Superhuman Speech Analysis?

Getting Broader, Deeper & Faster.

Björn W. Schuller
Head GLAM, Imperial College London
Chair EIHW, University of Augsburg
CEO audEERING

Superhuman?

Superhuman? ASR.

- **Human: ASR**

  Misses ~1-2 words in 20 → 5-10% “Word Error Rate” (WER)

  → 1 minute conversation ~16 words

- **Machine: ASR**

  Switchboard: 2.4k (260 hrs), 543 speakers

  1995: **43%** (IBM), **2004**: **15.2%** (IBM), **2016:** 8% (IBM), **6.3%** (Microsoft)

  2017: **5.5%** (IBM) **5.1%** (Microsoft/IBM)

  Human: 5.9% WER (single) **5.1% WER** (multiple pro transcribers)

  AM: CNN-BLSTM, LM: entire history of a dialog session

Linguistic Data Consortium, 1993/1997.

Superhuman? Paralings.

- **Speech Analysis (CP): Objective Tasks**

  Alcohol Intoxication

  16 speakers from ALC, 47 listeners: **71.7%** UAR (human)

  Interspeech 2011 Challenge full ALC: **72.2%** UAR (system fusion)

  Agglomoration (Weninger et al. 2011) >80%

  Heart Rate, Skin Conductance, Health State, …

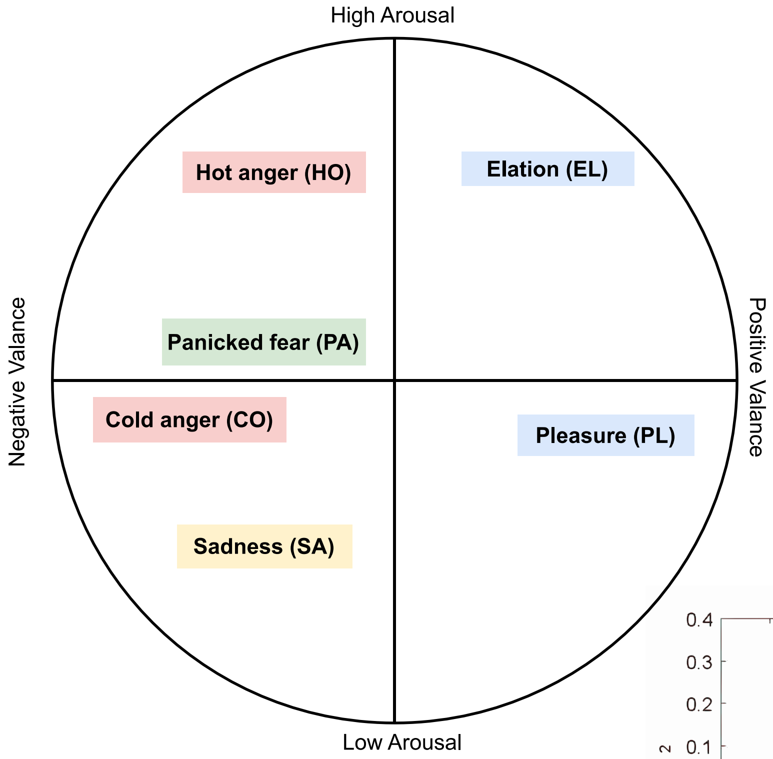
- **Speech Analysis (CP): Subjective Tasks**

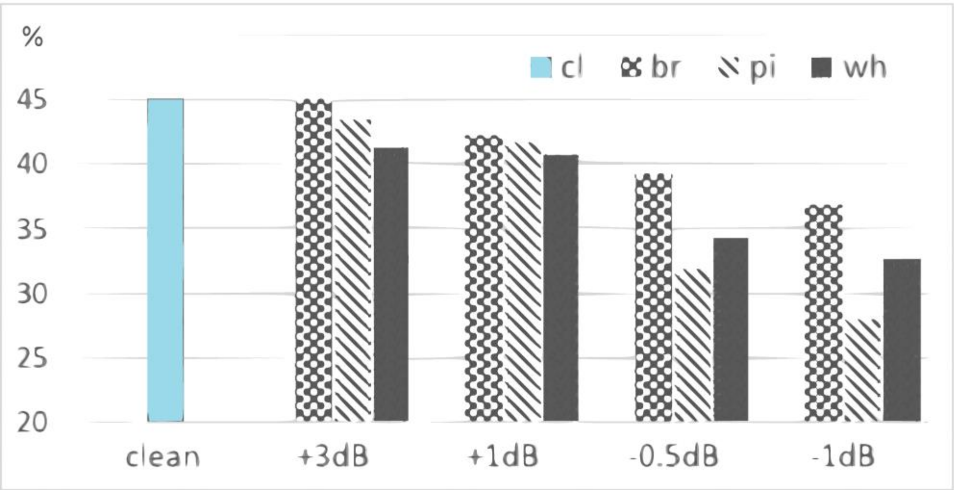
  Ground Truth?

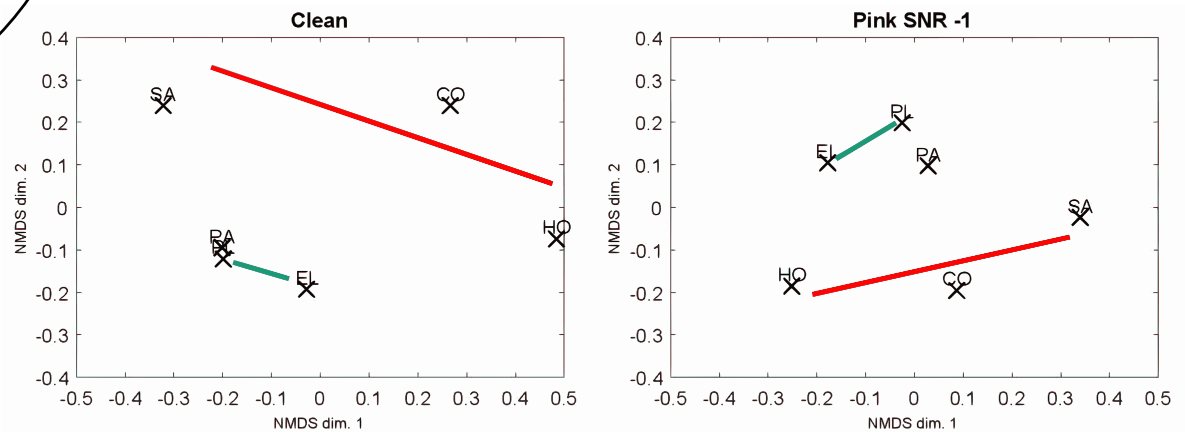
  Emotion, Personality, Likability, …?

Schiel: "Perception of Alcoholic Intoxication in Speech", Interspeech, 2011.

Human Performance?







“The Perception of noisified non-sense speech in the noise”, Interspeech, 2017.

Rett & ASC.

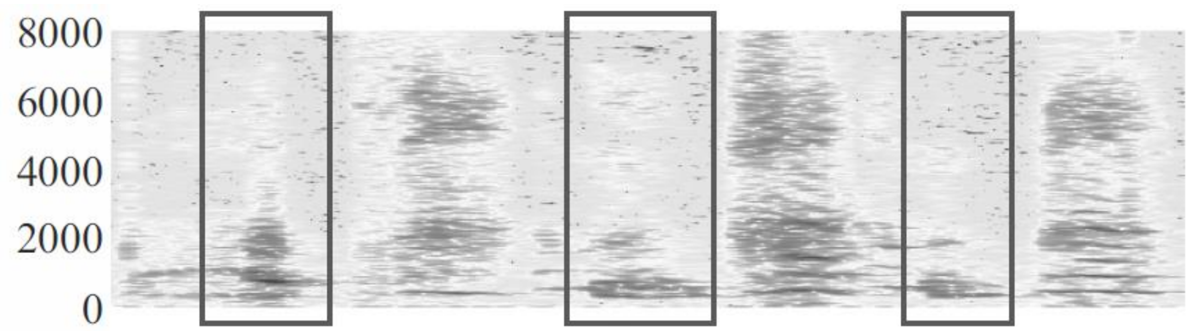
	%UA
Rett Syndrome	76.5
ASC	75.0

- **Rett & ASC Early Diagnosis**

  16 hours of home videos

  6-12 / 10 months

  Vocal cues: e.g., inspiratory vocalisation



"A Novel Way to Measure and Predict Development: A Heuristic Approach to Facilitate the Early Detection of Neurodevelopmental Disorders", Current Neurology and Neuroscience Reports, 2017.
"Earlier Identification of Children with Autism Spectrum Disorder: An Automatic Vocalisation-based Approach", Interspeech, 2017.

Getting Broader.

Speaker ID & Verification

Speech Recognition

Language Understanding

Deep Paralings

Speech Analysis

Sentiment Analysis

Gender Recognition

Broad Paralings

Emotion Recognition

Language ID

Health Classification

Speaker Diarisation

Personality Recognition

Paralings.

INTERSPEECH ComParE

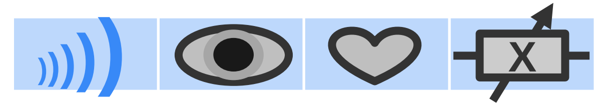
2018	# Classes	%UA/*AUC/$^+$CC
Affect: Atypical	[-1,1]	?
Affect: Self-Ass.	[-1,1]	?
Crying	3	?
Heart Beats	3	?

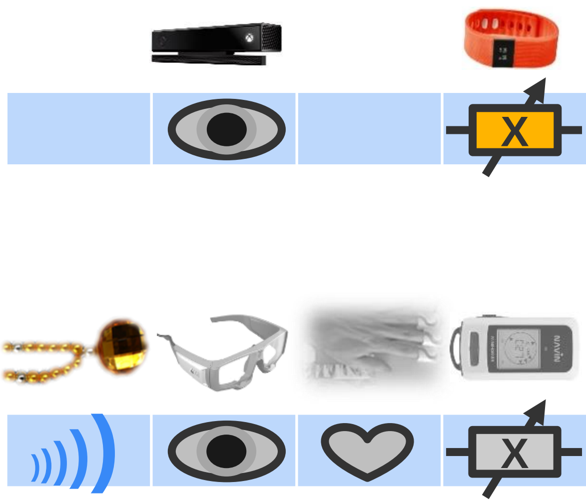
	# Classes	%UA/*AUC/$^+$CC
Addressee	2	70.6
Cold	2	72.0
Snoring	4	70.5
Deception	2	72.1
Sincerity	[0,1]	65.4+
	11	82.2
	[0,1]	43.3+
	[0,100]	54.0$^+$
	7	62.7
	3	61.6
Physical Load	2	71.9
Social Signals	2x2	92.7*
Conflict	2	85.9
Emotion	12	46.1
Autism	4	69.4

Age
Gender
Interest
Emotion
Negativity

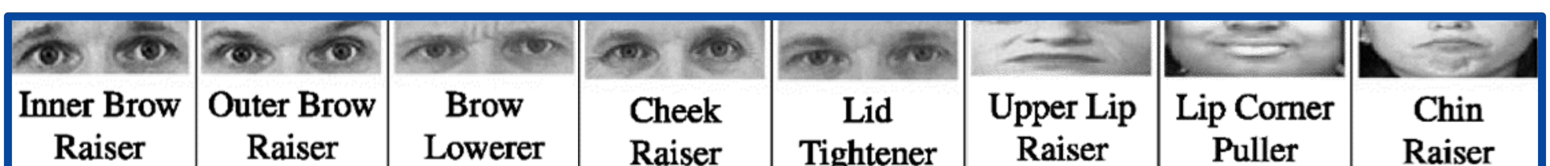
Broad Paralings.

- **Pseudo Multimodality**



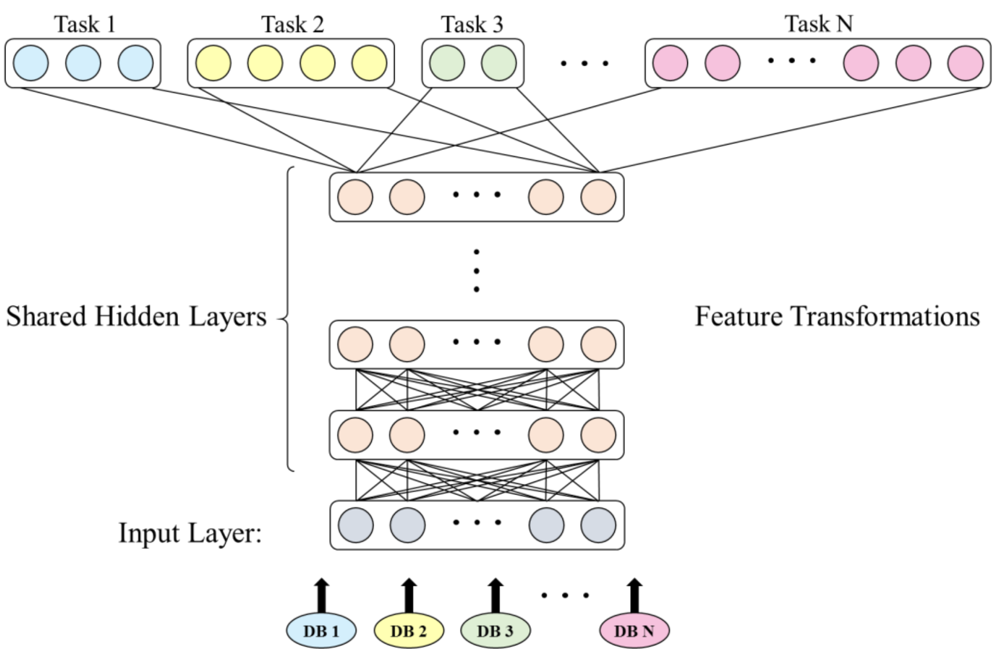


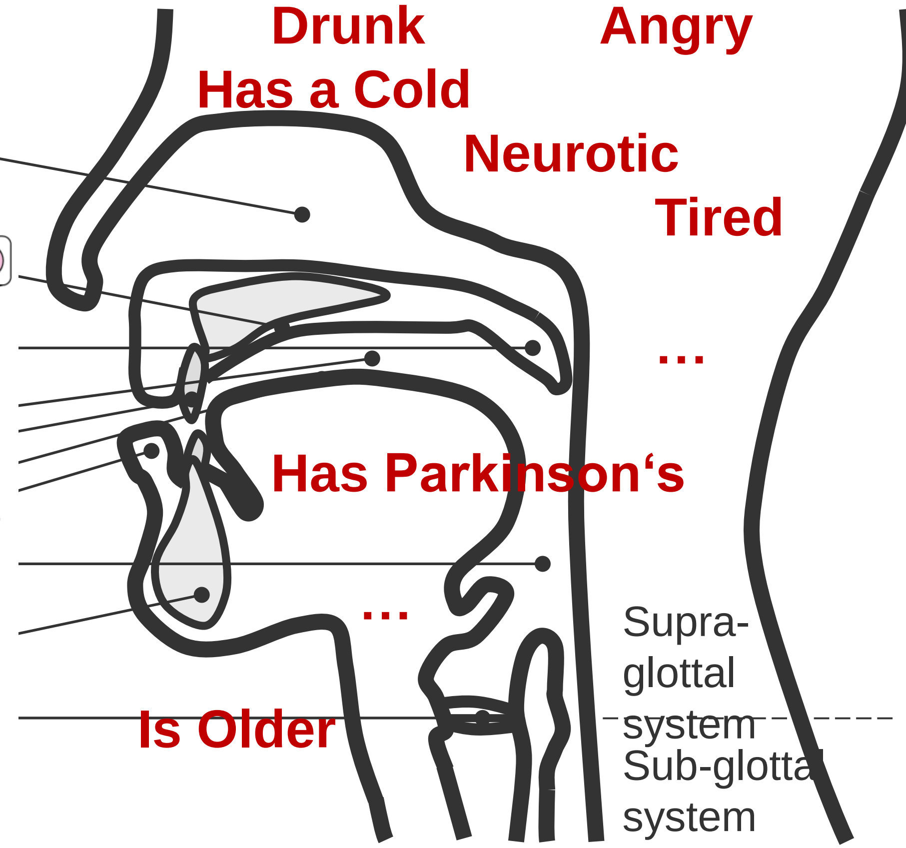
	*MAE +CC %UA
Heart Rate	8.4*
Skin Conductance	.908+
Facial Action Units	65.0
Eye-Contact	67.4



Broad Paralings.

- **Multiple-Targets**
- **1 Voice**



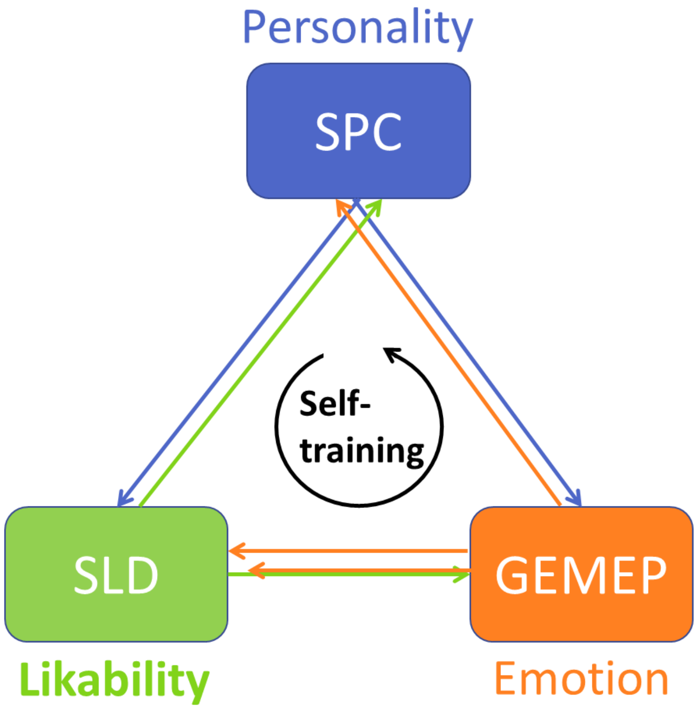


"Multi-task Deep Neural Network with Shared Hidden Layers: Breaking down the Wall between Emotion Representations", ICASSP, 2017.

Broad Paralings.

- **Cross-Task Self-Labelling**

%UA	Base	CTL
Extraversion	71.7	+1.8
Agreeableness	58.6	+4.5
Neuroticism	63.3	+3.0
Likability	57.2	+2.9

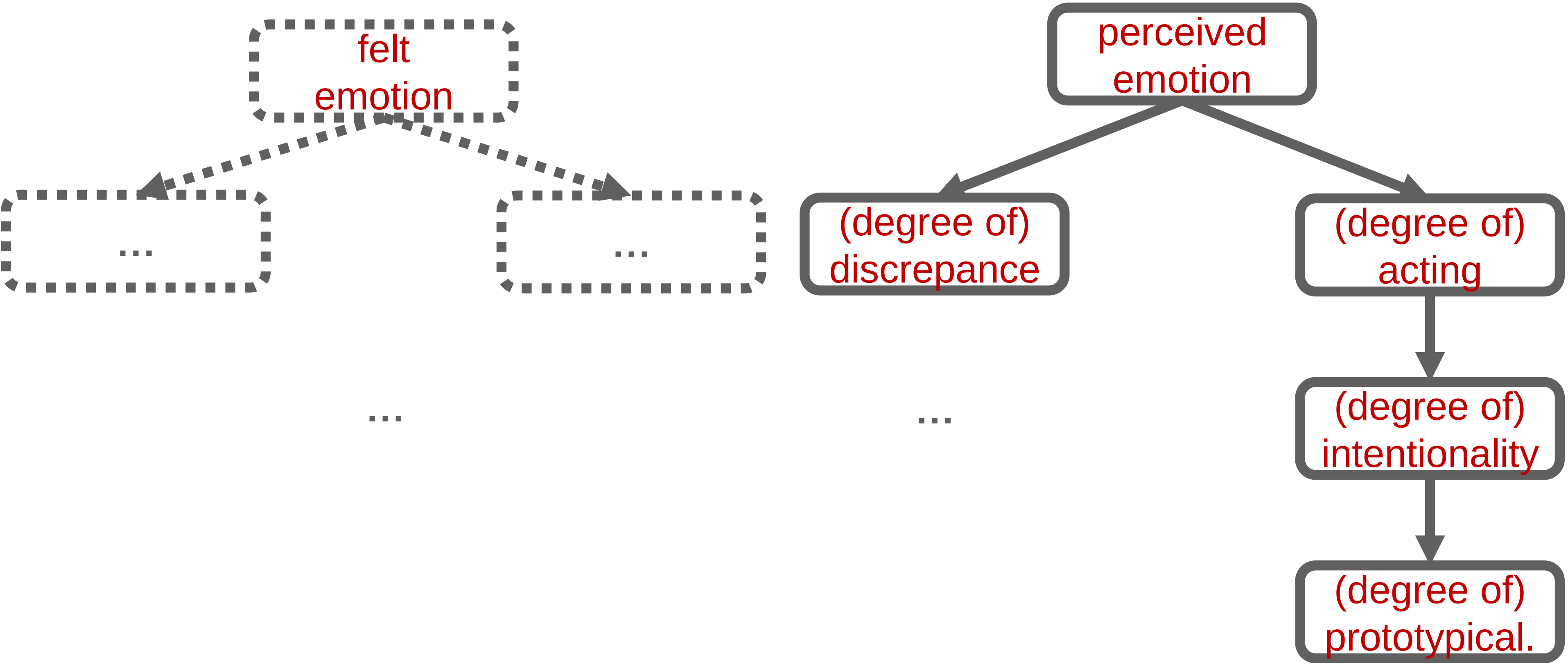


Algorithm: *Cross-Task Labelling*
Repeat for each task:
Repeat until $\mathcal{U} \in \{\}$:

1. (Optional) Upsample training set $\mathcal{L}$ to even class distribution $\mathcal{L}_D$
2. Use $\mathcal{L}/\mathcal{L}_D$ to train classifier $\mathcal{H}$, then classify $\mathcal{U}$
3. Select a subset $\mathcal{N}_{st}$ that contains those instances predicted with the highest confidence values
4. Remove $\mathcal{N}_{st}$ from the unlabelled set $\mathcal{U}$, $\mathcal{U} = \mathcal{U} \setminus \mathcal{N}_{st}$
5. Add $\mathcal{N}_{st}$ to the labelled set $\mathcal{L}$, $\mathcal{L} = \mathcal{L} \cup \mathcal{N}_{st}$

"Semi-Autonomous Data Enrichment Based on Cross-Task Labelling of Missing Targets for Holistic Speech Analysis", ICASSP, 2016.

Deep Paralings.

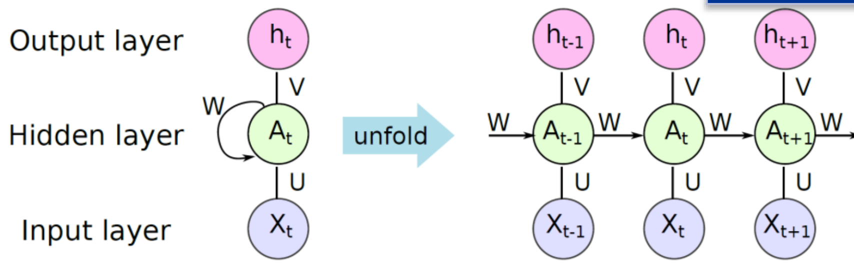


"Reading the Author and Speaker: Towards a Holistic and Deep Approach on Automatic Assessment of What is in One's Words", CICLing, 2017.

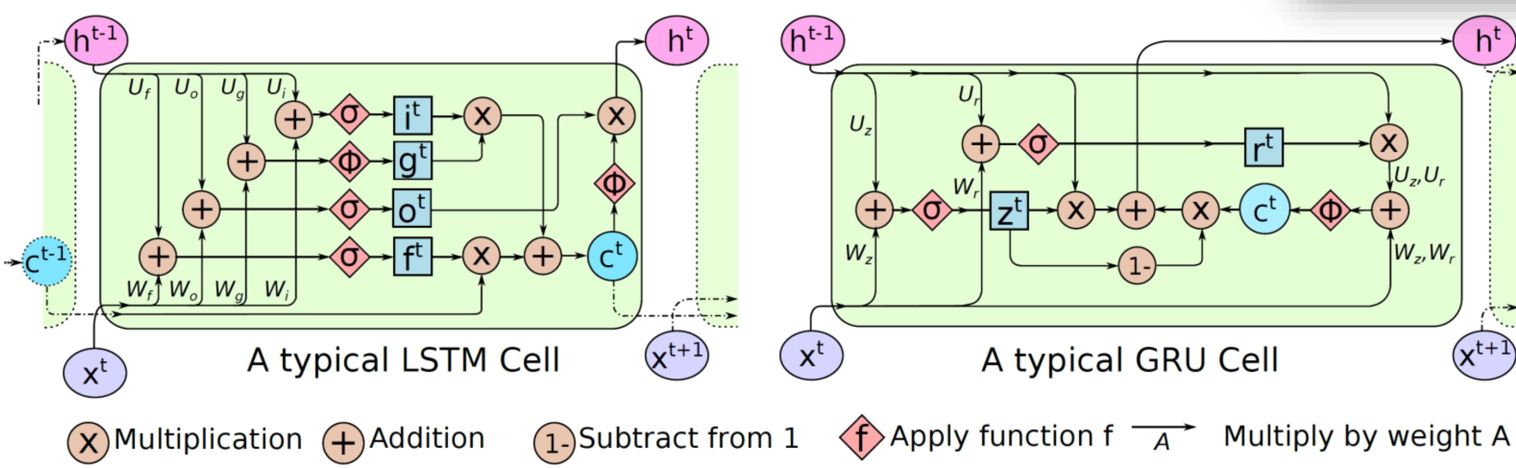
Getting Deeper.

Deep Recurrent Nets.

Arousal	CC
HMM	83.5
HMM+LSTM-RNN	87.2
(LSTM-RNN)	96.3

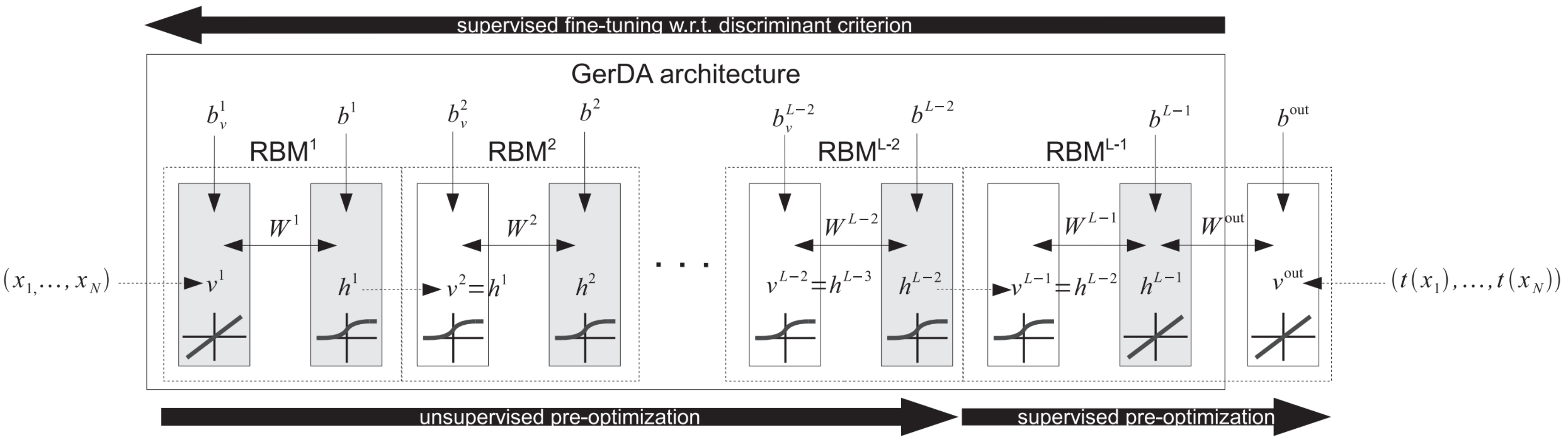


Recurrent neural network unfolded



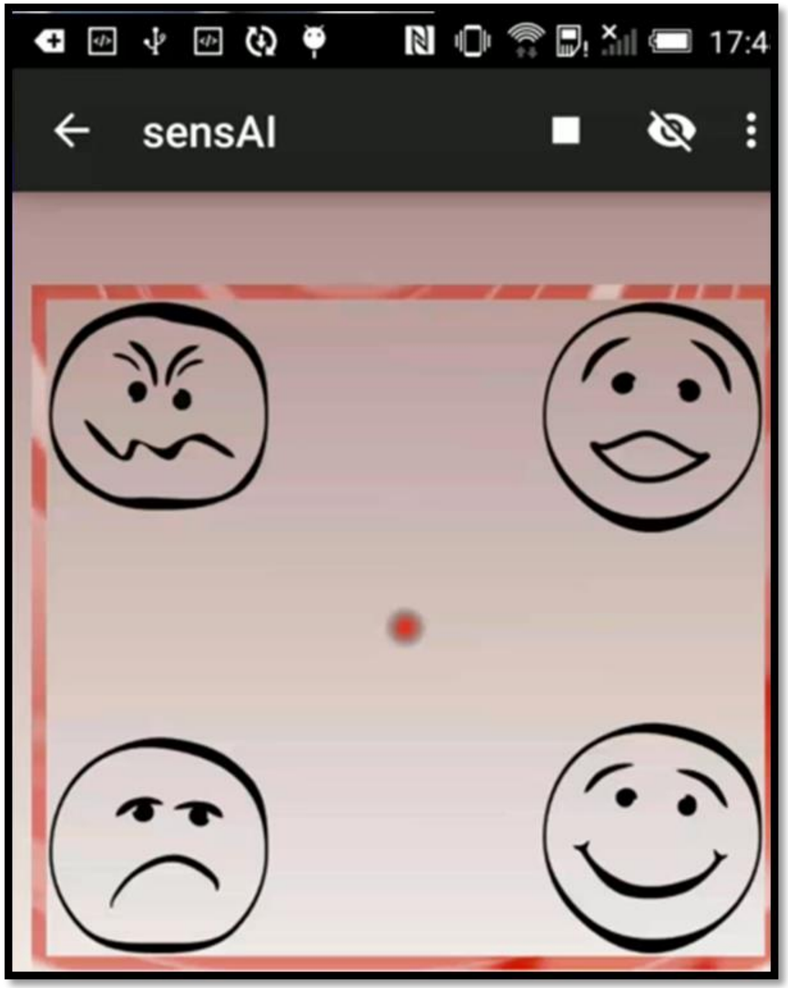
"A Combined LSTM-RNN-HMM Approach to Meeting Event Segmentation and Recognition", ICASSP, 2006.
"Abandoning Emotion Classes – Towards Continuous Emotion Recognition with Modelling of Long-Range Dependencies", Interspeech, 2008.

Deep Recurrent Nets.

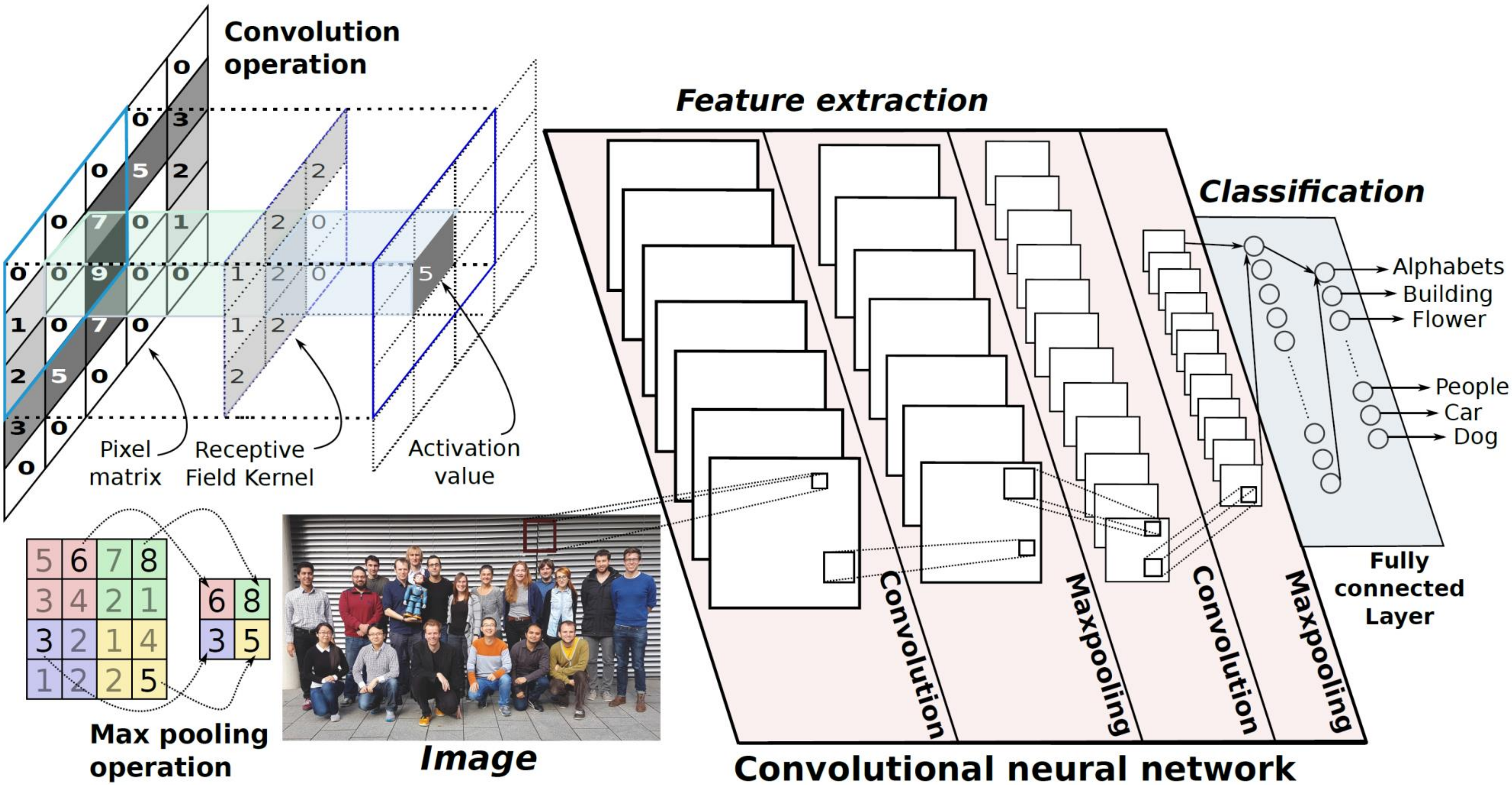


"Robust discriminative keyword spotting for emotionally colored spontaneous speech using bidirectional LSTM networks", ICASSP, 2009.
"Deep neural networks for acoustic emotion recognition: raising the benchmarks", ICASSP, 2011.

Deep Recurrent Nets.



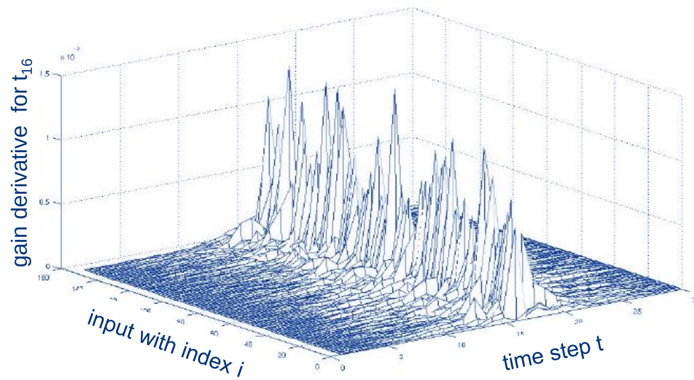
Convolutional Neural Nets.

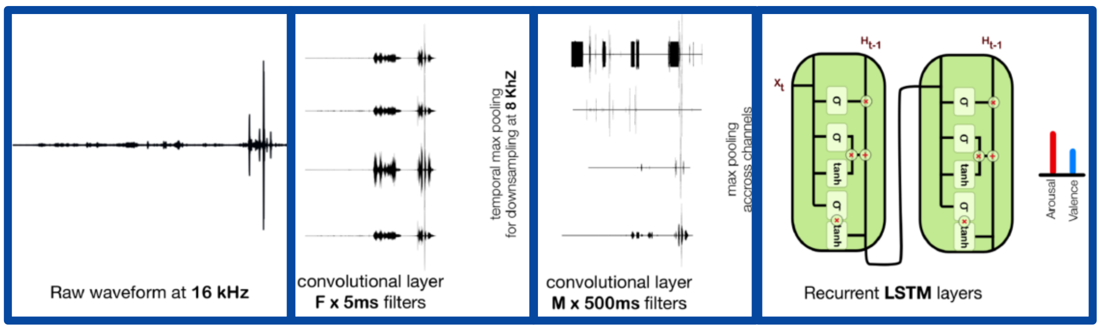


End-to-End.

CCC Recola	Arousal	Valence
ComParE+LSTM	.382	.187
e2e (2016)	.686	.261

- **CNN + LSTM RNN**

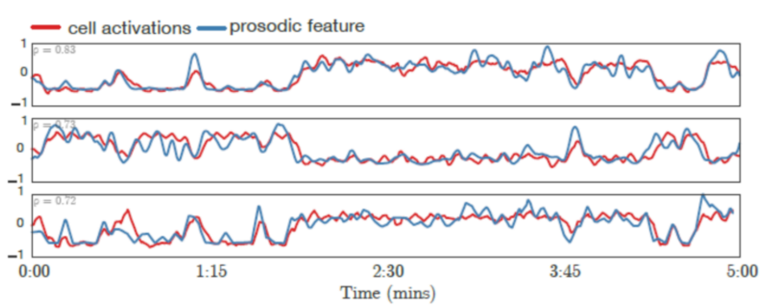




energy range (.77)

loudness (.73)

F0 mean (.71)

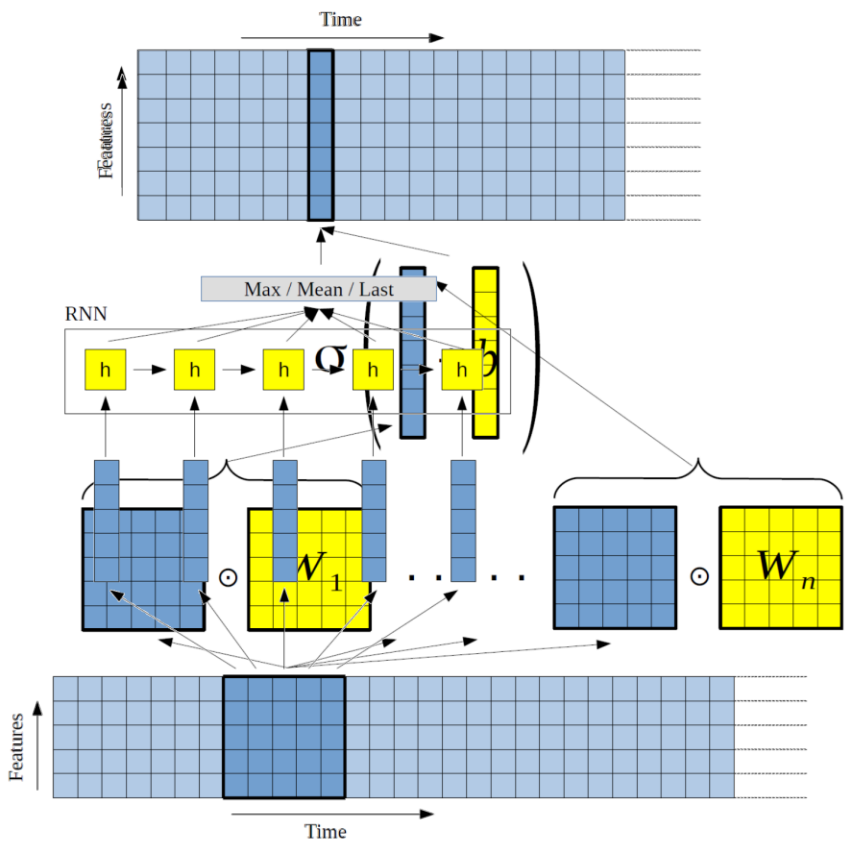


"Adieu Features? End-to-End Speech Emotion Recognition using a Deep Convolutional Recurrent Network", ICASSP, 2016.

End-to-End.

CCC Recola	Arousal	Valence
ComParE+LSTM	.382	.187
e2e (2016)	.686	.261

- **CNN + LSTM → CLSTM ?**



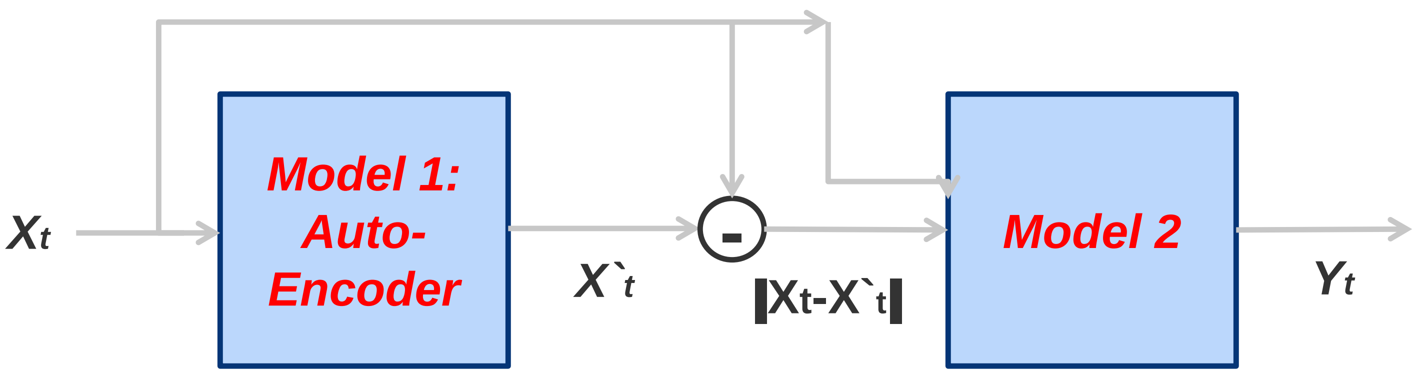
"Convolutional RNN: an enhanced model for extracting features from sequential data", IJCNN, 2016.

Learning by Errors.

CCC Recola	Arousal	Valence
ComParE+LSTM	.382	.187
e2e (2016)	.686	.261
Reconstruction Error	.729	.360

- **Reconstruction Error**

RE of Auto-Encoder as additional input feature



Either: Low Level Descriptors (LLD) or Statistical funtionals

Deep BLSTM RNN

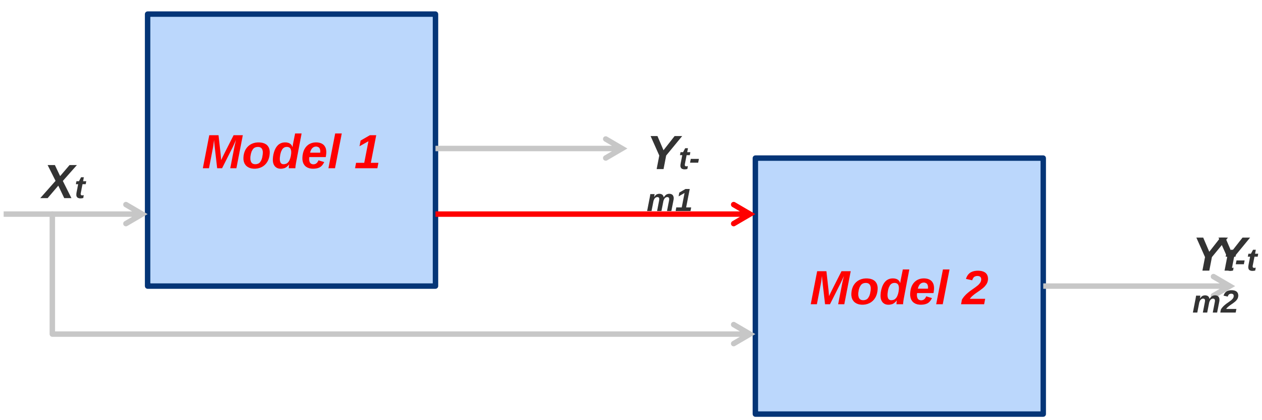
“Reconstruction-error-based Learning for Continuous Emotion Recognition in Speech”, ICASSP, 2017.

Prediction-based.

CCC Recola	Arousal	Valence
ComParE+LSTM	.382	.187
e2e (2016)	.686	.261
Reconstruction Error	.729	.360
Prediction-based	.744	.377

- **Tandem Learning**

  concatenate two models
  for combined strengths

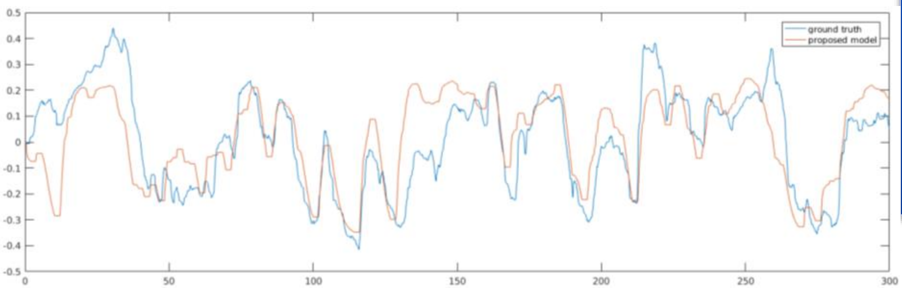


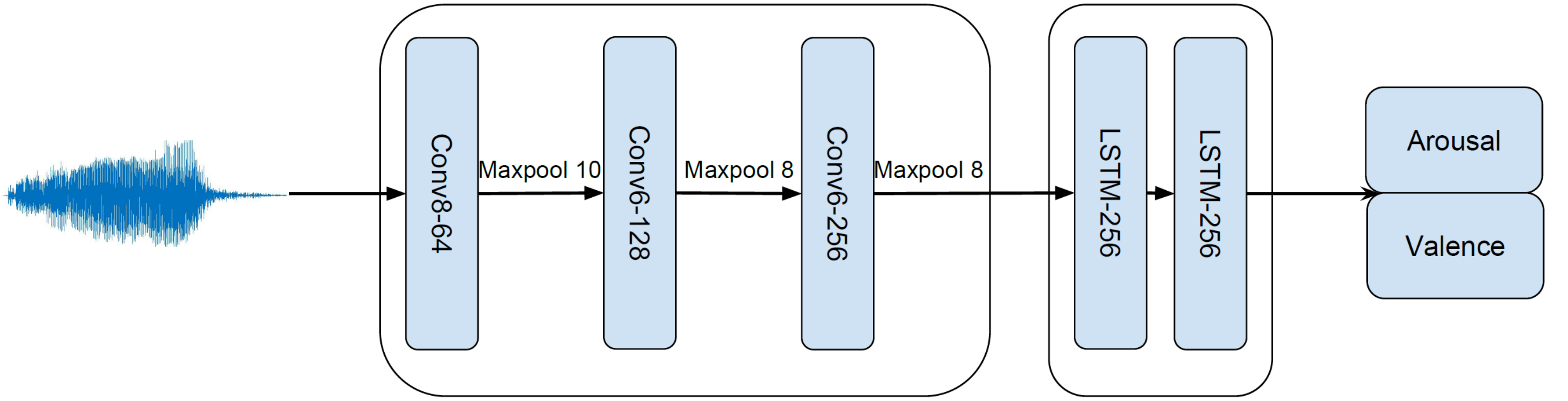
"Prediction-based Learning for Continuous Emotion Recognition in Speech", ICASSP, 2017.

End-to-End.

- **CNN + LSTM RNN**

CCC Recola	Arousal	Valence
ComParE+LSTM	.382	.187
e2e (2016)	.686	.261
Reconstruction Error	.729	.360
Prediction-based	.744	.377
BoAW	.753	.430
e2e (submitted)	.787	.440





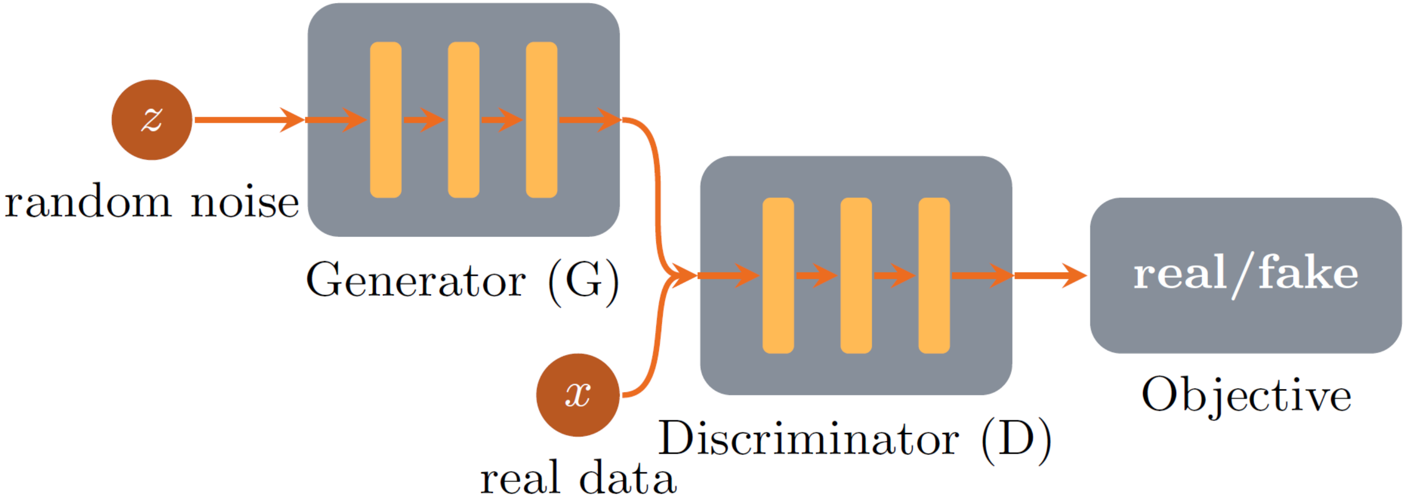
Input raw waveform at 16kHz — Convolution layers — Recurrent layers — Output label at 25Hz

“Affect Recognition by Brdiging the Gap between End-2-End Deep Learning and Conventional Features”, submitted.

Adversarial Nets.

- **Conditional Adversarial Nets**

CCC Recola	Arousal	Valence
ComParE+LSTM	.382	.187
e2e (2016)	.686	.261
Reconstruction Error	.729	.360
Prediction-based	.744	.377
BoAW	.753	.430
e2e (submitted)	.787	.440
CAN (submitted)	.737	.455



"Towards Conditional Adversarial Networks for Recognition of Emotion in Speech", submitted.

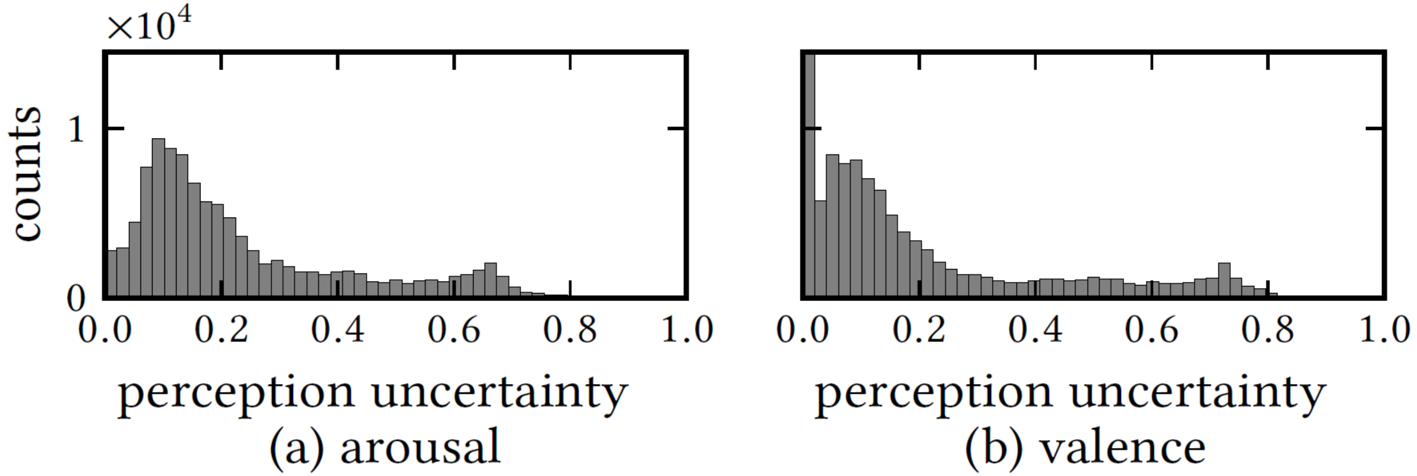
Co-Learning Trust.

- **Multi-task Learning of Subjective / Uncertain Ground Truth**

Example: Arousal / Valence (SEWA data of AVEC 2017)

Perception uncertainty (K ratings):

$$\sigma_n^{(i)} = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}\left(e_{n,k}^{(i)} - e_n^{\mathrm{MLE},(i)}\right)^2}$$



"From Hard to Soft: Towards more Human-like Emotion Recognition by Modelling the Perception Uncertainty", ACM Multimedia, 2017.

Co-Learning Trust.

CCC SEWA	Arousal	Valence
Single	.234	.267
Multiple (+conf)	.275	.292
Single (A/V)	.386	.478
Multiple (+conf, A/V)	.450	.515

(a) $E^{(V)} = 0.08,\ \sigma^{(V)} = 0.17$

(b) $E^{(V)} = 0.08,\ \sigma^{(V)} = 0.79$

(c) $E^{(V)} = 0.79,\ \sigma^{(V)} = 0.47$

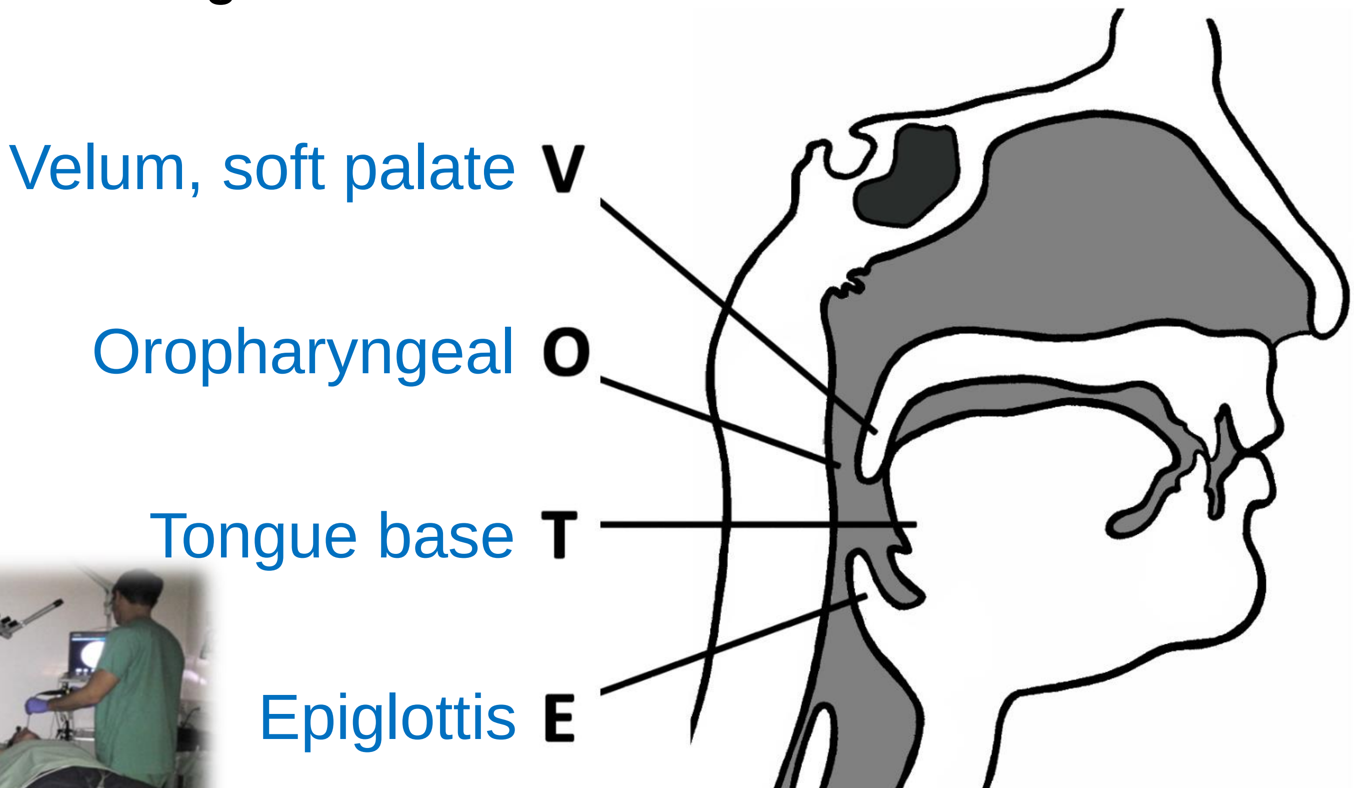
(d) $E^{(V)} = 0.69,\ \sigma^{(V)} = 0.70$

*"From Hard to Soft: Towards more Human-like Emotion Recognition by Modelling the Perception Uncertainty", **ACM Multimedia**, 2017.*

Audio = Images?

	%UA
CNN+LSTM	40.3
Functionals	58.8

- **VOTE Snoring Classification**

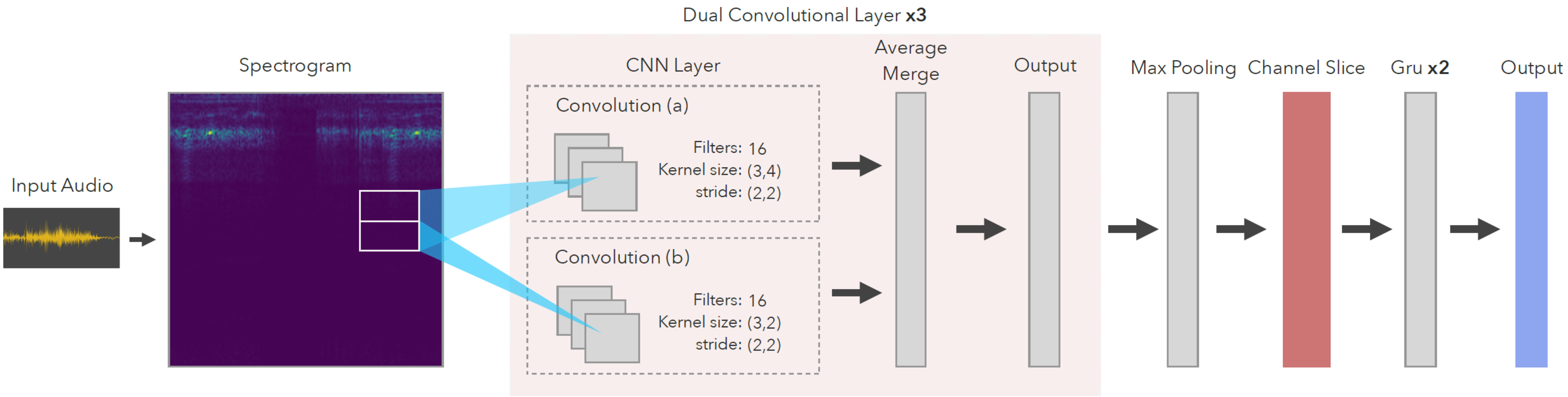


"Classification of the Excitation Location of Snore Sounds in the Upper Airway by Acoustic Multi-Feature Analysis", IEEE Transactions on Biomedical Engineering, 2017.

Audio = Images?

- **VOTE Snoring Classification**

	%UA
CNN+LSTM	40.3
Functionals	58.8
CNN+GRU	63.8

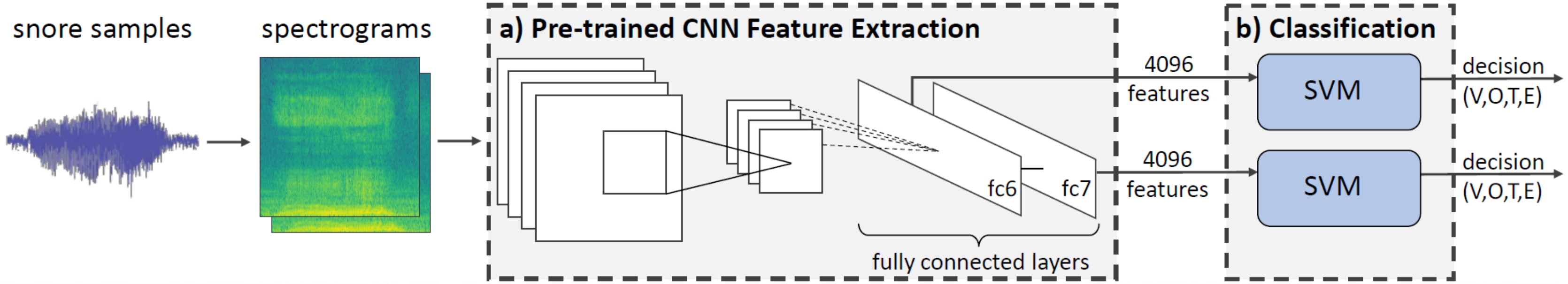


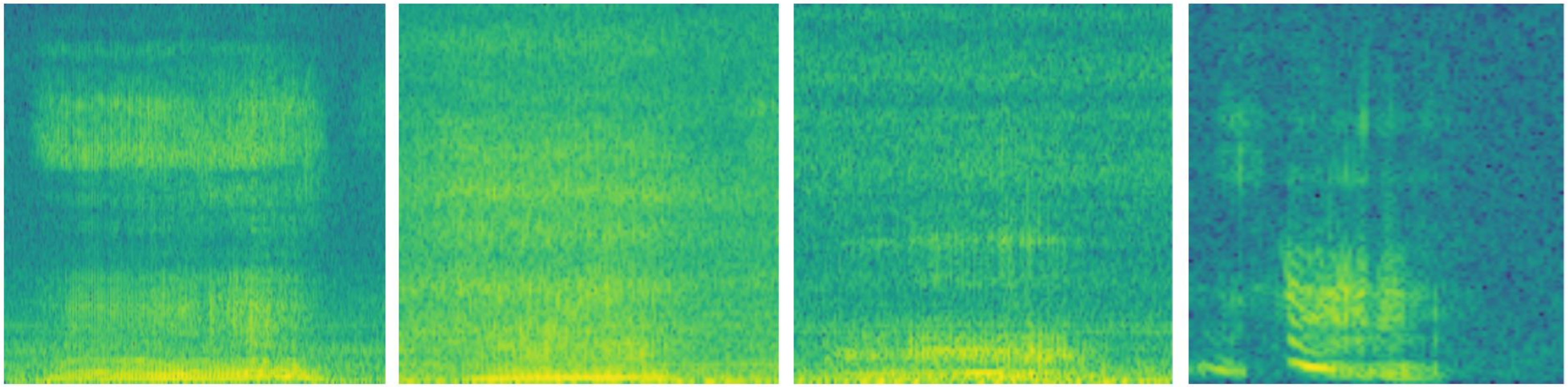
"A CNN-GRU Approach to Capture Time-Frequency Pattern Interdependence for Snore Sound Classification", submitted.

Audio = Images?

- **VOTE Snoring Classification**

	%UA
CNN+LSTM	40.3
Functionals	58.8
CNN+GRU	63.8
Deep Spec	67.0



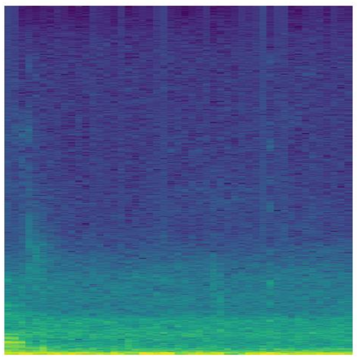


(a) *Velum* (b) *Oropharyngeal lateral walls* (c) *Epiglottis* (d) *Tongue*

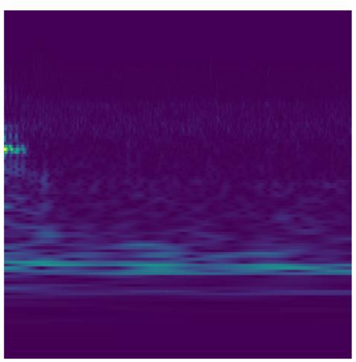
"Snore sound classification using image-based deep spectrum features", Interspeech, 2017.

Audio = Images?

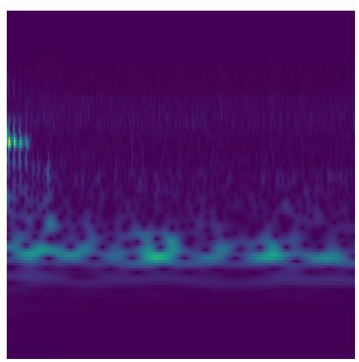
- **Wavelets vs STFT via VGG16**



(a) STFT



(b) *bump* wavelet



(c) *morse* wavelet

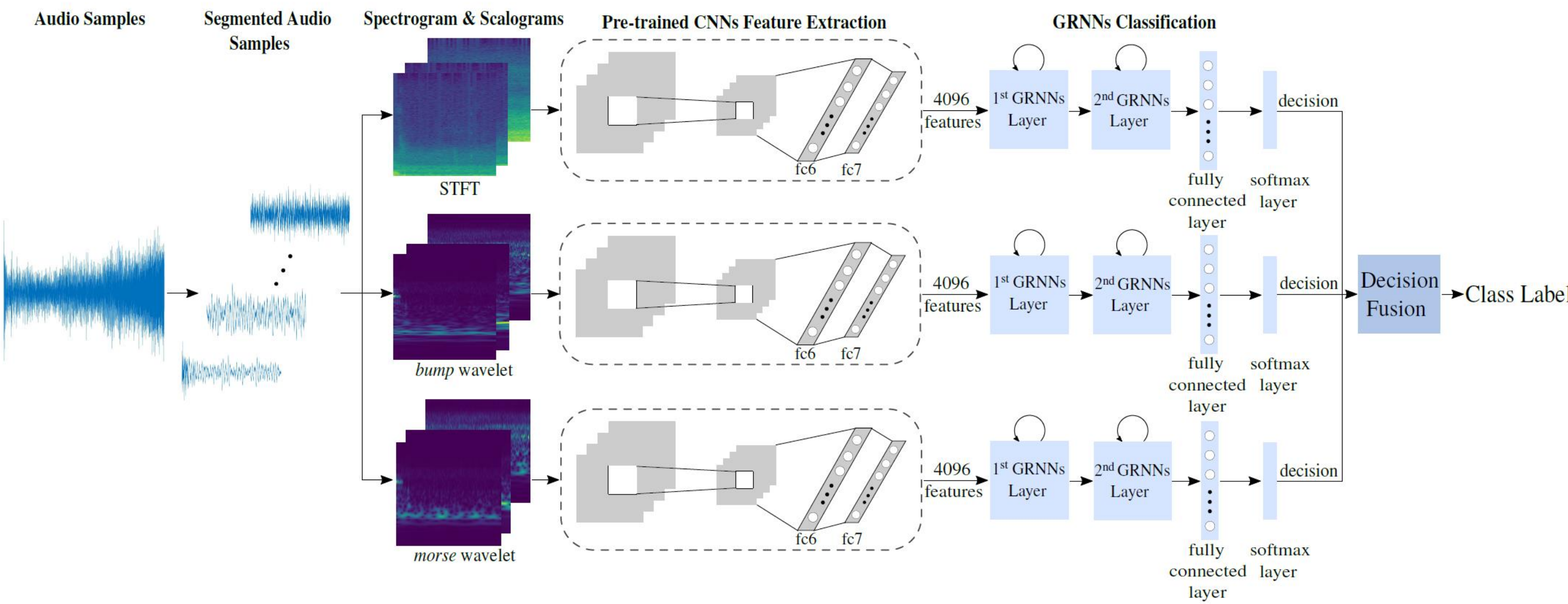
Input: 224×224 RGB image
2×conv size: 3; ch: 64 Maxpooling
2×conv size: 3; ch: 128 Maxpooling
3×conv size: 3; ch: 256 Maxpooling
3×conv size: 3; ch: 512 Maxpooling
3×conv size: 3; ch: 512 Maxpooling
Fully connected layer *fc6* with 4096 neurons Fully connected layer *fc7* with 4096 neurons Fully connected layer with 1000 neurons
Output: softmax layer of probabilities for 1000 classes

“Deep Sequential Image Features for Acoustic Scene Classification", DCASE, 2018.

Audio = Images?

- **Wavelets vs STFT via VGG16**

DCASE 2017	%WA
STFT	76.5
STFT+bump	79.8
STFT+morse	76.9
All	80.9

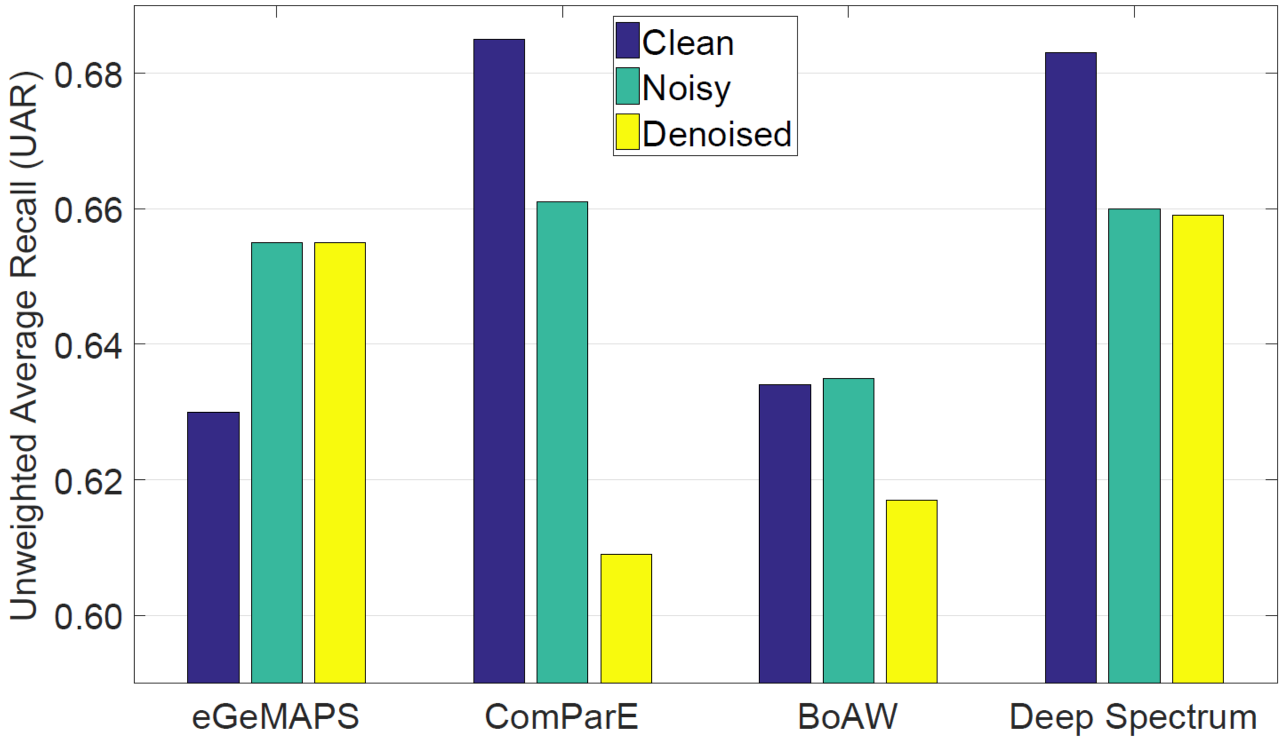


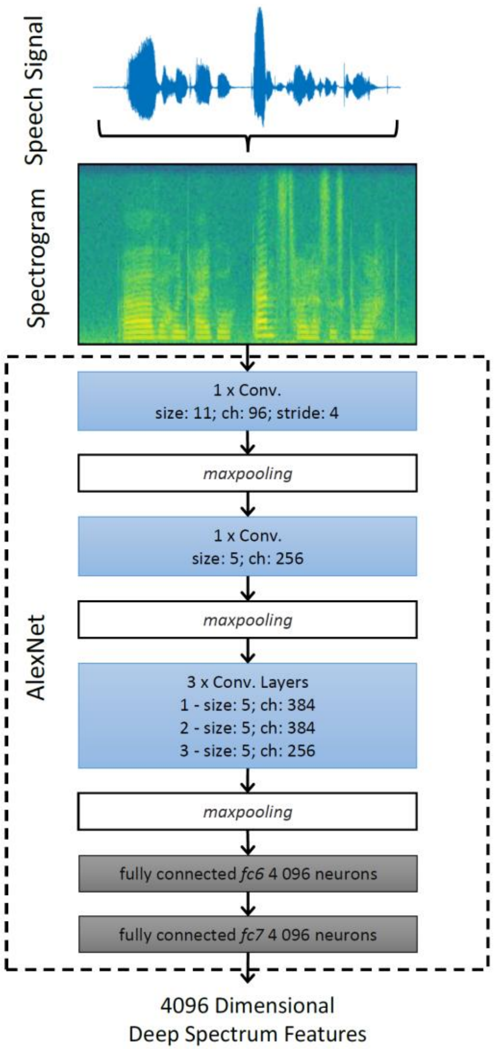
"Deep Sequential Image Features for Acoustic Scene Classification", DCASE, 2018.

Speech = Images?

- **Emotion with Image Nets**

  IS Emotion Challenge task – 2 classes



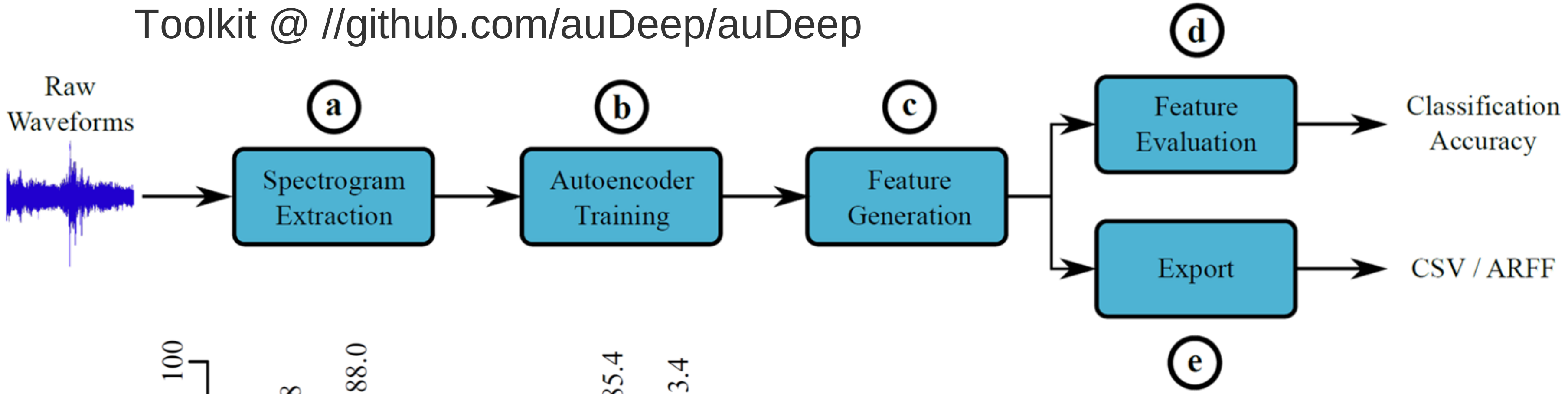


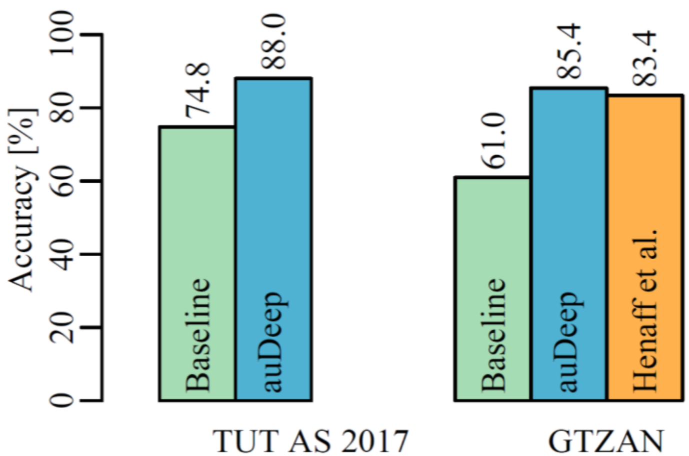
"An Image-based Deep Spectrum Feature Representation for the Recognition of Emotional Speech", ACM Multimedia, 2017.

Speech = Images?

- **audDeep**

  Toolkit @ //github.com/auDeep/auDeep





"auDeep: Unsupervised Learning of Representations from Audio with Deep Recurrent Neural Networks", arxiv.org, 2017.

Getting Faster.

Data?

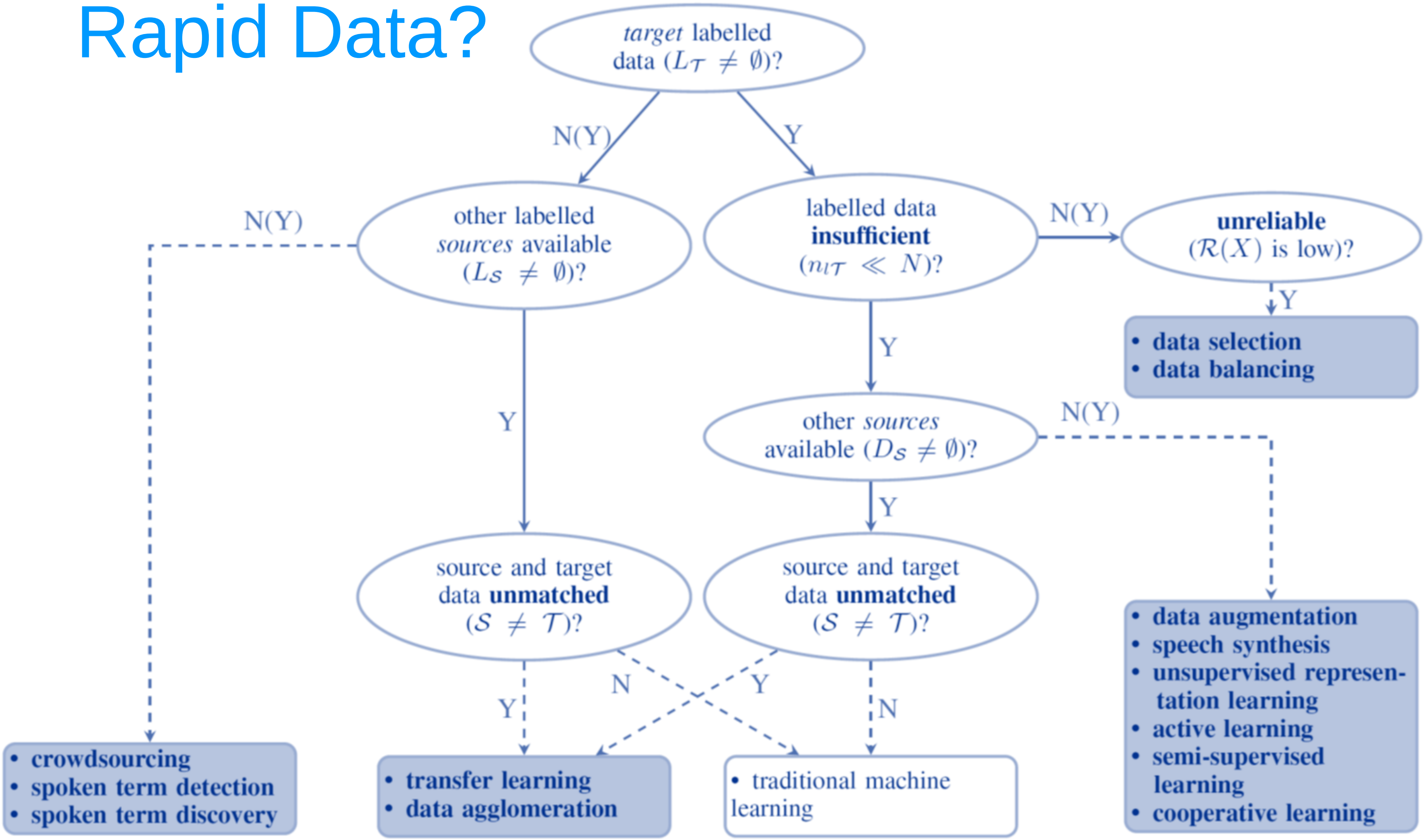
- **2.0 Yet?**

0-1 years:	**1 – 100** hrs	**ASA (~10 hrs)**
2-3 years:	~**1000** hrs	
10-x years:	**~10000** hrs	**ASR (2000+ hrs)**

→ Recognise states/traits independent of person, content, language, cultural background, acoustic disturbances at human parity?

R. Moore, "A Comparison of the Data Requirements of Automatic Speech Recognition Systems and Human Listeners", 2003.

Rapid Data?

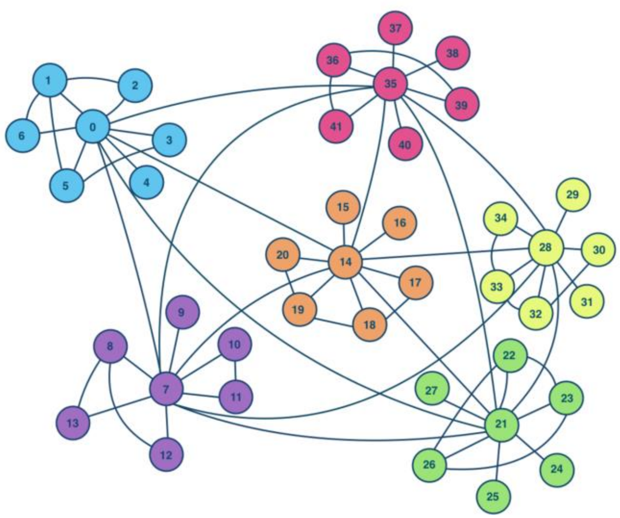


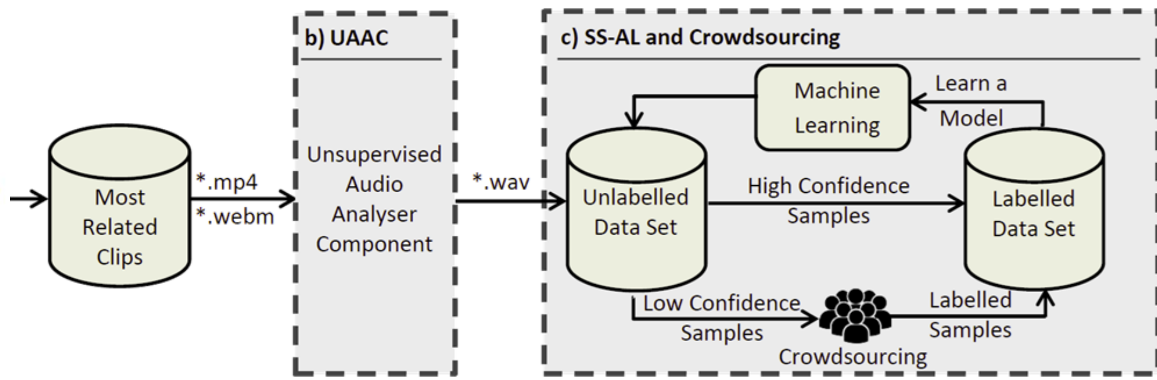
"Efficient Data Exploration for Automatic Speech Analysis: Challenges and State of the Art", IEEE Signal Processing Magazine, 2017.

Rapid Data.

- **YouTube?**
  300 h/min videos
  3k videos for new tasks
  only 3 h/task

%UA	oSMILE	oXBOW	CNN
Freezing	**70.2**	67.5	57.0
Intoxication	64.7	**72.6**	66.8
Screaming	89.2	**97.0**	89.2
Threatening	**73.8**	67.0	71.9
Coughing	95.4	**97.6**	95.4
Sneezing	79.2	79.8	**85.2**



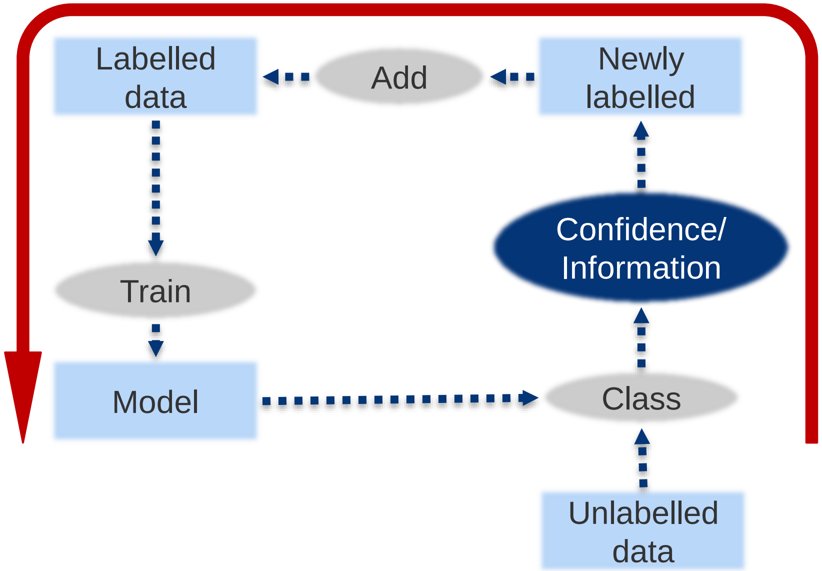


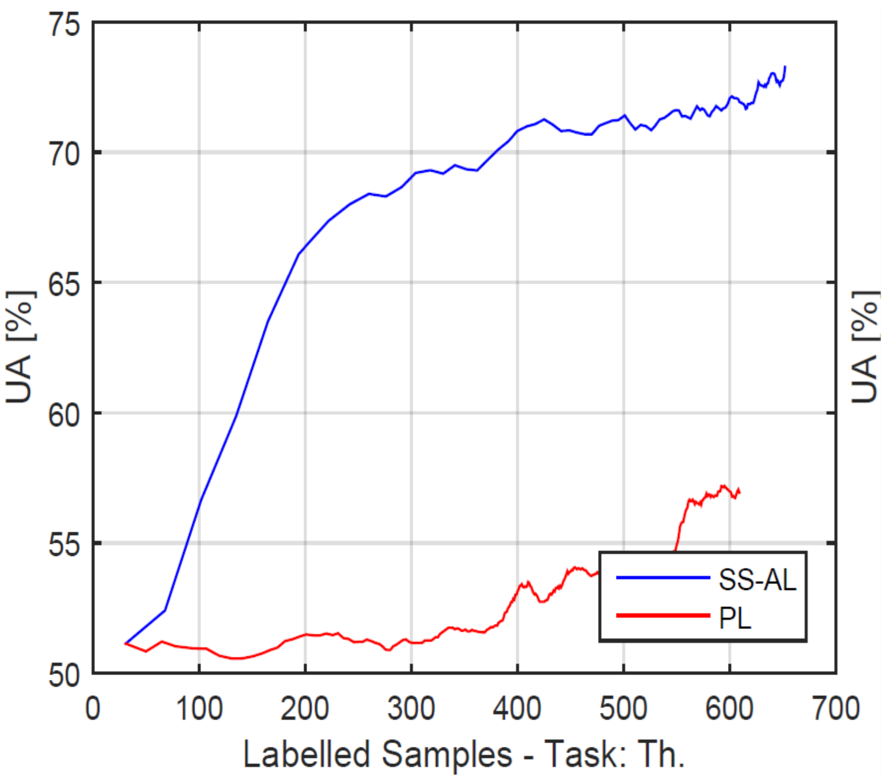
"CAST a Database: Rapid Targeted Large-Scale Audio-Visual Data Acquisition via Small-World Modelling of Social Media Platforms", ACII 2017.

Rapid Data.

iHEAR u PLAY

- **Intelligent Labelling**
  0) Transfer Learning
  1) Dynamic Active Learning
  2) Semi-Supervised Learning

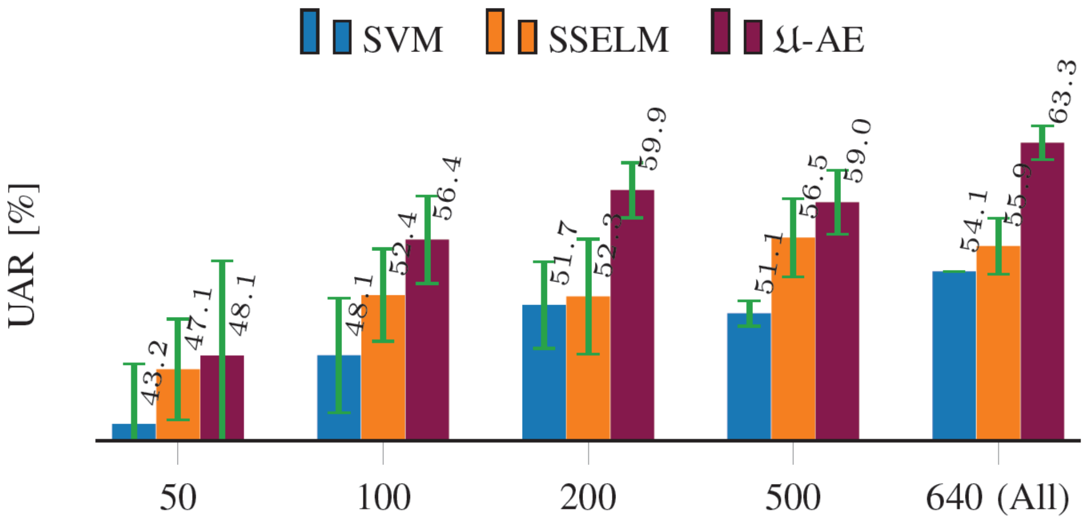


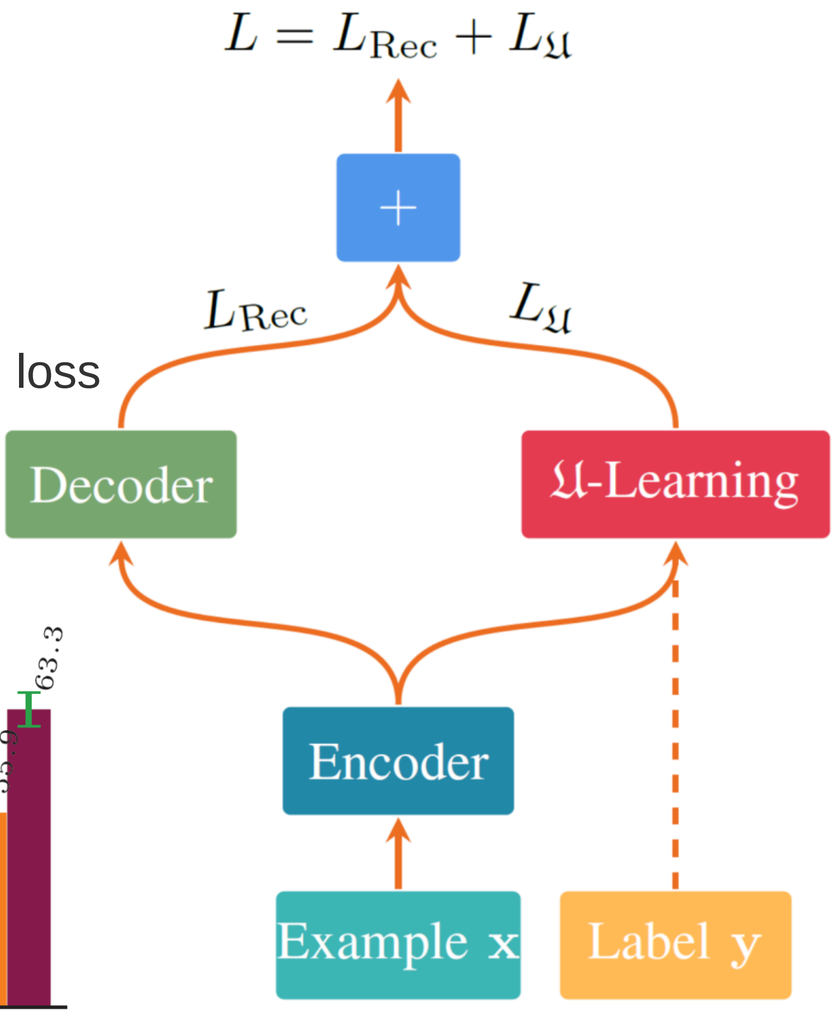


"Cooperative Learning and its Application to Emotion Recognition from Speech", IEEE Transactions Audio Speech & Language Processing, 2015.

Rapid Data: TL.

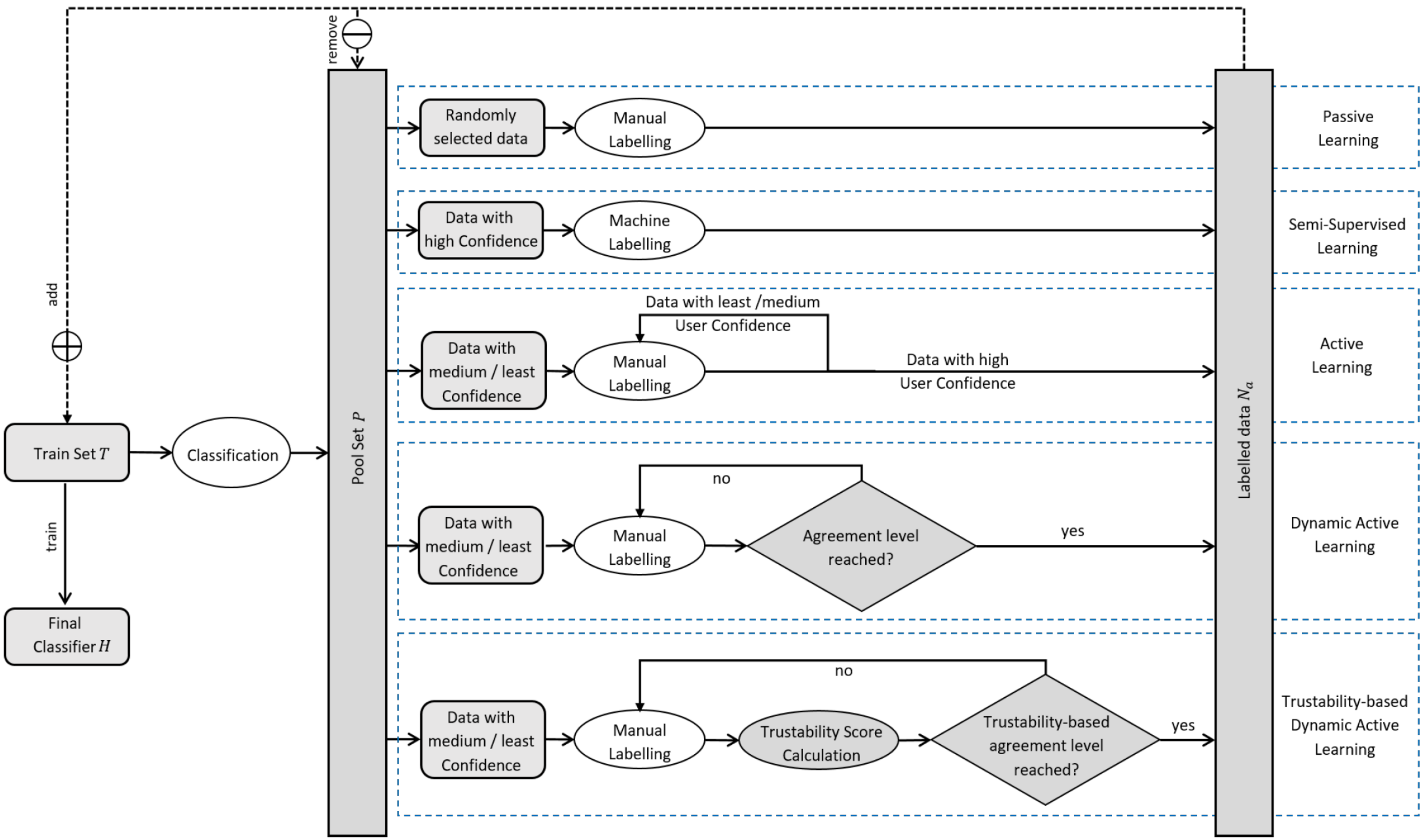
- **TL: Universum Autoencoders**
  Jointly minimise reconstruction error & universum (unlabelled dataset) learning loss
  Whispered → TRANSFER → normal
  GeWEC (4 class) + Unlabelled: ABC





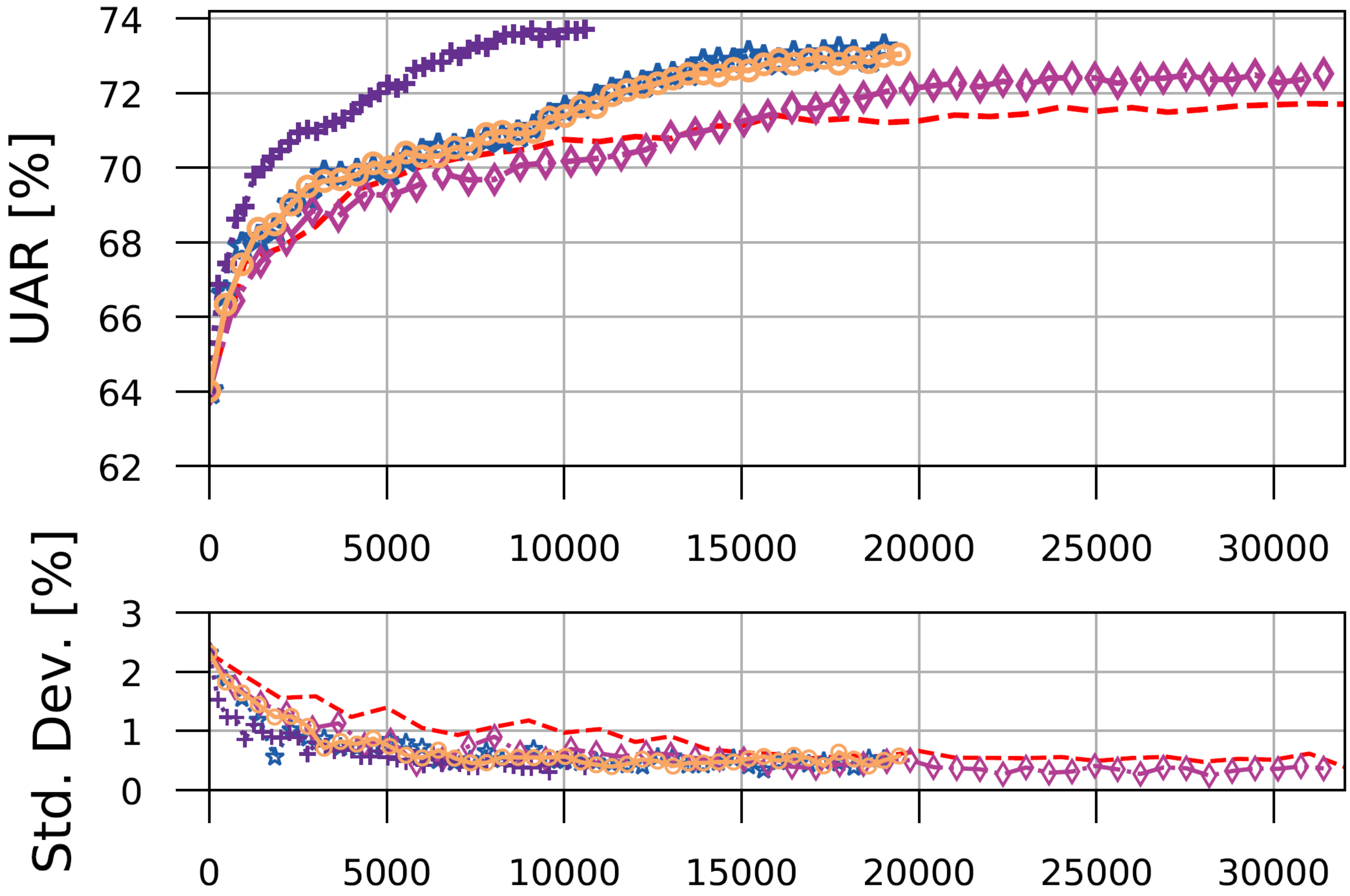
*"Universum Autoencoder-based Domain Adaptation for Speech Emotion Recognition", **Signal Processing Letters**, 2017.*

Rapid Data: AL.



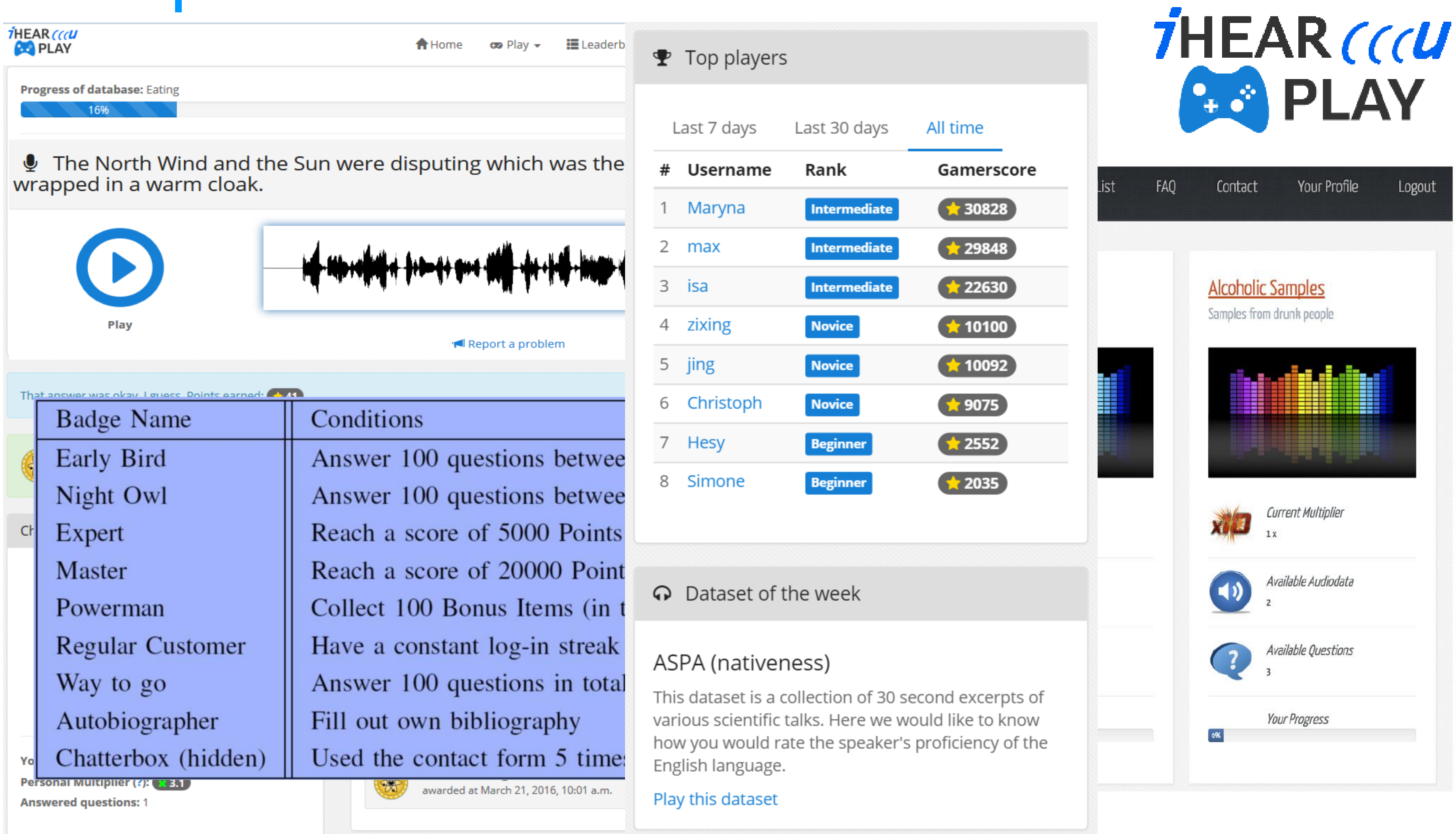
"Trustability-based Dynamic Active Learning for Intelligent Crowdsourcing Applications", submitted.

Rapid Data: AL.



"Trustability-based Dynamic Active Learning for Intelligent Crowdsourcing Applications", submitted.

Rapid Data: AL+CS.



"iHEARu-PLAY: Introducing a game for crowdsourced data collection for affective computing", WASA, 2015.

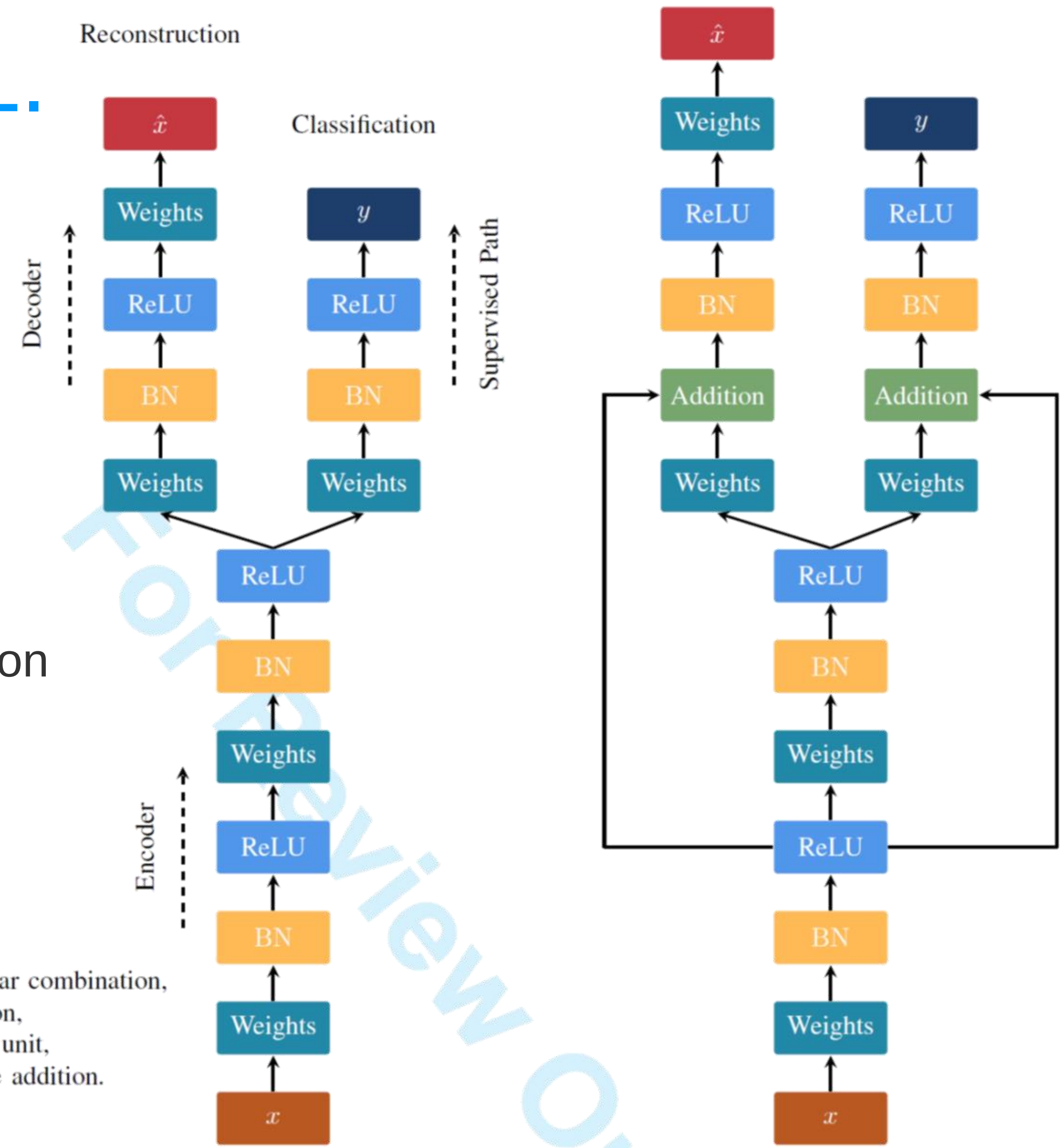
Rapid Data: SSL.

- **AEs for SSL**

  Supervised Learning:
  Keep only relevant info

  Unsupervised AEs:
  Keep all info for reconstruction

  w/o (left) or w/ (right)
  skip compensation



Weights: Weighted linear combination,
BN: Batch normalisation,
ReLU: Rectified linear unit,
Addition: Element-wise addition.

"Semi-Supervised Autoencoders for Speech Emotion Recognition", IEEE Transactions on Audio Speech and Language Processing, 2017.

Fast Transmission.

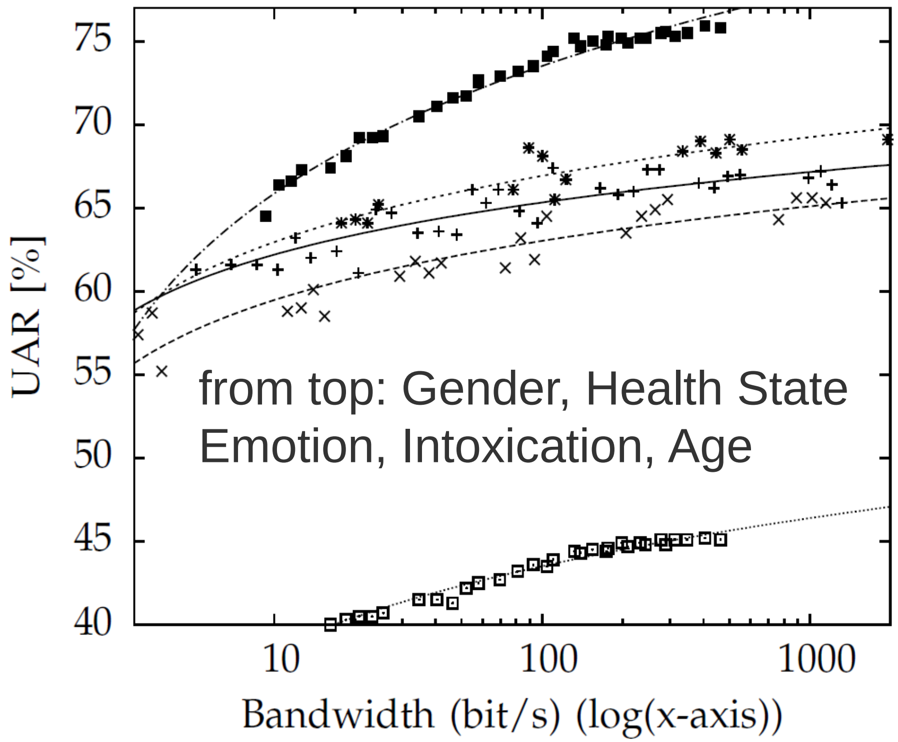
openXBOW –|)→

vector quantisation

1 2 3 4 5 …

2 2 2 1 1 1 1 3 3 1 4 4 5 4

dictionary



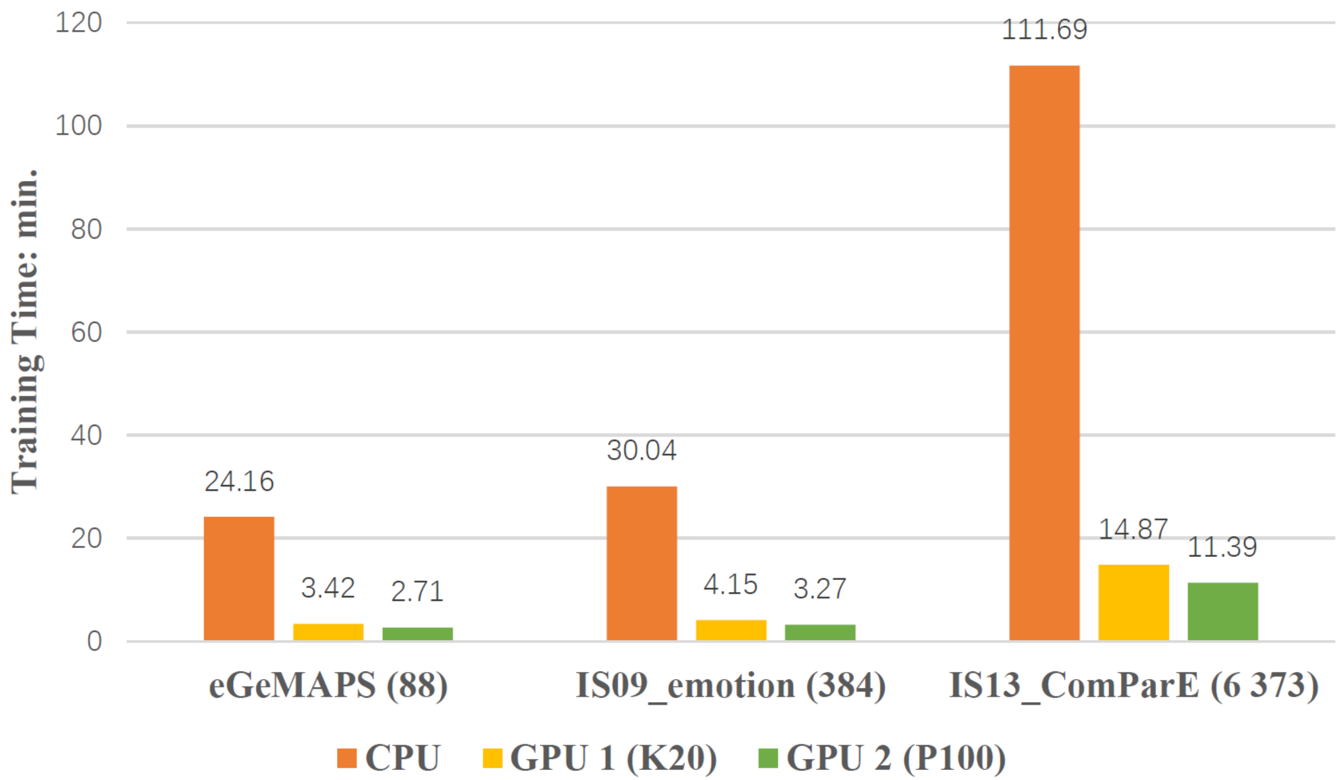
1:5 xxxxx
2:3 xxx
3:2 xx
4:3 xxx
5:1 x
…

histogram

"openXBOW – Introducing the Passau Open-Source Crossmodal Bag-of-Words Toolkit", Journal of Machine Learning Research, 2017.

Fast Processing.

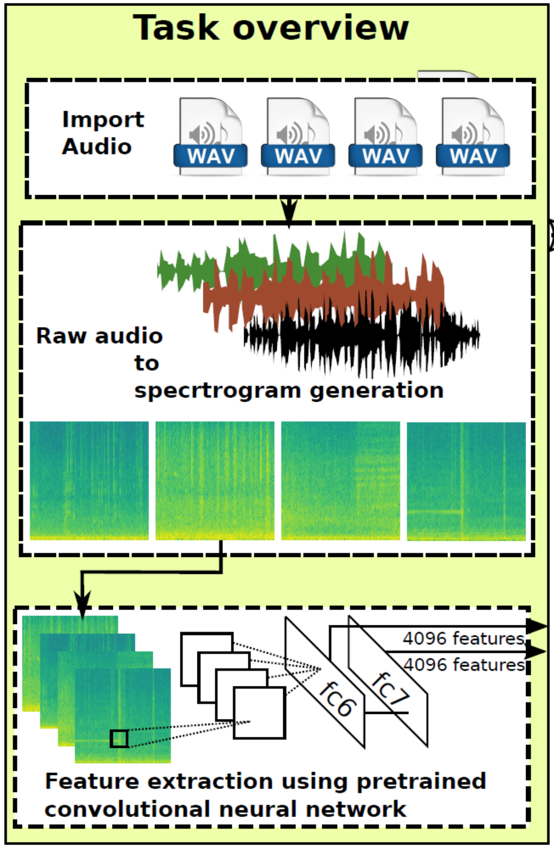
- **GPU feature extraction**

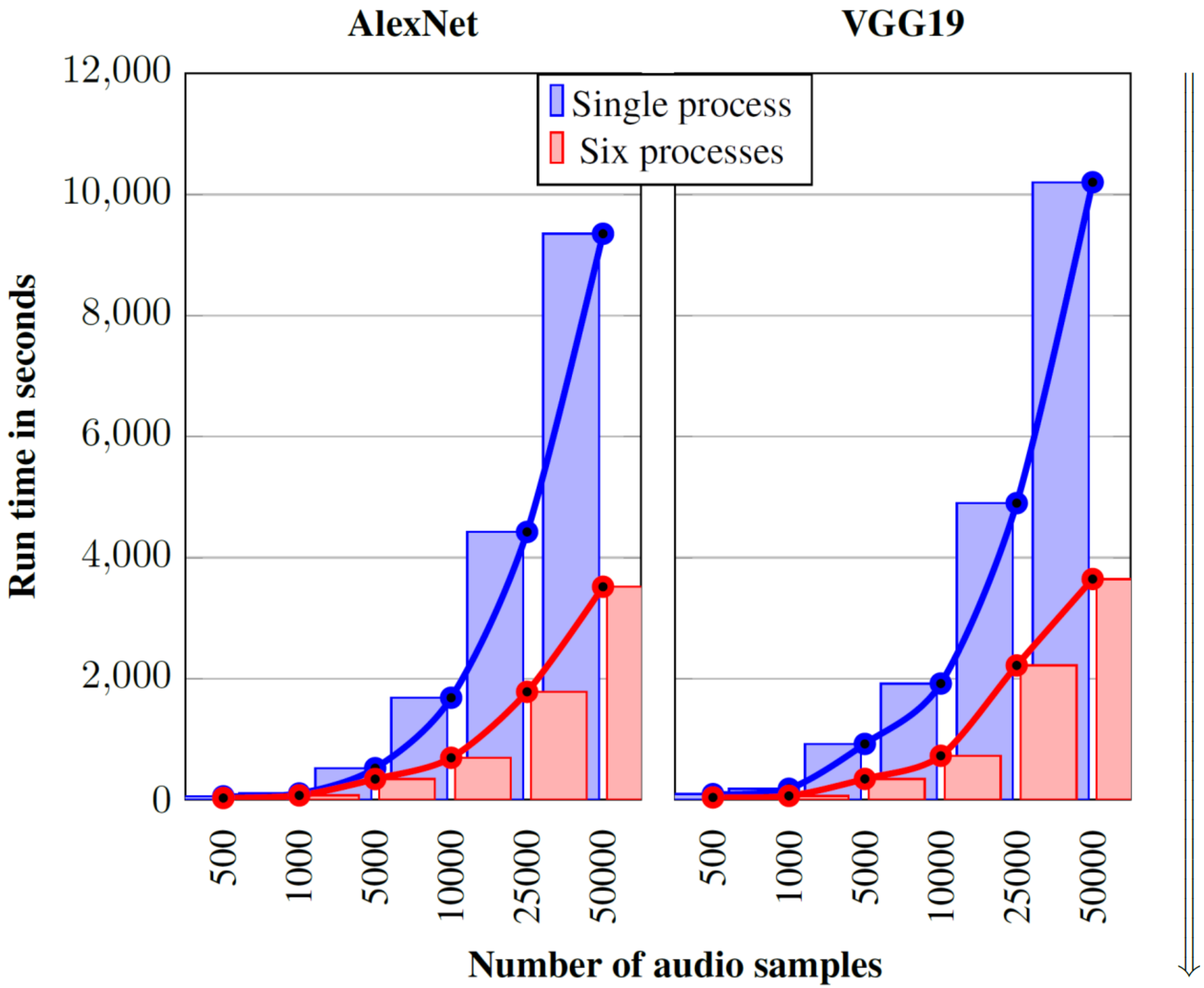


*"GPU-based Training of Autoencoders for Bird Sound Data Processing", **IEEE ICCE-TW**, 2017.*

Fast Processing.

- **Parallel…**

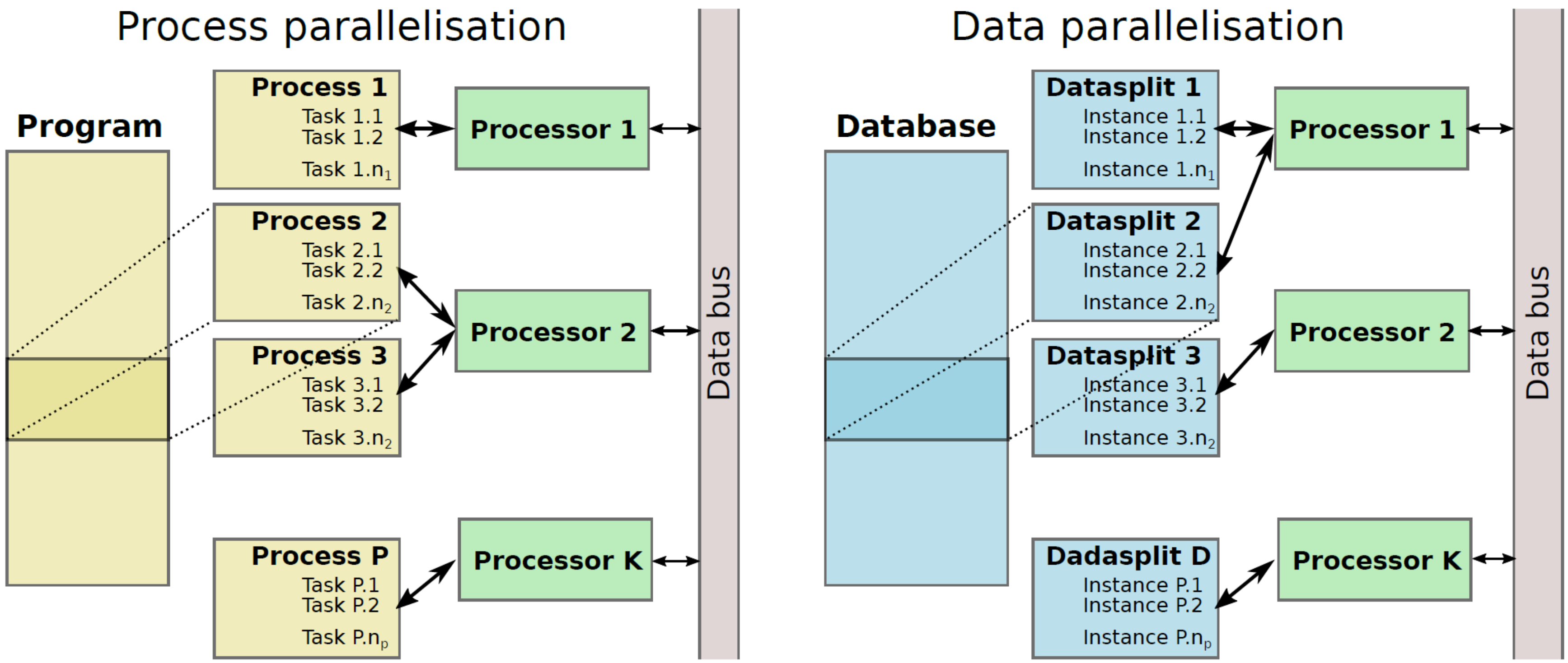




"Big Data Multimedia Mining: Feature Extraction facing Volume, Velocity, and Variety", to appear.

Fast Processing.

- **Parallelisation**



"Big Data Multimedia Mining: Feature Extraction facing Volume, Velocity, and Variety", to appear.

Application?

Speaker Verification.

- **Robstness against Paralinguistics?**
- **Example: Alcohol Intoxication**
  - Negative influence
  - Improvement by multi-condition training
  - Larger effect for female speakers

UBM	Tgt	True	Imp.	EER
S	S	S	S	8.1
S	S	A	S	12.9
S	S	A	A	12.3
S	A	S	S	10.9
S	A	A	S	8.1
S	A	A	A	7.9

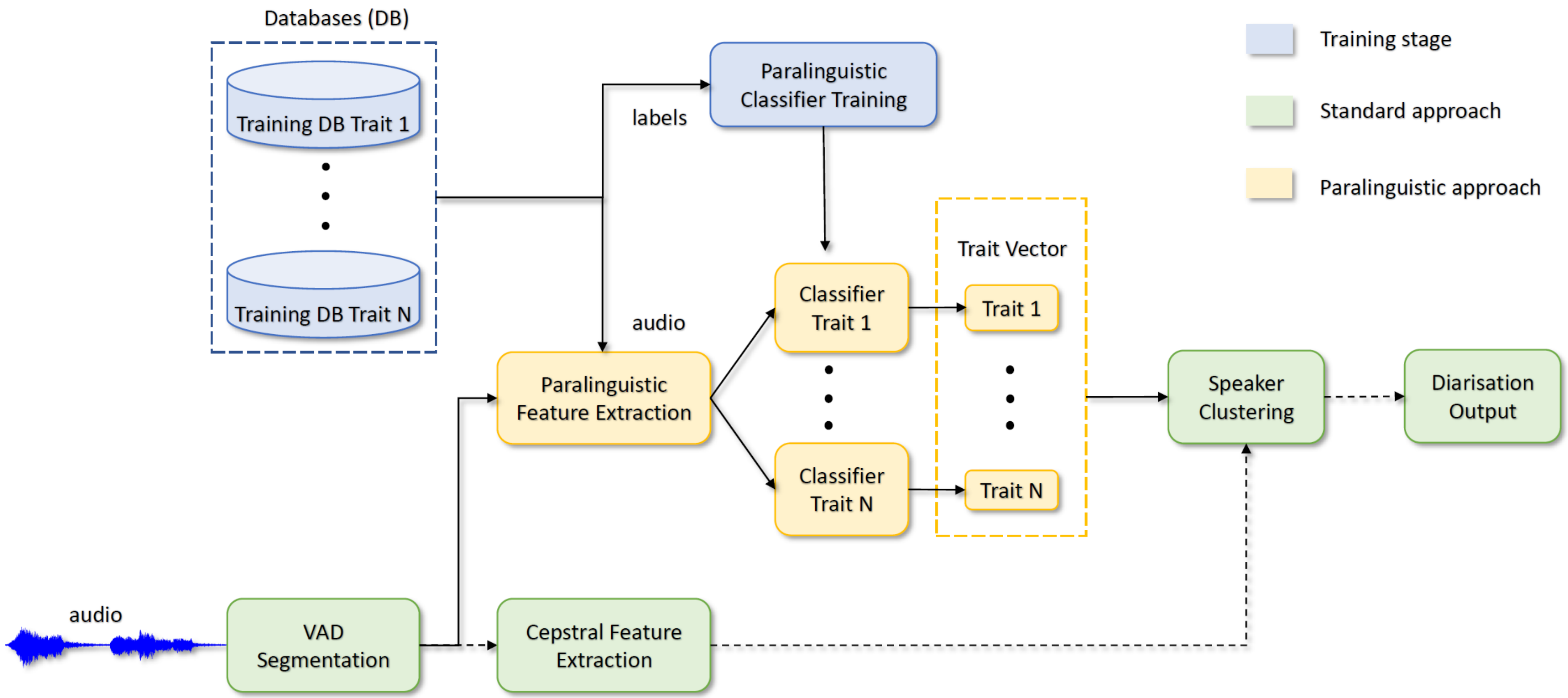
	sober	alcoholised
Spontaneous		
Command & control		

"On the Influence of Alcohol Intoxication on Speaker Recognition," AES, 2014.

Diarisation.

System	Miss	sperr
LIUM	6.3	39.0
sensAI	15.2	23.4
Paralings	6.3	38.0

- **Paralings for Diarisation**
  SEWA database



"A Paralinguistic Approach To Holistic Speaker Diarisation", ACM Multimedia, 2017.

So?

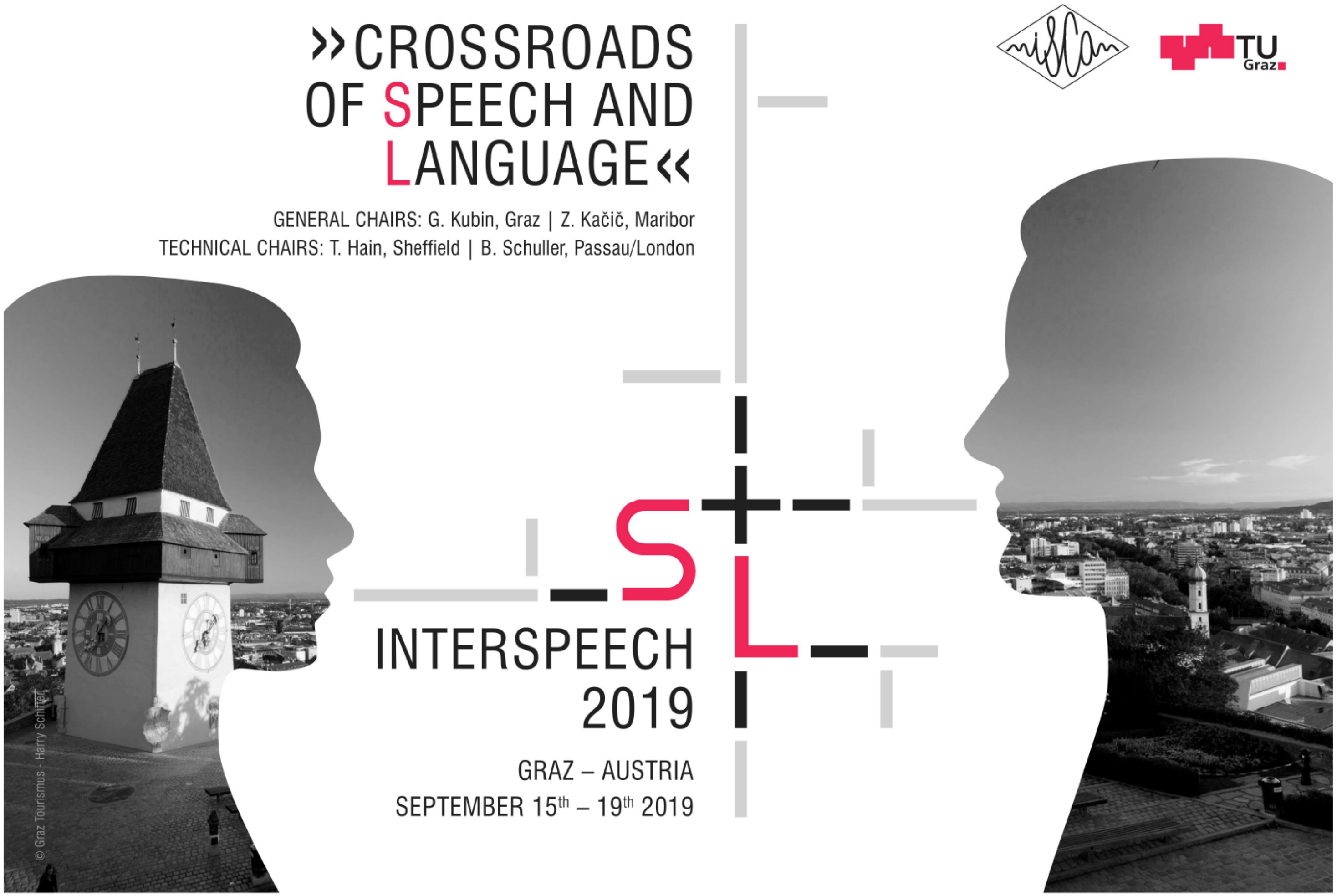
- **Superhuman in several objective tasks**
- **More (independent) perception studies for subjective tasks!**
- **Increased realism and performance (up to 2x)**
- **Good progress by improved Deep Architectures**
- **Still even many low hanging fruits!**

Vision.

- Tighter Coupling w/ Synthesis
- Embedding in Dialogues
- Reinforcement Learning
- NPU optimized Solutions

Thank You.

Events.


»CROSSROADS OF SPEECH AND LANGUAGE«
GENERAL CHAIRS: G. Kubin, Graz | Z. Kačič, Maribor
TECHNICAL CHAIRS: T. Hain, Sheffield | B. Schuller, Passau/London
TU Graz
INTERSPEECH 2019
GRAZ – AUSTRIA
SEPTEMBER 15th – 19th 2019
© Graz Tourismus - Harry Schiffer

Events.


CALL FOR PAPERS - ACII2019, Cambridge, UK

8th International Conference on Affective Computing and Intelligent Interaction

3-6 September, 2019

www.acii2019.org

Books.


Computational Social Sciences
Alexandra Balahur-Dobrescu
Maite Taboada
Björn W. Schuller
Computational Methods for Affect Detection from Natural Language
Springer
BJÖRN SCHULLER | AN
COMPUTATIO
PARALINGUIS
EMOTION, AFFECT AND PERS
SPEECH AND LANGUAGE P
WILEY
Signals and Communication Technology
Björn Schuller
Intelligent Audio Analysis
The Handbook of Multimodal-Multisensor Interfaces
Volume 1
Foundations, User Modeling, and Common Modality Combinations
Sharon Oviatt
Björn Schuller
Philip Cohen
Daniel Sonntag
Gerasimos Potamianos
Antonio Krüger

Abstract & CV

Human performance is often appearing as a glass ceiling when it comes to automatic speech and speaker analysis. In some tasks, such as health monitoring, however, automatic analysis has successfully started to break this ceiling. The field has benefited from more than a decade of deep neural learning approaches such as recurrent LSTM nets and deep RBMs by now; however, recently, a further major boost could be witnessed. This includes the injection of convolutional layers for end-to-end learning, as well as active and autoencoder-based transfer learning and generative adversarial network topologies to better cope with the ever-present bottleneck of severe data scarcity in the field. At the same time, multi-task learning allowed to broaden up on tasks handled in parallel and include the often met uncertainty in the gold standard due to subjective labels such as emotion or perceived personality of speakers. This talk highlights the named and further latest trends such as increasingly deeper nets and the usage of deep image nets for speech analysis on the road to 'holistic' superhuman speech analysis 'seeing the whole picture' of the person behind a voice. At the same time, increasing efficiency is shown for an ever 'bigger' data and increasingly mobile application world that requires fast and resource-aware processing. The exploitation in ASR and SLU is featured throughout.

Björn W. Schuller heads Imperial College London's/UK Group on Language Audio & Music (GLAM), is a CEO of audEERING, and a Full Professor at University of Augsburg/Germany in CS. He further holds a Visiting Professorship at the Harbin Institute of Technology/China. He received his diploma, doctoral, and habilitation degrees from TUM in Munich/Germany in EE/IT. Previous positions of his include Visiting Professor, Associate, and Scientist at VGTU/Lithuania, University of Geneva/Switzerland, Joanneum Research/Austria, Marche Polytechnic University/Italy, and CNRS-LIMSI/France. His 650+ technical publications (15000+ citations, h-index 59) focus on machine intelligence for audio and signal analysis. He is the Editor in Chief of the IEEE Transactions on Affective Computing, a General Chair of ACII 2019, and a Technical Chair of Interspeech 2019 among various further roles.