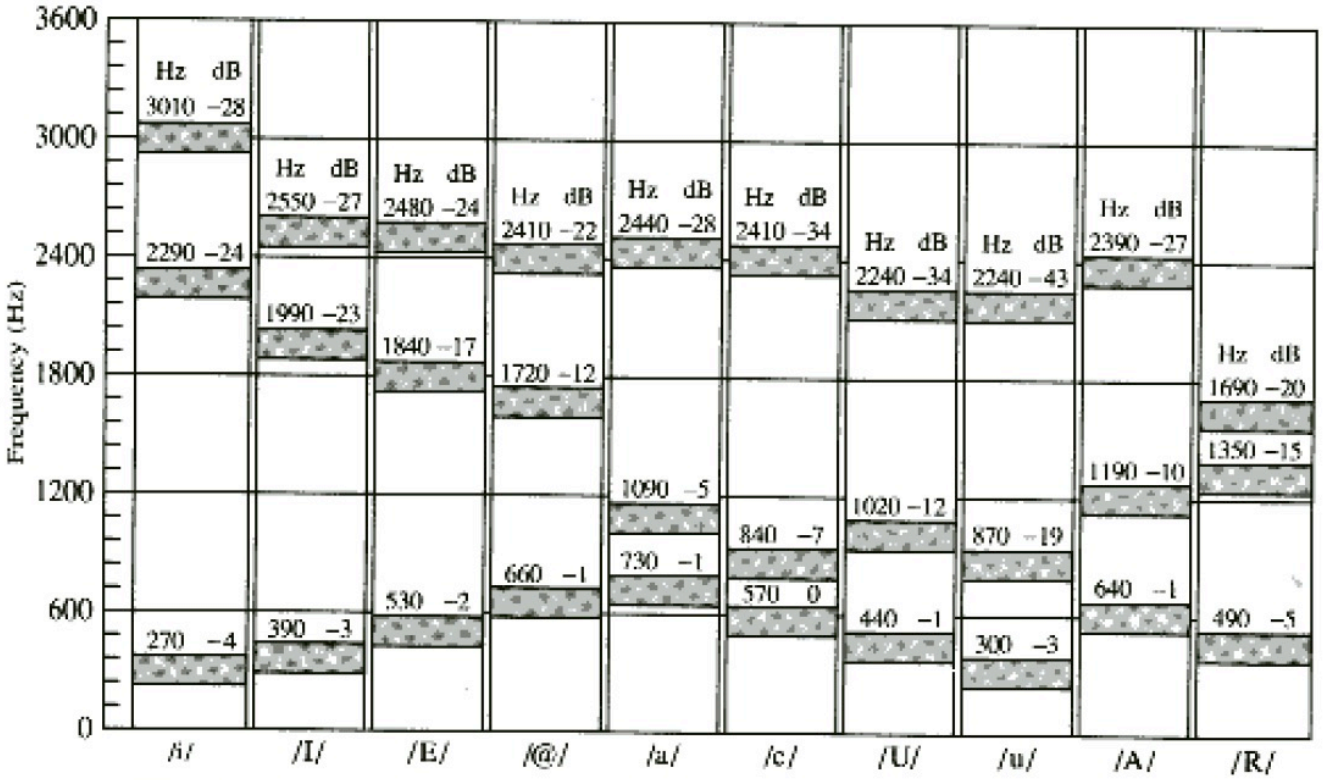


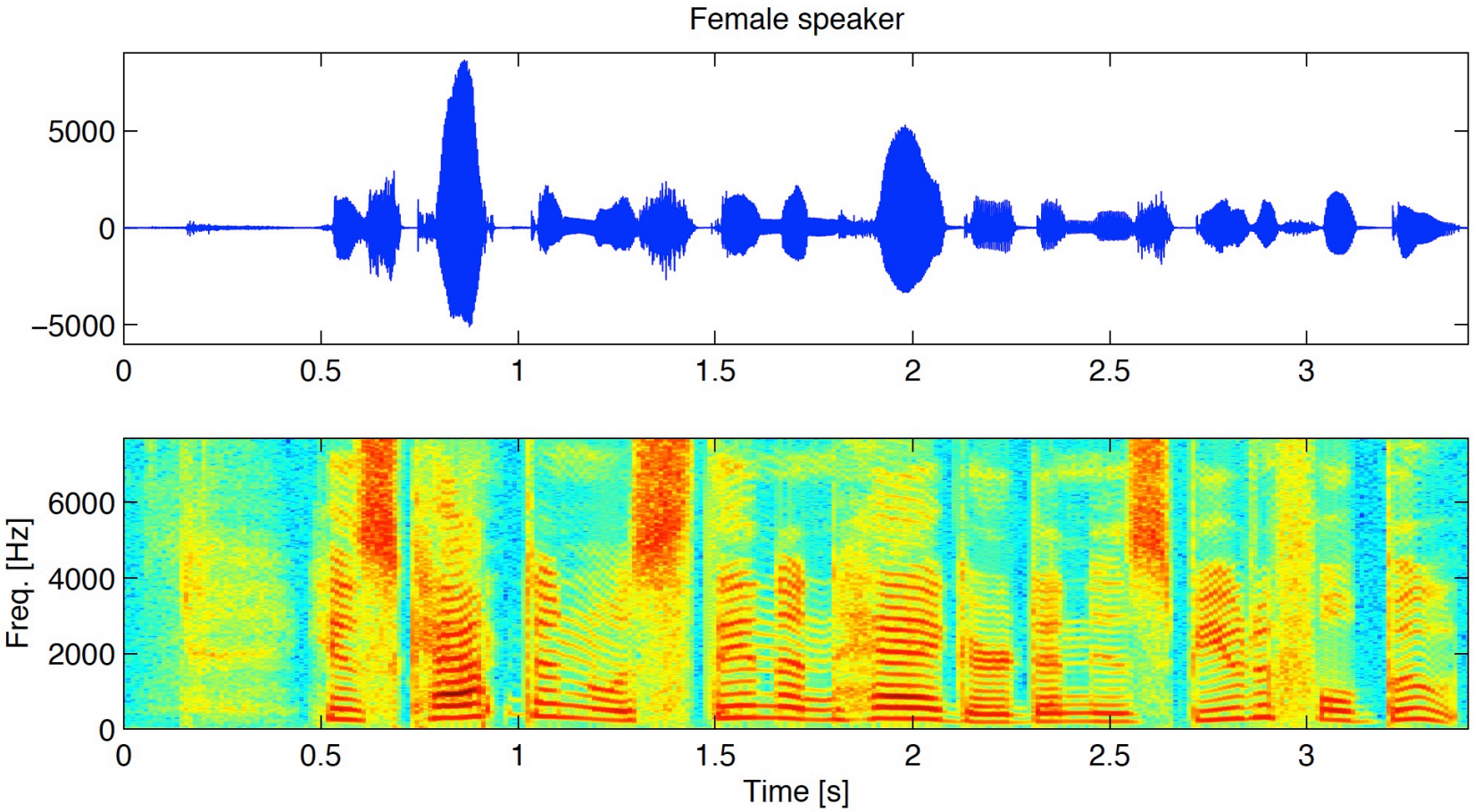
FIGURE 2.11. Average formant locations for vowels in American English (Peterson and Barney, 1952).

From *Deller, Hansen, and Proakis: "Discrete-Time Processing of Speech Signals", IEEE Press, 2000.*

TU Delft

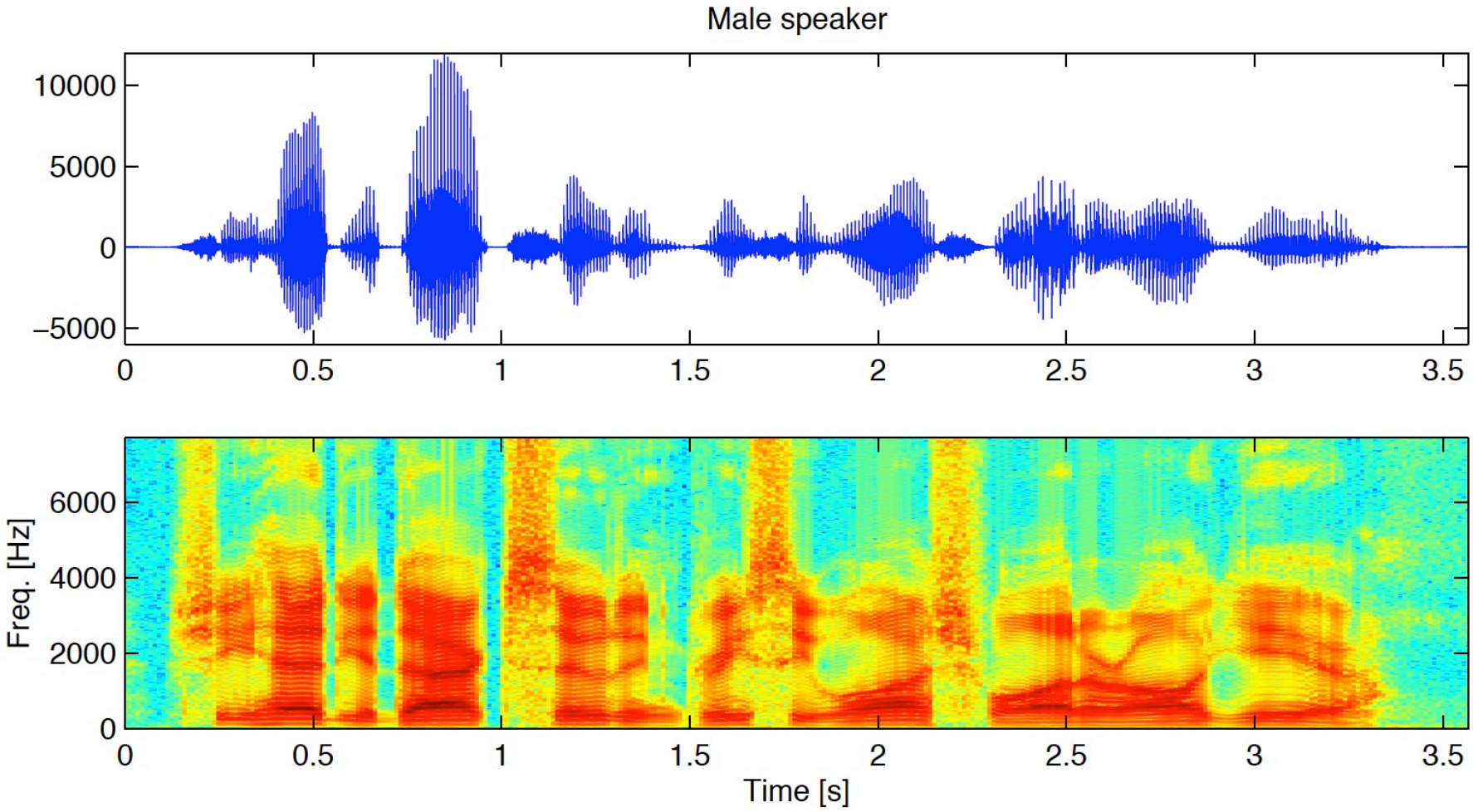
Categorization of Speech

Spectrogram: Time-vs-Freq-vs-Spectral Magnitude (no phase!).
"His captain was thin and haggard and his beautiful boots..."



TU Delft

Categorization of Speech



TU Delft

Speech Production - General Model

General acoustical (short-time) model:

$$S(f) = U(f)H(f)R(f),$$

where

$S(f)$ is Fourier transform of speech signal.

$U(f)$ is Fourier transform of source excitation.

$H(f)$ is vocal tract transfer function.

$R(f)$ describes radiation effects at lips.

Speech Production - General Model

General acoustical (short-time) model:

$$S(f) = U(f)H(f)R(f),$$

Model assumptions:

- $S(f)$ constant with time (model valid for short time duration).
- $U(f), H(f)$, and $R(f)$ are linear.
- $U(f), H(f)$, and $R(f)$ are separable (no coupling between sub-systems).

Our goal: Discrete-time model:

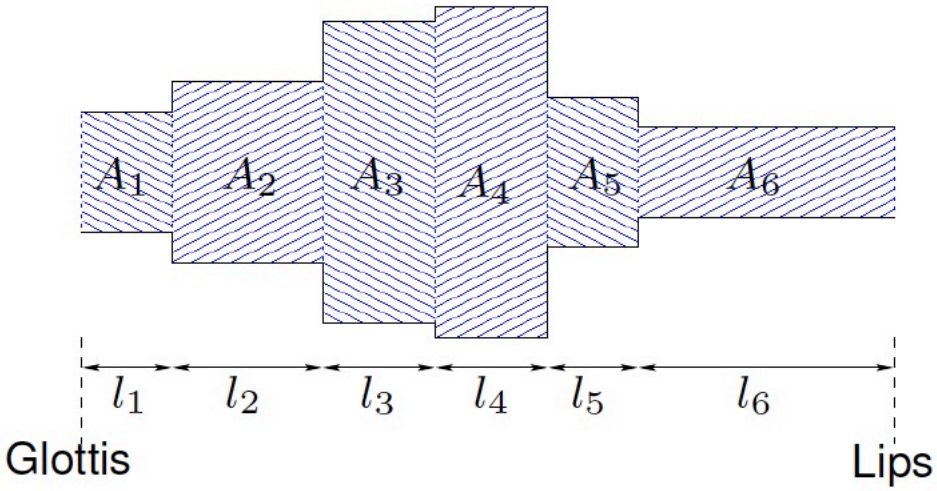
$$S(z) = \theta_0 U(z)H(z)R(z), \qquad \theta_0 \text{ is gain factor.}$$

Speech Production - Vocal Tract

Multitube lossless (time-continuous) model of vocal tract $H(f)$:

General assumptions:

- Vocal tract is series of concatenated acoustic tubes.
- Tubes are hardwalled (no wall friction, no wall vibration), and lossless (no heat conduction).
- All air-particles within same cross-sectional area A_k have same velocity.

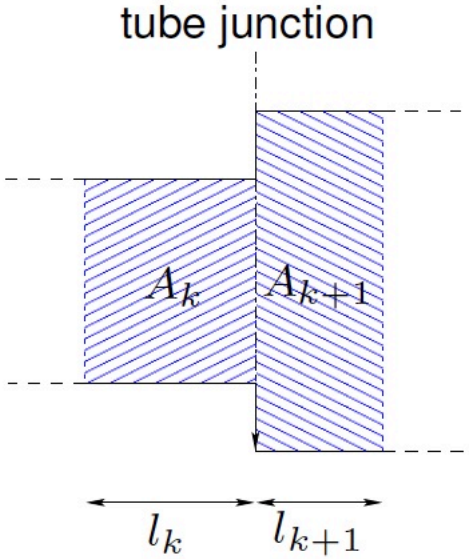


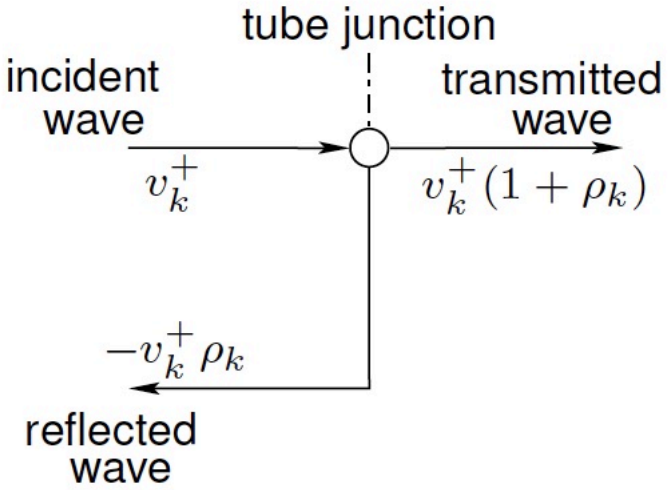
TU Delft

Speech Production - Vocal Tract

Reflection at tube junctions:

- Reflection coefficient: $\rho_k = \frac{A_{k+1} - A_k}{A_{k+1} + A_k}$
- Value of ρ_k depends on area mismatch
  - $A_k \approx A_{k+1} => \rho_k \approx 0$
  - $A_k \gg A_{k+1} => \rho_k \approx -1$
  - $A_k \ll A_{k+1} => \rho_k \approx 1$





Speech Production - Vocal Tract

Signal flow graph (2-tube model):

Model consists of additions, multiplications and delays ⇒ Filtering operations!

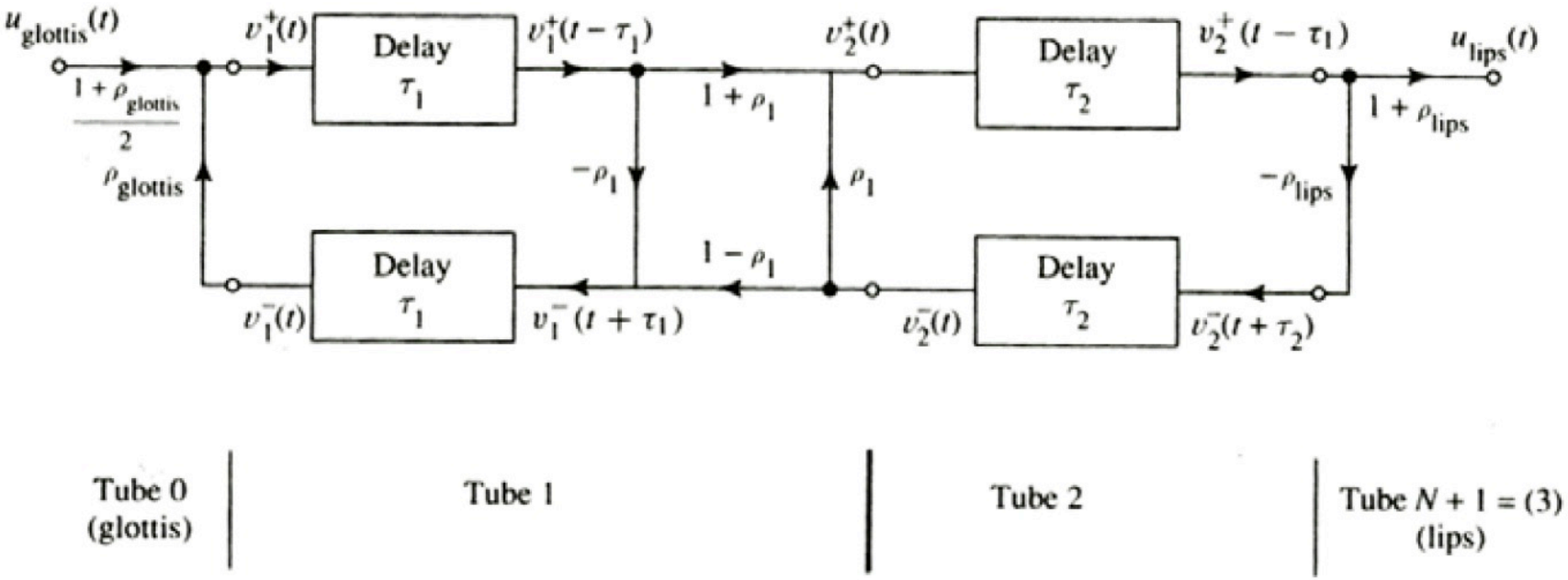


FIGURE 3.22. Overall signal flow diagram for a two-tube acoustic model of speech production.

From *Deller, Hansen, and Proakis: "Discrete-Time Processing of Speech Signals", IEEE Press, 2000.*

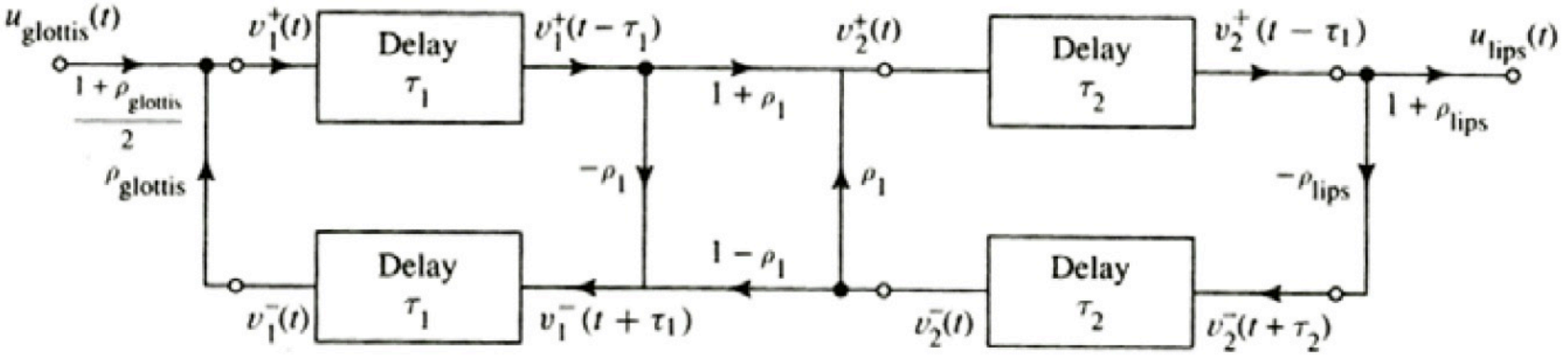
TU Delft

Speech Production - Vocal Tract

Signal flow graph (2-tube model):

Model consists of additions, multiplications and delays $\Rightarrow$ Filtering operations!

From *Deller, Hansen, and Proakis: "Discrete-Time Processing of Speech Signals", IEEE Press, 2000.*



Two tube section (with glottis input, lips output):

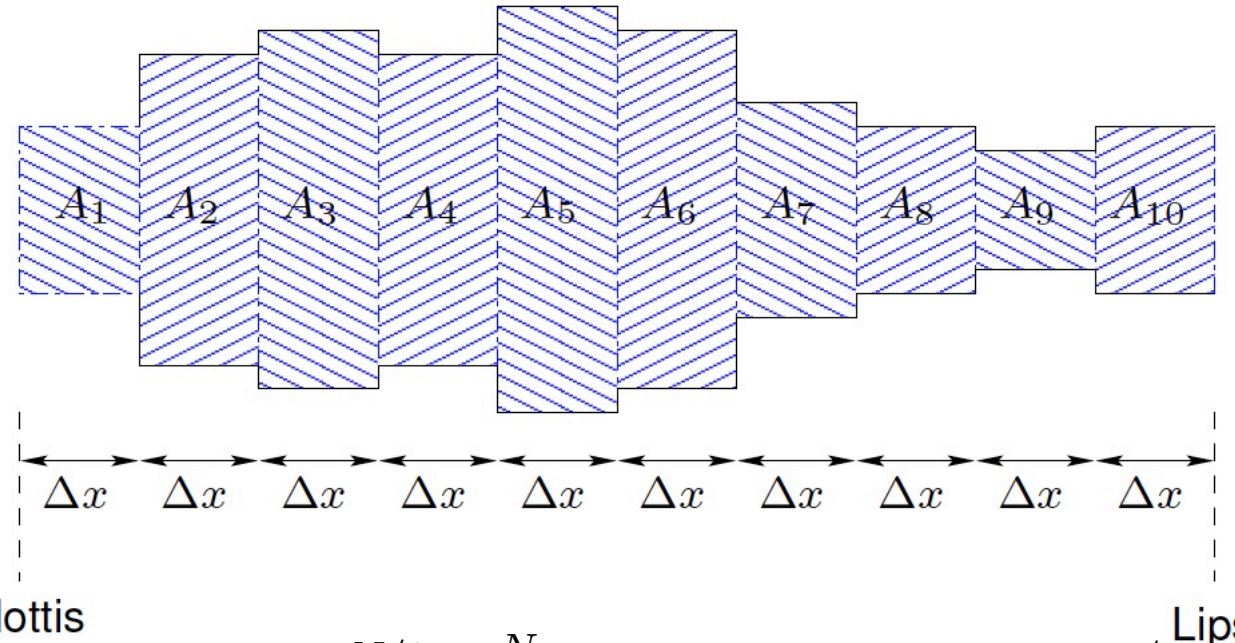
- $H_{2tube}(z) = \dfrac{U_{lips}(z)}{U_{glottis}(z)} = \dfrac{\frac{1+\rho_{glot}}{2} z^{-2/2}(1+\rho_1)(1+\rho_{lips})}{1+(\rho_1\rho_{glottis}+\rho_1\rho_{lips})z^{-1}+\rho_{glottis}\rho_{lips}z^{-2}} =$
  $\dfrac{Gz^{-1}}{1-\sum_{k=1}^{2} b_k z^{-k}}$ with $G = \frac{(1+\rho_{glot})(1+\rho_1)}{2(1+\rho_{lips})}$,
  $b_1 = -(\rho_1\rho_{glottis}+\rho_1\rho_{lips})$ and $b_2 = -\rho_{glottis}\rho_{lips}$.

TU Delft

Speech Production - Vocal Tract

Discrete-time model $H(z)$:

- Assume $l_1 = l_2 = \cdots = l_N = \Delta x.$



- $H(z) = \dfrac{1+\rho_{glot}}{2} \dfrac{z^{-N/2}\prod_{k=1}^{N}(1+\rho_k)}{1-\sum_{k=1}^{N} b_k z^{-k}} = \dfrac{G z^{-N/2}}{1-\sum_{k=1}^{N} b_k z^{-k}}.$

TU Delft

Speech Production - Vocal Tract

We note

- $H(z)$ has N poles.
- $H(z)$ has $N/2$ zeros at $z = 0$.

The vocal tract discrete-time model used in practice is the following all-pole simplification:

$$H(z) = \frac{G}{1 - \sum_{k=1}^{N} b_k z^{-k}} = \frac{G}{\prod_{k=1}^{N} (1 - p_k z^{-1})},$$

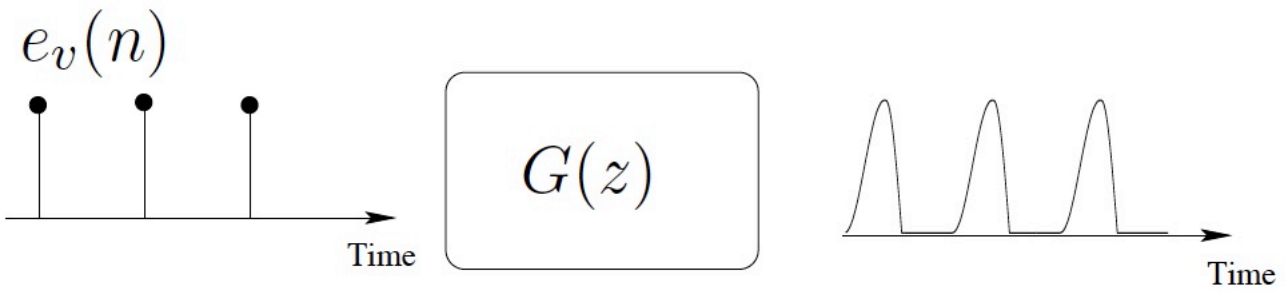
where p_k denote (complex-valued) pole locations.

Speech Production – U(z) and R(z)

- Voiced excitation:

$$S_v(z) = \theta_0 \overbrace{E_v(z)G(z)}^{U_v(z)} H(z)R(z),$$

where $e_v(n)$ is discrete-time impulse train and $G(z)$ is glottal shaping filter (usually assumed to be low-order, all-pole).



TU Delft

Speech Production – U(z) and R(z)

- Unvoiced excitation:

$$S_u(z) = \theta_0 \overbrace{E_u(z)}^{U_u(z)} H(z) R(z),$$

  where $e_u(n)$ is normally modeled using white noise sequence.

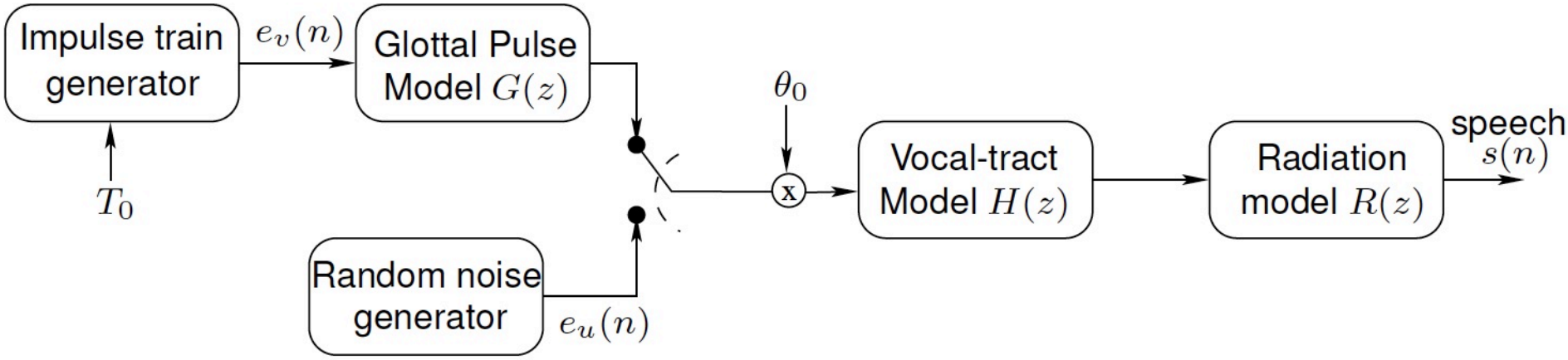
- Lip radiation $R(z)$ is often modeled as:

$$R(z) = 1 - z_0 z^{-1}, \qquad z_0 \approx 1 \ (z_0 < 1)$$

  It can be shown that

$$R(z) = \frac{1}{\prod_{k=0}^{\infty} z_0^k z^{-k}} \approx \frac{1}{\prod_{k=0}^{L} (1 - b_k z^{-k})} \text{ for some finite } L.$$

Speech Production - Source-Filter Model

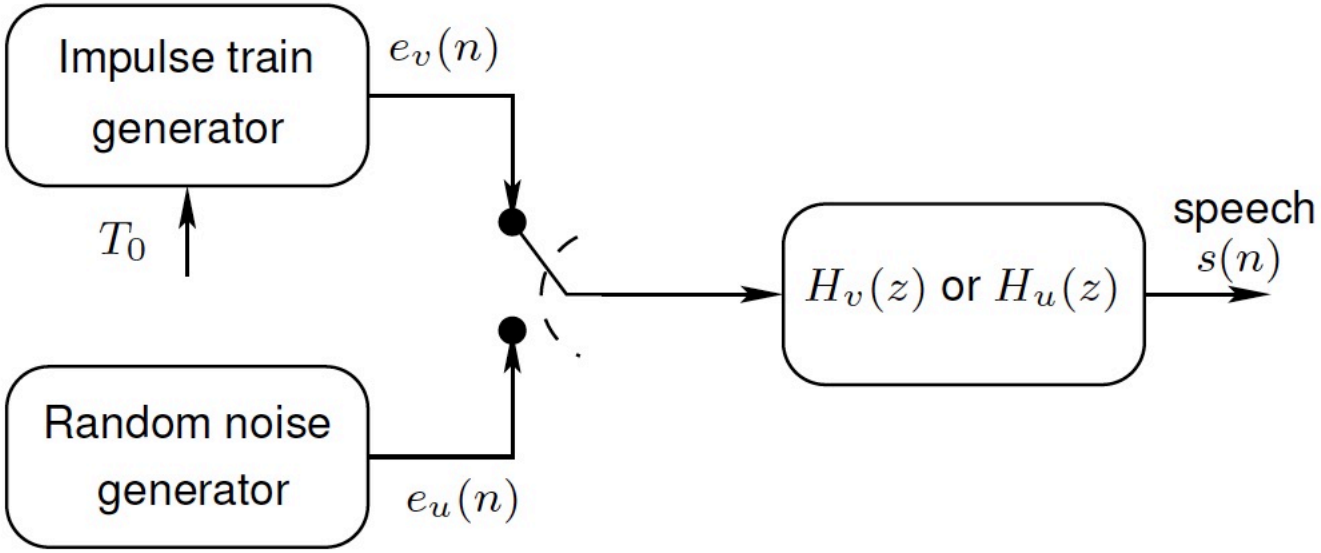


$$S(z) = \begin{cases} \theta_0 E_v(z) G(z) H(z) R(z) & \text{if voiced} \\ \theta_0 E_u(z) H(z) R(z) & \text{if unvoiced} \end{cases}$$

We have argued that $H(z)$, $G(z)$, and $R(z)$ are/can be approximated as all-pole.

TU Delft

Speech Production - Source-Filter Model

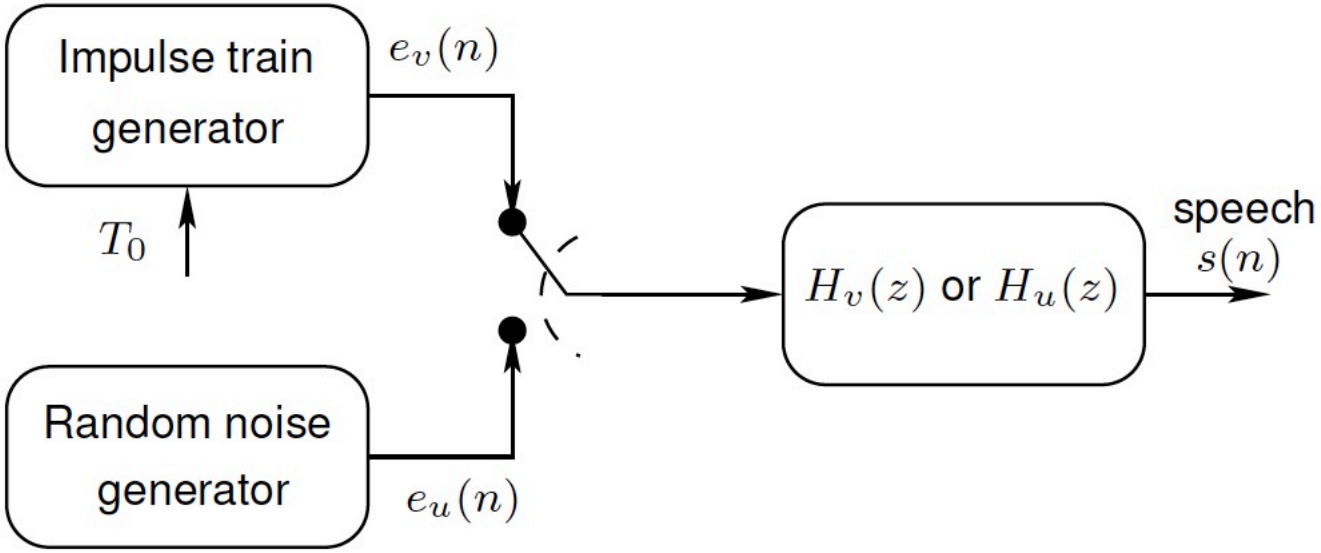


$$S(z) = \begin{cases} E_v(z)H_v(z) & \text{if voiced} \\ E_u(z)H_u(z) & \text{if unvoiced} \end{cases}$$

where $H_v(z) = \frac{\theta_v}{\prod_{k=0}^{P_v}(1-p_k z^{-k})}$ and $H_u(z) = \frac{\theta_u}{\prod_{k=0}^{P_u}(1-p_k z^{-k})}$.

TU Delft

Speech Production - Examples

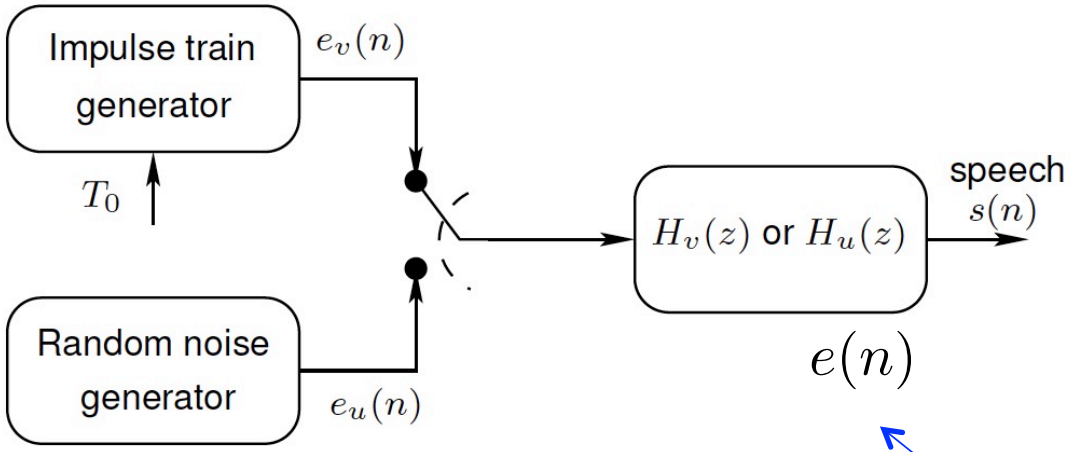


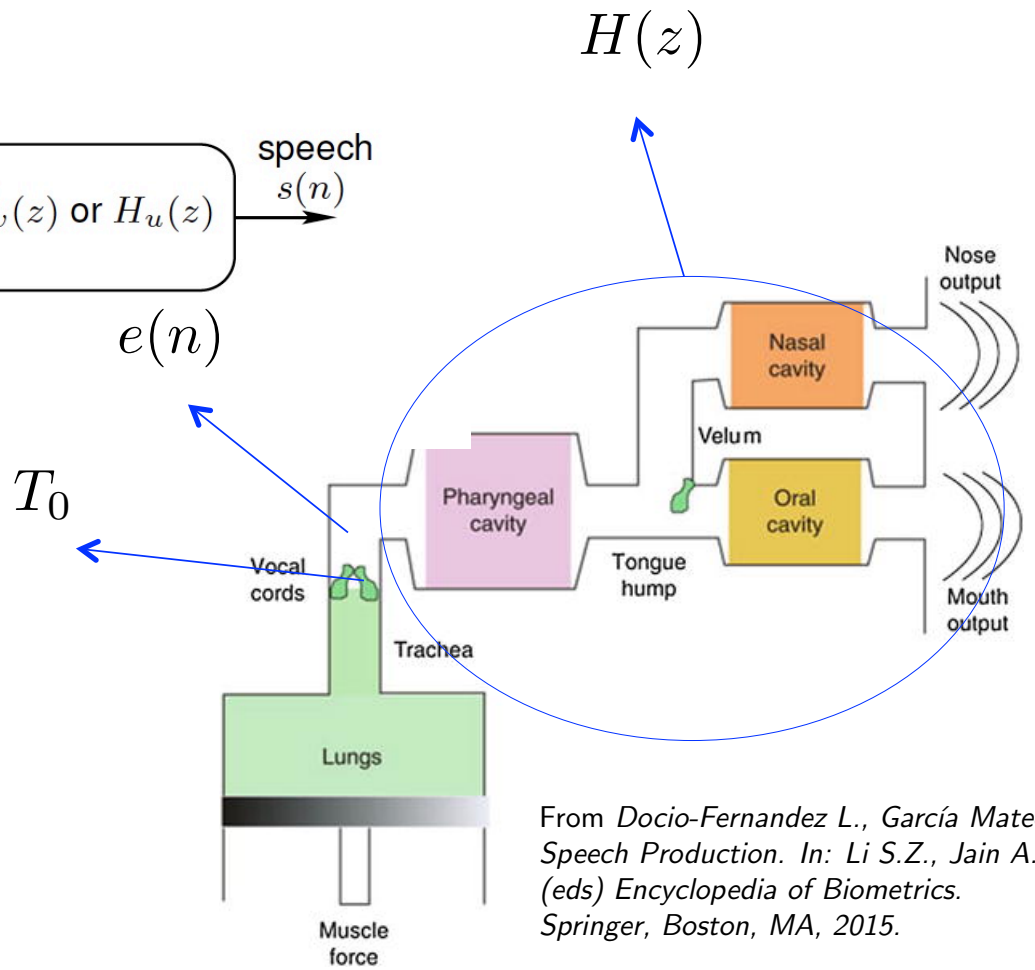
Example: "The birch canoe slid on the smooth planks"

- Original speech.
- Modelled speech.
- Modelled speech assuming only unvoiced speech.

TU Delft

Speech Production - Source-Filter Model

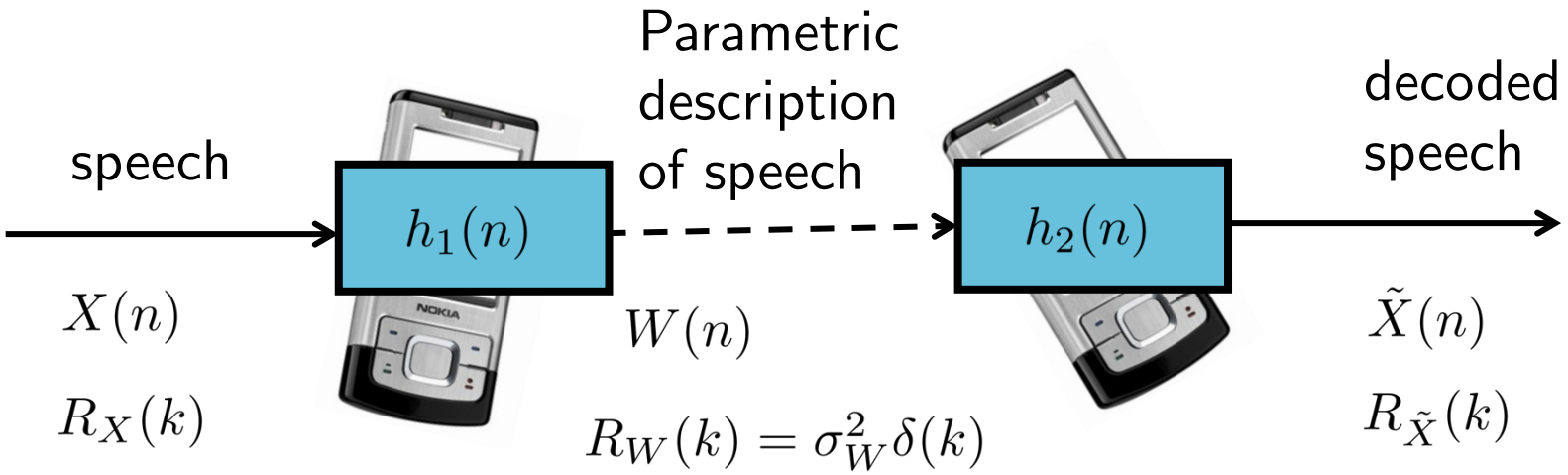




From *Docio-Fernandez L., García Mateo C. Speech Production. In: Li S.Z., Jain A.K. (eds) Encyclopedia of Biometrics. Springer, Boston, MA, 2015.*

TU Delft

Speech Coding - Example



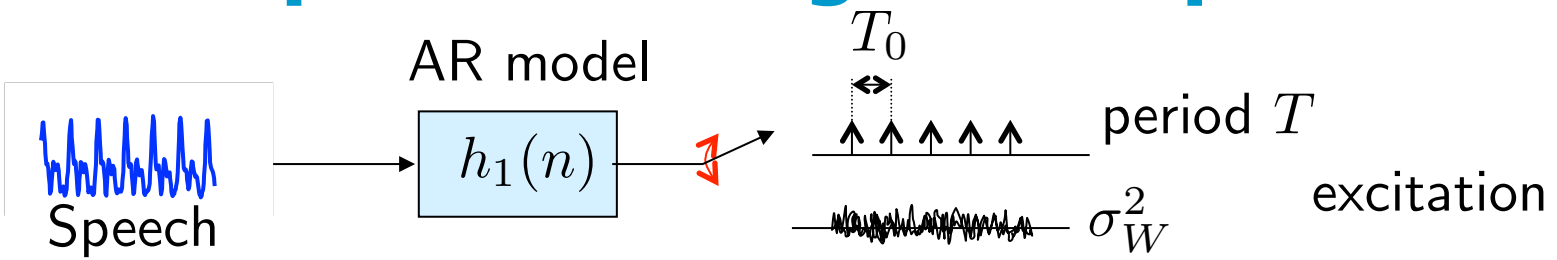
What to transmit?

The statistical description of the speech process.

Choose $h_1(n)$ such that the output is uncorrelated with minimum variance.

$W(n) = h_1(n) * X(n)$
Then transmit $h_1(n)$ (= inverse filter of $h_2(n)$) and σ_W^2

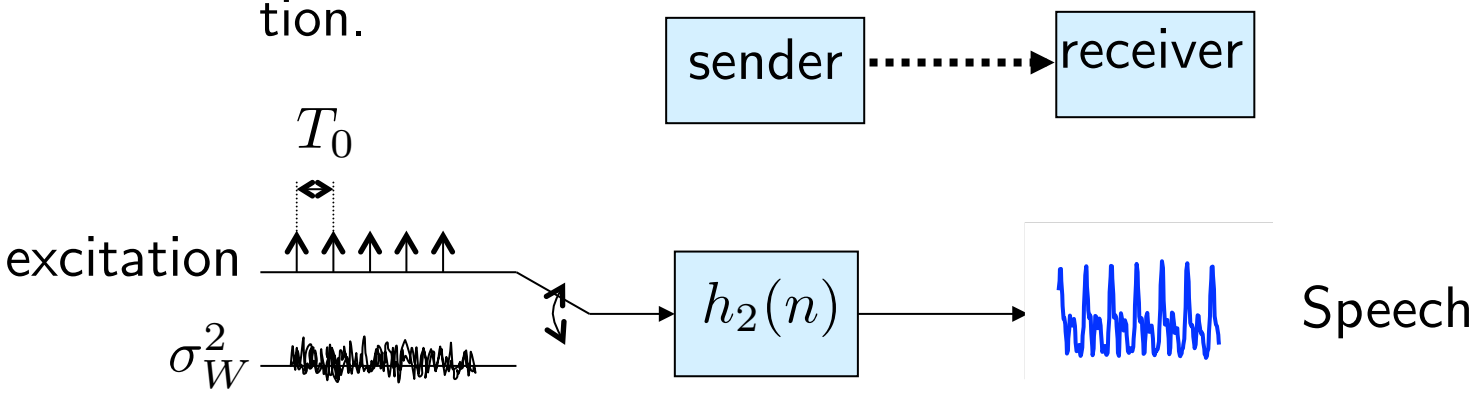
Speech Coding - Example



1 Determine filter $h(n)$ using Yule – Walker equations and error variance.

2 Determine whether speech is unvoiced or voiced

3 Transmission/Quantization of $h(n)$ and voiced/unvoiced information.



4 Synthesis of speech based on received filter $h(n)$ and voiced/unvoiced information.