Speech Synthesis

Akihiro Kato
University of Eastern Finland

Lecturer’s Profile

Akihiro Kato

1997-2001: Chuo University in Japan

Electric and electronic engineering

2001-2012: ALPS Electric Co., Ltd. In Japan

Bluetooth headsets / car kits

2012-2016: University of East Anglia in the United Kingdom

Speech and audio-visual processing using machine learning

2017- University of Eastern Finland

Overview

- Concepts of model-based speech synthesis
- High level and low level modelling
- Source-filter model
- Sinusoidal model

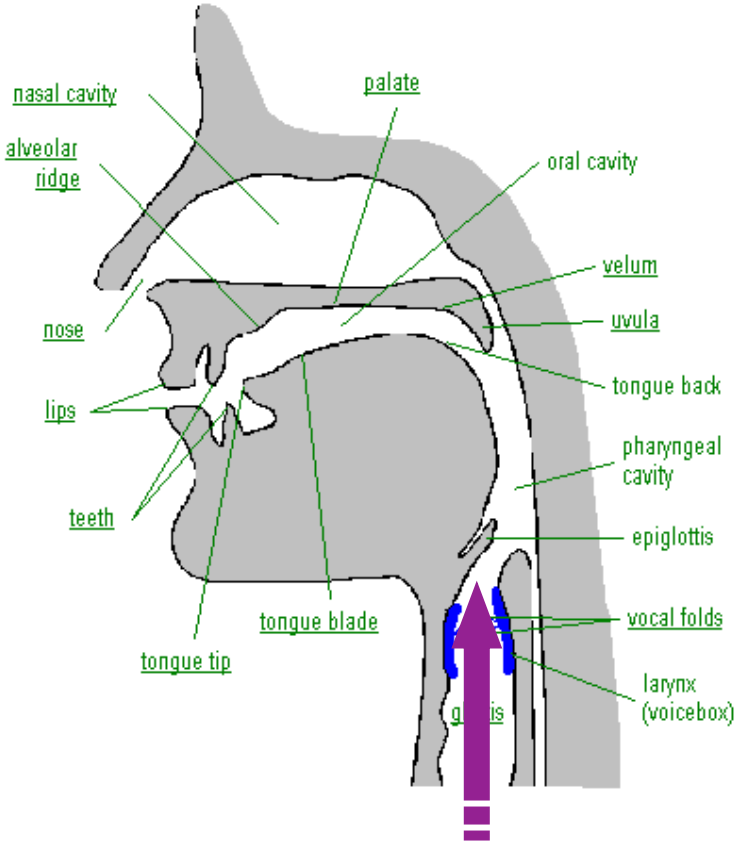
Model-Based Speech Synthesis

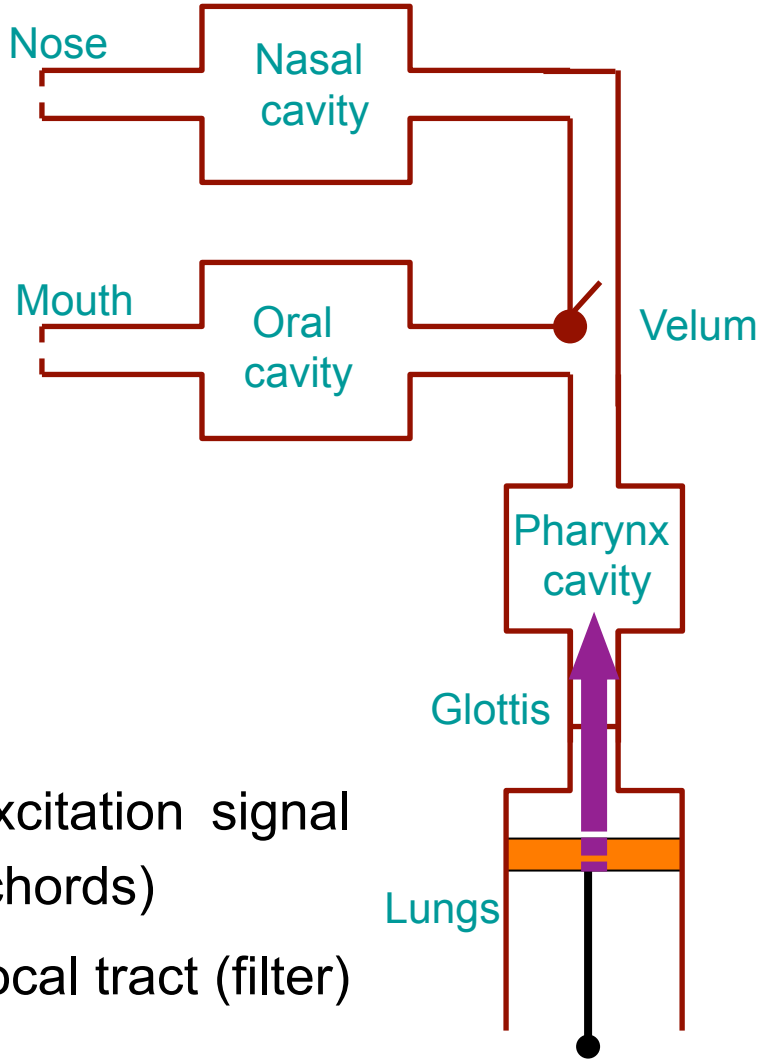
- In model-based speech synthesis, the properties of the signal that is to be synthesised are modelled in some way by a set of parameters
- A highly sophisticated model of the sound generation process can be modelled ➔ Low level modelling
- Alternatively, Broader parameters of the signal also can be modelled ➔ High level modelling
- Low level modelling has the potential to provide a more realistic and better synthesis of the sounds - however, it is usually more computationally complex and obtaining true values of model parameters is difficult
- High level modelling uses less complex models and can generally synthesise an adequate quality of the sound and usually requires fewer parameters

Model-Based Speech Synthesis

- In this lecture of model-based speech synthesis we will focus on two models of speech
  - –Source-filter model
  - –Sinusoidal model
- The sinusoidal model is a high level model that requires relatively few parameters that are easy to measures
- The source-filter model can operate at different levels of modelling
- The filter component of the source-filter model is generally a high level model while the source part can be modelled at either low level or high level
- These models also can be applied to other waveforms such as music

Model-Based Speech Synthesis

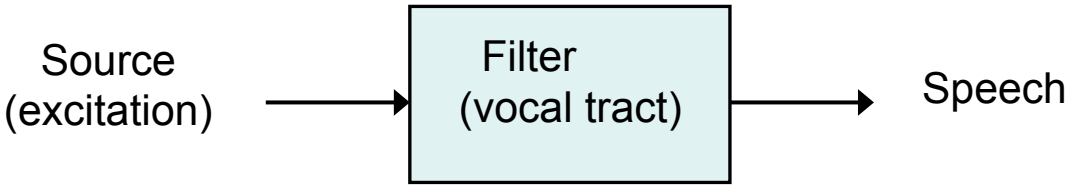




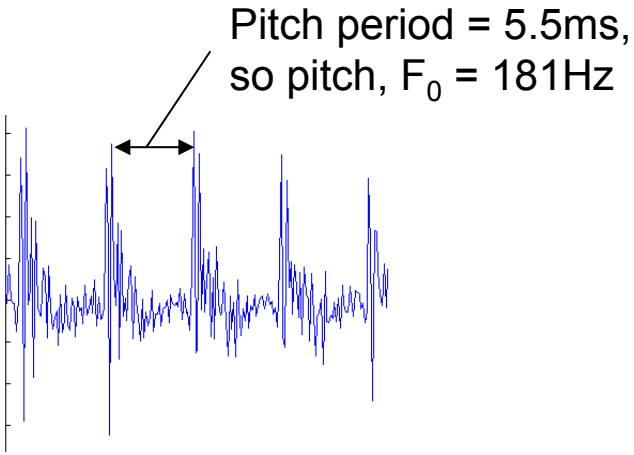
- From air expelled by the lungs, an excitation signal (source) is created at the glottis (vocal chords)
- The excitation signal then excites the vocal tract (filter) and produces speech

Source-Filter Model of Speech

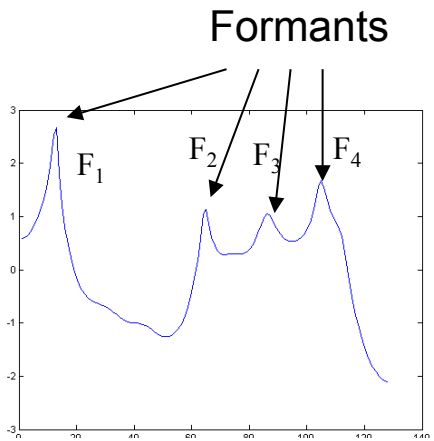
The source-filter model assumes that a source signal (excitation) is generated by the lungs and vocal chords. Then this is filtered by the vocal tract



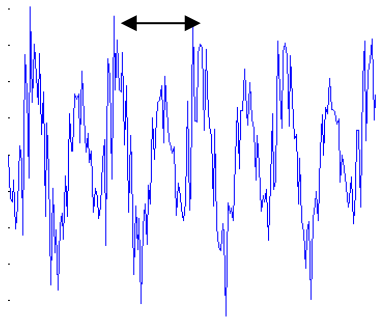
Voiced Speech



Excitation signal



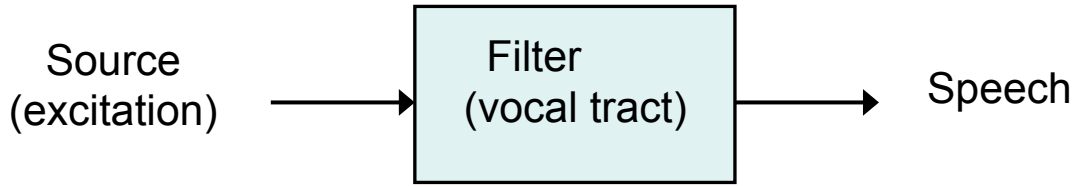
Frequency response of vocal tract



Speech signal

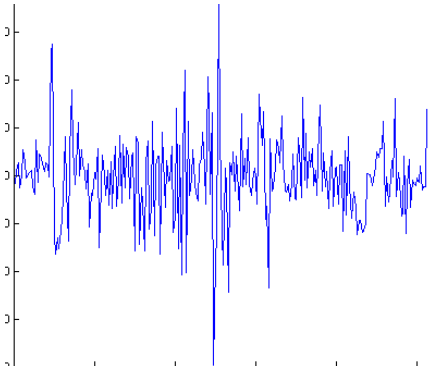
Source-Filter Model of Speech

The source-filter model assumes that a source signal (excitation) is generated by the lungs and vocal chords. Then this is filtered by the vocal tract

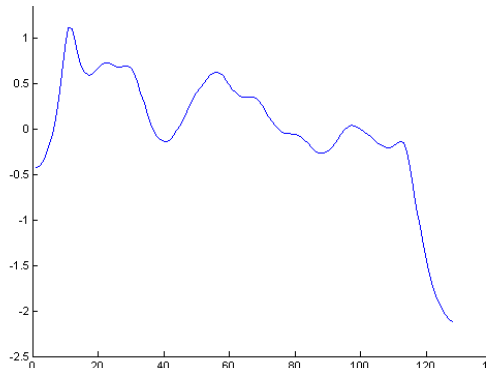


Unvoiced Speech

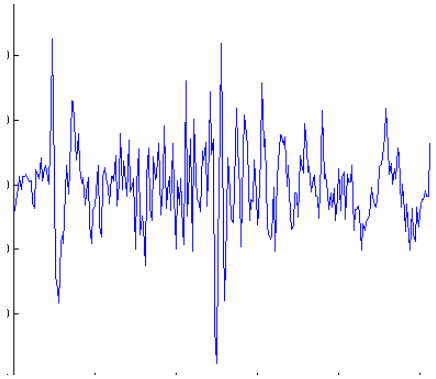
No pitch pulses



Excitation signal



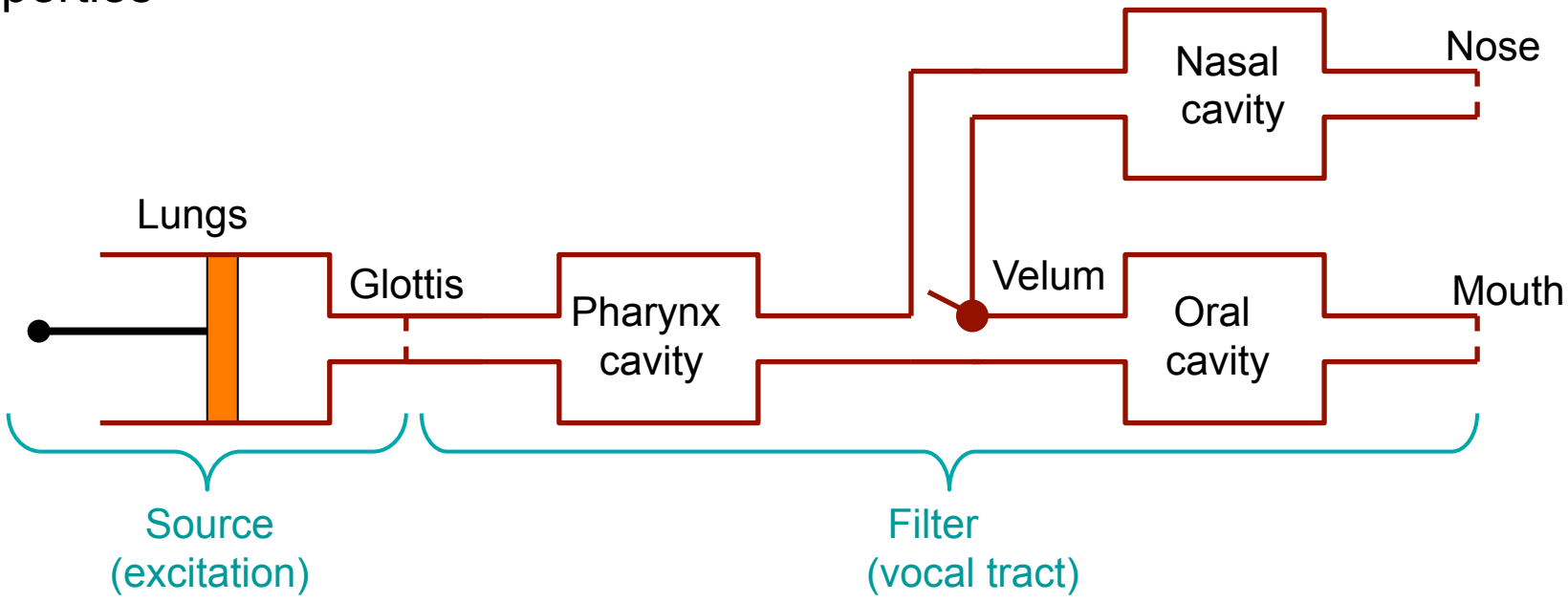
Frequency response of vocal tract



Speech signal

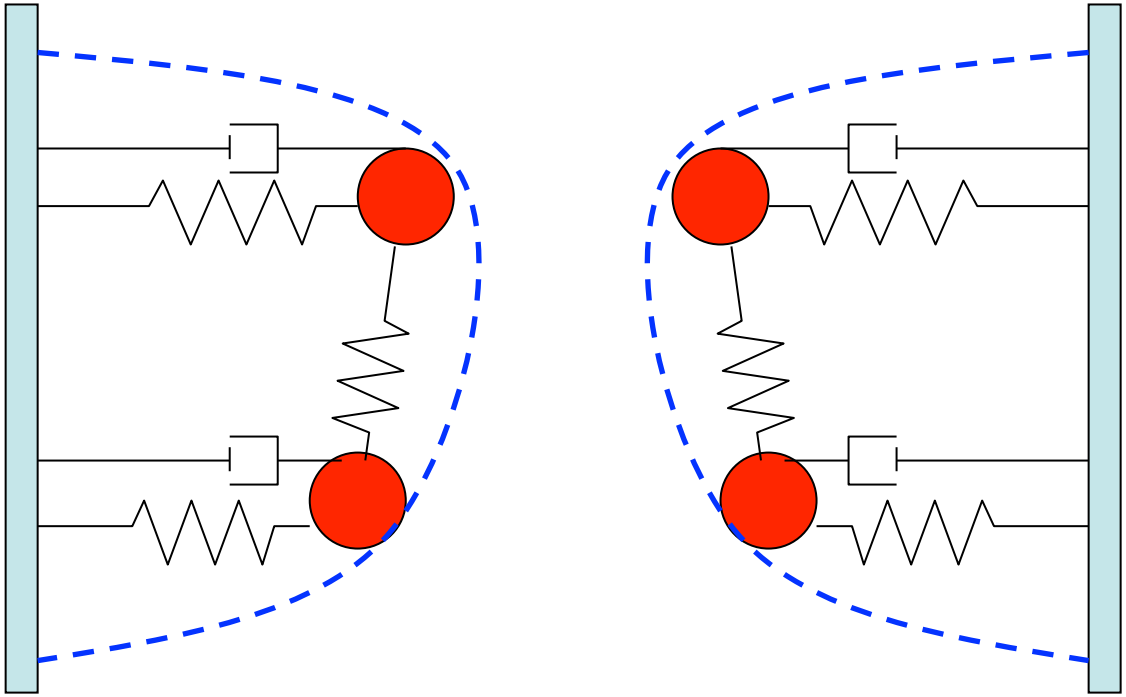
Source-Filter Model of Speech

- To model speech using the source-filter model, we must model two components - source (excitation) and filter (vocal tract)
- We can model the source at
  - –Low level using a physical model of the vocal chords
  - –High level using a model of the shape of the excitation signal
- The vocal tract filter uses high level model that is based on its spectral properties



Physical Model of Larynx

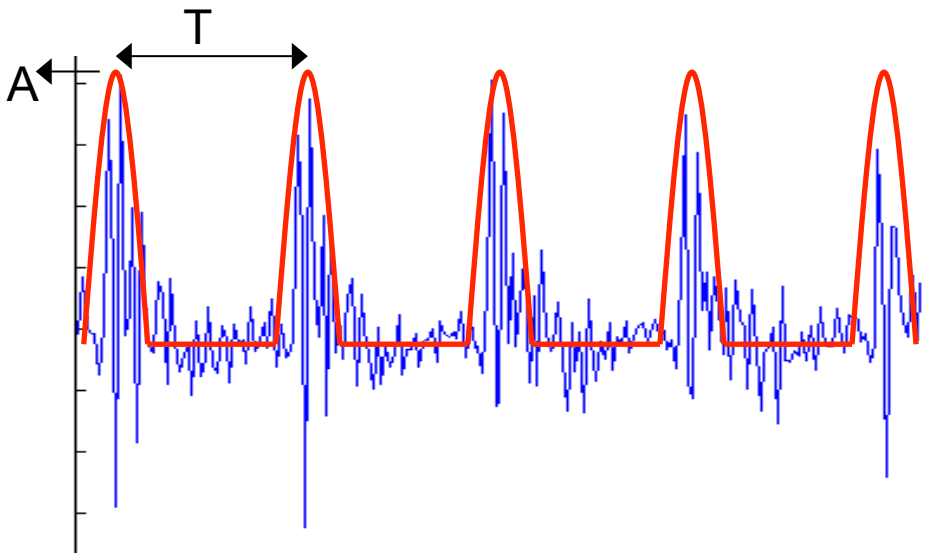
- Two-mass model of larynx
- On either side of the larynx two weights are connected by springs and dampers to model the shape of larynx
- Model parameters must be chosen carefully to give the correct movement of the weights representing larynx



- Model parameters include
  - –M1 lower mass
  - –M2 upper mass
  - –K1 lower spring stiffness
  - –K2 upper spring stiffness
  - –Kc couple spring stiffness
  - –E1 damping ratio
  - –E2 damping ratio
  - –Lg glottal length

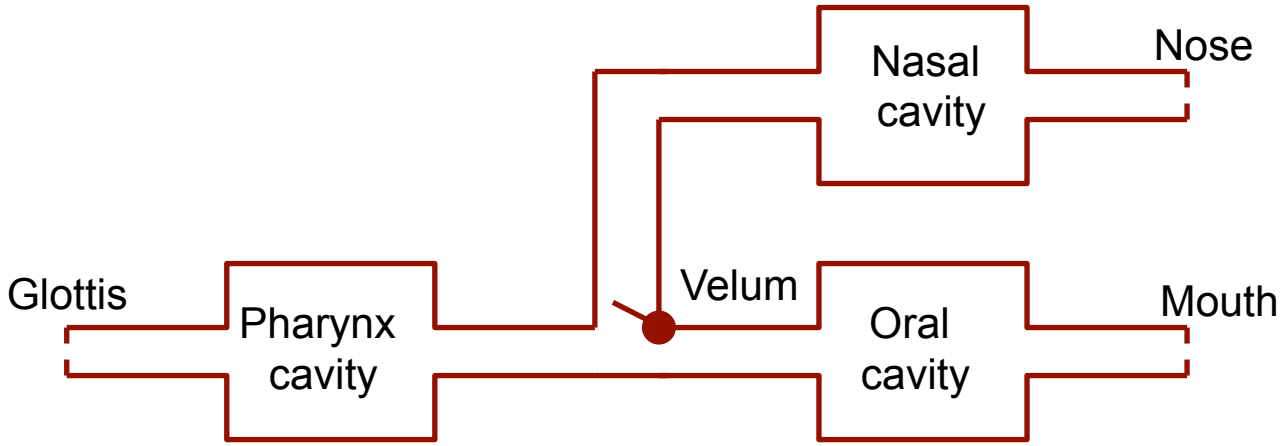
High Level Model of Source Signal

- High level model of source signal is much more simple and requires substantially fewer parameters - period T, amplitude A
- However, the produced waveform of excitation source is much simplified and can only create the general shape of the source signal
- Exhibits many smaller errors - no noise-like component



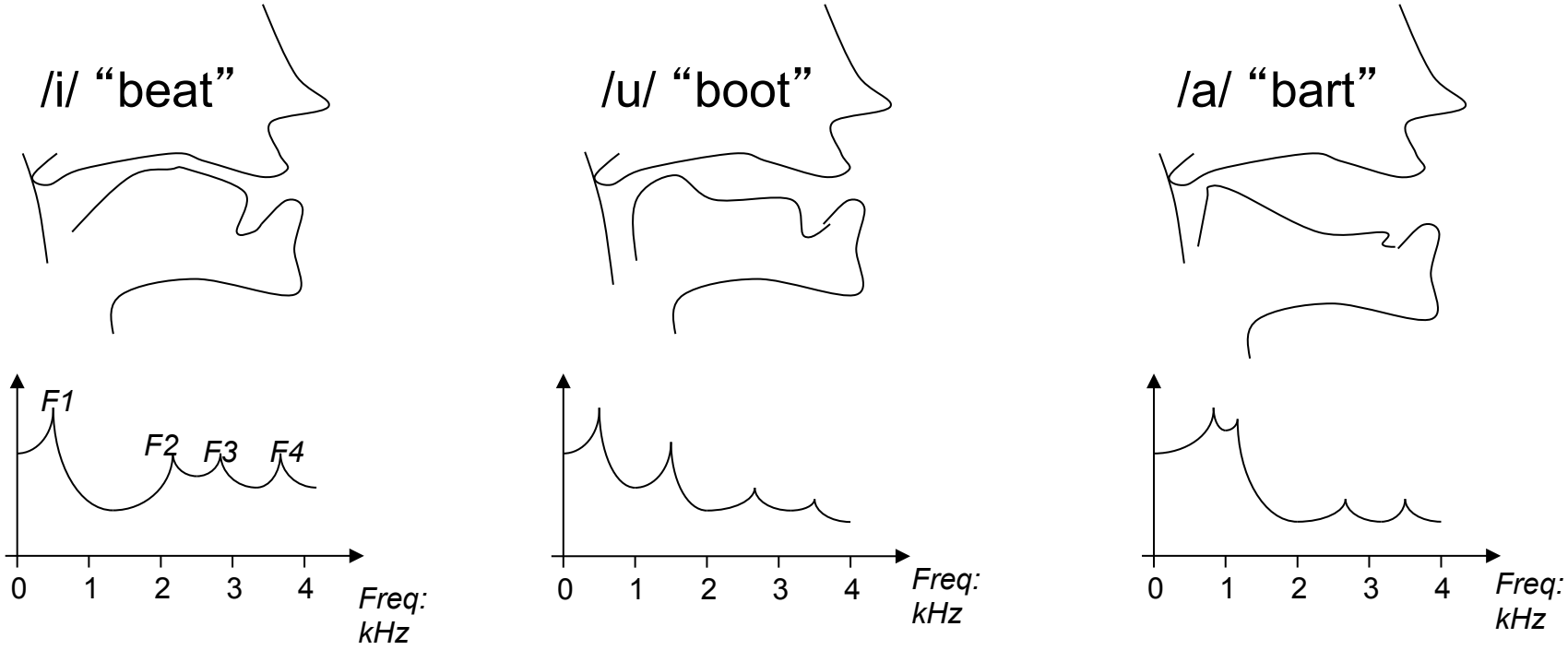
Vocal Tract Filter Modelling

- Vocal tract comprises a series of resonant cavities - pharynx cavity, nasal cavity and oral cavity
- A valve (velum) opens and closes the nasal cavity
- Size of cavities is mainly determined by tongue position in the mouth



Vocal Tract Filter Modelling

Changes in the shape of the vocal tract causes the changes of the resonant frequencies of the vocal tract which produces different speech sounds

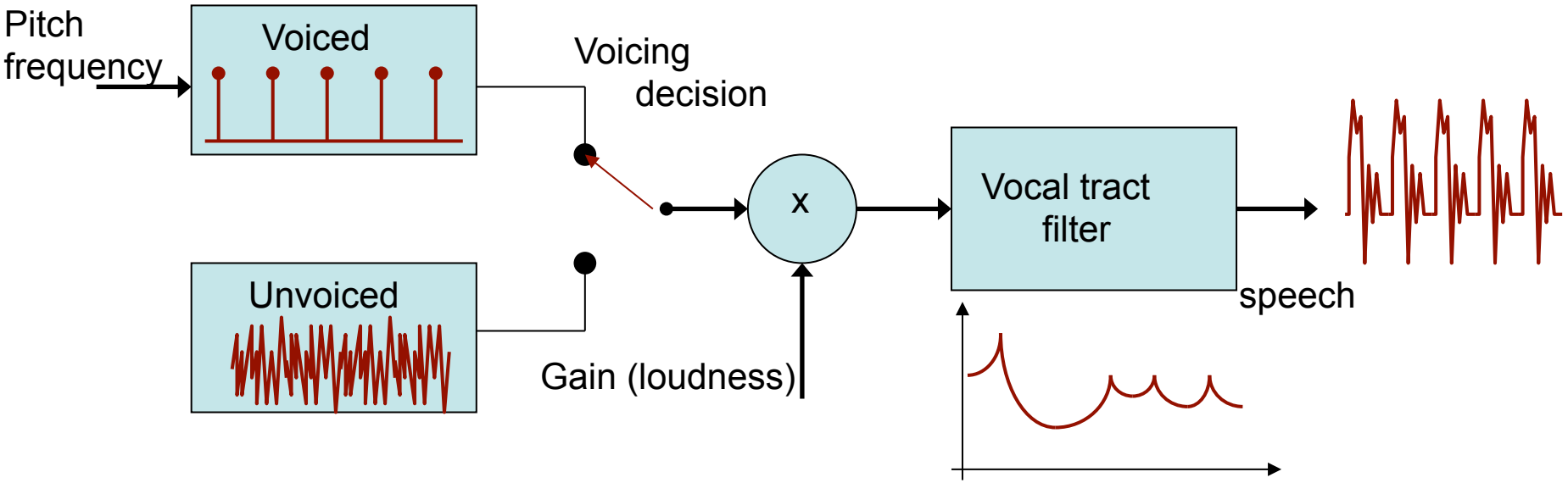


Tongue divides vocal tract into two cavities. Rear cavity determines first formant frequency, F_1. Front cavity determines second formant frequency, F_2.

Larger cavity means lower formant frequency

Source-Filter Model of Speech

Combining the source and filter model leads to a speech production model which can reproduce both voiced and unvoiced speech



Parameters of model:

Voiced/unvoiced decision

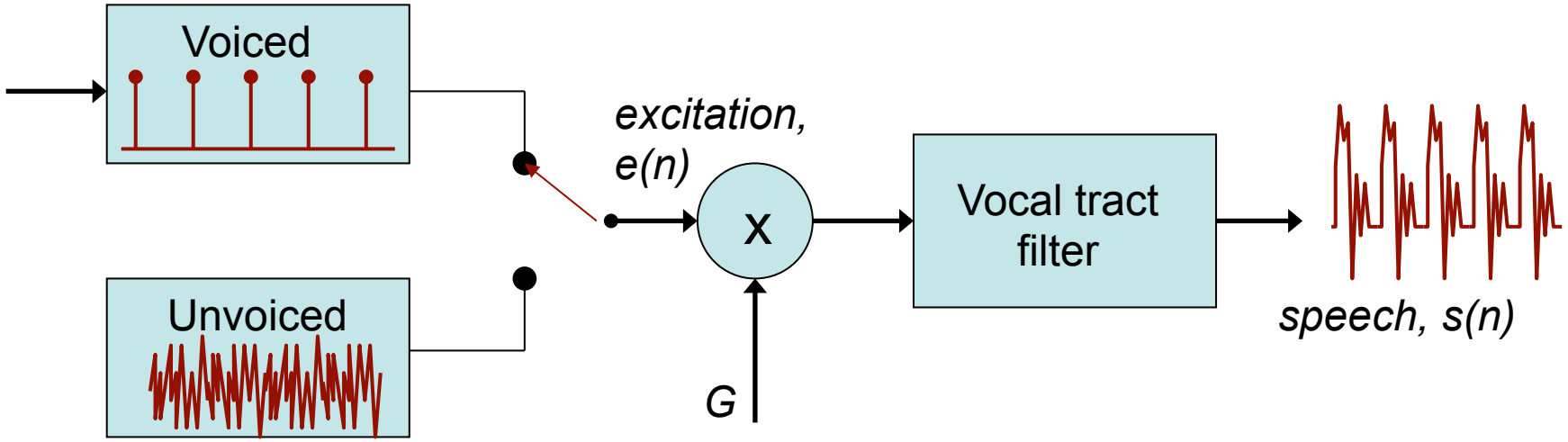
Pitch frequency

Gain (loudness)

Vocal tract shape (frequency response)

excitation

Calculating Model Parameters



Vocal tract is modelled by a digital filter - *all pole model*,

$$H(z) = \frac{1}{1 - \sum_{k=1}^{P} a_k z^{-k}}$$

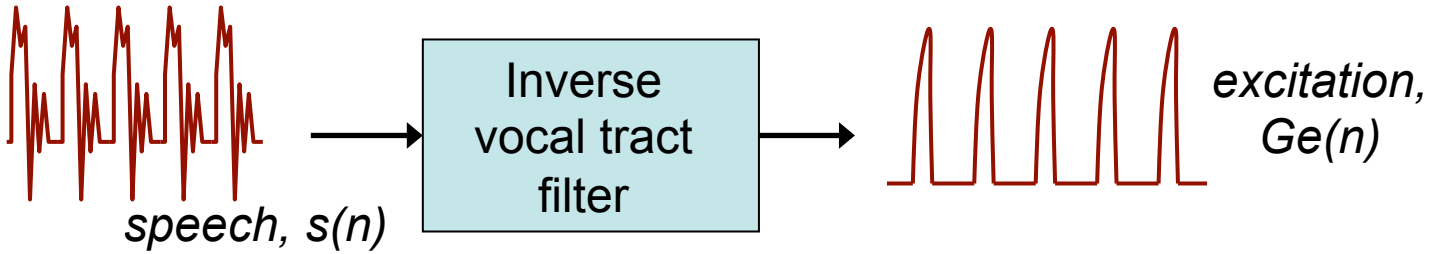
Coefficients, a_k, determine the frequency characteristics of the vocal tract model. Changing the a_k*s* allows the formant frequency, amplitude and bandwidth to vary.

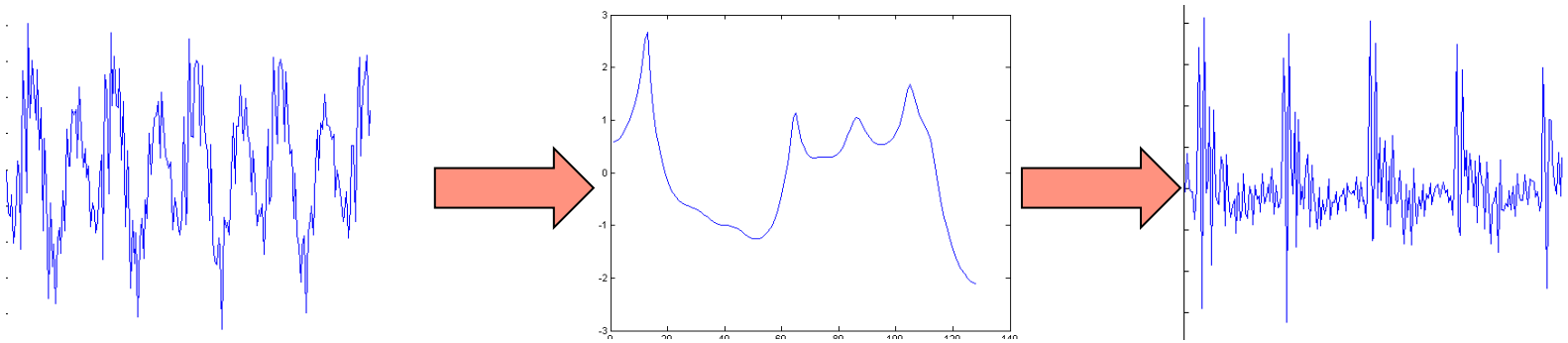
P is the order of the filter and determines how many formants can be modelled - *P/2* roughly indicates the number of formants.

Calculating Model Parameters

Coefficients, a_k, are calculated from the speech signal, *s(n)* by using Linear Prediction – Linear Predictive Coding (LPC).

Excitation signal *e(n)* is found as a residual of LP.

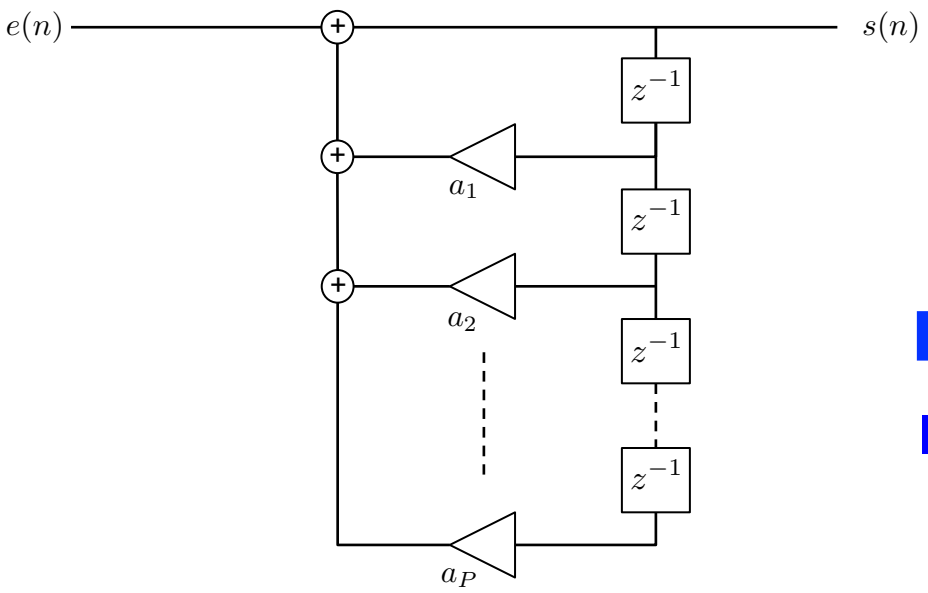




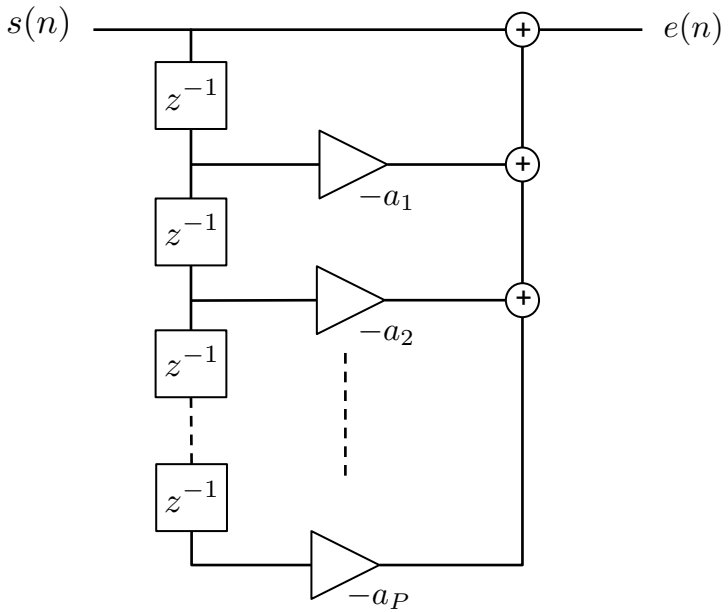
Calculating Model Parameters

$$S(z) = \frac{1}{1 - \sum_{k=1}^{P} a_k z^{-k}} E(z)$$

$$E(z) = S(z) - \sum_{k=1}^{P} a_k z^{-k} S(z)$$



Inverse



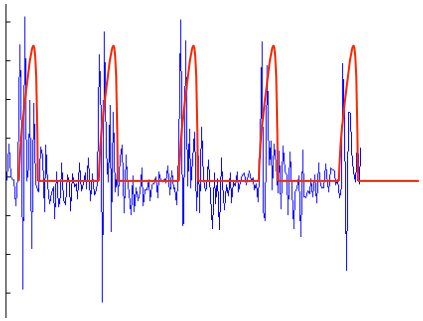
All pole model

Linear Prediction

Coefficients of the Vocal tract filter: Applying LP to the speech signal
Excitation: Residual of the prediction

Practical Use of Source-Filter Model

- Source-filter model provides a useful way to compress speech for communication and storage purposes ➔ vocoder
- Speech is windowed into 16 ms frames for analysis because speech is assumed to be stationary over this period, i.e. 62.5 Hz.
- Typically, for 10th order vocal tract filter, use 40 bits/frame
- Pitch information requires a few bits per frame, typically 4 bits
- Resultant bit rate = 62.5 x (40 + 4) = 2.75kbit/s
- Resulting quality tends to be poor mainly due to simplification of excitation signal - replacing true excitation with series of artificial pulse
- A vocoder standard is US DoD Federal Standard LPC-10, or as known now FS-1015 (slightly enhanced) - bit rate of 2.4kbit/s and MOS of 2.3



Original

Vocoder

Second Speech Model- Sinusoidal Model

- Sinusoidal analysis shows that a signal can be represented in terms of a combination of different sinusoids
- The sinusoidal model exploits this fact to allow a waveform to be synthesised from a sum of sinusoids
- This can be applied to speech, music and other signals with a harmonic structure
- The basic synthesised signal, *s(n),* takes the form

$$s(n) = A_1 \cos(\omega_1 n + \theta_1) + A_2 \cos(\omega_2 n + \theta_2) + \ldots + A_L \cos(\omega_L n + \theta_L)$$

$$s(n) = \sum_{l=1}^{L} A_l \cos(\omega_l n + \theta_l)$$

A_l *is amplitude*

ω_l *is frequency*

θ_l *is phase offset*

L *is the number of sinusoids*

Second Speech Model- Sinusoidal Model

To synthesise or encode real waveforms, the challenge is how to find the parameters of the sinusoidal model

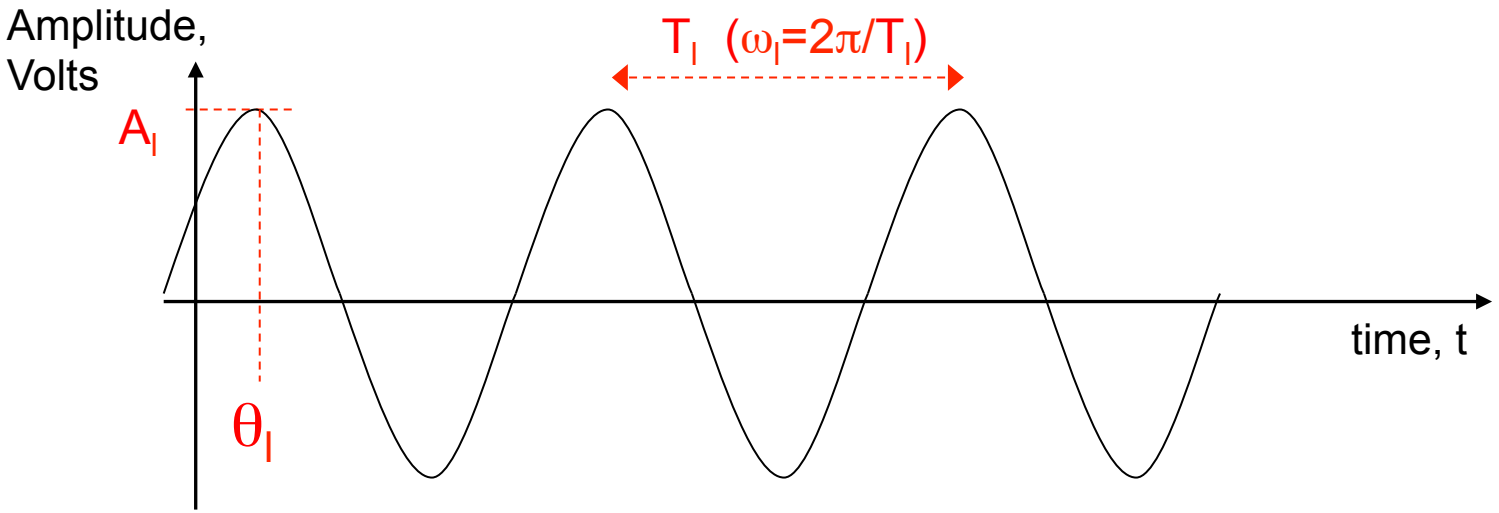
L - the number of sinusoids

A_l - the amplitude of each sinusoid

ω_l - the frequency of each sinusoid

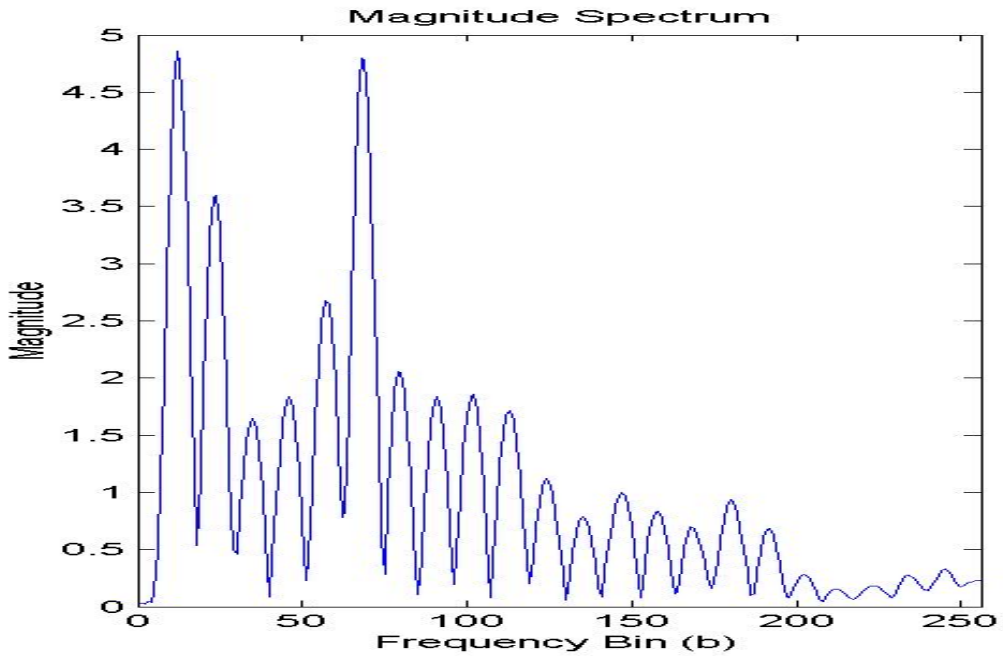
θ_l - the phase of each sinusoid

$$s(n) = \sum_{l=1}^{L} A_l \cos(\omega_l n + \theta_l)$$



Sinusoidal Model- Parameter Estimation

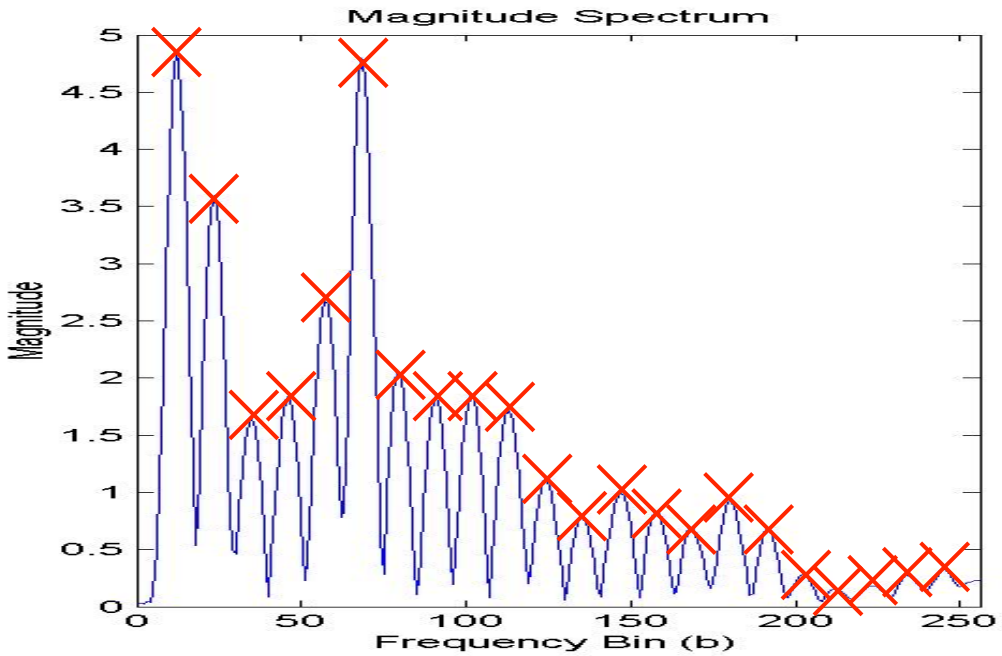
Parameter estimation begins by converting time-domain waveform into amplitude and phase spectrum



Identify peaks in the resulting amplitude spectrum to find the frequency of sinusoids for the sinusoidal model

Sinusoidal Model- Parameter Estimation

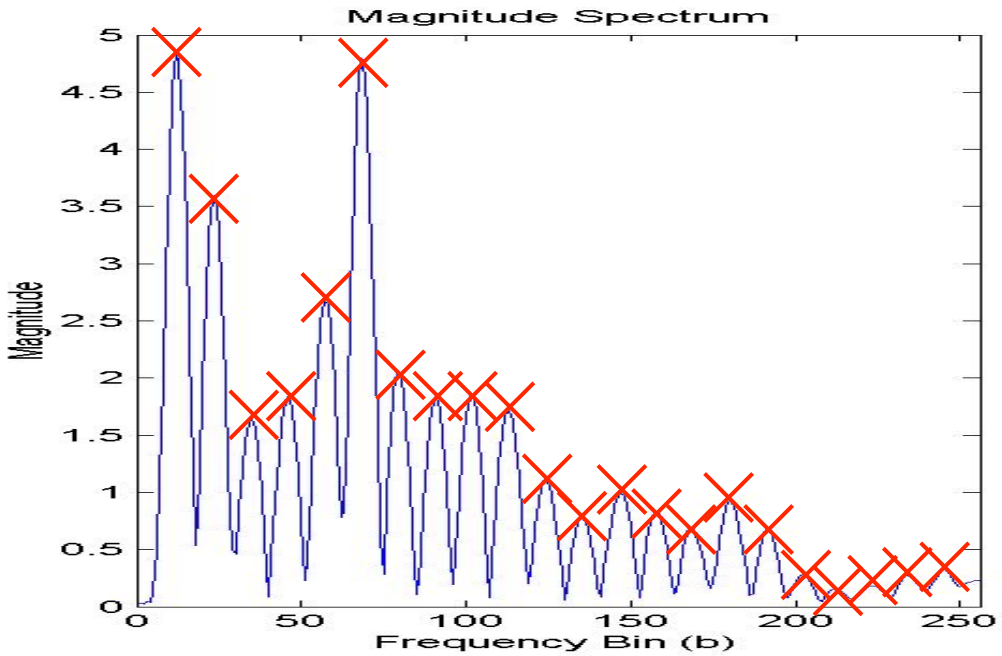
Parameter estimation begins by converting time-domain waveform into amplitude and phase spectrum



Identify peaks in the resulting amplitude spectrum to find the frequency of sinusoids for the sinusoidal model

Sinusoidal Model- Parameter Estimation

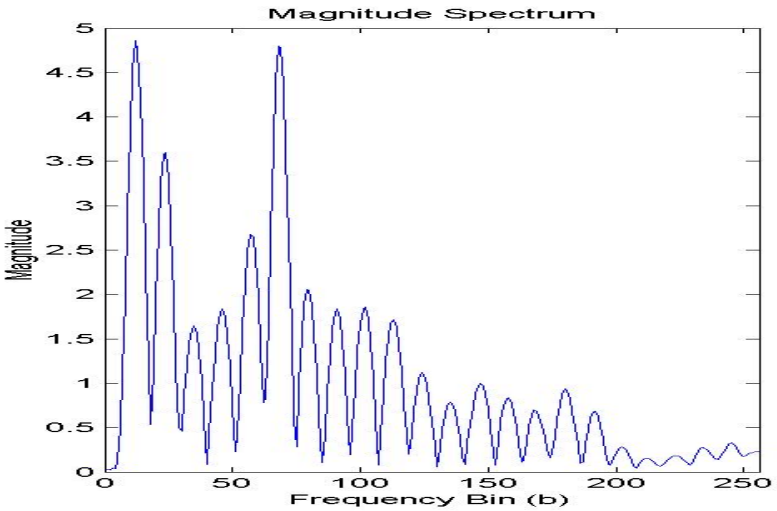
From knowledge of the location of sinusoids, determine the number of sinusoids and their frequency, amplitude and phase



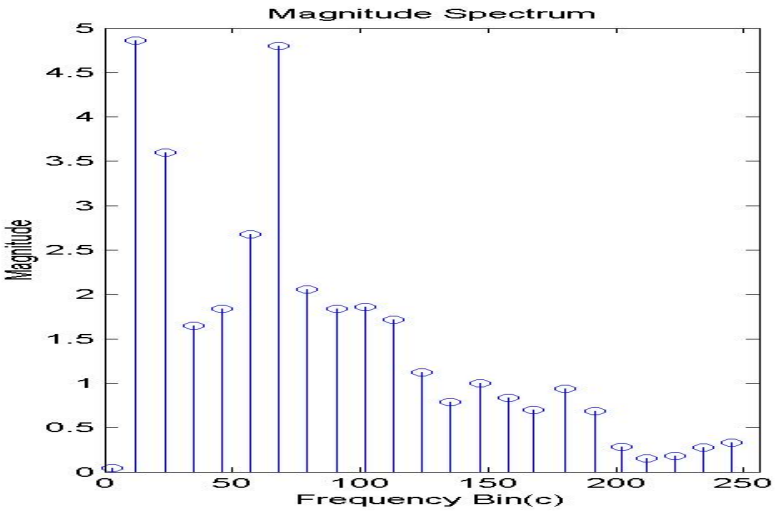
Sinusoidal Model- Parameter Estimation

Using the set of parameters, a sinusoidal model can synthesise a speech signal

$$s(n) = \sum_{l=1}^{L} A_l \cos(\omega_l n + \theta_l)$$



Original spectrum

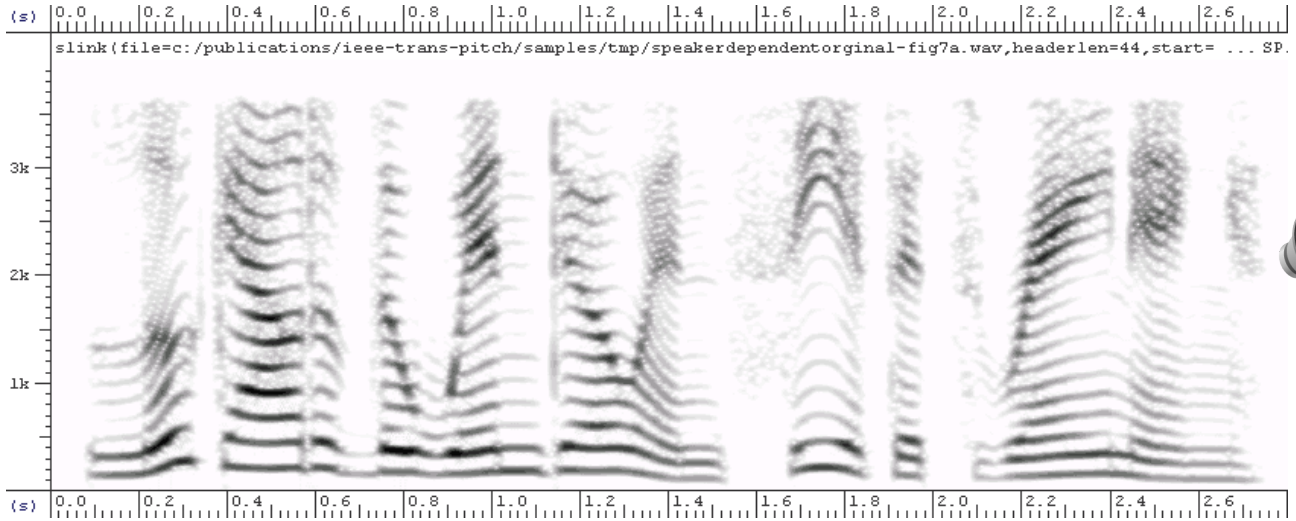


Sinusoidal model spectrum

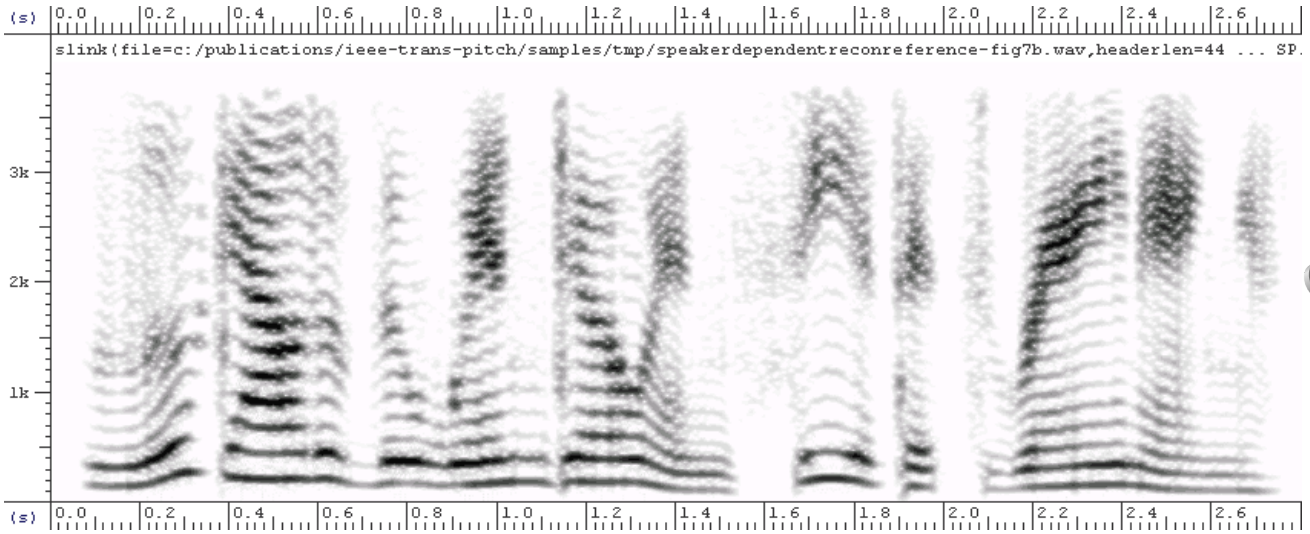
Examining the amplitude spectrum of the synthesised signal turns out that it's made up of just sinusoids at spectral peaks in the original signal

No energy exists in between the sinusoids - but for many signals, this still provides an adequate quality and intelligibility of the speech signal

Sinusoidal Model - Examples

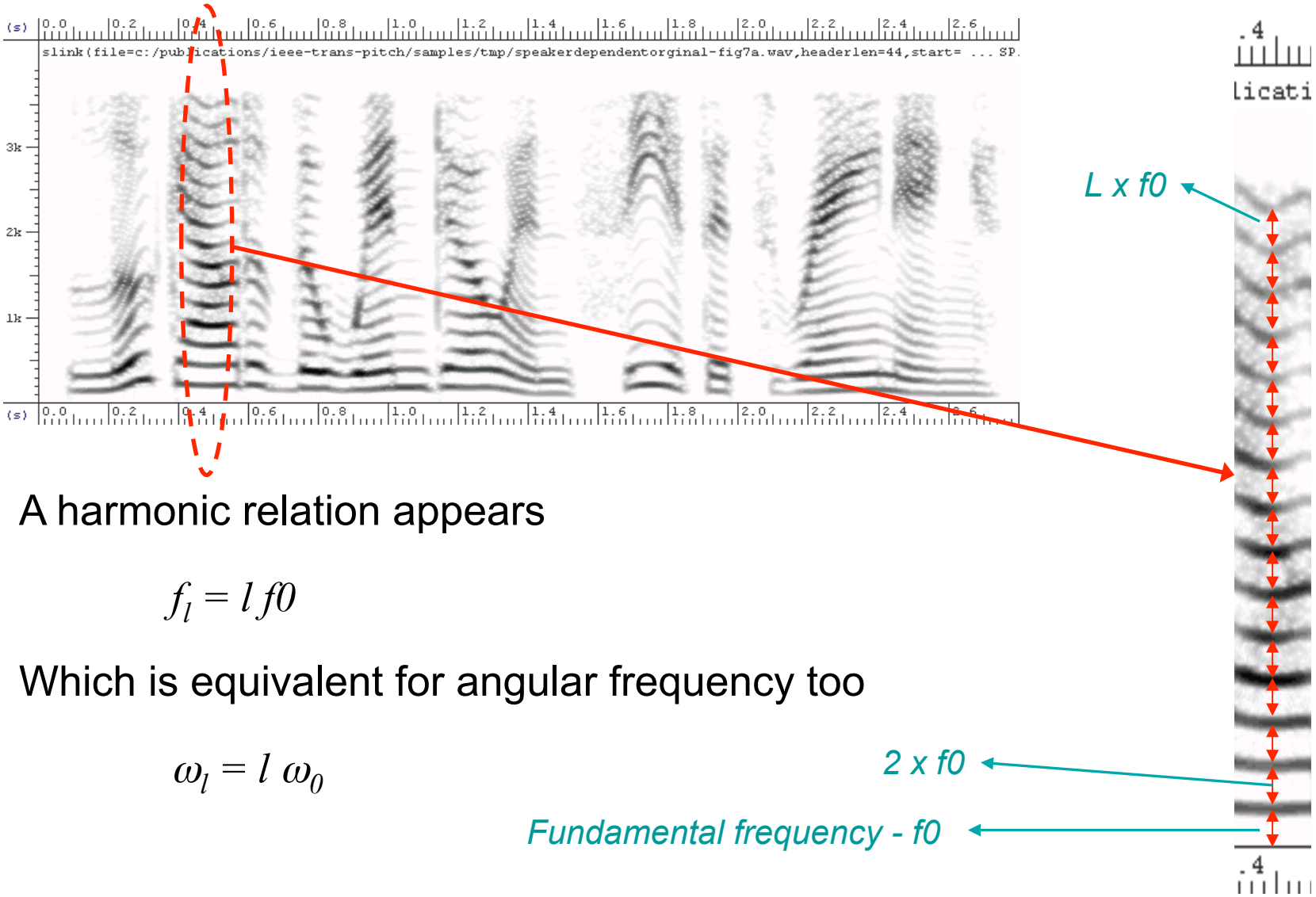


Natural Speech



Sinusoidal Model

Sinusoidal Model- Harmonic Approximation



A harmonic relation appears

$$f_l = l\, f0$$

Which is equivalent for angular frequency too

$$\omega_l = l\, \omega_0$$

Sinusoidal Model- Harmonic Approximation

This harmonic approximation can be used to simplify the sinusoidal model and reduce the number of parameters needed

Using the approximation

$$\omega_l = l\ \omega_0$$

The sinusoidal model can be simplified

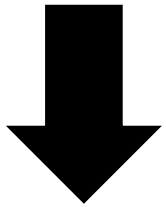
$$s(n) = \sum_{l=1}^{L} A_l \cos(\omega_l n + \theta_l)$$

$$s(n) = \sum_{l=1}^{L} A_l \cos(l\omega_0 n + \theta_l)$$

Now only require knowledge of ω_0, rather than ω_1, ω_2, …, ω_L

Sinusoidal Model - Noise

Sinusoidal model is ideal for modelling strongly harmonic sounds - but what about more noise-like sounds and also the energy that exists in between sinusoid frequencies



An extension of the sinusoidal model exists which is known as the *harmonic plus noise model (HNM)* - this combines sinusoidal components with a noise signal, *d(n)*, to allow both harmonic (sinusoidal) components and a noise component

$$s(n) = \sum_{l=1}^{L} A_l \cos(l\omega_0 n + \theta_l) + d(n)$$

Summary

- This lecture has considered model-based synthesis of signals - concentrated mainly on speech signals although other signals (e.g. music) can also be modelled
- Low-level modelling tends to model the physics of the generation process - if done accurately, this can lead to very good synthesis but this is very difficult to be implemented
- Higher-level modelling is more practical - from a computation and estimation perspective and generally leads to higher quality synthesis
- For higher level modelling we have examined source-filter model and sinusoidal model and seen how they can be applied to speech signals
- Within the source-filter model, we have looked at both a low level and high level modelling of the excitation (source) signal