Advanced Signal Processing 2

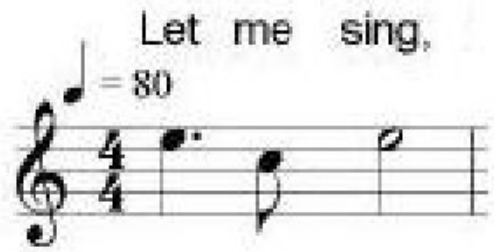
Synthesis of Singing

Outline

- Features and requirements of signing synthesizers
- HMM based synthesis of singing
- Articulatory synthesis of singing
- Examples

Requirements of a singing synthesizer

- Integration of musical score
  - Note properties
    - Pitch
    - Duration
  - Integration can be done manually or automatically



- Synthesis of (singing) speech sounds
  - Direct synthesis of singing
  - Conversion from spoken synthetic speech
- Modeling singing effects
  - Vibrato
  - Overshoot, etc.
- Addition / Improvement of naturalness

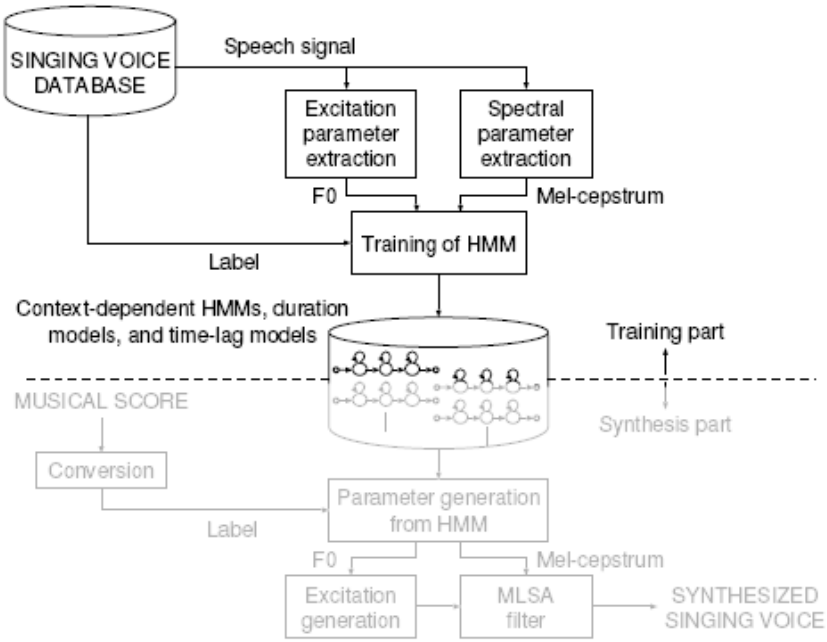
HMM-based synthesis of singing

- Based on [1] and [2]

- Unit-selection for singing would require vast amount of recorded data
- HMM-based system by relatively little training data
- System is similar to HMM-based speech synthesis
- Two main differences:
  - Contextual factors
  - “Time-lag” - modeling

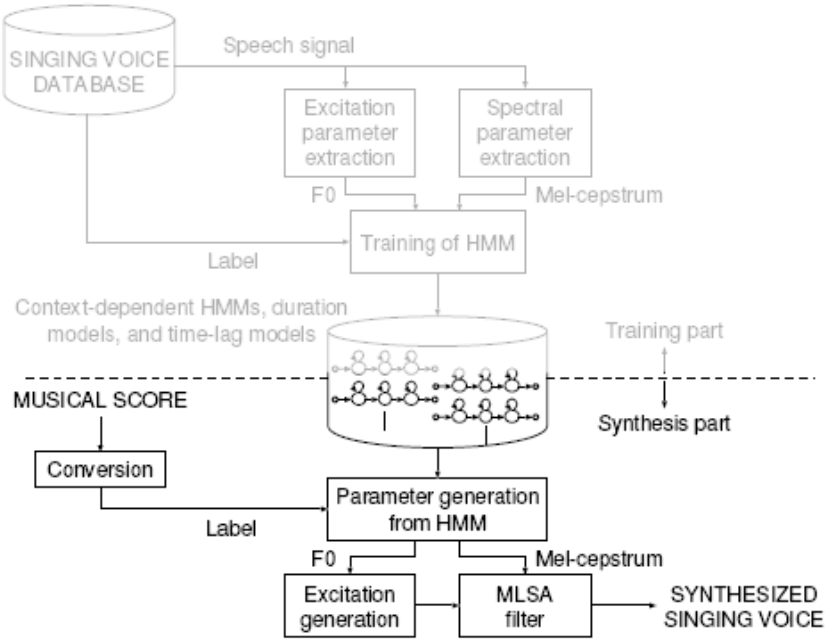
HMM-based synthesis – Overview (Analysis)

- Parameter extraction
  - Mel-cepstral coefficients
  - Fundamental frequencies
- Training/Estimation of
  - Context-dependent HMMs
  - State duration models
  - Time-lag models



HMM-based synthesis – Overview (Synthesis)

- Musical score → context-dependent label sequence
- Song HMM = concatenation of context-dep. HMMs
- Determination of state durations (time-lag models!)
- Generate speech parameters from HMMs
- Synthesize speech by MLSA-filter

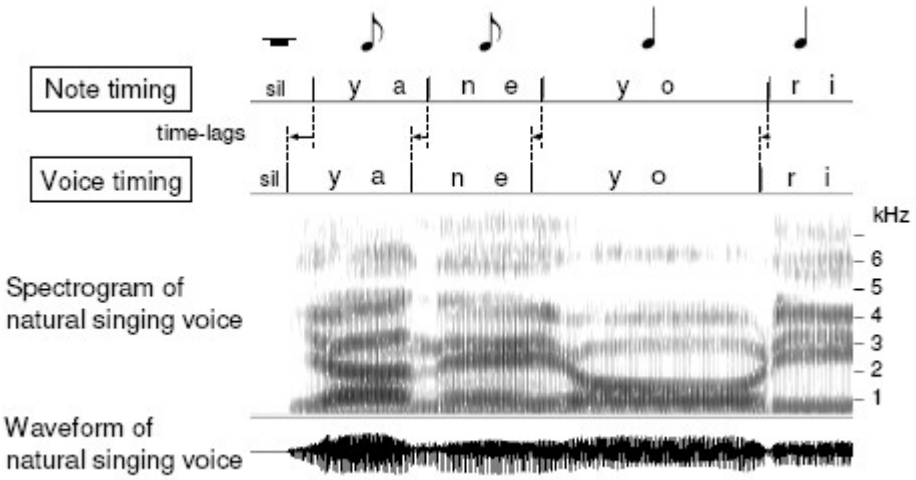


HMM-based synthesis - Contextual factors

- Different from those in synthesis of reading speech
- This method uses:
  - Phoneme
  - Tone (musical notes like “A4” or “C5#”)
  - Duration of notes (in units of 100ms)
  - Position in the current musical bar
  - For all of them: preceding, current and succeeding one is taken into account
- Determined automatically from score and lyrics

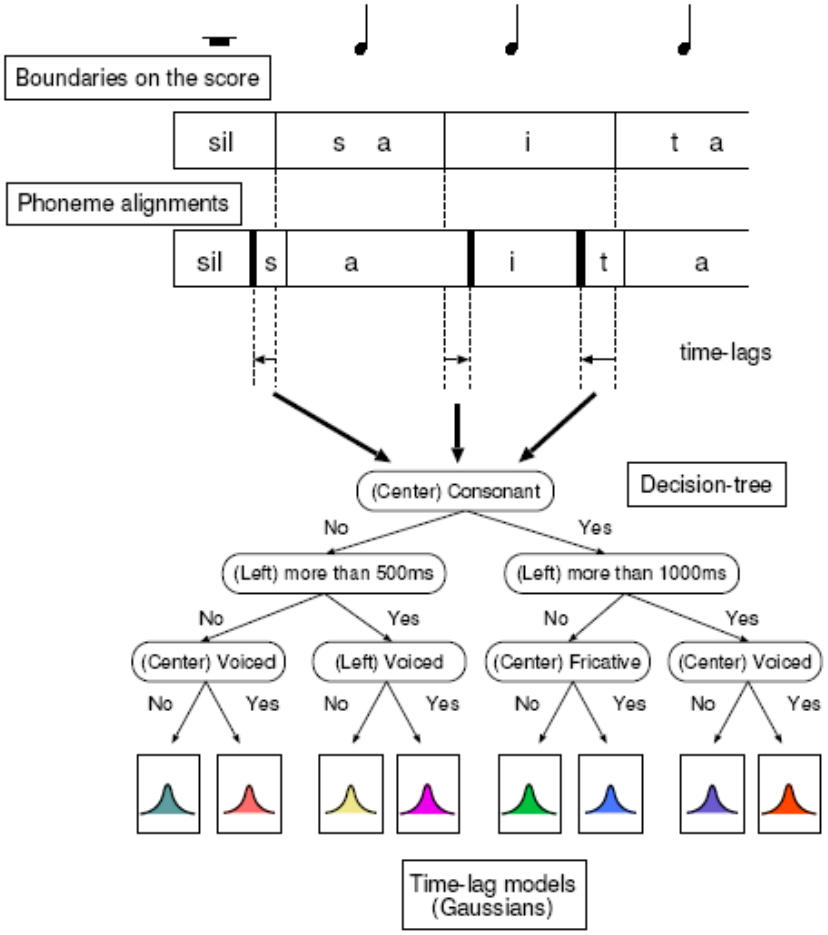
HMM-based synthesis - Time-lag modeling (1)

- Strictly following score will sound unnatural
- Lags between start of notes and speech



HMM-based synthesis - Time-lag modeling (2)

- Context-dependent labels are assigned
- Clustered by decision tree
  - Result: Decision tree-clustered context-dependent time-lag models
  - One-dimensional Gaussians



HMM-based synthesis - Time-lag modeling (3)

- At synthesis stage:
- Determination of each note duration from score
- Simultaneously determine time-lags and state durations
- The joint probability has to be maximized

$$P(\underline{d}, \underline{g}|\underline{T}, \underline{\Lambda}) = P(\underline{d}|\underline{g}, \underline{T}, \underline{\Lambda}) \cdot P(\underline{g}|\underline{\Lambda})$$
$$= \prod_{k=1}^{N} P(\underline{d_k}|T_k, g_k, g_{k-1}, \Lambda) \cdot P(g_k|\Lambda)$$

- d_k - state durations of k^{th} note, g_k – time-lag (of start timing of $k+1^{th}$ note), T_k – duration of k^{th} note from score
- → leads to a set of linear equations

HMM-based synthesis – Experimental evaluation

- Self-recorded singing database with manual corrections
- Results: “smooth and natural-sounding”
  - Time-lag models substantially improved quality
  - Characteristics of original singer found in synthesized voice
- Samples

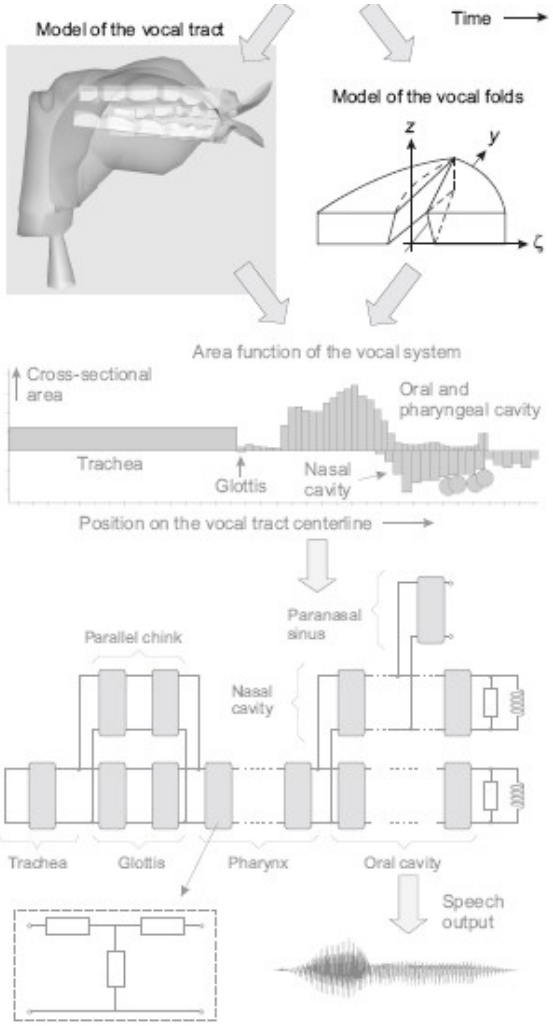
Articulatory synthesis of singing

- Based on [3] and [4]

- Completely different approach
- Main features:
  - Complex three-dimensional model of vocal tract
  - Sound synthesis by simulation of this model
  - Input of the system is a “gestural score”
- Extension of an existing speech synthesizer
  - Transformation of musical score into gestural score
  - Pitch-dependent articulation of vowels

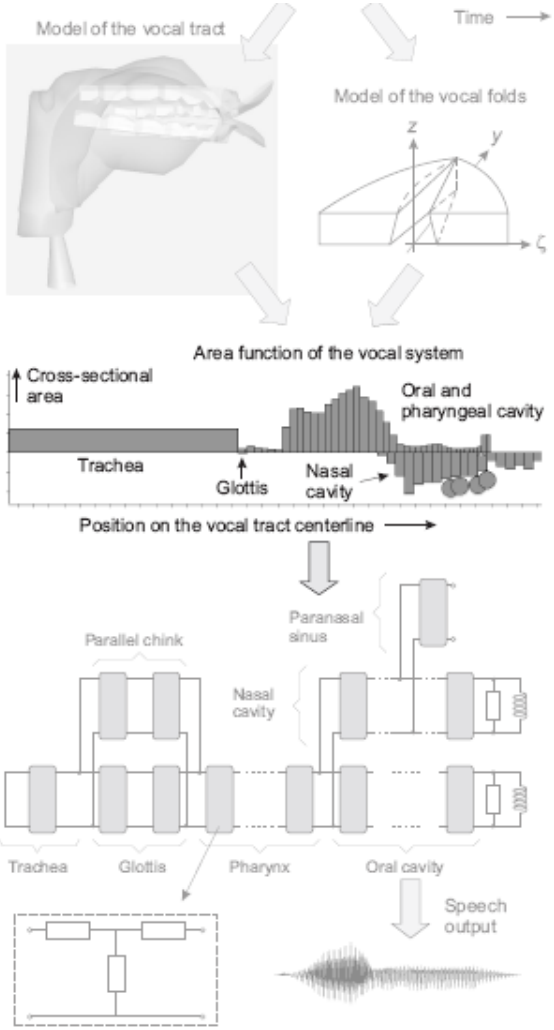
Articulatory synthesis – Overview (1)

- 3D wireframe representation of male vocal tract
- Parameters determined by MRI-images for German vowels and consonants
- Shape and position of movable structures is a function of 23 parameters
- Dynamic MRI-data used for coarticulated consonants



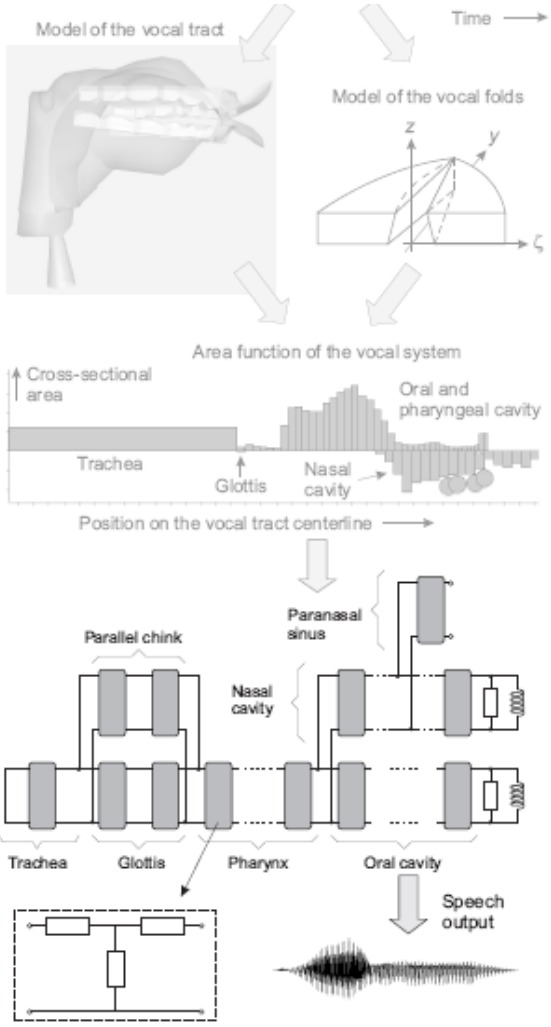
Articulatory synthesis – Overview (2)

- Acoustical simulation by branched tube model
- Short abutting elliptical tube sections
- Represented by an area function and a discrete perimeter function



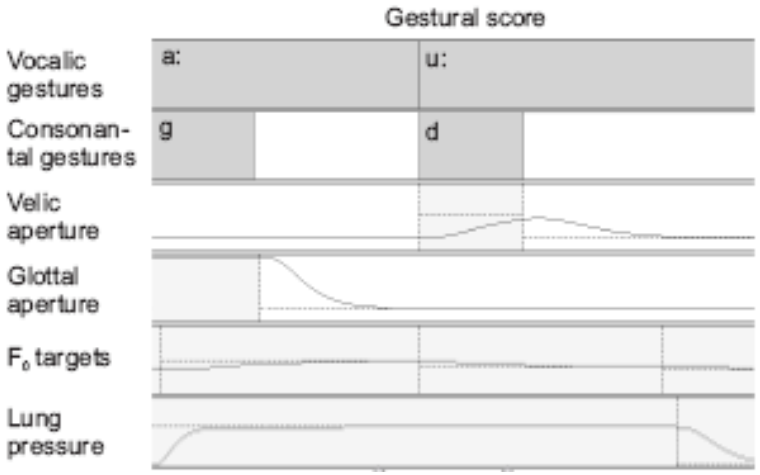
Articulatory synthesis – Overview (3)

- Analogy of acoustical and electrical transmission
- Branched tube model represented by inhomogeneous transmission line circuit with lumped elements
- Each tube section → Two-port T-type network, elements are function of tube geometry
- Simulated by finite difference equations in time domain
- Additional techniques to simulate several types of losses
- All major speech sounds possible



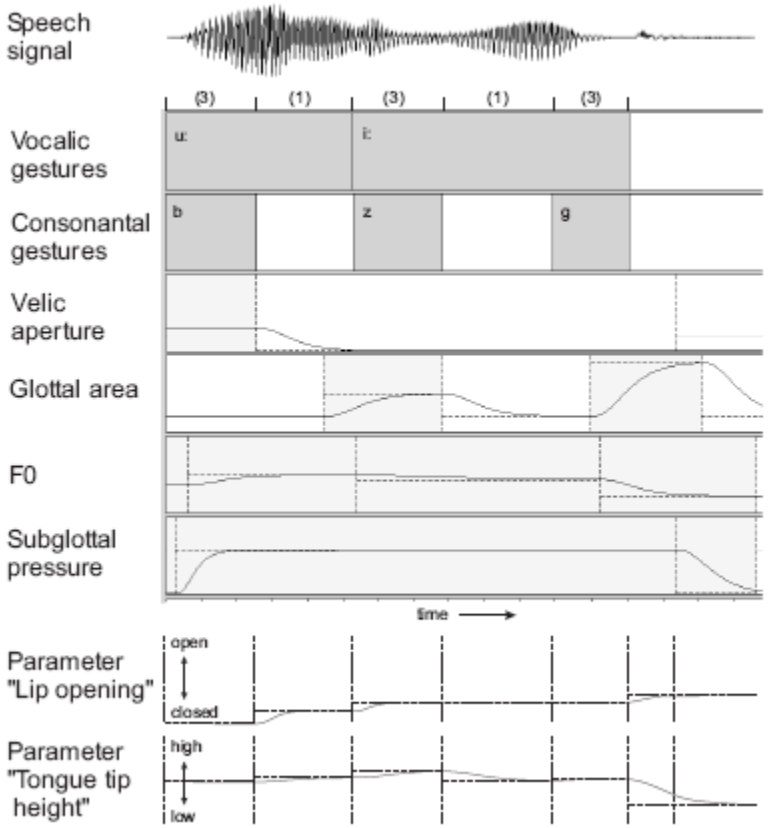
Articulatory Synthesis – Gestural Score (1)

- Is the input of the synthesizer
- Generation of parameters for vocal tract model
- Utterances represented by patterns of articulatory gestures
- Gestures are “goal-oriented articulatory movements” (What has to be done by vocal tract, but not how)
- Six types of gestures
- How to obtain: Transformation of defined XML-format for songs



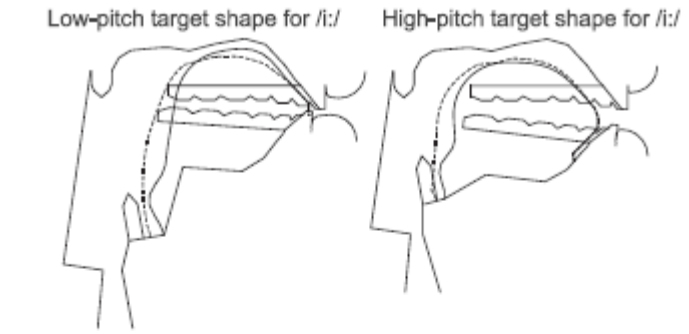
Articulatory Synthesis – Gestural Score (2)

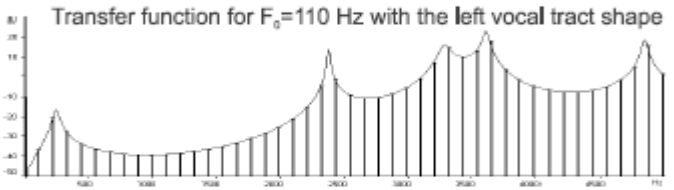
- Example: [mu:zi:k]
- Only one configuration for the group {[b], [p], [m]}
- Consonant and vowel intervals overlap
  → coarticulation
- Two lowest rows are examples for target functions of vocal parameters
- They are called “motor commands”
- Realized by third-order dynamical systems

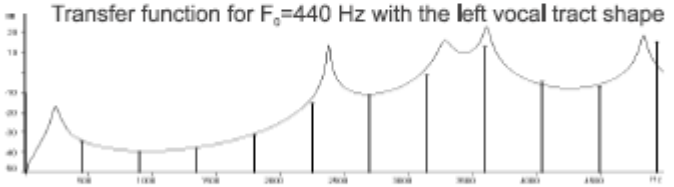


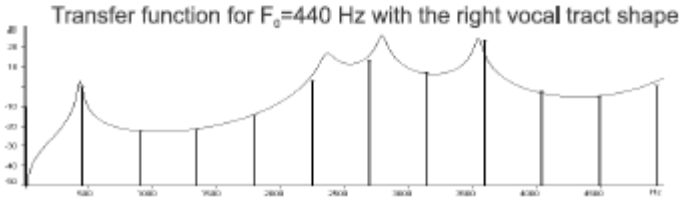
Articulatory Synthesis – Pitch dependent vocal tract target shapes

- Vocal tract shape for the same vowel depends on pitch
- Vowels at higher pitches are sung more “open”
- Tuning of vocal tract formants is necessary
- First formant should match first harmonic voice source
- Two “extreme” shapes for 110 Hz and 440 Hz
- Linear interpolation in between
- Low pitch shape is the one for speech synthesis









Conclusion

- The challenge of synthesizing singing
- Giving a speech synthesizer the ability to sing
- Two very different approaches
  - HMM-based
  - Articulatory synthesis

Singing samples

- HMM-based Singing Synthesis
- Synthesis of Singing Challenge 2007, Belgium

References

- [1] Keijiro Saino, Heiga Zen, Yoshihiko Nankaku, Akinobu Lee, Keiichi Tokuda: “An HMM-based Singing Voice Synthesis System”, 2006
- [2] Takayoshi Yoshimura, Keiichi Tokuda, Takashi Masuko, Takao Kobayashi, Tadashi Kitamura: “Simultaneous Modeling of Spectrum, Pitch and Duration in HMM-based Speech Synthesis”, 1999
- [3] Peter Birkholz: “Articulatory Synthesis of Singing”, 2007
- [4] Peter Birkholz, Ingmar Steiner, Stefan Breuer: “Control Concepts for Articulatory Speech Synthesis”, 2007

Thank you for your attention!