

Emotional speech synthesis

Technologies and research approaches

Marc Schröder, DFKI

with contributions from

Olivier Rosec, France Télécom R&D

Felix Burkhardt, T-Systems

HUMAINE WP6 workshop, Paris, 10 March 2005

Overview

- ◆ Speech synthesis technologies
 - ➔ formant synthesis
 - ➔ HMM synthesis
 - ➔ diphone synthesis
 - ➔ unit selection synthesis
 - ➔ voice conversion
- ◆ Research on emotional speech synthesis
 - ➔ straightforward approach (and why not to do it)
 - ➔ systematic parameter variation: Burkhardt (2001)
 - ➔ non-extreme emotions: Schröder (2004)

Overview

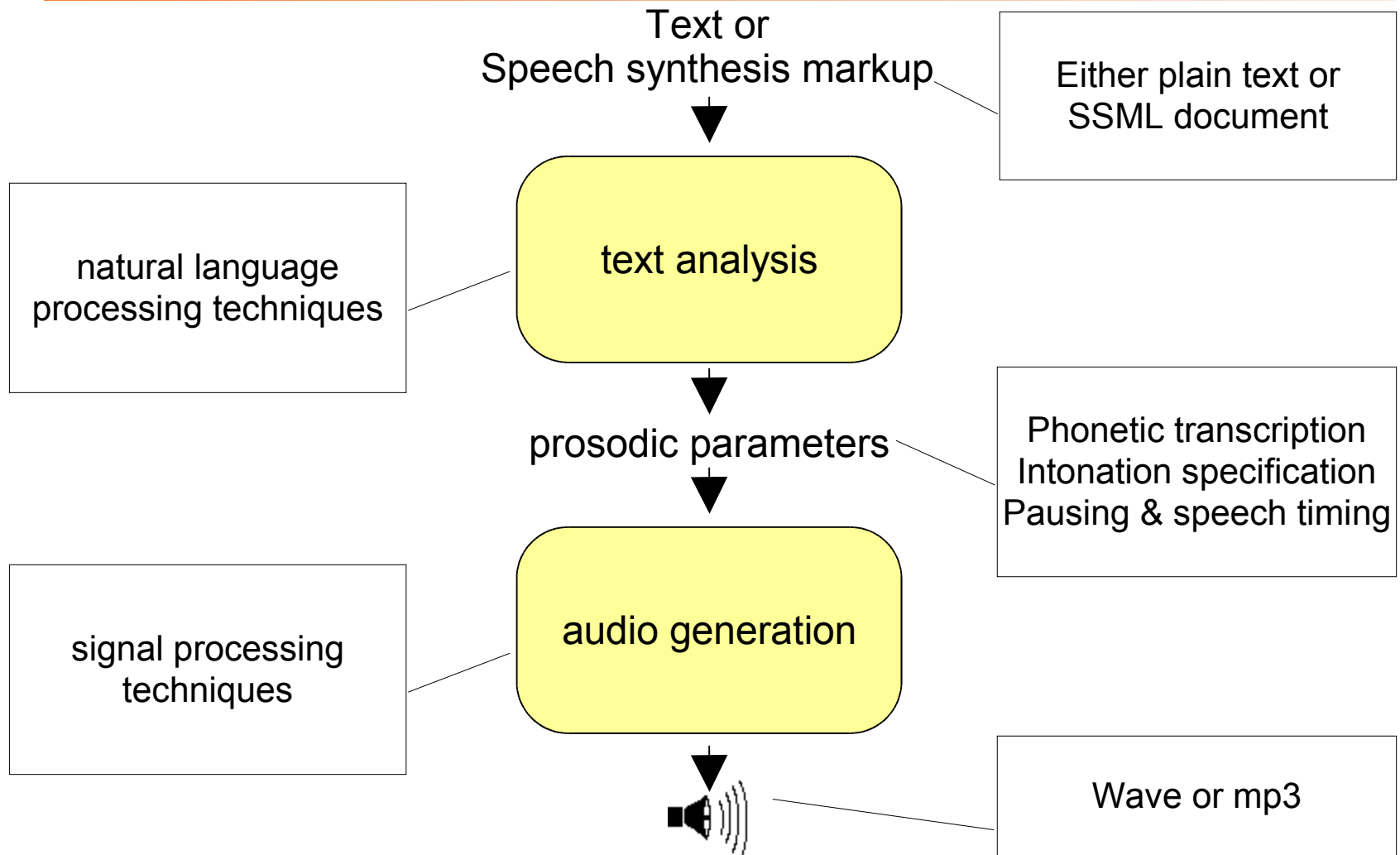
◆ Speech synthesis technologies

- formant synthesis
- HMM synthesis
- diphone synthesis
- unit selection synthesis
- voice conversion

◆ Research on emotional speech synthesis

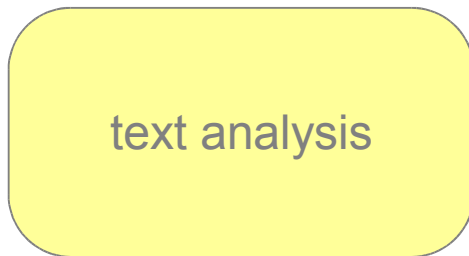
- straightforward approach (and why not to do it)
- systematic parameter variation: Burkhardt (2001)
- non-extreme emotions: Schröder (2004)

Speech synthesis



Speech synthesis technologies

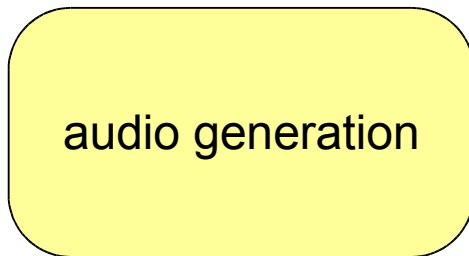
Text or
Speech synthesis markup



text analysis



prosodic parameters



audio generation



- ◆ Formant synthesis
- ◆ HMM-based synthesis
- ◆ Diphone synthesis
- ◆ Unit selection synthesis
- ◆ Voice conversion

Speech synthesis technologies

Formant synthesis

- ◆ Acoustic modelling of speech
- ◆ Many degrees of freedom, can potentially reproduce speech perfectly
- ◆ Rule-based formant synthesis: Imperfect rules for acoustic realisation of articulation
=> robot-like sound

Examples:

Janet Cahn (1990):

- [angry](#)
- [happy](#)
- [sad](#)
- [fearful](#)

Felix Burkhardt (2001):

- [neutral](#)
- [angry](#)
- [happy](#)
- [sad](#)
- [fearful](#)

Speech synthesis technologies

HMM synthesis

- ◆ Hidden Markov Models trained from speech database(s)
- ◆ synthesis using acoustic model (MLSA)
=> robot-like sound

Examples:

Miyanaga et al. (2004):	trained from corpus:		trained from corpus:	
Parametrise HMM output parameters using a “style control” vector	<u>neutral</u>	<u>0.5 joyful</u>	<u>joyful</u>	<u>1.5 joyful</u>
		interpolated		interpolated

Speech synthesis technologies

Diphone synthesis

- ◆ Diphones = small units of recorded speech
 - from middle of one sound to middle of next sound
 - e.g. [grEIt] = _-g g-r r-EI EI-t t-__
- ◆ Signal manipulation to force pitch (F0) and duration into a target contour
 - Can control prosody, but not voice quality

Examples:

[neutral](#)

[angry](#)

[angry](#)

Marc Schröder (1999):

[happy](#)

Ignasi Iriondo (2004):

[happy](#)

[sad](#)

[sad](#)

[fearful](#)

[fearful](#)

Speech synthesis technologies

Diphone synthesis

◆ Is voice quality indispensable?

→ Interesting diversity of opinions in the literature

→ Tentative conclusion: “It depends!”

- ...on the emotion (Montero et al., 1999)

- prosody conveys surprise, sadness

- voice quality conveys anger, joy

- ...on speaker strategies (Schröder, 1999)

[angry1](#) [orig_angry1](#)

[angry2](#) [orig_angry2](#)

Speech synthesis technologies

Diphone synthesis

- ◆ Partial remedy: Record voice qualities
- ◆ Schröder & Grice (2003): Diphone databases with three levels of vocal effort
 - male: [loud](#) [modal](#) [soft](#)
 - female: [loud](#) [modal](#) [soft](#)
- ◆ Voice quality interpolation: Turk et al. (in prep.)
 - female: [loud](#) [1](#) [2](#) [modal](#) [3](#) [4](#) [soft](#)
- ◆ Not yet successful: smiling voice
 - [modal1](#) [smile1](#)
 - [modal2](#) [smile2](#)

Speech synthesis technologies

Unit selection synthesis

- ◆ Select small speech units out of very large speech corpus (e.g., 5 hours of speech)
- ◆ Avoid signal manipulation to maintain natural prosody from the units
 - ➔ **Cannot** control prosody or voice quality
 - ➔ Very good “playback” quality with emotional recordings

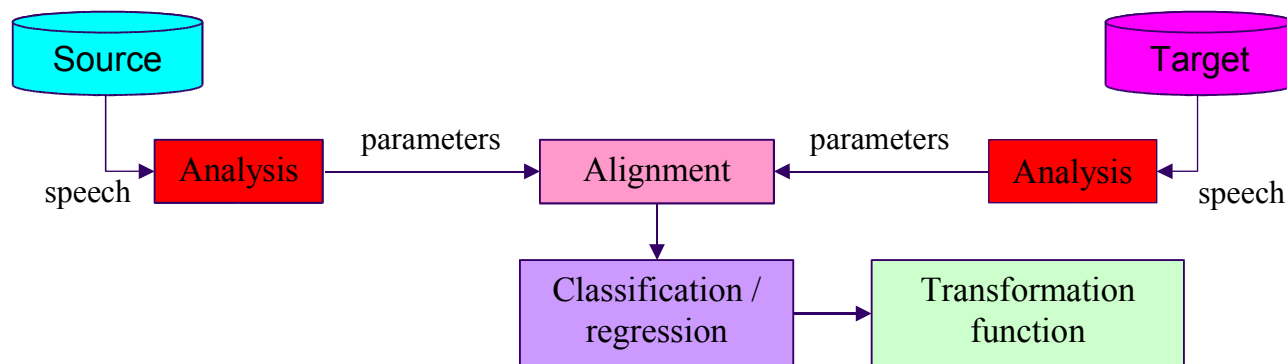
Examples:

Akemi Iida (2000): [angry](#)
[happy](#)
[sad](#)

Ellen Eide (IBM, 2004): [good news](#)
[bad news](#)

Speech synthesis technologies

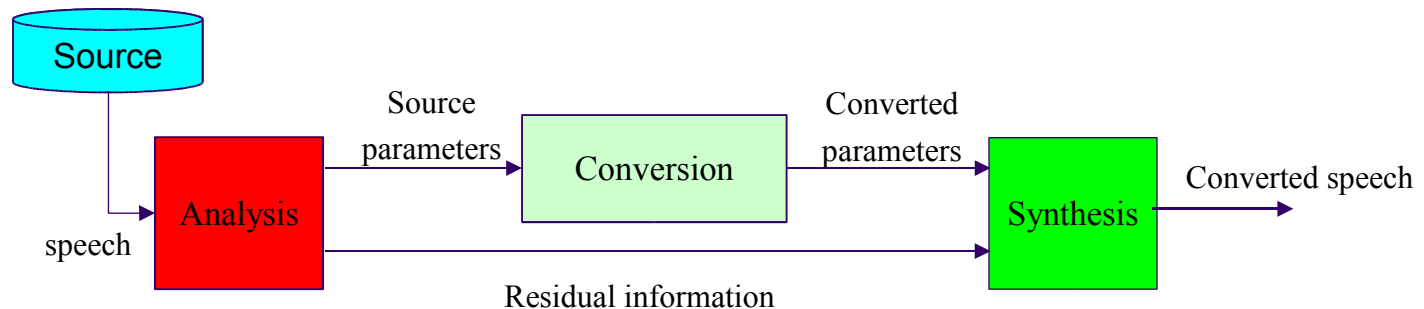
Voice conversion – How to learn a new voice ?



- ❖ Learning data needed: about 5 minutes
- ❖ Transformed parameters: timbre and F_0
- ❖ Conversion techniques: VQ, GMM, ...
- ❖ Potential application to emotion
 - source = neutral speech
 - target = emotional speech

Speech synthesis technologies

Voice conversion – Transformation step



❖ Analysis / synthesis: LPC, formant or HNM

❖ Output quality of the converted speech

- ➔ Can be fairly good in terms of speaker (/emotion?) identification
- ➔ Degradation of naturalness

Example for speaker transformation:

France Télécom
speech synthesis team:

[source](#)

[target](#)

[conversion](#)

Speech synthesis technologies: Summary

◆ Current choice:

➔ “Explicit modelling” approaches

- low naturalness
- high flexibility, high control over acoustic parameters
- explicit models of emotional prosody

➔ “Playback” approaches

- high naturalness
- no flexibility, no control over acoustic parameters
- emotional prosody implicit in recordings

◆ Technical challenge over next years: combine the best of both worlds!

Overview

◆ Speech synthesis technologies

- formant synthesis
- HMM synthesis
- diphone synthesis
- unit selection synthesis
- voice conversion

◆ Research on emotional speech synthesis

- straightforward approach (and why not to do it)
- systematic parameter variation: Burkhardt (2001)
- non-extreme emotions: Schröder (2004)

Research on emotional speech synthesis

The “straightforward” approach

(and why not to do it)

◆ The “straightforward” approach

- ➔ record one actor with four emotions
 - anger, fear, sadness, joy (+neutral)
- ➔ measure acoustic correlates
 - overall pitch level + range, tempo, intensity
 - copy synthesis or prosody rules, synthesise
- ➔ forced-choice perception test with “neutral” text
 - overall recognition rates
- ➔ ...and then?

“there has been neither continuity nor cumulativeness in the area of the vocal communication of emotion”

(Scherer, 1986, p. 143)

Research on emotional speech synthesis

The “straightforward” approach

(and why not to do it)

May not be representative

“straightforward” approach

Why these four?
Applications don't need “basic” emotions

record one actor with four emotions

Needed: quality control (e.g., expert rating)

Emotion words too ambiguous – use frame stories when recording

lose local effects

measure acoustic correlates

lose interaction with linguistic structure

overall pitch level + range, tempo, intensity
more and different parameters needed: voice quality!

forced-choice perception test with “neutral” text

unexpected percepts?

overall recognition rates
...and then?

Untypical for applications

Applications need better continuity, not recognition of emotion

How bad are errors?
Need measure of semantic similarity of states

(Scherer, 1986, p. 143)

Research on emotional speech synthesis

The “straightforward” approach

(and why not to do it)

May not be representative

“straightforward” approach

Why these four?
Applications don't need “basic” emotions

→ record one actor with four emotions

Needed: quality control (e.g., expert rating)

Emotion words too ambiguous – use frame stories when recording

lose local effects

measure acoustic correlates

overall pitch level + range, tempo, intensity

lose interaction with linguistic structure

more and different parameters needed: voice quality!

→ forced-choice perception test with “neutral” text

unexpected percepts?

overall recognition rates

Untypical for applications

...and then?

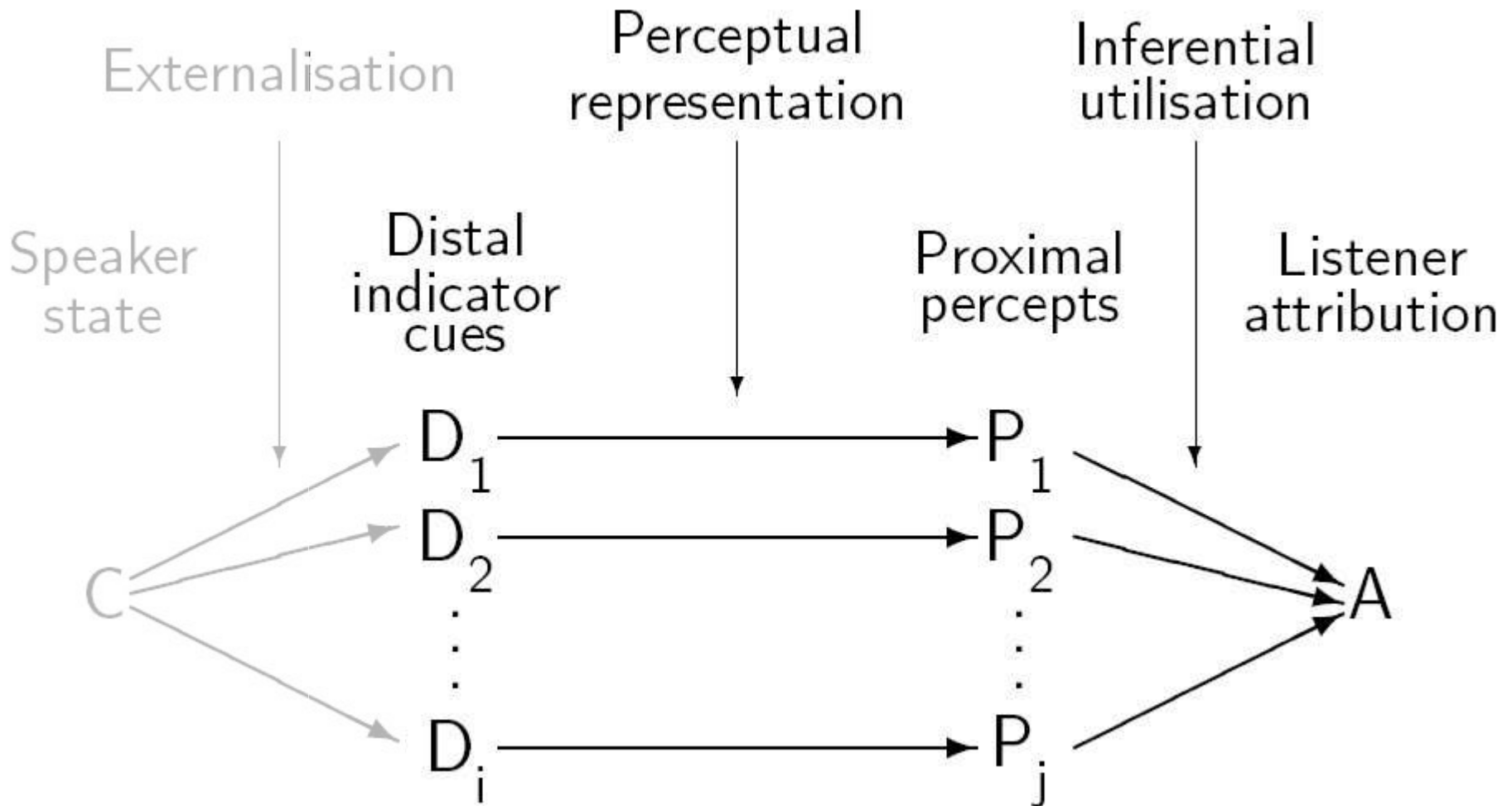
Applications need better continuity, not recognition of emotion

How bad are errors?
Need measure of semantic similarity of states

(Scherer, 1986, p. 143)

Emotional speech synthesis research

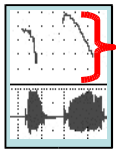
Listener-centred orientation



Emotional speech synthesis research

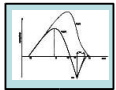
Listener-centred approach: Burkhardt (2001)

- ◆ Stimuli: systematically varied selected acoustic features using formant synthesis

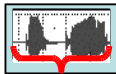


→ pitch height (3 variants)

→ pitch range (3 variants)



→ phonation (5 variants)



→ segment durations (4 variants)



→ vowel quality (3 variants)

- ◆ one semantically neutral sentence

→ Complete factorial design would be >2000 stimuli

→ tested three groups of parameters combinations

- Pitch/Phonation: 45 stimuli, Pitch/Segmental: 108 stimuli,

- Phonation/Segmental: 60 stimuli

Listener-centred approach: Burkhardt (2001)

◆ Forced choice perception test

- neutral, fear, anger, joy, sadness, boredom

=> Perceptually optimal values for each category

◆ Second step:

- varied additional acoustic parameters
- further differentiation into subcategories:
 - hot/cold anger, joy/happiness, despair/sorrow

◆ Goals

- ➔ Model many, gradual states on a continuum
- ➔ Allow for gradual changes over time
- ➔ Model many acoustic parameters, including voice quality

◆ Success criterion

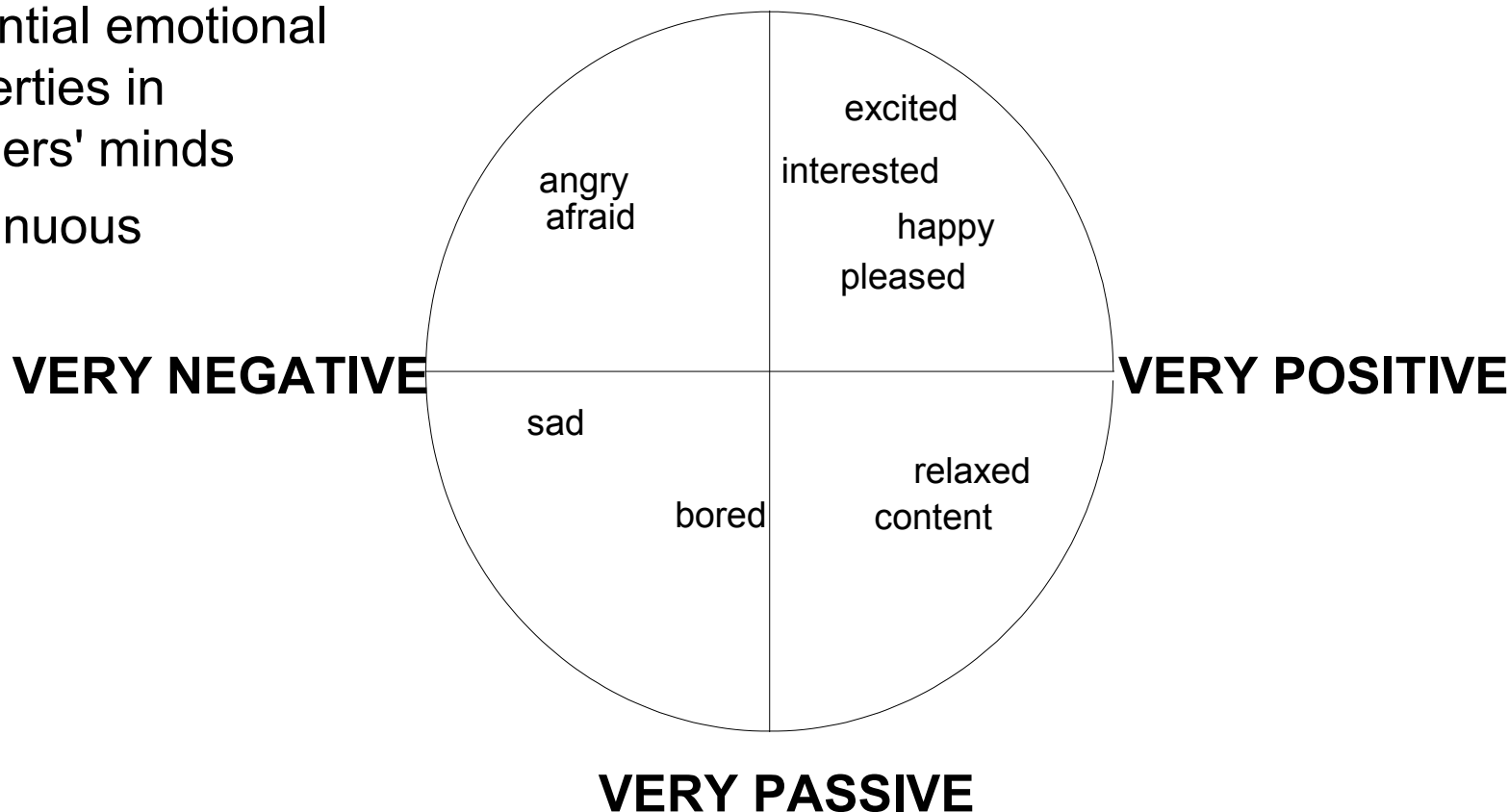
- ➔ Voice “fits with” the text

Emotional speech synthesis research

Dimensional approach: Description system

- Representation of emotional states in a 2-dim. space, activation-evaluation space: **VERY ACTIVE**

- Essential emotional properties in listeners' minds
- Continuous



Emotional speech synthesis research

Dimensional approach: Emotional prosody rules

Database analysis

- ➔ Belfast Naturalistic Emotion Database: 124 speakers, spontaneous emotions
- ➔ Search for correlations between emotion dimensions and acoustic parameters

Activation

- ➔ numerous, robust correlations

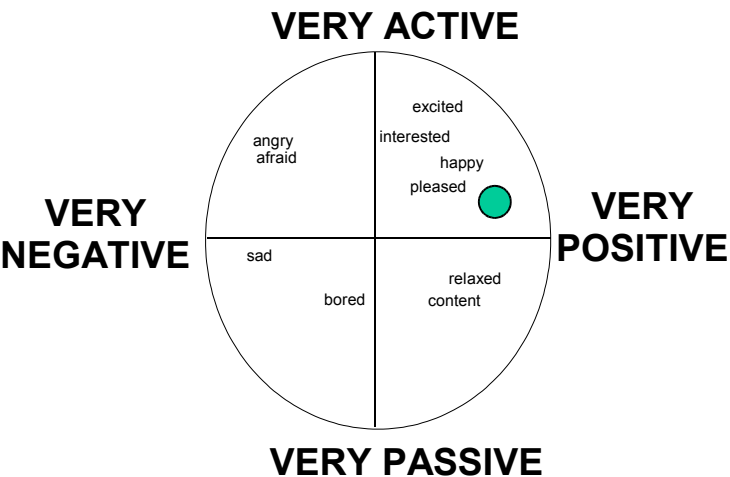
Evaluation and Power

- ➔ fewer, weaker correlations

Acoustic variable	Correlations						
	Activation		Evaluation		Power		
	female	male	female	male	female	male	
fundamental frequency	F0 median	↑↑	↑↑	↓	↑	↓	↓
	F0 range	↑↑	↑↑	↓			
	med. magn. F0 rises	↑↑	↑↑				
	range magn. F0 rises	↑↑	↑↑	↓			
	med. magn. F0 falls	↑↑	↑↑	↓	↑		
	range magn. F0 falls	↑↑	↑↑	↓			↓
	med. dur. F0 rises	↑		↑	↑		
	rng. dur. F0 rises	↑	↑		↑		
	med. dur. F0 falls	↑			↑		
	rng. dur. F0 falls	↑	↑	↓	↑		↑
	med. slope F0 rises	↑↑	↑↑	↓	↓		↓
	med. slope F0 falls	↑↑	↑↑	↓	↓		↓
	F0 rises p. sec.	↓	↓		↓		
	F0 falls p. sec.	↓	↓	↓	↓		
tempo	duration pauses	↓	↓			↓	
	'tune' duration	↑	↑		↑		↓
	intensity peaks p. sec.						↓
	fricat. bursts p. sec.	↓			↑		
intens.	intensity median			↓			
	intensity range						
	dynamics at peaks	↑	↑	↓	↓		↑
voice quality	spectral slope non-fric.	↑↑	↑↑	↓	↑		↓
	Hamm. 'effort'		↑↑		↓		↑
	Hamm. 'breathy'	↓			↓		↑
	Hamm. 'head'	↓	↓			↓	↓
	Hamm. 'coarse'	↓	↓		↓		↓
	Hamm. 'unstable'	↓	↓		↑		↓

Emotional speech synthesis research

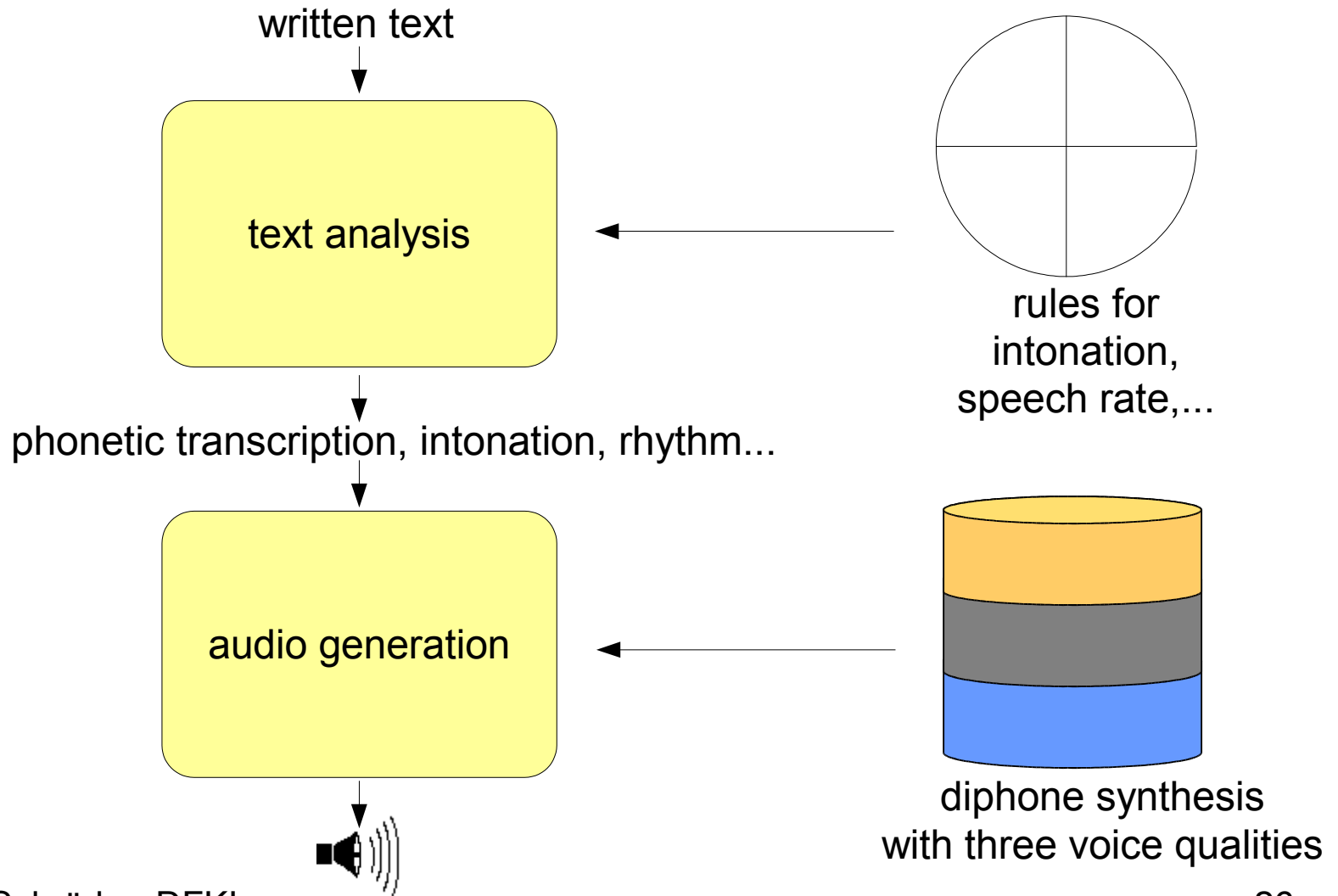
Dimensional approach: Synthesis method



- ❖ Rules map each point in emotion space onto its acoustic correlates
- ❖ Flexibility: gradual build-up of emotions, non-extreme emotional states
- ❖ Emotions are not fully specified through the voice
 - ➔ complementary information required: verbal content, visual channel, situational context

Emotional speech synthesis research

Dimensional approach: Realisation in the system

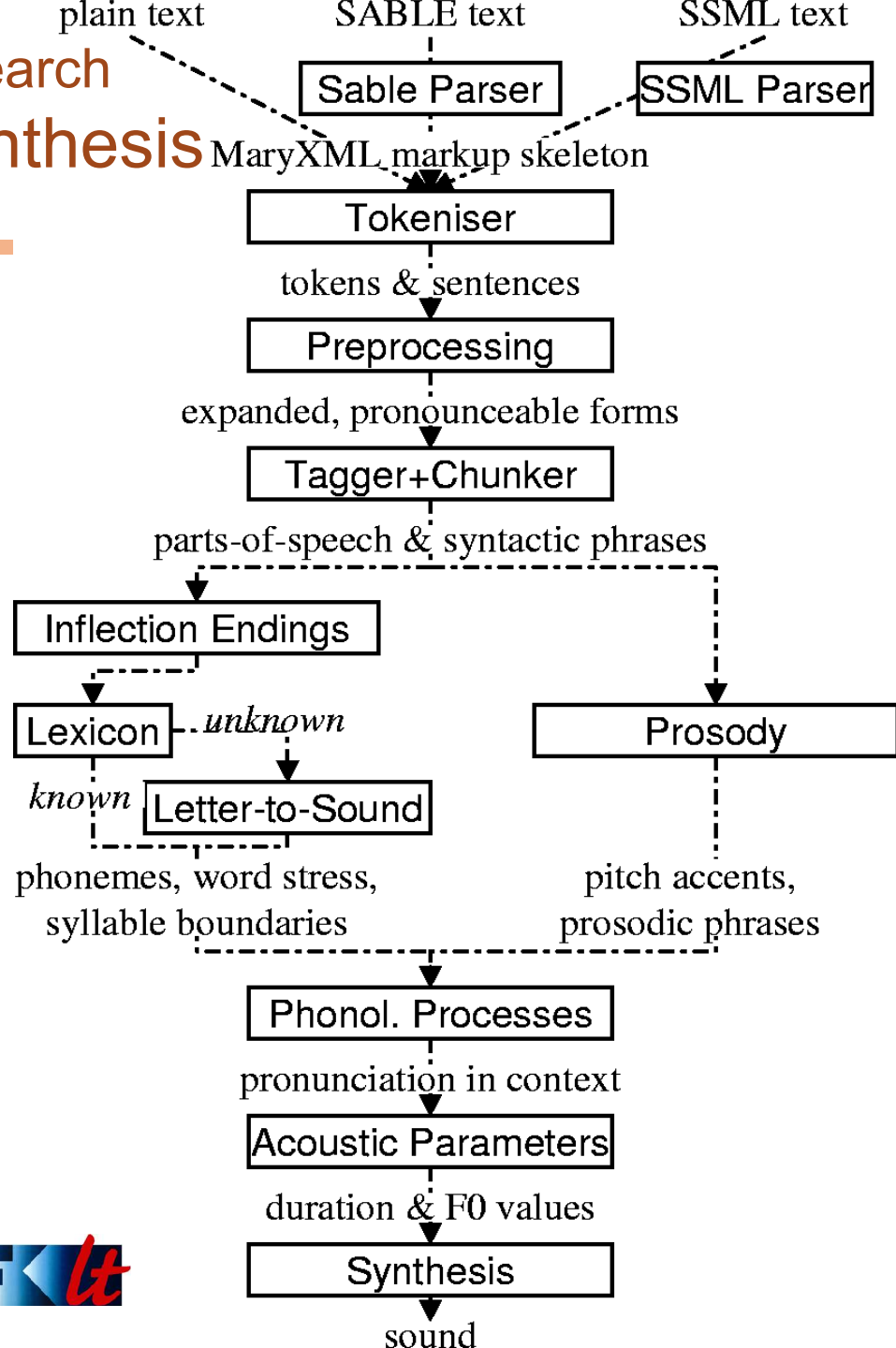


Emotional speech synthesis research

MARY: DFKI's speech synthesis

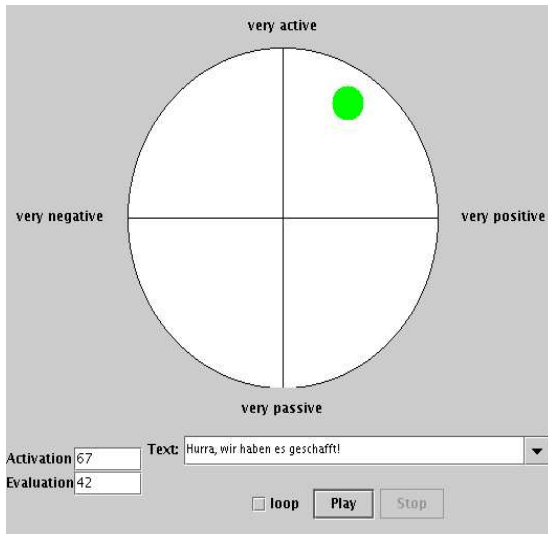
<http://mary.dfki.de>

- ◆ Developed in cooperation with Institute of Phonetics, Saarland Univ.
- ◆ Languages: German, English
- ◆ Transparent and flexible
 - ➔ Modular
 - ➔ Internal MaryXML format
 - ➔ Input/output possible at all intermediate processing steps ⇒ allows for fine-grained control



Emotional speech synthesis research

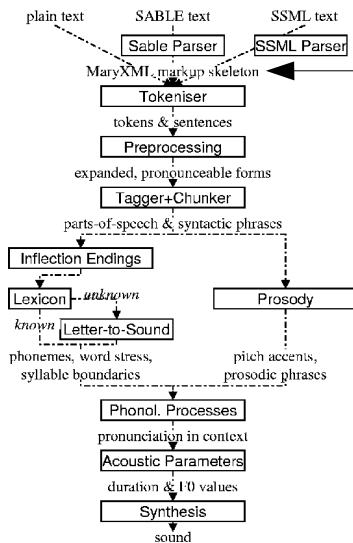
Dimensional approach: Technical realisation



```
<emotion activation="67" evaluation="42">
Hurra, wir haben es geschafft!
</emotion>
```

Emotional prosody rules
(XSLT stylesheet)

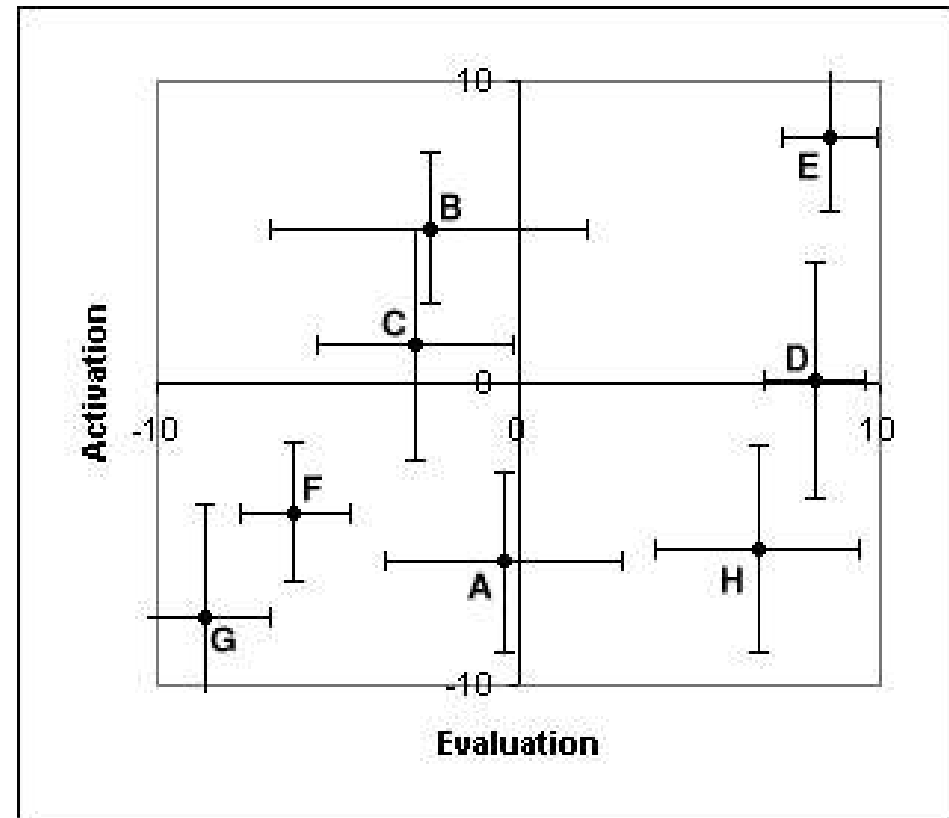
```
<maryxml>
<prosody accent-prominence="+13%"
accent-slope="+46%" fricative-duration="+21%"
liquid-duration="+13%" nasal-duration="+13%"
number-of-pauses="+47%" pause-duration="-13%"
pitch="134" pitch-dynamics="+5%"
plosive-duration="+21%"
preferred-accent-shape="alternating"
preferred-boundary-type="high" range="52"
range-dynamics="+40%" rate="+42%" volume="72"
vowel-duration="+13%">
Hurra, wir haben es geschafft!
</prosody>
</maryxml>
```



Emotional speech synthesis research

Dimensional approach: Listening test

- ◆ Eight emotion-specific texts
- ◆ Prosodic parameters predicted for each of the eight emotional states
- ◆ Factorise text x prosody => 64 stimuli
- ◆ Listeners evaluate stimuli on a scale
“How well does the sound of the voice fit to the text spoken?”



Situationsbeschreibung

Er erzählt rückblickend, wie er seine Frau kennengelernt hat. „Kennengelernt haben wir uns, als wir beide zur Uni gegangen sind. Ich weiß noch, wie ich sie das erste Mal gesehen habe: **sie war am anderen Ende eines total überfüllten Raums, und sie hatte diese wundervollen großen braunen Augen.**“

Wie gut passt der Klang der Stimme zum Inhalt des Textes?
passt optimal

1

3

4

5

6

7

8

2

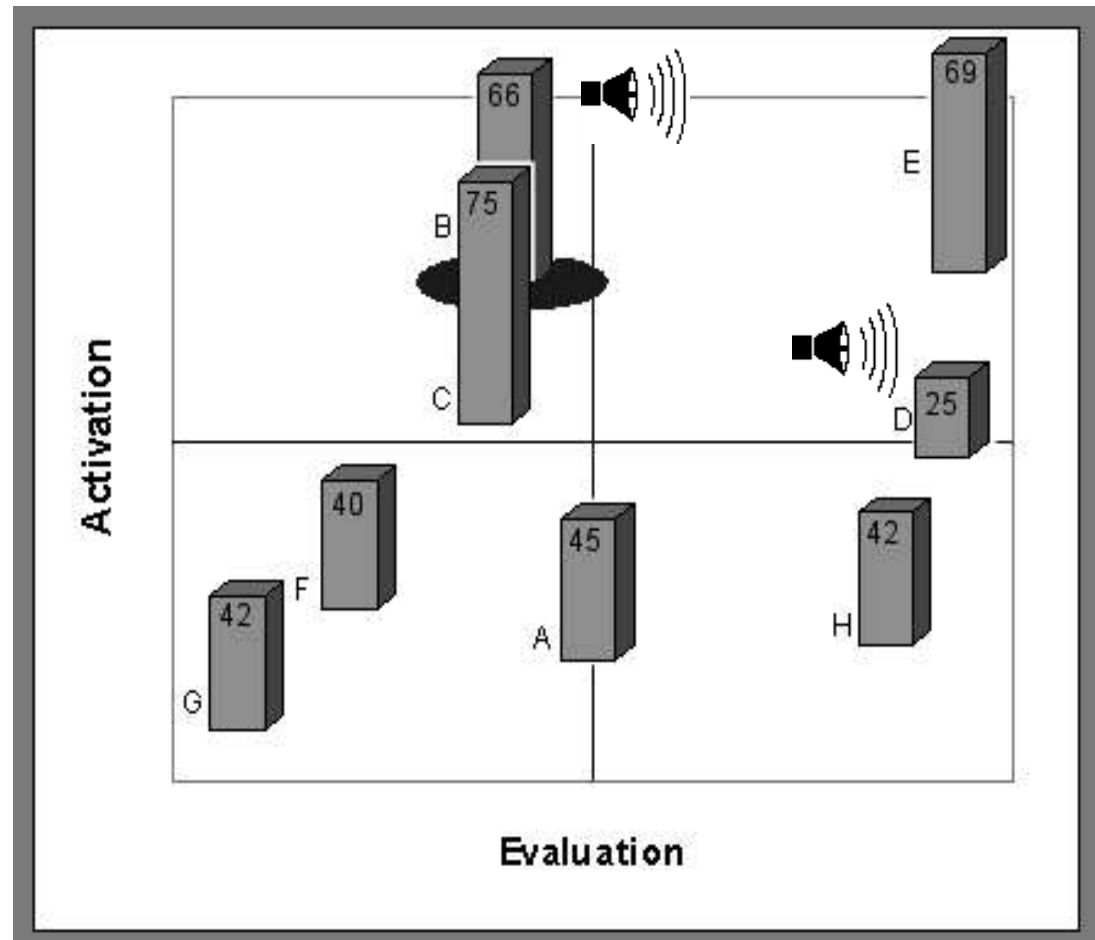
passt überhaupt nicht

Fertig

Emotional speech synthesis research

Dimensional approach: Listening test results

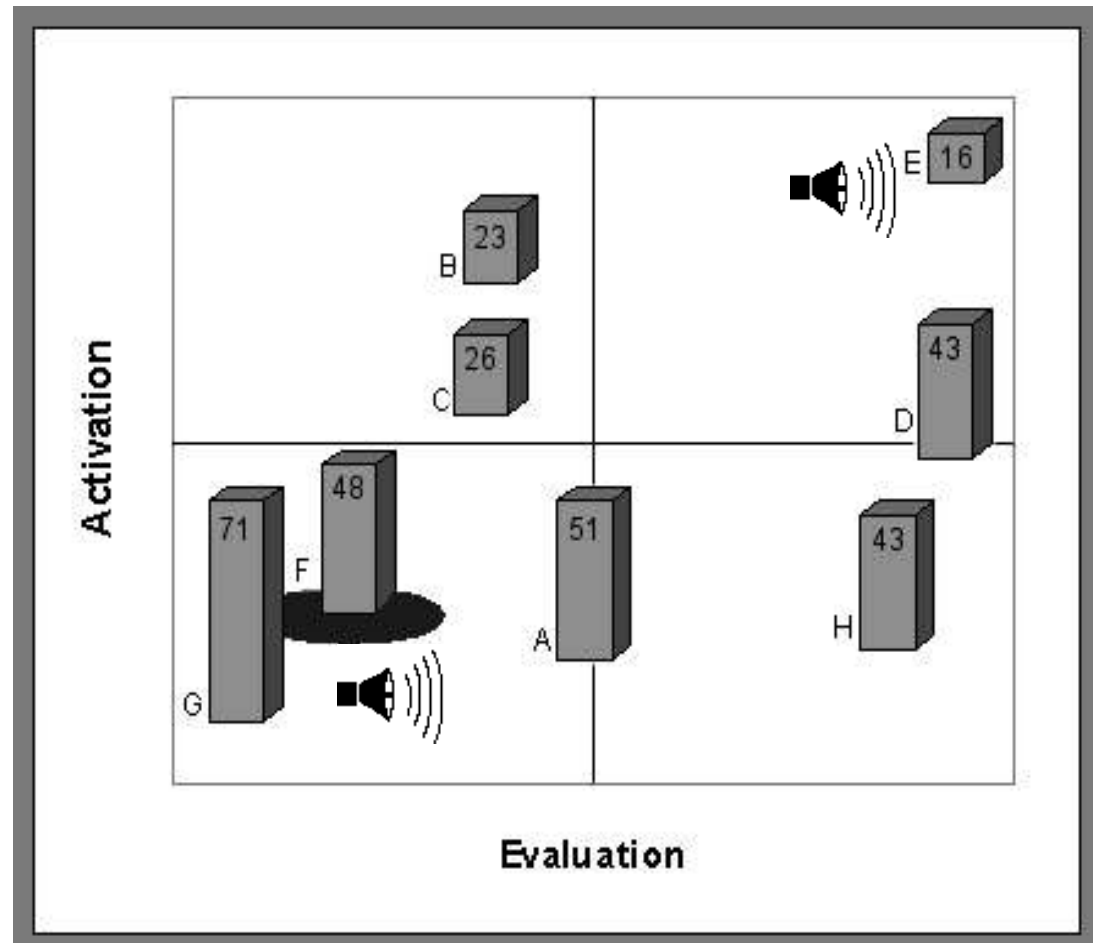
- ✦ *Activation dimension* successfully conveyed / perceived as intended
- ✦ *Evaluation dimension* less successful
- ✦ Acceptability is gradual: Neighbouring states more acceptable than distant states



Emotional speech synthesis research

Dimensional approach: Listening test results

- ✦ *Activation dimension* successfully conveyed / perceived as intended
- ✦ *Evaluation dimension* less successful
- ✦ Acceptability is gradual: Neighbouring states more acceptable than distant states



Emotional speech synthesis research

Dimensional approach: Summary

- ◆ Flexible framework
- ◆ Successful in expressing degree of activation
- ◆ Failure to express evaluation
 - sound of the smile?!
 - specialised modalities?
 - text => evaluation
 - voice => activation
- ◆ Emotional prosody rules not fine-tuned
 - only global evaluation so far

Summary

- ◆ Speech synthesis technology
 - data-driven or flexible
- ◆ Research on emotional prosody rules
 - listener-centred task
 - database analyses to be validated perceptually

Outlook: Speech synthesis research in HUMAINE

◆ Capability 5.2: Speech expressivity

- address the dilemma of data-driven vs. flexible
- investigate suitable measures for prosody and voice quality in controlled recordings
- attempt copy synthesis using different technologies
- attempt voice conversion
- evaluate success of different methods