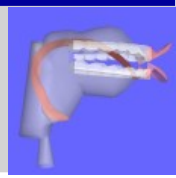
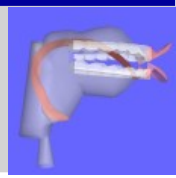


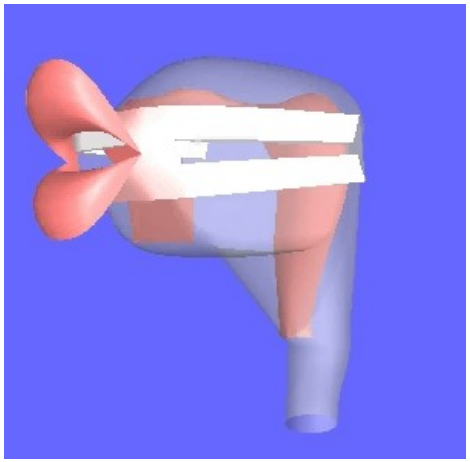
Introduction to Articulatory Speech Synthesis

Eva Lasarcyk, M.A.

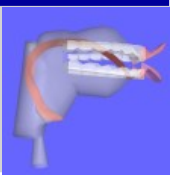
January 25, 2010



Guten Tag, liebe Zuhörer.
(Hello, dear listeners.)

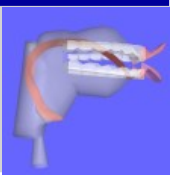


Why speech synthesis?



- Applications
  - Machine reads aloud text for you
    - handicapped people
    - for authors to check their texts
  - Avatars
  - Telephone dialog systems
  - Natural interaction with service robots
  - Part of "Speech-To-Speech" translation systems
- Research – phonetic applications
  - Imitate, manipulate, and understand speech production
  - And perception

How can we create synthetic speech?

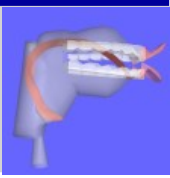


- 3 main strategies
  - Imitate acoustics directly – Formant synthesis
  - Record speech, chop it up, regroup – Concatenative synthesis
  - Imitate, simulate speech production process – Articulatory synthesis

Most systems nowadays use this technique

- Long history
- Some recent major improvements

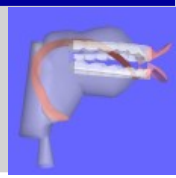
Concatenation of speech segments



- Record speech, chop it up, regroup – Concatenative synthesis
- Goal: Record a LOT to manipulate LITTLE
- Trend: Huge databases with intelligent selection of units
- Advantages
  - Sounds quite natural
  - You need little phonetic knowledge, it's more a signal processing task
  - High quality can be obtained by using a LOT of speech data
- Disadvantages
  - Data recording costly (time/money)
  - Speaker-dependent, post-hoc manipulations decrease quality, structurally new words may easily sound "funny"

Willkommen beim Tag der offenen Tür.

… "ideal" synthesis should be able to …



Cf.: Christine H. Shadle and Robert I. Damper (2001). Prospects for Articulatory Synthesis: A Position Paper. In: Proceedings 4th International Speech Communication Association (ISCA) Workshop on Speech Synthesis, Pitlochry. 121-126.

- sound as natural & intelligible as a human
- recreate a specific voice
- create "generic" voices

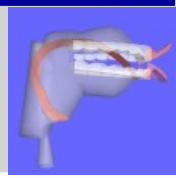
- highly complex
- simulation time intensive
- high quality
  hard to achieve

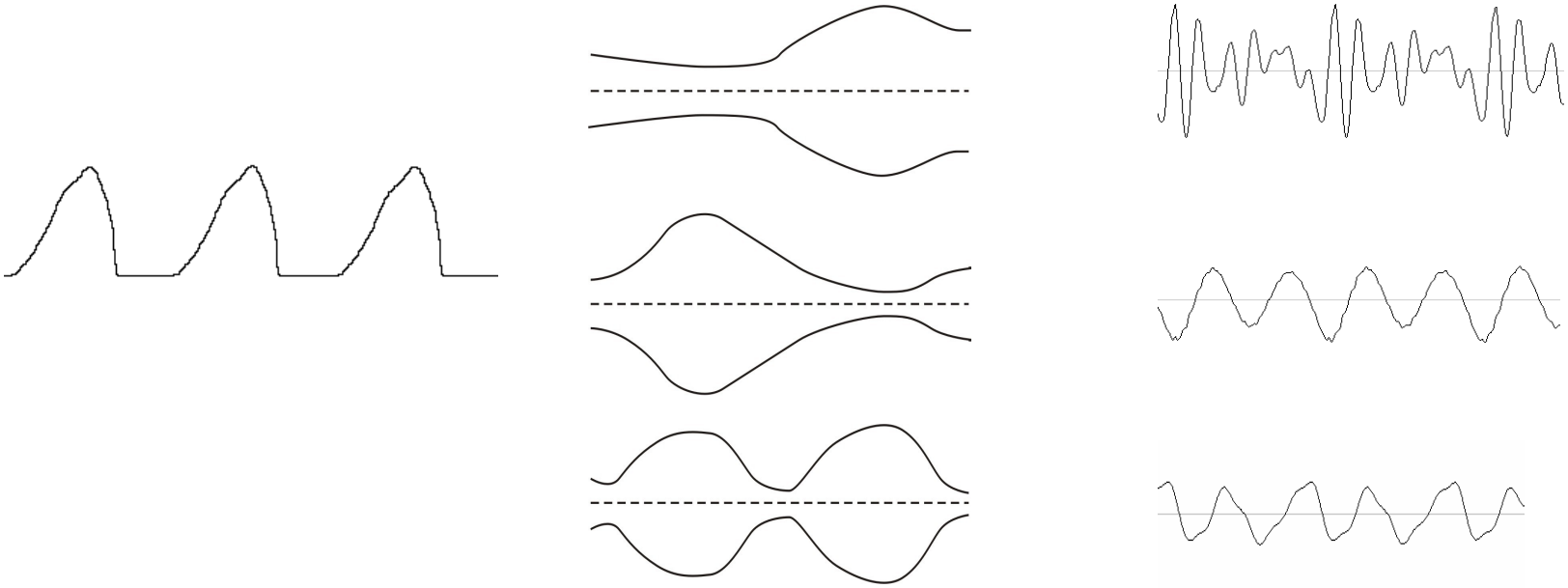
- sound like extraordinary speakers (opera singer, alien)
- speak any language with any emotion without much effort
- … be freely controllable
- … allow us insights into speech production and perception ☺

Do it yourself: Imitate speech production

- Physical simulation of sound with an articulatory model

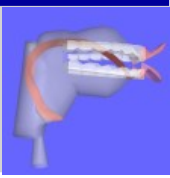
How are speech waves created?





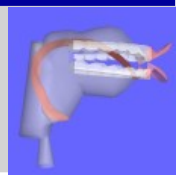
Source	+	Filter	=	Speech signal
Vocal folds		Vocal Tract		Speech

The source: Vocal fold oscillation



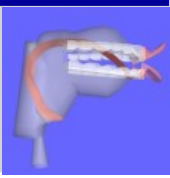
- Different default positions for breathing, speaking and e.g. whispering.
- Oscillation is not only "open-close" but has a vertical component, too.

The filter – resonance cavity shapes



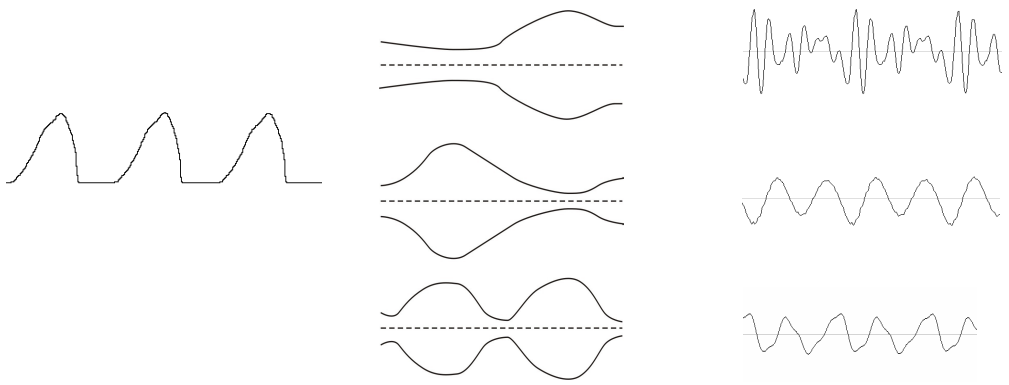
- x-ray movie showing articulation movements during speaking

Filter: Tongue position of vowels



- Chart of vocal tract shapes for different vowels
- Depending on the vowel, the tongue has different shapes

Now we've almost all we need ...



Source + Filter = Speech signal

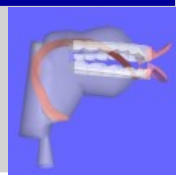
Vocal folds

Vocal Tract

Speech

… to create speech sounds ourselves!

Mechanical speaking machine



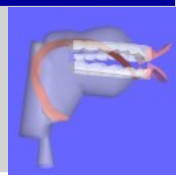
Wolfgang von Kempelen

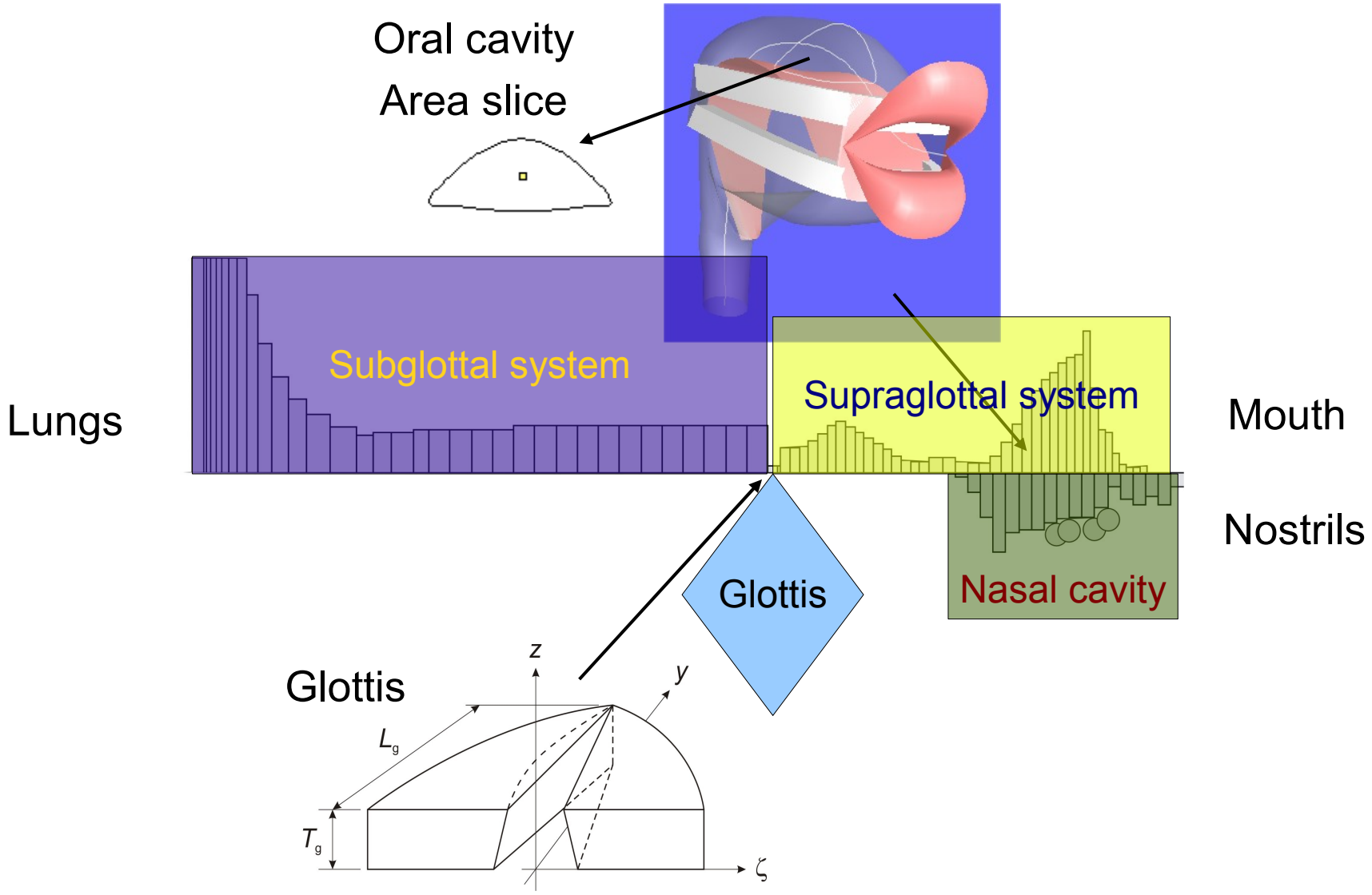
- 1791: "Mechanismus der menschlichen Sprache nebst der Beschreibung einer sprechenden Maschine."
- One of the first attempts to recreate human speech

- image see e.g. http://www.acoustics.hut.fi/publications/files/theses/lemmetty_mst/chap2.html

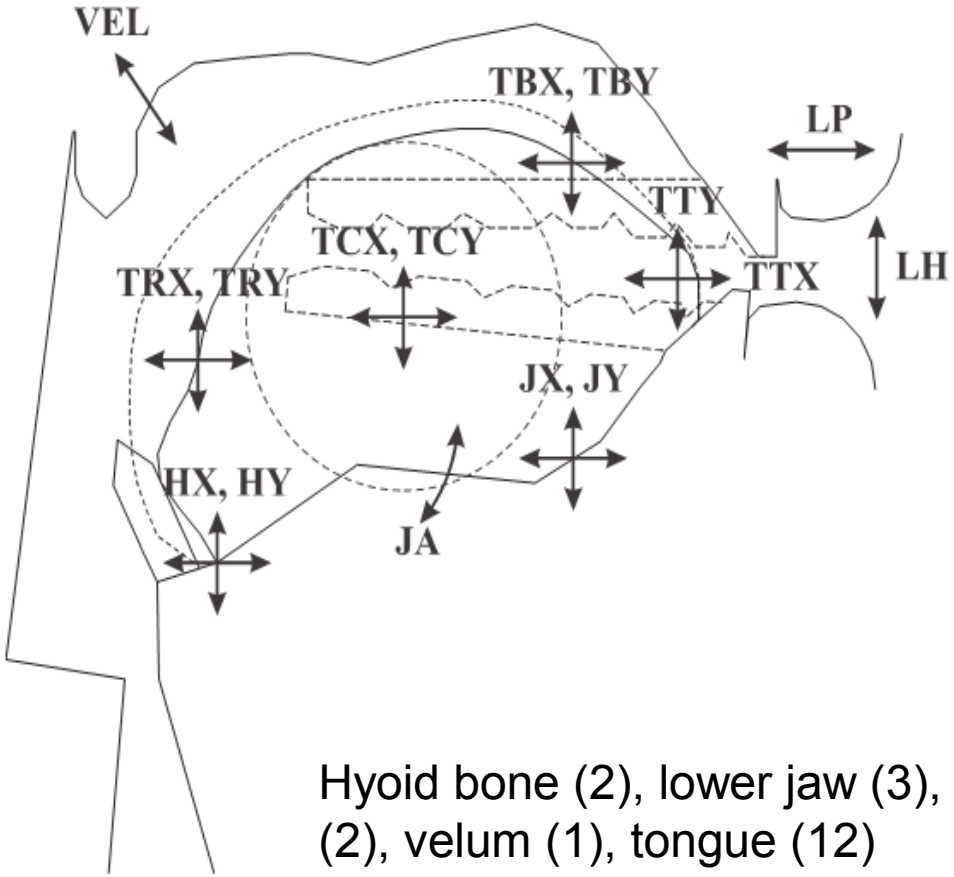
Available in the Phonetics department

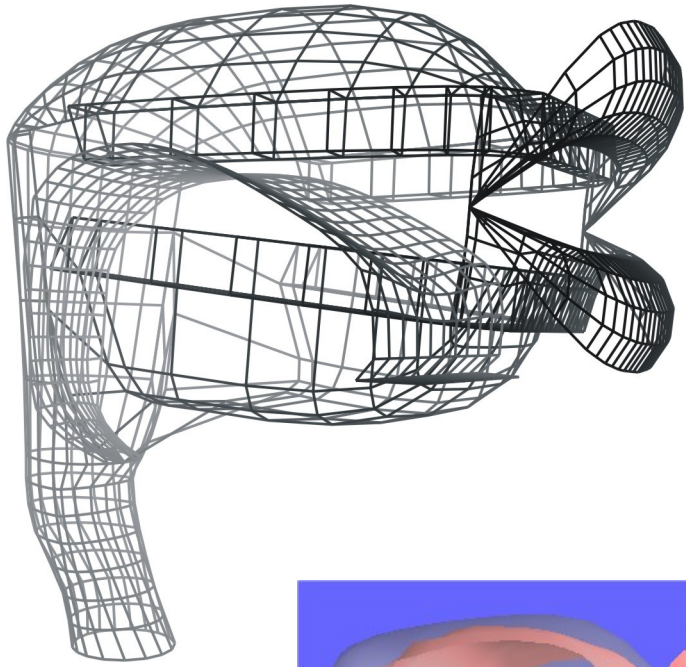
Vocal tract: Geometrical model



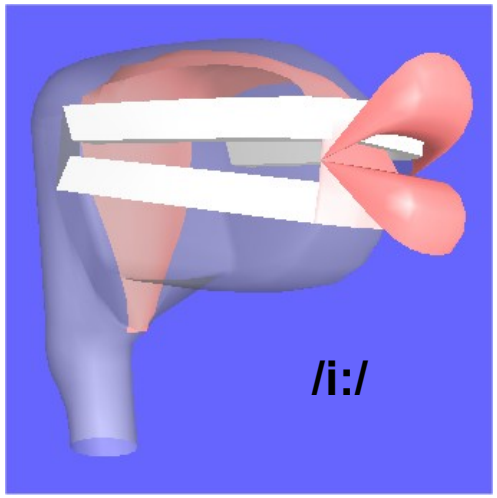


Supraglottal system



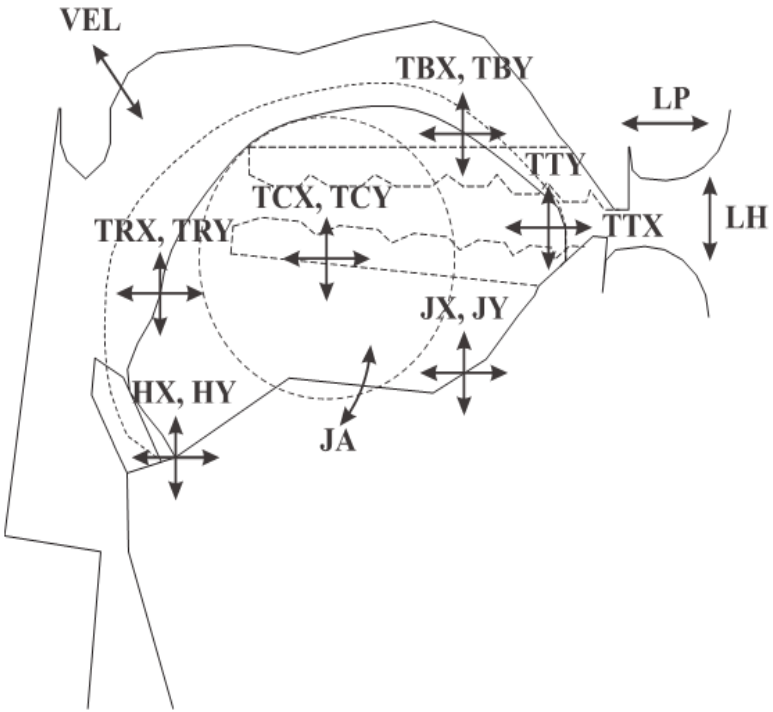


/a:/



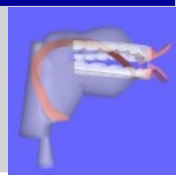
Hyoid bone (2), lower jaw (3), lips (2), velum (1), tongue (12)

Computer speaking machine – control...

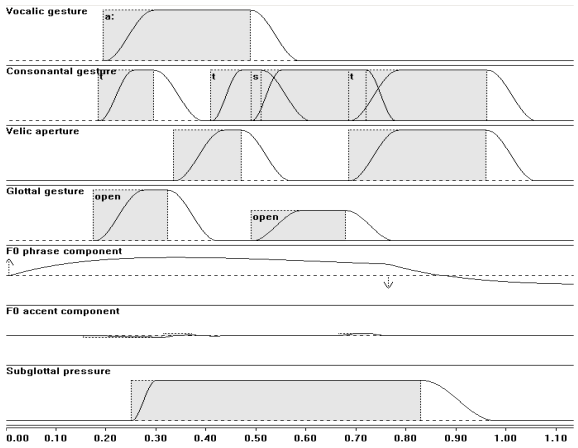


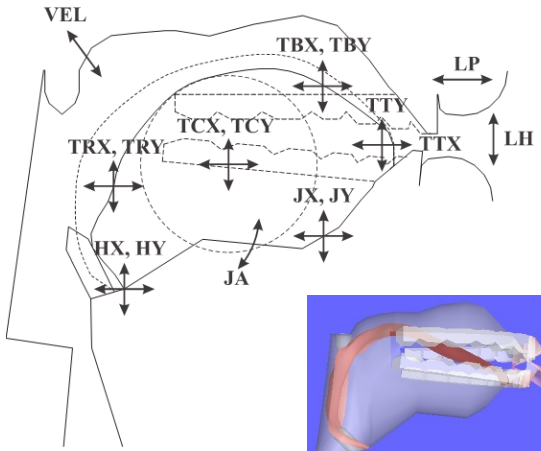
- Temporal coordination of gestures needs to be controlled
- A "brain" needs to give the instructions
- In this synthesis system it is realized by the "gestural score"

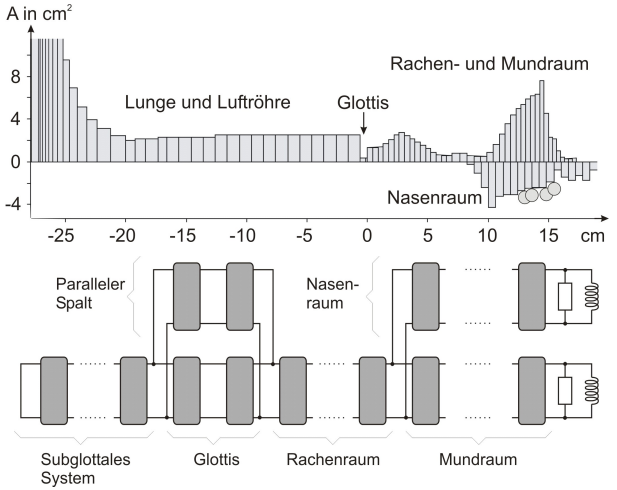
3D articulatory speech synthesizer

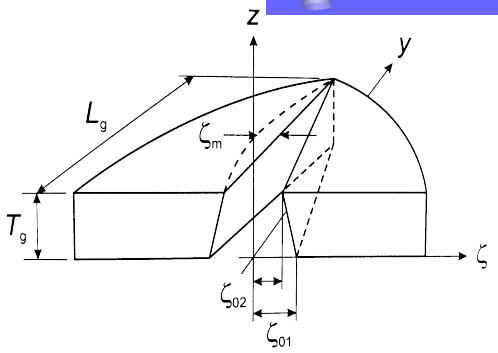


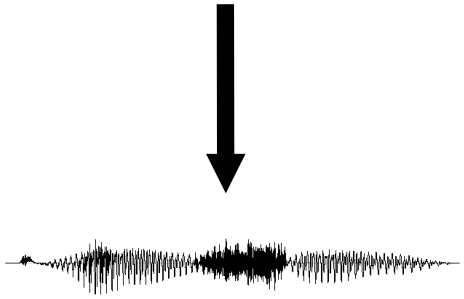
Gestural score → 3D model vocal tract; glottis → Aerodynamic-acoustic simulation







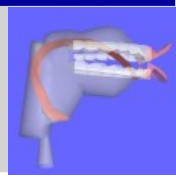




Main advantage over other synthesis strategies: Speech production becomes transparent

VocalTractLab by Peter Birkholz, University Hospital Aachen, www.vocaltractlab.de

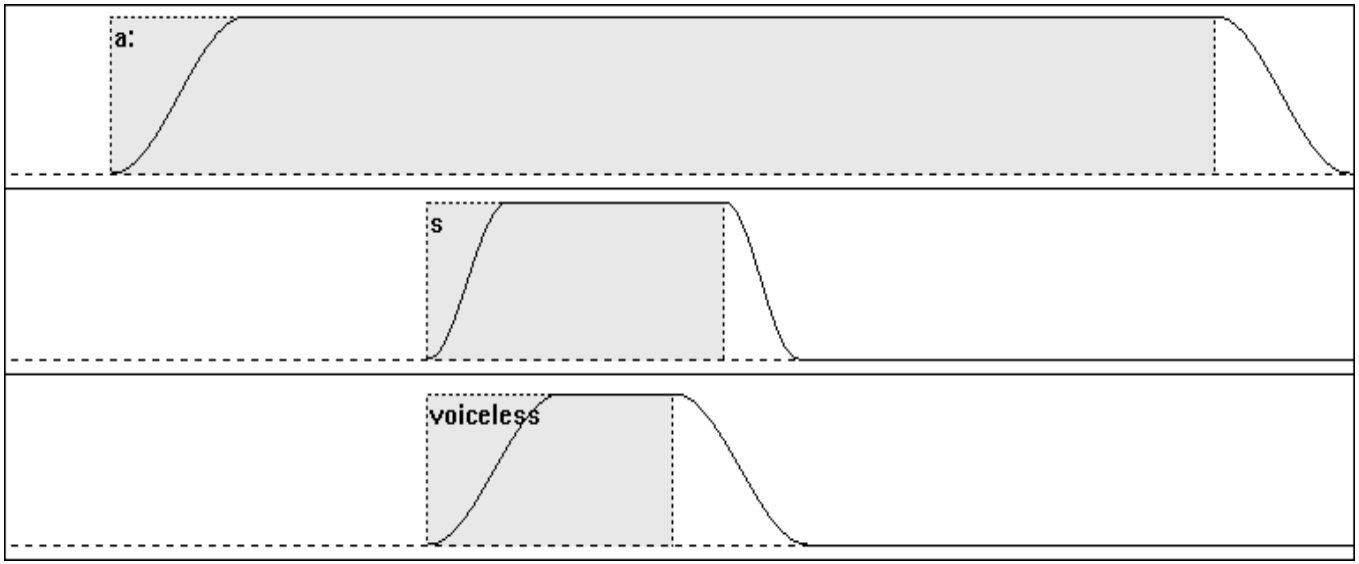
Consonants and vowels



vocalic gesture

consonantal gesture

glottal gesture

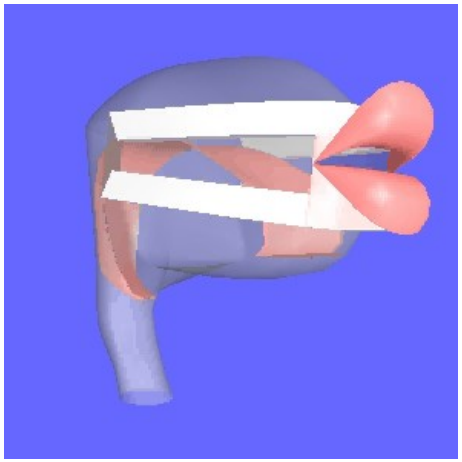


- Only the **targets** are specified, the transitions are calculated automatically. Sometimes the target **realizations** change due to the phonetic context (e.g. [g] target in [i:gi:] vs. [u:gu:])

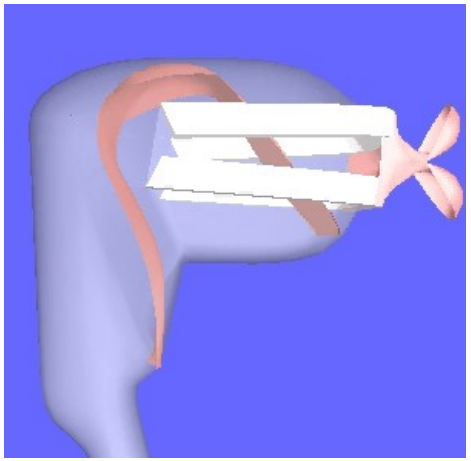
- [a:sa: i:si: u:su:]
- [aSa iSi uSu]

- more examples on simple gesture patterns ...

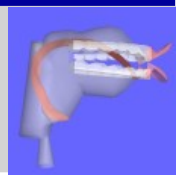
Single gestures: Lips



Single gestures: Velum



Gestural score



vocalic gestures

consonantal gestures

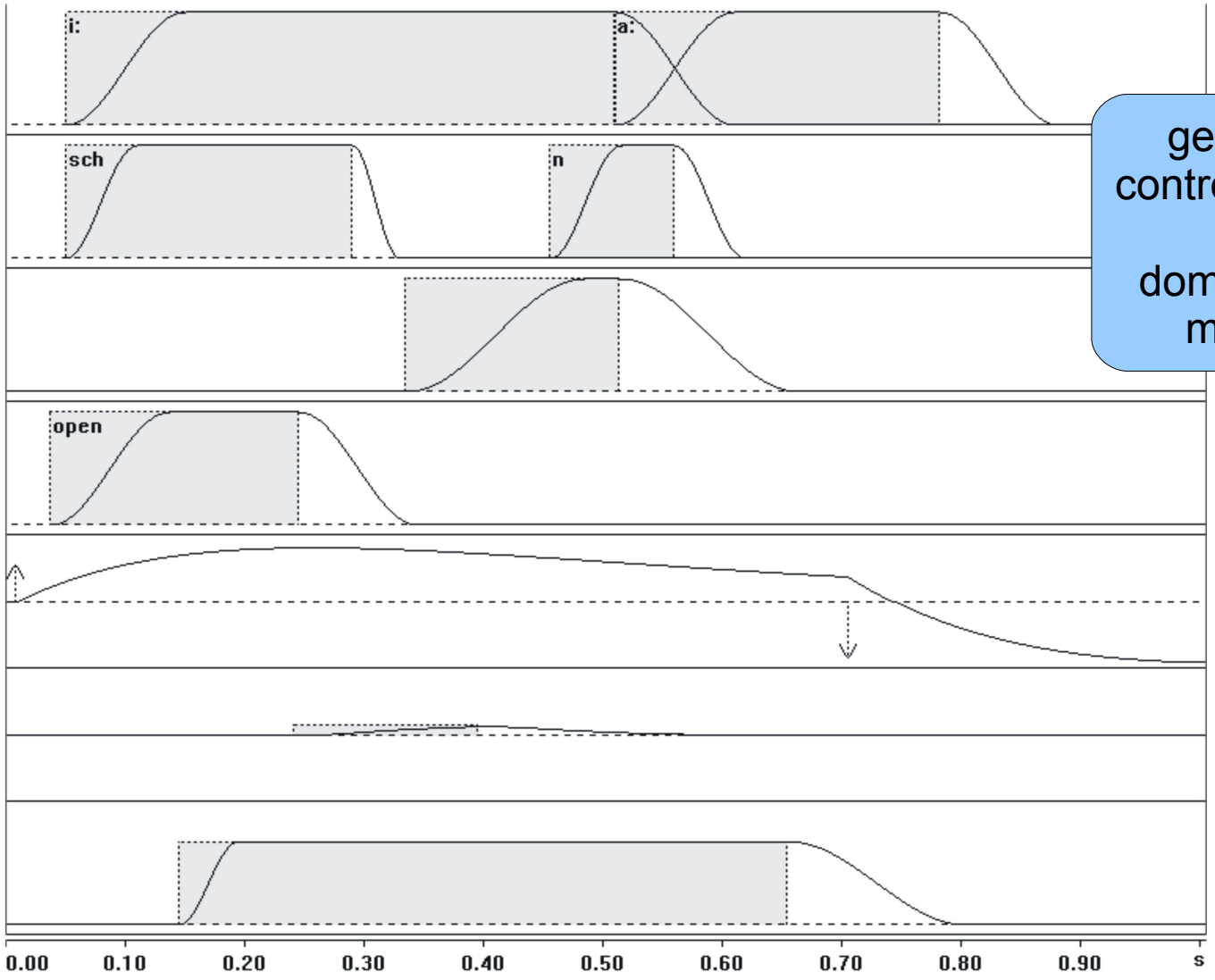
velic gestures

glottal gestures

F0 (pitch) gestures

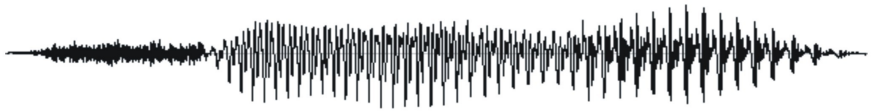
F0 (pitch) gestures

pulmonic gestures

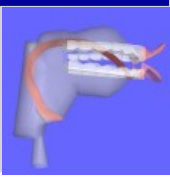


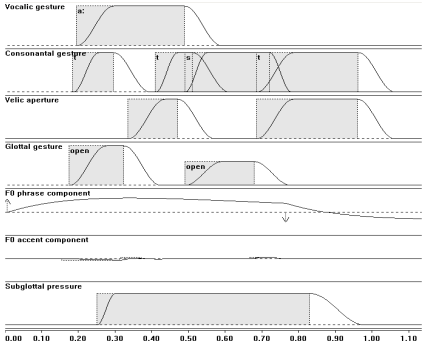
gestural control model + dominance model

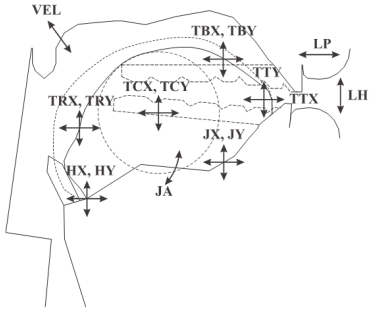
"China"



Look behind the graphical surface





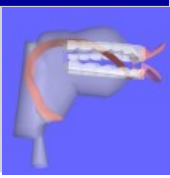


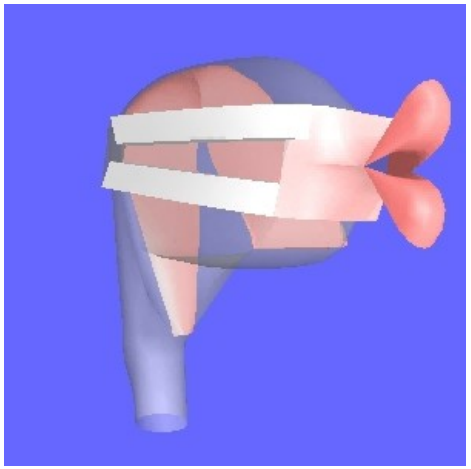

```
<gestural-score>
  <gesture time="0.1850" dur="0.5970"
     amp="800.0000" t-on="0.0500" t-
     off="0.1500„ type="PRESSURE"
     desc="" />
  <gesture time="0.1000" dur="0.8260"
     amp="1.0000" t-on="0.1000" t-off="0.1000"
     type="VOCALIC"
     desc="a:" />
  <gesture time="0.3700" dur="0.1300"
     amp="1.0000" t-on="0.1000" t-off="0.0700"
     type="CONSONANTAL" desc="p" />
  <gesture time="0.0150" dur="0.0000"
     amp="0.3960" t-on="2.0944"
     t-off="0.0000" type="F0-PHRASE"
     desc="test3" />
  <gesture time="0.7750" dur="0.0000"
     amp="-0.3000" t-on="2.0000"
     t-off="2.0000" type="F0-PHRASE" desc="" />
  <gesture time="0.4150" dur="0.0730"
     amp="-0.1000" t-on="0.1500" t-off="0.0500"
     type="F0-ACCENT" desc="" />
  <basis-f0 f0="80.0" />
</gestural-score>
```

```
<phoneme name="p">
  <param name="HX" value="0.4515" domi="0.0" />
  <param name="HY" value="-4.1888" domi="0.0" />
  <param name="JX" value="-0.0314" domi="75.0" />
  <param name="JY" value="-1.5691" domi="25.0" />
  <param name="JA" value="-0.0511" domi="25.0" />
  <param name="LP" value="0.0459" domi="50.0" />
  <param name="LH" value="-1.0000" domi="100.0" />
  <param name="VA" value="-0.8070" domi="100.0" />
  <param name="TCX" value="-0.7166" domi="25.0" />
  <param name="TCY" value="-1.9459" domi="25.0" />
  <param name="TCR" value="1.6955" domi="50.0" />
  <param name="TTX" value="3.9277" domi="50.0" />
  <param name="TTY" value="-2.0057" domi="50.0" />
  <param name="TBX" value="1.8430" domi="50.0" />
  <param name="TBY" value="-1.2070" domi="50.0" />
  <param name="TRE" value="-0.3822" domi="25.0" />
  <param name="TS1" value="0.0000" domi="50.0" />
  <param name="TS2" value="0.0000" domi="50.0" />
  <param name="TS3" value="0.0600" domi="50.0" />
  <param name="TS4" value="-0.0200" domi="50.0" />
  <param name="MA1" value="0.2000" domi="100.0" />
  <param name="MA2" value="0.2000" domi="100.0" />
  <param name="MA3" value="0.0000" domi="100.0" />
</phoneme>
```

[aba]

Illustrations of usage

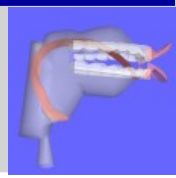




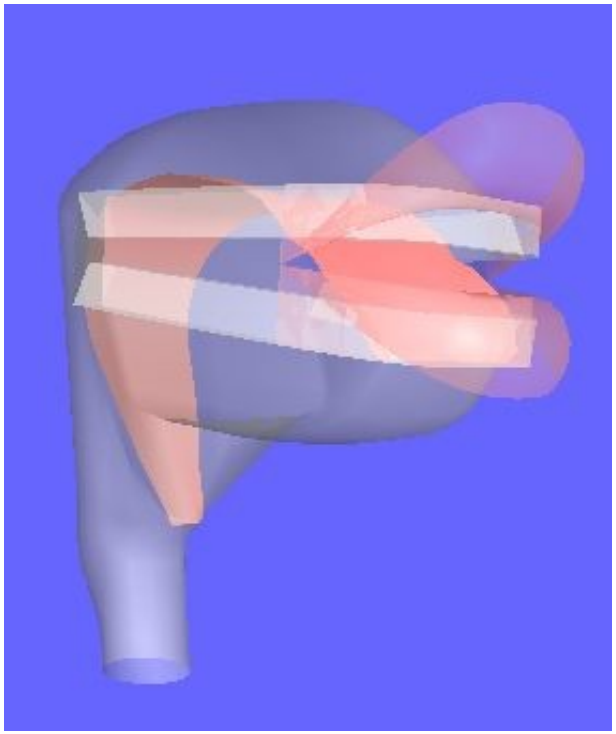
Der Zug hat eine Stunde Verspätung.

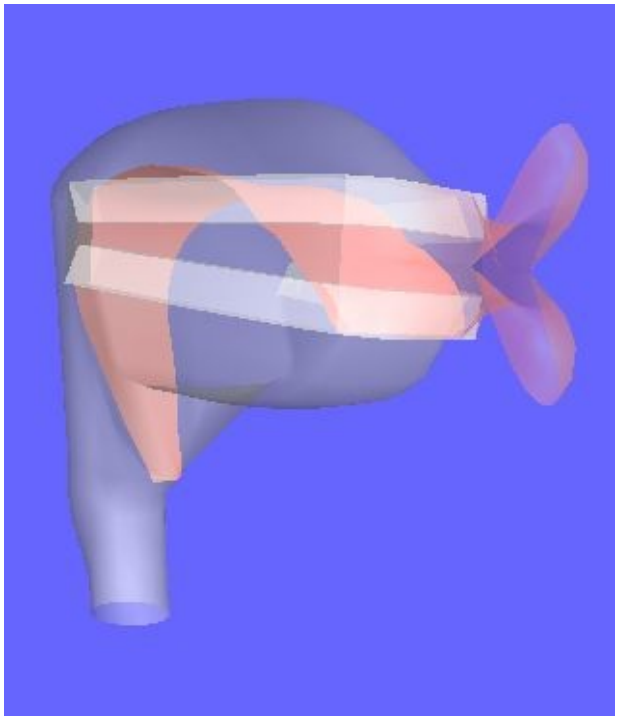
The train has a one hour delay.

Variation in speaking: Lip rounding/spreading



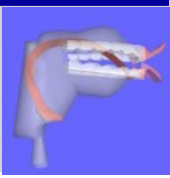
Wie geht‘s Ihnen?
(How are you?)





- happy/sad impression?

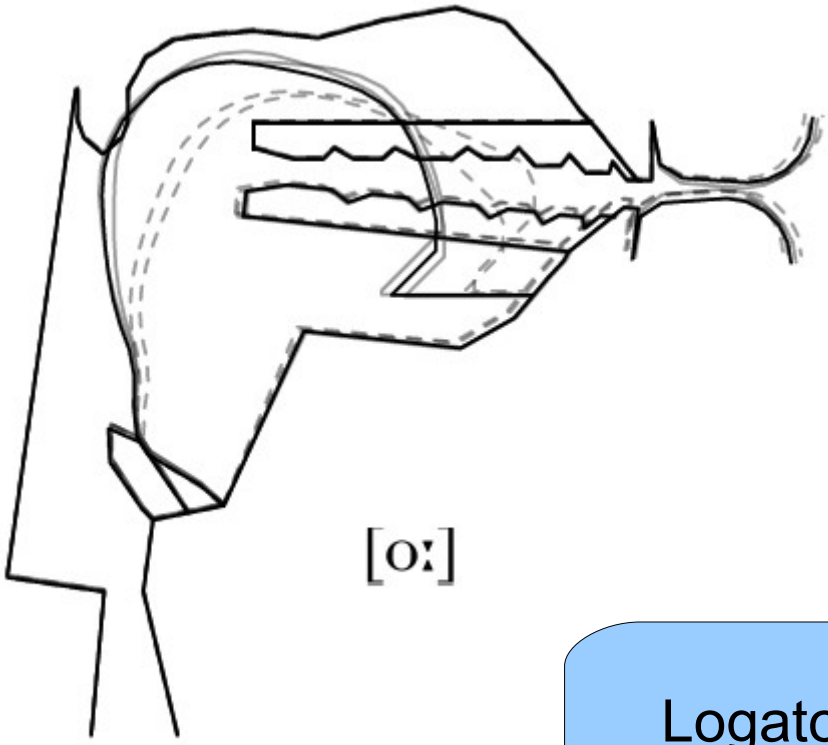
Variation in articulation: Regional accents



Real speaker
<< loben >>

Region 1

Region 2

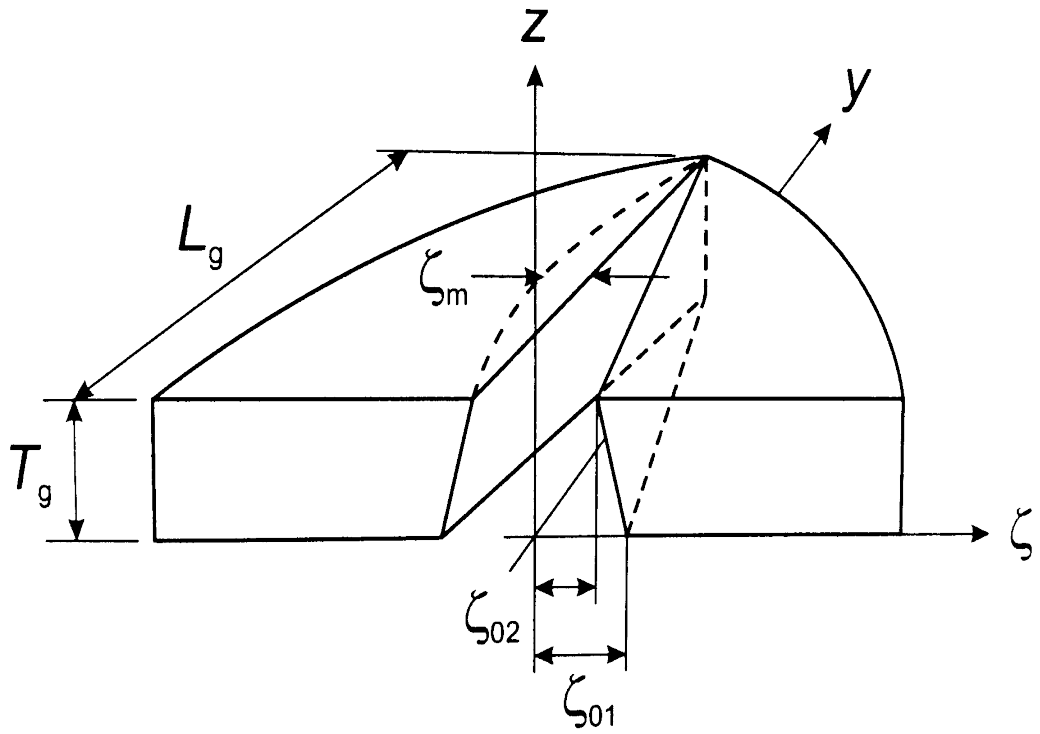


Logatome imitation

Region 1

Region 2

Variation in the voice (source): Aging

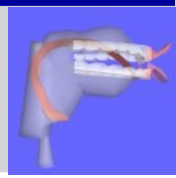


Age group 1

Age group 2

Age group 3

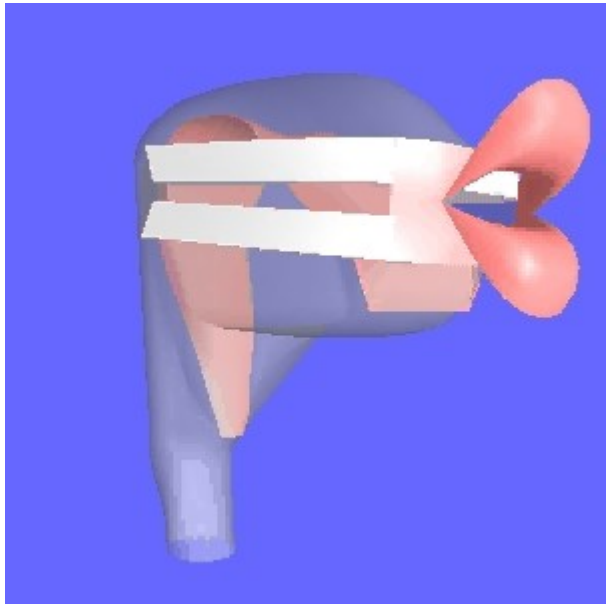
Different speaking rates

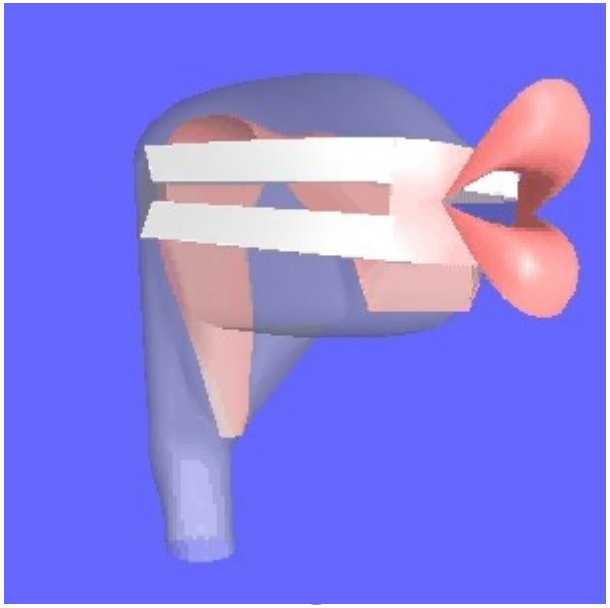


- change the time scale of the gestural score

Der Zug hat eine Stunde Verspätung.

The train has a one hour delay.

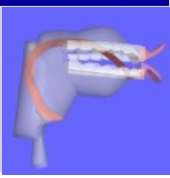




Visualization (speech therapy)

Yet somewhat slower...

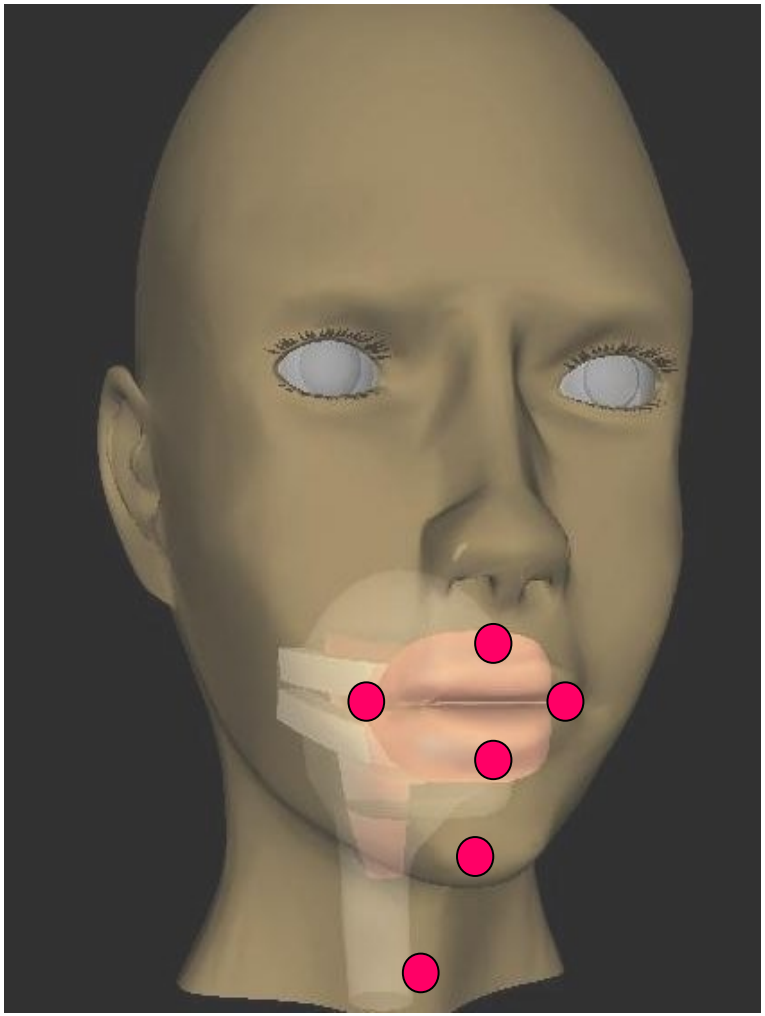
Singing



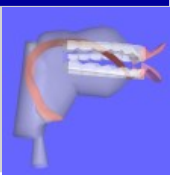
Dona Nobis Pacem

W. A. Mozart

Integration into animated faces

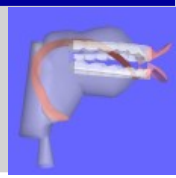


More adaptations

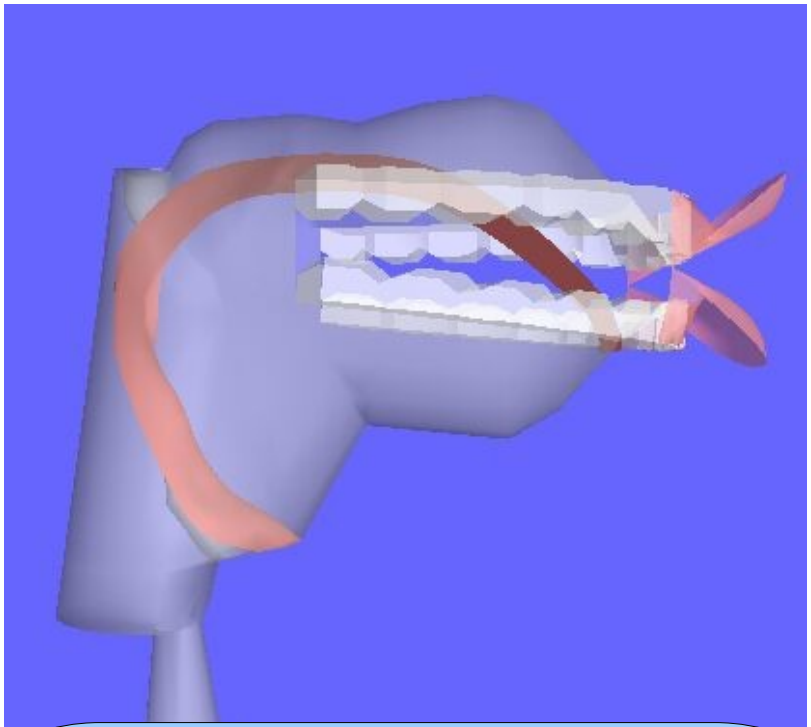


- More individual speakers
- Speaking styles
- Automatic Text-To-Speech component (gestural coordination)

All wishes coming true?



- Freely controllable – many parameters
  - speaking styles
  - emotions
  - speaking rate
  - specific speakers
  - any language
  - children's voices
  - singing
  - facial animation
- Sounds okay, intelligible ...



Still a lot to discover and develop.

Research tool.

Commercial synthesis (future?)

VocalTractLab

Towards high-quality articulatory speech synthesis

vocaltractlab.de
by
Peter Birkholz

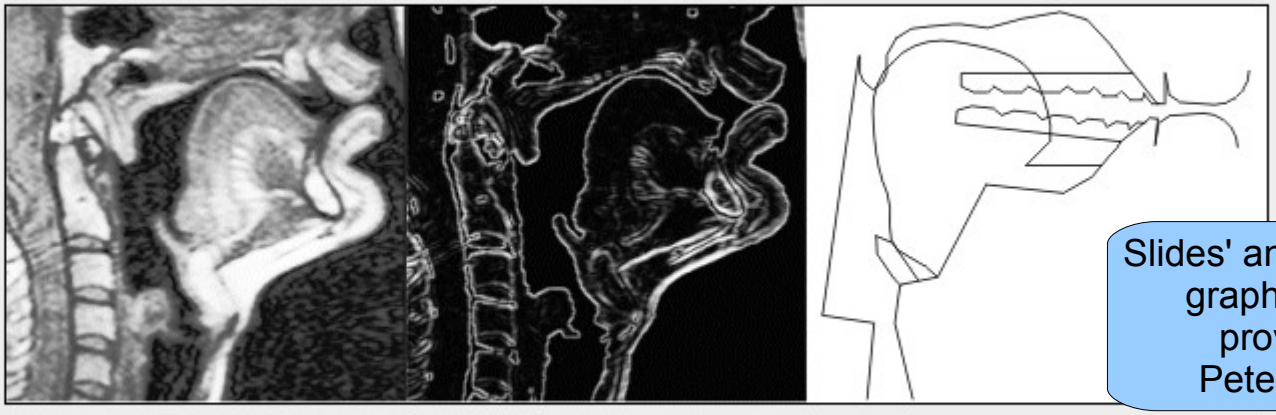
Content

Welcome to VocalTractLab

This web site features:

- **VocalTractLab:** An interactive multimedia application for the simulation and study of speech production. This software contains a complete articulatory speech synthesizer!
- **Image3D:** A tool for the exploration of volumetric magnetic resonance images and the tracing of outlines in these images.
- **Background information** about articulatory speech synthesis and the models and methods implemented in VocalTractLab.

We hope that this web site and the software made available here will help interested people to improve their understanding of the human vocal system and the principles of speech production.



Slides' animations and graphics mainly provided by Peter Birkholz

This web site was created by Peter Birkholz who is working in the field of articulatory speech synthesis since about 1999. The software featured at this site was originally developed as a research tool and is continually extended due to ongoing research.

ty 2010