TTÜ KÜBERNEETIKA INSTITUUT
Institute of Cybernetics at TUT

Towards Audiovisual TTS in Estonian

Einar MEISTER [a], Sascha FAGEL [b] and Rainer METSVAHI [a]

[a] Institute of Cybernetics at Tallinn University of Technology, Estonia
[b] zoobe message entertainment GmbH, Berlin, Germany

Background

- The nature of human speech is essentially bimodal involving auditory and visual information
- The contribution of visual information in the perception of different sounds is variable – visual information is most important in distinguishing labiodentals and bilabials (e.g. /t/ and /p/), but e.g. the bilabials /p/ and /m/ cannot be distinguished visually
- An example of the complementary nature of multimodal speech perception is the **McGurk illusion** in which an audio /ba/ paired with a video /ga/ is heard as /da/

TTÜ KÜBERNEETIKA INSTITUUT
Institute of Cybernetics at TUT

Background

- Audiovisual text-to-speech synthesis (AVTTS) is a technique for automatic generation of voice-synchronized facial animations from an arbitrary text
- AVTTS is applied as virtual talking heads, animated tutors, aids for the hearing impaired, multimodal interfaces, etc
- The list of languages involved include English, German, French, Italian, Spanish, Swedish, Mandarin Chinese, Japanese, Danish, Czech, etc
- Previously no work on AVTTS has been done for Estonian

Methods of AVTTS

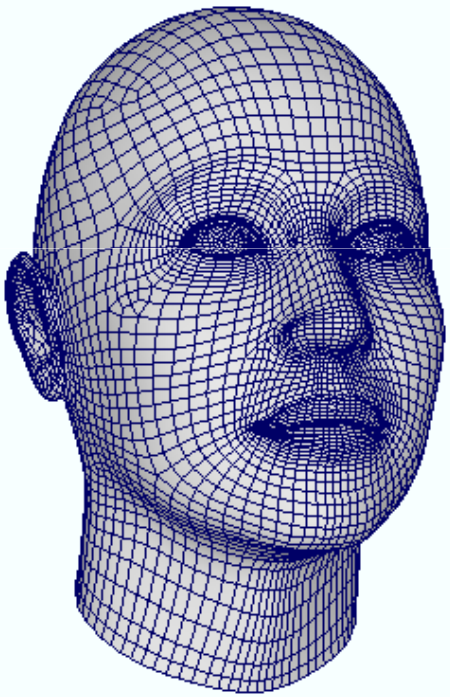
AVTTS involves two main components:

(1) a text-to-speech module to produce synthetic audio of the input text

(2) a face model producing visible articulatory movements synchronized with audio

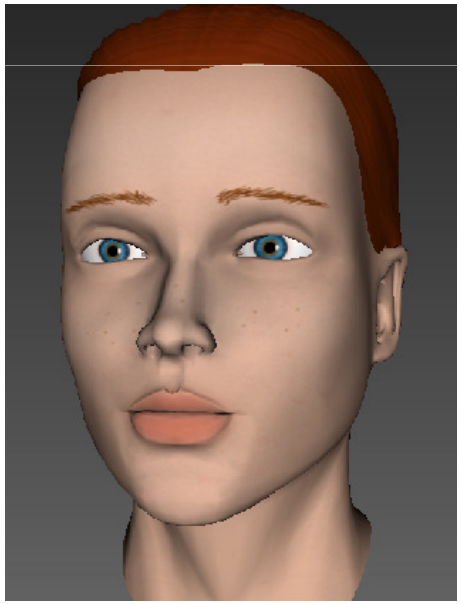
Parametric models

- **Parke's model** (1982) – implemented as a limited number of meshes representing the face topology, and controlled by a small set of parameters to move the lips and the jaw



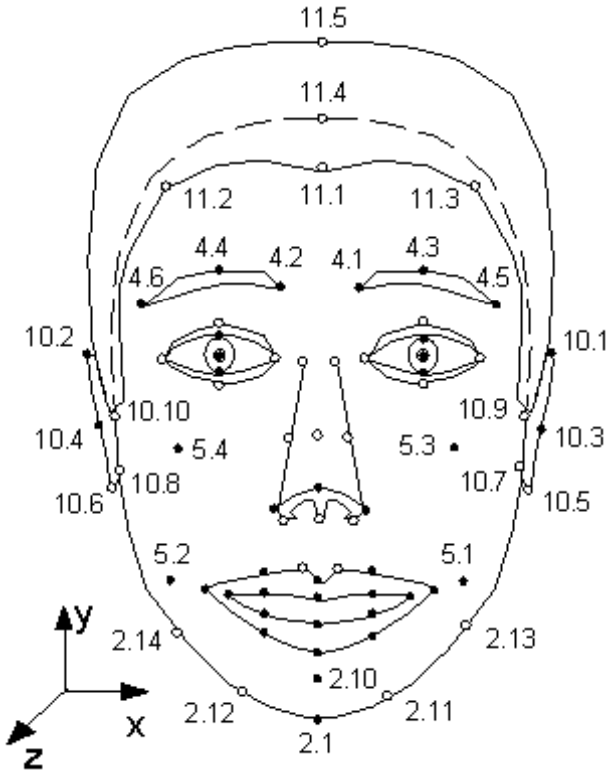
Descendants of Parke's model

BALDI, MASSY, LUCIA, models at KTH


Parametric models

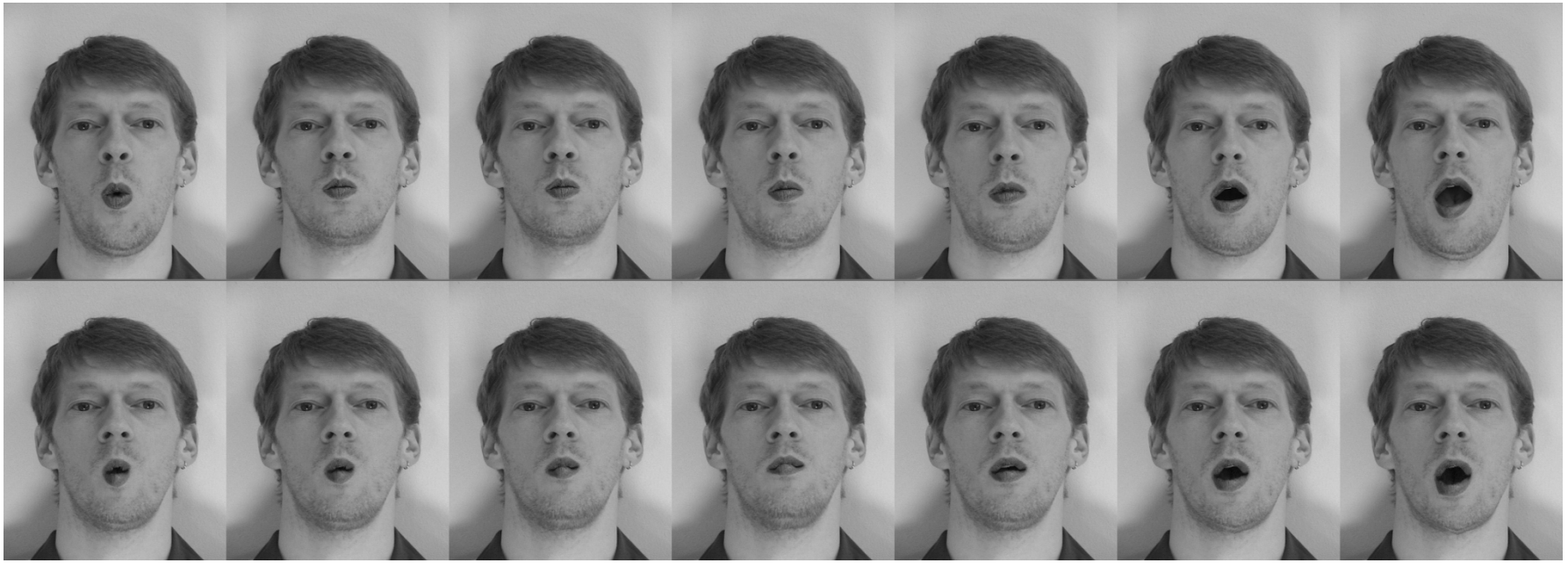
- Some models are compatible with the industrial **MPEG-4 standard** that has defined 84 Feature Points (FPs) to enable the animation of a face model
- The FPs are controlled by a set of Facial Animation Parameters (FAPs) which represent a set of basic facial actions – head motion, tongue, eye and mouth control
- FAPs do not provide enough freedom to control the lips and the jaw



TTÜ KÜBERNEETIKA INSTITUUT
Institute of Cybernetics at TUT

Image-based approach

- Exploits a large set of annotated video recordings and concatenates single images or video sequences to produce the appropriate visual output



Corpus-based AV synthesis

- Extends the unit selection strategy developed initially for audio speech synthesis to audiovisual speech
- Needs a large appropriately annotated bimodal corpus in which the acoustic and visual components are kept together
- The system searches in the corpus for the most suitable sequence of audiovisual segments that match the target speech
- This results in maximal coherence between the two components of bimodal speech and avoids perceptual ambiguity

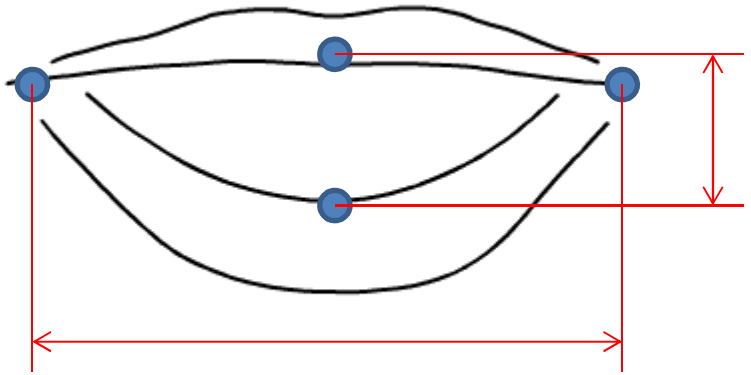
Phonemes and visemes

- Acoustically (and articulatorily) close phones are grouped to **phonemes**
- Visually (nearly) indistinguishable phones are grouped to **visemes** – one viseme represents the configuration of articulators corresponding to one or more phones
- Estonian has 9 vowel and 17 consonant phonemes
- How many visemes and how they map to phonemes?

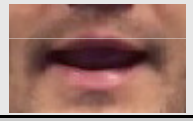
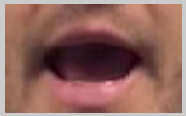
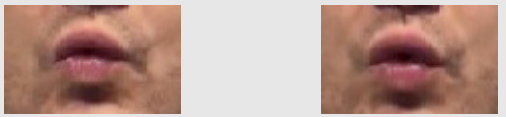
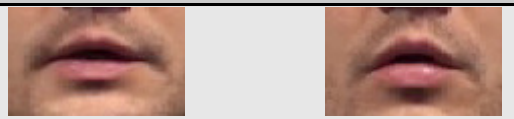
TTÜ KÜBERNEETIKA INSTITUUT
Institute of Cybernetics at TUT

Preliminary phoneme-to-viseme mapping

- Carried out using video recordings of a native male speaker
- The speaker read a list of two-syllable CVCV words containing CVC and VCV combinations of Estonian phonemes
- For each phoneme a frame representing the typical visual pattern of the mouth and the lips was extracted
- Lip width and mouth opening were measured
- 12 visemes defined

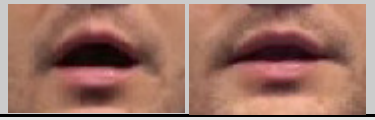
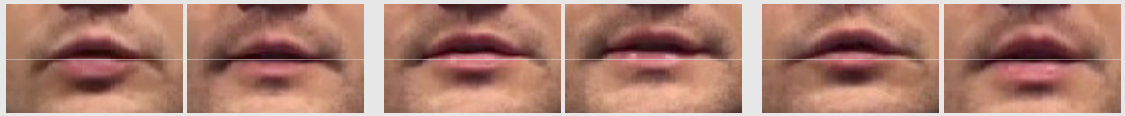
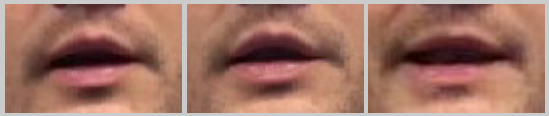
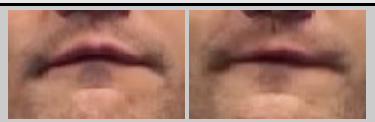
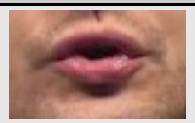


Preliminary phoneme-to-viseme mapping

Viseme number	Phonemes in SAMPA	Typical mouth and lip patterns
1	i, j	
2	e	
3	{	
4	u, y	
5	o, 2	
6	7, k	

TTÜ KÜBERNEETIKA INSTITUUT
Institute of Cybernetics at TUT

Preliminary phoneme-to-viseme mapping

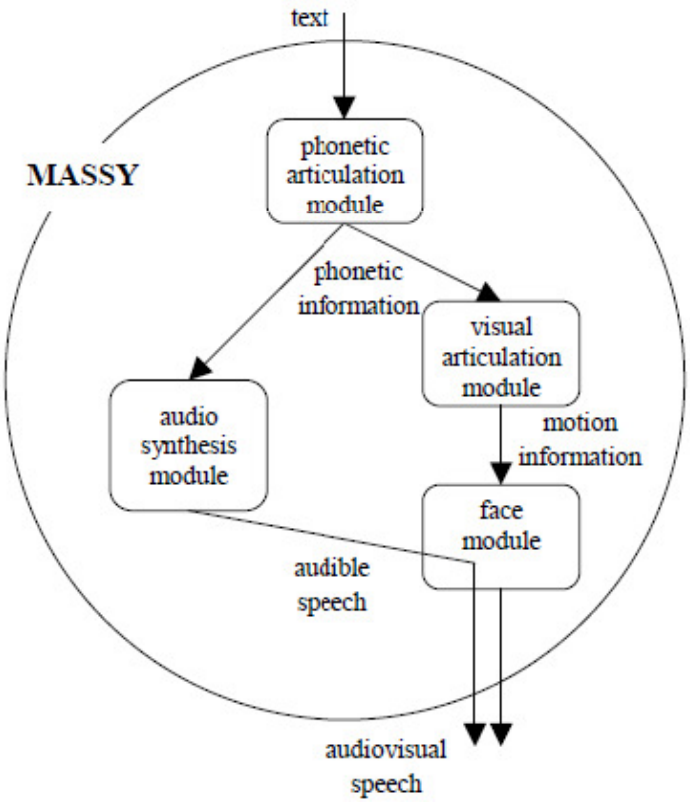
Viseme number	Phonemes in SAMPA	Typical mouth and lip patterns
7	A, h*	
8	t, t', s, s', n, n'	
9	r, l, l'	
10	m, p	
11	v, f	
12	S	

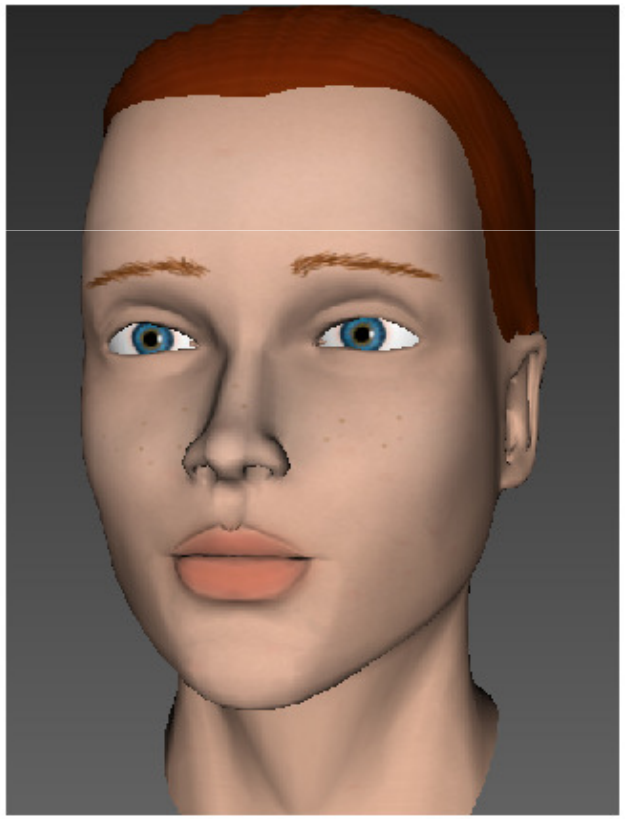
TTÜ KÜBERNEETIKA INSTITUUT
Institute of Cybernetics at TUT

MASSY model

- Developed by Sascha Fagel, originally for German, has been adapted to English, as well
- Four basic modules:
  - text-to-phoneme ⇐ Estonian txt2pho
  - audio synthesis (MBROLA) ⇐ Estonian diphone DB
  - visual articulation ⇐ Estonian viseme DB
  - virtual 3D face
  - 3D face model is implemented in VRML (Virtual Reality Modeling Language)

MASSY model

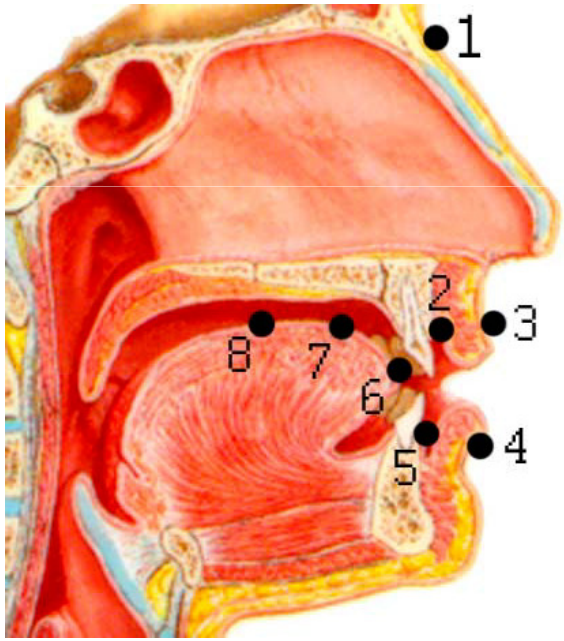

text
MASSY
phonetic articulation module
phonetic information
visual articulation module
audio synthesis module
motion information
face module
audible speech
audiovisual speech



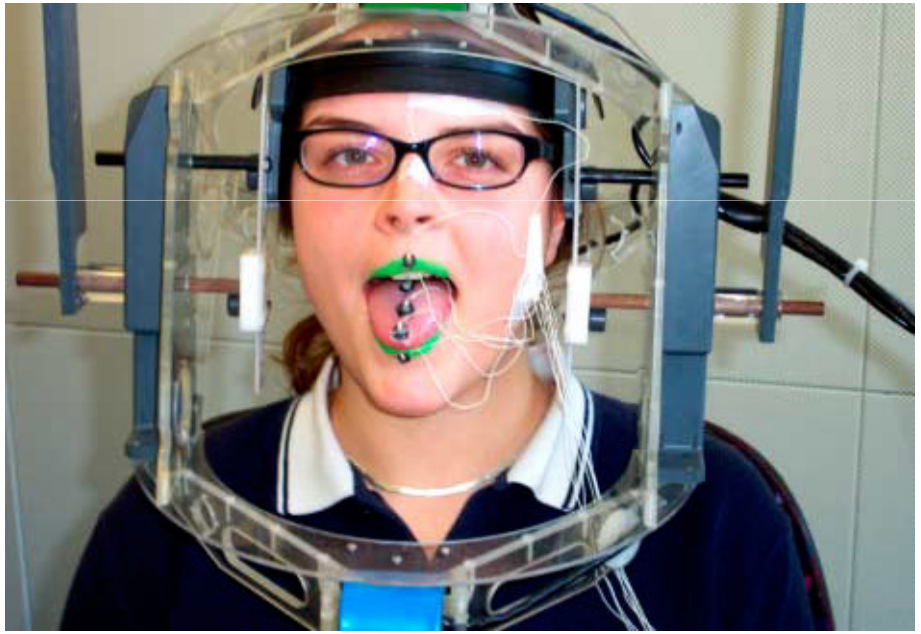
Visual articulation module

- Viseme targets are defined by six motion parameters of the face model:
  - lip width
  - jaw height
  - lip height
  - tongue tip height
  - tongue back height
  - lower lip retraction

EMA measurements of target positions



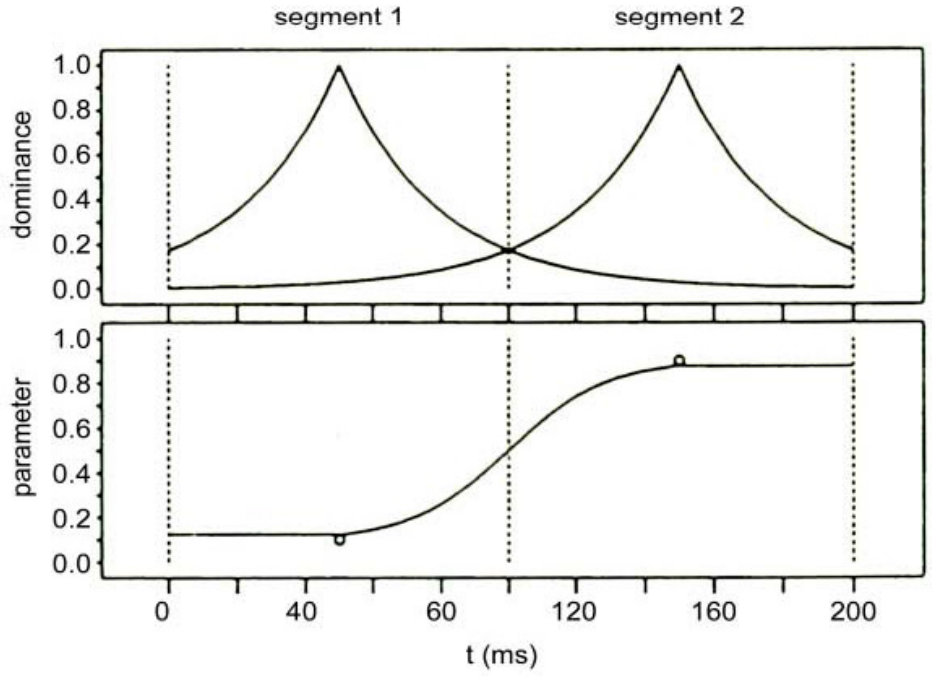
Sensor positions



EMA Carstens AG 100

Dominance model

- Motion parameters are generated with dominance model taking into account the target position and the strength (the dominance)

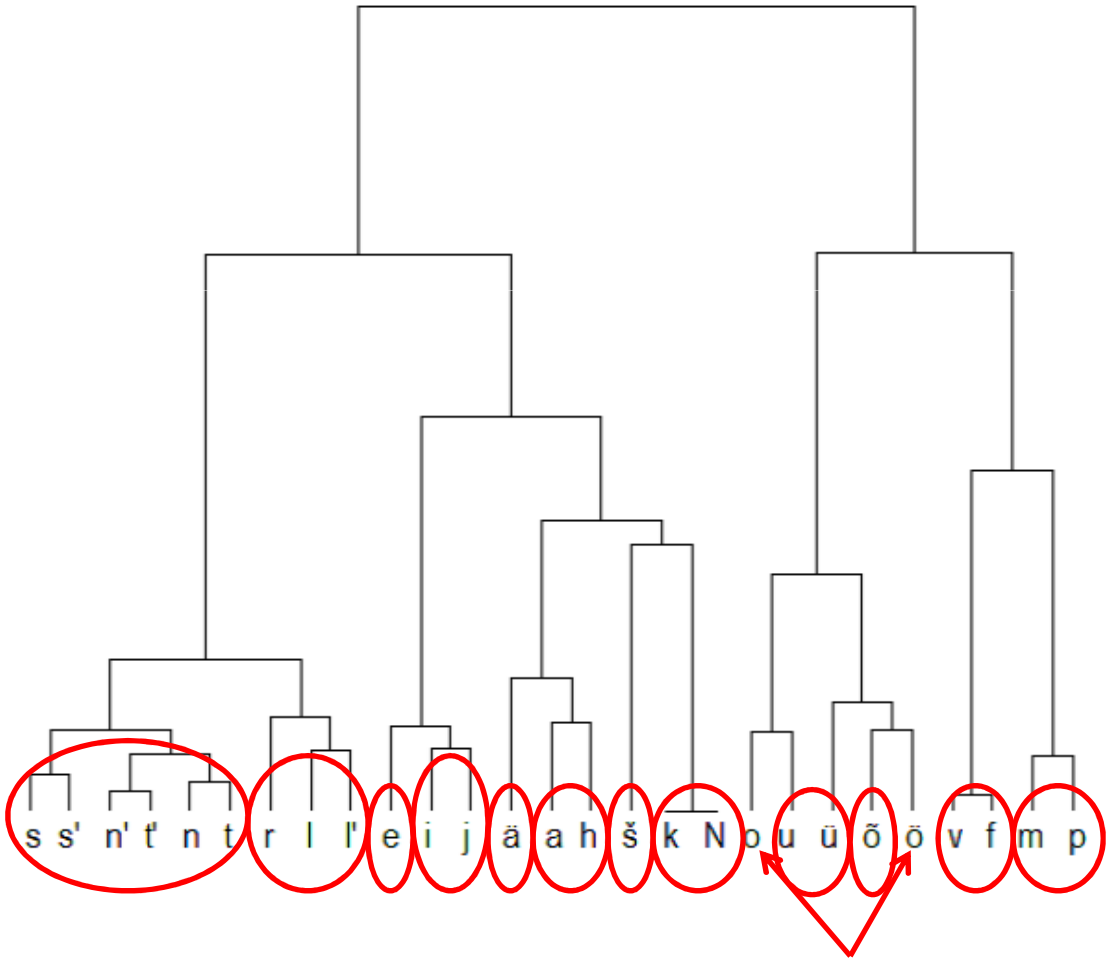


Adapting MASSY for Estonian

- Target positions of six motion parameters for Estonian visemes were determined:
  - lip width and jaw height (mouth opening) were measured from video recordings
  - other parameters derived from the German dominance model
  - further tuning was carried out in a series of live experiments involving different vowel and consonant combinations synthesized with the adapted prototype

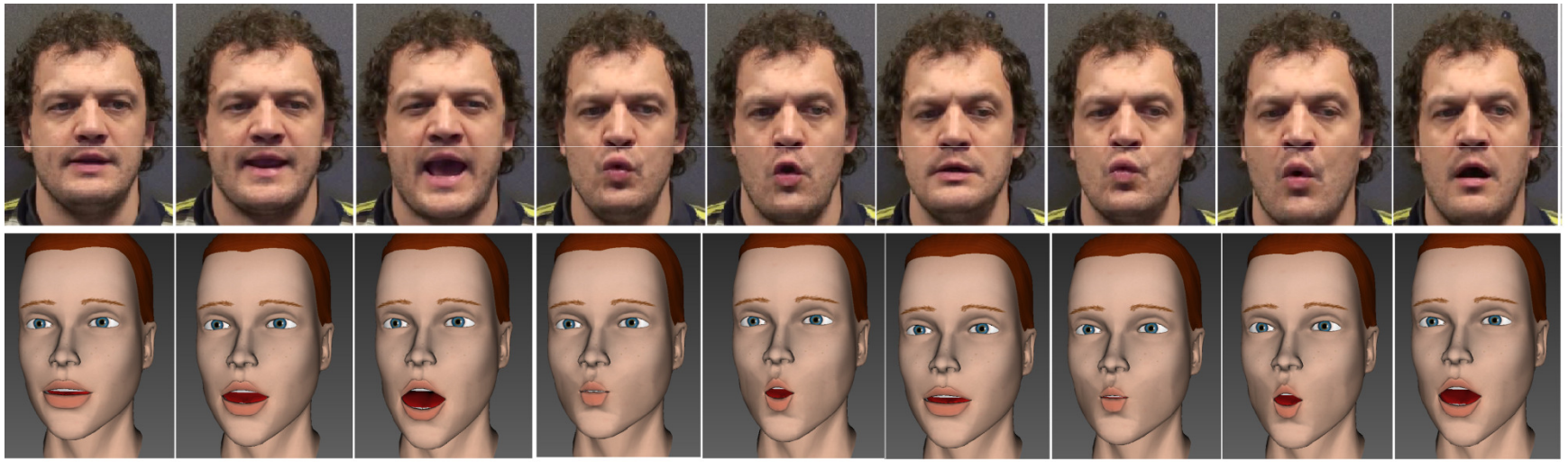
TTÜ KÜBERNEETIKA INSTITUUT
Institute of Cybernetics at TUT

Cluster analysis of Estonian phonemes on the basis of articulatory features



TTÜ KÜBERNEETIKA INSTITUUT
Institute of Cybernetics at TUT

Face patterns of Estonian vowels



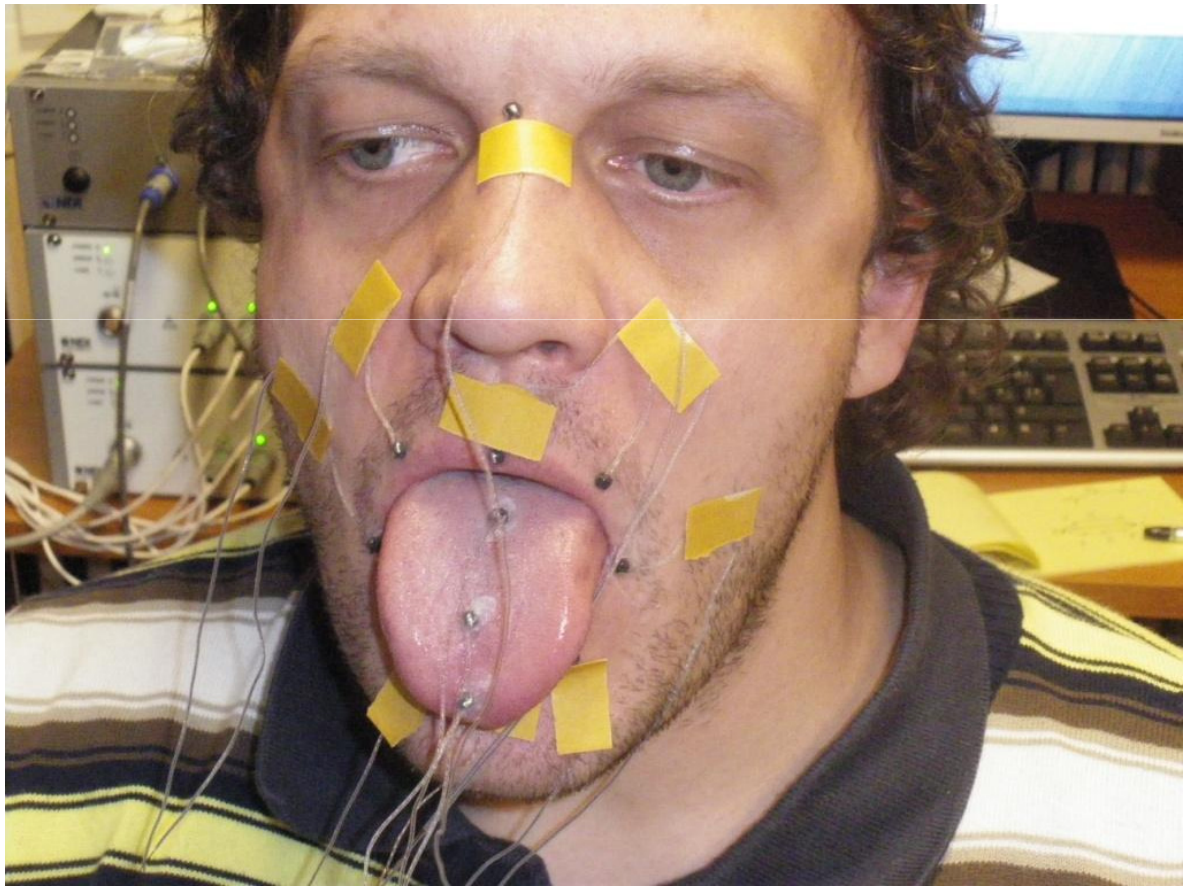
TTÜ KÜBERNEETIKA INSTITUUT
Institute of Cybernetics at TUT

Further steps

- EMA 3D recordings with a native speaker
- Annotation and segmentation
- Measurements of articulatory targets from 3D data
- Analysis of motion data of different sensors
- Improving the Estonian dominance

TTÜ KÜBERNEETIKA INSTITUUT
Institute of Cybernetics at TUT

EMA: Wave Speech Research System



Further steps

- **Adapting LUCIA model:**
  - MPEG4 animated talking face
  - compatible with FESTIVAL and MBROLA speech synthesis
  - enables expressive/emotive synthesis
  - open source code
  - limited documentation ☹
- **New faces:**
  - **Blender, Xface, etc**
  - FaceGen Modeller

TTÜ KÜBERNEETIKA INSTITUUT
Institute of Cybernetics at TUT

Thank you!