Engineering Paralinguistics: And Next ... the Transparent Speaker?

Björn Schuller

Technische Universität München

ZefiS - Zentrum für interdisziplinäre Sprachforschung, Bergische Universität Wuppertal

21 December 2011, 6:15 PM

Outline

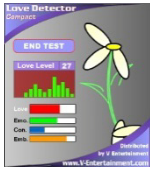
Introduction	Speech Processing
Computational Intelligence	Vision

Introduction

Computational Paralinguistics

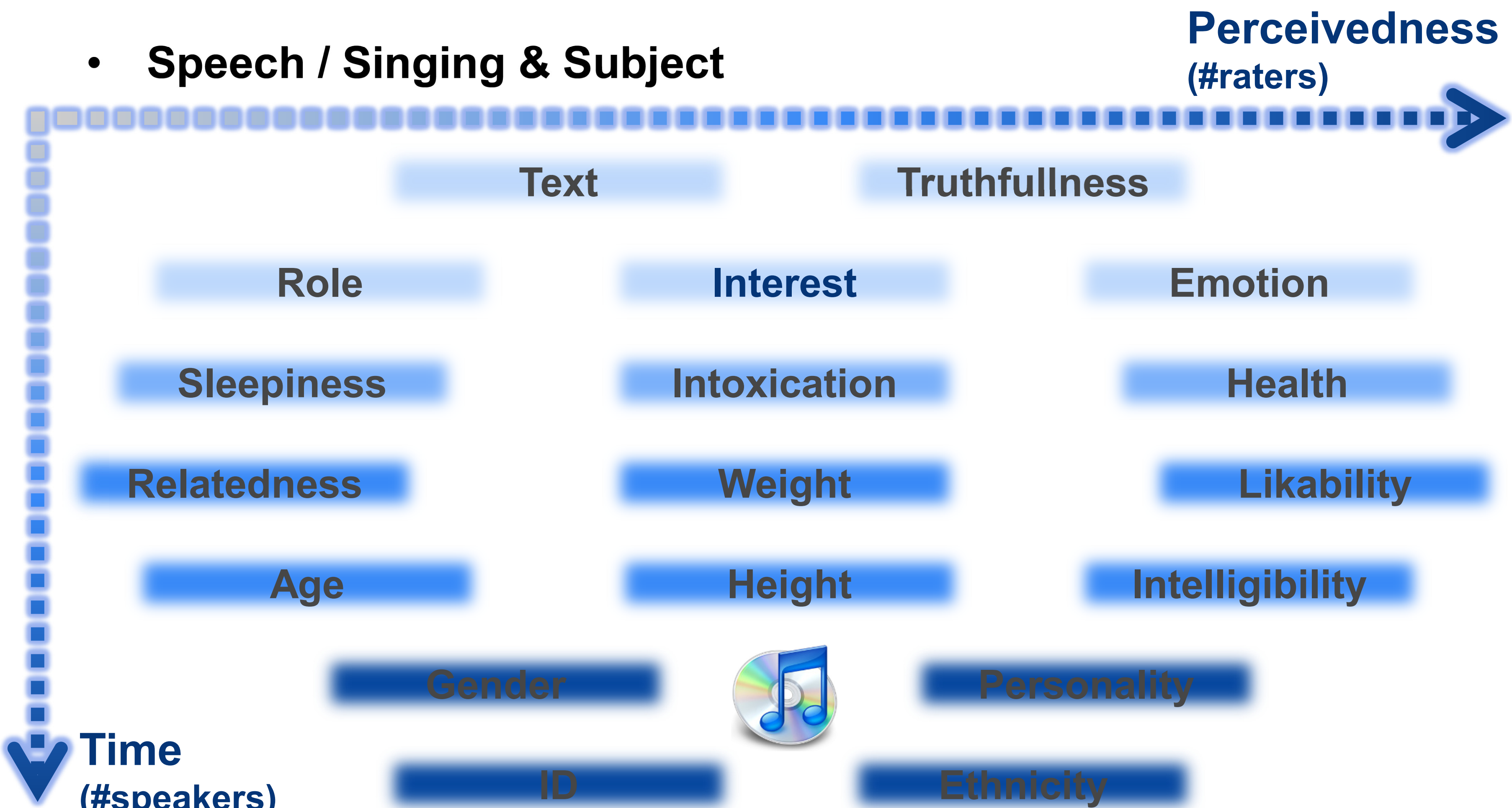
(Best Technical Demo IEEE ACII 2009)

- **Encoding**
  Semi-symbolic representation
- **Analysis / Editing / Synthesis**
  Voice conversion
- **Media Retrieval**
  Search by speaker attribute
- **Natural Interaction**
  Social competence
- **Monitoring**
  Threat detection, Customer monitoring
- **Voice Coaching**
  Interactive Emotion Games



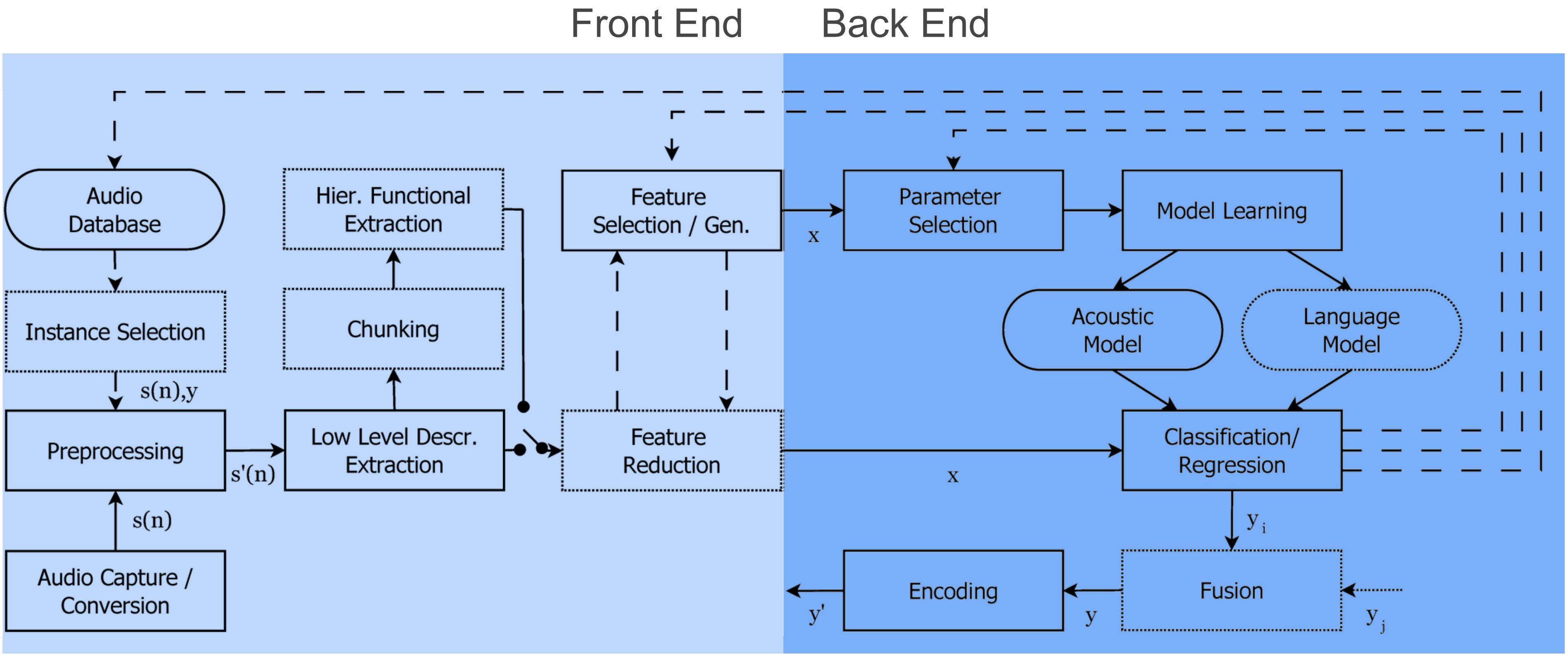
Speech / Singing

- **Speech / Singing & Subject**



Computational Paralinguistic Analysis

- **openEAR**



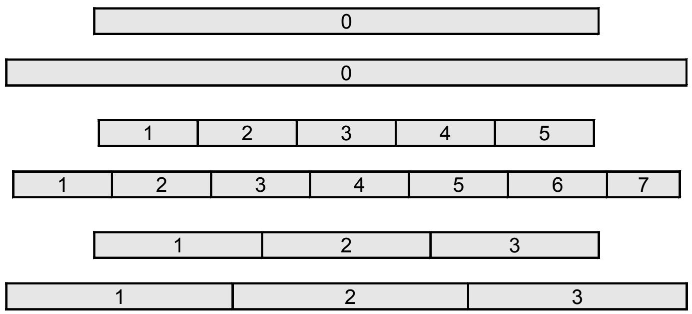
*"Intelligent Audio Analysis", **Springer** (to appear).*

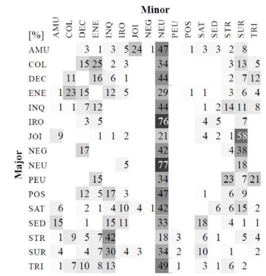
…In the Real Life

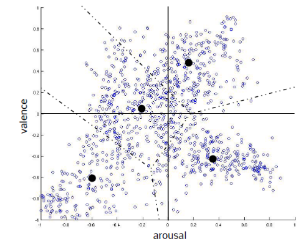
- **Data**
  - Monaural
  - Non-prototypical
  - Non-preselected

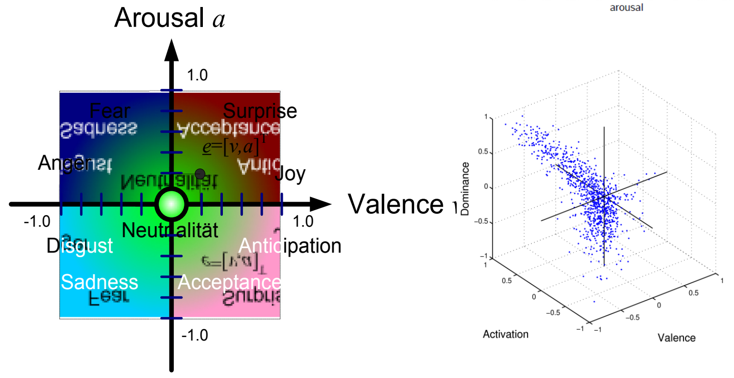
- **Processing**
  - Fully automatic chunking
  - Meta-data from web
  - No optimisation on test
  - Independence

- **Task Formulization**
  - Influence on Gold Standard
  - Self-learnt?



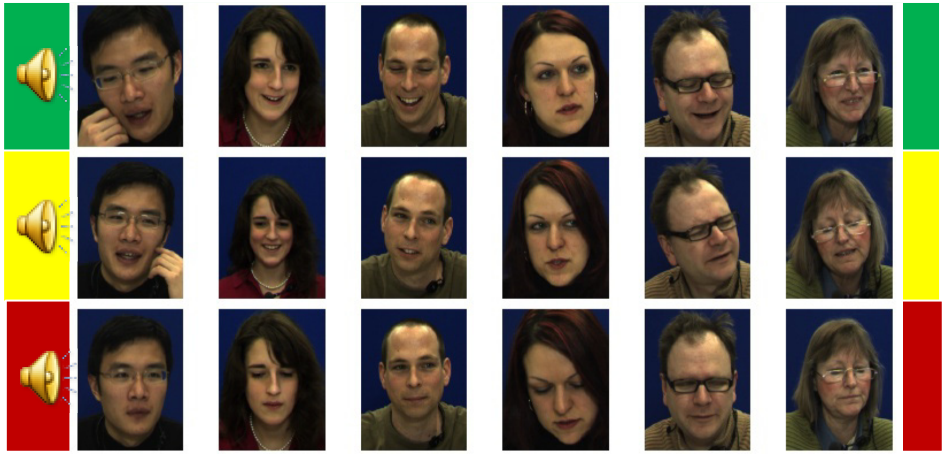




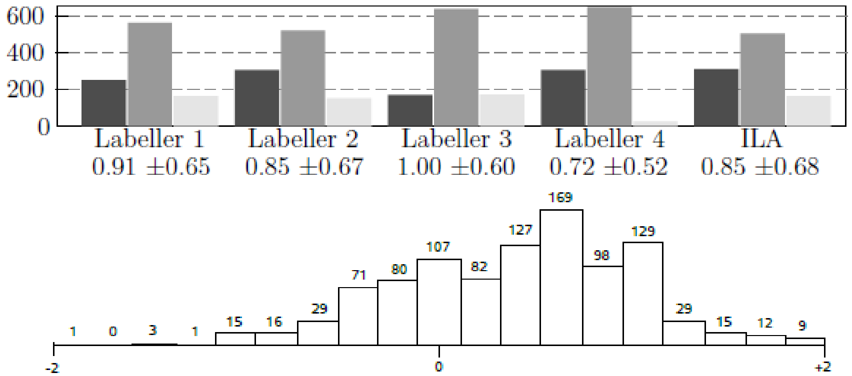


Speech Data

- **TUM AVIC**
  - Conversational speech
  - 21 subjects, 11,414 turns
- **Annotation**
  - Text, non-linguistic vocalizations
  - Neutrality / Interest / Curiosity
  - 4 annotators



K	L1	L2	L3	L4
Labeller 1	1.00	0.86	0.62	0.61
Labeller 2		1.00	0.72	0.71
Labeller 3			1.00	0.44
Interlabel	**0.89**	**0.97**	**0.75**	**0.74**



"Being Bored? Recognising Natural Interest by Extensive Audiovisual Integration for Real-Life Application", ***Image and Vision Computing***, *27(12): 1760-1774, 2009.*

Speech Data

- **Speech In Minimal Invasive Surgery**
- **Collection**
  29 operations 37.4 h,
  Segmentation: 16% speech
- **Annotation**
  Emotion
  Text
  Noise by type

  4 Annotators

Emotion	[m:s]	#Turns	[%]
Neutral	235:49	6189	67.4
Joy	34:20	894	9.8
Anger	22:28	539	6.4
Impatience	29:26	856	8.4
Confusion	27:58	818	8.0
Total	350:01	9,299	100

Speech Data

- **Community Based Labelling**

  Amazon Mechanical Turk?

- **Data Synthesis**

  Example: Emotion in Speech

  Cross-corpus testing, 3 levels of valence, 6 databases

  Training with real speech / synthesized speech

*“Learning with Synthesized Speech for Automatic Emotion Recognition”, **ICASSP**, 2010.*
(Pending European Patent)

Test with real speech

Train	% WA
Human	64.8
Synth.	**75.4**
Human + Synth.	79.5

Speech Data

- **Data Pooling & Unsupervised Learning**
  - Example: Emotion Recognition
  - 6 databases (ABC, AVIC, DES, eNTERFACE, SAL, VAM):
    - 8k sounds, 6h speech 7 classes
  - Leave-One-Corpus-Out, pooling of data, binary arousal / valence
  - Unsupervised learning
  - Significance: $p < 0.001$

Train	% UA Arousal	% UA Valence
Labelled 3	62.6	55.6
Labelled 3 + Unlabelled 2	**63.2**	**57.1**
Labelled 5	63.9	58.4

*“Unsupervised Learning in Cross-Corpus Acoustic Emotion Recognition”, **IEEE ASRU**, 2011.*

Speech Data

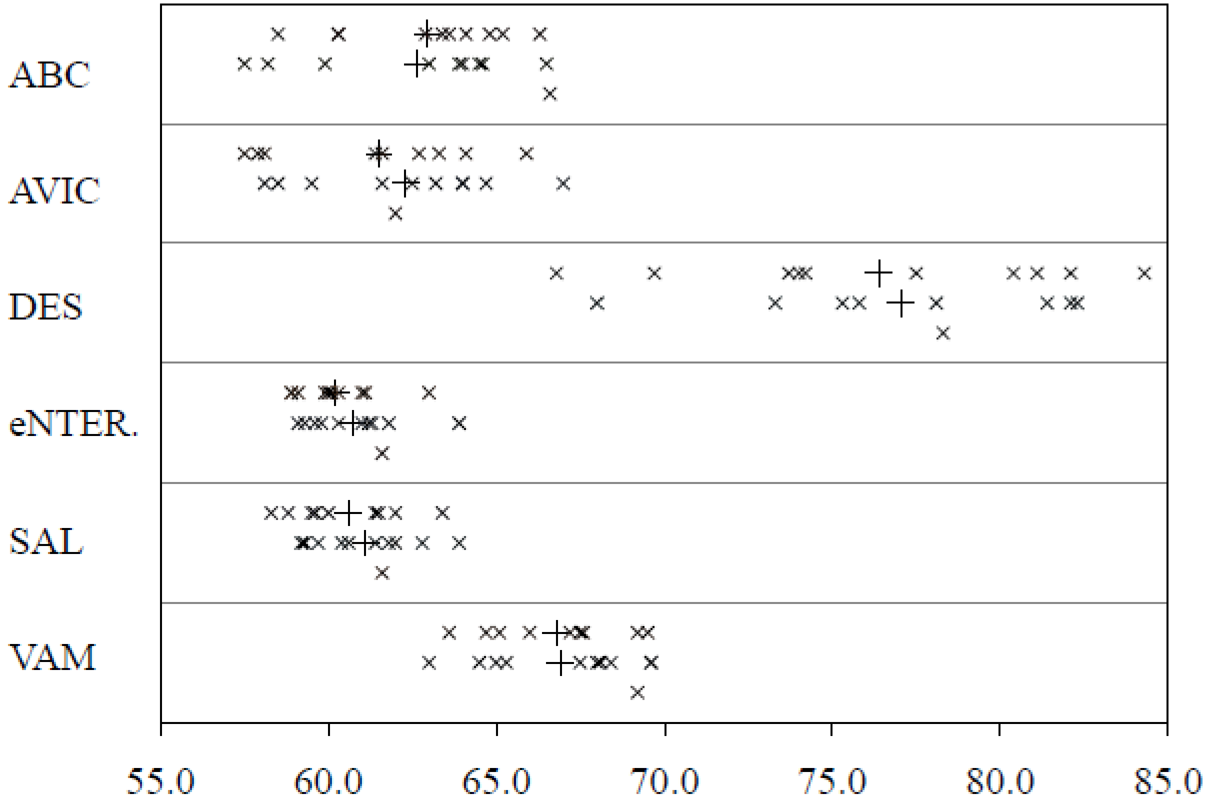
- **Data Pooling & Unsupervised Learning**

Arousal

Labelled 3

+ Unlabelled 2

Labelled 5



Speech Data

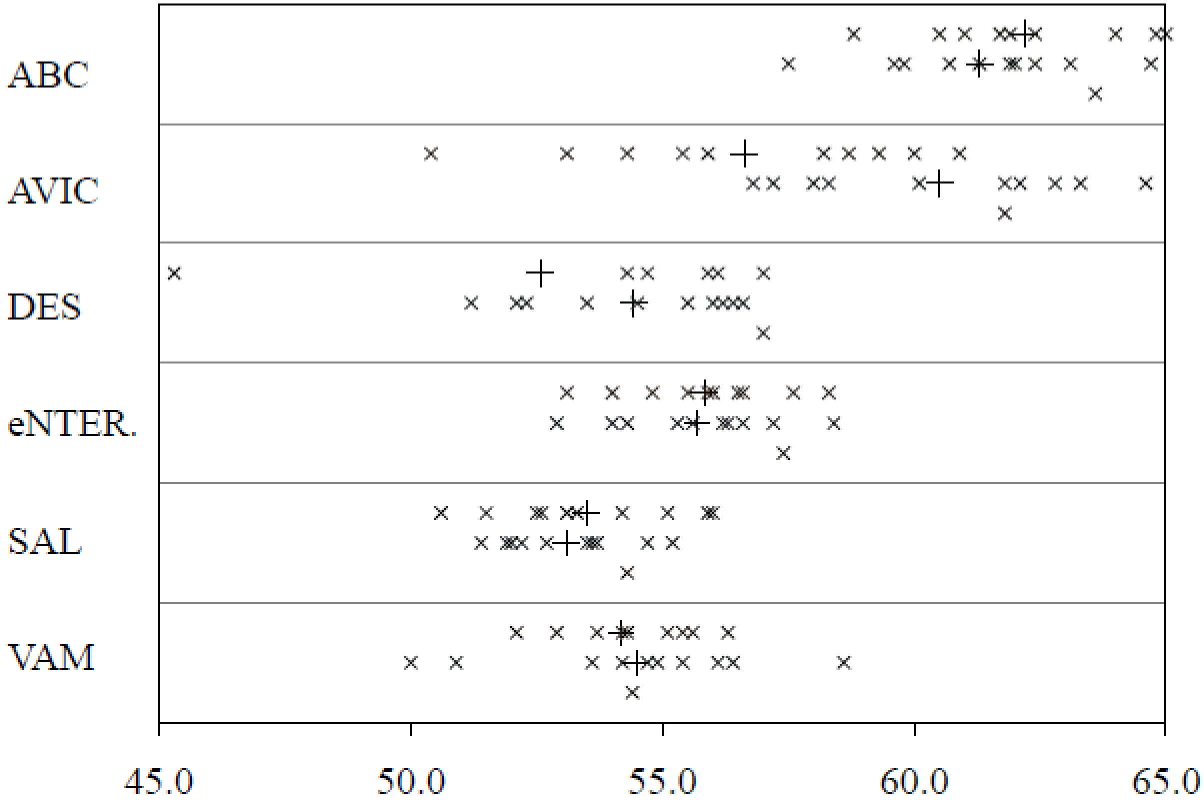
- **Data Pooling & Unsupervised Learning**

Valence

Labelled 3

+ Unlabelled 2

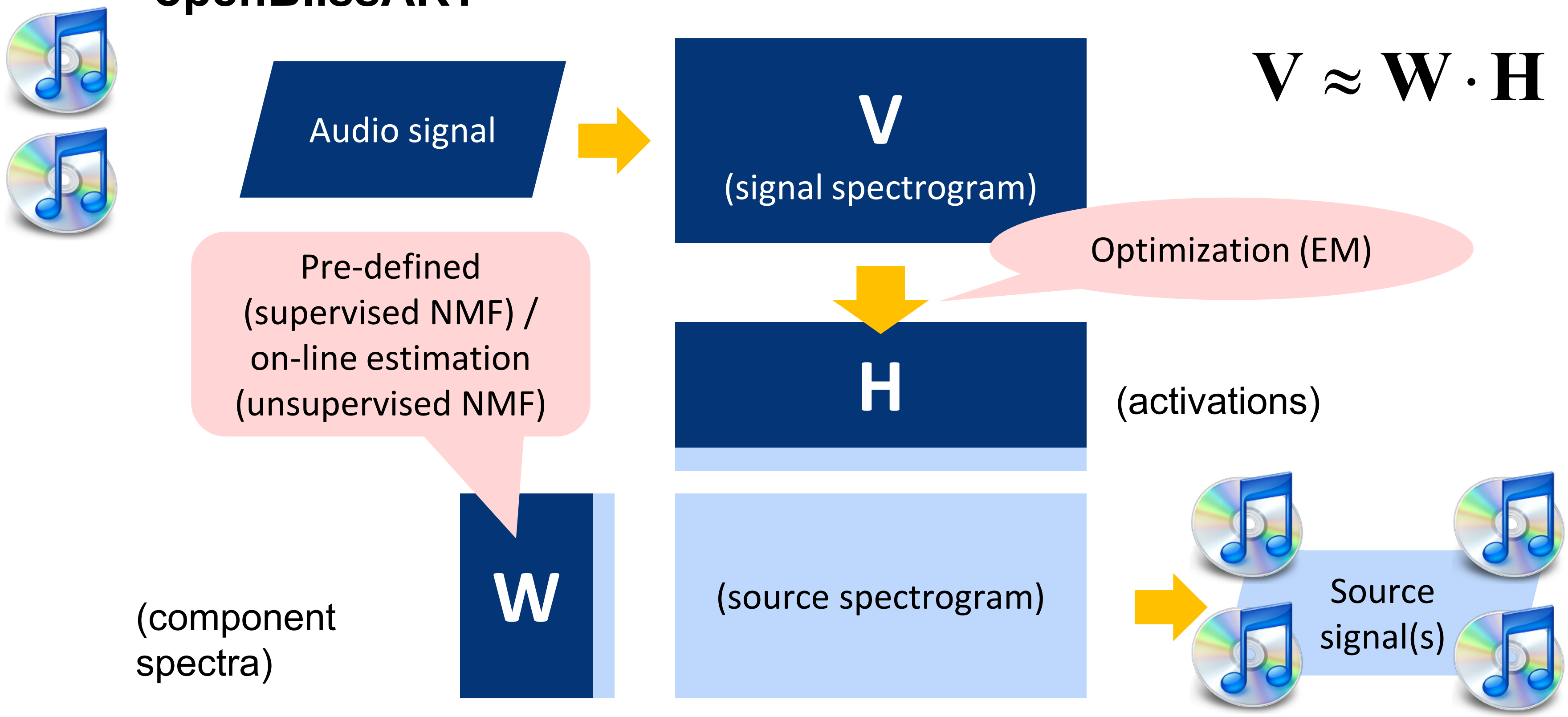
Labelled 5



Speech Processing – The Front End

Audio Source Separation

- **openBlissART**



"openBliSSART: Design and Evaluation of a Research Toolkit for Blind Source Separation in Audio Recognition Tasks", **ICASSP**, 2011.

Audio Editing

- **Audio Editing (adMIRe)**

  Separation, Chorus, Chords, Key, Onsets, Down-Beats, Key-Shift, Stretch

Canon in D (Johann Pachelbel, English Chamber Orchestra – Raymond Leppard) *– D major, 87.5bpm*

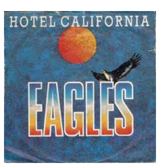
D A Bm F#m G D G A

Original Clicks

Basket Case (Green Day, Billie Joe Armstrong (Vocals)) *- Eb major, 84.9bpm*

Eb Bb Cm G Ab Eb Bb

Original **Vocals** Rest Clicks

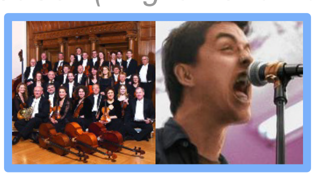


Hotel California (Eagles, Don Henley (Drums)) *- B minor, 150.3bpm*

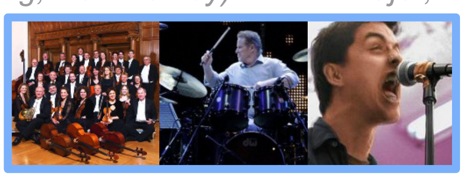
Original **Drums** Rest Clicks

"The Canon Hotel Case" *(English Chamber Orchestra, Billie Joe Armstrong, Don Henley)* *- D major, 120 bpm*

Mix 1



Mix 2



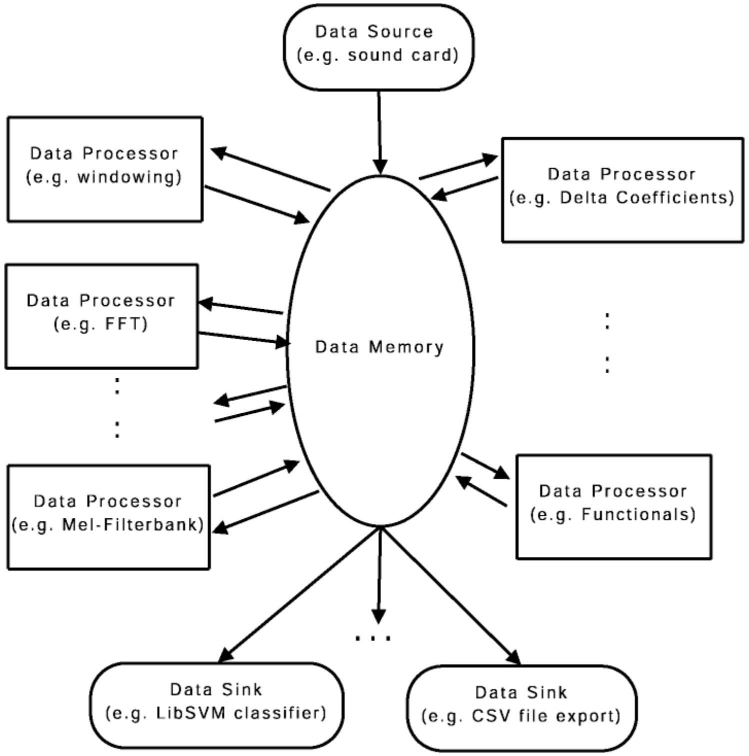
Speech Features

- **openSMILE**

  Speech & Music Interpreation
  by Large Space Extraction

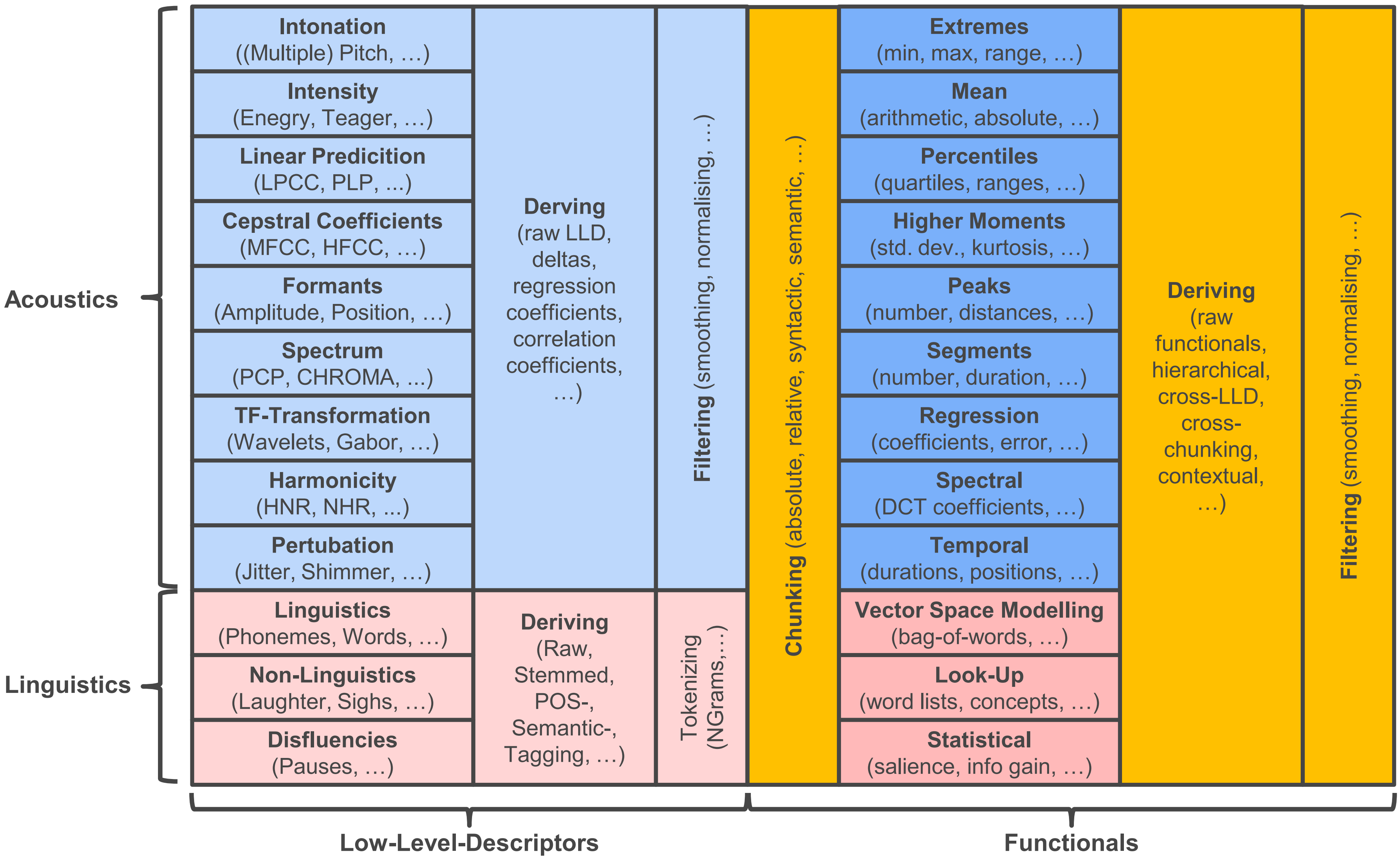
  Low-Level-Descriptors
  (Hierarchical) Functionals
  Standard feature sets

  Multithreading
  Memory efficient
  Fully configurable



#features	RTF
10k	.02
500k	**.03**

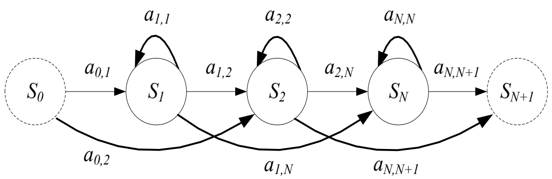
"openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor", **ACM Multimedia**, 2010.
(3rd place ACM MM Open Source Software Competition)


Acoustics
Linguistics
Intonation ((Multiple) Pitch, …)
Intensity (Enegry, Teager, …)
Linear Predicition (LPCC, PLP, ...)
Cepstral Coefficients (MFCC, HFCC, …)
Formants (Amplitude, Position, …)
Spectrum (PCP, CHROMA, ...)
TF-Transformation (Wavelets, Gabor, …)
Harmonicity (HNR, NHR, ...)
Pertubation (Jitter, Shimmer, …)
Linguistics (Phonemes, Words, …)
Non-Linguistics (Laughter, Sighs, …)
Disfluencies (Pauses, …)
Derving (raw LLD, deltas, regression coefficients, correlation coefficients, …)
Deriving (Raw, Stemmed, POS-, Semantic-, Tagging, …)
Filtering (smoothing, normalising, …)
Tokenizing (NGrams,…)
Chunking (absolute, relative, syntactic, semantic, …)
Extremes (min, max, range, …)
Mean (arithmetic, absolute, …)
Percentiles (quartiles, ranges, …)
Higher Moments (std. dev., kurtosis, …)
Peaks (number, distances, …)
Segments (number, duration, …)
Regression (coefficients, error, …)
Spectral (DCT coefficients, …)
Temporal (durations, positions, …)
Vector Space Modelling (bag-of-words, …)
Look-Up (word lists, concepts, …)
Statistical (salience, info gain, …)
Deriving (raw functionals, hierarchical, cross-LLD, cross-chunking, contextual, …)
Filtering (smoothing, normalising, …)
Low-Level-Descriptors
Functionals

Intelligence – The Back End

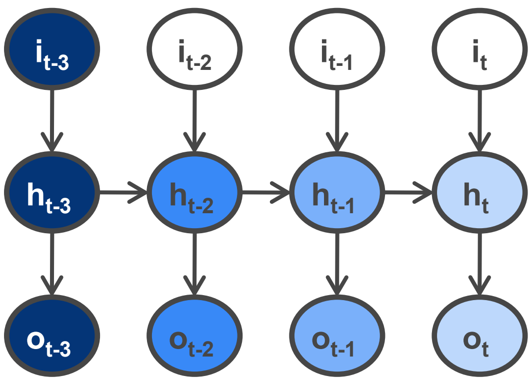
Computational Intelligence

- **Sequence Learning**
  Audio is sequential

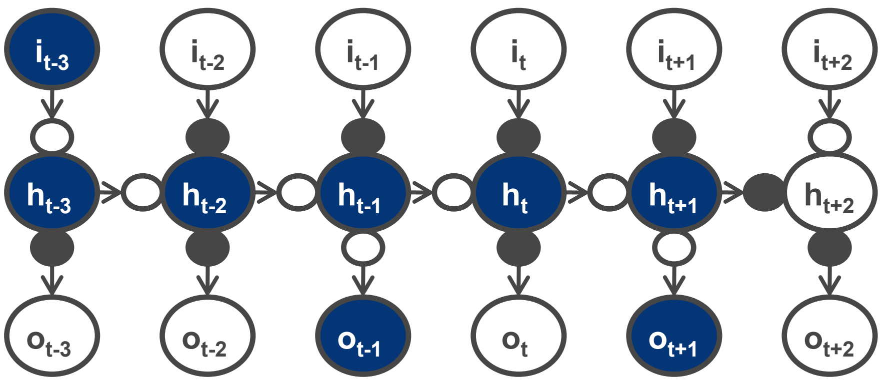


- **Vanishing Gradient Problem**

Recurrent Neuronal Network



Long Short-Term Memory RNN



*"Combining Long Short-Term Memory and Dynamic Bayesian Networks for Incremental Emotion-Sensitive Artificial Listening", **IEEE Journal of Selected Topics in Signal Processing**, 4(5): 867-881, 2010.*

*"Tandem Decoding of Children's Speech for Keyword Detection in a Child-Robot Interaction Scenario", **ACM Transactions on Speech and Language Processing**, 7(4), 2011.*

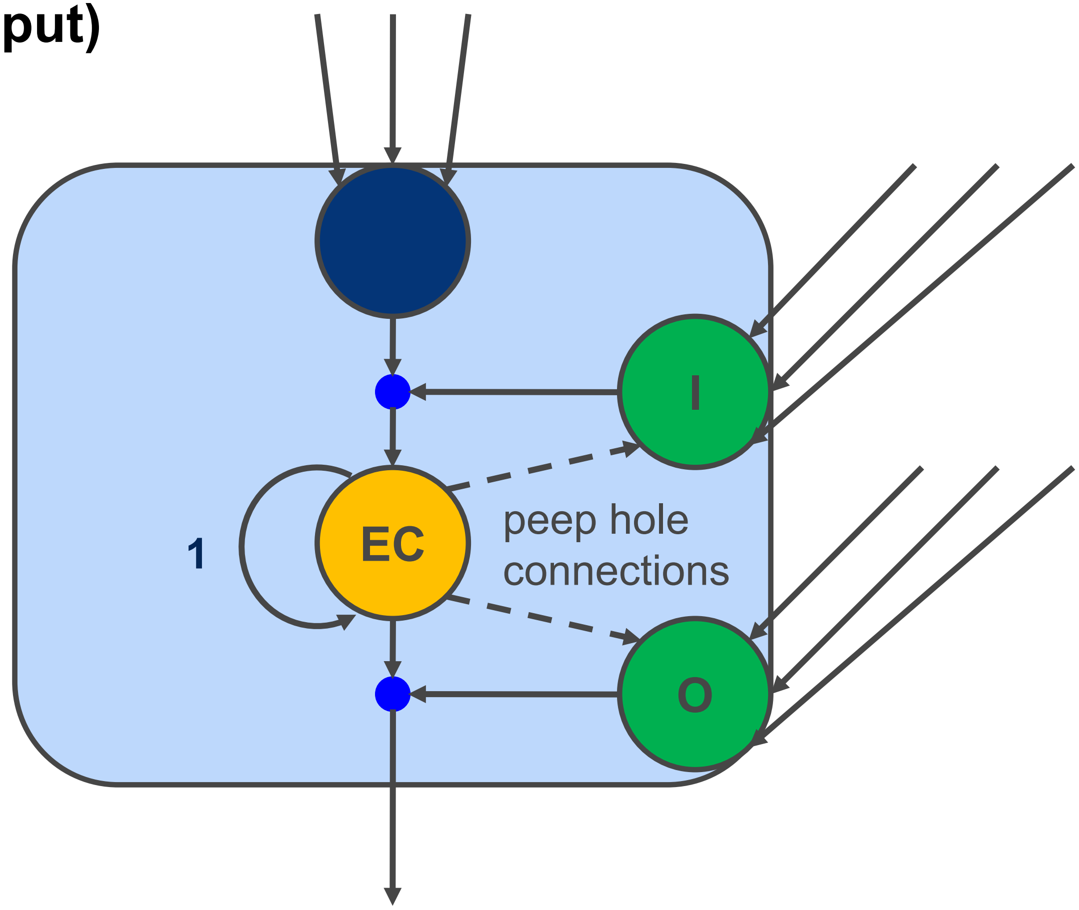
Long Short-Term Memory

- **Original Cell (Input, Output)**

Linear unit
Auto-weight 1
"error carousel"

Non-linear gate
Input / output

Multiplicative
opening or shut-down



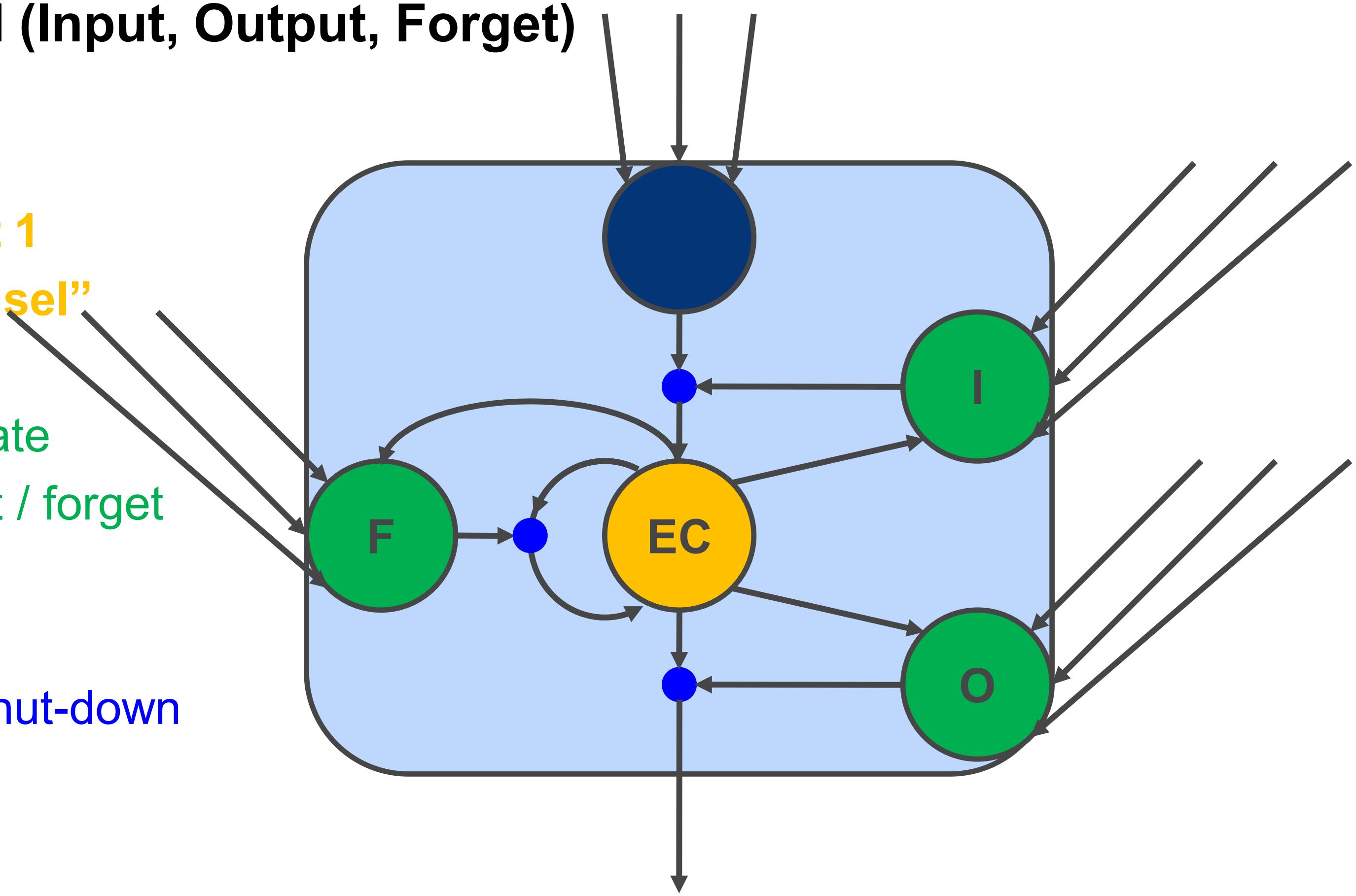
Long Short-Term Memory

- **Current Cell (Input, Output, Forget)**

Linear unit
Auto-weight 1
“error carousel”

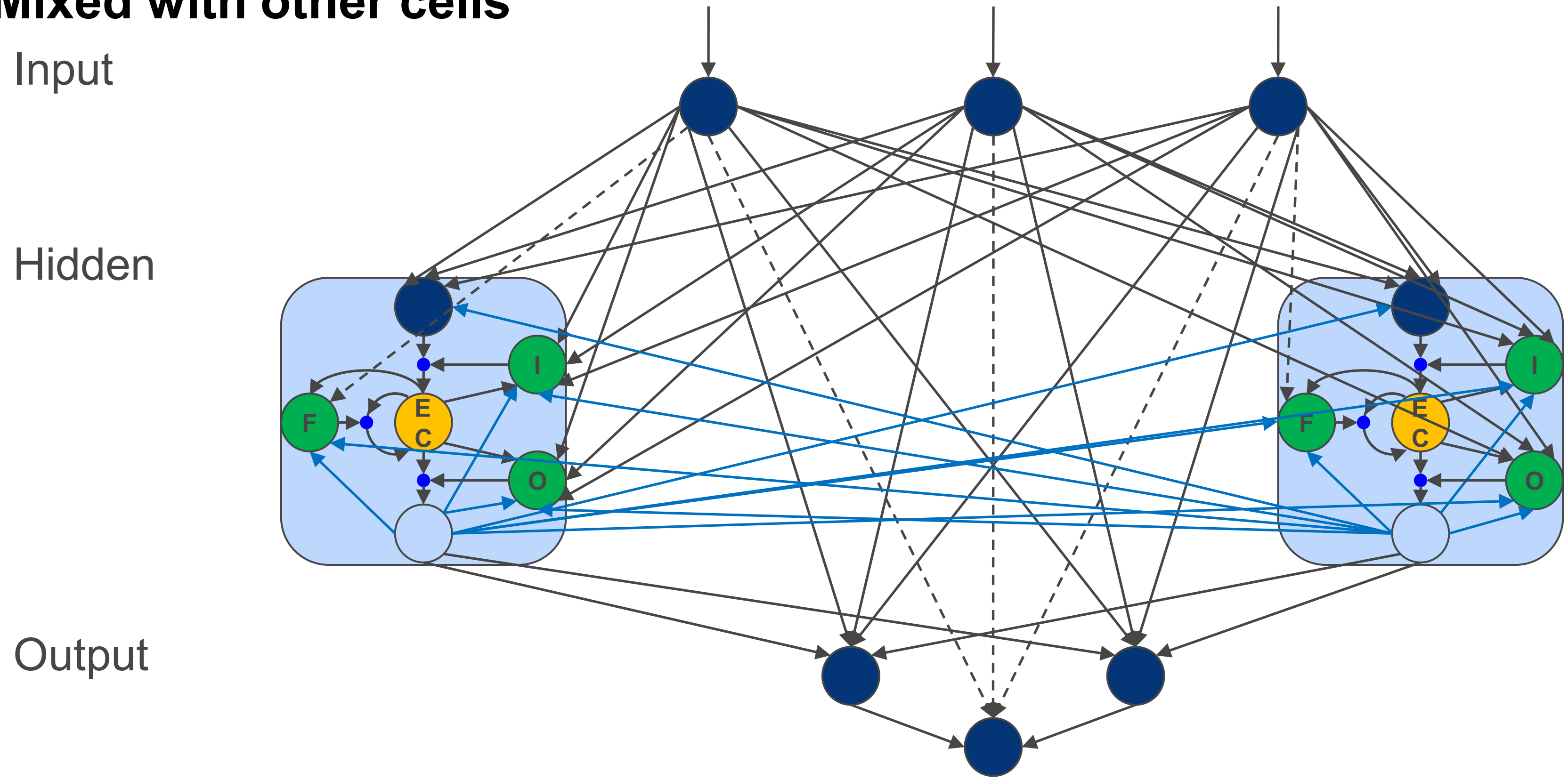
Non-linear gate
Input / output / forget

Multiplicative
opening or shut-down



Long Short-Term Memory

- **Mixed with other cells**

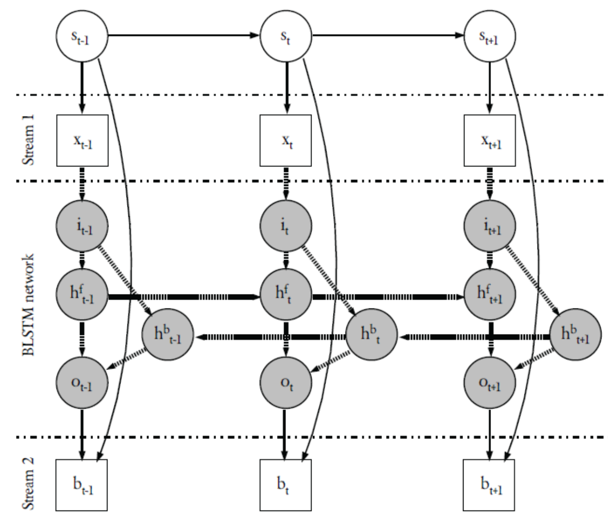


Keywords

- **Example: CHiME Challenge 2011**
  - Grid corpus (voice commands)
  - Add. noise, reverberation, home environment
  - Convolutive NMF (openBliSSART)
  - BLSTM-RNN (openSMILE), Multistream HMM

-6 dB
0 dB
6 dB

$$V \otimes \left(\frac{W_{sp} H_{sp}}{WH} \right)$$



% WA	Base	NMF
CHiME Keywords	55.9	**91.9**

"The Munich 2011 CHiME Challenge Contribution: NMF-BLSTM Speech Enhancement and Recognition for Reverberated Multisource Environments", **CHiME**, 2011.

ASR of Emotional Speech

- **Example: FAU Aibo**

  MFCC, polyphones, SC-HMM, full covariances

  Back-off bigrams

  Testing: E > A > N > M

  Training (AM): N > E > A > M

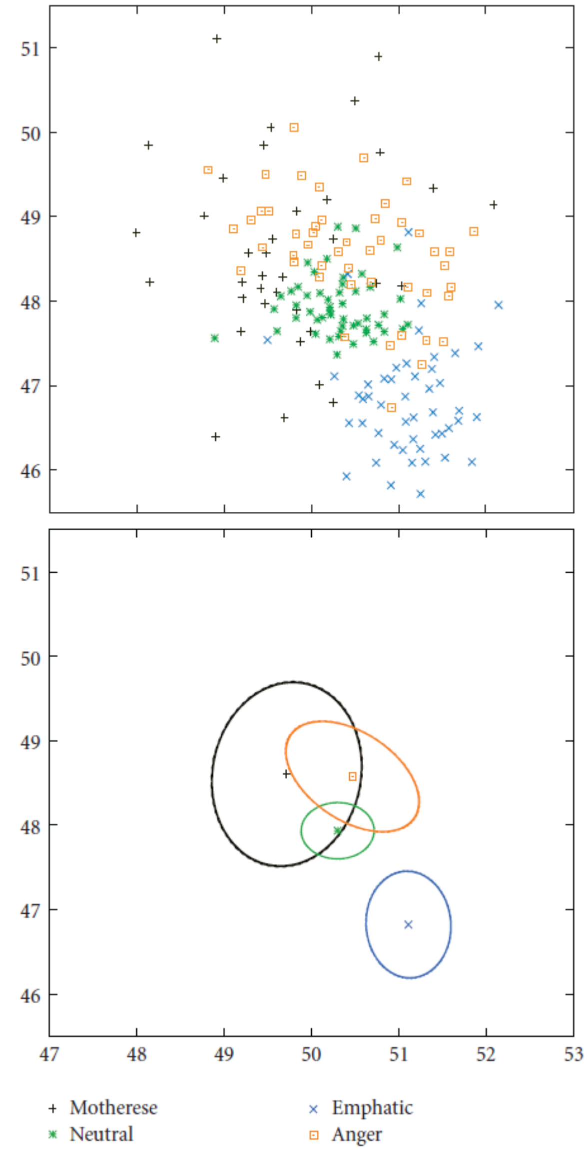
- **Explanation**

  Sammon transformation:

  High dispersion, neutral in the center

  Neutral words per turn

Mother.	Neutral	Emphat.	Anger
44.2%	94.4%	56.7%	29.7%



ASR of Emotional Speech

- **Adapting ASR Models**

  AM, LM, both

  Word accuracy

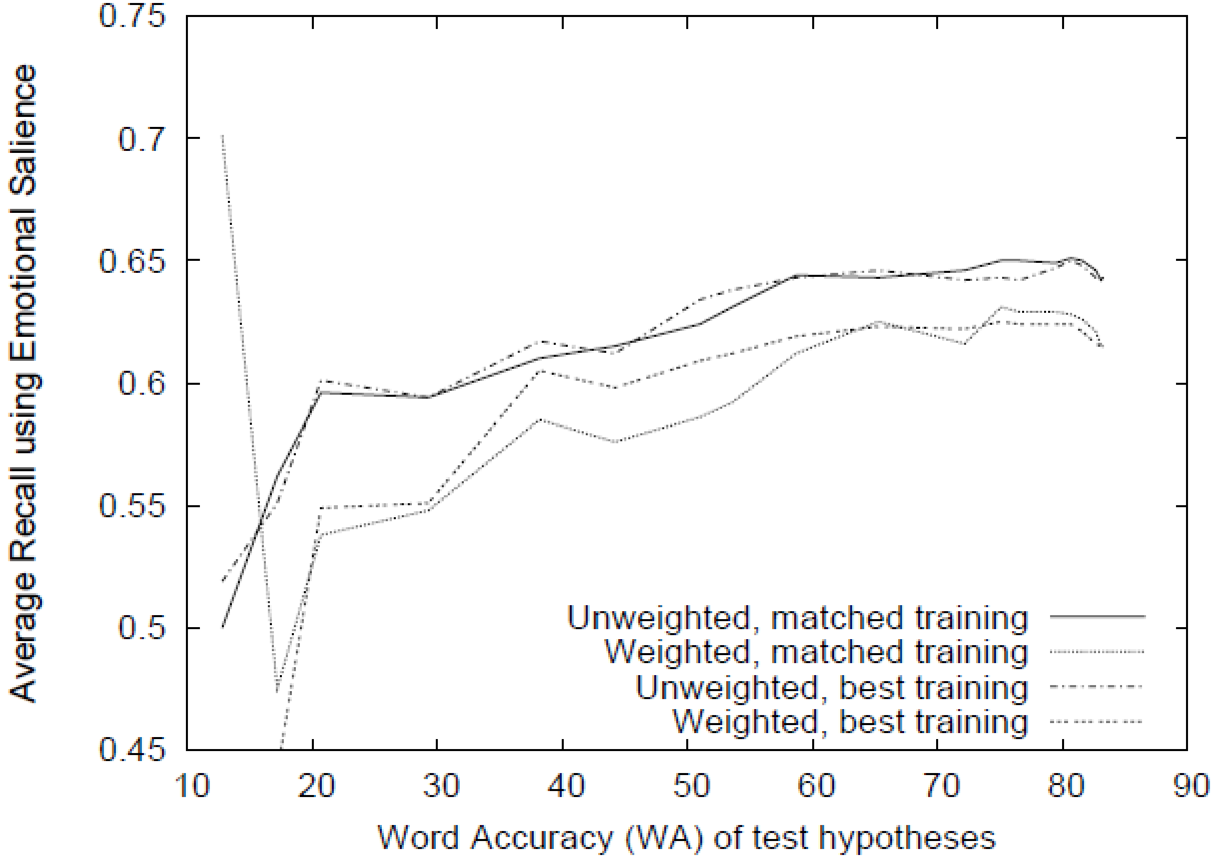
  Significance

	M	E	A
Baseline system	65.0	81.0	79.2
Adapted systems			
Acoustic models	64.5 ○ ○ ○ ○ ○	83.1 • ○ ○ ○ ○	83.6 • • • • •
Linguistic models	65.9 ○ ○ ○ ○ ○	81.6 ○ ○ ○ ○ ○	81.6 • • • • •
Both	65.9 ○ ○ ○ ○ ○	84.4 • • • • •	85.1 • • • • •

*"On the Impact of Children's Emotional Speech on Acoustic and Language Models", **EURASIP Journal on Audio Speech and Music Processing**, 2010.*

ASR and AER

- **ASR Influence**
  Salience
  Emotion Challenge
  2-class Task



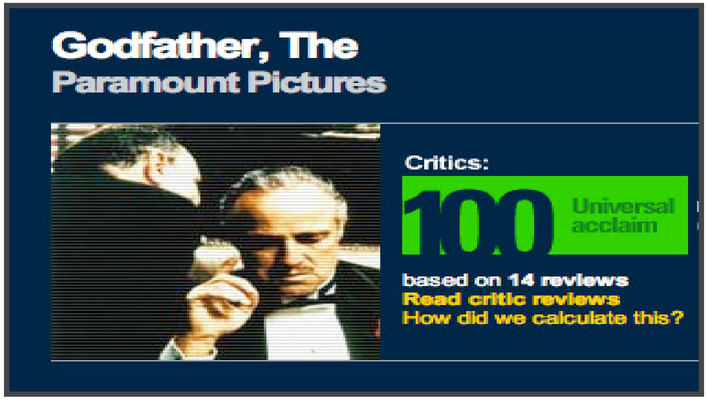
*"Emotion Recognition using Imperfect Speech Recognition", **Interspeech**, 2010.*

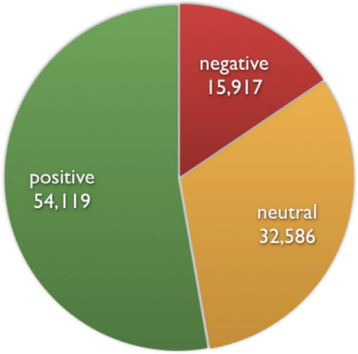
Sentiment

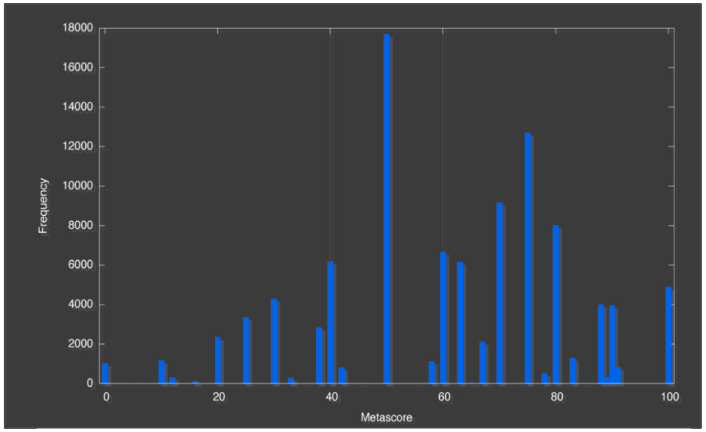
- **2-class Valence of Movie Critic (Metacritic Corpus)**

  4,901 movies, over 100 k reviews

Meaning	Score	Color	Reviews
Universal Acclaim	81 - 100	green	15 353
Generally Favorable	61 - 80	green	38 766
Mixed or Average	40 - 60	yellow	32 586
Generally Unfavorable	20 - 39	red	13 194
Overwhelming Dislike	0 - 19	red	2 723







Sentiment

- **2-class Valence of Movie Critic (Metacritic Corpus)**

Sentiment

- **2-class Valence of Movie Critic (Metacritic Corpus)**

  Bag-of-NGrams

g_{min}	g_{max}	**Accuracy**
1	1	75.61%
1	2	76.76%
1	**3**	**77.33%**
1	4	76.46%
1	5	76.91%
2	2	69.43%
2	3	70.65%
2	4	71.16%
2	5	72.45%
3	3	70.92%
3	4	71.23%
3	5	71.32%

Sentiment

- **2-class Valence of Movie Critic (Metacritic Corpus)**

  Bag-of-NGrams

Transformation	**Accuracy**
f_{ij}	76.53%
norm(f_{ij})	76.63%
TF	76.90%
norm(TF)	76.85%
IDF	76.53%
norm(IDF)	77.16%
TFIDF	76.89%
norm(TFIDF)	**77.33%**

*"Learning and Knowledge-based Sentiment Analysis in Movie Review Key Excerpts", **Springer LNCS**, 6456: 448-472, 2011.*

Sentiment

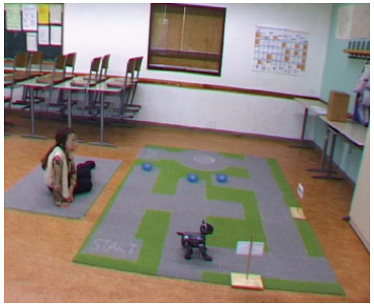
- **2-class Valence of Movie Critic (Metacritic Corpus)**
  Bag-of-Ngrams vs. On-Line Knowledge Source
  General Inquirer, ConceptNet, WordNet

% UA	Learnt	OKS
2-calss positive / negative	**77.33**	**68.61**
Recall positive	77.00	75.61
Recall negative	78.41	45.46

*"Learning and Knowledge-based Sentiment Analysis in Movie Review Key Excerpts", **Springer LNCS**, 6456: 448-472, 2011.*

Emotion

- **INTERSPEECH 2009 Emotion Challenge**
  FAU AIBO: 51 children, 9h speech, 18k turns
  .4k openSMILE features, SVM



% UA	Base	Vote
5-class: Anger, Emphatic, Neutral, Pos., Rest	38.2	**44.0**
2-class: Negative, Idle	67.7	**71.2**

"Recognising Realistic Emotions and Affect in Speech: State of the Art and Lessons Learnt from the First Challenge",
***Speech Communication**, 53(9/10): 1062-1087, 2011.*

Age, Gender, Interest

- **INTERSPEECH 2010 Paralinguistic Challenge**
  aGender: 954 speakers, 47h speech, 65k turns
  TUM AVIC: 21 speakers, 2h speech, 4k turns
  1.6k openSMILE features, SVM / RSS-REP

% UA	**Base**	**Vote**
4-class: Child, Youth, Adult, Senior	48.9	**53.6**
3-class: Child, Female, Male	81.2	**85.7**

CC	**Base**
Level of Interest [-1,1]	.421

"Paralinguistics in Speech and Language - State-of-the-Art and the Challenge",
Computer, Speech, and Language *(to appear).*

Intoxication & Sleepiness

- **INTERSPEECH 2011 Speaker State Challenge**
  ALC: 154 speakers, 39h speech, 12k turns
  SLC: 99 speakers, 21h speech, 9k turns
  4k openSMILE features, SVM

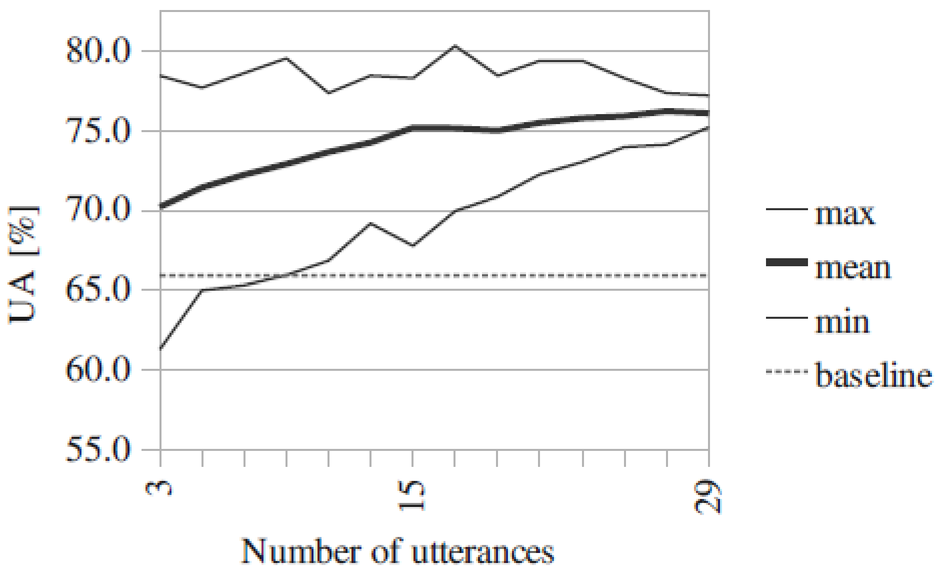
% UA	**Base**	**Vote**
2-class: above/below 0.5 per mill	65.9	**72.2**

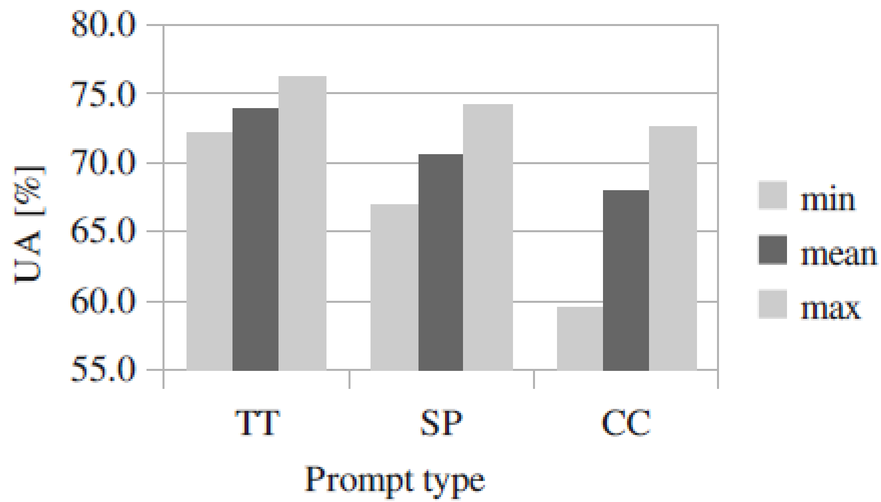
% UA	**Base**	**Vote**
2-class: above/below 7.5 Karolinska SS	70.3	**72.5**

*"The INTERSPEECH 2011 Speaker State Challenge", **Interspeech**, 2011.*

Intoxication & Sleepiness

- **INTERSPEECH 2011 Speaker State Challenge**
  Intoxication: Using several speech turns (left)
  Focusing on Tongue Twisters (TT), Spontaneous (SP), or C&C speech





*"Fusing Utterance-Level Classifiers for Robust Intoxication Recognition from Speech", **ACM ICMI**, 2011.*

Personality, Likability, Pathology

- **INTERSPEECH 2012 Speaker Trait Challenge**

  SPC: speaker, 2h speech, .6k turns

  SLD: 800 speakers, 2h speech, .8k turns

  NCSC: 55 speakers, 2h speech, 2.4k turns

  6k openSMILE features, Random Forests / (SVM)

  *priliminary

% UA	**Base**
2-class: above/below mean openness	57.0*
2-class: above/below mean conscientiousness	79.6*
2-class: above/below mean extraversion	75.8*
2-class: above/below mean agreeableness	56.1*
2-class: above/below mean neuroticism	68.2*
Mean	**67.3***

*"The INTERSPEECH 2012 Speaker Trait Challenge", **Interspeech**, 2012.*

Personality, Likability, Pathology

- **INTERSPEECH 2012 Speaker Trait Challenge**

  SPC: 330 speakers, 2h speech, .6k turns

  SLD: 800 speakers, 1h speech, .8k turns

  NCSC: 55 speakers, 3h speech, 2.4k turns

  6k openSMILE features, Random Forests / (SVM)

  *priliminary

% UA	**Base**
2-class: above/below mean likability	67.6*

% UA	**Base**
2-class: above/below mean intelligibility	66.7*

*"'Would You Buy A Car From Me?' – On the Likability of Telephone Voices", **Interspeech**, 2011.*

*"The INTERSPEECH 2012 Speaker Trait Challenge", **Interspeech**, 2012.*

Height

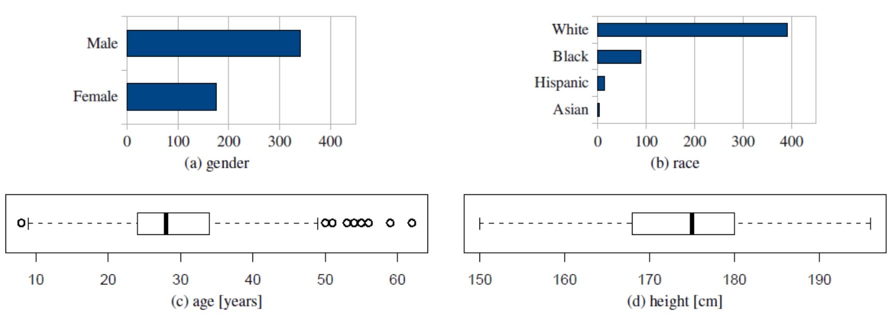
- **TIMIT Age, Gender, Height (Dialect, Education, Race)**
  TIMIT corpus: 630 speakers, 6k turns
  1.6k openSMILE features, SVR

Task	Context	CC	MAE [years/cm]
Height	–	0.2956	7.05
	R	0.2861	7.09
	G	0.2992	7.01
	A	0.3139	6.94
	A,G	**0.3171**	**6.91**
	A,R	0.3023	7.00
	G,R	0.2904	7.05
	A,G,R	0.3035	6.98
	All	0.3063	7.07

*"Semantic Speech Tagging: Towards Combined Analysis of Speaker Traits", **AES**, 2011.*

Singer Traits

- **Gender, Race, Age, Height**
  - UltraStar Database: 516 singers, 586 tracks, 37h music, 423k beats
  - 46 openSMILE features, bi-directional LSTM RNN
  - Blind Voice Separation (VS)
  - Harmonics Enhancement by NMF
  - Lead Voice Isolation by source / filter model + NMF



% UA	Base	VS
2-class: voice, none	74.6	**75.7**
2-class: Female, Male	86.9	**89.6**
2-class: White, Other	52.8	**64.4**
2-class: Above / Below 30 years	54.5	**58.9**
2-class: Above / Below 175 cm	64.7	**72.1**

*"Automatic Assessment of Singer Traits in Popular Music: Gender, Age, Height and Race", **ISMIR**, 2011.*

AVEC

- **AVEC Corpus**

  Solid-SAL part of SEMAINE, Challenge: 24 recordings

  ~4 character conversation sessions / recording

Audio Sub-Challenge: word level

Video Sub-Challenge: frame level

Audiovisual Sub-Challenge: word level

# / (h:m:s) / [ms]	**Train**	**Development**	**Test**	**Total**
Sessions	31	32	32	95
Frames	501 277	449 074	407 772	1 358 123
Words	20 183	16 311	13 856	50 350
Total duration	2:47:10	2:29:45	2:15:59	7:32:54
Avg. word duration	262	276	249	263

AVEC

- **Correlation**

  All correlations have p-value << 0.01

CC [%]	Word level E	Word level P	Word level V	Frame level E	Frame level P	Frame level V
Activation	-3.2	22.4	20.7	-3.2	24.5	24.9
Expectation		-35.8	-10.4		-37.3	-7.7
Power			29.7			29.6

- **Baselines**

  Face Registration: position by OpenCV's VJ face detector

  Eye-localization by OpenCV's Haar-cascade object detector

  Image rotation, scaling to 100 pixels between eyes, cropping to 200 x 200

  LBP responses in 59 dim. histograms over face (10 x 10 blocks): 5.9k

  1.9k openSMILE audio features

AVEC

- **Baselines**

  SVM, posteriors per word / modality, binary above / below mean

  Challenging amount of data (> 1 M frames, 5 908 features / frame)

  Video: Sampling 1k frames, audio: 1/3 from training / development

Accuracy [%]	Activity WA	Activity UA	Expectation WA	Expectation UA	Power WA	Power UA	Valence WA	Valence UA	Mean WA
Audio Sub-Challenge									
Development	63.7	64.0	63.2	52.7	65.6	55.8	58.1	52.9	62.7
Test	**55.0**	57.0	**52.9**	54.5	**28.0**	49.1	**44.3**	47.2	**45.1**
Video Sub-Challenge									
Development	60.2	57.9	58.3	56.7	56.0	52.8	63.6	60.9	59.5
Test	**42.2**	52.5	**53.6**	49.3	**36.4**	37.0	**52.5**	51.2	**46.2**
Audiovisual Sub-Challenge									
Test (*A*)	51.2	51.2	59.2	49.5	52.7	45.9	55.8	46.5	54.7
Test (*V*)	77.1	77.2	36.8	45.5	53.7	52.9	60.8	47.6	57.1
Test (*AV*)	**67.2**	67.2	**36.3**	48.5	**62.2**	50.0	**66.0**	49.2	**57.9**

*"AVEC 2011 – The First International Audio/Visual Emotion Challenge", **Springer LNCS**, 6975(II): 415–424, 2011.*

Non-Verbals

- **Types**

  Laughter, Sigh

  Hesitation, Consent

- **Shape & Appearance**

  Register & crop faces from all subjects

  20 tracked facial fiducial points

  4 eye corners and tip of nose (stable, invariant to facial deformations)

  Transform to warp each face to reference frames

  Finally, all faces re-sampled to 64 x 64

  Appearance by first 30 PCs of image gradients

Non-Verbals

- **Audio Features**

Acoustic Low-level Descriptors (9)
Perceptual Linear Prediction Cepstral Coefficients (PLP-CC) 1–5
Logarithmic Energy
Loudness
Fundamental Frequency (F_0)
Probability of Voicing

Functionals (7)
Extremes (maximum, minimum value)
Range (maximum – minimum value)
Arithmetic mean
Standard deviation
Skewness, Kurtosis

Non-Verbals

- **Results on TUM AVIC**

[%] Features	LSTM UAR	LSTM WAR	SVM UAR	SVM WAR
Appear	32.5	50.0	31.8	60.0
Shape	48.4	56.1	39.6	60.2
Shape+Appear	40.8	51.8	39.2	58.2
Audio	64.6	73.5	59.1	72.4
Audio+Appear	60.3	64.2	59.4	72.1
Audio+Shape	**72.0**	**73.5**	60.5	72.4
Audio+Shape+Appear	64.3	63.1	62.7	74.2

*"Audiovisual Classification of Vocal Outbursts in Human Conversation Using Long-Short-Term Memory Networks", **ICASSP**, 2011.*

Animals

- **Animals & Birds**
  HU-ASA: 6h audio, 1.4k turns
  IS09 openSMILE features, SVM / cyclic HMM / LSTM-RNN



% WA	**SVM**	**cHMM**	**LSTM**
5-class: Pass., Non-P., Canidae, Felidae, Primates	56.0	**64.0**	**62.3**
2-class: Passeriformes, Non-Passeriformes	75.6	**79.6**	**81.3**


*"Audio Recognition in the Wild: Static and Dynamic Classification on a Real-World Database of Animal Vocalizations", **ICASSP**, 2011.*

Vision

Summary

- **Recent Avenues towards Computational Paralinguistics**
  - High Realism
  - Standardisation
- **Audio Data**
  - Synthesis
  - Unsupervised Learning
- **Audio Signal Processing**
  - Source Separation by NMF (openBlissART)
  - Feature Brute-Forcing (openSMILE)
- **Computational Intelligence**
  - Temporal evolution by LSTM (openEAR)

Where to Go from Here

- **Separation and Multi-task Processing of Real-Life Streams**
- **Massive Unsupervised Learning of Space and Models**
- **Closing Gap between Analysis & Synthesis**
- **New Challenges…**

“I hear a mother – guess around mid-forties – talk to a young boy in a friendly tone. Seems not be her child, though. He seems to be a rather open nature, yet tired and maybe not truthful.”

Holistic Unsupervised Computational Paralinguistics

Abstract

Recently, an increasing number of speaker states and traits is adressed in research on automatic speaker classification. Examples comprise personality traits, likability, height, and intoxication of a person derived from characteristics of the voice and the spoken content. This talk aims to provide an overview on the dominant methodology used, benchmark accuracies reached as manifested by research Challenges the speaker held, and concludes with recent trends in the field and new avenues to overcome data sparseness and unreliability.