Speech Synthesis by Articulatory Models

Advanced Signal Processing Seminar

Helmuth Ploner-Bernard

hamlet@sbox.TUGraz.at

Speech Communication and Signal Processing Laboratory

Graz University of Technology

Overview

- Introduction
- Articulators and (Co-)Articulation
- Sound Wave Propagation in the Vocal Tract
- The Acoustic Tube Model
- Articulatory Models
- The “Inverse” Problem of Parameter Estimation

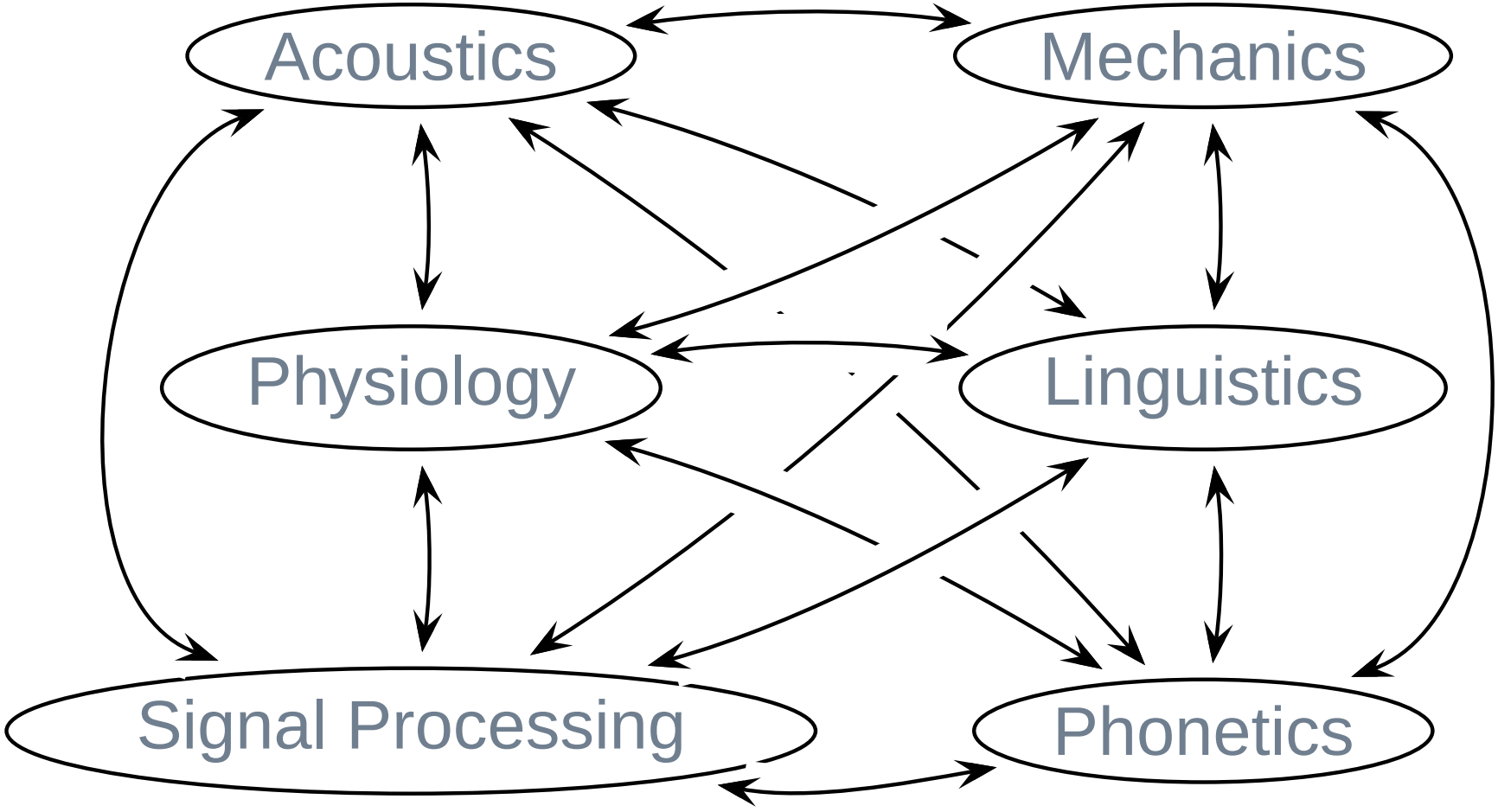
We are . . . here!

- Introduction
- Articulators and (Co-)Articulation
- Sound Wave Propagation in the Vocal Tract
- The Acoustic Tube Model
- Articulatory Models
- The “Inverse” Problem of Parameter Estimation

Introduction – Articulatory Models

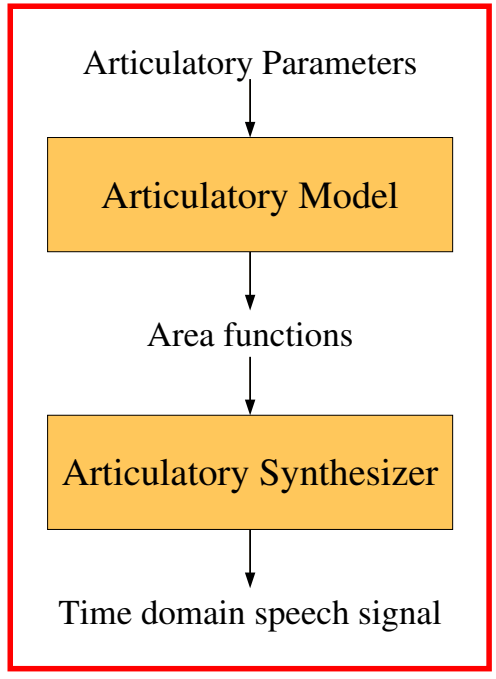
- Fields of application
  - (Most natural sounding) Speech synthesis
  - Low bit-rate coding
  - Speech recognition
  - Understanding of human speech production
- Attempt to describe the actual speech production mechanisms
  - Set of slowly time-varying *physiological* parameters

Introduction – Knowledge of …



Introduction

How does speech synthesis with articulatory models work?

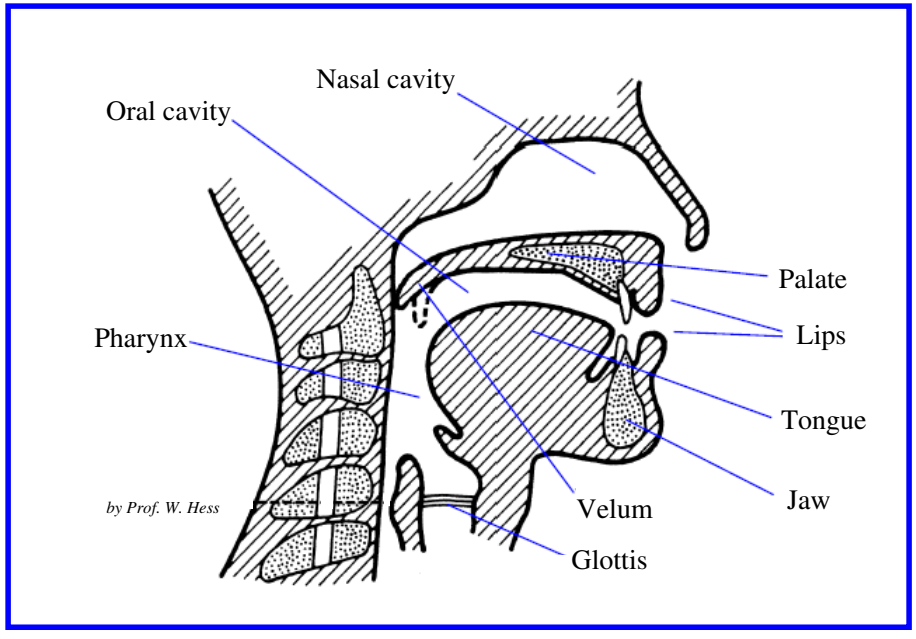


- Source-tract interaction can be accounted for quite easily

We are . . . here!

- Introduction
- Articulators and (Co-)Articulation
- Sound Wave Propagation in the Vocal Tract
- The Acoustic Tube Model
- Articulatory Models
- The “Inverse” Problem of Parameter Estimation

Articulators (Speech-Organs)



- Source-filter model
  - Excitation
  - Vocal tract does the filtering
- Acoustic differences between sounds from different
  - manners and
  - places

  of articulation

(Co-)Articulation

- *Articulation* of an (isolated) phoneme involves
  - “Critical” articulators, essential for correct production
  - “Non-critical” articulators, place and manner unspecified
- *Co-articulation* in fluent speech
  - Target positions of articulators strongly affected by each other
  - Dependent on phonetic context

(Co-)Articulation

- Associate priorities with parameters of articulatory model and let your controller exploit them
- Incorporate realistic physiological and dynamic constraints (cf. functional models)
- $\longrightarrow$ more natural sounding speech

We are . . . here!

- Introduction
- Articulators and (Co-)Articulation
- Sound Wave Propagation in the Vocal Tract
- The Acoustic Tube Model
- Articulatory Models
- The “Inverse” Problem of Parameter Estimation

Wave Propagation

Acoustic theory of speech production by FANT

- Vocal tract $\longrightarrow$ acoustic tube
- Infinitely high sound impedance, rigid walls
- Lossless planar wave propagation governed by WEBSTER's horn equation:

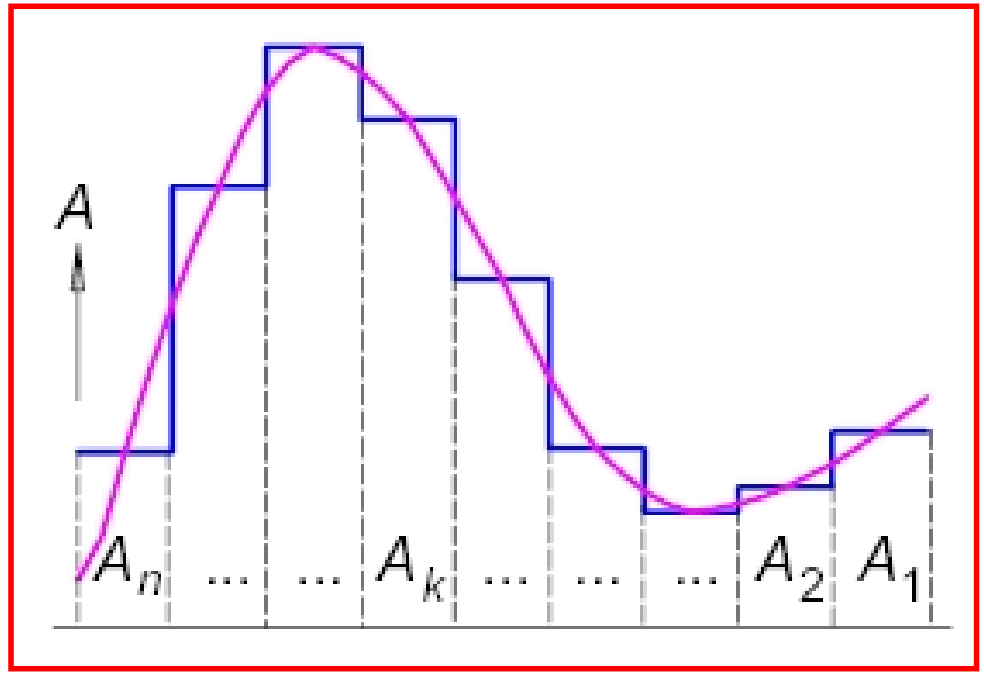
$$\frac{\partial^2 v}{\partial x^2} + \frac{1}{A}\frac{dA}{dx}\frac{\partial v}{\partial x} = \frac{1}{c^2}\frac{\partial^2 v}{\partial t^2}$$

x . . . Direction of traveling wave	t . . . Time
v . . . Sound particle velocity	c . . . Velocity of wave propagation

A . . . **Area function, wait until next slide**

Wave Propagation – Area function

- Cross-sectional areas as a function of position between glottis and lips
- Time-varying shape, depending on specific positions of articulators



(figure by *Prof. W. Hess*)

Wave Propagation – Neutral vowel

- /ə/: assume $A(x,t) \equiv \text{const}\ \forall x,t$
- Cylindrical acoustic tube
- Resonance frequencies f_k at

$$f_k = \frac{(2k-1)c}{4l}, \quad k = 1, 2, \ldots .$$

- l is the total length of the vocal tract
- For a male speaker $f_k \approx 500, 1500, \ldots$ Hz
- Comparable f_k's for bent pipes

Wave Propagation

- Horn equation cannot be solved for arbitrary area function
- Changes in vocal tract shape lead to changes in Eigenfrequencies

Wave Propagation

- Horn equation cannot be solved for arbitrary area function
- Changes in vocal tract shape lead to changes in Eigenfrequencies
- At $f = 3.5\,\text{kHz}$ first cross-modes in vocal tract
- ☺ most of the energy in speech signals concentrated in region below this frequency

Wave Propagation

- Horn equation cannot be solved for arbitrary area function
- Changes in vocal tract shape lead to changes in Eigenfrequencies
- At $f = 3.5\,\text{kHz}$ first cross-modes in vocal tract
- ☺ most of the energy in speech signals concentrated in region below this frequency

---

- Nasal cavity separate tube of fixed length parallel to the vocal tract

We are . . . here!

- Introduction
- Articulators and (Co-)Articulation
- Sound Wave Propagation in the Vocal Tract
- The Acoustic Tube Model
- Articulatory Models
- The “Inverse” Problem of Parameter Estimation

The Acoustic Tube Model

- Starting point: Short acoustic tube of *constant* cross-sectional area
- The horn equation

$$\frac{\partial^2 v}{\partial x^2} + \frac{1}{A}\frac{dA}{dx}\frac{\partial v}{\partial x} = \frac{1}{c^2}\frac{\partial^2 v}{\partial t^2}$$

The Acoustic Tube Model

- Starting point: Short acoustic tube of *constant* cross-sectional area
- The horn equation

$$\frac{\partial^2 v}{\partial x^2} + \frac{1}{A}\frac{dA}{dx}\frac{\partial v}{\partial x} = \frac{1}{c^2}\frac{\partial^2 v}{\partial t^2}$$

can be simplified to the form

$$\frac{\partial^2 v}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2 v}{\partial t^2}$$

The Acoustic Tube Model

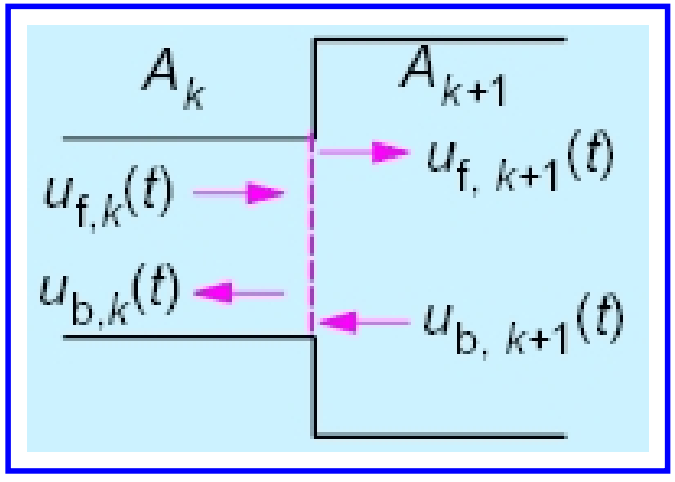
- Equation has a general solution of the form

$$u(x,t) = u_{\mathsf{f}}\left(t - \frac{x}{c}\right) - u_{\mathsf{b}}\left(t + \frac{x}{c}\right)$$

  where $u = vA$ is the volume velocity

- Combination of two waves traveling in opposite directions
  - forward
  - backward

The Acoustic Tube Model



(figure by *Prof. W. Hess*)

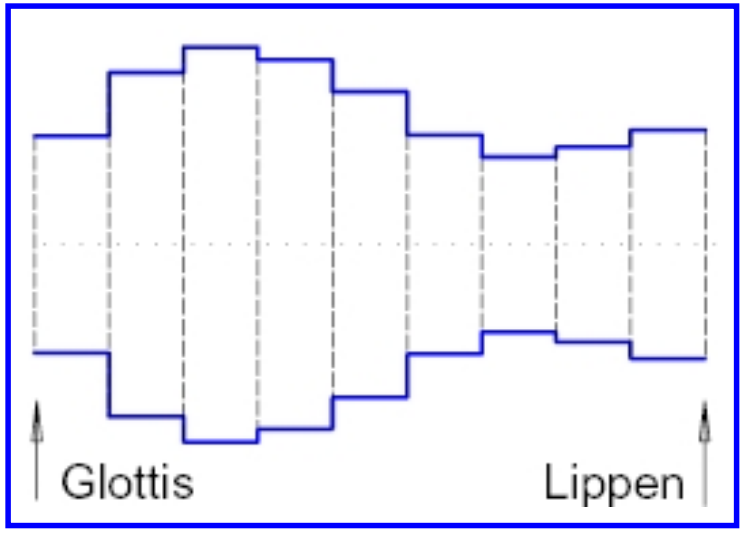
- Approximate continuous area function A by concatenation of homogeneous acoustic tubes
- At junctions, part of the traveling wave is reflected

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

- r_k reflection coefficient

- FANT chooses 2-4 sections of *variable* length

The Acoustic Tube Model

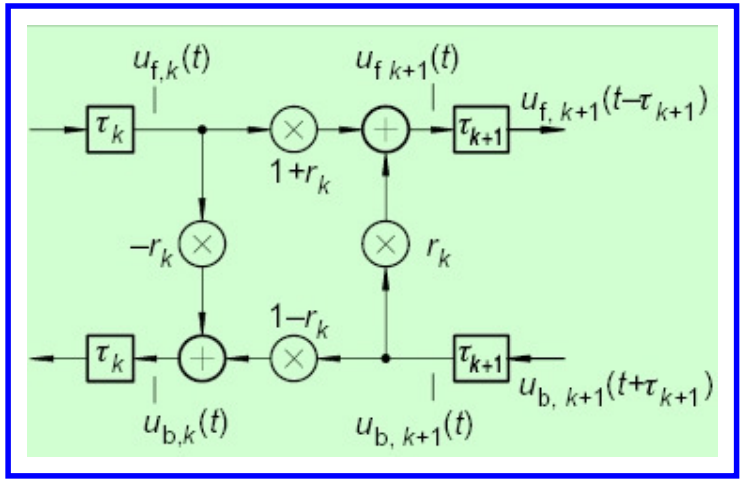


(figure by *Prof. W. Hess*)

- Toward a digital implementation, convenient to take equidistant samples of $A(x)$
- Delay through each segment

$$\tau = \frac{\Delta x}{c}$$

The Acoustic Tube Model



(figure by *Prof. W. Hess*)

- KELLY-LOCHBAUM structure
- About 20 segments
- Idealized, lossless model

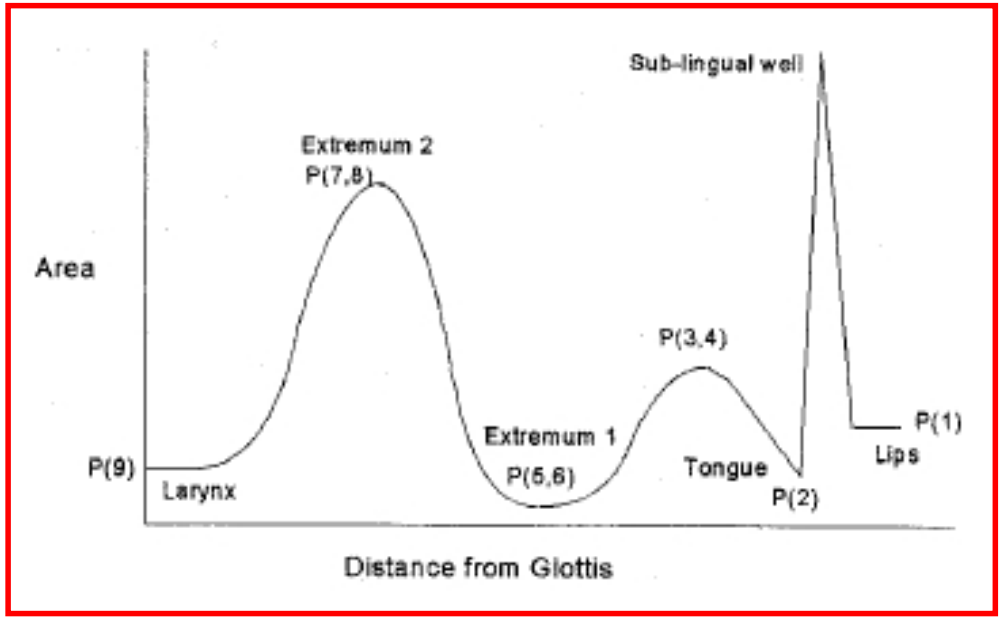
The Acoustic Tube Model – Losses

- In reality, losses occur due to
  - Resonances of yielding walls
  - Viscous and thermal losses along the path of propagation $\longrightarrow$ add multipliers
  - Radiation at the lips $\longrightarrow$ insert additional segment in front of the lips
- Freeze delay τ to any given sampling interval
  - Wave digital filters

We are . . . here!

- Introduction
- Articulators and (Co-)Articulation
- Sound Wave Propagation in the Vocal Tract
- The Acoustic Tube Model
- Articulatory Models
- The “Inverse” Problem of Parameter Estimation

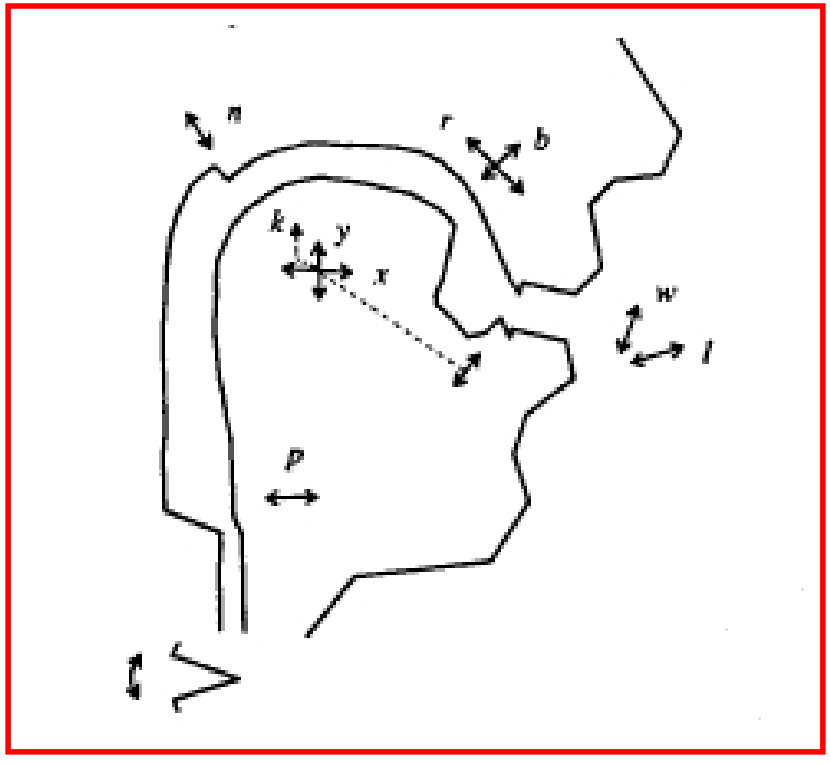
Articulatory Models – Static



- Vocal tract described in terms of area functions
- Motion is succes-sion of stationary shapes

Example shows nine-parameter model

Articulatory Models – Dynamic



- Set up equation of motion for every articulator
- Articulators are elastic
- Have masses and an inertia
- Constraints regarding positions, velocities and accelerations

COKER's model

We are . . . here!

- Introduction
- Articulators and (Co-)Articulation
- Sound Wave Propagation in the Vocal Tract
- The Acoustic Tube Model
- Articulatory Models
- The “Inverse” Problem of Parameter Estimation

Parameter Estimation (1)

- “Inverse” problem
- Acquire model parameters directly or indirectly from speech signal
- Most difficult
- Non-unique, i. e. more than one vocal tract shape can produce signal with identical spectrum

Parameter Estimation (2)

- Required:
  - Good acoustic matching
  - Smooth evolution of area functions or articulatory parameters
  - Anatomical feasibility
- Most methods are unable to determine vocal tract length

Parameter Estimation – MRI (1)

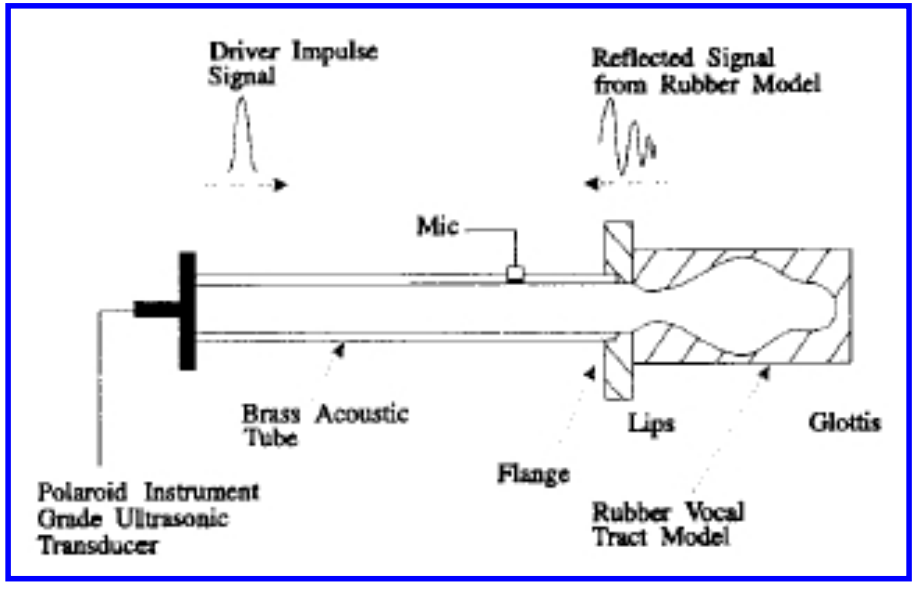
- Most intuitive way
- “Measure” vocal tract shape directly
- Several scans necessary for 3D-model (how can we represent /l/ with mid-sagittal area functions?)
- Much signal processing to be done here
- Costly, time consuming and noisy

Parameter Estimation – MRI (2)

Parameter Estimation – LPC

- Simple, cheap method
- Evaluate reflection coefficients from LEVINSON-DURBIN algorithm for Linear Predictive Coding
- Characterize an idealized acoustic tube model
- Obtained from real world lossy signals
- ↯ Inaccurate results

Parameter Estimation – Impedance



- Acoustic impedance measurement
- Special acoustic volume velocity impulse sent toward the lips
- Shaped in vocal tract, reflected at the closed glottis
- Cheap, fast, for many shapes
- What about the nasal cavity?
- How to account for losses

Parameter Estimation – ABS

- ABS: Analysis by Synthesis
- Method for automated parameter identification from natural utterances
- Algorithm:
  - Extract descriptive parameters from signal
  - Look up “best matching” articulatory parameters in codebook
  - Re-synthesize with articulatory parameter set
  - Compare re-synthesized signal to target speech signal (original)
  - Iteratively optimize parameters

Parameter Estimation – ABS

- Segmentation
  - Phoneme basis, variable length
  - Fixed frame lengths
- Time alignment, pitch synchronous analyses to avoid influence of glottal excitation
- Descriptive parameters
  - LPC-coefficients
  - Mel frequency cepstral coefficients
  - Coefficients of any spectral transformation

Parameter Estimation – ABS

- Remember: Mapping is non-unique
- Find other shapes of vocal tract according to a cost function
- Components of cost function
  - Distance between spectra
  - Smoothness of area function
  - Smooth evolution of parameters between adjacent frames
  - Signal energy
- Improvement: multi-frame optimization

Optional: Generation of the codebook

- Random sampling
  - Iterate through various configurations of articulatory parameters
  - Store along with their corresponding descriptive parameters
  - Huge amount of items
  - Unnecessary data not used in language or by a speaker
- “Inching” approach
  - Start out at extreme articulatory parameters
  - Interpolations on trajectories in articulatory space
  - Attention to sparsely populated areas

Summary

- Wave propagation in the vocal tract
- Area function responsible for different sounds
- Co-articulation with priority parameters
- Non-unique acoustic-to-articulatory mapping
- Tube model, KELLY-LOCHBAUM structure, WDF
- Static models, dynamic models
- Parameter estimation: MRI, LPC, Impedance measurement, ABS

References

- http://www.ikp.uni-bonn.de/dt/lehre/materialien/aap/aap_1f.pdf
- http://www.radiologyinfo.org/
- J.W. Devaney and C. C. Goodyear. A comparison of acoustic and magnetic resonance imaging techniques in the estimation of vocal tract area functions. International Symposium on Speech, Image Processing and Neural Networks, pages 575–578, April 1994.
- A. R. Greenwood and C. C. Goodyear. Articulatory speech synthesis using a parametric model and a polynomial mapping technique. International symposium on speech, image processing and neural networks, pages 595–598, April 1994
- S. Parthasarathy and C.H. Coker. Phoneme-level parametrization of speech using an articulatory model. International Conference on Acoustics, Speech and Signal Processing, pages 337–340, April 1990
- Peter Vary, Ulrich Heute, and Wolfgang Hess. Digitale Sprachsignalverarbeitung. B.G. Teubner Stuttgart, 1998

Thank you for your attention!

Have a look at the accompanying paper on the web!