Speaker Classification

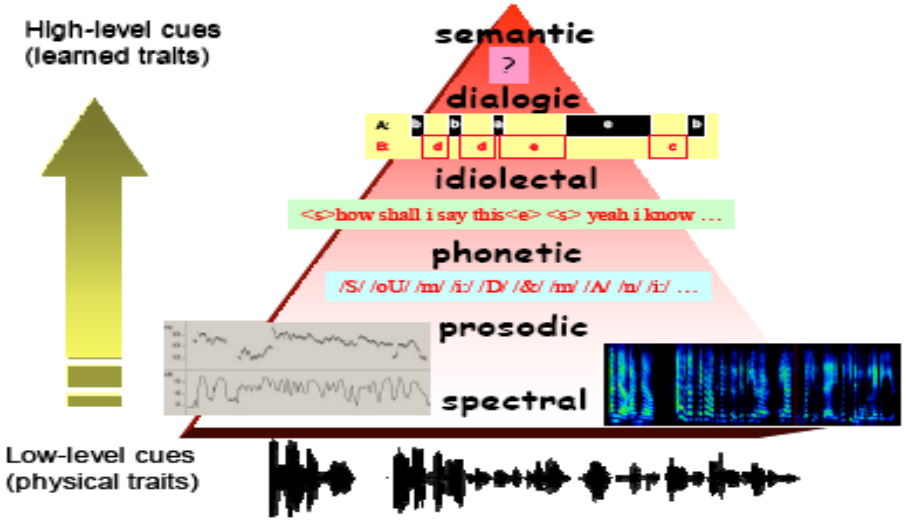
Felix Burkhardt

Contents of talk

- Speaker classification
  - Applications
  - Technical approaches
  - Evaluation

Speech features

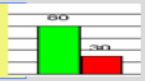
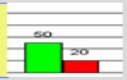
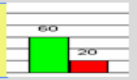
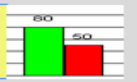
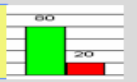
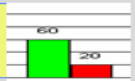
- Telephone Speech: digital signal, 8 kHz sample rate, 16 bit linear PCM, usually decoded from GSM or a-law
- Time domain vs frequency domain > Fast Fourier Transform (FFT)
- Mel Frequency Cepstral Coefficients (MFCC) used in Automatic Speech Recognition (ASR) to model short frame (10-20ms) frequency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



graphic: Reynolds et al 2003

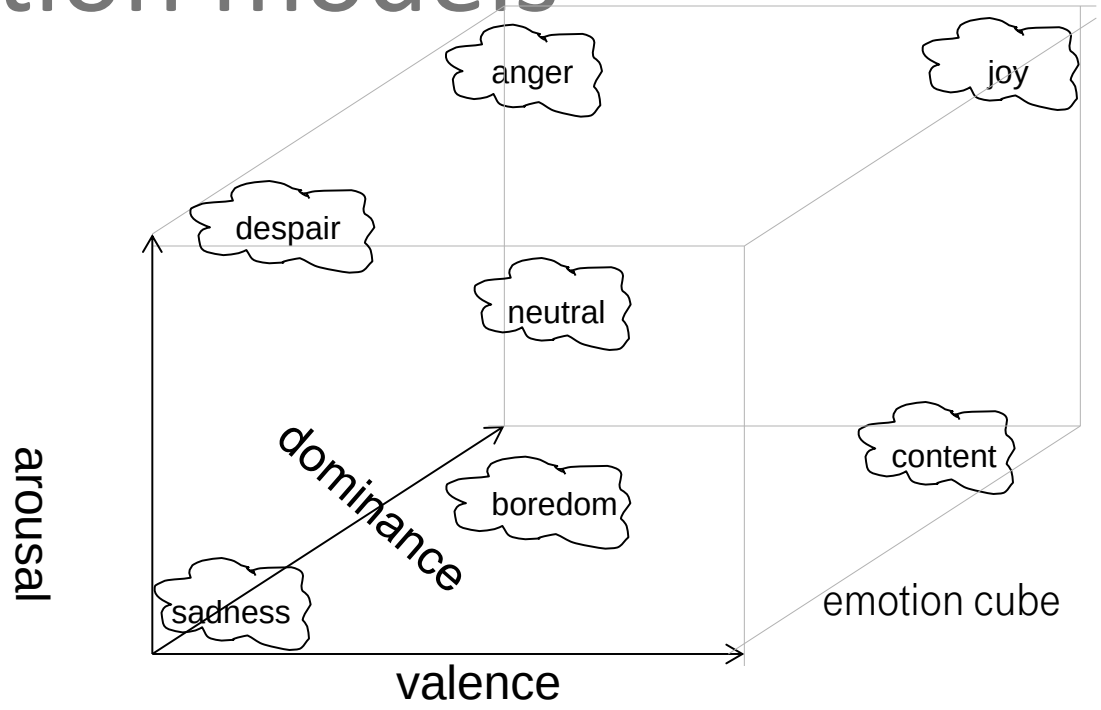
Open source feature extraction: OpenEar, Praat

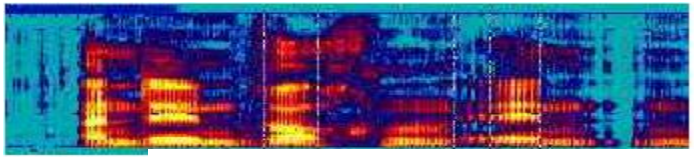
Speech Based Classification

constant features	Age	BC for ternary classification and multiword good quality input		Gender	BC for ternary classification and multiword good quality input	
origin related	Sociolect	BC for clearly socialily distinguished user groups (binary)		Dialect	E.g. bavarian. BC for binary classification (about 10% over chance level)	
language related	Language	WC for closely related languages, short utterances and bad audio		Accent	E.g. french accent. BC for binary classification of far-related languages.	
health related	Physical Health	BC for binary classification with clearly distinguished user-groups		Mental Health	E.g. depression recgnition. BC for clearly distinguishe d user groups.	
mood related	Short-term Condition	E.g. stress, tiredness. BC for binary classification and good quality multi-word input.		Mood Emotion	E.g. anger, joy. BC for binary classification and clearly expressed emotion (e.g. hot anger).	

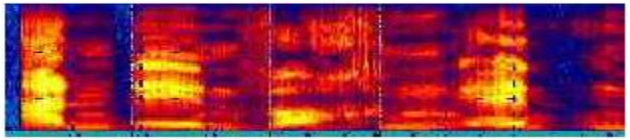
Emotion models

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

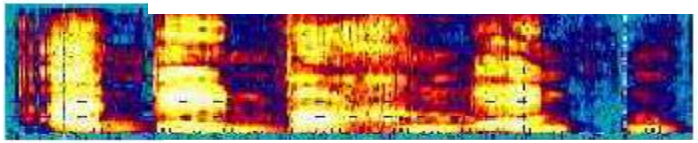




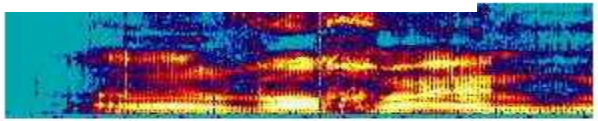
neutral



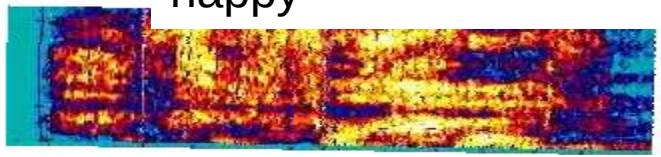
angry



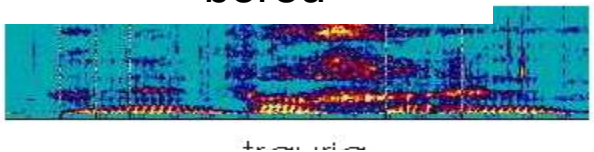
happy



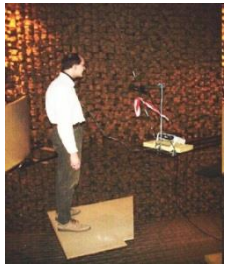
bored



frightened



sad



spectrograms from emotional acted speech

source: TUB emotional database

Applications

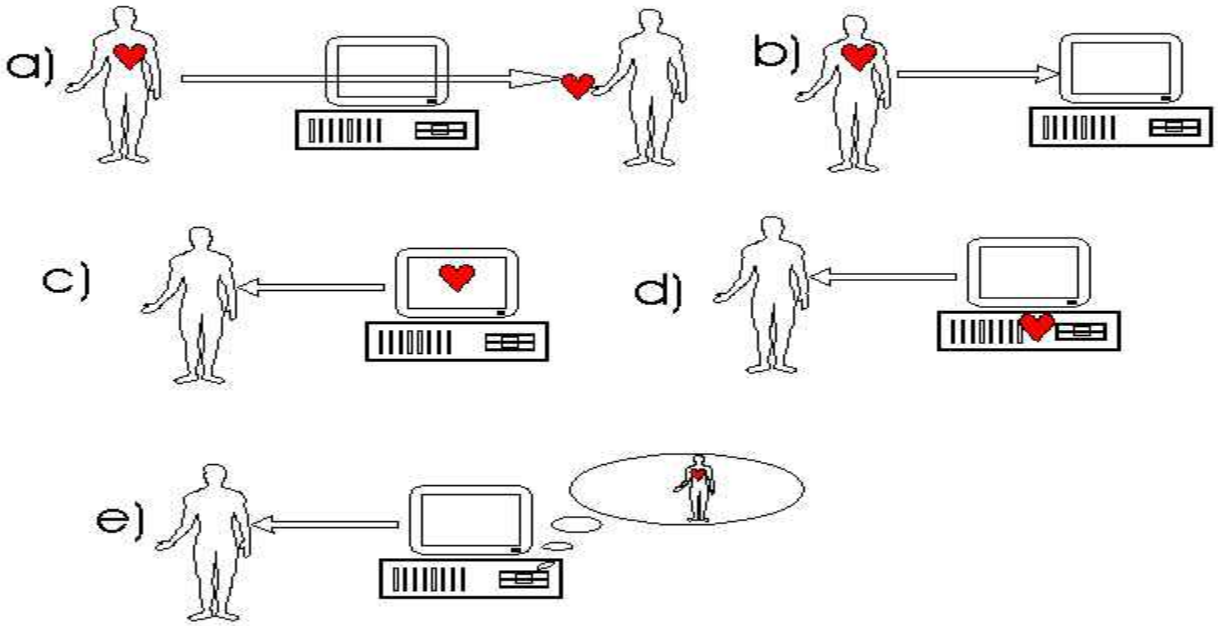
a) Mediated emotion

b) Affect recognition

c) Affect simulation

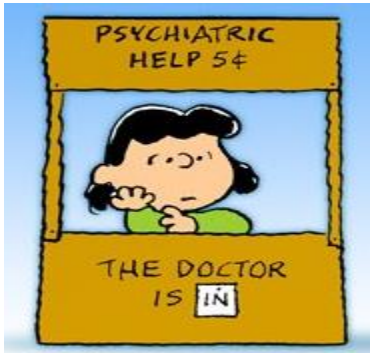
d) Modeling Emotional intelligence

e) Modeling human emotional behaviour



Applications speaker characteristics recognition

- Dispatching callers to trained agents
- Fun applications, e.g. “how enthusiastic do I sound”
- Call center quality management
- Match caller and agent
- Adapted dialog and/or persona design
- Target group specific advertising
- Market analysis of target groups
- Alert systems, i.e. analysis of urgency in speaker’s voice
- ….
- believable agents, artificial human



Applications speaker characteristics simulation

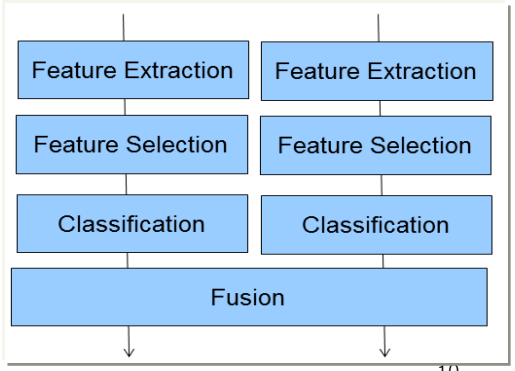
- Fun, e.g. emotional greetings
- Prosthesis
- Emotional chat avatars
- Gaming, believable characters
- Adapted dialog design
- Adapted persona design
- Target-group specific advertising
- …
- Believable agents, artificial humans

Classification: technical approach

- Basic approach:
  - extract features,
  - select best ones,
  - classify features,
  - fuse classifier outputs
- Principal Component Analysis (PCA) or Linear Discriminant Analysis (LDA) – preprocessing step to reduce feature vector dimension
- Gaussian Mixture Models: model training data as Gaussian densities
- Artificial Neural Networks (ANN), e.g. Multi Layer Perceptron.
- Support Vector Machines (SVM): use „kernel functions“ to separate non-linear decision boundaries
- Classification and Regression Trees (CART)
- Hidden Markov Models (HMMs) used to model temporal structure

- New Trend: Deep Neural Networks operate directly on base features

Open source classification: Weka



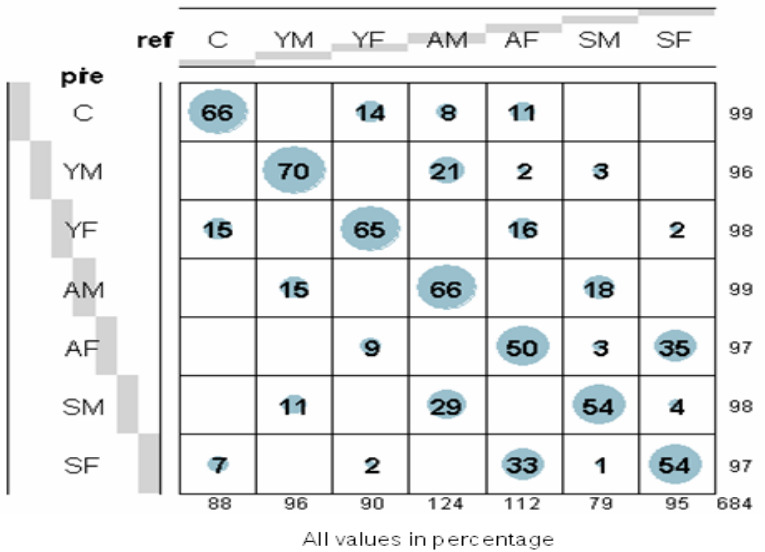
Evaluation: performance

- Performance depends strongly on application.
- Some rules of thumb:
  - Human performance gives a good estimate: machines show similar results.
  - Number of classes: Seperate Men from Women is easier than age in years.
  - Longer speech material helps: whole dialogs are easier than a “Hallo?” of one second duration.
  - The more the feature is expressed n the speech. The better the performance: foreign languages are easier than age.
- Examples:
  - Gender recognition with longer samples: nearly100%
  - Anger recognition with rare false alarm: about 40%

Evaluation example: Age and gender classification 2007

- Compared a parallel phone recognizer, derived from an automatic language identification system;
- a system using dynamic Bayesian networks to combine several prosodic features
- a system based solely on linear prediction analysis;
- Gaussian mixture models based on MFCCs for separate recognition of age and gender.

- Based on laboratory data: SpeechDat II, long sentences



Source: Metze et al 2007

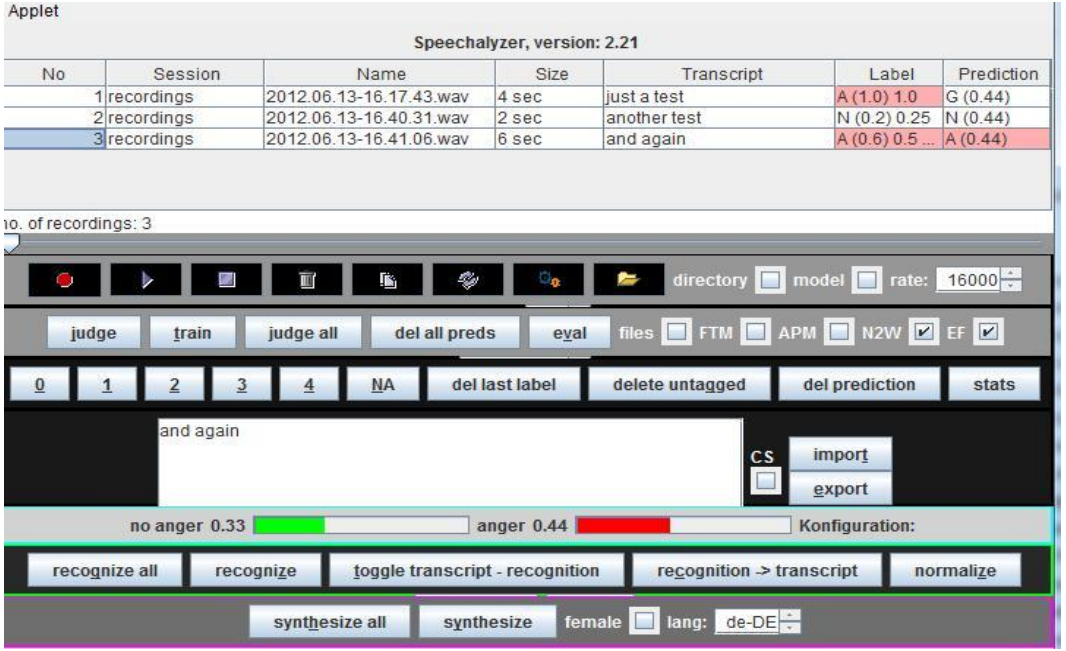
Outlook Speaker classification

- Challenges
  - Difficult, noisy data
    - Unclear what the target class is (emotion)
    - Feature not well presented in speech (age)
  - Adapting systems implicate intelligence
- Strategies
  - First step non-critical applications

- But
  - Humans do it, machines can do it

Tool: Speechalyzer

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



Thanks

Felix.Burkhardt@telekom.de