KBM – Audio - Application

Frank Nack

Outline

- Last lecture
- Sound/tracks
- Animated Conversation
- Voice responsive head

Audio – summary

Investigated

- Audition
- Listening and vocal sound production
- Emotional layer of speech

Example

Pitch, rhythm, dynamics, primary and secondary emotions,
audio emotion recognition system

Findings

- Audition is a sign system but it is more complicated to identify the smallest sign unit, as the temporal nature of sound is essential.
- The properties of the acoustic signal are enough to identify a sound BUT individual characteristics are important (thresholds of frequency spaces)
- The interpretation of sound relies on context (other clues, such as facial expression or gestures for for spoken language, are of importance)
- Audio stimulates primary emotions, though the identification for each individual depends on its thresholds for arousal and valence.

sound/tracks

Real-Time Synaesthetic Sonification and Visualisation of Passing Landscapes

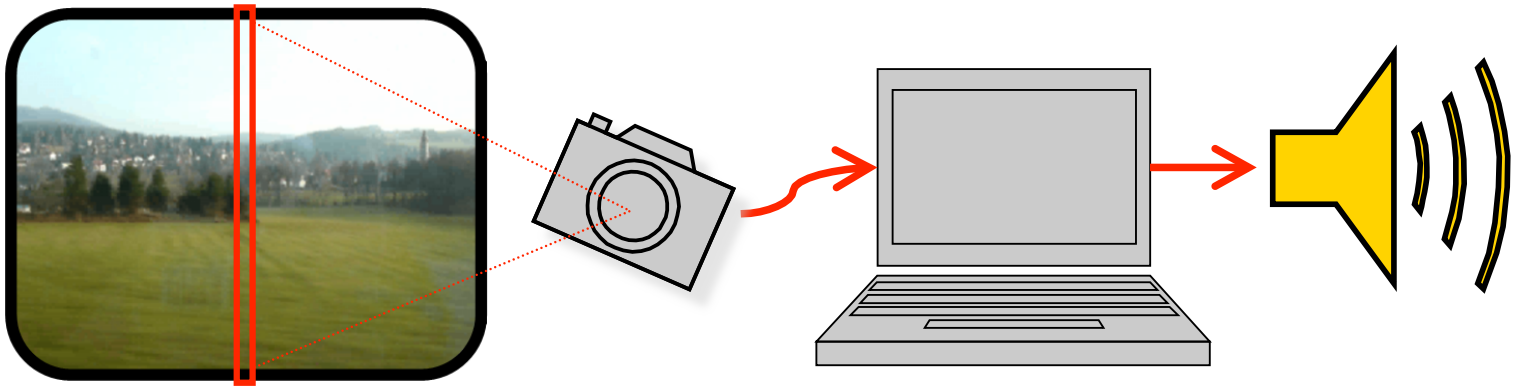
- Tim Pohle, **Peter Knees**, and Gerhard Widmer

  www.cp.jku.at/soundtracks

On a Train Journey…

Slides by Tim Pohle, **Peter Knees**, and Gerhard Widmer

Basic Idea

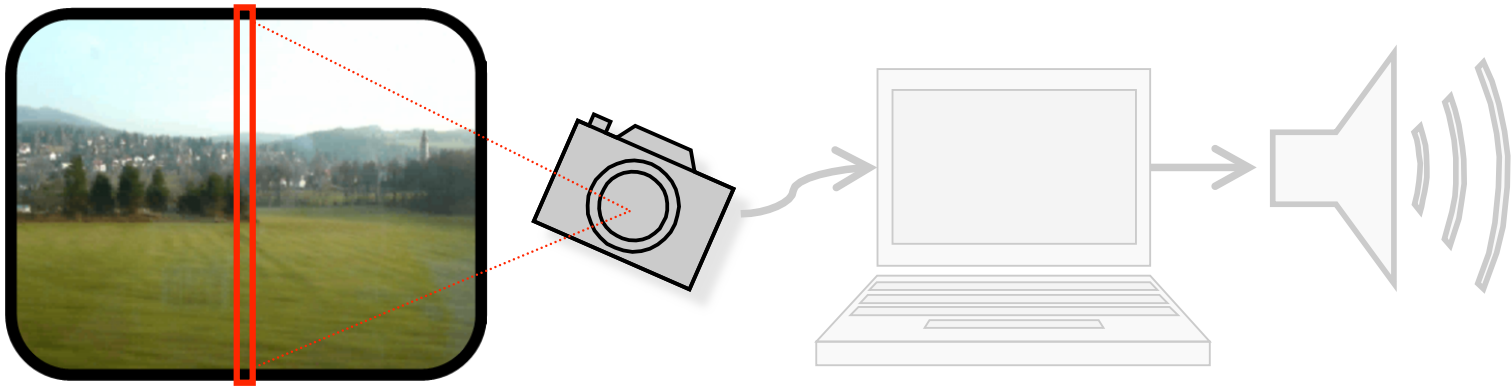


View outside train captured by camera
(e.g., Web-cam, camera built into mobile phone)

Image data is transformed to sound data
(using a Laptop or mobile phone)

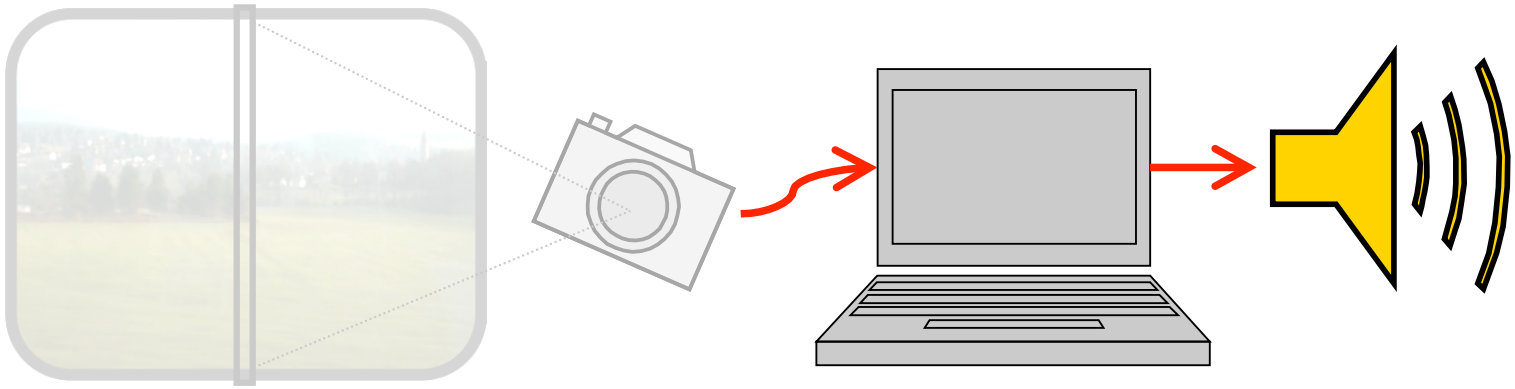
Sound is immediately played to the user

Technical Realisation – Step 1



- Capture at constant rate of 7 fps → “rhythm”
- Only central pixel column is processed

Technical Realisation – Step 2

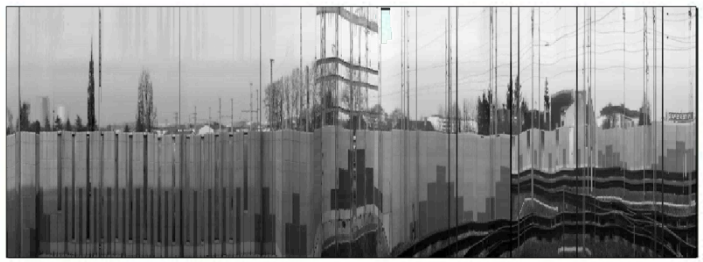


Sonification

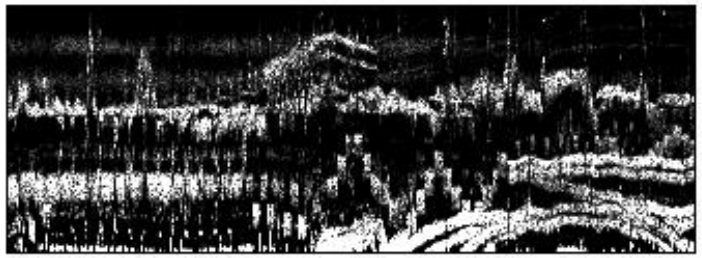
- **Filter Approach**
- **Piano Roll Approach**
- **Colour-based Approaches**
  - Historically-Inspired
  - Psychologically-Inspired

Filter Approach

Idea: Interpret pixels as spectrogram (→ iFFT)



time →



Spectrogram of synthesised sound

- Spatial height
  → **pitch**
- Pixel brightness (grayscale value)
  → intensity of frequency/**loudness**

Piano Roll Approach

Idea: Interpret pixels as piano roll (→ MIDI)

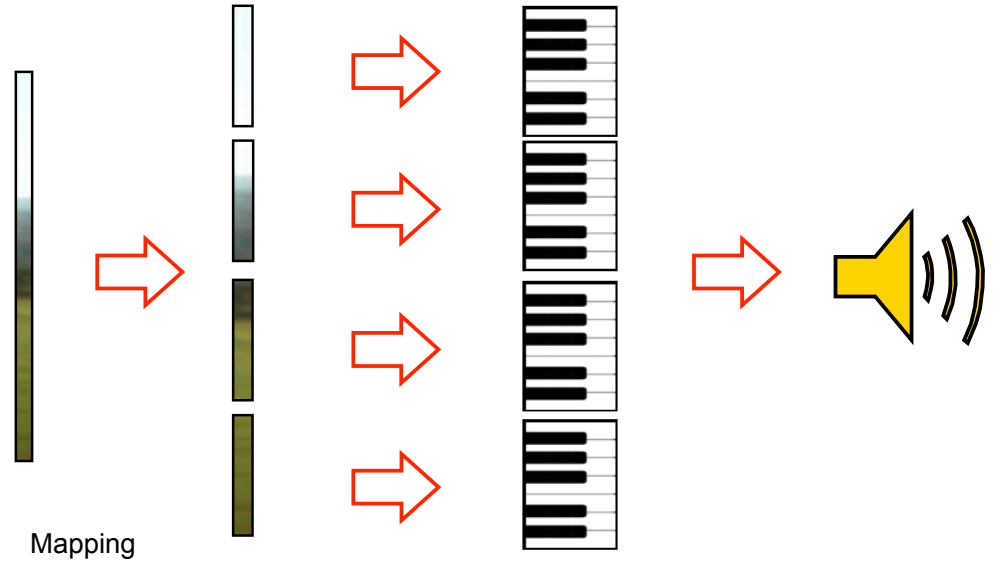
time →

Piano roll

- Spatial height → **pitch** (MIDI note number)
- Pixel brightness → **loudness** (velocity of MIDI note)
- Colour → **timbre** (MIDI instrument)

Colour-based Approaches

Idea: Colours are mapped to notes (piano)



- Spatial height → **octave**
- Pixel brightness → velocity/**loudness**
- Colour → **note**

Colour-to-Tone Mappings

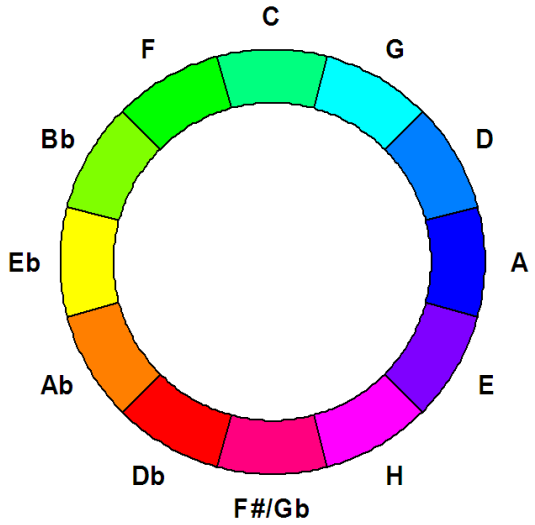
Historically-inspired

"Clavier à Lumières" by Alexander Skrjabin
(Russian composer and "synaesthete")

Psychologically-Inspired

Determine "root-note" → tonality

Between frames move along circle-of-fifths



The Score Image

Slides by Tim Pohle, **Peter Knees**, and Gerhard Widmer

Implementation

Java version
for Laptops

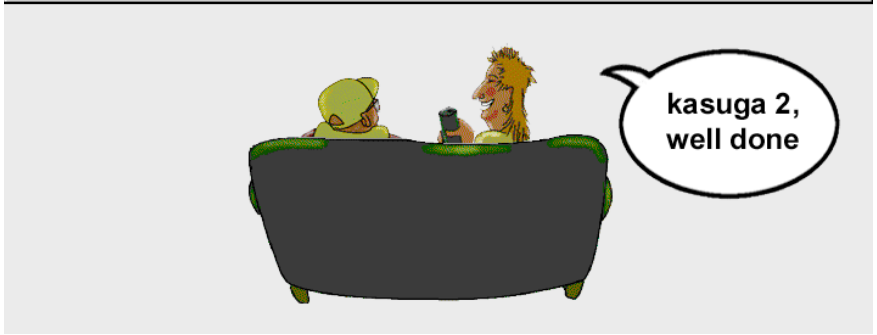
Python/C++ version
for Symbian phones

Slides by Tim Pohle, **Peter Knees**, and Gerhard Widmer

How to get it…

www.cp.jku.at/soundtracks

Animated conversation – Sport Commentator



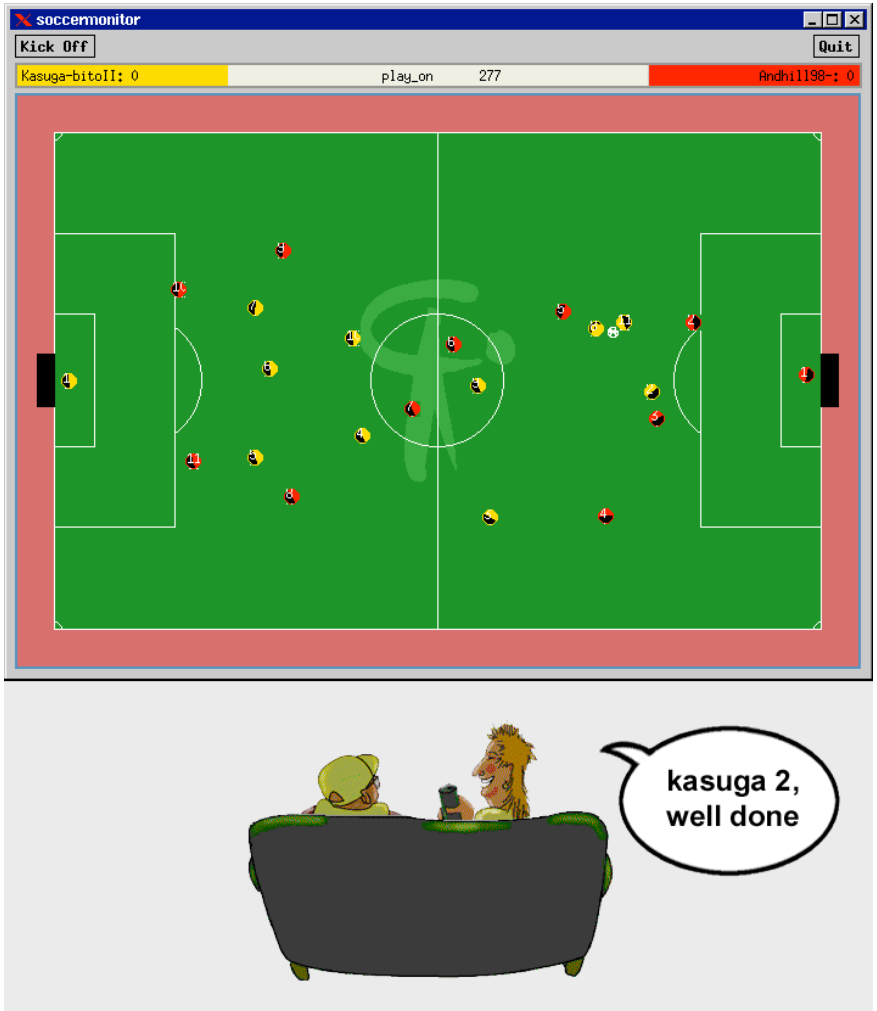
Trying to imitate the skills of human presenters, this work aims at using the notion of presentation teams, which convey information in the style of performances to be observed by him or her.

Andre & Rist (2000)

The following example is based on the work by Andre and Rist (2000)

On BB there is a video for the salesman example. This provides an idea of the actual speeach output.

Animated conversation – Sport Commentator



Framework

Dialogue Type: jointly watched events

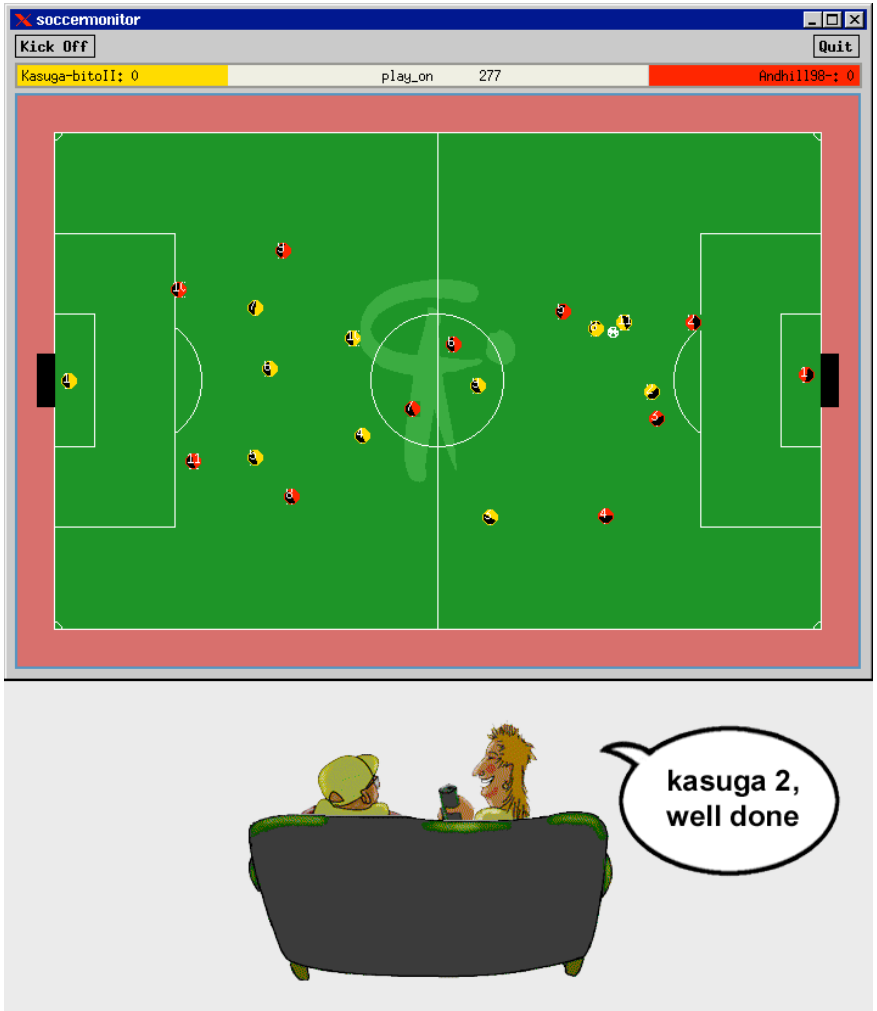
Character role: proponent, opponent

Character characteristics: psychological traits, such as introversion, openness or agreeableness => represented as a vector [discrete value , p-trait; .....]

Emotion is influenced by psychological trait

e.g. balanced character is lees angry about a missed goal of own team than an unbalanced character

Animated conversation – Sport Commentator



Autonomous actors

Each agents will be assigned

- a set of communicative goals which they try to achieve
- a repertoire of dialogue strategies

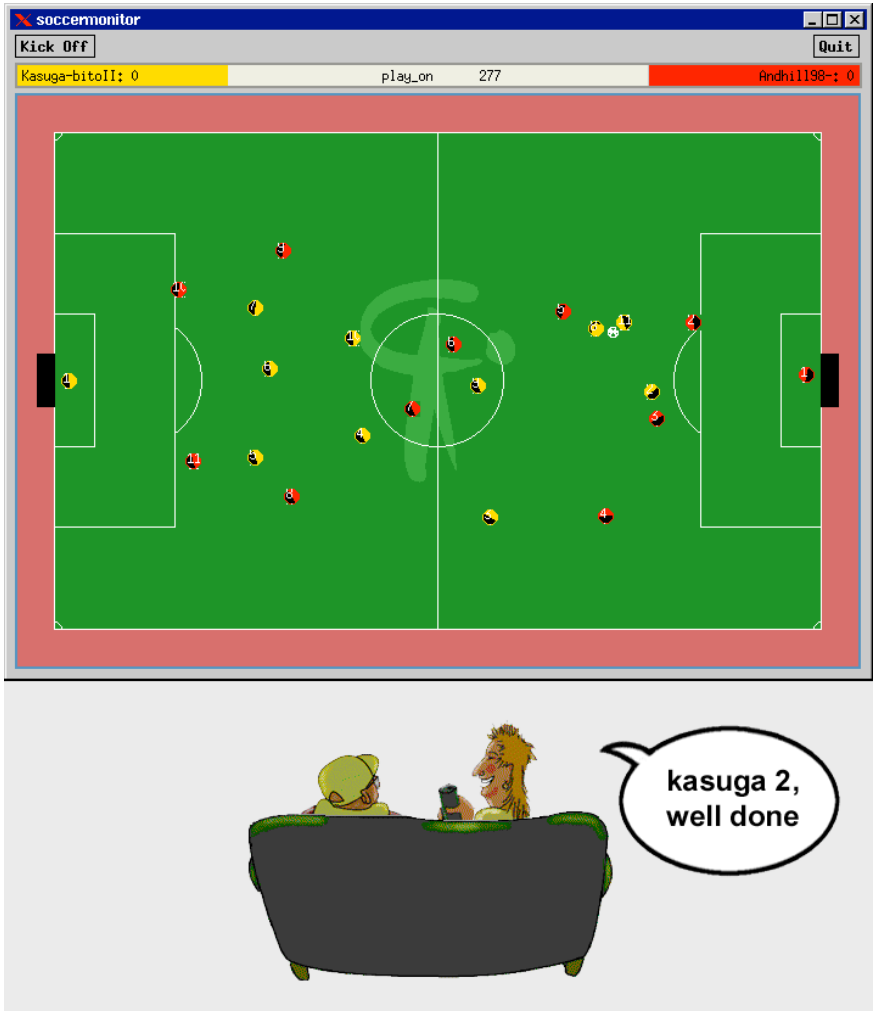
=> Determination and assignment of dialogue contributions is handled by the agent.

=> High demands on reactive capabilities (limited knowledge on what other agents say)

=> Difficult to ensure the coherence of the dialogue (synchronisation).

Potential solution: assigning each agent its own reactive planner. The agents' dialogue strategies are then realized as operators of the single planners.

Animated conversation – Sport Commentator



Instantiation

Based on speech act theory
(by saying something, we do something)

Input from Rocco II

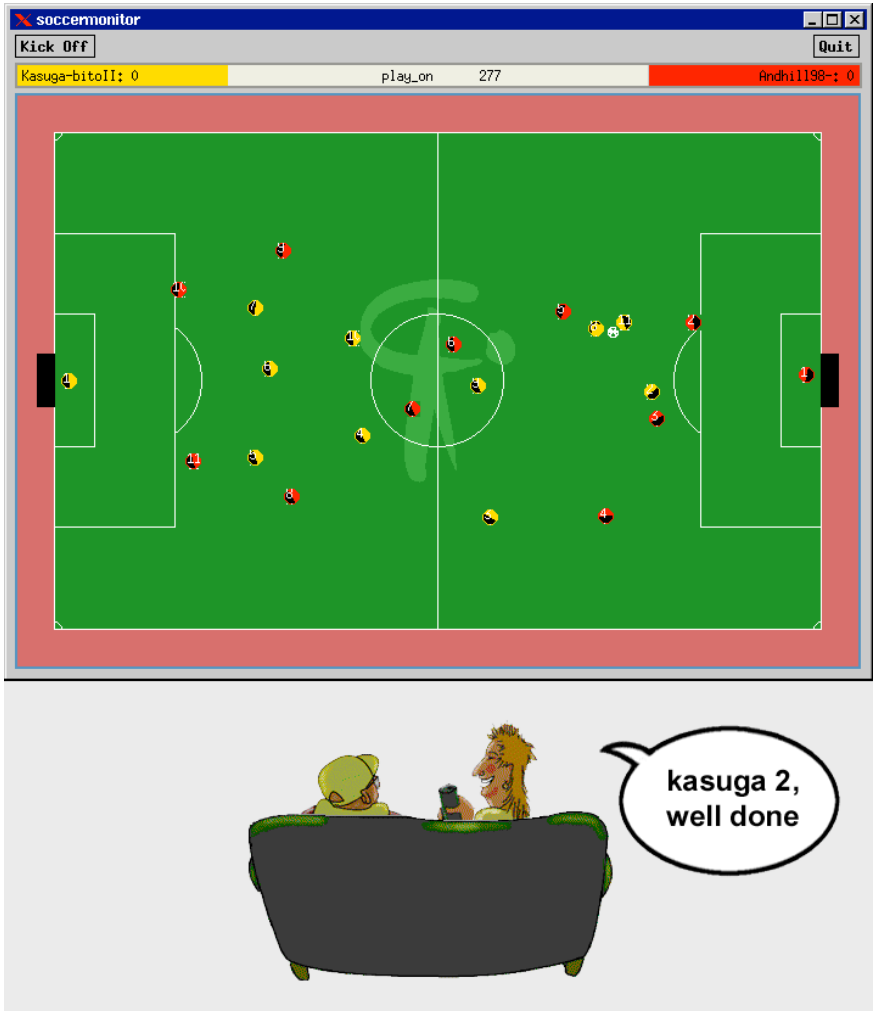
- player location and orientation (for all players),
- ball location
- game score
- play modes (e.g. goal kicks)

Character profiles:

P-Trait: extraversion (extravert, neutral, introvert)
openness (open, neutral, not open).

Emotion set: Arousal (calm, neutral, excited)
Valence (positive, neutral, negative)

Animated conversation – Sport Commentator



Speech generation

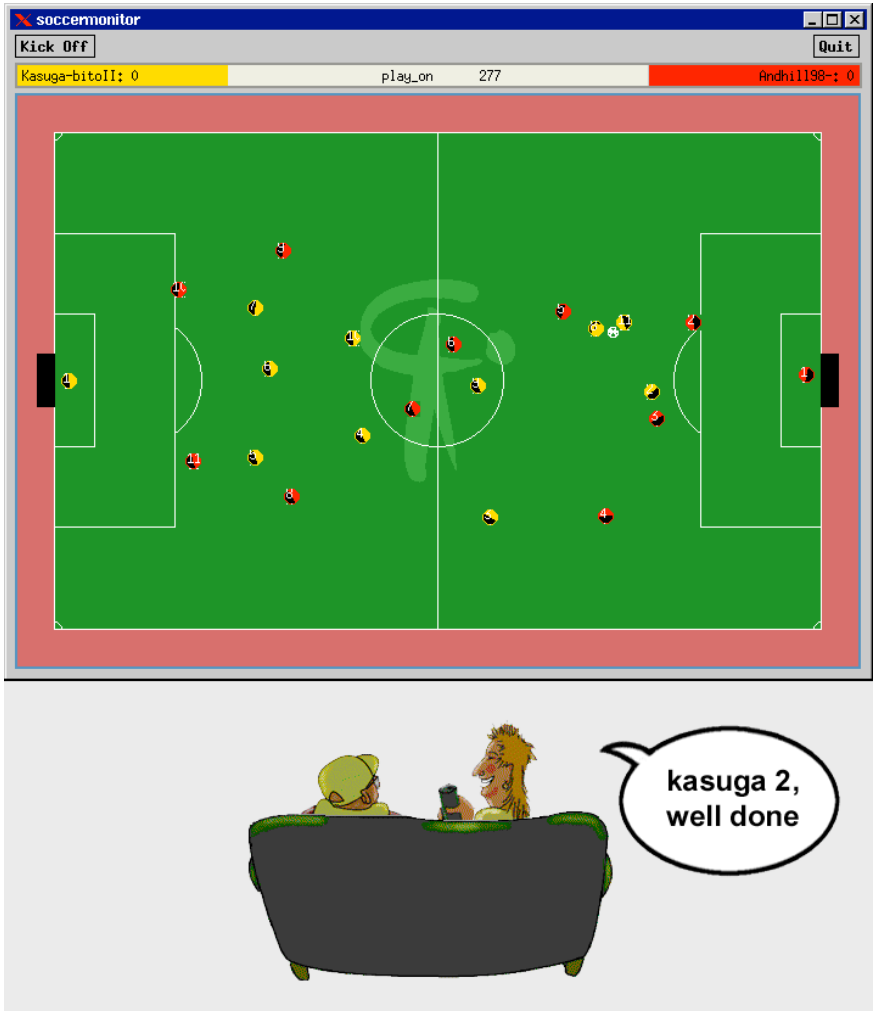
Based on a parametrized template-based generator.

13.5 hours of TV soccer reports in English => 300 basic templates.

Template parameter:

- Verbosity (length)
- Specificity (detail)
- Force: powerful, normal, hesitant,
- Floridity: dry, normal, flowery
- Formality: formal, colloquial, slang
- Bias: negative, neutral, positive.

Animated conversation – Sport Commentator



Speech generation II

Template selection process:

- Accommodate for the temporal constraints of a real-time live report.
- Avoid repetition of templates
- Communicate the speaker's attitude
- Considers the speakers' personality.

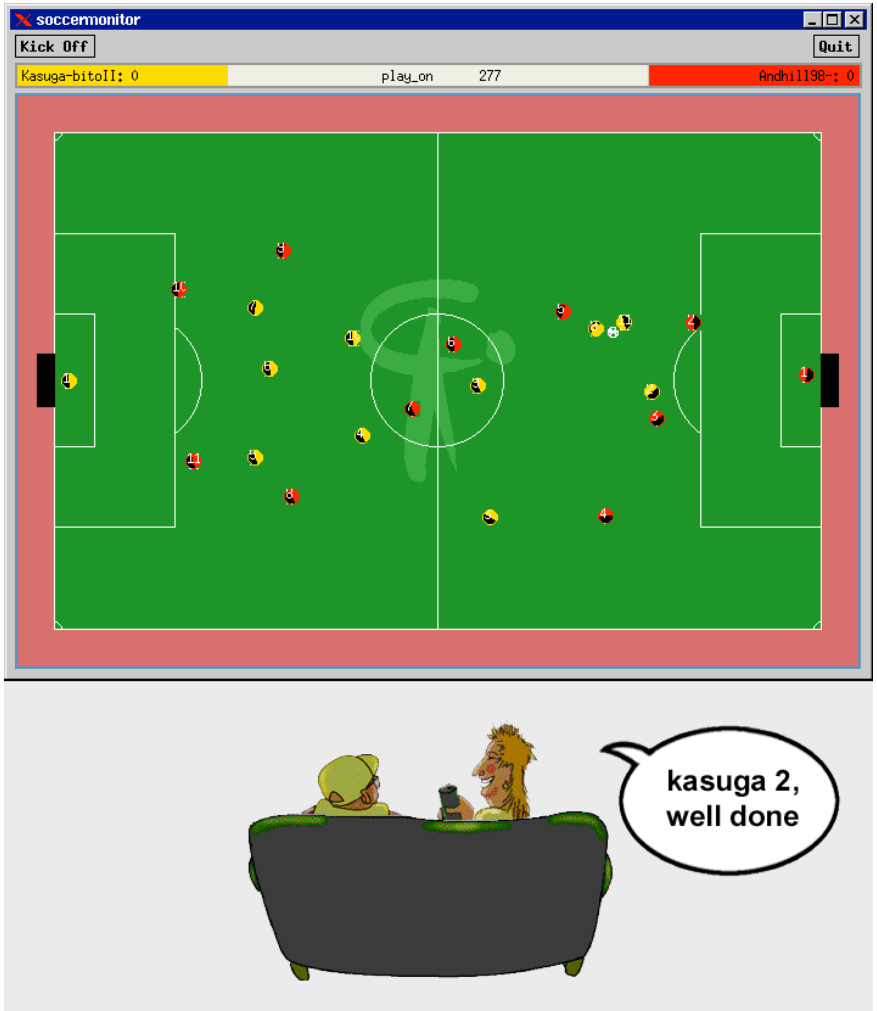
Acoustic realisation inspired by Cahn

Uses the TrueTalk speech synthesizer

Applies: pitch accent, pitch range and speed.

Excitement:- higher talking speed and pitch range.

Animated conversation – Sport Commentator

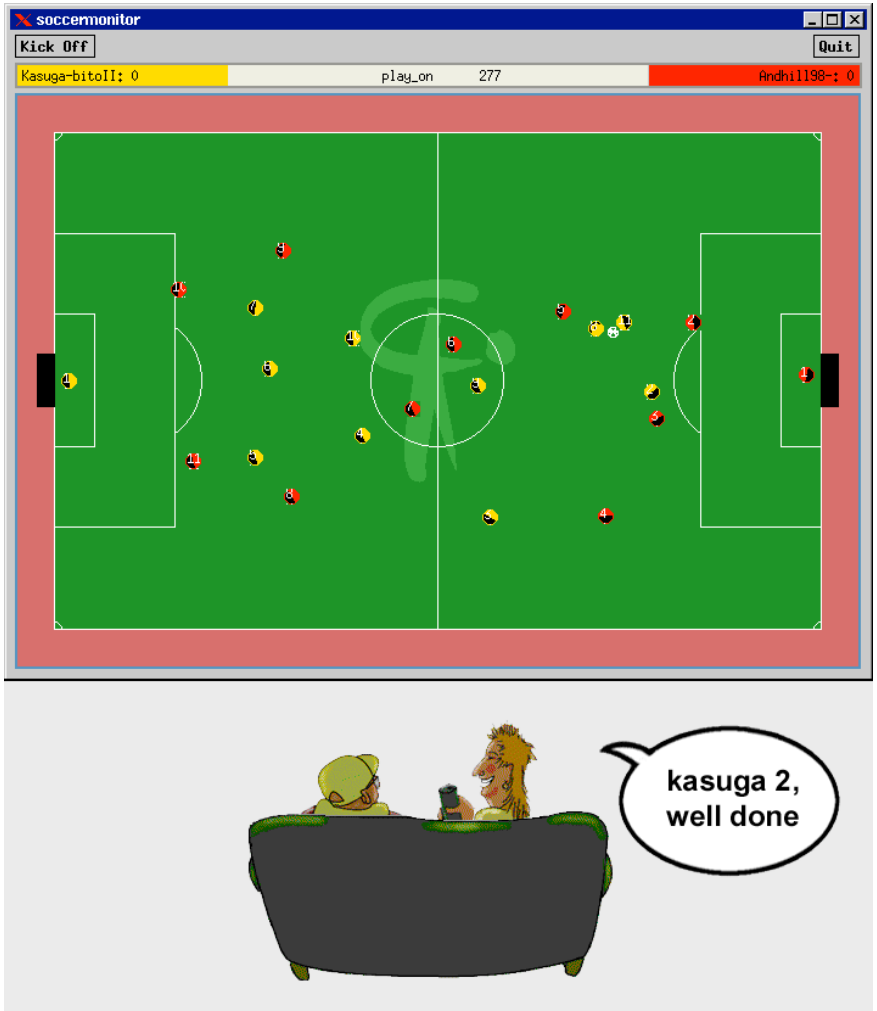


Generated speech

Agent	Attitude	Personality factors
Gerd	in favor of team Kasuga	extravert, open
Matze	neutral	introvert, not open

Gerd: *Kasuga kicks off*
;;; recognized event: kick off
Matze: *Andhill 5*
;;; recognized event: ball possession, time pressure
Gerd: *We're live from an exciting game, team Andhill in red versus Kasuga in yellow*
;;; time for background information
Matze: *Now Andhill 9*
;;; recognized event: ball possession
Gerd: *Super interception by yellow 4*
;;; recognized event: loss of ball, attitude: pro Kasuga,
;;; forceful language because it is extravert
still number 4
;;; recognized event: ball possession, number 4 is
;;; topicalized
Matze: *Andhill 9 is arriving*
;;; recognized event: approach
Gerd: *ball hacked away by Kasuga 4*
;;; recognized event: shot, flowery language since it is
;;; creative

Animated conversation – Sport Commentator



Results

First informal system tests were encouraging. Even though it was not the intention to make use of humor people found both scenarios entertaining and amusing.

Furthermore, people were very eager about to test various role castings in order to find out which effect this would have on the generated presentations.

These observations suggest that people possibly learn more about a subject matter because they are willing to spend more time with a system.

Voice responsive head

University of Augsburg: Emotion recognition from speech

University of Bielefeld: Anthropomorphic robot BARTHOC Jr.
Robot recognizes emotional content (happiness, fear and neutral) from speech and mirrors it by facial expressions.

Voice responsive head

University of Augsburg: Emotion recognition from speech
Greta agent from Pelachaud et al.
http://perso.telecom-paristech.fr/~pelachau/Greta/

Aim

The agent does not analyse the meaning of the user's verbal utterances, but instead just interprets the user's emotive cues from speech and responds to them emotionally range.

See for more info and paper:
https://www.informatik.uni-augsburg.de/de/lehrstuehle/hcm/projects/tools/emovoice/index.html

Voice responsive head

University of Augsburg: Emotion recognition from speech
Greta agent from Pelachaud et al.
http://perso.telecom-paristech.fr/~pelachau/Greta/

Emotions

Fear, anger, joy, boredom, sadness, disgust, neutral.

Segmentation

Voice activity detection with no in-between pauses longer than 1000 ms

Advantage: very fast
Disadvantage: segment might not be linguistically sound

Feature space

20 features related to pitch, MFCCs and energy
http://www.fon.hum.uva.nl/praat/
http://patrec.cs.tu-dortmund.de/cms/en/home/Research/ESMERALDA/index.html

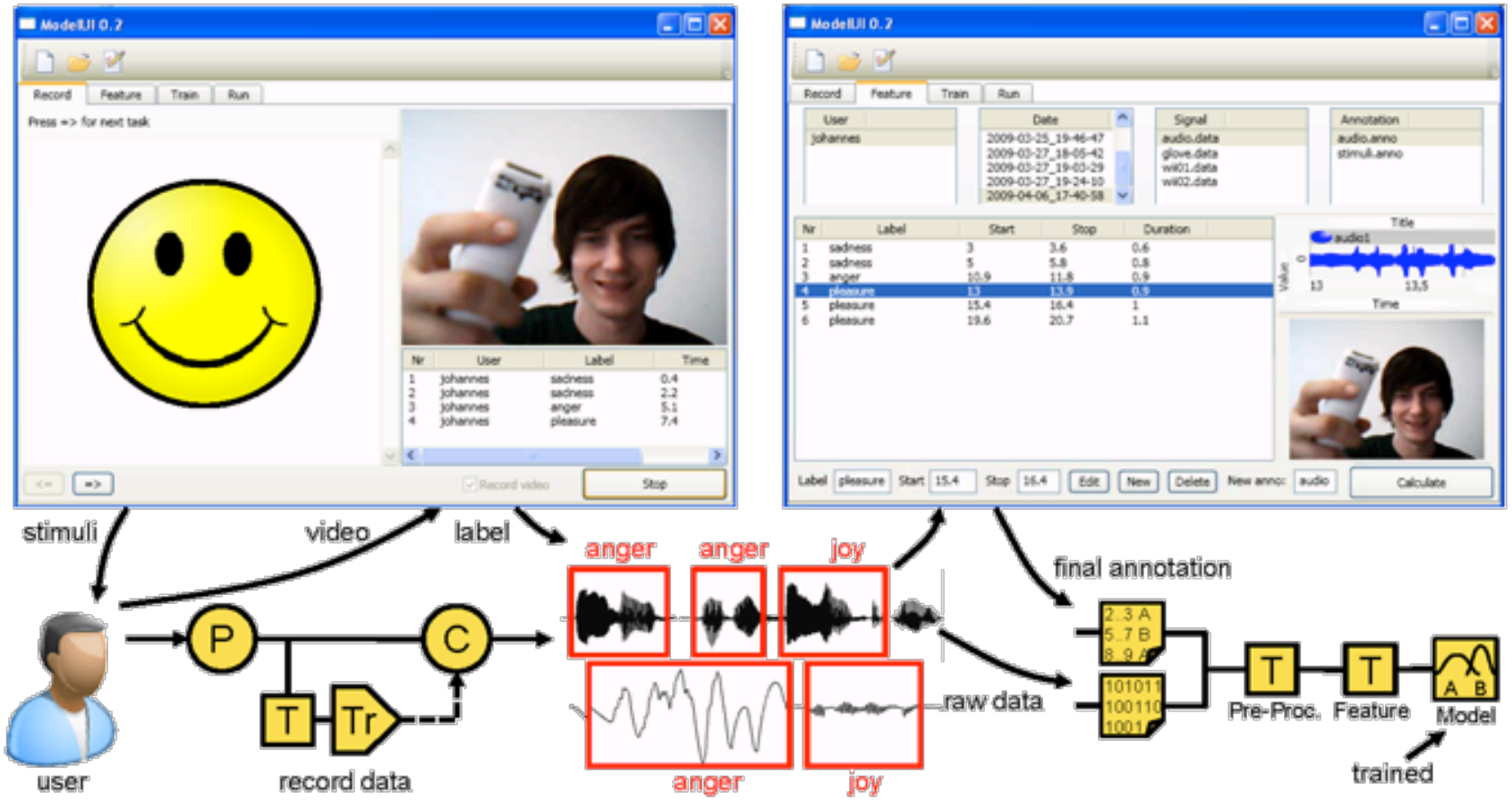
Classification

Naive Bayes classifier => simple but fast

Training

Berlin emotional speech database (very prototypical, acted emotions)

Voice Sensor Toolkit



University of Augsburg: Smart Sensor Integration (SSI)
http://mm-werkstatt.informatik.uni-augsburg.de/ssi.html

Audio Applications – References

Audio Applications – References

Andre, E. & Rist, T. (2000). Presenting through Performing: On the Use of Multiple Lifelike Characters in Knowledge-Based Presentation Systems. IUI 2000 New Orleans LA USA

Cahn, J. The Generation of Affect in Synthesized Speech. Journal of the American Voice I/O Society, 8: I- 19, 1990.

Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., Douville, B., Prevost, S., Stone, M. (1994) "Animated Converstaion: Rule-Based Generation of Facial Expression, Gesture and Spoken Intonation for Multiple Conversational Agents." Proceedings of SIGGRAPH '94, pp. 413-420.

W.S. Condon and W.D. Osgton. Speech and body motion synchrony of the speaker-hearer. In D.H. Horton and J.J. Jenkins, editors, The perception of Language, pages 150–184. Academic Press, 1971.

de Rosis, F., Pelachaud, C., Poggi, I., Carofiglio, V., de Carolis, B.: From Greta's mind to her face: modelling the dynamics of affective states in a conversational embodied agent. International Journal of Human-Computer Studies 59 (2003) 81–118

Pierrehumbert, Janet: 1980, The Phonology and Phonetics of English Intonation, Ph.D dissertation, MIT. (Dist. by Indiana University Linguistics Club, Bloomington, IN.)

Klaus R. Scherer. The functions of nonverbal signs in conversation. In H. Giles R. St. Clair, editor, The Social and Physhological Contexts of Language, pages 225–243. Lawrence Erlbaum Associates, 1980.

Scott Prevost and Mark Steedman. Generating contextually appropriate intonation. In Proceedings of the 6th Conference of the European Chapter of the Association for Computational Linguistics, pages 332–340, Utrecht, 1993.