Emotional Speech Synthesis

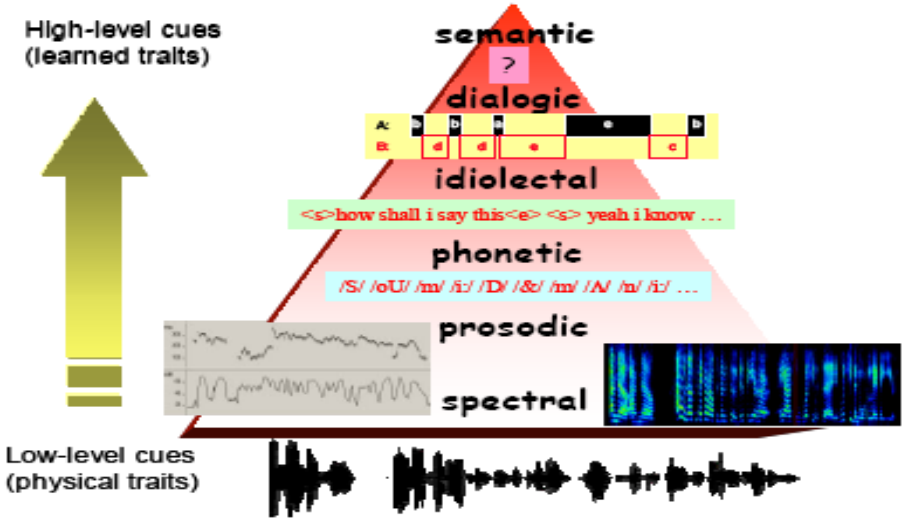
Felix Burkhardt

Contents of talk

- Emotional Speech Synthesis
  - Technology
  - Applications
  - Examples

Speech features

- Telephone Speech: digital signal, 8 kHz sample rate, 16 bit linear PCM, usually decoded from GSM or a-law
- Time domain vs frequency domain > Fast Fourier Transform (FFT)
- Mel Frequency Cepstral Coefficients (MFCC) used in Automatic Speech Recognition (ASR) to model short frame (10-20ms) frequency spectra speaker independent
- Prosody
- Microprosody, e.g.
- Jitter/Shimmer
- Pitch Durations Energy
- Functionals: Mean, max/min, deviation, regression, …,

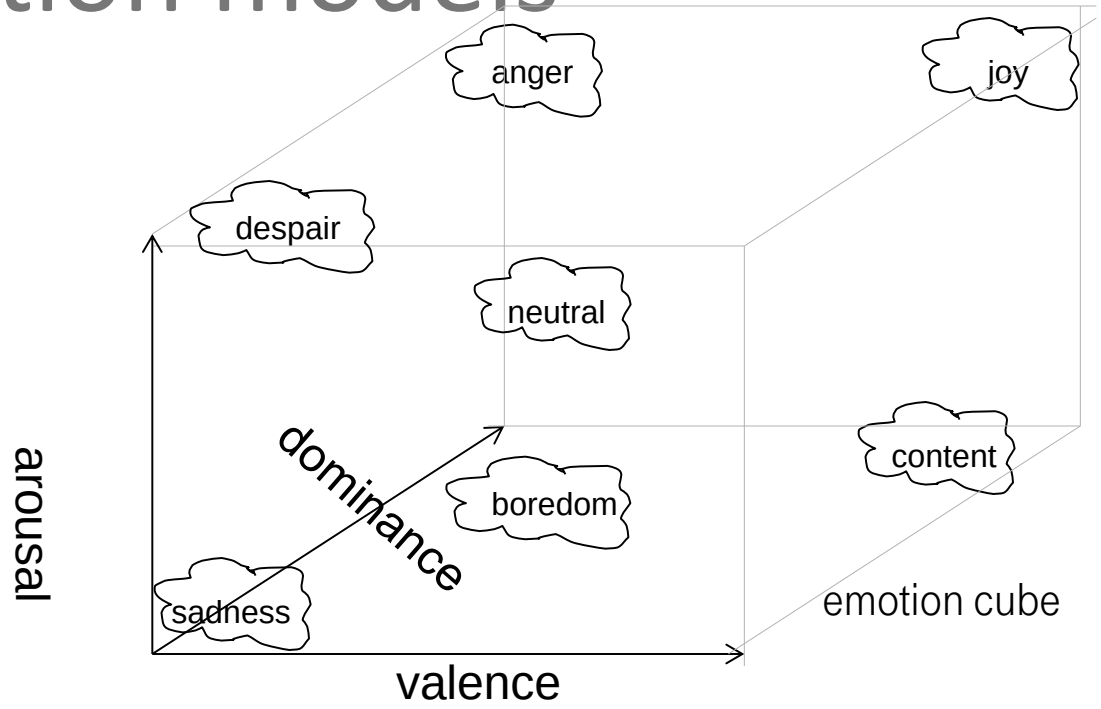


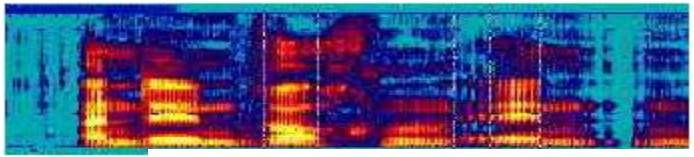
graphic: Reynolds et al 2003

Open source feature extraction: OpenEar, Praat

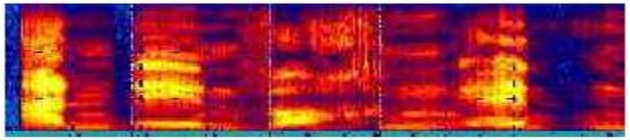
Emotion models

- categories, e.g. anger, joy, …
- dimensions, e.g. activation, dominance, valence
- appraisals, e.g. novelty, intrinsic pleasantness, relevance, coping potential,

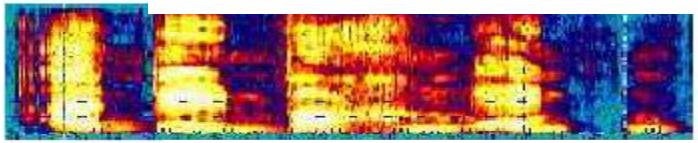




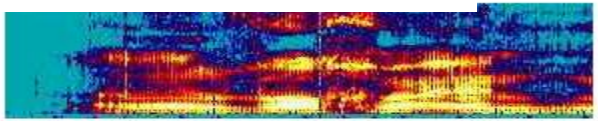
neutral



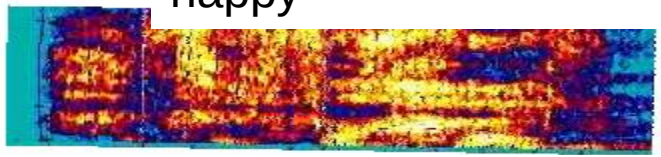
angry



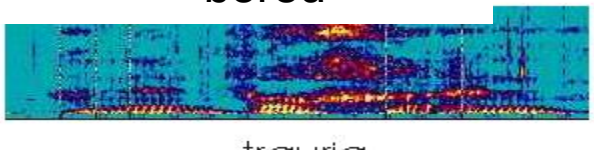
happy



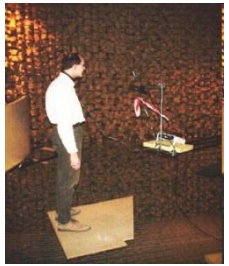
bored



frightened



sad



spectrograms from emotional acted speech

source: TUB emotional database

Applications

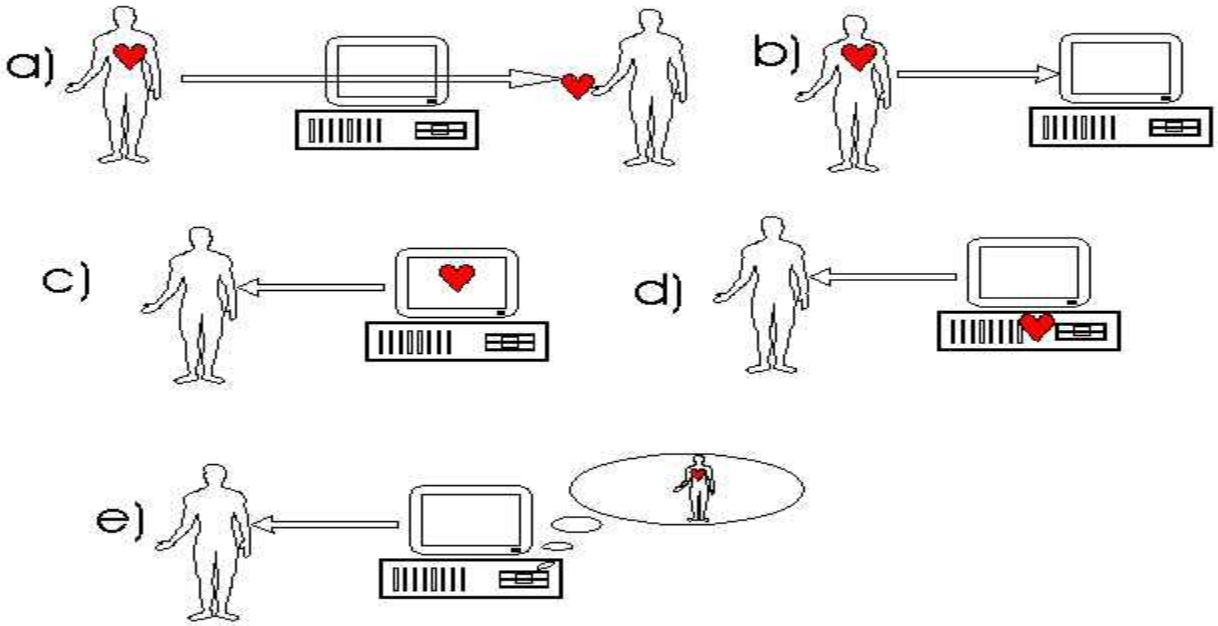
a) Mediated emotion

b) Affect recognition

c) Affect simulation

d) Modeling Emotional intelligence

e) Modeling human emotional behaviour

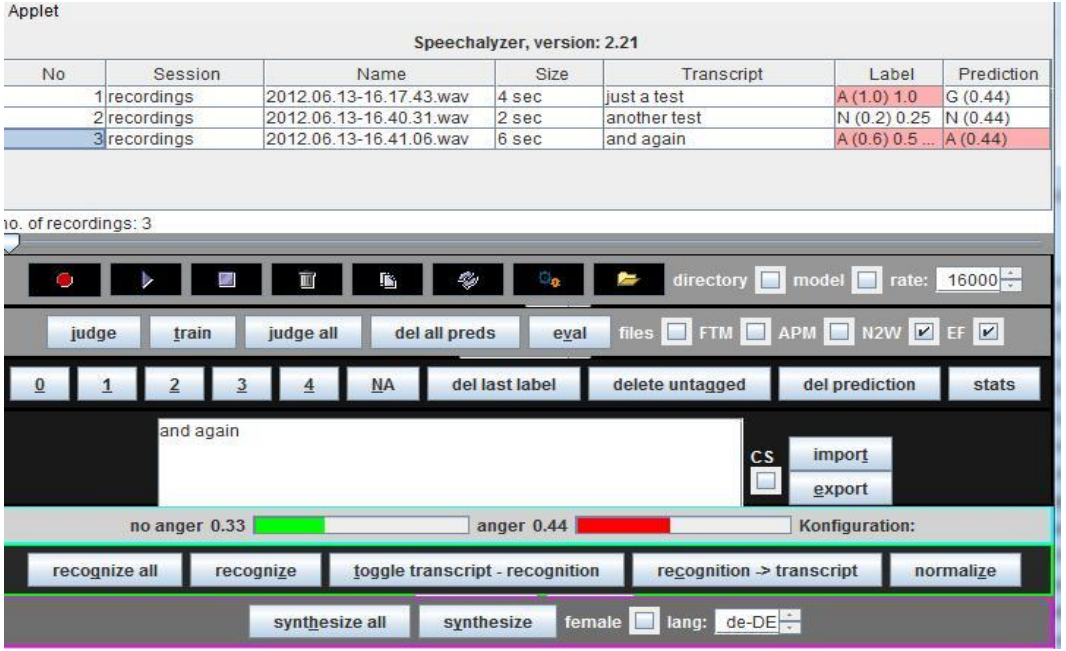


Applications speaker characteristics simulation

- Fun, e.g. emotional greetings
- Prosthesis
- Emotional chat avatars
- Gaming, believable characters
- Adapted dialog design
- Adapted persona design
- Target-group specific advertising
- …
- Believable agents, artificial humans

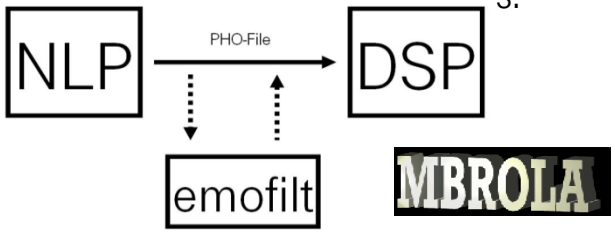
Tool: Speechalyzer

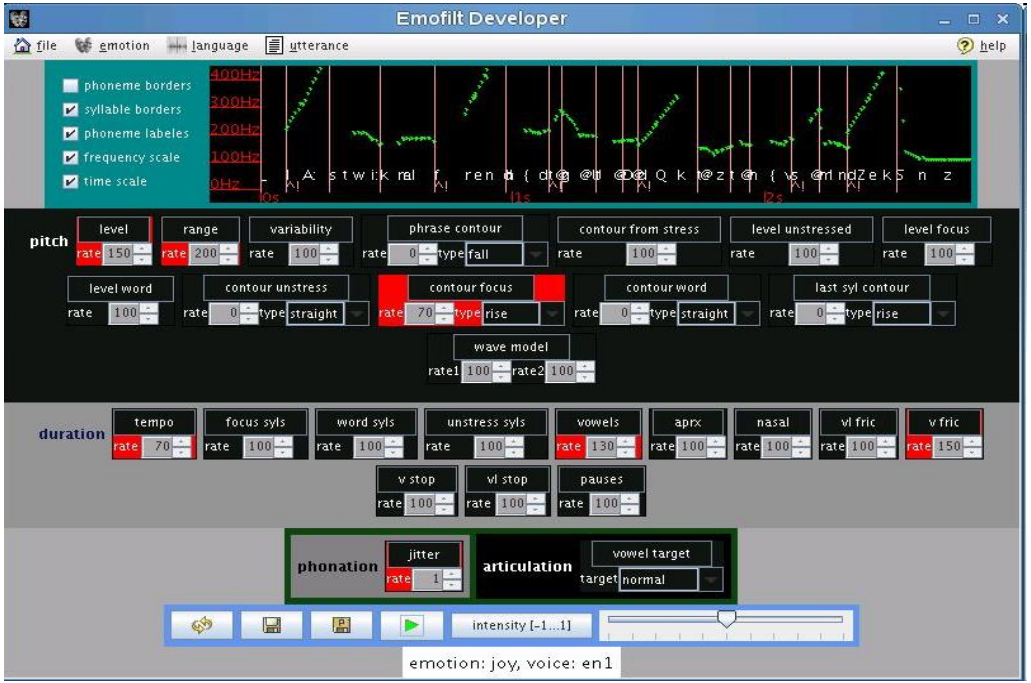
- Tool for Voice
  - transcription,
  - categorization,
  - synthesis
  - recognition
  - tuning
  - evaluation
- Open source since 7/2012
- Implements EmotionML Standard



Tool: Emofilt

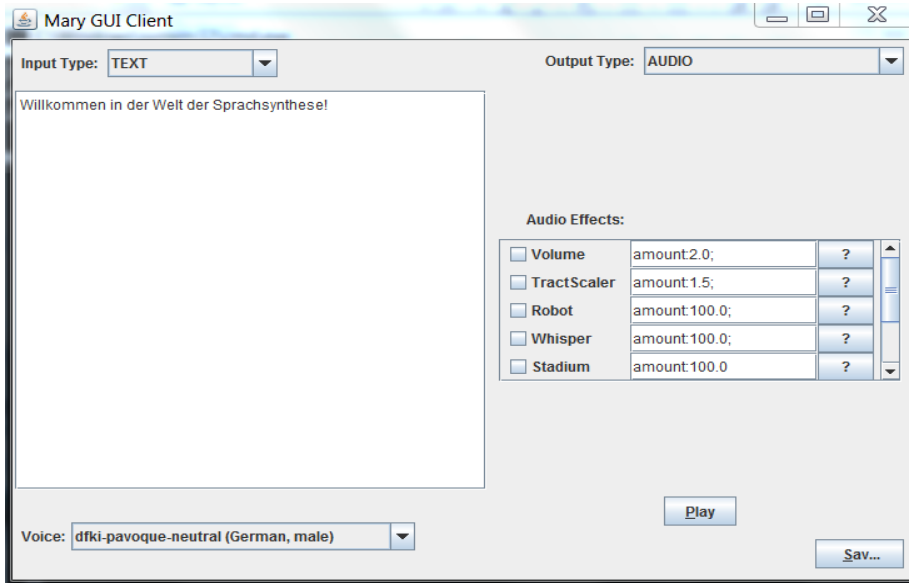
- Emofilt is an open source Java program to simulate emotional arousal with speech synthesis based on the free-for-non-commercial-use MBROLA synthesis engine.
- NOT a complete text-to-speech system
- Emofilt's language-dependent modules are controlled by external XML-files and it is as multilingual as MBROLA which currently supports 35 languages.



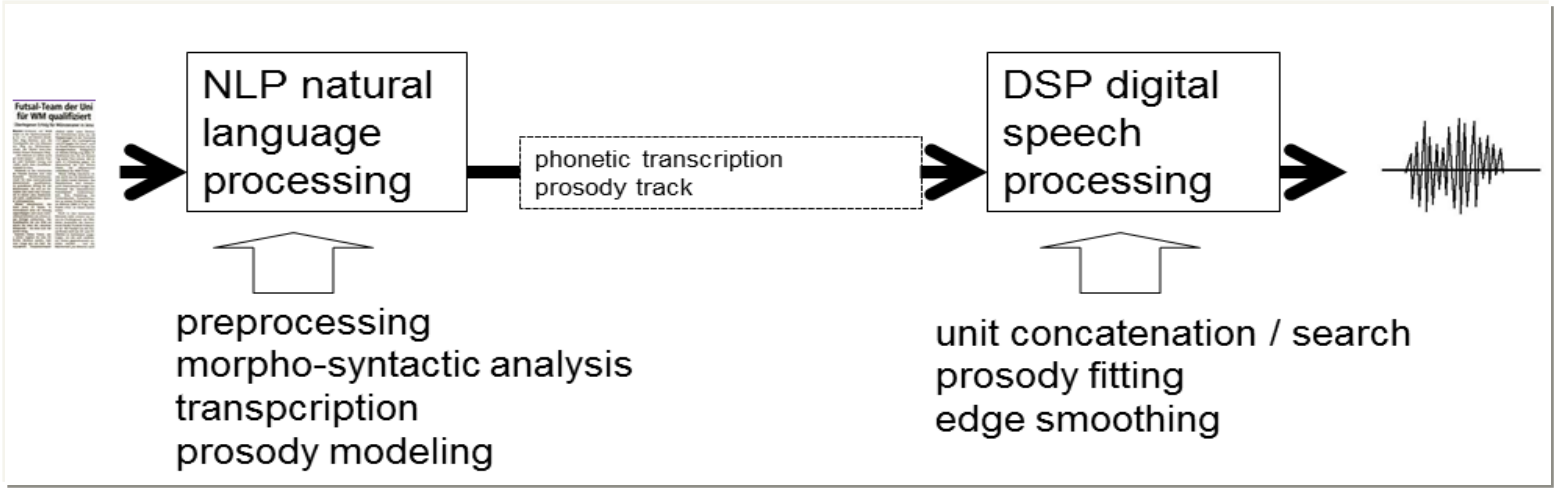


Tool: Mary Speech synthesizer

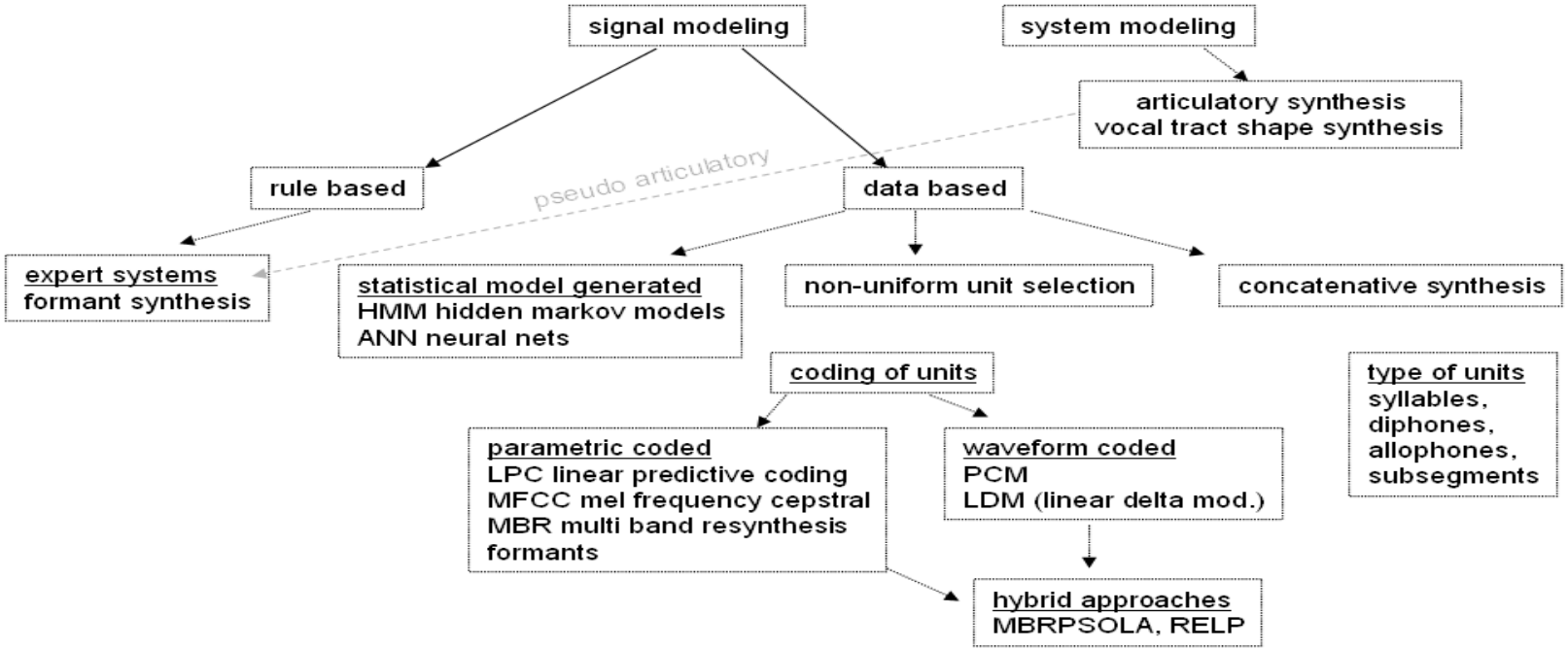
- Mary TTS is a text to speech system originally developed by Marc Schröder / Jürgen Trouvain at the Univ. Saarland / DFKI.
- It's open source, free and a great tool to learn about Speech synthesis
- It's a modular system written in Java for several processing steps, each process result can be viewed in XML format.
- We will try it out
- Bring a laptop with Java latest version on it and an internet connection



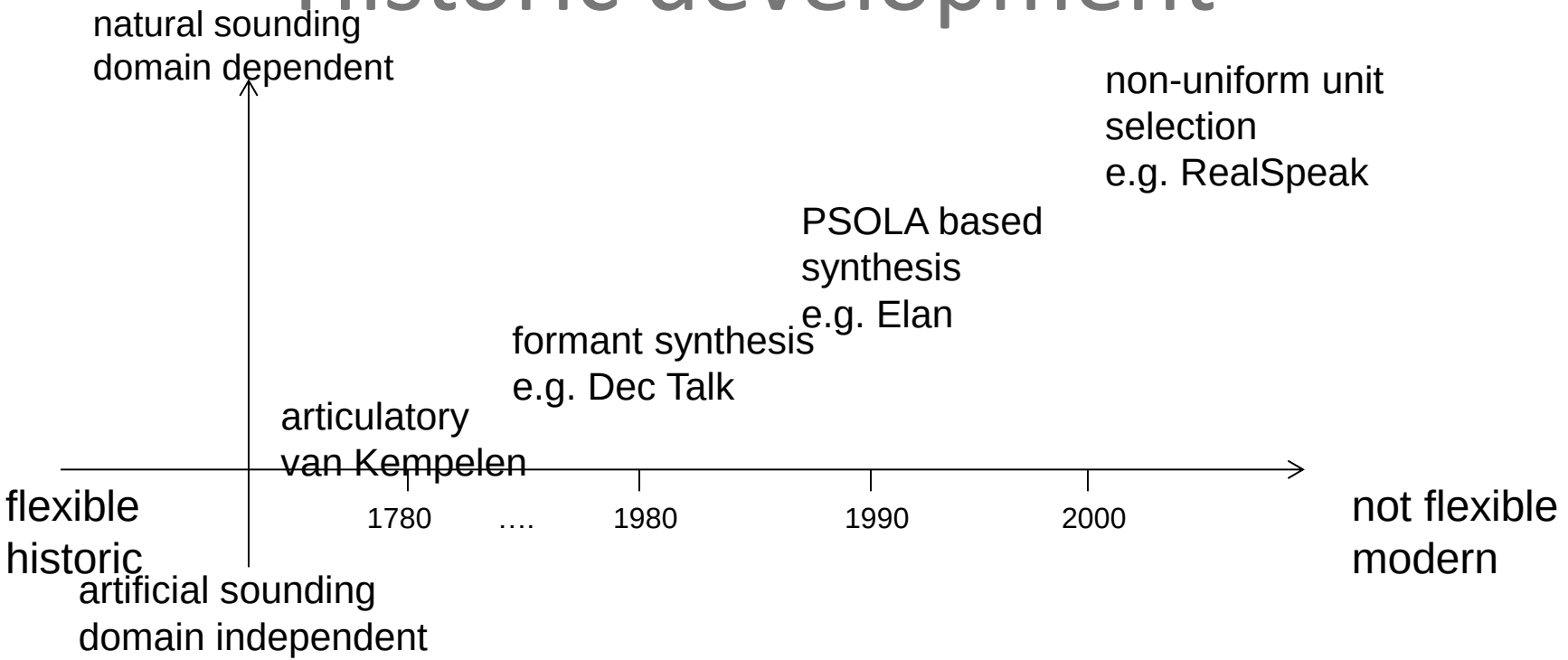
Speech synthesis approach



Speech synthesis: technology



Historic development



Simulation: technology

	natural sounding	flexible modification	prosody manipulation	voice quality manipulation	
formant synthesis	inherent buzziness due to inadequate source-filter modeling	very flexible due to full parametric control	rule based flexibility	rule based flexibility	a)
articulatory synthesis	only copy synthesis possible	theoretically very flexible due to direct physical modeling of muscle tension	without limitation	yes	b)
diphone-based synthesis	limited naturalness due to many concatenation points	only with respect to prosodic variation	yes, but high jumps introduce audible distortions	only by using several databases	c)
non-uniform unit selection	quite natural within the domain of the database	no flexibility outside the scope of the database	only if emotional models are part of the concatenation function	only if emotional models are part of the concatenation function	d)
HMM-based synthesis	introduces slight buzziness	yes, even linear interpolation between styles possible	yes, if emotional data is part of the database	yes, due to underlying source-filter model	e)

a) Emofilt, b) VoiceTract c) EmoFilt, d) E. Eide (IBM), e) Kobayashi Lab

)

Simulation: further examples

- Laughter synthesis (mass spring model)
- Kismet robot (formant synthesis)
- Voice transformation (LF and harmonic model)

a)

b)

c)

a) S. Sundaram, SAIL b) C. Breazeal, MIT c) Y. Agiomyrgiannakis and O. Rosec, Orange Labs

Simulation: evaluation

a) emotion recognition

	Neut.	Zorn	Aerg.	Zufr.	Freu.	wein.Tr.	st.Tr.	Angst	Lang.
Neut.	55.4	0.6	0	21.4	0	1.2	7.1	1.6	12.5
Zorn	4.8	28.6	38.1	4.8	7.1	7.1	0	9.5	0
Aerg.	19	11.9	59.5	7.1	0	2.4	0	0	0
Zufr.	11.9	0	0	61.9	14.3	0	0	0	11.9
Freu.	2.4	4.8	4.8	4.8	81	0	0	2.4	0
w.Tr.	2.4	0	0	7.1	0	69	9.5	11.9	0
s.Tr.	2.4	0	4.8	0	0	26.2	38.1	2.4	26.2
Ang.	2.4	4.8	7.1	4.8	11.9	14.3	2.4	52.4	0

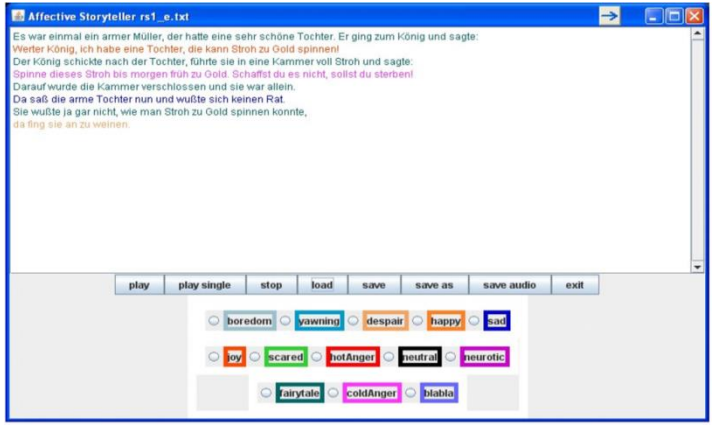
a) F. Burkhardt 2000

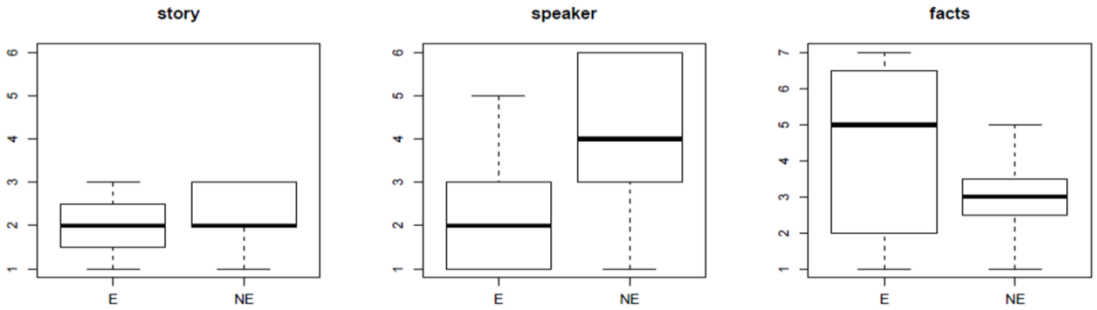
b) user studies

iCat role	Measure	neutral	face only	speech & face
companion	% correct recognition	78.1	88.4	84.4
	reaction time (in sec)	5.7	5.9	5.6
educator	% correct answers	75.0	66.7	66.7
	reaction time (in sec)*	5.9	5.9	4.7
motivator	z-score	-0.28	0.03	0.25

b) J. Kessens et al ACII 2009

Evaluation: storyteller





F. Burkhardt: An Affective Spoken Story Teller Interspeech, 2011

Outlook Simulation

- Unified models from non-uniform unit selection and HMM based synthesis
- Models for extralinguistic speech sounds
- Enhanced voice source – vocal tract (source filter) models for parametric synthesis
- Speaking style variations (informal, news-style, flirtatious)
- Fast building of new speech databases
- New voice transformation techniques

Thanks

Felix.Burkhardt@telekom.de