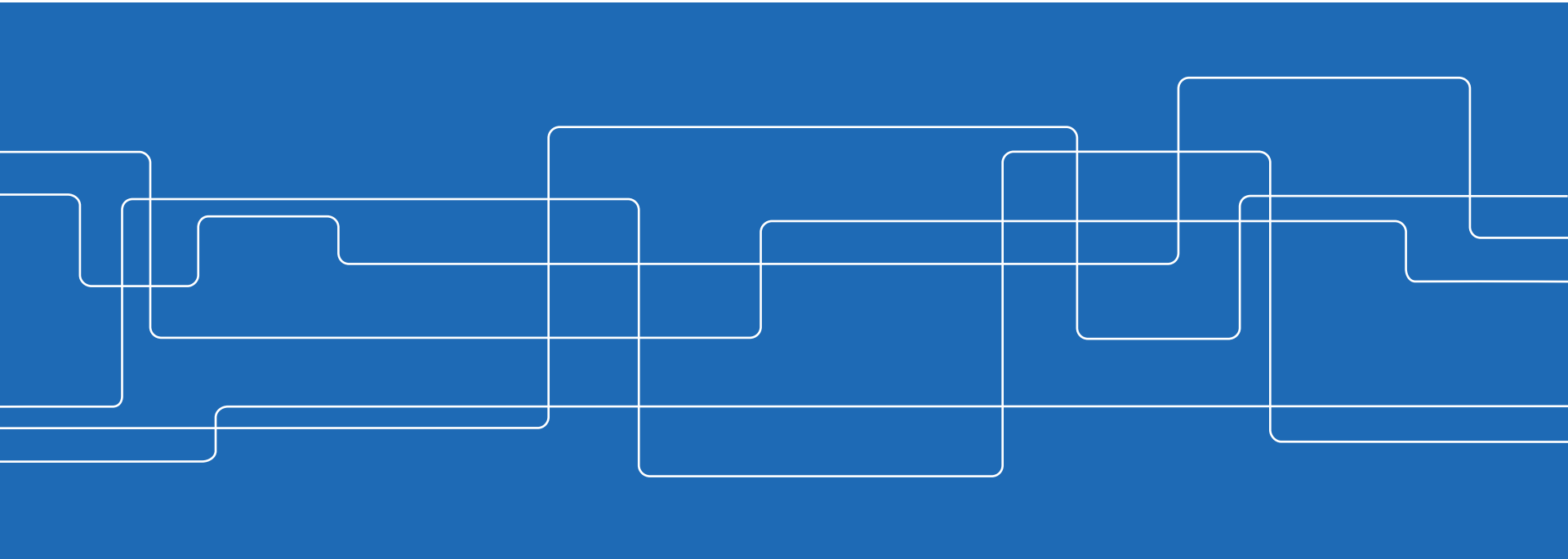
Speech Synthesis

Olov Engwall
Professor in Speech Communications

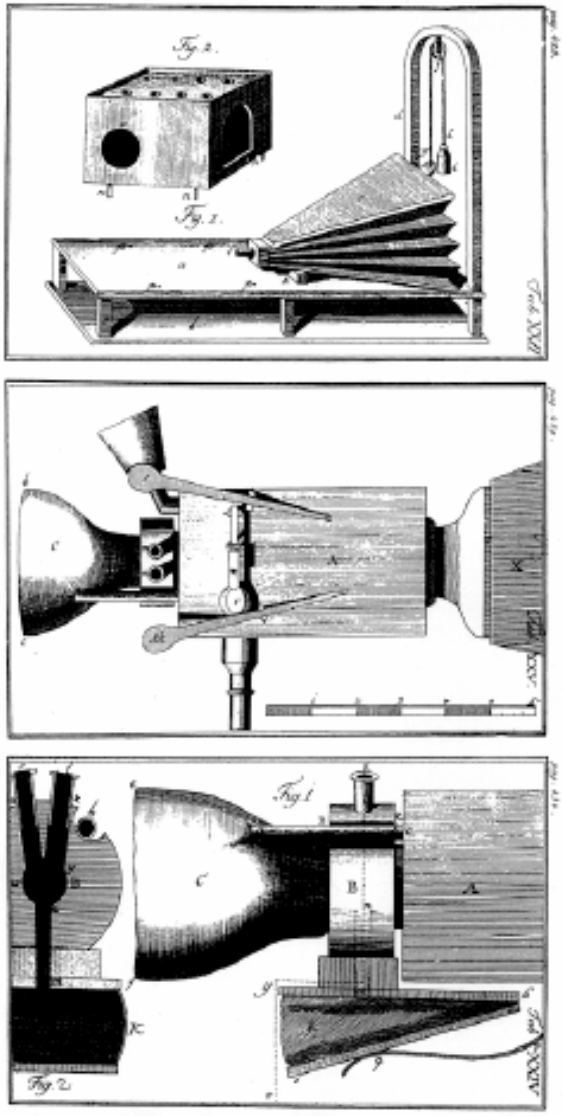


The dawn of speech synthesis

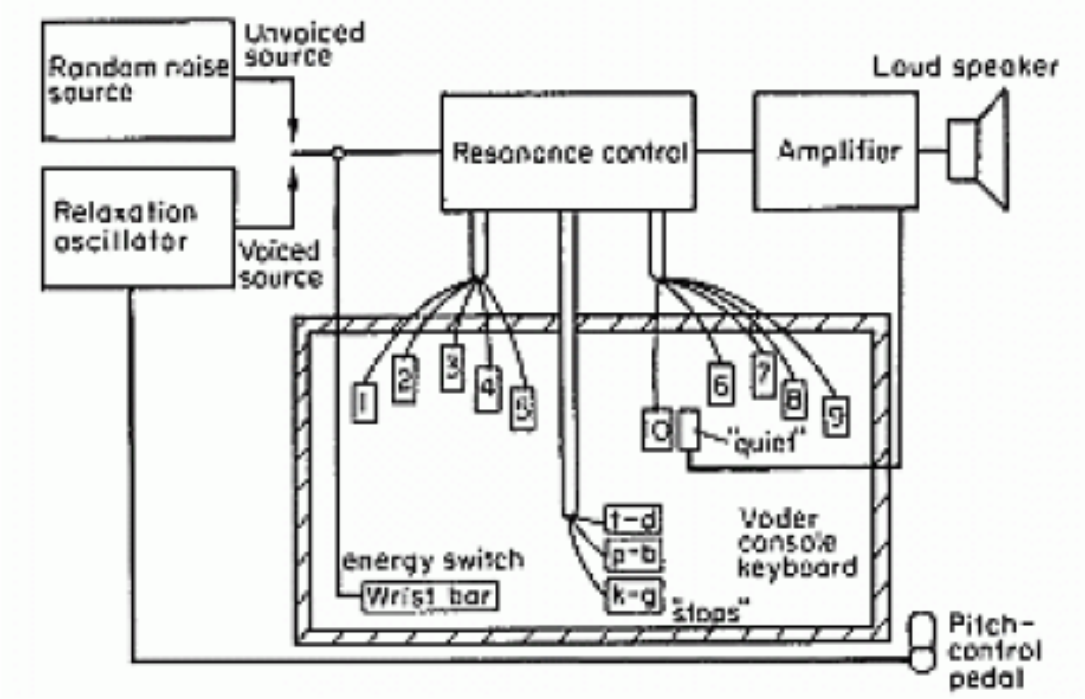
Wolfgang von Kempelen's book *Mechanismus der menschlichen Sprache nebst Beschreibung einer sprechenden Maschine (1791).*

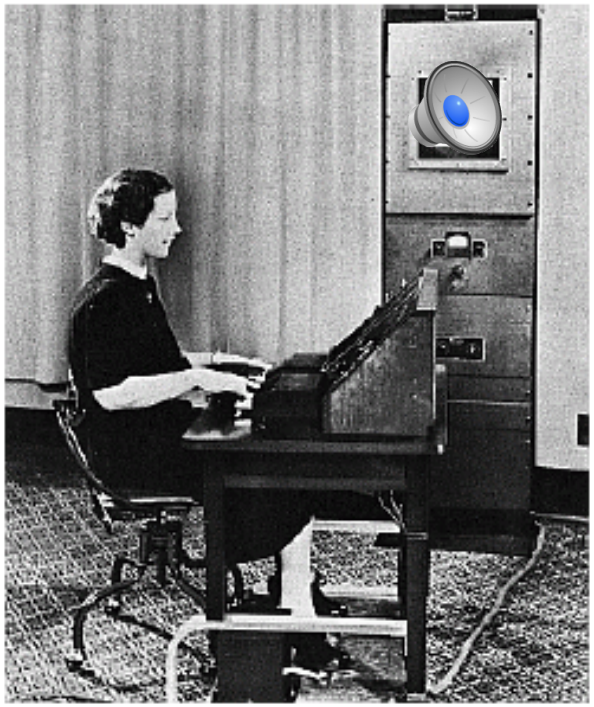
The essential parts
- pressure chamber = lungs,
- a vibrating reed = vocal cords,
- a leather tube = vocal tract.

The machine was
- hand operated
- could produce whole words and short phrases.



First electronic synthesis





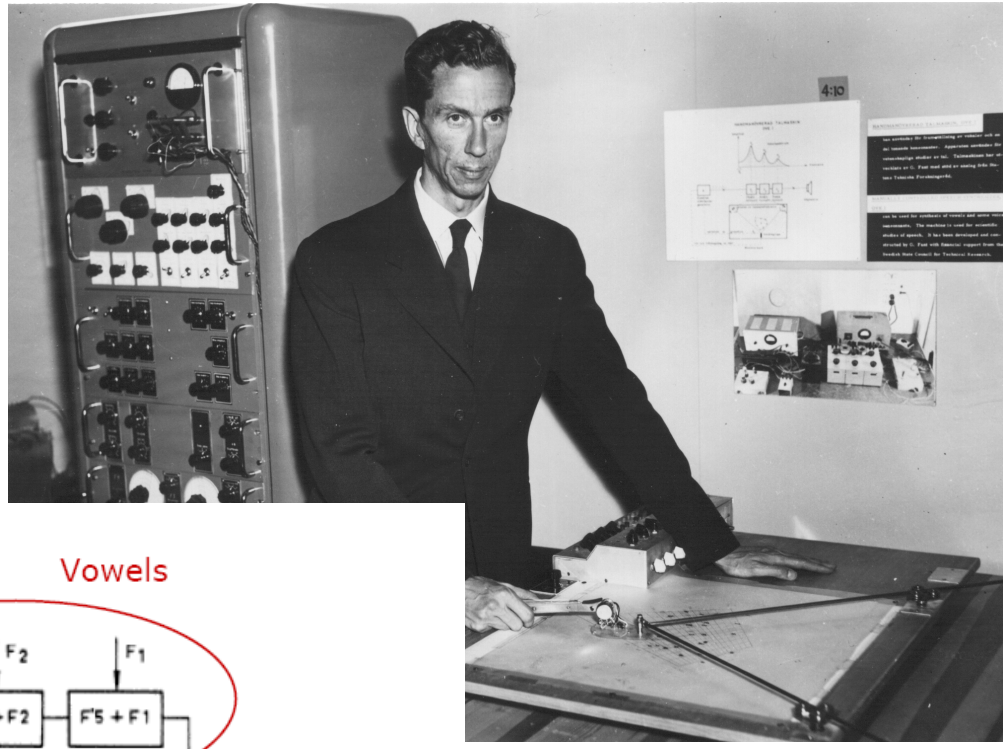
- Homer Dudley presented VODER (Voice Operating Demonstrator) at the World Fair in New York in 1939

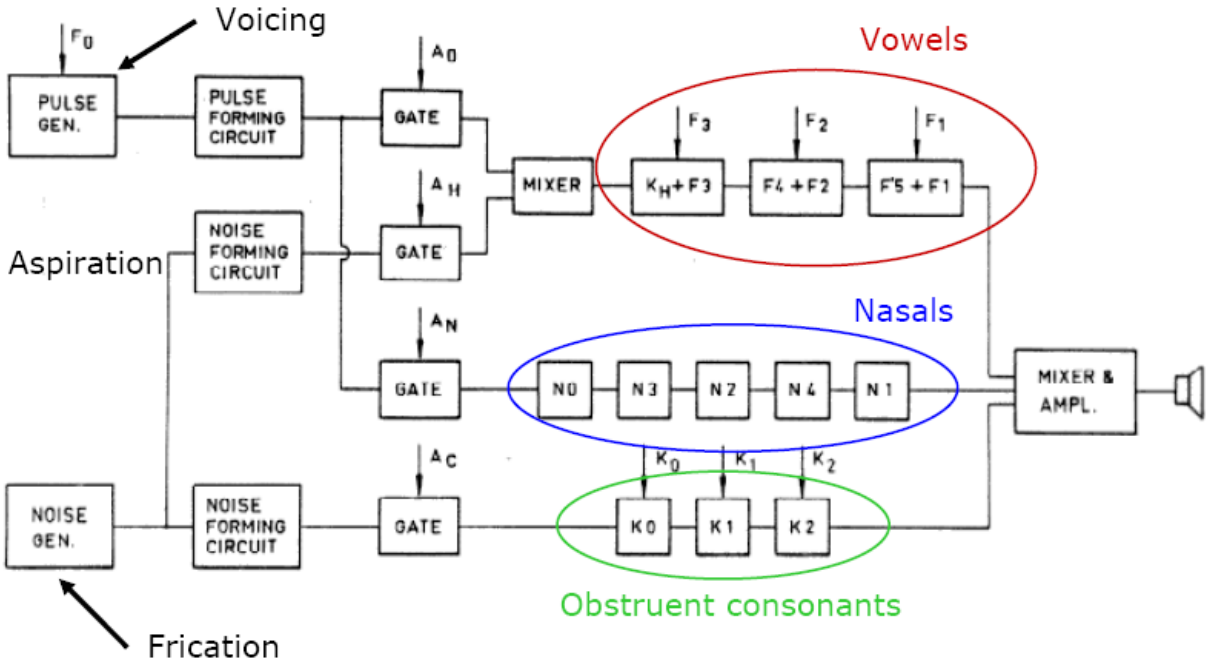
- The device was played like a musical instrument, with voicing/noise source on a foot pedal and ten bandpass filters.

OVE @ KTH

OVE I (1953)

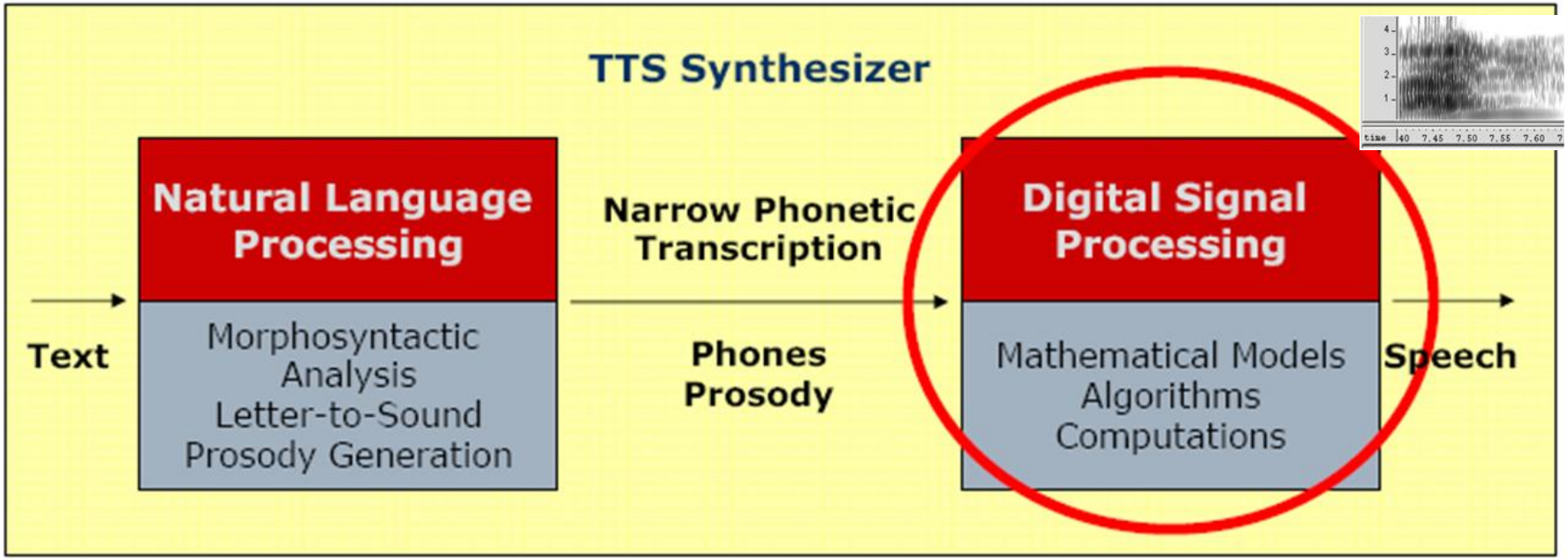
OVE II (1962)





Text-To-Speech synthesis (TTS)

Our focus in this course!

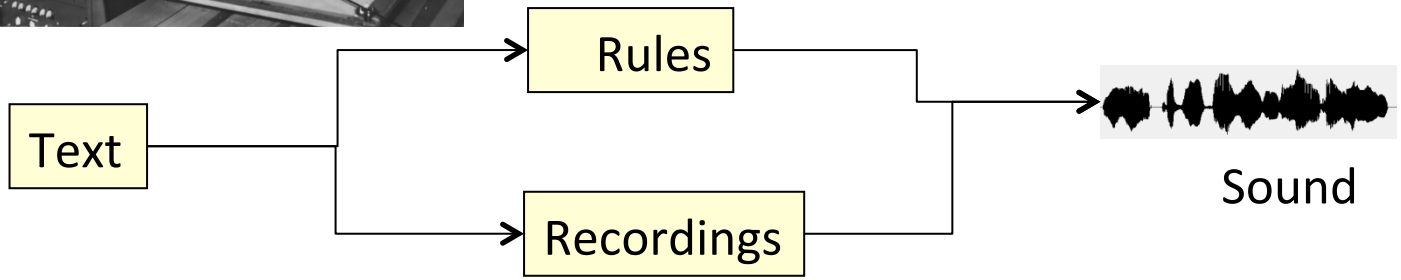


The *automatic* generation of synthesized sound or visual output from *any* phonetic string.

Synthesis approaches

Parametric synthesis

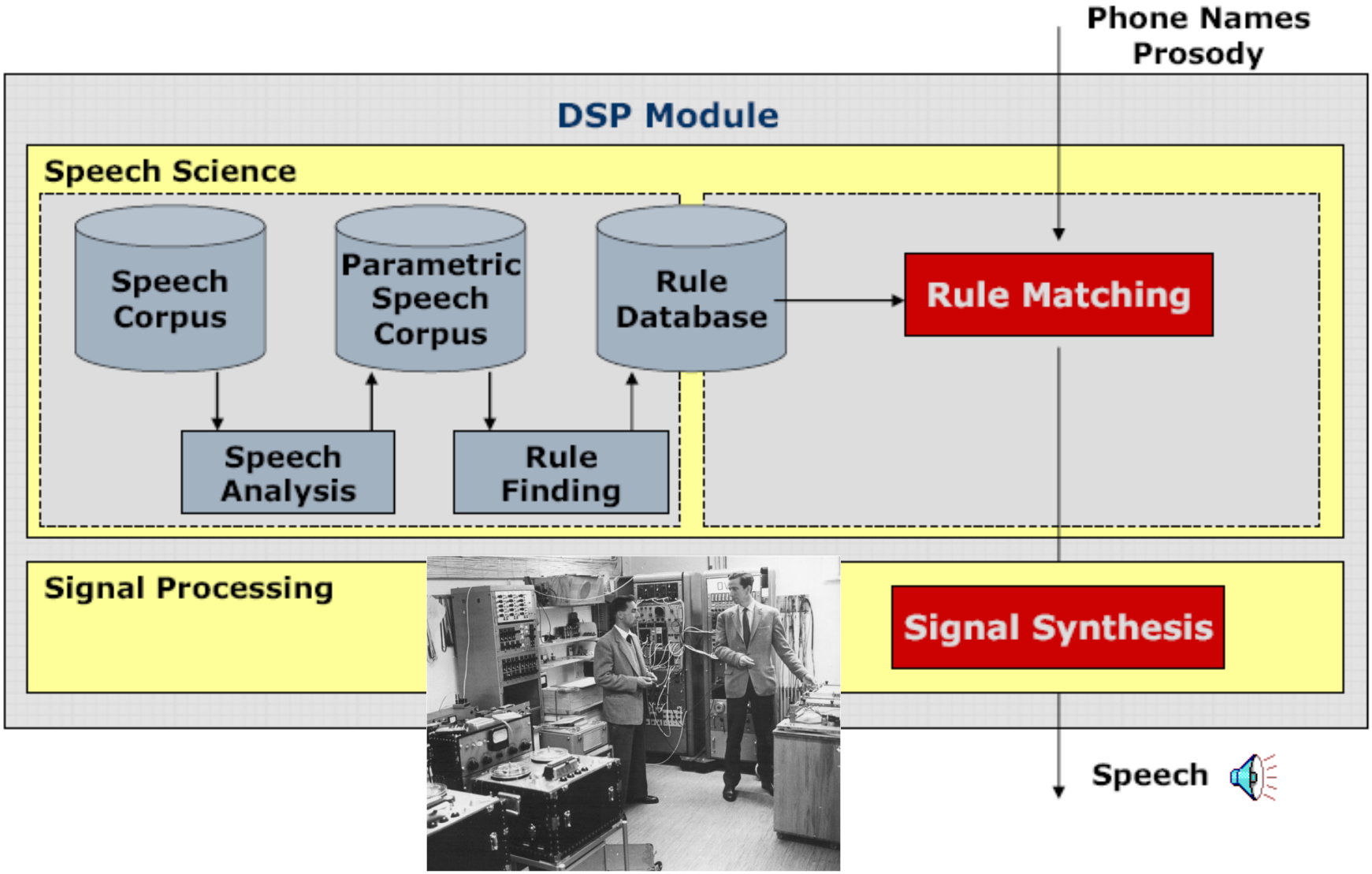
Mathematical rules describe the process
Less natural, but can say anything.



Concatenative synthesis

What should the synthesis be able to say?
How big should these parts be?
How do you get the parts to fit?

Synthesis by Rule/Parametric

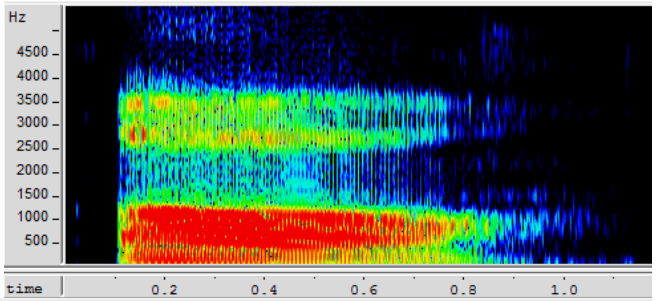


Replicating the properties of speech

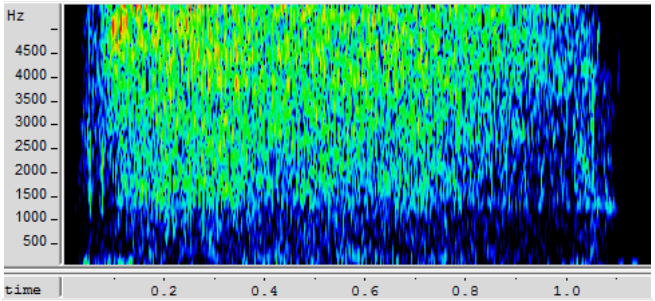
Vowels

Consonants

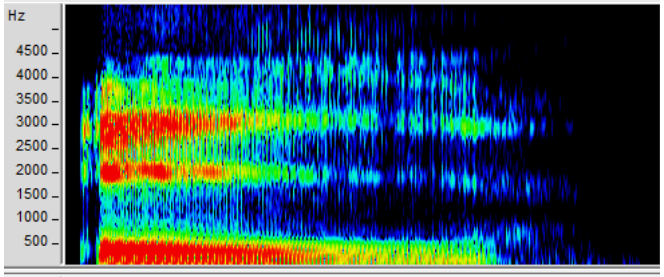
A



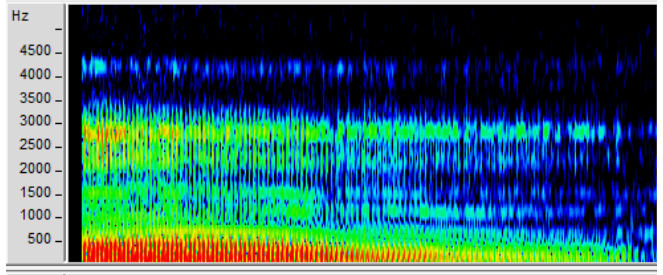
S



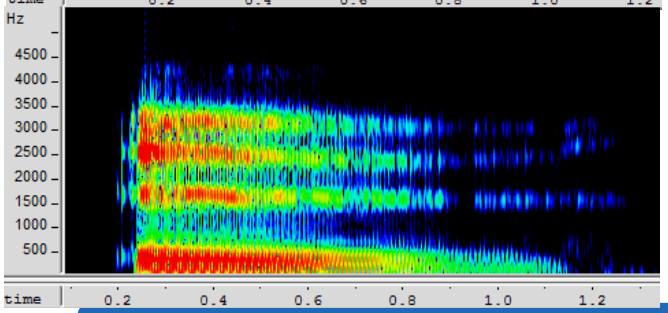
I



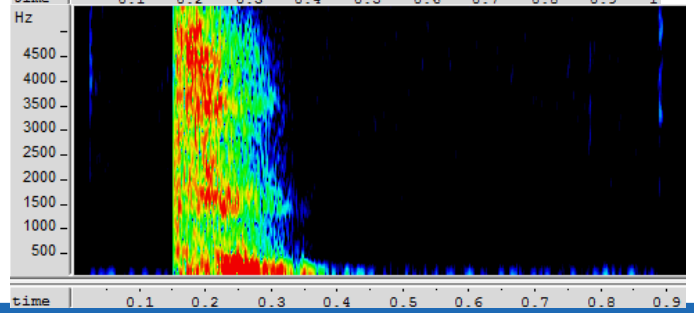
N



U

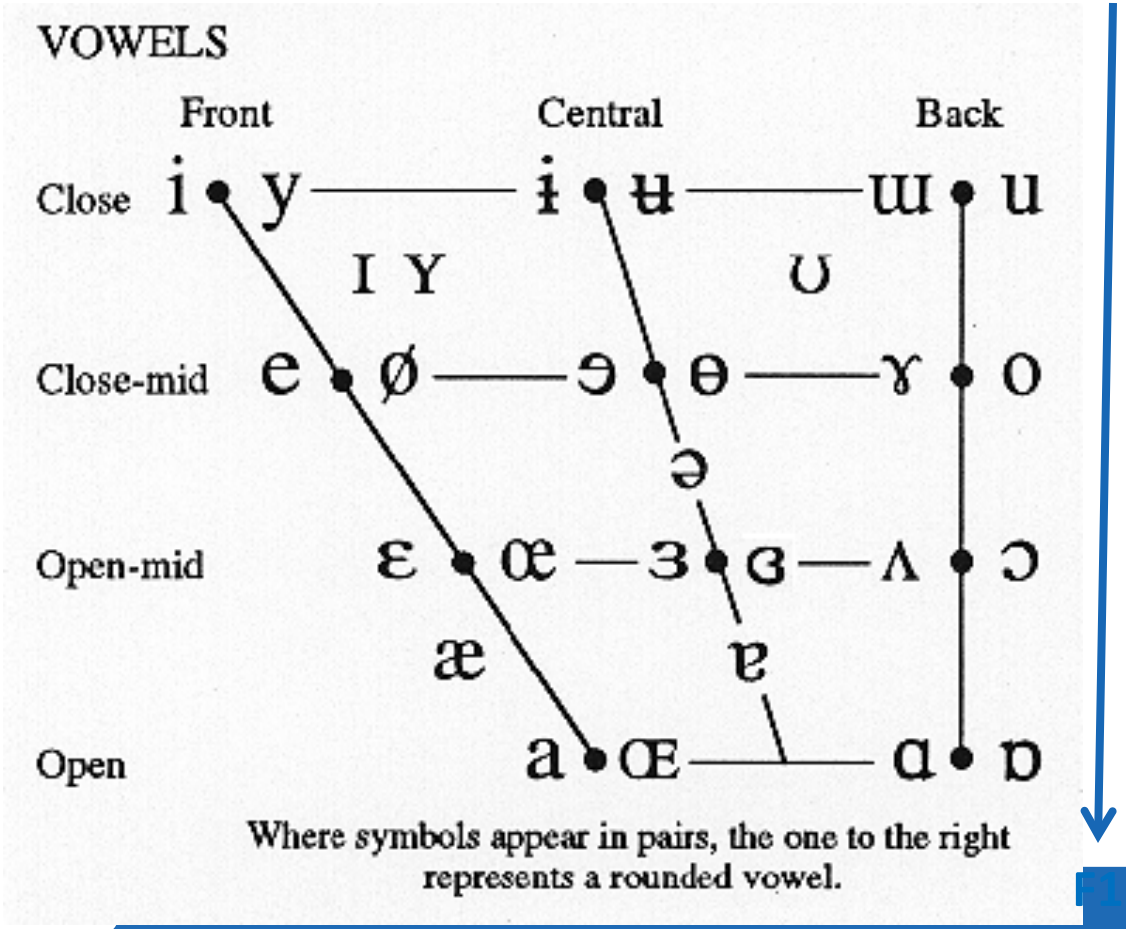


T

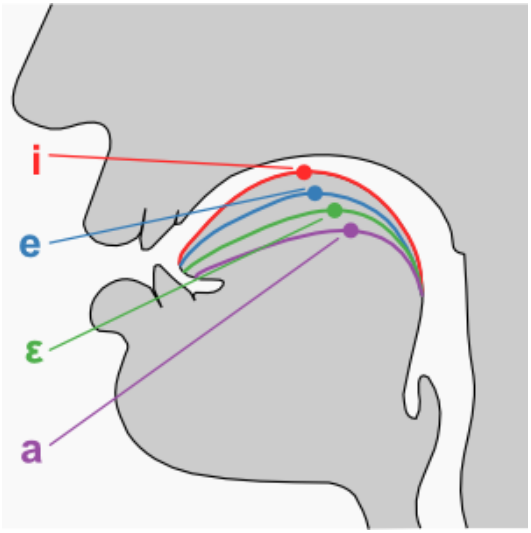


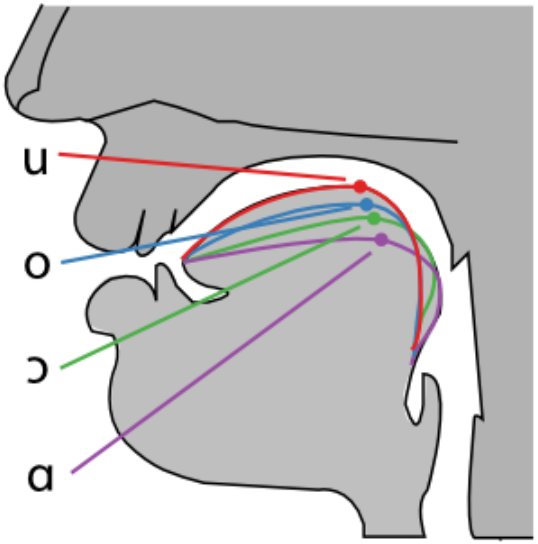
The vowel space

F2



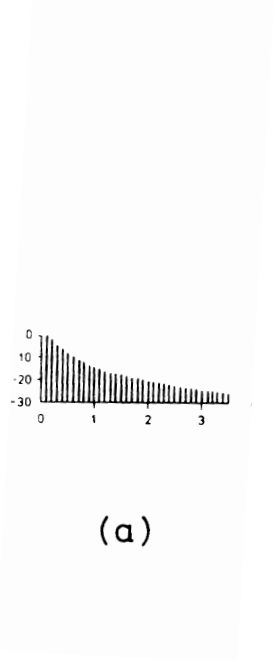
F1

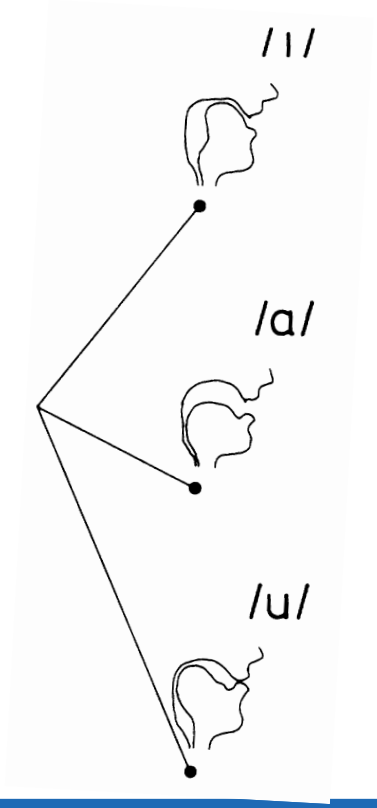


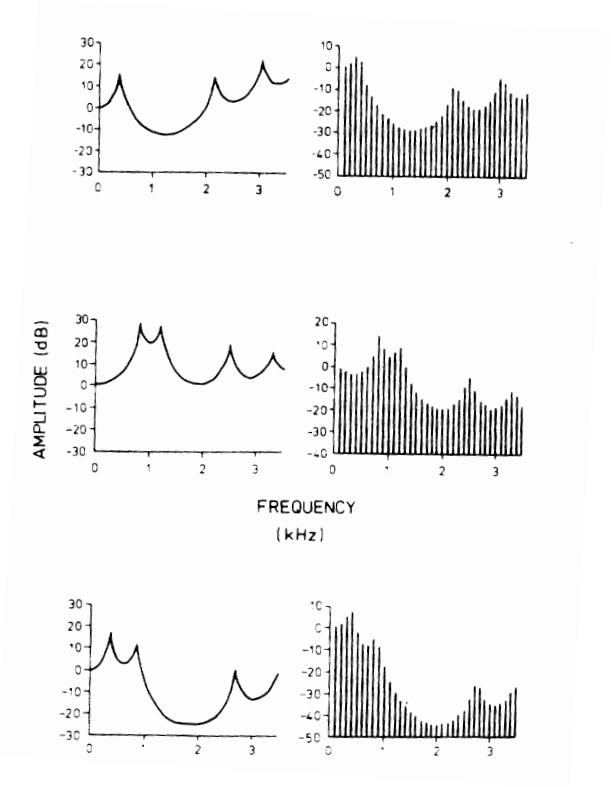


The source-filter theory

The **signal** is the result of a **linear filter** excited by one or several sources.

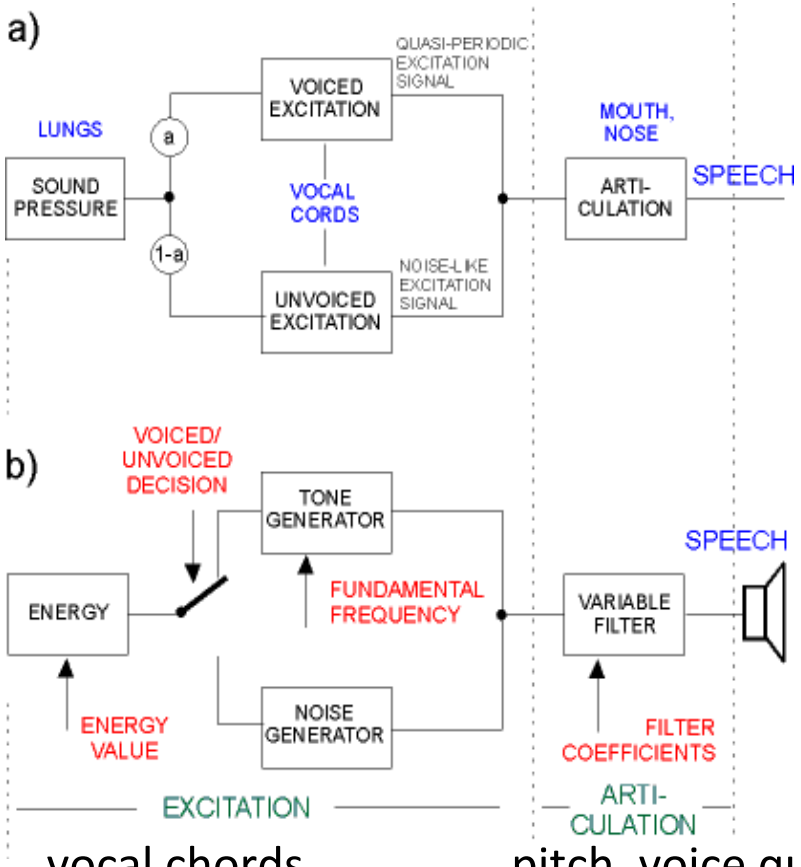






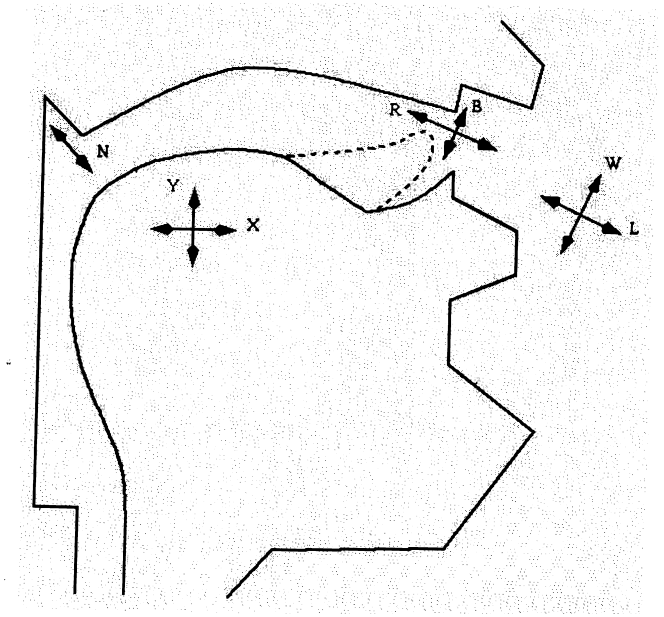
Models of speech production

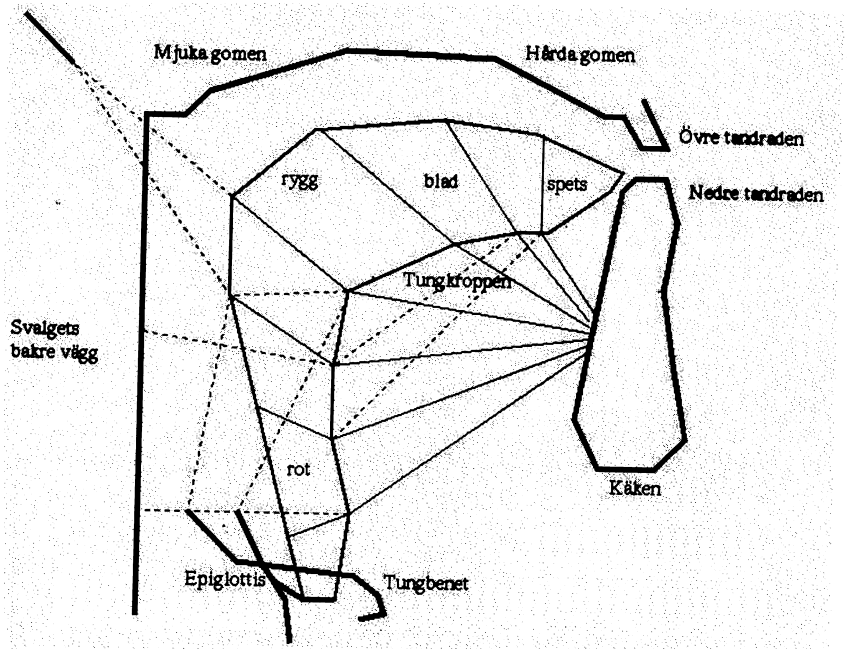
a) humans b) machines



Carrier	vocal chords,	pitch, voice quality
Modulation	vocal tract, nose, mouth	formants

Filter I: articulatory synthesis

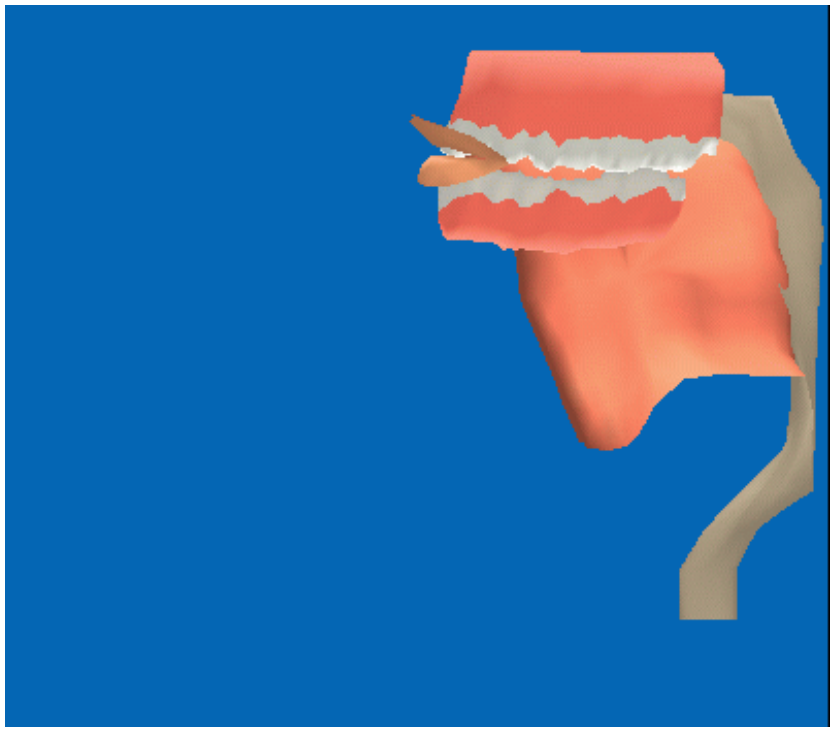




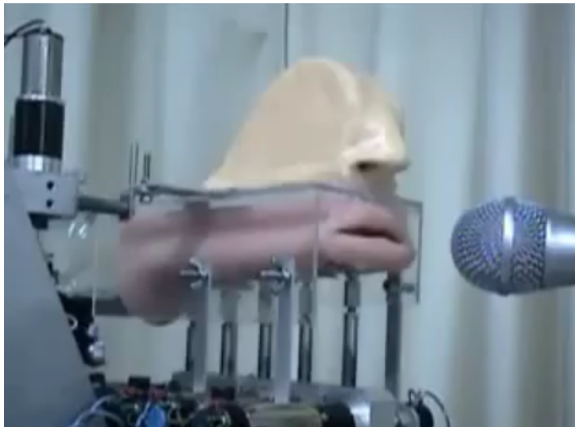
Functional
Geometric parameters control the different parts of the tongue, jaw, lips etc.

Physiological
Muscle model. Articulations are created through activation of different muscles.

Articulatory Synthesis

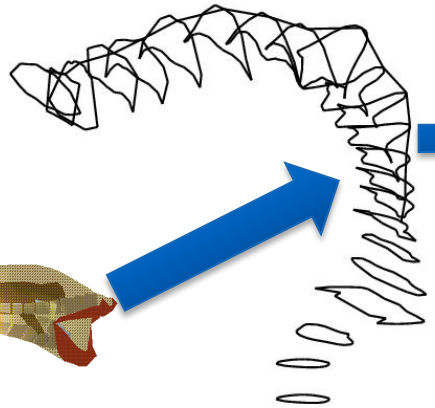


- Jaw opening
- Lip rounding
- Lip Protrusion
- Tongue position
- Tongue height
- Tongue tip
- Velum

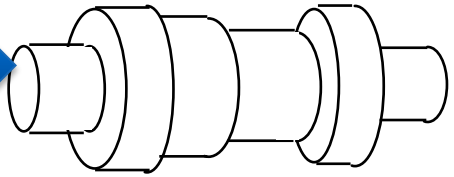


Synthesis from vocal tract shapes

Vocal tract model

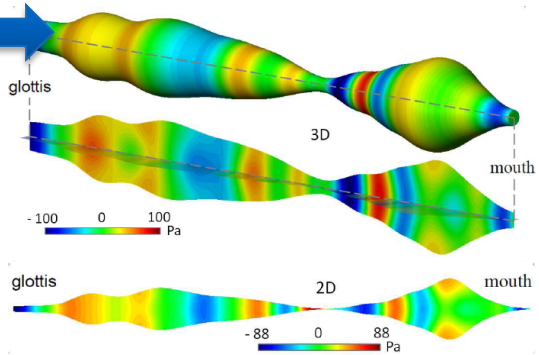


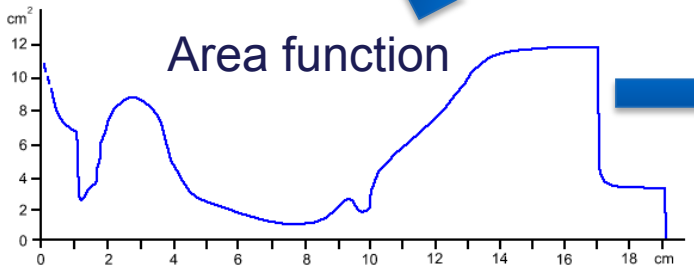
Cross-sections



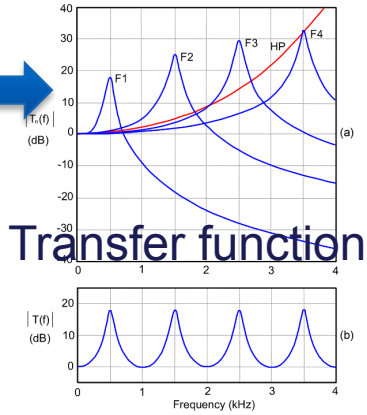
Tubes

3D air flow calculations

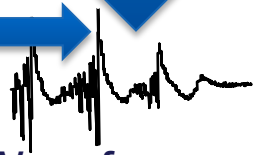




Area function

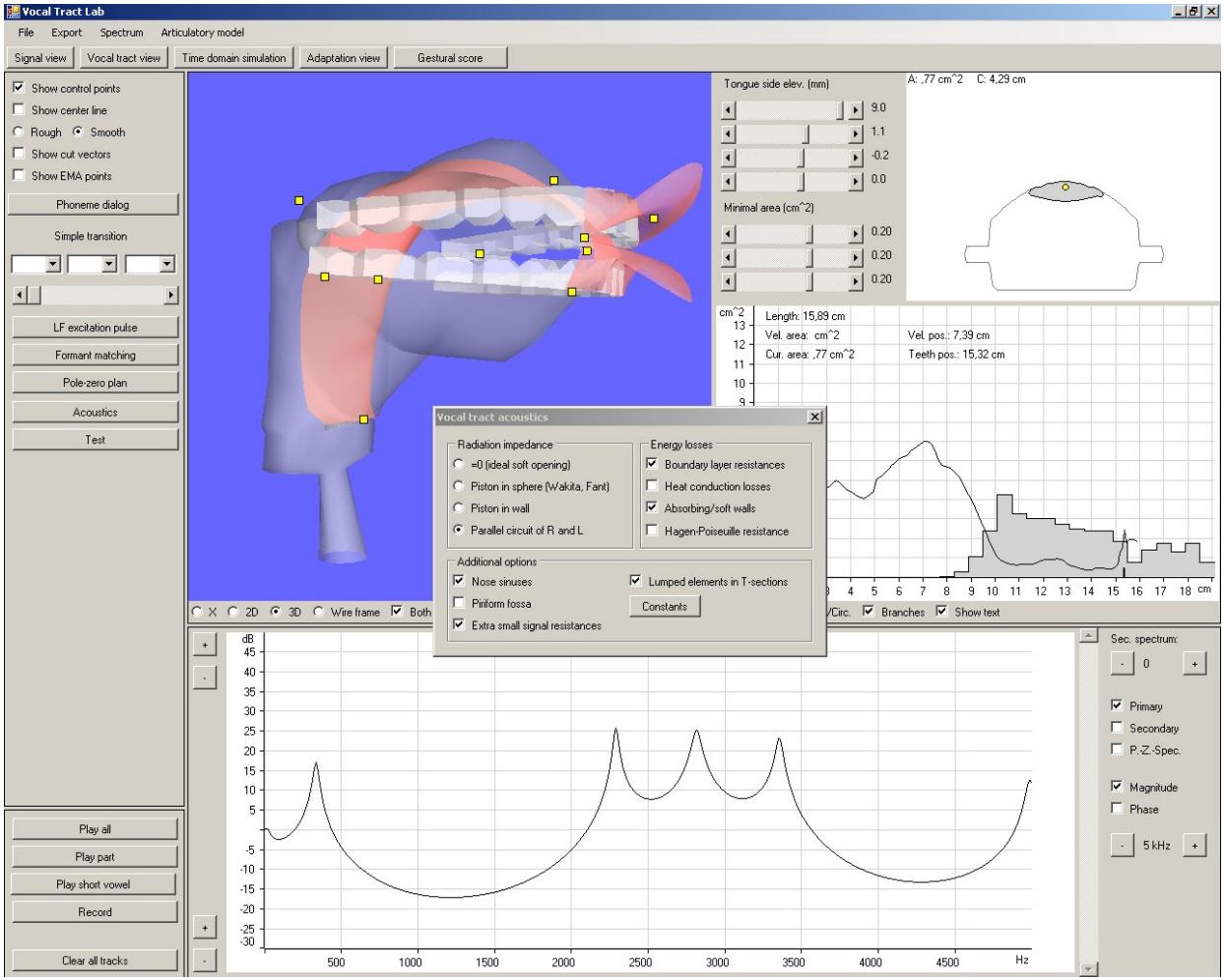


Transfer function

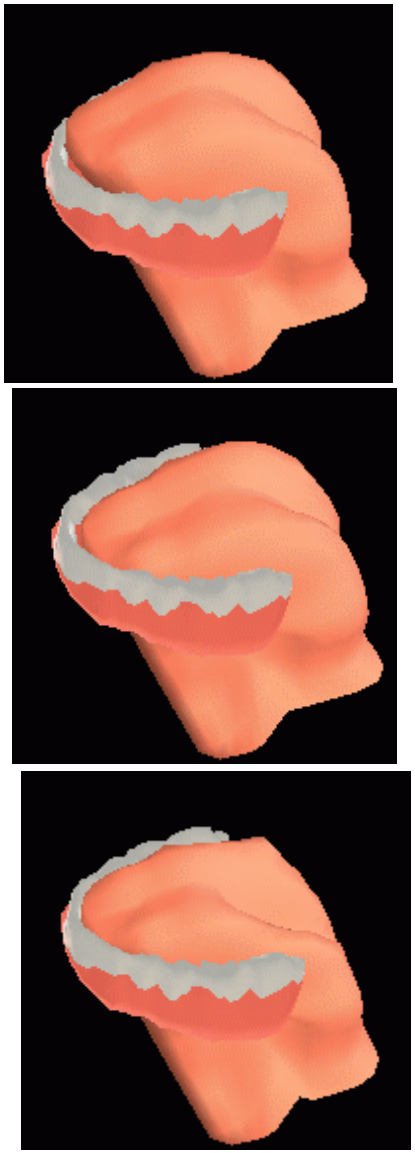


Waveform

Try yourself: www.vocaltractlab.de



Summary: Articulatory Synthesis



Benefits:

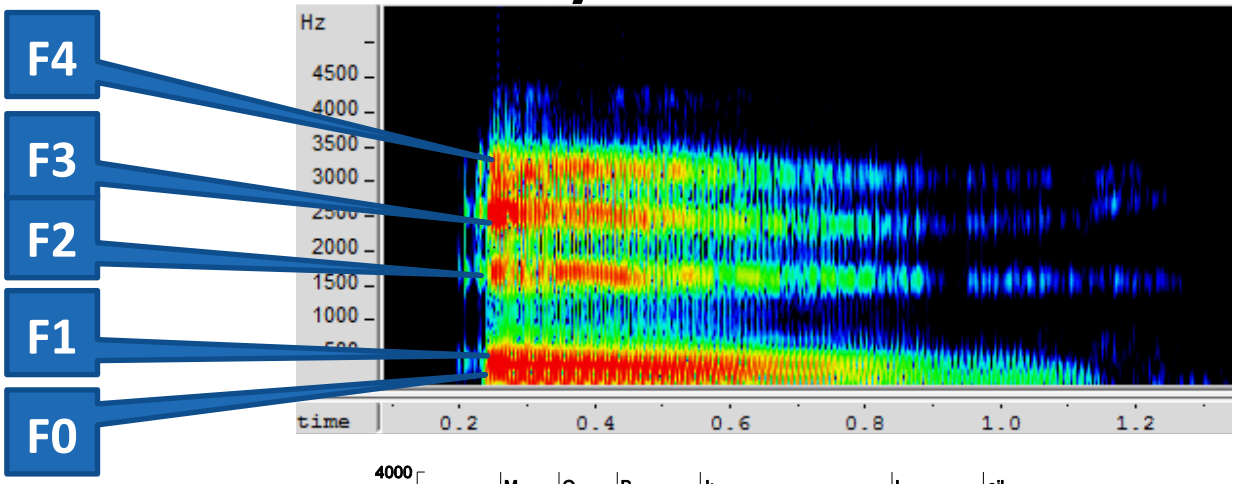
- Speech production in the same way as humans
- Can be made with very few parameters
- The changes are intuitive
  - (raise the tongue tip, round the lips)

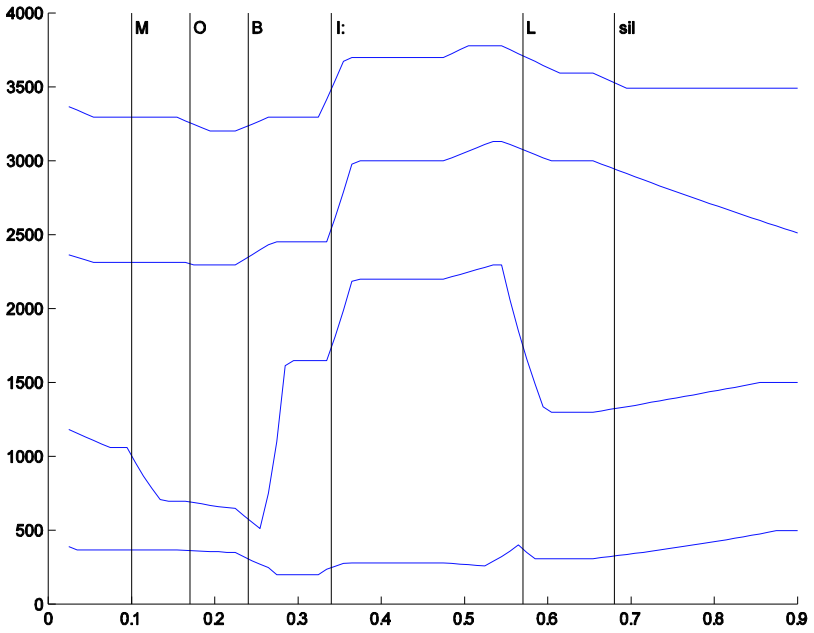
Disadvantages:

- Computationally demanding
- Problems with consonants
- Articulatory measurements required
- State-of-the-art articulatory synthesis still sounds bad

Jaw height

Filter II: Rule-driven formant synthesis





Parameters are generated by rule (RULSYS)

Formant values are generated by interpolating between target frequencies

Parameters are fed to a synthesizer for the source (GLOVE)

Summary: Formant synthesis

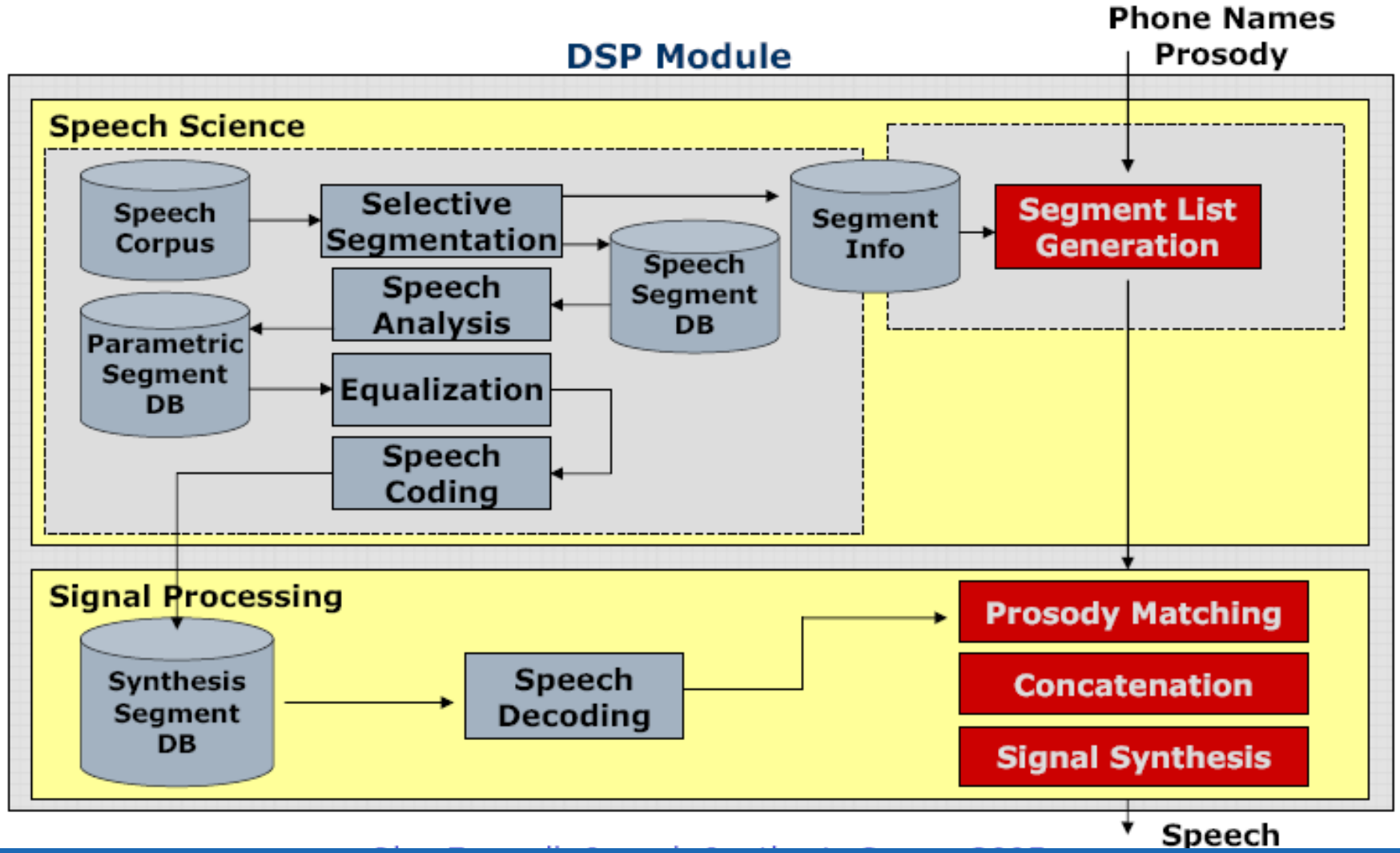
Benefits:

- Possible to change the voice to get different:
  - speakers
  - emotions
  - voice qualities
- Small footprint

Disadvantages:

- Hard to achieve naturalness in voice source
- Some consonant sounds are hard to model with formants (bursts)

Synthesis by Concatenation



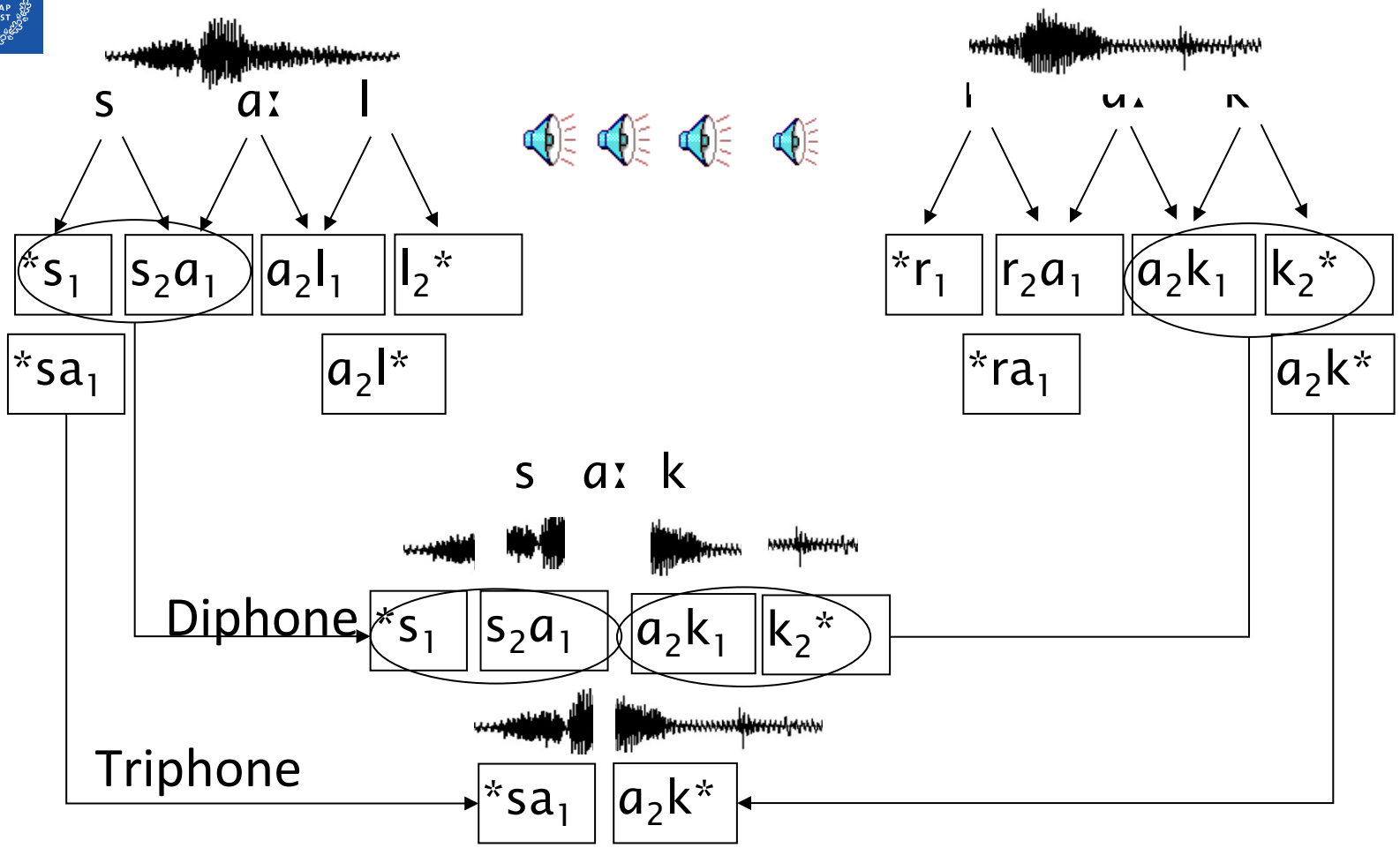
Concatenative synthesis database preparation I

1. Choose the speech units
   - Phone, Diphone, Sub-word unit, unit selection 
2. Compile all units needed
   - Allophones, allowed transitions CV VC VV CC, <sil>
3. Choose context
   - Carrier phrases, non-sense words or natural phrases
4. Select speaker
   - Record many (ATT: 200!) professional speakers,
   - Select by listening test of synthesis attempts
5. Record utterance
   - Use synthesized prompts to guide speaker
   - Read with constant pitch, power and duration
   - Similar recording conditions

Concatenative synthesis database preparation II

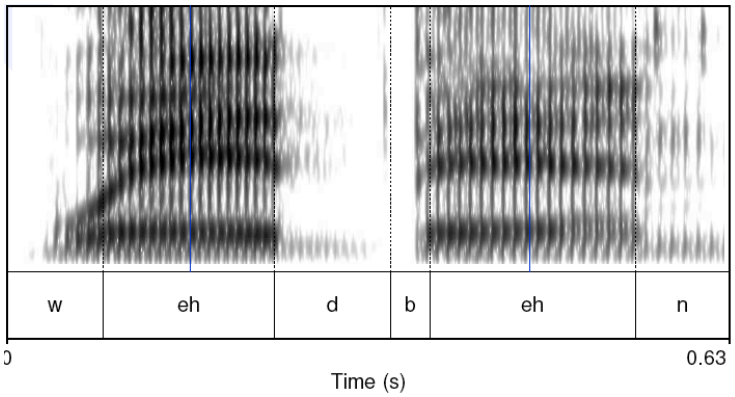
6. Segment signal and extract speech units
    - Manually or automatically, using forced alignment (text + ASR)
    - Find stable part
    - Manually check for errors
7. Store segment waveforms (along with context) and information in a database: Dictionary, waveform, pitch mark
8. Extract parameters & create parametric segment database
    - for data clustering
    - prosody matching
9. Perform amplitude equalization (prevents mismatches)

Diphone & Triphone synthesis



Sequences of a particular sound/phone in all its environments of all/most two-phone sequences occurring in a language

DYO Diphone synthesis



Excercise:
Diphone ”synthesis”; cut and paste

Rationale: the **center’of** a phonetic realization is the most **stable** region, whereas the transition from one segment to another contains the most interesting phenomena, and is thus the hardest to model.

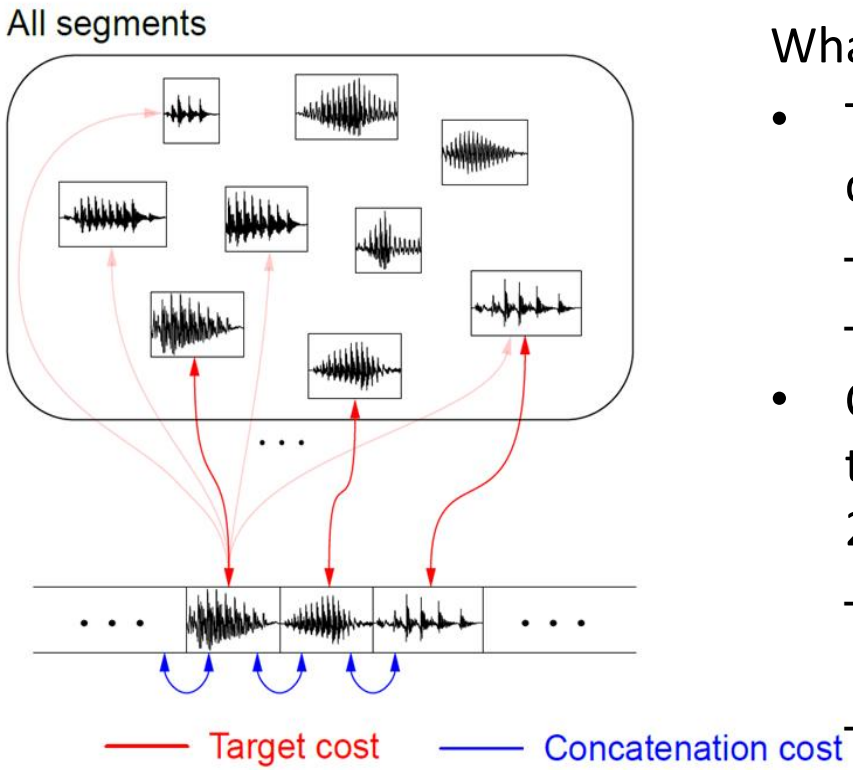
From Diphone Synthesis to Unit Selection Synthesis

Diphone synthesis does not allow enough variation to give natural sounding speech.

“There’s no data like more data”

- Lots of copies of each unit mean you can choose just the right one for the context
- Larger units mean you can capture wider effects
- Units nearest in this space to the targets will be chosen and will require only minor modification
- The corpus is segmented into phonetic units, indexed, and used as-is
- The trend is towards longer and longer units


Unit Selection Synthesis



- Find the unit in the database that is the best to synthesize this target segment

What does “best” mean?

- Target cost: Closest match to the target description, in terms of
  - Phonetic context
  - Pitch, power, duration, phrase position
- Concatenation cost: The difference between the end of diphone 1 and the start of diphone 2:
  - Matching formants + other spectral characteristics
  - Matching energy
  - Matching F0

Summary: Unit Selection

Advantages

- Quality is far superior to diphones
- Natural prosody selection sounds better
- Non-linguistic features of the speakers voice built in

Disadvantages:

- Fixed voice
- Quality can be very bad in places
  - HCI problem: mix of very good and very bad is quite annoying
- Large footprint, it is computationally expensive
- Can't synthesize everything you want:
  - Diphone technique can move emphasis
  - Unit selection gives good (but possibly incorrect) result

From Unit selection to HMM synthesis

Problems with Unit Selection Synthesis

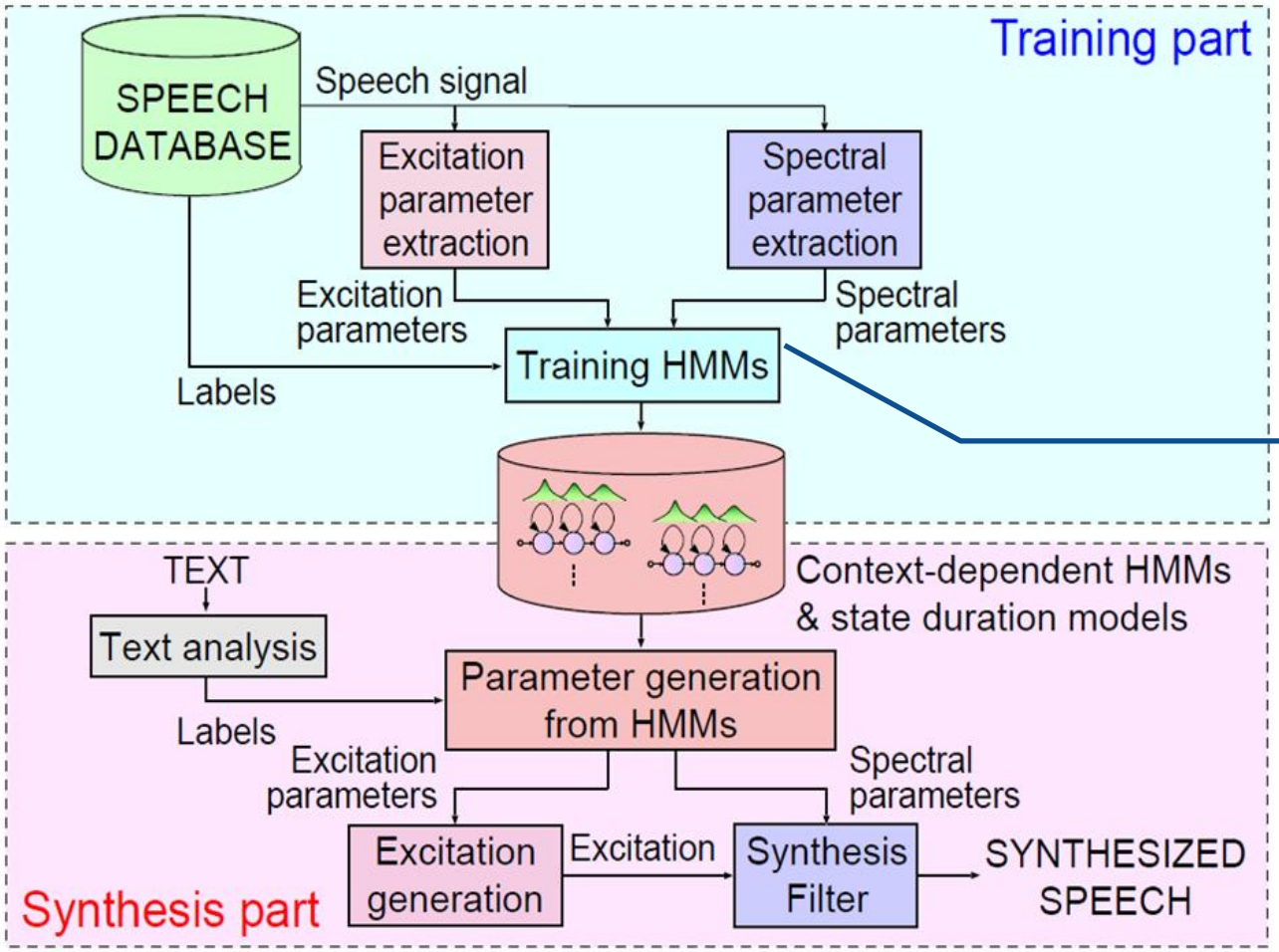
- Discontinuities: Can't modify signal
- Hit or miss: database often doesn't have exactly what you want
- Fixed voice

Solution: HMM (Hidden Markov Model) Synthesis

- Stable, Smooth and easy to create multiple voices
- Sounds unnatural to researchers, but naïve subjects prefer it

- Example: Nina as unit selection and HMM synthesis voice


HMM Synthesis



Segment features:
- Context,
- position in syllable

Syllable features:
- Stress and lexical accent, position in word and phrase

Word features
- number of syllables
- position in phrase

Phrase features
- phrase length

Utterance features:
- length in syllables

Speaker features:
- Dialect,
- speaking style,
- emotion

HMMs in synthesis

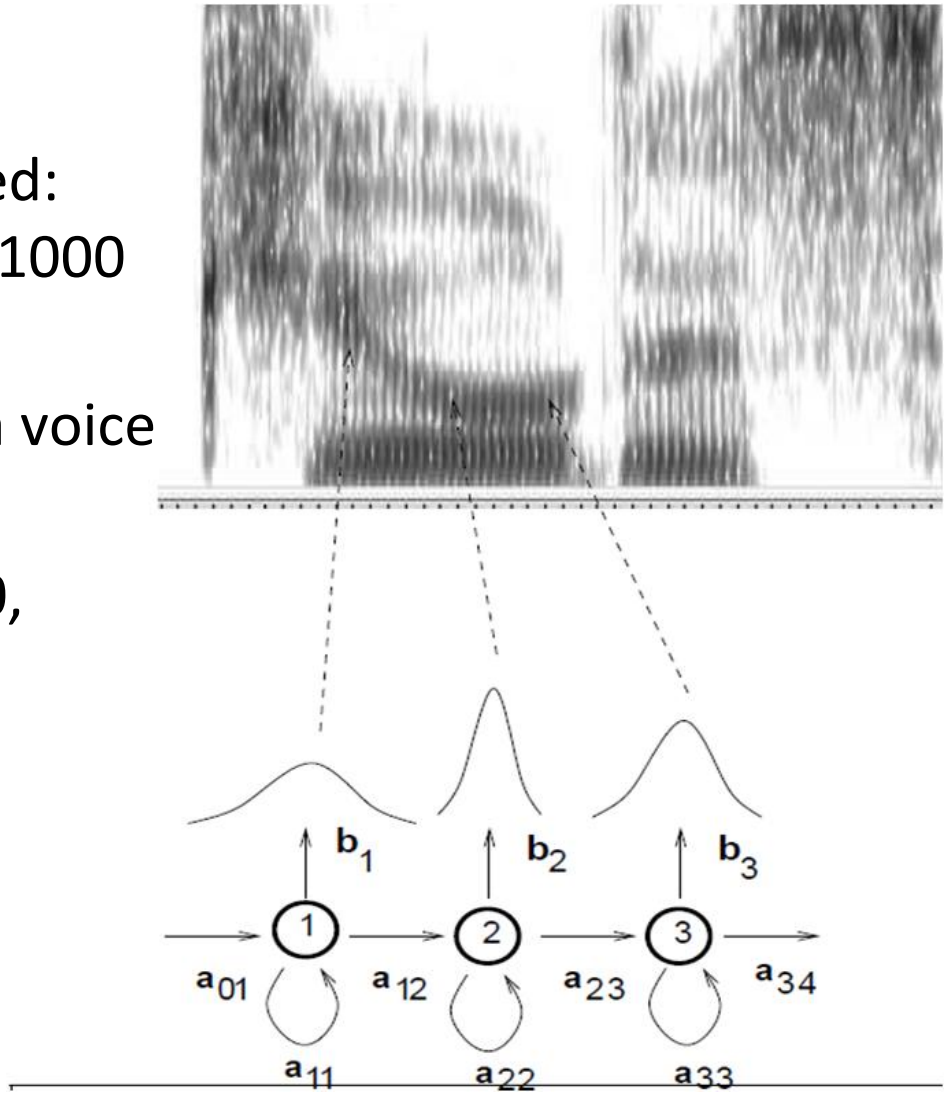
The training is automatic. You need:

- The text + recordings of about 1000 sentences
- takes 24 hours and generates a voice of less than 1 MB

Separate HMMs for: Spectrum, F0, Duration

Training in two steps:

1. Context independent models
2. Use these models to create context dependent models.

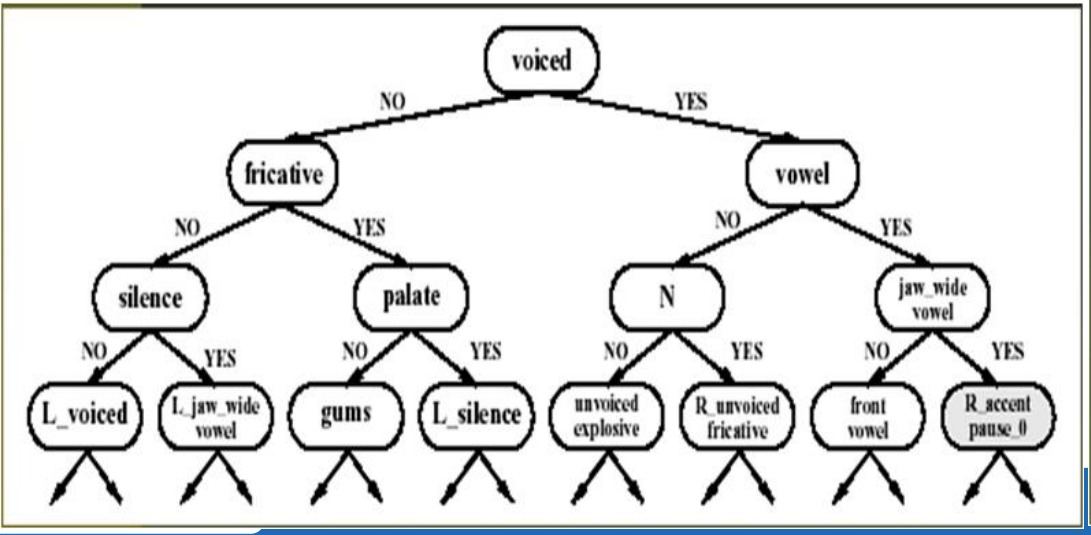


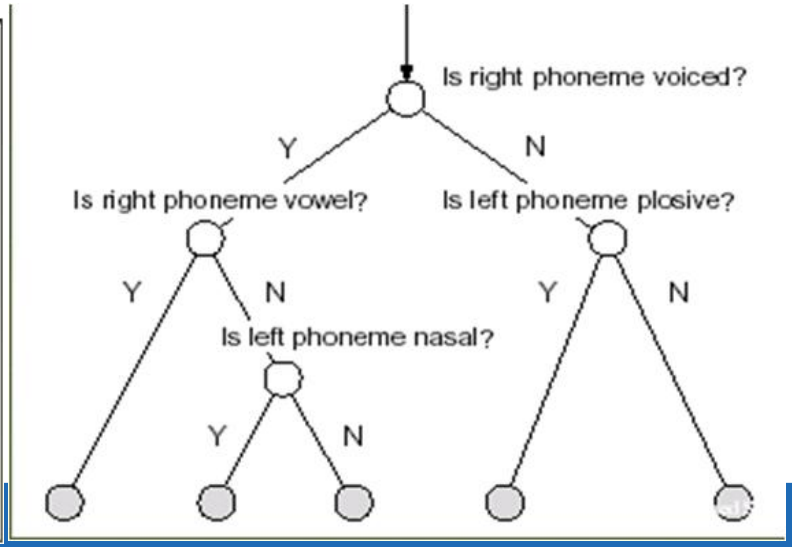
Clustering

- Storing all contexts requires much space
- It may be difficult to find alternatives for missing models
- Many models are very similar = redundancy

Groups a large database into clusters
Three decision trees: Duration, F0 and Spectrum
Division based on yes/no questions

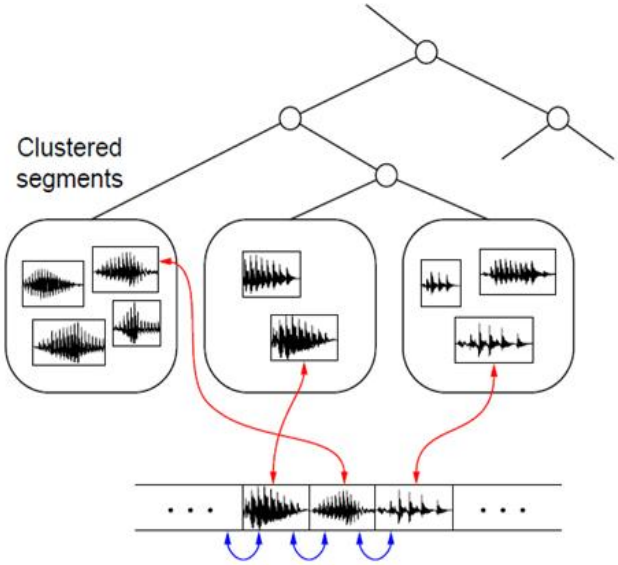
- Grouping acoustic similar phonemes
- Features.
- Context.



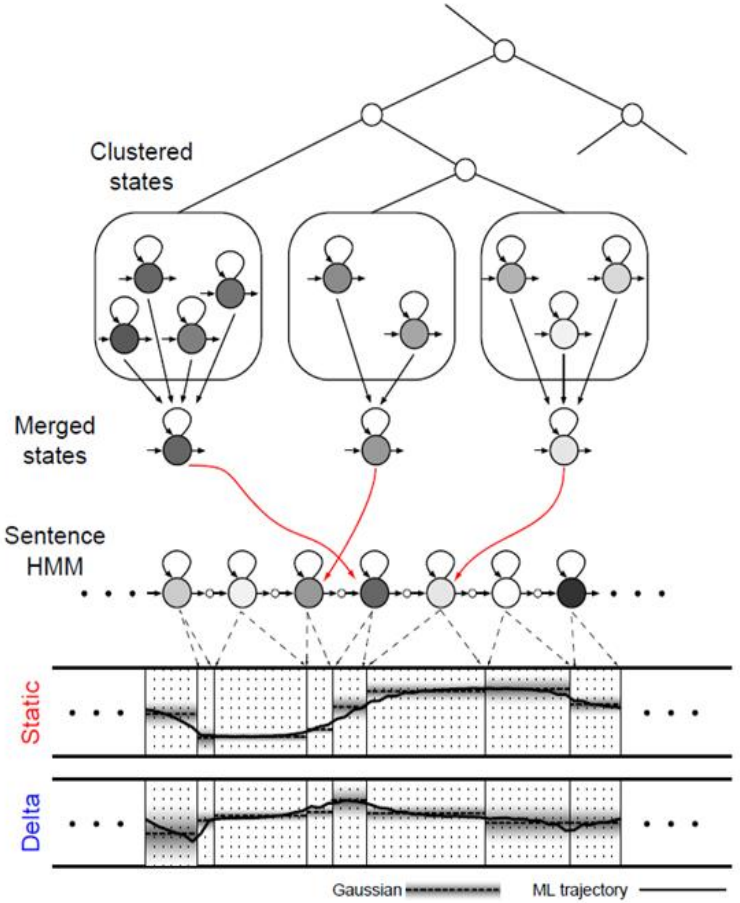


Compare unit selection and HMM synthesis

Unit selection

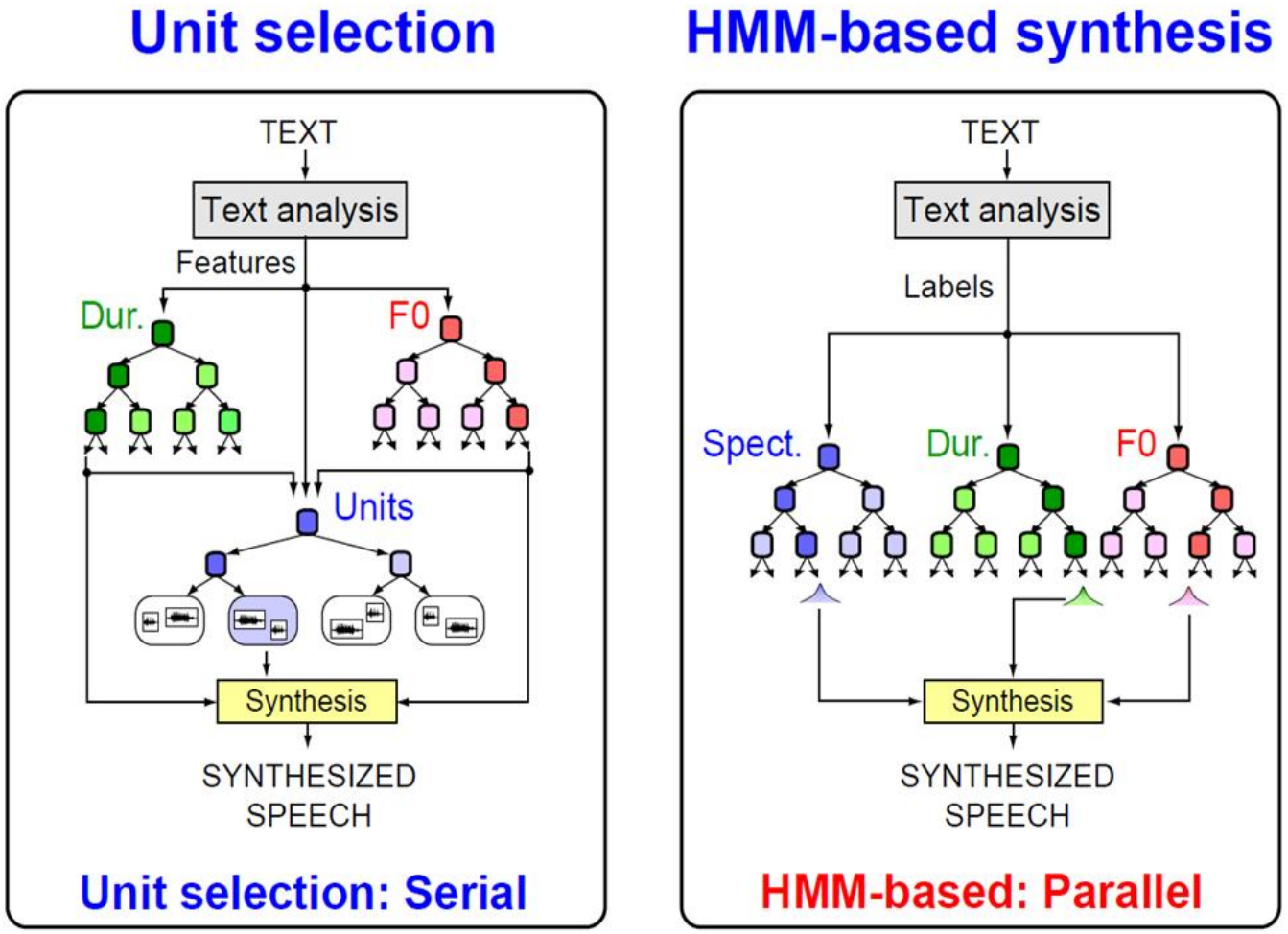


HMM-based synthesis



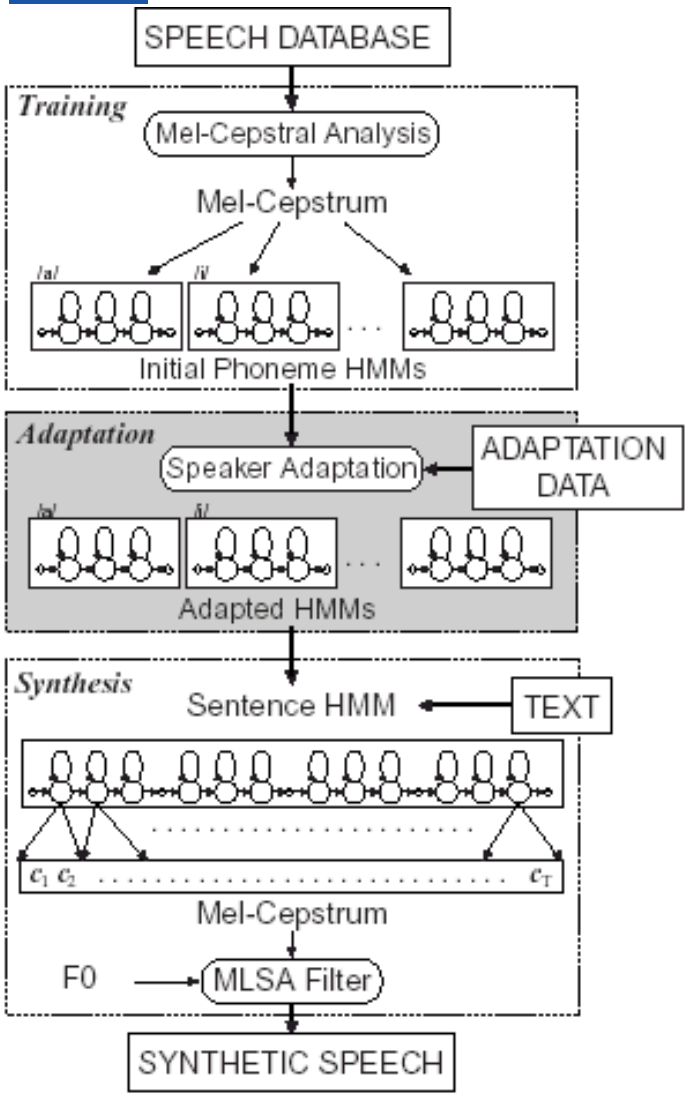
Slide from Tokuda

Compare unit selection and HMM synthesis 2



Slide from Tokuda

Speaker adaptation



Norrland	Skåne	Gotland	Svealand
vart tar universum slut?			
centerpartister och kristdemokrater menar dock att brudparet kan slippa betala det kan bådas föräldrar göra			
gula sidorna finns från och med i fredags sökbart via mobiltelefonen			
om man till exempel tar en telefon och frågar hur den fungerar så svarar vetenskapsmannen med att lyfta på luren och slå numret			

Text-to-Speech Synthesis Evolution

Poor Intelligibility; Poor Naturalness, Small footprint

Formant Synthesis

Bell Labs; Joint Speech Research Unit; MIT (DEC-Talk); Haskins Lab; KTH

→

Good Intelligibility; Poor Naturalness

LPC-Based Diphone/Dyad Synthesis

Bell Labs; CNET; Bellcore; Berkeley Speech Technology

→

Good Intelligibility; Customer Quality Naturalness (Limited Context)

Unit Selection Synthesis

ATR in Japan; CSTR in Scotland; BT in England; AT&T Labs (1998); L&H in Belgium

→

Multi-speaker training, speaker adaptation; Naturalness, generative, Small footprint

HMM Synthesis

HTS in Japan; CSTR in Scotland;

1962 1967 1972 1982 1987 1992 1997 2002

Year

Practical System Issues

Size of typical system (Rhetorical rVoice):

- ~300M

Speed:

- For each diphone, average of 1000 units to choose from, so:
- 1000 target costs
- 1000x1000 join costs
- Each join cost, say 30x30 float point calculations
- 10-15 diphones per second
- 10 billion floating point calculations per second

But commercial systems must run ~50x faster than real time

Heavy pruning essential:

- 1000 units -> 25 units

What the voice conveys

The linguistic component
(the words that are said)

The extralinguistic component
(the identity of the speaker)

The paralinguistic component
(the attitude of the speaker)

The dialog control component
(selection of the next speaker)

Prosody

Prosody = melody, rhythm, “tone” of speech

Not what words are said, but how they are said 

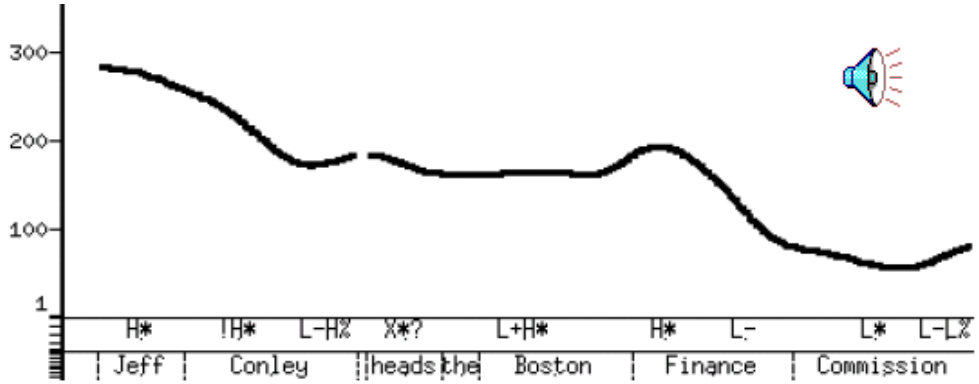
Prosody is conveyed using:

- Pitch
- Phone durations
- Energy 

Human languages use prosody to convey:

- phrasing and structure (e.g. sentence boundaries)
- disfluencies (e.g. false starts, repairs, fillers) 
- sentence mode (statement vs question)
- emotional attitudes (urgency, surprise, anger)

Intonation: F0 contour



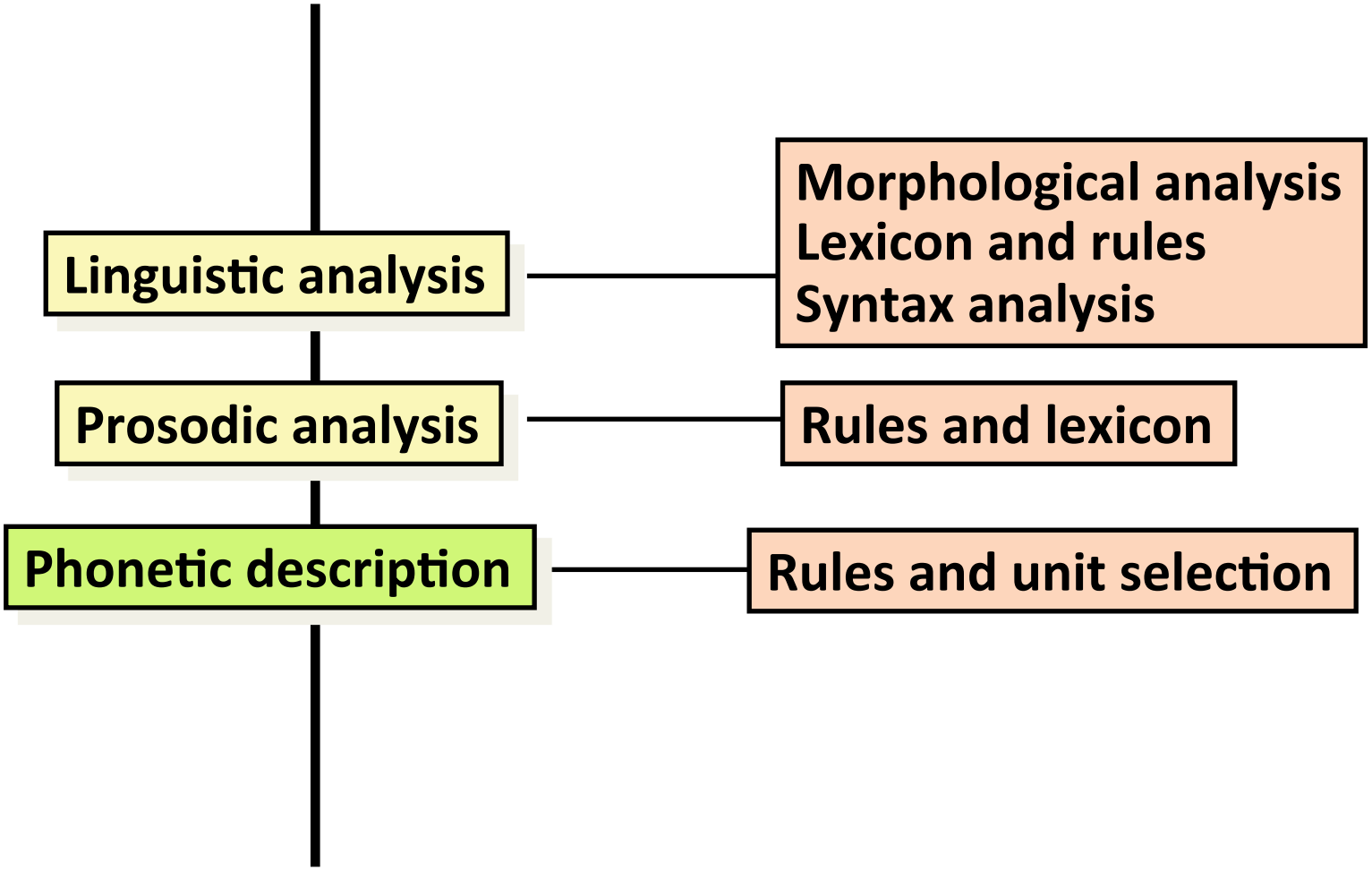
Large pitch range (female)
Authoritive (final fall)
Emphasis for Finance (H*)
Final has a raise – more information to come

Word stress and sentence intonation

- each word has at least one syllable which is spoken with higher prominence
- in each phrase the stressed syllable can be accented depending on the semantics and syntax of the phrase

Prosody relies on syntax, semantics, pragmatics: personal reflection of the reader.

From text


Linguistic analysis
Morphological analysis
Lexicon and rules
Syntax analysis
Prosodic analysis
Rules and lexicon
Phonetic description
Rules and unit selection

Synthesis requires understanding

Text-to-speech Synthesis

Natural Language Generation

Homographs: Tomten ökade på stegen för att komma till stegen på tomten.

Numbers: I Februari 2009 fanns 2009 Boeing 747 på 747 orter.

Abbreviations: St Göran, St Essingen, High St

Expressions: “Hoppa på bussen”

Letters are pronounced differently depending on

- Context (kula/kyla, öga/öra, barn)
- Origin (jeans, James Bond)
- Speaker state (emotions etc)
- Not “one character = one phoneme” (‘x’= /ks/, ‘thought)

Preprocessing

Sentence end detection (semicolon, period – ratio, time and decimal point, sentence ending respectively)

Abbreviations (e.g. – for instance)

Changed to their full form with the help of lexicons

Acronyms (I.B.M – these can be read as a sequence of characters, or NASA which can be read following the default way)

Numbers (Once detected, first interpreted as rational, time of the day, dates and ordinal depending on their context)

Idioms (e.g. “In spite of”, “as a matter of fact”– these are combined into single FSU using a special lexicon)

Grapheme-to-phoneme conversion

Dictionary:

- Store a maximum of phonological knowledge into a lexicon.
- Compounding rules describe how the morphemes of dictionary items are modified.
- Hand-corrected, expensive
- The lexicon is never complete: needs out of vocabulary pronouncer, transcribed by rule.

Rules:

- A set of letter to sound (grapheme to phoneme) rules.
- Words pronounced in a such a particular way that they have their own rule are stored in exceptions directory.
- Fast & easy, but lower accuracy

Machine learning:

- Cart tree
- Analogy pronunciation