Tampere University

Speech production

SGN 14007

Lecture 7

Annamaria Mesaros

Subjects covered in the next 4 lectures

- Speech production and phonetics
  - Articulatory phonetics (how sounds are produced)
  - Acoustic phonetics (modeling speech production)
- Speech processing and features
- Speech recognition
- Speech synthesis

Speech processing

- Speech is a natural way for humans to communicate
- Before writing, only way to pass knowledge to next generation
- Speech production apparatus is a part of the motor system for respiration and alimentation

- Speech processing applies common signal processing tools:
  - Short-time Fourier transform (STFT)
  - Filters to model speech production.

- Speech-specific tools:
  - LPC analysis
  - Cepstral analysis
  - Pitch estimation …

Applications of speech processing

- coding (for example mobile phones)
- enhancement: (noise reduction, hearing aids)
- synthesis (text to speech, voice interfaces)
- recognition (speech to text, speaker identity)
- modification (make a person sound like someone else)

- Big leaps during last years due to machine learning
- Speech is becoming a popular user interface
- Amazon Echo, Google assistant, Siri

Phonetics

- Phonetics studies speech:
  - Production -> ARTICULATORY
  - Acoustic realization -> ACOUSTIC
  - Perception -> AUDITORY

Perception of speech

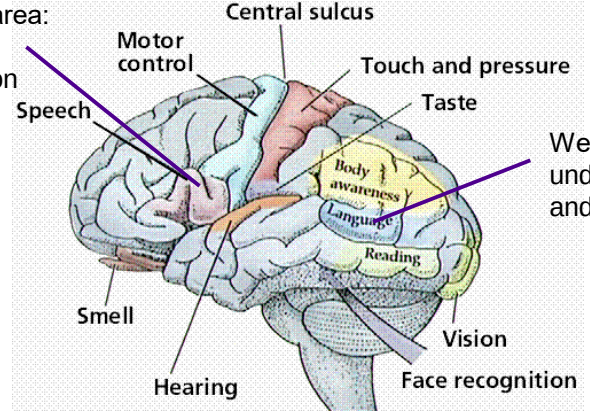
- Frequency range of 200 Hz – 5.6 kHz is most relevant for speech perception
  - Matches the range of greatest auditory sensitivity
- Speech is highly redundant
  - By clipping the speech waveform into binary values is still understandable.
  - Redundancy helps to understand speech in adverse conditions
- Listeners use context to understand
  - Anticipate words based on speaker and context of conversation, and general knowledge

Vocal organs

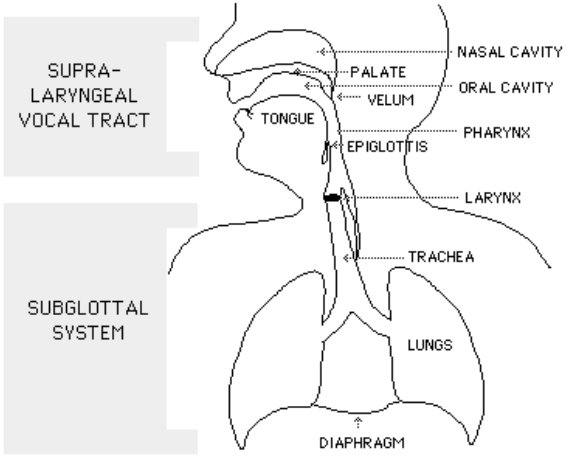
Central

Peripheral

Broca's area: speech production



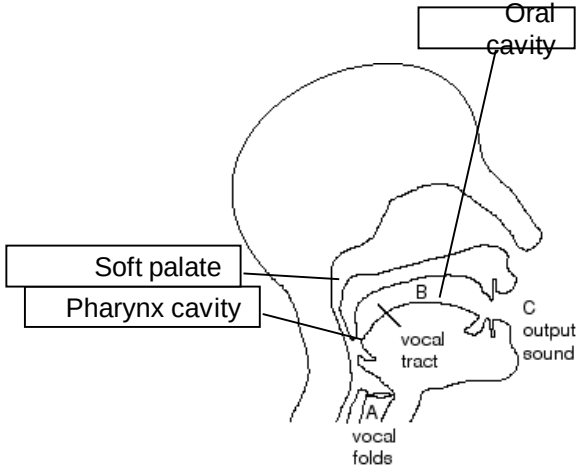
Wernicke's area: understanding of written and spoken language



Vocal tract

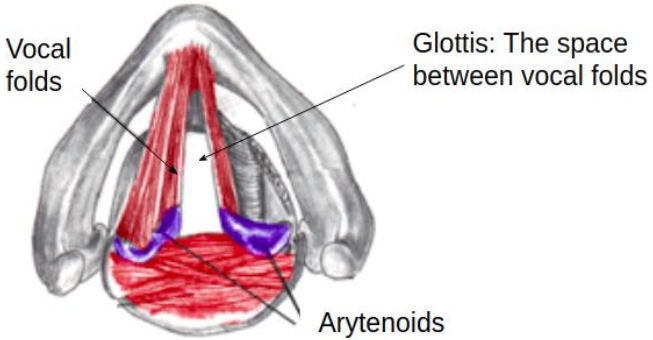
- **Vocal tract** refers to vocal organs after the larynx
- Divided into following sections:
  - Pharynx cavity (throat)
  - Nasal cavity
  - Oral cavity

- Organs of vocal tract that move to produce various speech sounds (”articulators”):
  - Tongue
  - Soft palate (velum) -> opens/closes path to nasal cavity
  - Lower jaw
  - Lips

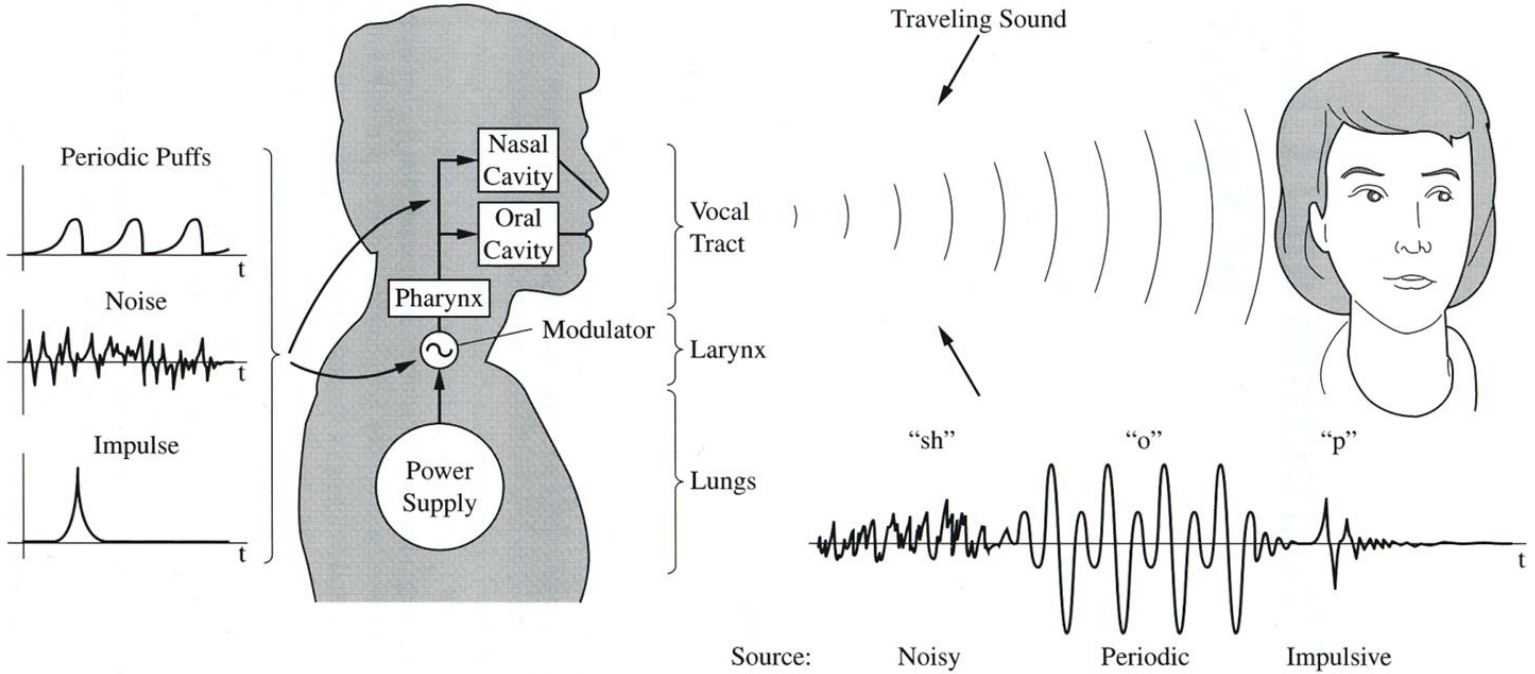


Glottis (in larynx)

- **Larynx** (voice box): an organ in the neck involved in breathing, sound production, and protecting the trachea against food aspiration.
- From the speech production viewpoint, the role of larynx is to turn the silent flow of air from the lungs into audible sound
- **Glottis** is the space between **vocal folds**
- The arytenoid cartilages are a pair of small three-sided pyramids which form part of the larynx, to which the vocal folds (vocal cords) are attached



Three sources of sound energy



Quatieri: Discrete –Time Speech Signal Processing Principles and Practice

Three sources of sound energy

1 - Periodic: phonation, vocal folds vibration

2 - Noisy: Turbulence (noise)

- Air moving quickly through a small hole
- Fricative or unvoiced sounds
- E.g. tongue/teeth (“ss” in “hiss”)

3 - Impulsive (explosion)

- Release of pressure build up
- E.g. behind lips (“p” in “peak”) or tongue (“t” in “tell”)
- Plosive sounds

Compare: “b” in “bat” (voiced plosive) with “p” in “pat” (unvoiced plosive)

Phonation (vocal folds vibration)

- Caused by pressurized air passing through the membranous portion of the narrowed glottis.
- Causes **repeated opening and closing of the glottis at frequency F0** (fundamental frequency or pitch)
- Formation of voiced sounds in this way is called phonation

- Typical values of F0: (during normal conversation)
  - males 120Hz, females 200Hz, children 300 Hz.

- Frequency of vibration: F0 can be altered with muscles from
  - 80-400 Hz for males, 120-800 Hz for females
- F0 changes depending on two factors:
  - regulation of the length of the vocal folds and
  - adjustment of aerodynamic factors result in vocal fold vibration
    - E.g. larger sub-glottal pressure increases F0

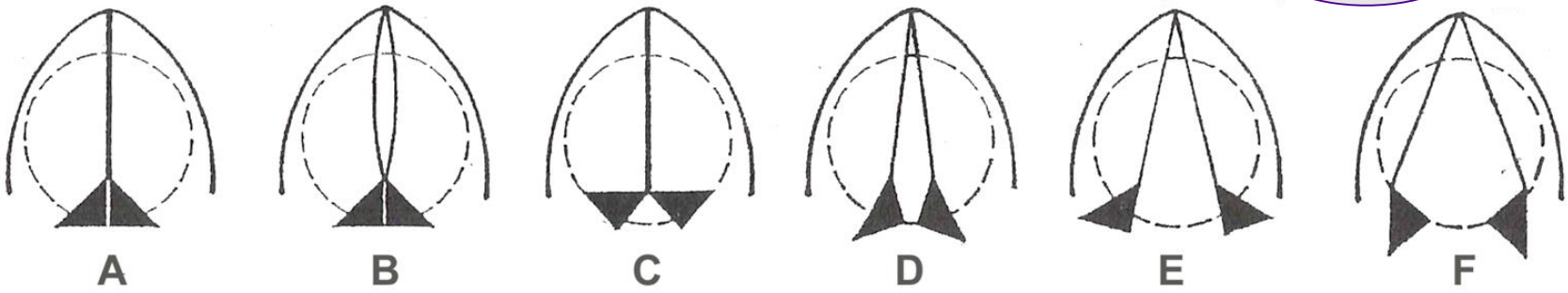
For those who want more details, not important for exam

Phonation types

- **The modal register** (“normal speech”):
  - Virtually all speech employs the modal register.
  - Speech uses an F0 range about one octave: for males 80-160 Hz.
  - Singers often use a two-octave range.
    - Higher F0 requires more effort.
- **Breathy or murmur phonation**
  - Combines voicing and whispering (arytenoids open)
- **Creaky voice**: irregular and low F0 and weak intensity.
  - Like a door creaking on its hinges
- **Vocal fry or “pulse register phonation”**
  - Sub-category of creaky voice: regular, low and audible F0 (3-50Hz)
- **Falsetto or loft**: High F0.
  - Vocal folds are thin, only the center part of the vocal folds vibrates. The vocal cords might not completely close.

For those who want more details, not important for exam

Function of the vocal folds



A: vocal folds and arytenoids closed -> **glottal closure (no airflow)**

B: Vocal folds vibrating, arytenoids closed -> **phonation, f0; voicing**

C: Vocal folds close, arytenoids open-> **whisper**

D: glottal constriction -> weak unvoiced noise, **glottal fricative** [h]

E: rest/breathing position -> **unvoiced consonants**

F: deep-breath position (sigh / breathlessness) -> **not used for speech**

http://www.youtube.com/watch?v=wjRsa77u6OU

Thinking break (2min)

Or watch how vocal chords work

https://www.youtube.com/watch?v=DD_oJ8G3Bhs

http://www.youtube.com/watch?v=wjRsa77u6OU

Articulatory phonetics

Phonemes

- **Phoneme**: the smallest linguistic unit which may change the meaning (kill vs. kiss).
  - Phonemes are combined to form larger entities such as words.
- Articulatory phonetics: describes phonemes based on how they are produced
- Phonetic alphabets:
  - International phonetic alphabet (IPA)
  - Represents sounds with symbols.
  - For notational reasons (ASCII-based) others are used too, e.g. Arpabet

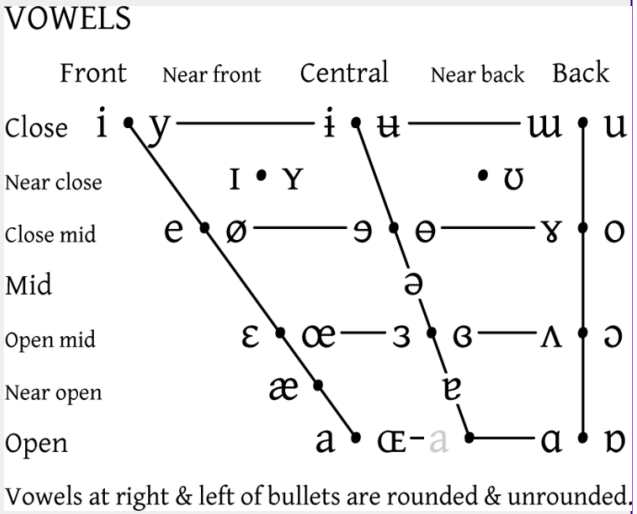
- Speech sounds
  - Consonant vs. vowel:
    - Consonants involve an obstruction in air stream above the glottis.
  - Voiced vs. voiceless:
    - Voiced if vocal chords vibrate
  - Nasal vs. oral
    - Nasal if air travels through nasal cavity and oral cavity closed
  - Lateral vs. non-lateral
    - In lateral phonemes, air stream passes through the sides of the oral cavity (”ball”, ”lateral”) and not through the middle

Vowels

- Vowels are voiced phonemes, where the vocal tract is open.

- Vowels are characterized by using articulation features:
  - Open-Close dimension: refers to how close the tongue is to the roof of the mouth. The more closer to palate the more ”closed” the the vowel is.
  - Front-Back dimension: refers to position of articulation by means of tongue positions: the narrowest point of the vocal tract is essential.
  - Lip roundedness (binary value), right vs. left of bullet: rounded • unrounded. e.g. i • y
  - Nasalization: When the velum is open airflow gets to the nasal cavity and a nasal phoneme is produced. When the velum is closed, an oral phoneme is produced.

http://www.internationalphoneticalphabet.org/ipa-sounds/ipa-chart-with-sounds/



Details

Consonants

- In most consonants, the airflow is obstructed at some point

- Consonants are characterized by:
  - Voicing – voiced or unvoiced
    - Determined by the vibration of the vocal folds
  - Manner of articulation
    - How freely the air stream flows when the consonant is produced
  - Place of articulation
    - Where is the primary constriction along the vocal track

- **Voicing**: A consonant can be voiced or unvoiced
  - In English, voiced consonants include [v] (van), [z] (zip), [ʒ] (confusion), [b], [d], [g], [dʒ] (gin)
  - Unvoiced consonants include: [f], [s], [p], [t], [k], [h], [s], [tʃ]

Consonants' manners of articulation

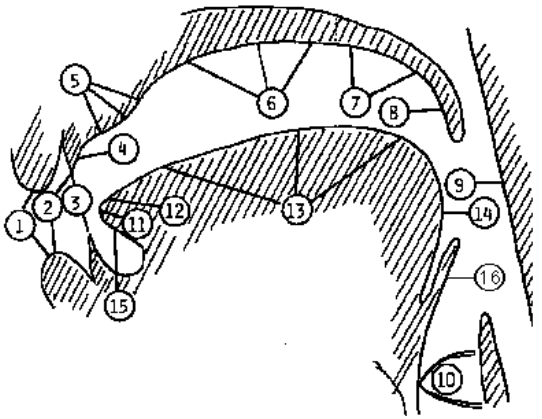
- Main variation in the manner of articulation regards the question **how freely the air stream flows** when the consonant is produced

- Sonorants: continuous, non-turbulent airflow in the vocal tract
  - vowel-like sounds with greater constriction or less loud (e.g. /m/, /n/, /w/, /l/)
- Obstruent: airflow is partly or completely obstructed
  - Plosive: Complete block + sudden release: (e.g. /p/,/t/)
  - Fricative: articulators close together, turbulent airflow produced. Usually most of the energy at high frequencies (e.g./f/,/s/,/z/,/sh/,/h/)
- Flaps and trills:
  - Articulator vibrates once or rapidly.

Details

Consonants' places of articulation

- Place of articulation tells where is the primary constriction along the vocal track:
  - **bilabial** (1): made with the two lips (P,B,M)
  - **labio-dental** (2): lower lip & upper front teeth (F,V)
  - **dental** (4): tongue tip/blade&upper front teeth (TH,DH)
  - **alveolar** (5): tongue tip/blade & alveolar ridge (T,D,N)
  - **retroflex**: tongue tip & back of the alveolar ridge (R)
  - **palato-alveolar**: tongue tip&back of the alveolar ridge (SH)
  - **palatal** (6): front of the tongue & hard palate (Y,ZH)
  - **velar** (7): back of the tongue & soft palate (K,G,NG)
  - **uvular**: (8) back of the tongue against or near the uvula.
  - **pharyngeal**: (9) in the pharynx
  - **glottal**: (10) in the glottis



(you do not have to remember the above latin words)

For those who want more details, not important for exam

IPA – international phonetic alphabet

THE INTERNATIONAL PHONETIC ALPHABET (revised to 2005)

CONSONANTS (PULMONIC)

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

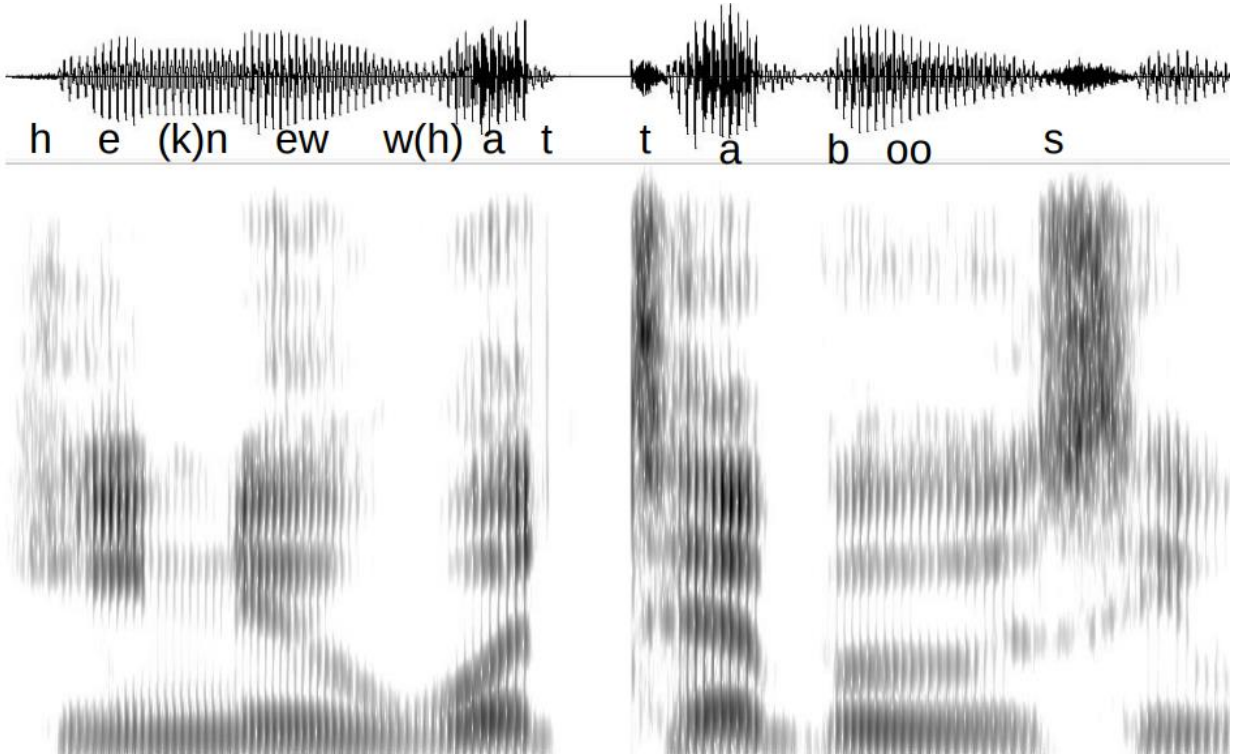
Pronunciation of IPA consonants: http://teaching.ncl.ac.uk/ipa/consonants-pulmonic.html

Details, not important for exam

Other phonetics terms

- **Phoneme**: the smallest linguistic unit which may change the meaning (kill vs. kiss). Phonemes are combined to form larger entities such as words. Noted in text with slashes e.g. /i/
  - The English language has 42 phonemes
- **Phone**: individual spoken **realization** of a phoneme
  - In principle all phones are different
  - Different speech sounds that are realizations of the same phoneme are known as **allophones** (t in tree, toe, catnip, button, etc)
  - noted in text with brackets e.g. [i]
- **Coarticulation**: vocal organs move in a continuous manner and therefore (conceptually isolated) speech sound is influenced by, and becomes more like, a preceding or following speech sound.
- **Diphone**: the time-span between the middle-part of a phone until the middle part of the following phone. Includes phone transition.
- **Triphone**: a temporal unit that covers two diphones.

Example sentence



Prosody

- Prosody refers to **longer-term properties** of speech
- **Rhythm**: varying the temporal length of syllables (or some other units**)**
- **Stress**: relative emphasis of syllables in a word or certain words in a sentence, manifested in higher/lower pitch or dynamics (loudness)
- **Intonation**: variation of pitch over a segment of multiple words (e.g. Sentence) that may:
  - indicate the attitudes and emotions of the speaker
  - signal the difference between statement and question
  - focus attention on the important words

Break

Consider production of different phonemes by noticing mouth shape, place of articulation, etc

Chech the IPA chart with sounds

Speech production models

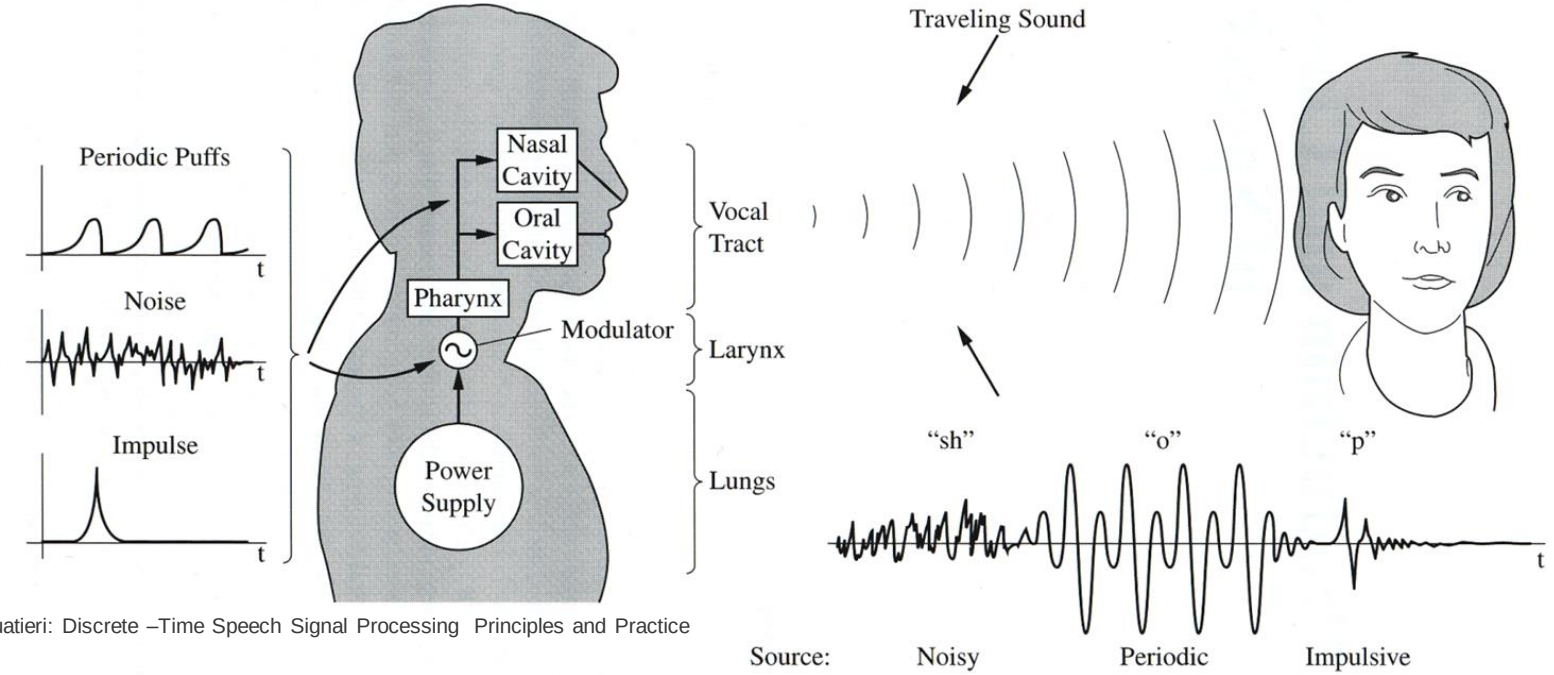
Modeling of speech production

- A common mathematical approximation of the acoustic production of speech is the source-filter model
  - Allows to efficiently code and synthesize speech
  - Also to understand speech mechanisms

- The source-filter model consists of two parts:
  - **Source signal generator** (glottal excitation)
    - Voiced sounds: periodic glottal vibration
    - Unvoiced sounds:
      - Turbulent noise for fricatives
      - Pressure release for bursts
  - **Filter** that acts on the signal
    - Models the vocal tract (and lips)

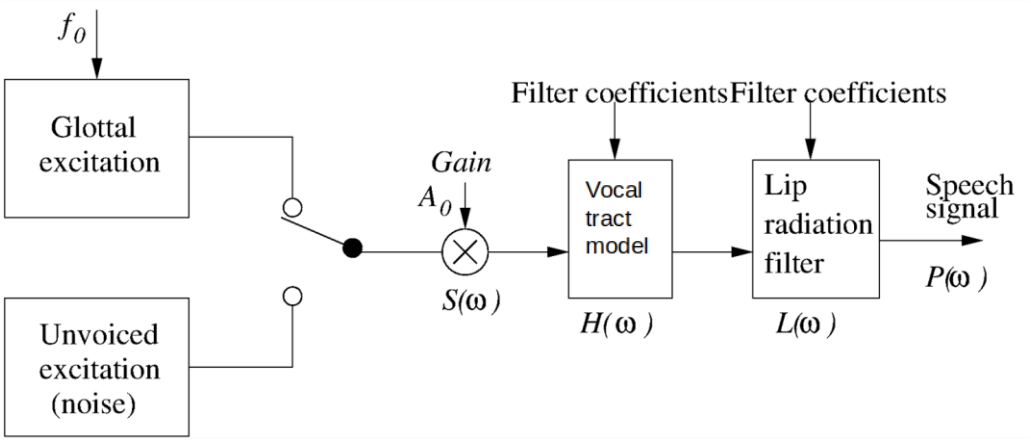
Source-filter model

Illustration of different source signals and the parts of the filter, and the resulting waveform.



Quatieri: Discrete –Time Speech Signal Processing Principles and Practice

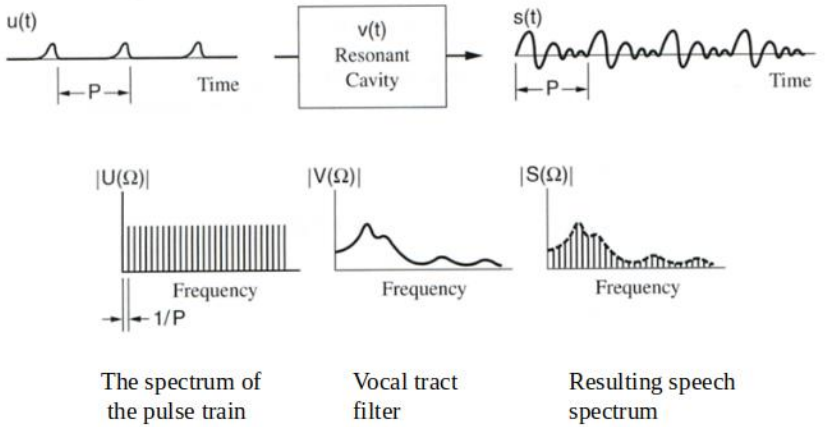
Source-filter signal processing model



$$P(\omega)=S(\omega)\cdot H(\omega)\cdot L(\omega)$$

Source-filter model

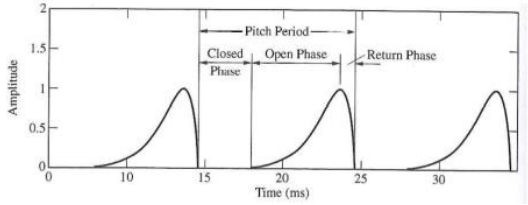
- Figure: voiced excitation signal u(t) is filtered
- Top row: time-domain signals and operations
- Bottom: corresponding frequency domain magnitude responses



Quatieri: Discrete –Time Speech Signal Processing Principles and Practice

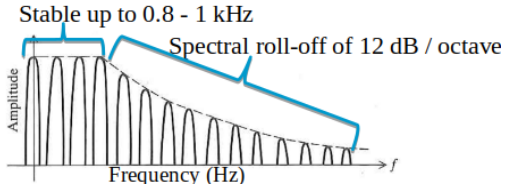
Glottal pulse-train during phonation

Glottal pulse train is created by airflow passing vocal cords

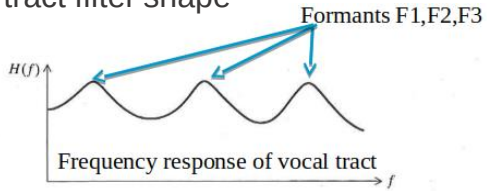


Remember: Phonation = voiced speech

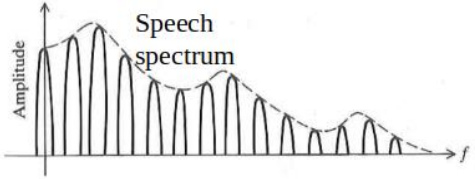
Ideal spectrum of a glottal pulse-train



Vocal tract filter shape



Resulting speech signal spectrum (bottom right)

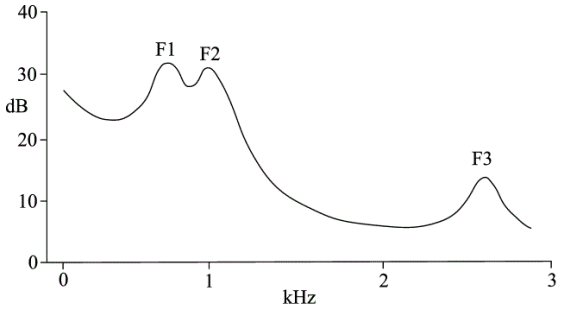


Formants: resonances of vocal tract

- The most important characteristic of the vocal tract are its **resonances (formants)**
  - Due to standing waves in the vibrating air column
  - Formants (F1, F2, ...) can usually be seen in the spectrum as boosted frequency regions
  - In addition to frequency, a formant is characterized by its intensity and bandwidth
  - Different vocal tract configurations correspond to different formant frequencies

  → **All vowels can be classified based on formants**

spectrum of phoneme /a/



Formants

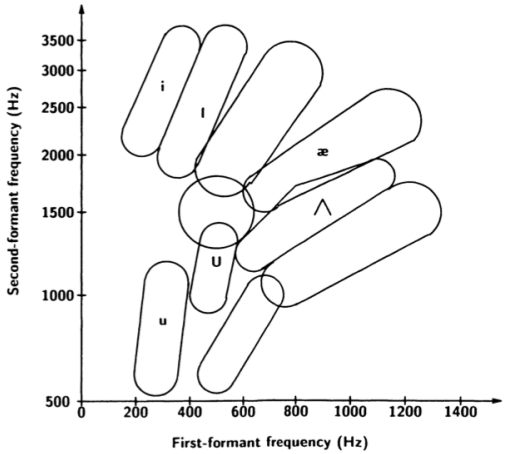
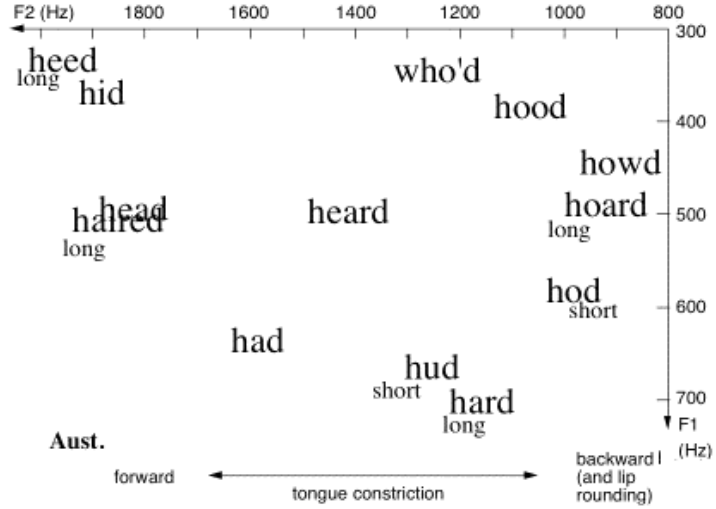
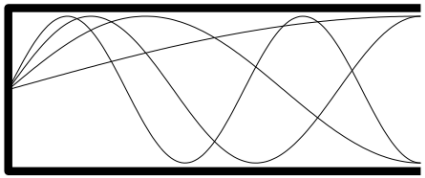


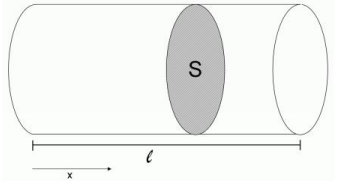
Figure 3.13 Plot of F1 vs F2 for vowels spoken by 60 speakers. (After Peterson and Barney [31].)

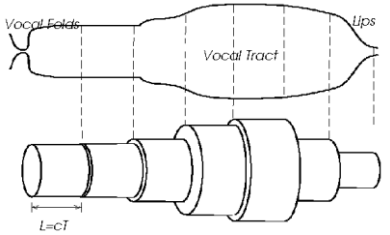


Modeling the vocal tract with simple tubes

- Resonances of the vocal tract are due **to standing waves in the vibrating air column** (similar to e.g. wind instruments)
- In a tube of uniform cross-sectional area, the wavelengths λ of standing waves are $4\ell, \frac{4}{3}\ell, \frac{4}{5}\ell, \frac{4}{7}\ell, \ldots$
- Substituting the typical vocal tract length (male 0.17m, female 0.15m), the frequencies of the resonances would be 500Hz, 1500Hz, 2500Hz
- In vocal tract, the cross-sectional area S varies and thus resonance frequencies vary, but as a rule of thumb, there is roughly one resonance per 1 kHz

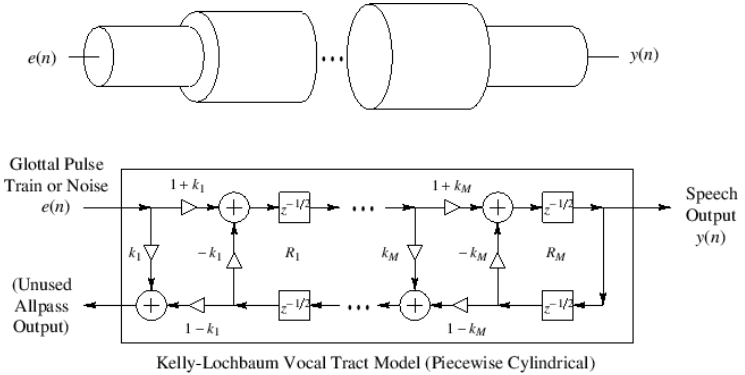






Lattice structure

- When two simple tubes are joined, reflections occur at the boundary
- Wave propagation and reflections modeled using the Kelly-Lochbaum lattice structure



Kelly-Lochbaum Vocal Tract Model (Piecewise Cylindrical)

The lattice-structured filter resulting from the Kelly-Lochbaum equations is an **all-pole filter:**

- its transfer function $H(z) = \frac{g}{A(z)}$ includes only poles (= zeros of denominator A(z))

Physical tube models

- Length of one tube section with this model:
  - 340[m/s]/16000[1/s] = 2 cm
- Vocal tract length: 17cm(male), 15cm(female)
  - 8 or 7 tube sections in the vocal tract with 16kHz

- Watch a proof-of-concept:

  http://www.splab.net/APD/G200/

Thinking break (2 min)

Watch Samuli Siltanen’s video https://youtu.be/3mwElfO4x8A?t=3m50s

Linear system models

Linear system models

- Linear models can be categorized as one of the following (in speech processing, input x(n) refers to glottis excitation and output y(n) to measured speech):
- **Autoregressive moving average model (ARMA)**

$$\hat{y}(n) = -\sum_{k=1}^{p} a(k)y(n-k) + \sum_{k=0}^{q} b(k)x(n-k)$$

Corresponds to a **generic linear recursive filter**

$$H(z)_{ARMA} = \frac{B(z)}{A(z)}$$

- **Moving average model (MA)**

$$\hat{y}(n) = \sum_{k=0}^{q} b(k)x(n-k)$$

Corresponds to a **FIR filter**

$$H(z)_{MA} = B(z)$$

- **Autoregressive model (AR)**

$$\hat{y}(n) = gx(n) - \sum_{k=1}^{p} a(k)y(n-k)$$

Corresponds to an **all-pole filter** (gain g is constant)

$$H(z)_{AR} = \frac{g}{A(z)}$$

AR model for modeling the vocal tract

- Typically the AR model is used in speech processing, because:
  - The lattice-structured model for the vocal tract corresponds to an all-pole filter (AR model).
    - In other words, the vocal tract is (with certain assumptions) theoretically an all-pole filter
  - The input signal x(n) is not known (glottal excitation).
  - AR model parameters a(k) can be computed efficiently.
  - A higher-order AR model can (to some extent) represent also the more generic ARMA model

- AR model parameters are estimated from the input signal using **linear prediction analysis**
  - Linear prediction is a good method for estimating the parameters of the vocal tract
  - Linear prediction is one of the most important tools in speech processing
  - Acronyms: LP (linear prediction), LP-analysis, LPC (linear predictive coding)
- From the speech processing viewpoint, the most important property of LP is its ability to model the vocal tract

Linear prediction analysis

- LP analysis finds the filter coefficients (a0, a1, a2,…, ap) that best predict the signal samples according to the AR model (discarding x(n) term):

$$\hat{y}(n) = gx(n) - \sum_{k=1}^{p} a(k)y(n-k)$$

$$\hat{y}(n) = -\sum_{k=1}^{p} a_k y(n-k)$$

- The coefficients of the predictive filter are chosen so that the squared prediction error is minimized in the analysis window:

$$\bar{a} = \arg\min_{\bar{a}} \sum (y(n) - \hat{y}(n))^2$$

- The resulting vector contains the **linear prediction coefficients**
- Keep in mind that speech is processed in short frames
  - Also LP analysis is done approx. every 10-30 ms in partly overlapping frames

Linear prediction and Yule-Walker equation

- **Optimal** filter parameters (LP coefficient) a(1),a(2),...,a(p) are found by setting **the partial derivatives of E with respect to each parameter a(n) to zero**.
- Represented with the autocorrelation function ,

$$r(\tau) = \sum_{n=0}^{N-1} x(n)x(n+\tau)$$

- The zeros of the partial derivates can be written in matrix form as **Yule-Walker equation**

$$\begin{bmatrix} r(0) & r(1) & r(2) & \cdots & r(p-1) \\ r(1) & r(0) & r(1) & \cdots & r(p-2) \\ r(2) & r(1) & r(0) & \cdots & r(p-3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r(p-1) & r(p-2) & r(p-3) & \cdots & r(0) \end{bmatrix} \begin{bmatrix} a(1) \\ a(2) \\ a(3) \\ \vdots \\ a(p) \end{bmatrix} = - \begin{bmatrix} r(1) \\ r(2) \\ r(3) \\ \vdots \\ r(p) \end{bmatrix}$$

- Solution:
  - For example the matrix inverse; computationally expensive
  - Levinson-Durbin recursion: (uses a(i) to obtain a(i+1)) is a fast algorithm

Linear prediction analysis

- The error criteria (mean squared error) affects how the spectrum is modeled
- The mean-squared error criteria is equivalent to minimizing the integral:

$$D = \int_{-\pi}^{\pi} \frac{|Y(\omega)|^2}{|H(\omega)|^2}$$

  - H(ω) is the response of the LP filter
  - Y(ω) is the signal spectrum

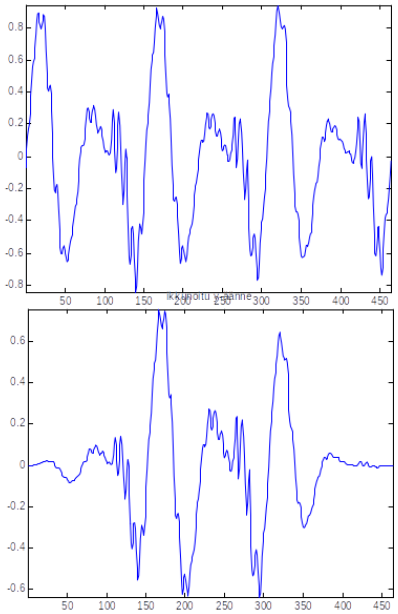
- The solution will model the peaks closely, and not model the low values.
- Low values of $|Y(\omega)|^2$ do not contribute to the error D

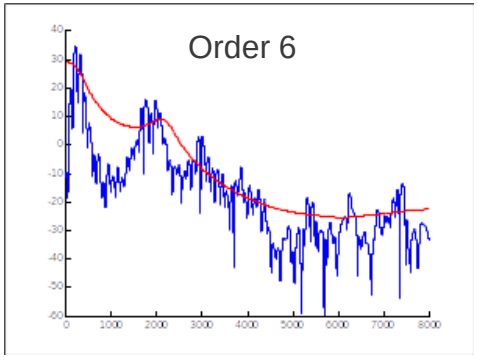
Choosing the model order

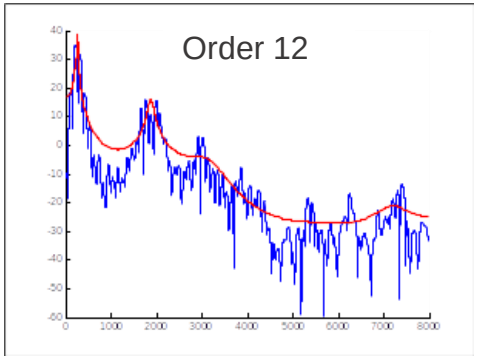
- There is usually one formant per kHz
  - Model order p can be estimated as the sampling rate in kHz
- For example
  - sampling rate 8kHz → model order 8
  - sampling rate 16kHz → model order 16
- However to compensate for model inaccuracies, usually a slightly higher model order is selected
  - sampling rate 8kHz → model order 10 or 12
  - sampling rate 16kHz → model order 18 or 20

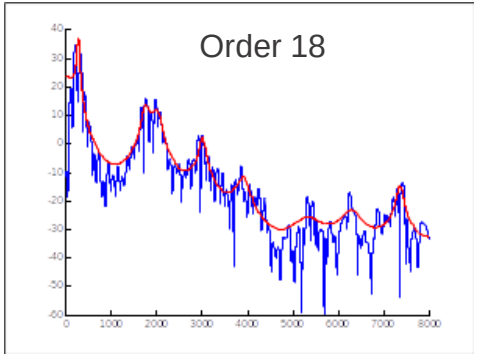
Choosing the model order

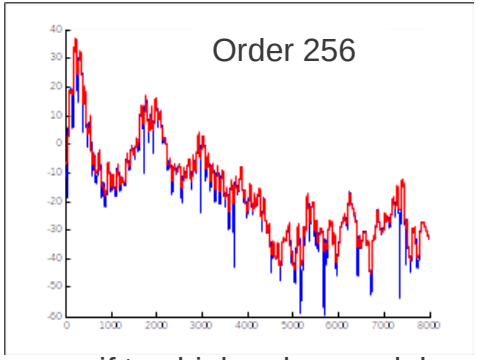
Phoneme /y/ example









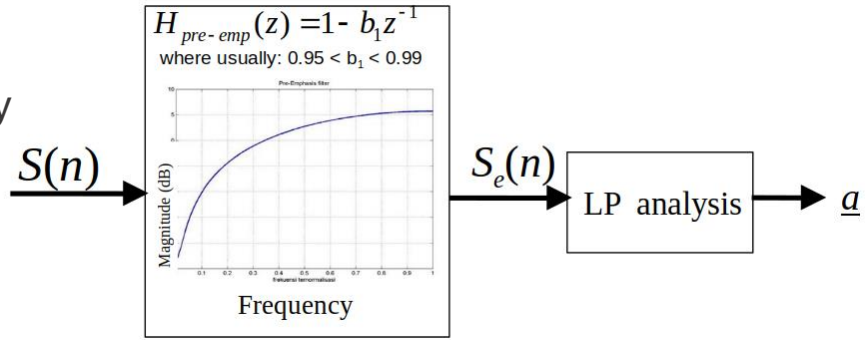


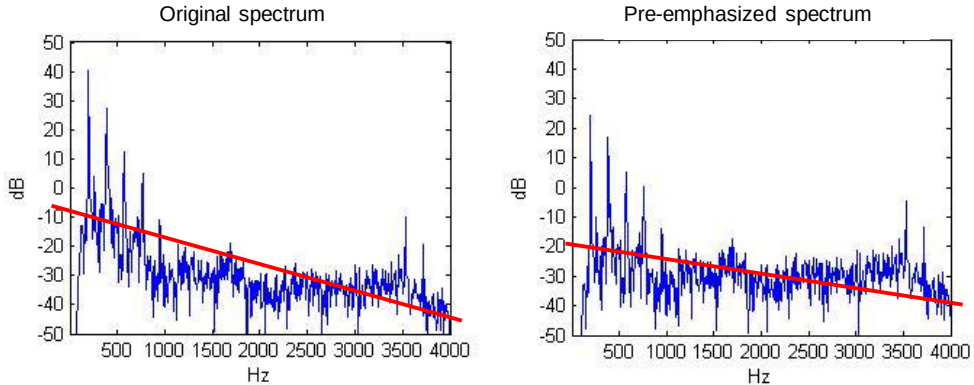
At sampling rate 16kHz, a good choice for model order would be 18. What happens if too high or low model order is chosen?

Pre-emphasis of high frequencies

- As can be seen in the previous figures, **speech spectrum has much less energy at high frequencies than at low frequencies**. That may have the consequence that LP analysis will not find any of the higher formants.

- To address the issue, usually a **pre-emphasis filter** is used that flattens the spectral tilt before the LP analysis:





Where is LP analysis used?

- Speech coding: enables the separate coding of excitation and vocal tract parameters
- Speech recognition: provides information about the speech spectrum (and therefore about the phoneme identity)
- Speech synthesis: allows separate control of the excitation and vocal tract parameters
- In MATLAB, LP analysis can be done with the command lpc
- In Python, `audiolazy` and `scikits.talkbox` provide LP implementation

Formant estimation

- Formants are resonances of the vocal tract
- A straightforward way to estimate formants is to factorize the LP polynomial

$$A(z) = 1 + a_1 z^{-1} + \ldots + a_p z^{-p}$$

into factors

$$A(z) = (1 - z_1 z^{-1})(1 - z_2 z^{-1})\ldots(1 - z_p z^{-1}),$$

z_1, z_2, .. z_p are the roots of the LP polynomial

Magnitude response of a pole pair: frequency

Pair of poles $re^{\pm j\theta}$ has transfer function:

$$\frac{1}{(1-re^{j\theta}z^{-1})(1-re^{-j\theta}z^{-1})}$$
$$=\frac{1}{1-r(e^{j\theta}+e^{-j\theta})z^{-1}+r^2e^{j\theta}e^{-j\theta}z^{-2}}$$
$$=\frac{1}{1-2r\cos(\theta)z^{-1}+r^2z^{-2}}$$

So the coefficients of the transfer function are:

$a_0 = 1$

$a_1 = -2r\cos(\theta)$

$a_2 = r^2$

On the unit circle $z = e^{j\omega}$ in the complex plane, the transfer function can be written as:

$$\frac{1}{(1-re^{j\theta}z^{-1})(1-re^{-j\theta}z^{-1})}$$
$$=\frac{1}{(1-re^{j\theta}e^{-j\omega})(1-re^{-j\theta}e^{-j\omega})}$$
$$=\frac{1}{(1-re^{j(\theta-\omega)})(1-re^{-j(\theta+\omega)})}$$

The magnitude response (absolute value of the transfer function) gets its maximum value when $(1-re^{j(\theta\pm\omega)})$ gets its minimum value, and that happens when $e^{j(\theta\pm\omega)}=1$, from which we get $\omega=\pm\theta$

→corresponding to frequency $\pm\frac{\theta}{2\pi}Fs$

Magnitude response of a pole pair: bandwidth

- Format bandwidth expresses how wide a formant is: if a formant is steep, its bandwidth is small, and vice versa
- Bandwidth is defined as the frequency range between the points where the magnitude response has dropped 3dB from its maximum value

$$BW_\omega = -2\ln r \qquad BW_f = -2\ln r \frac{F_s}{2\pi}$$

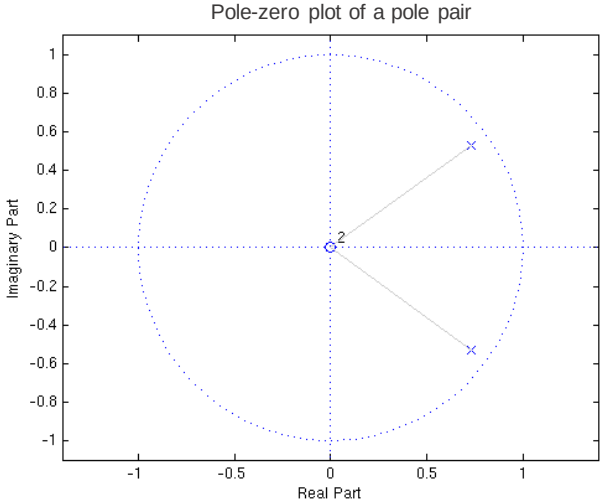
- The bandwidth of a pole pair $re^{\pm j\theta}$ depends on the distance of the pole from the origin

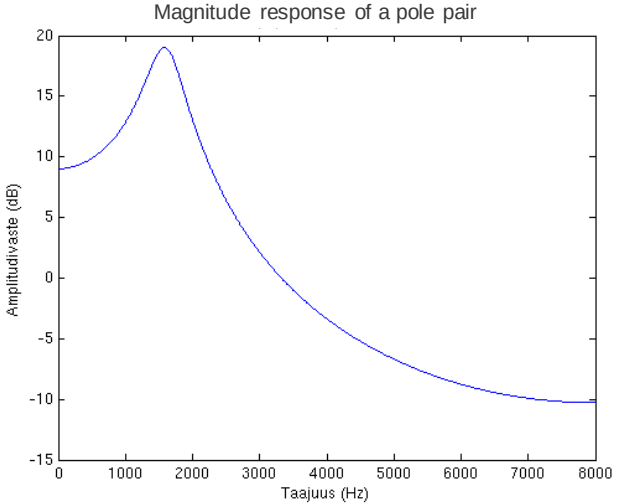
Magnitude response of a pole pair: example

Pole pair $z = 0.95e^{\pm i\,0.12\pi}$

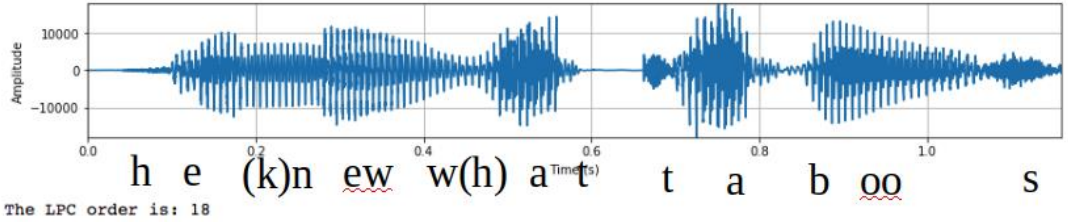
$$f = \frac{0.1 \cdot 2\pi}{2\pi} 16000Hz = 1600Hz$$

$$BW_f = -2\ln(0.95) \cdot \frac{16000}{2\pi} = 260\,Hz$$

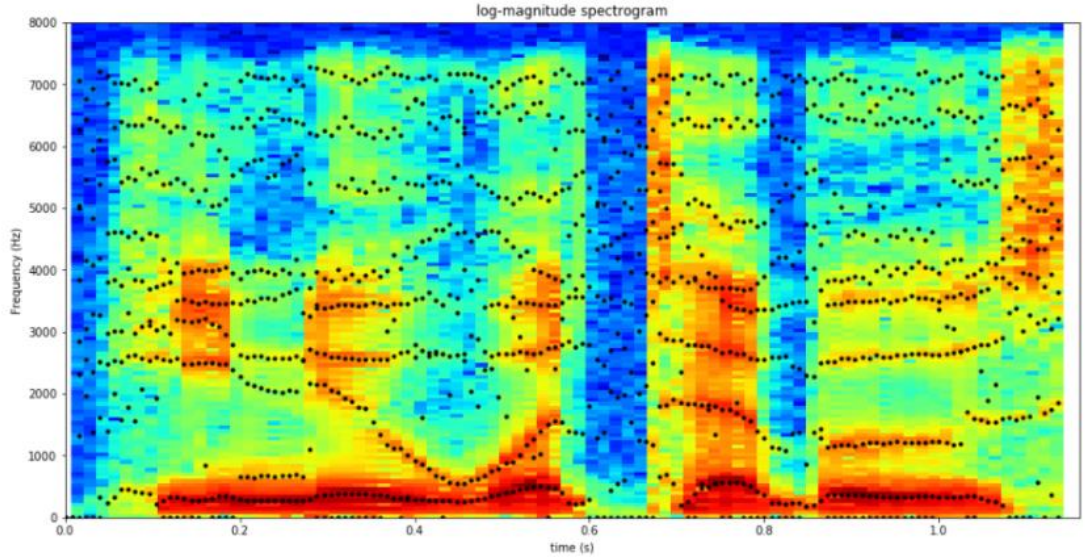




Example




```
The LPC order is: 18
```



Summary

Speech production

- Speech production organs and their role
  - Larynx, glottis, vocal tract
- Articulation of vowels and consonants
  - How we move articulatory organs to produce these
- Basic concepts
  - Phoneme, phone, prosody…

Source-filter model

- Understand the “acoustic role” of the glottis and vocal tract in the source-filter model
- Concept of formants
- Vocal tract modeling with different cross-sectional concatenated tubes
- All-pole filter
- Linear model (linear prediction) to estimate the vocal tract filter parameters
  - How to use it, selecting parameters, interpreting results