Combining Semantic and Acoustic Features for Valence and Arousal Recognition in Speech

Seliz Karadogan

Jan Larsen

Cognitive Systems Section

Dept. of Informatics and Mathematical Modelling

Technical University of Denmark

DTU Informatics

Department of Informatics and Mathematical Modeling

OVERALL GOAL

EXTRACT COGNITIVE INFORMATION TO DESIGN EFFICIENT AND NATURAL INTERACTIVE SYSTEMS

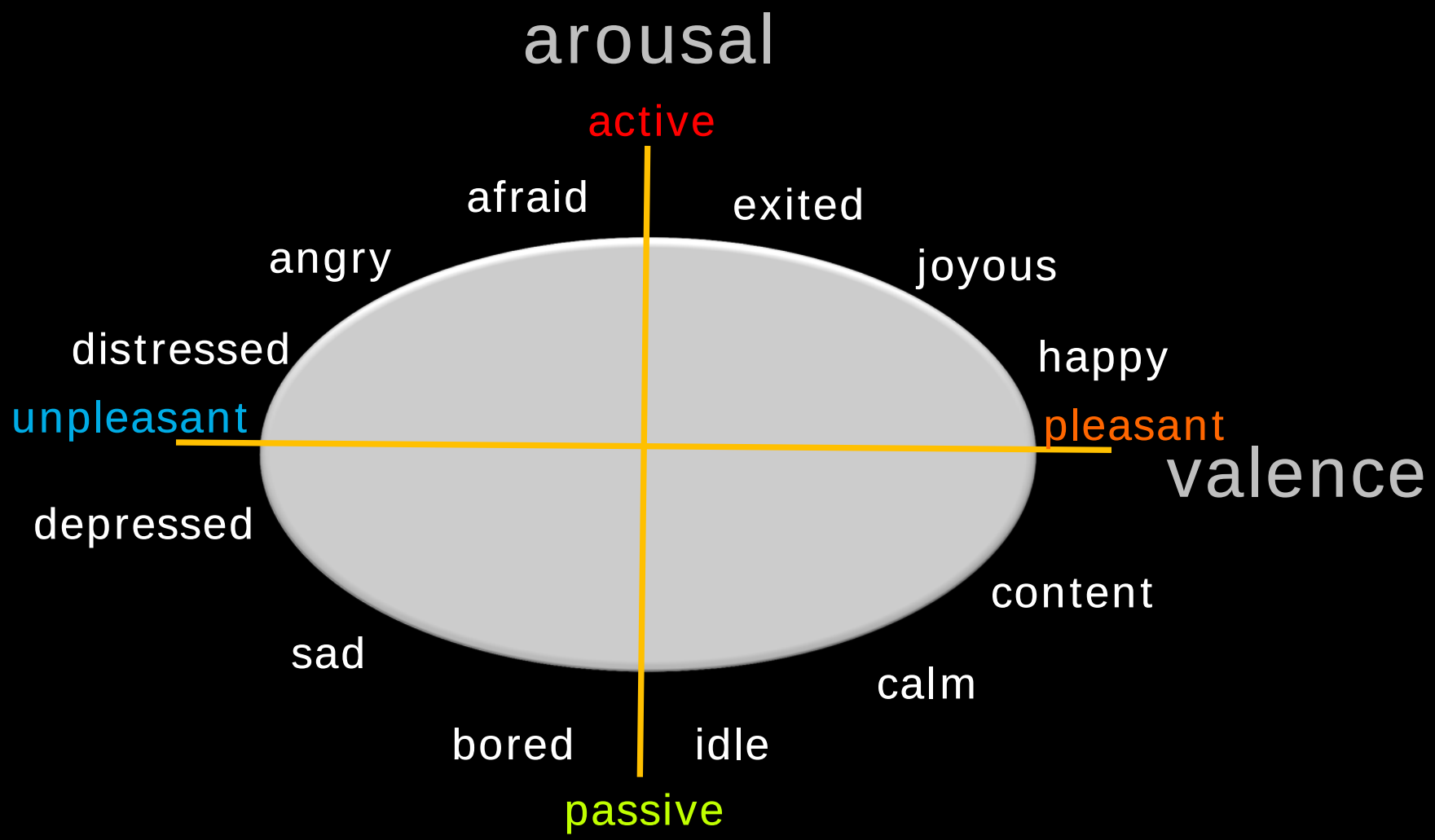
Approaches to emotional modeling

- Categorical: happiness, sadness, surprise etc.
- Dimensional: valence, arousal, dominance etc.

Emotions lie in audio-visual information

- Visual: Gesture, mimics
- Audio: Speech (acoustical and textual information)

Emotional spaces



The focus of emotion recognition in speech has been towards categorical approach using acoustic information

Scientific question: How do acoustical and textual/semantic information from speech influence the dimensional (valence and arousal) speech affect recognition?

Engineering question: Can we design a prediction model for AV values

B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04)*, vol. 1. IEEE, 2004, pp. I–577.

Z. Chuang and C. Wu, "Multi-modal emotion recognition from speech and text," *International Journal of Computational Linguistics and Chinese Language Processing*, vol. 9, no. 2, pp. 45–62, 2004.

M. El Ayadi, M. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

H. Gunes and M. Pantic, "Automatic, dimensional and continuous emotion recognition," *International Journal of Synthetic Emotions*, vol. 1, no. 1, pp. 68–99, 2010.

DTU

DTU

DR

Syntonetic

Musikzonen

Geckon

COSOUND

A COGNITIVE SYSTEMS APPROACH TO ENRICHED AND ACTIONABLE INFORMATION FROM AUDIO STREAMS

UCL

Royal School of Library and Information Science

Hindenburg Systems

B&O

Queen Mary University of London

Aalborg University

Copenhagen University

University of Glasgow

State and University Library

Vision

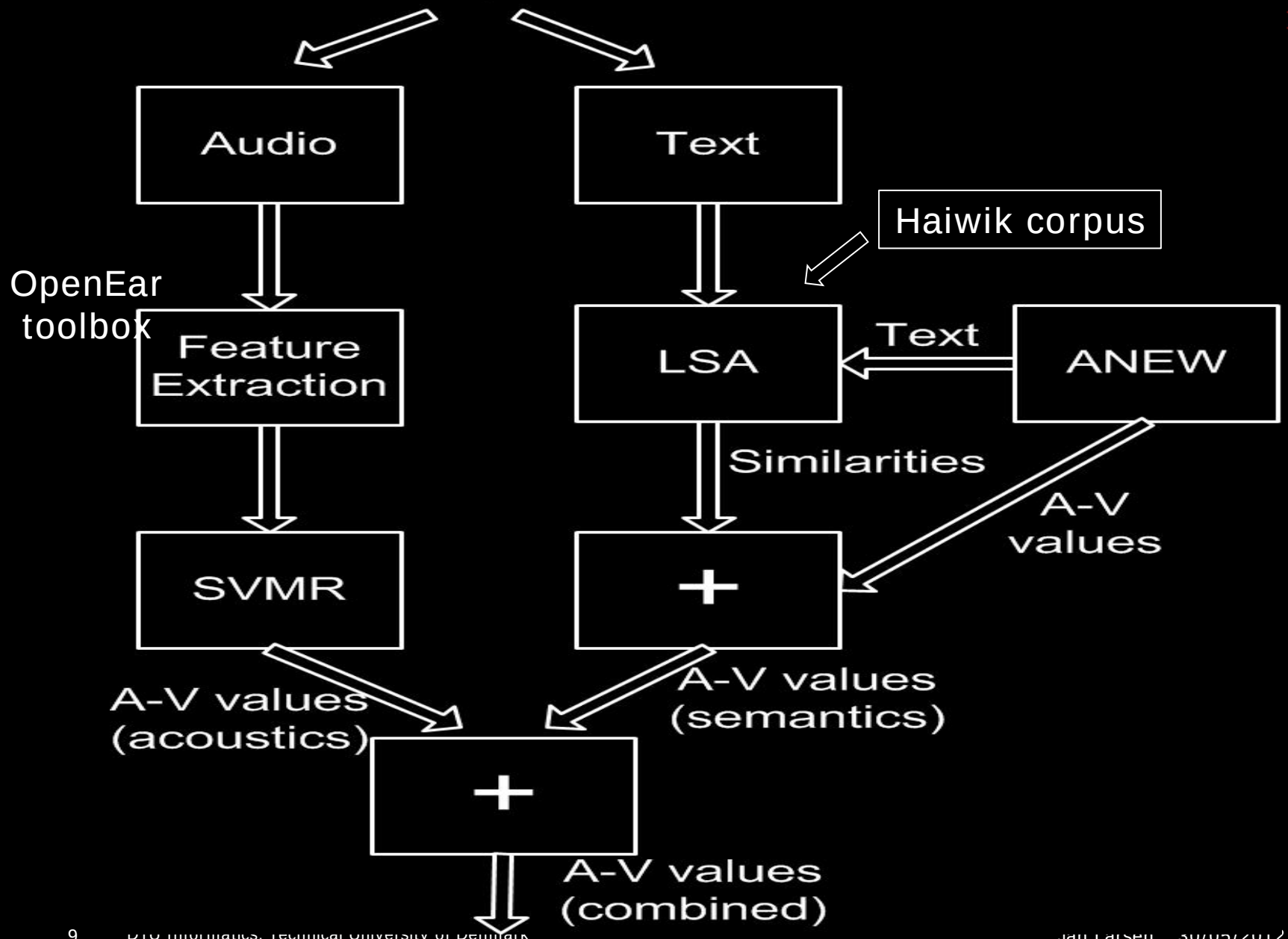
The overall vision is to foster truly participatory, collaborative, and cross-cultural tools for enrichment of audio streams which can improve interactivity, findability, experienced quality, ability to co-create, and boost productivity in a broad sense.

users in the loop framework – required to study and evaluate interactive and participative (crowd) designs

CoSound Hypothesis

The main hypothesis is that the integration of bottom-up data derived from audio streams and top-down data streams from users can enable actionable cognitive representations, which will positively impact and enrich user interaction with massive audio archives, as well as facilitating new commercial success.

We will test the hypothesis at three different functionality levels: 1) personalized audio streams; 2) task driven navigation and organization; 3) sharing of enriched audio streams through editing and co-creation.

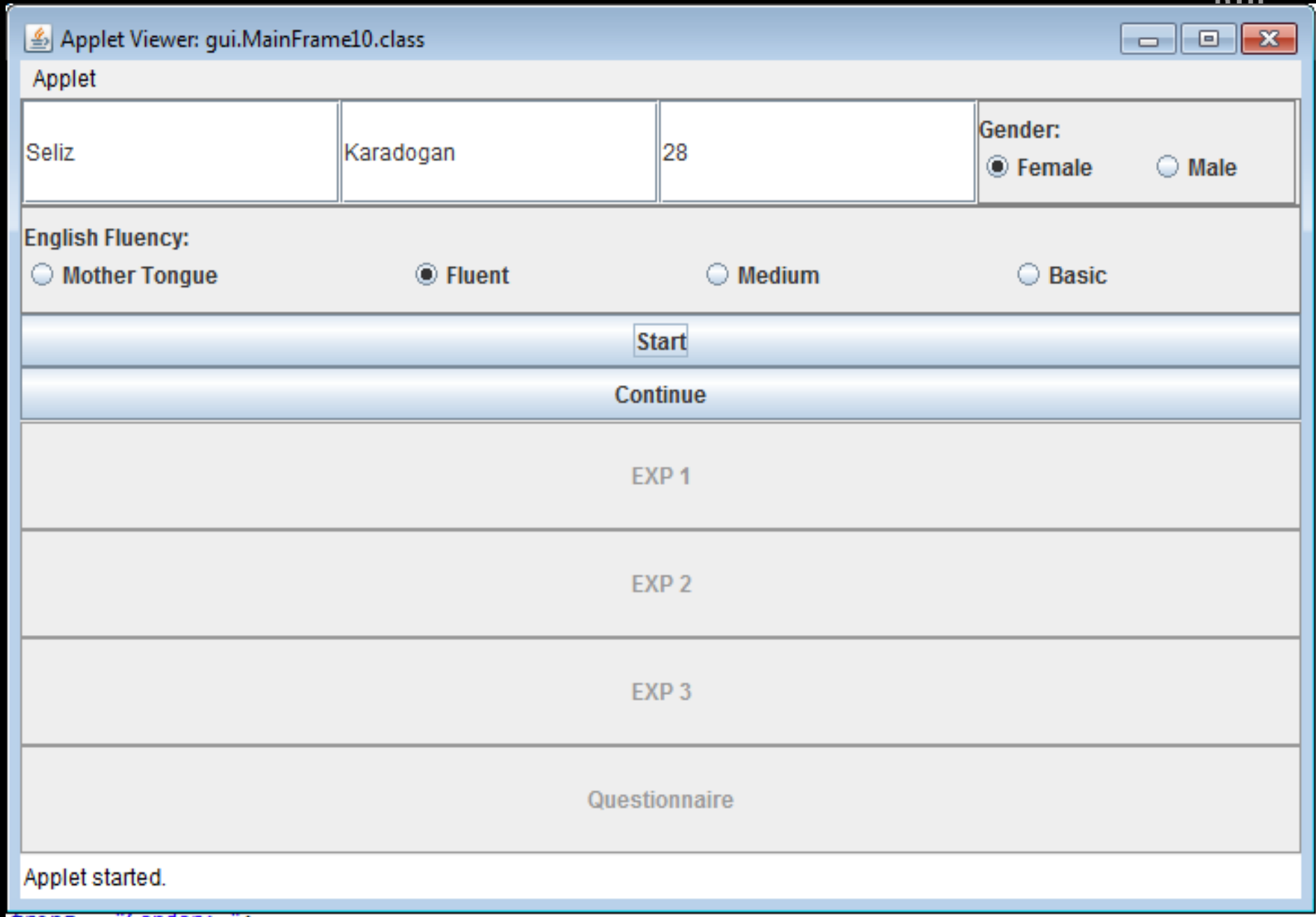

Movie clips
Audio
Text
Haiwik corpus
OpenEar toolbox
Feature Extraction
LSA
Text
ANEW
Similarities
A-V values
SVMR
+
A-V values (acoustics)
A-V values (semantics)
+
A-V values (combined)

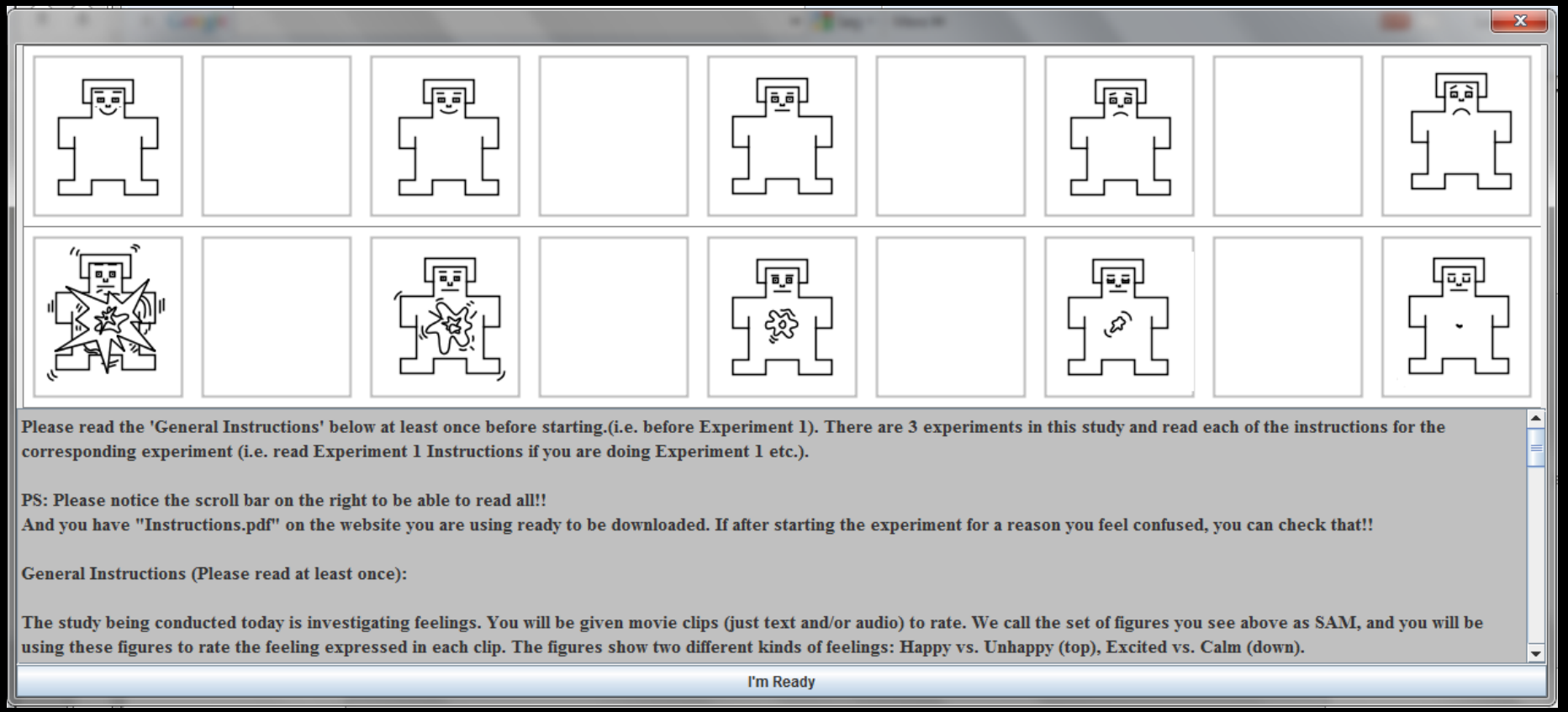
Challenge: Lack of available speech emotional databases

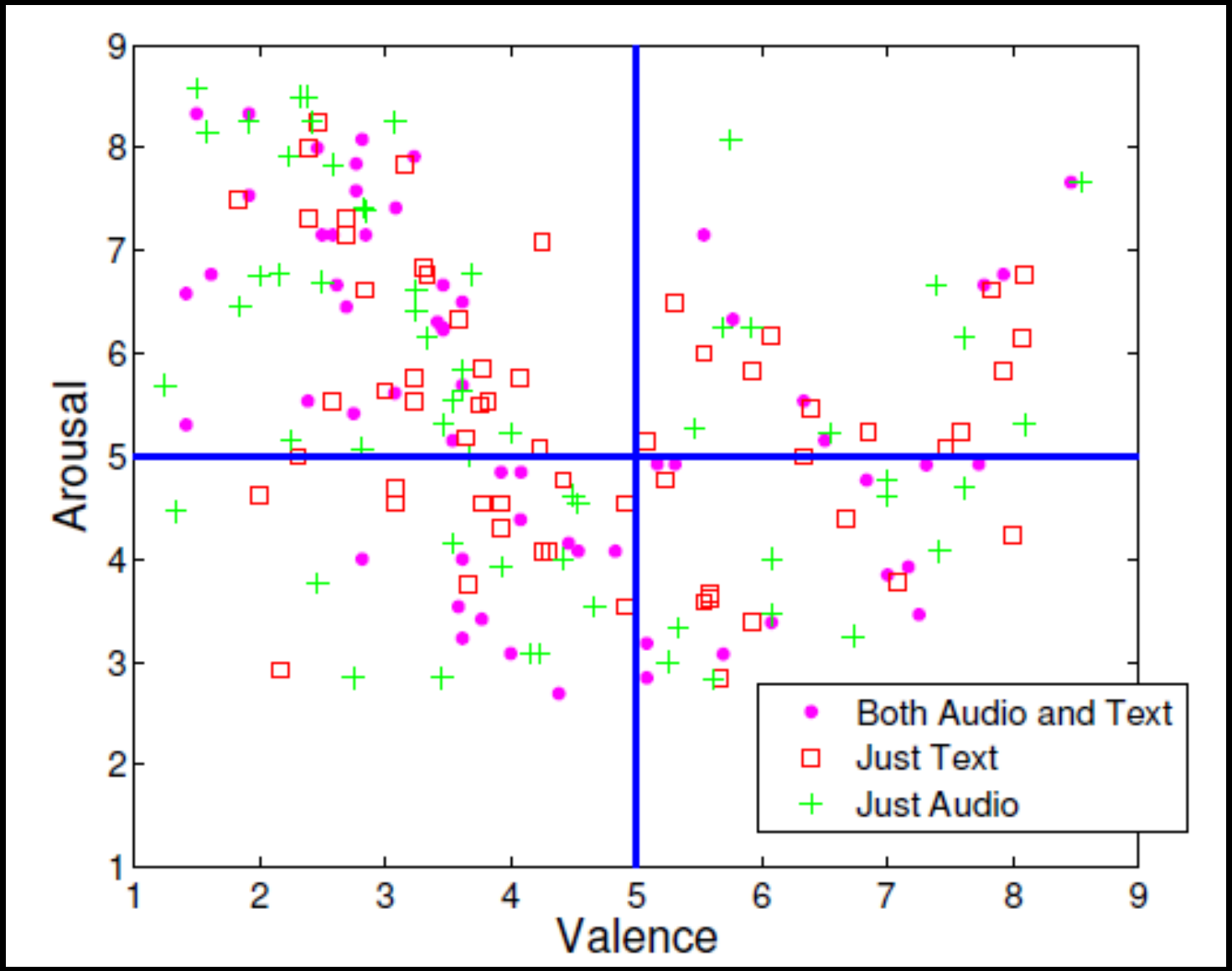
Solution: Design of a new database

- 59 clips (29 train, 30 test) in total from 11 movies between 5 and 25 seconds.
- Annotators: 13 people (7 female, 6 male) between ages 19 and 28
- Three experiments:
  - Just text
  - Just audio
  - Both text and audio
- A java applet has been created available

  http://www.student.dtu.dk/~seka/


Applet Viewer: gui.MainFrame10.class
Applet
Seliz
Karadogan
28
Gender:
Female
Male
English Fluency:
Mother Tongue
Fluent
Medium
Basic
Start
Continue
EXP 1
EXP 2
EXP 3
Questionnaire
Applet started.


Please read the 'General Instructions' below at least once before starting.(i.e. before Experiment 1). There are 3 experiments in this study and read each of the instructions for the corresponding experiment (i.e. read Experiment 1 Instructions if you are doing Experiment 1 etc.).
PS: Please notice the scroll bar on the right to be able to read all!!
And you have "Instructions.pdf" on the website you are using ready to be downloaded. If after starting the experiment for a reason you feel confused, you can check that!!
General Instructions (Please read at least once):
The study being conducted today is investigating feelings. You will be given movie clips (just text and/or audio) to rate. We call the set of figures you see above as SAM, and you will be using these figures to rate the feeling expressed in each clip. The figures show two different kinds of feelings: Happy vs. Unhappy (top), Excited vs. Calm (down).
I'm Ready


Arousal
Valence
1
2
3
4
5
6
7
8
9
Both Audio and Text
Just Text
Just Audio

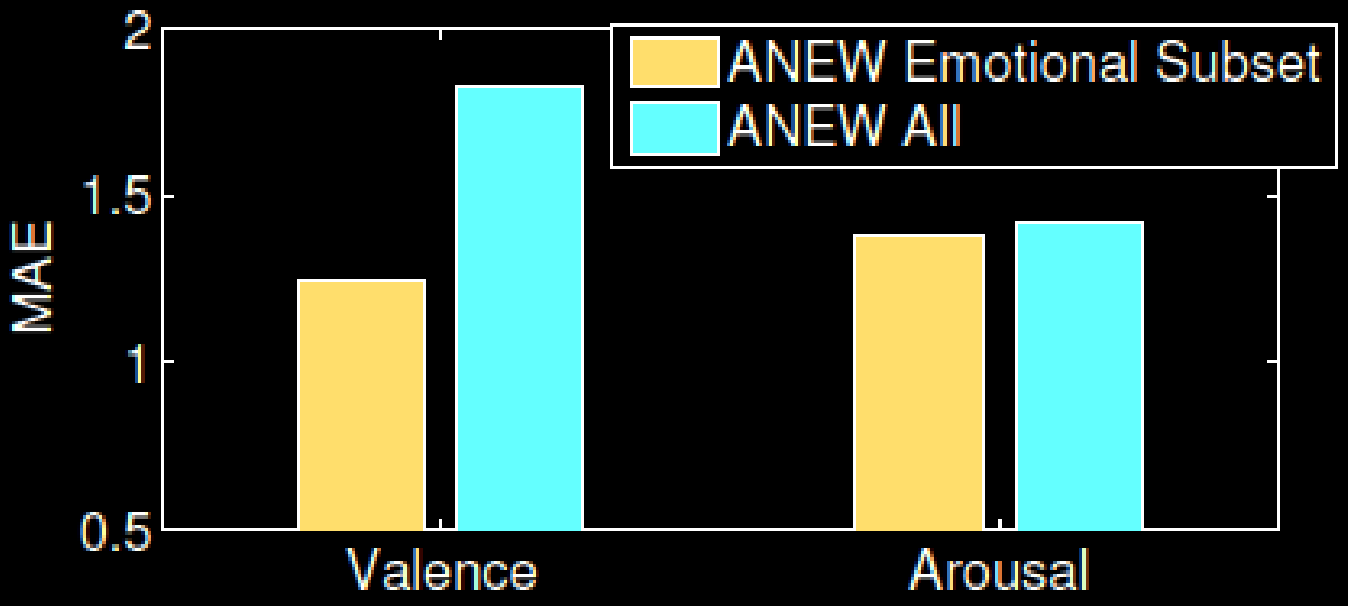
Semantic and Acoustic analysis are done independently and then combined

$$A\text{-}V = \beta * A\text{-}V_{semantics} + (1-\beta)\, A\text{-}V_{acoustics}$$

A-V represents arousal OR valence values while

β represents the weight of semantics

Pruning of ANEW database



Mean absolute error test performance

	Textual information	Acoustical informtion	β (weight of textual information)	Combined result
Valence (MAE)	1.45	1.98	0.80	1.40
Arousal (MAE)	1.39	1.29	0	1.29

Conclusions

- combining acoustic and semantic information improves the recognition results
- the valence dimension is recognized better using semantic features while the arousal dimension using acoustic features

the valence dimension is more about what we say, while the arousal dimension is more about how we say it

Acknowledgments and references


Bjørn Sand Jensen


Jens Brehm Nielsen


Jens Madsen


Seliz Karadogan

- S. Karadogan, J. Larsen: ”Dimensional Emotion Recognition from Speech Combining Semantic and Acoustic Features, submitted for IEEE Transaction on Affective Computing, 2012.
- B.S. Jensen, J.S. Gallego, J. Larsen: ”A Predictive Model of Music Preference Using Pairwise Comparisons,” IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP2012), Kyoto, Japan, March 2012.
- J. Madsen, J.B. Nielsen, B.S. Jensen, J. Larsen: ”Modeling Expressed Emotions in Music using Pairwise Comparisons,” 9’th International Conference on Computer Music Modellng and Retrieval (CMMR2012), Queen Mary University, UK, June 2012.
- J. Madsen, B.S. Jensen, J. Larsen, J.B. Nielsen, : Towards Predicting Expressed Emotion in Music from Pairwise Comparisons, Sound and Music Computation SMC2012, Copenhagen, July 2012.
- J.B. Nielsen, B.S. Jensen, J. Larsen: ”On Sparse Multi-Task Gaussian Process Priors for Music Preference Learning,” In NIPS CMPL workshop, pages 1-8, December 2011.
- B. S. Jensen and J. B. Nielsen: “Pairwise Judgments and Absolute Ratings with Gaussian Process Priors,” Technical report, DTU November 2011.
- J.B. Nielsen, B.S. Jensen, J. Larsen: “Learning from Pairwise Observations with Gaussian Processes: Review and Extensions,’” in preparation for submission 2012.
- B.S. Jensen, J.B. Nielsen, J. Larsen: ”Efficient Preference Learning with Pairwise Continuous Observations and Gaussian Processes,” IEEE Workshop on Machine Learning for Signal Processing (MLSP2011), Beijing, China, September 2011.