Big Data, Deep Learning at the Edge of x-Ray Speaker Analysis

SPECOM / ICR 2017

Björn W. Schuller

Imperial College London
Universität Passau
audEERING intelligent Audio Engineering

Data?

- **2.0 Yet?**

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

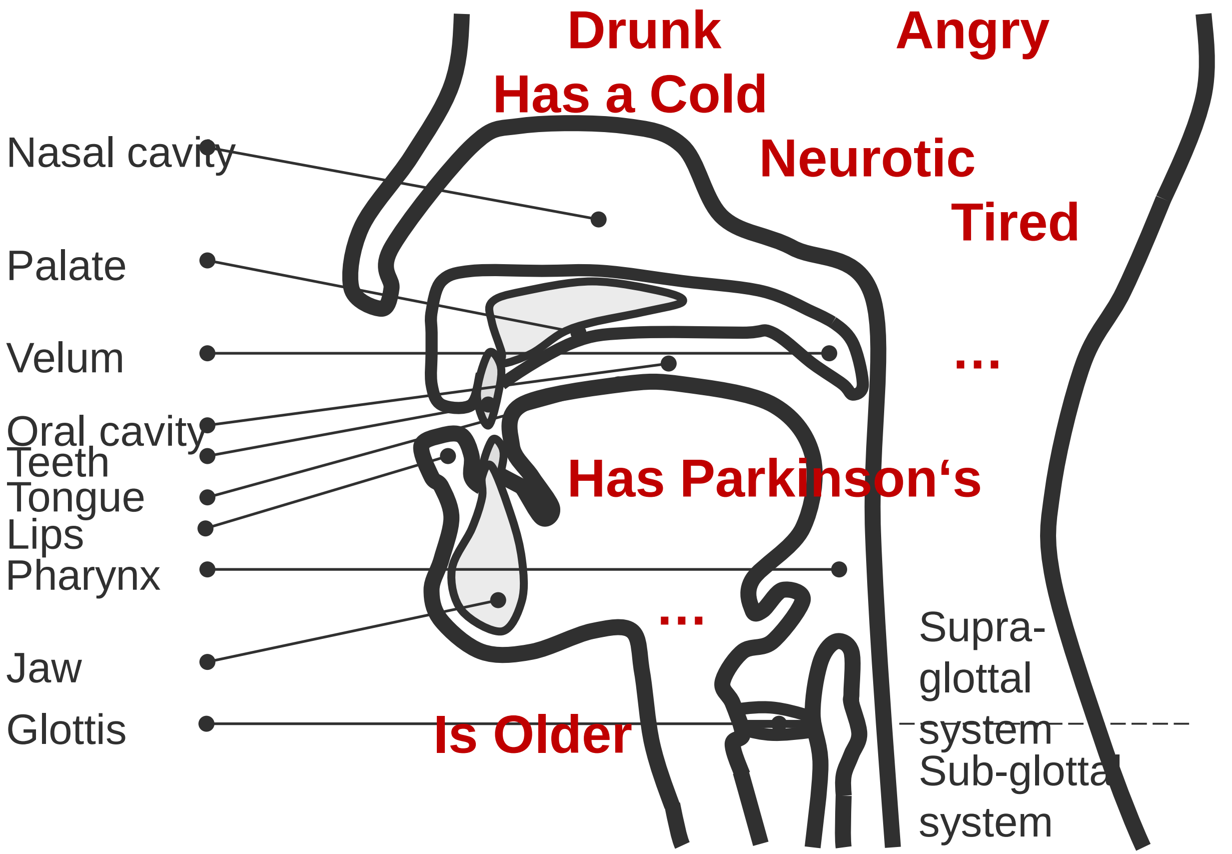
R. Moore, "A Comparison of the Data Requirements of Automatic Speech Recognition Systems and Human Listeners", 2003.

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

Holism.

- **Multiple-Targets**

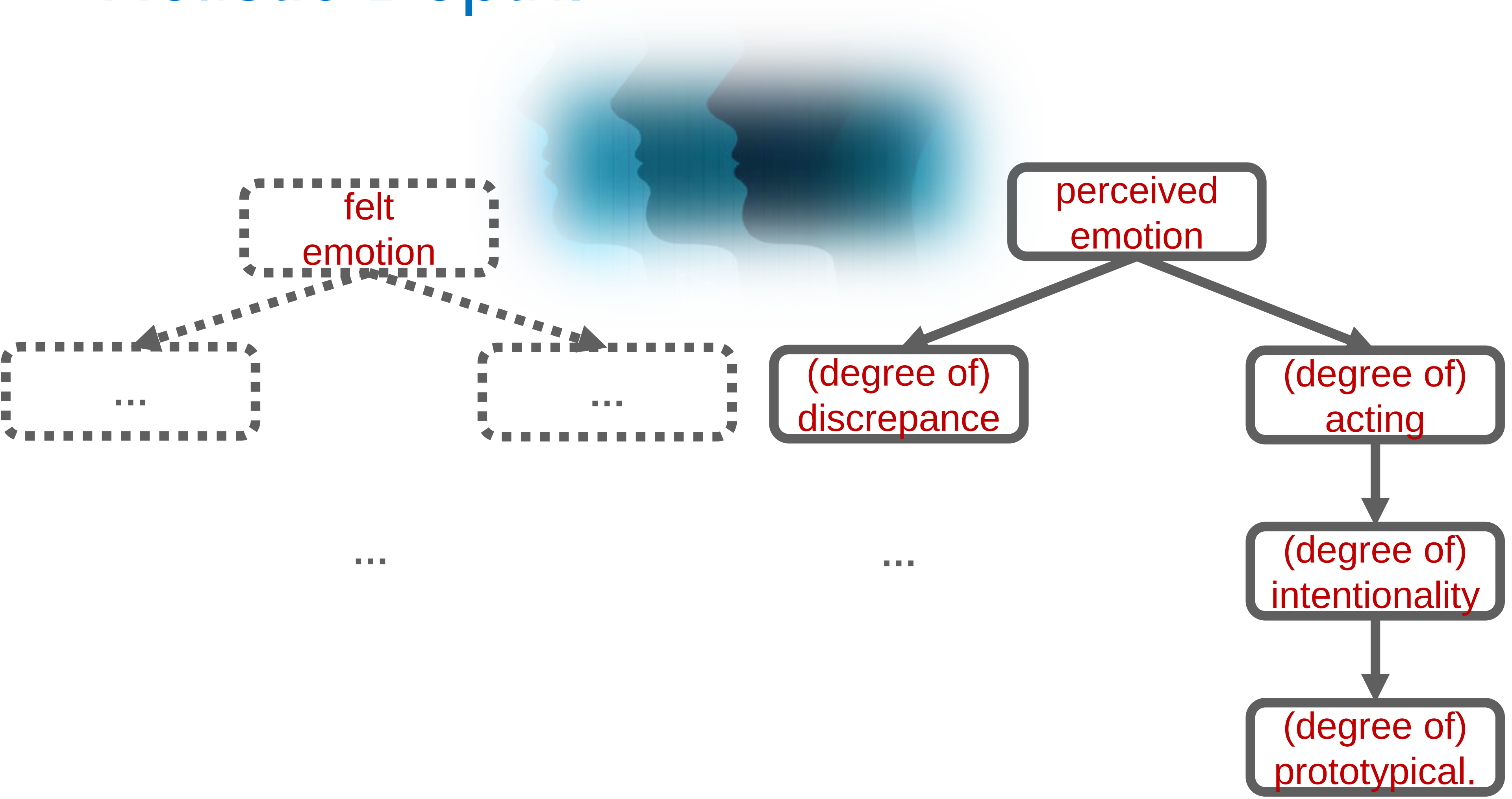
- **1 Voice**



Depth.

state						trait
spontaneous						acted
complex						simple
measured						assessed
continuous						categorical
felt						perceived
intentional						instinctual
consistent						discrepant
private						social
prototypical						peripheral
universal						culture-specific
uni-modal						multi-modal

Holistic Depth.



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

Big Data.

Going Larger.

- **Epilepsy, MS & Depression**
  - Big data RMT platform
  - Monitoring sleep, activity, gait, speech, social connectivity, e-health records, …
  - Data visualisation for patients / clinicians
  - Real-time
  - 100k participants in UK

- **Child Language Development**
  - Analyzing Child Language Experiences
  - around the World
  - Child vocalisation maturity
  - Child/Adult directed speech

Big Data.

MixedEmotions

Volume	Velocity	Variety
13 TB / 8300 h	GB/h	video, audio
350 mio tweets/day	real-time	diff. resol. / format
1.3 bio users	crawled	social data feed
130 mio web pages	every 48 h	text

- **Big Vs**

  Volume – e.g., 300 hours videos / min (YouTube, Dec 2014)

  Velocity – e.g., 500 mio Tweets / day (Twitter, Aug 2013)

  Variety – e.g., text, audio, video, sensors, diverse formats

- **Challenges**

  Unstructured, HW limits (data: ~ x2/1.5 years, disk speed: linear…)

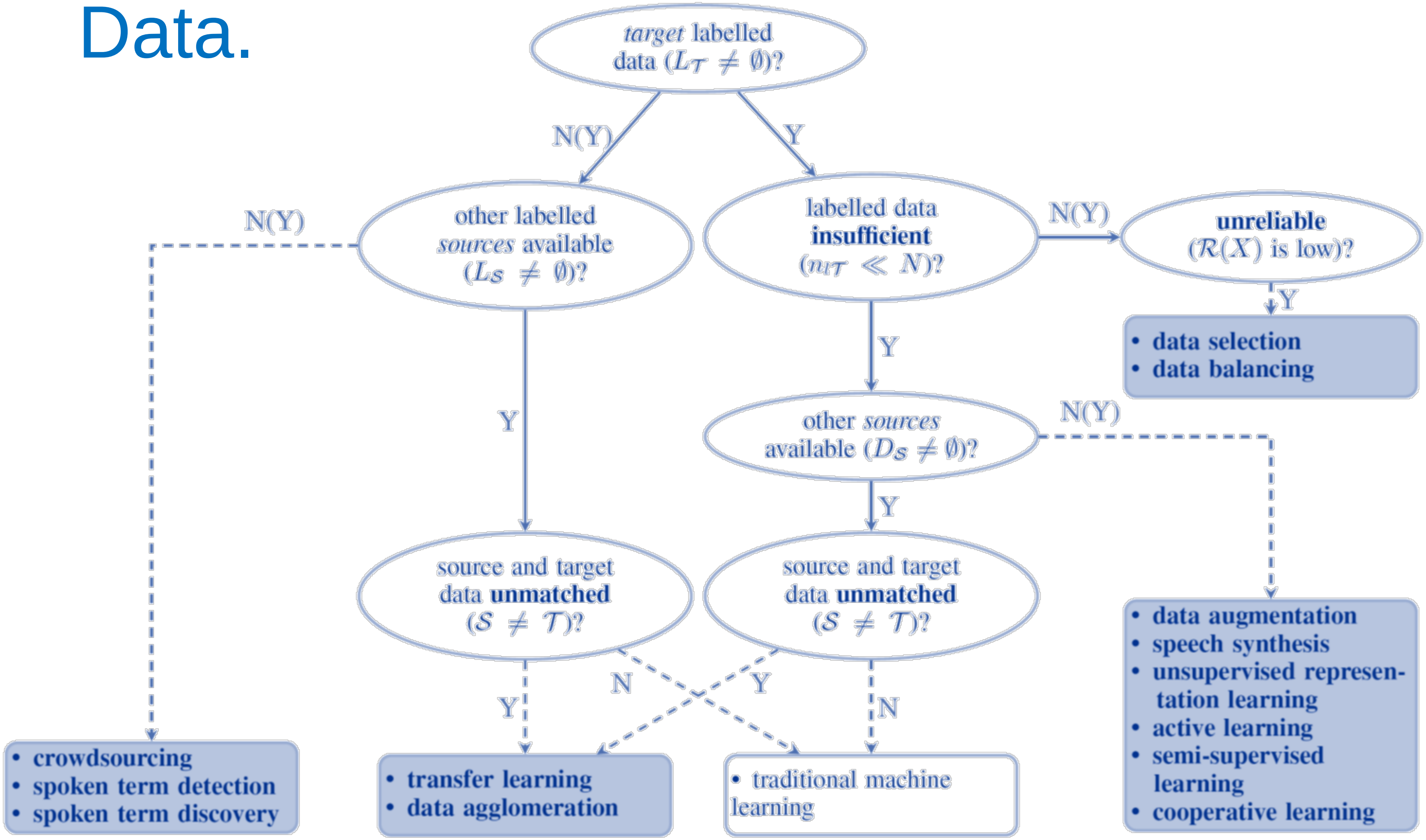
  Scaling, Visualisation, Privacy, Ethics…

- **Chances**

  Parallelisation (GPGPUs, multicore, etc.)

  Distribution (Cloud MapReduce, Disco, Hadoop, Skynet, etc.)

Data.

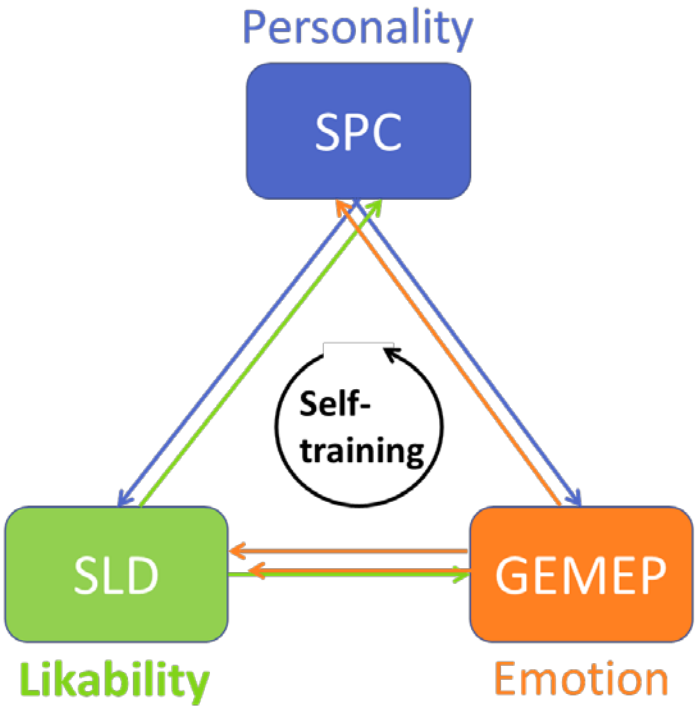


"Efficient Data Exploration for Automatic Speech Analysis: Challenges and State of the Art", IEEE Signal Processing Magazine, 2017.

Cross-Task.

- **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 $\mathcal{U} \in \{\}$:

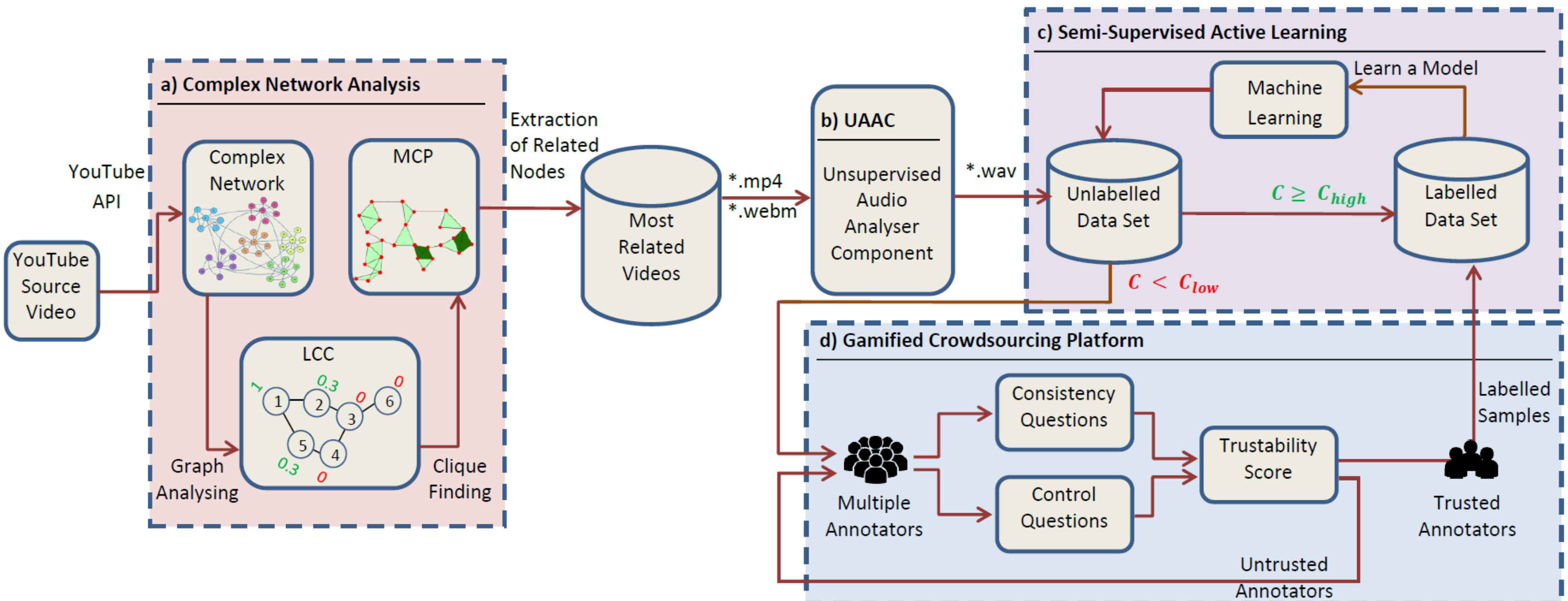
1. (Optional) Upsample training set $\mathcal{L}$ to even class distribution $\mathcal{L}_D$
2. Use $\mathcal{L}/\mathcal{L}_D$ to train classifier $\mathcal{H}$, then classify $\mathcal{U}$
3. Select a subset $\mathcal{N}_{st}$ that contains those instances predicted with the highest confidence values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

Big Data?

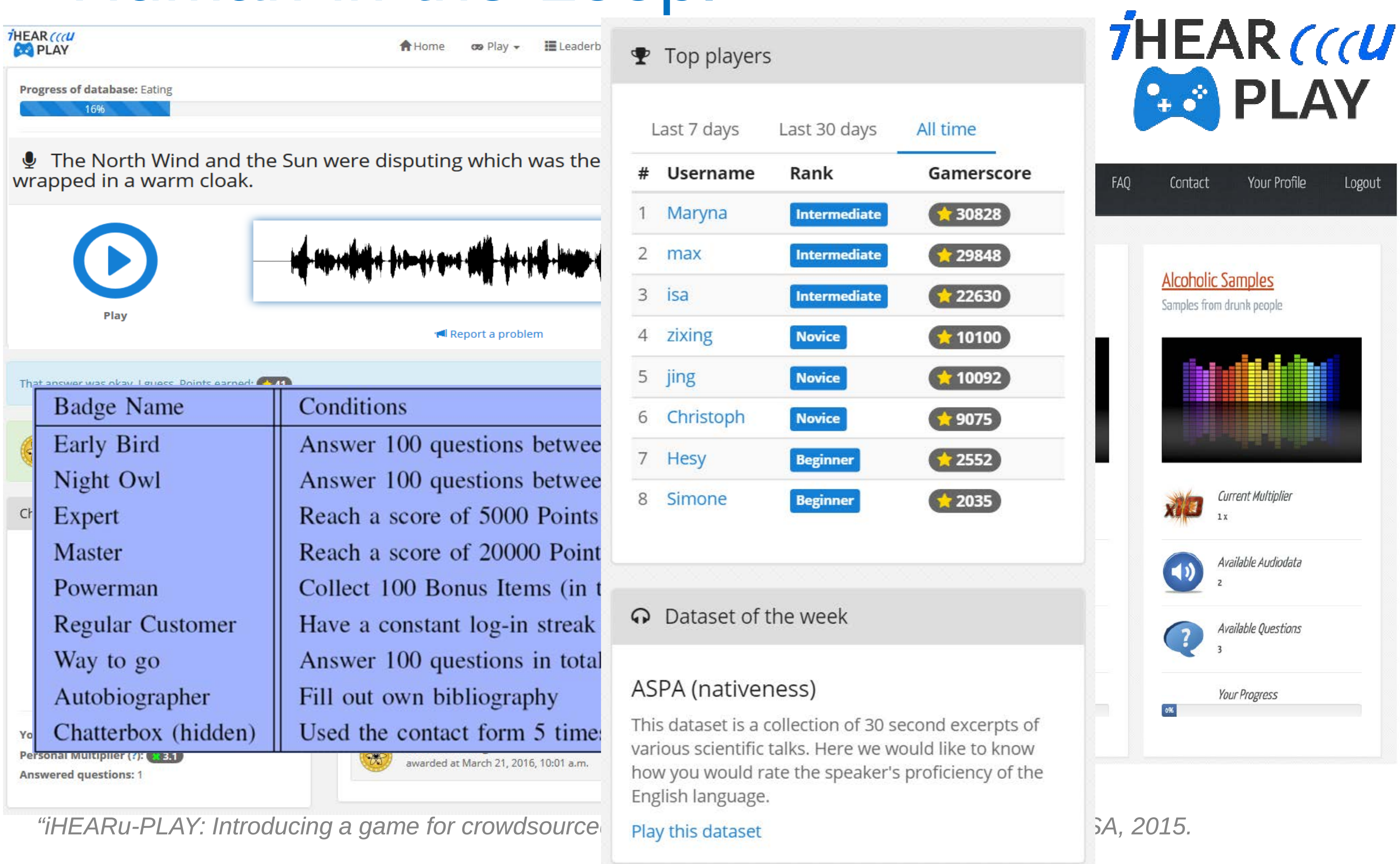
- **Targeted Data Acquisition**
  - Small World Modelling: find highly related videos
  - Local Clustering Coeff.+Maximum Clique Problem
  - Example: 3k videos for rapid training of new tasks

Task	%UA	BEST
Freezing	70.2	func.
Coughing	97.6	BoAW
Sneezing	85.2	NN
Intoxicat.	72.6	BoAW



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

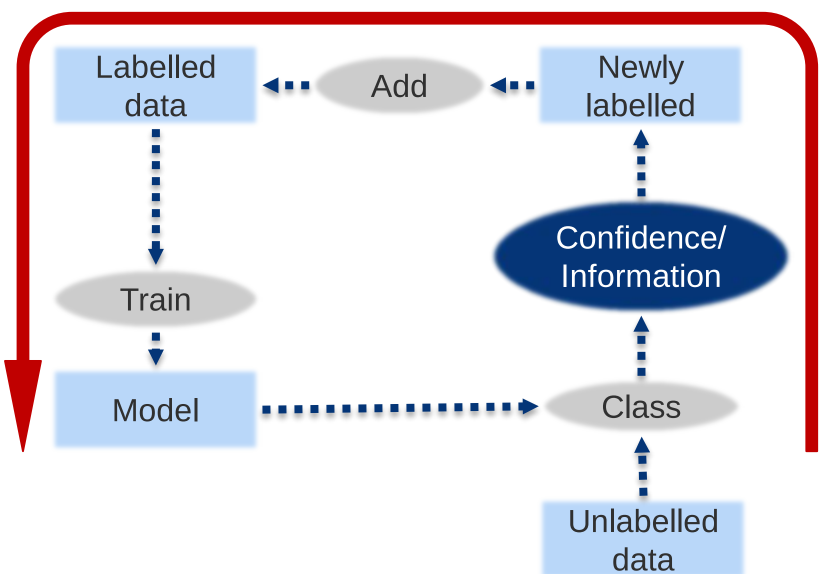
Human-in-the-Loop.

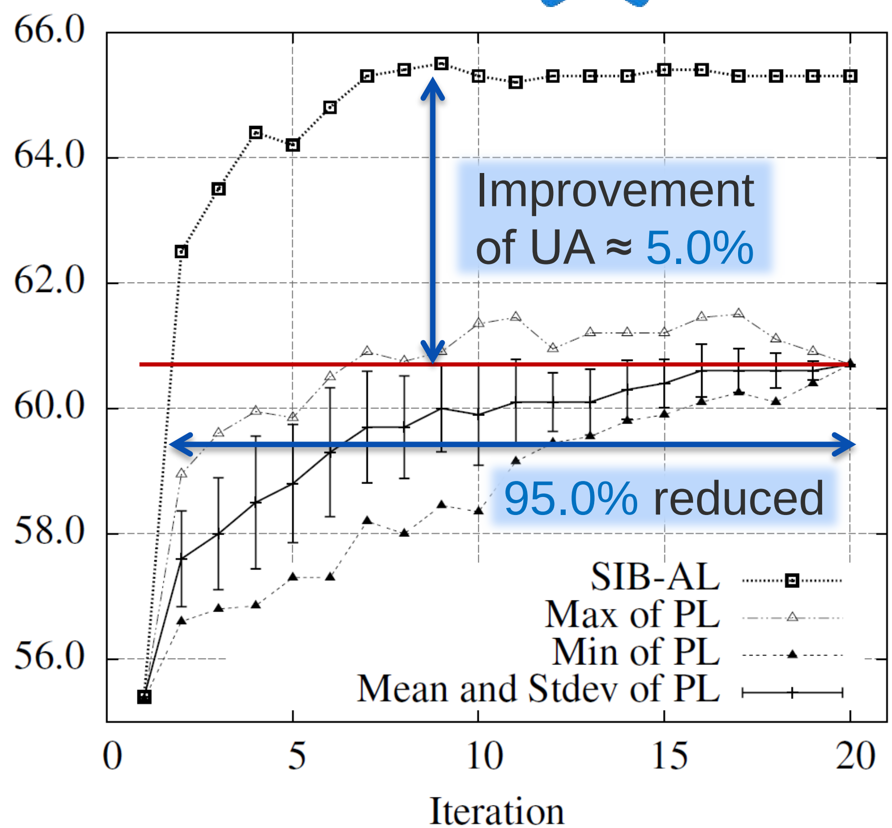


"iHEARu-PLAY: Introducing a game for crowdsource … SA, 2015.

Big Data?

- **Cooperative Learning in aRMT**
  0) Transfer Learning
  1) Dynamic Active Learning
  2) Semi-Supervised Learning

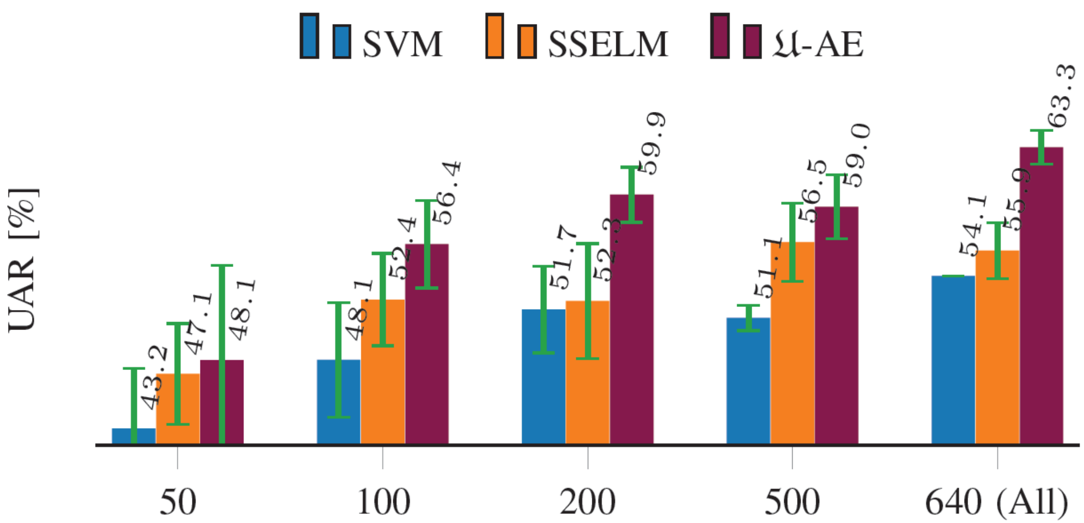


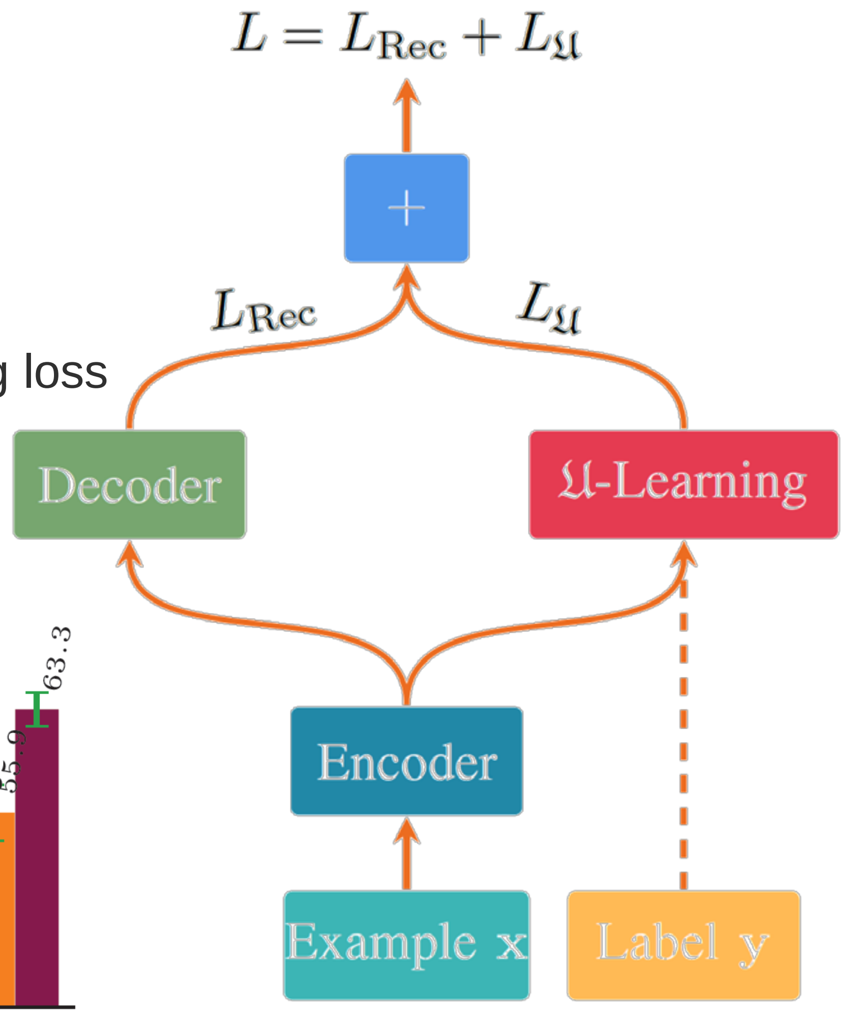


"Cooperative Learning and its Application to ESR", IEEE Transactions on Audio Speech and Language Processing, 2015.

TL.

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

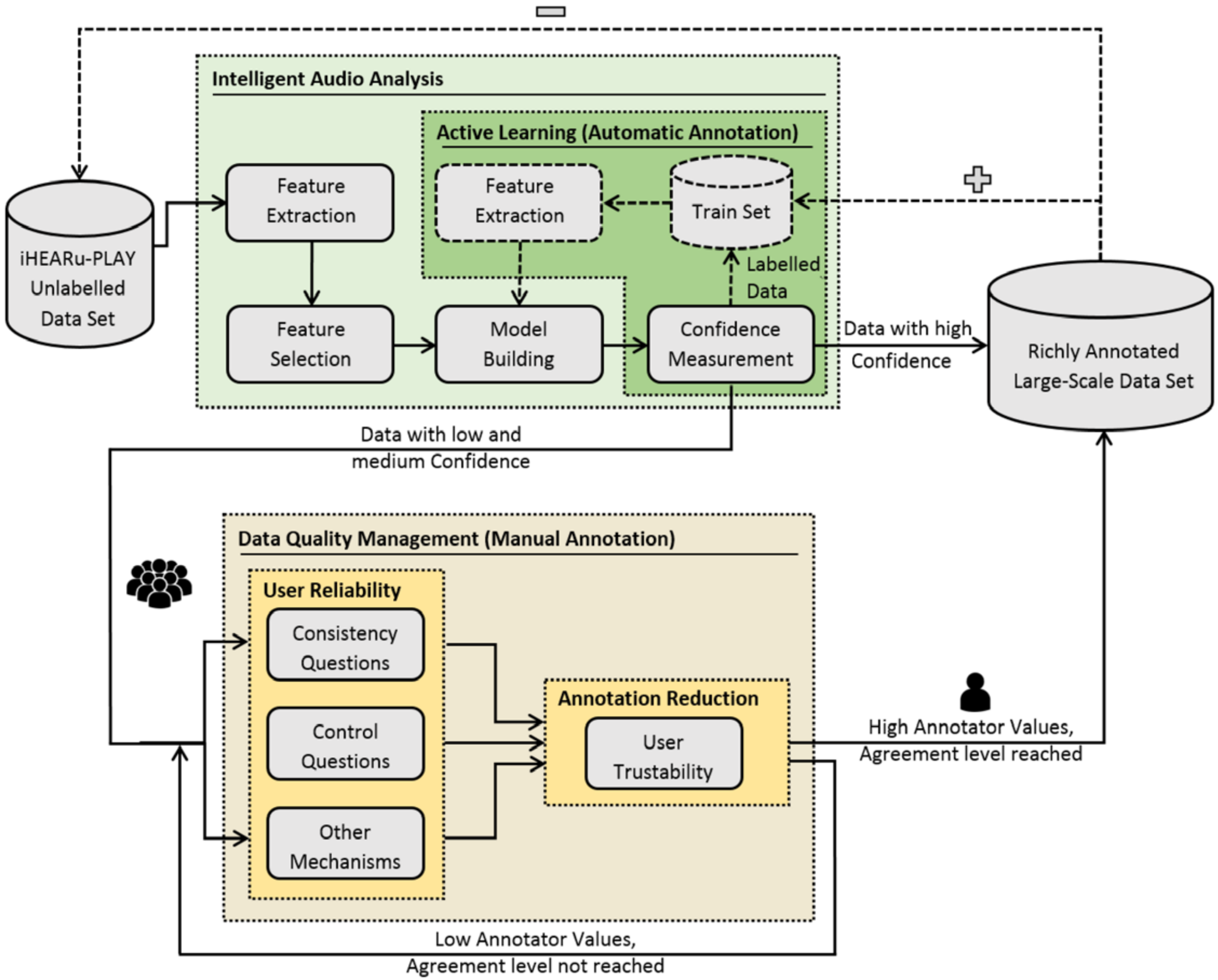




*"Universum Autoencoder-based Domain Adaptation for Speech Emotion Recognition", **Signal Processing Letters**, 2017.*

AL.

- **Reality…**

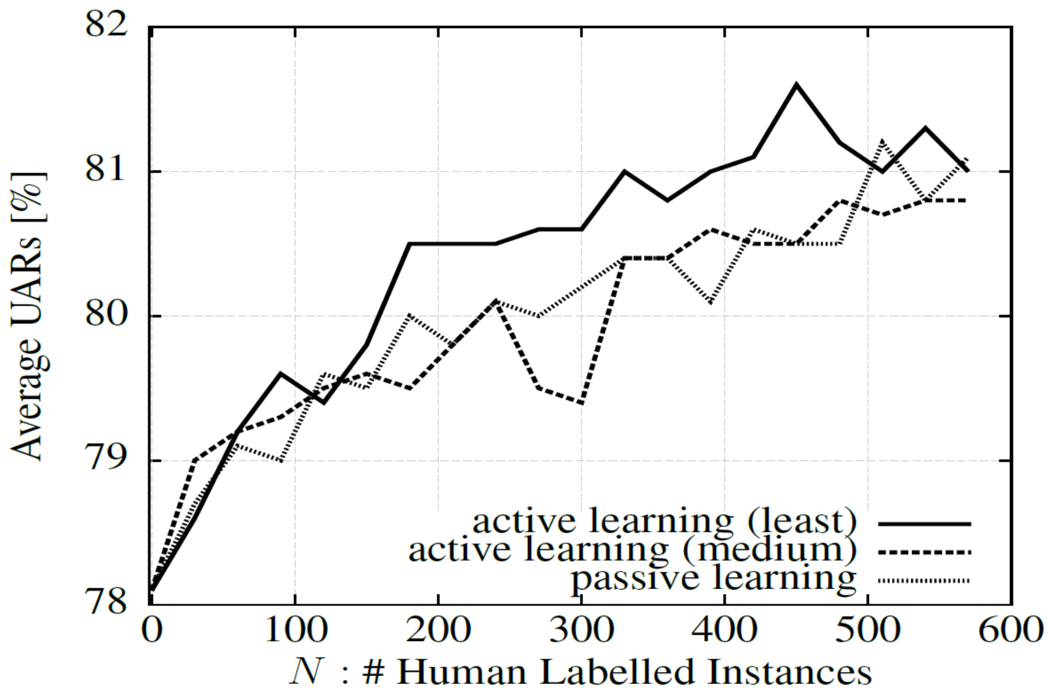


"Towards Intelligent Crowdsourcing for Audio Data Annotation: Integrating Active Learning in the Real World", Interspeech 2017.

AL.

- **Reality…**

  FAU AIBO Arousal AL



(a) *arousal*

"Towards Intelligent Crowdsourcing for Audio Data Annotation: Integrating Active Learning in the Real World", Interspeech 2017.

SSL.

- **AEs for SSL**

  Supervised Learning:

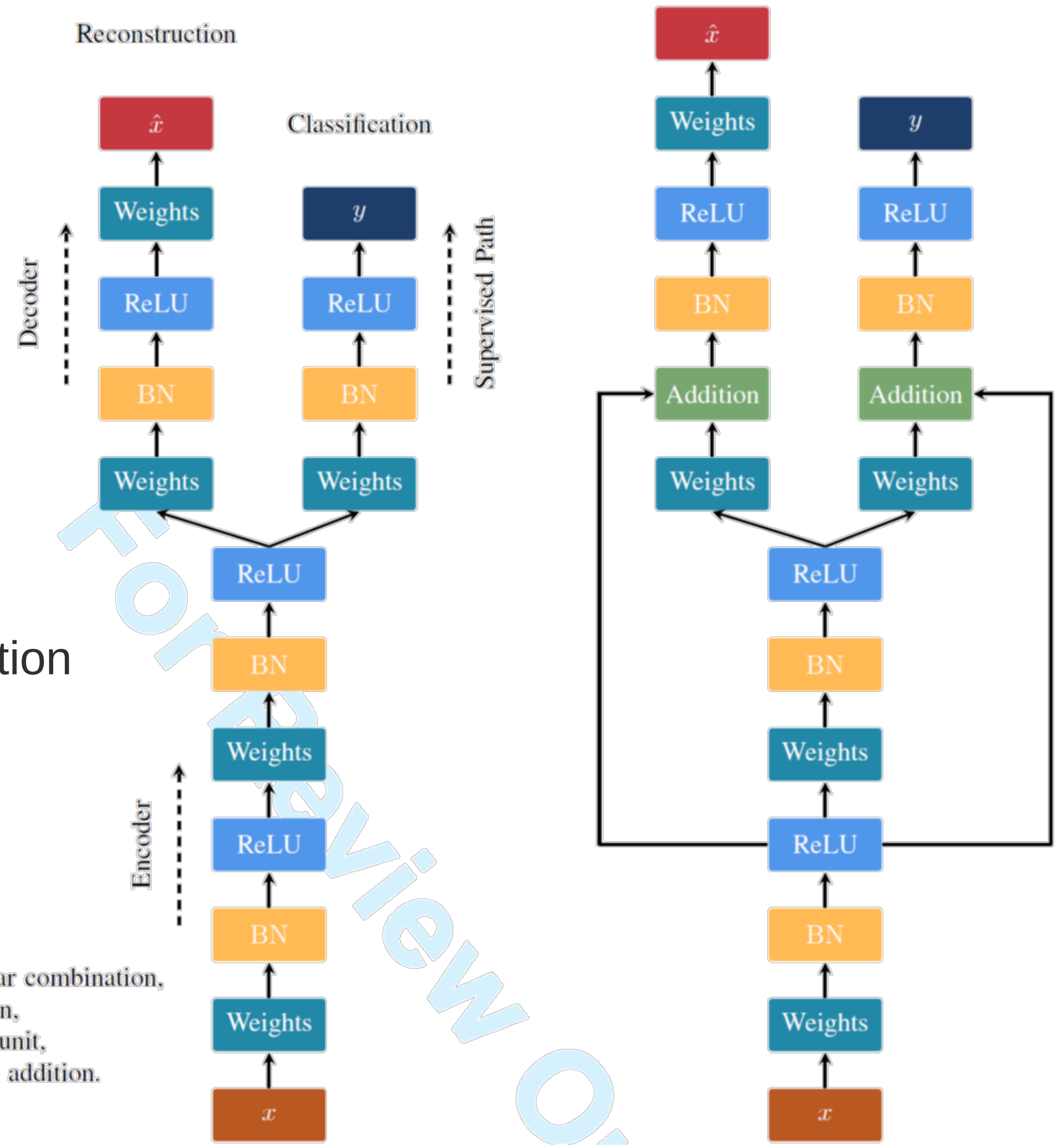
  Keep only relevant info

  Unsupervised AEs:

  Keep al info for reconstruction

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

  skip compensation

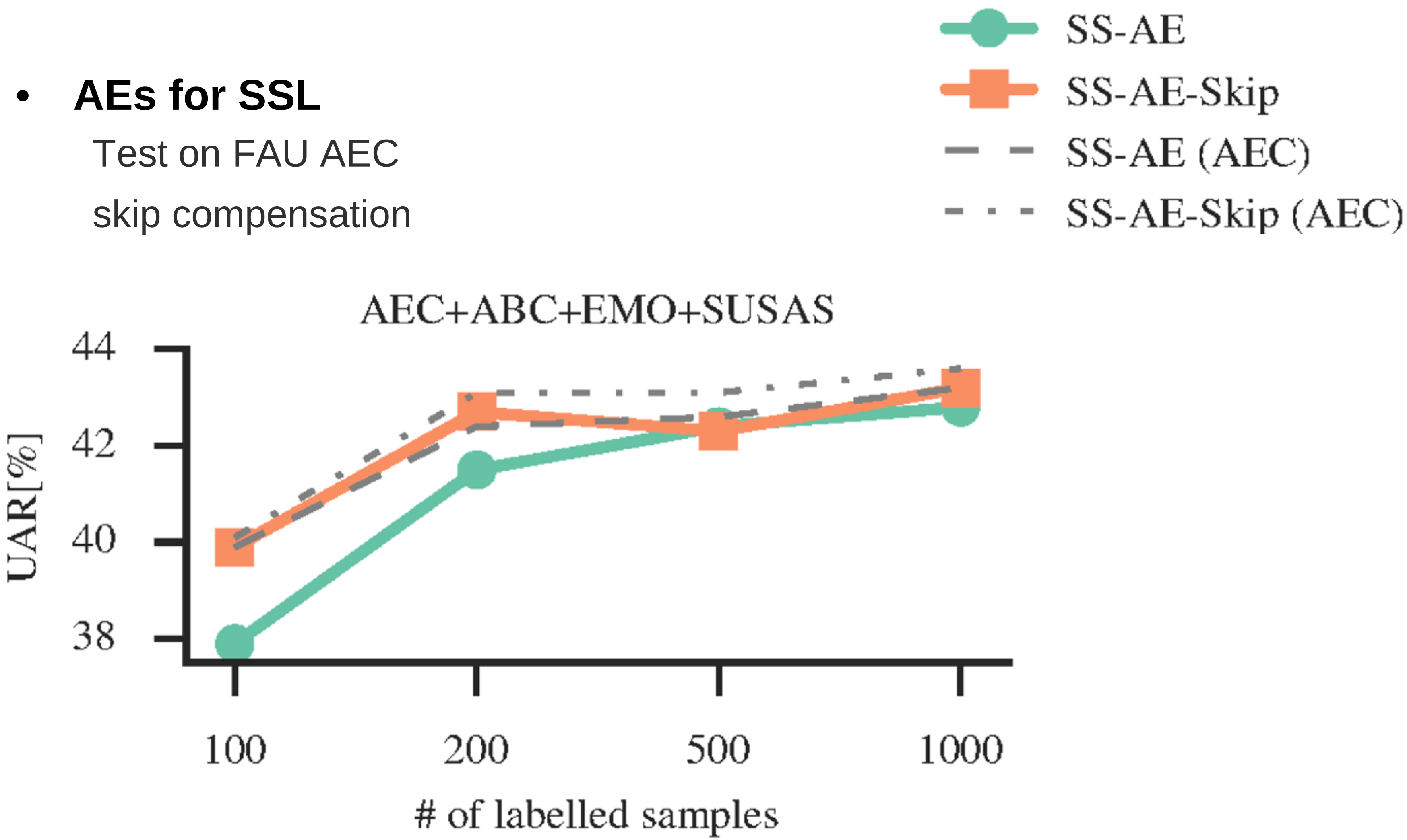


Weights: Weighted linear combination,
BN: Batch normalisation,
ReLU: Rectified linear unit,
Addition: Element-wise addition.

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

SSL.

- **AEs for SSL**
  Test on FAU AEC
  skip compensation



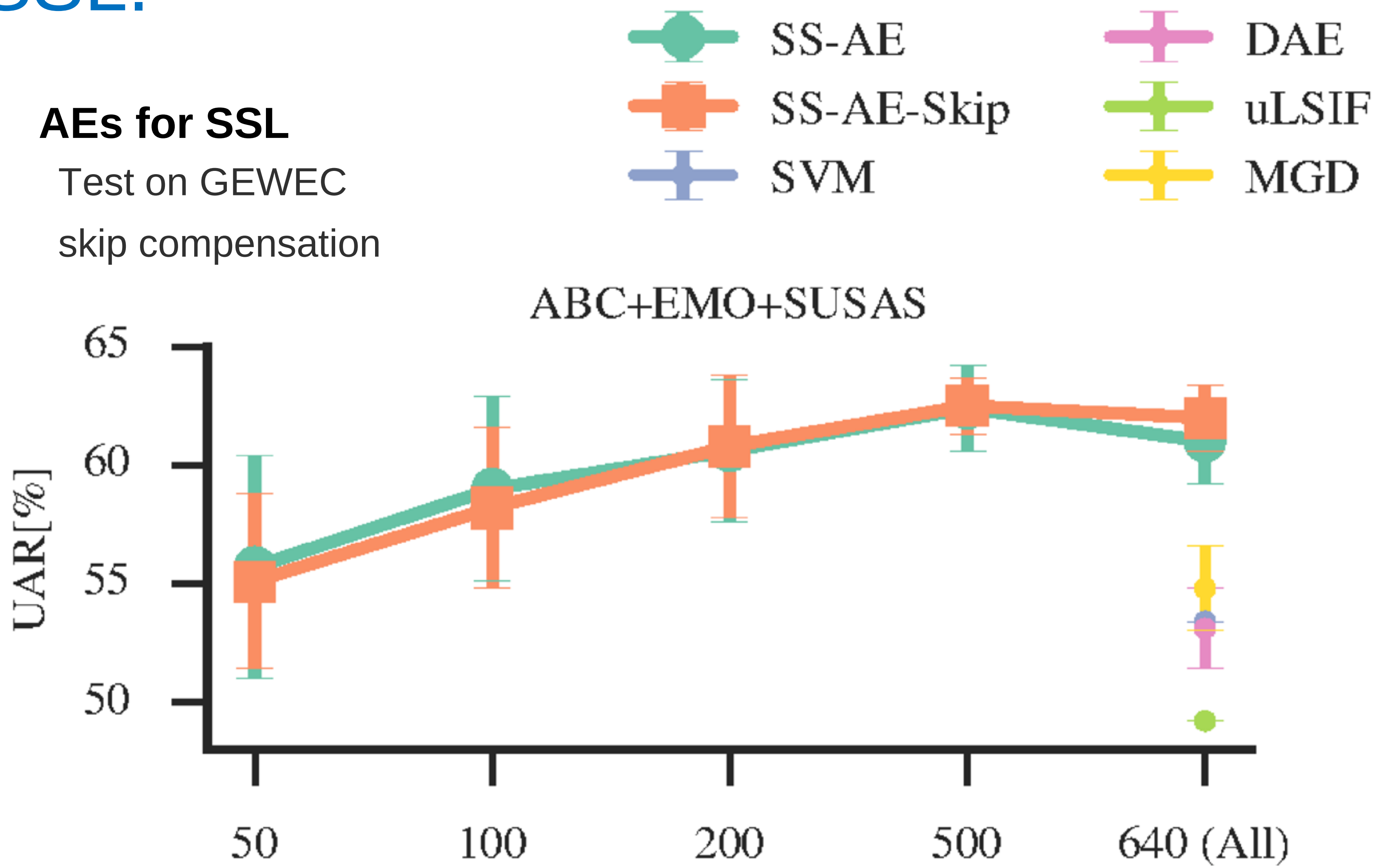
"Semi-Supervised Autoencoders for Speech Emotion Recognition", IEEE Transactions on Audio Speech and Language Processing, 2017.

SSL.

- **AEs for SSL**

  Test on GEWEC

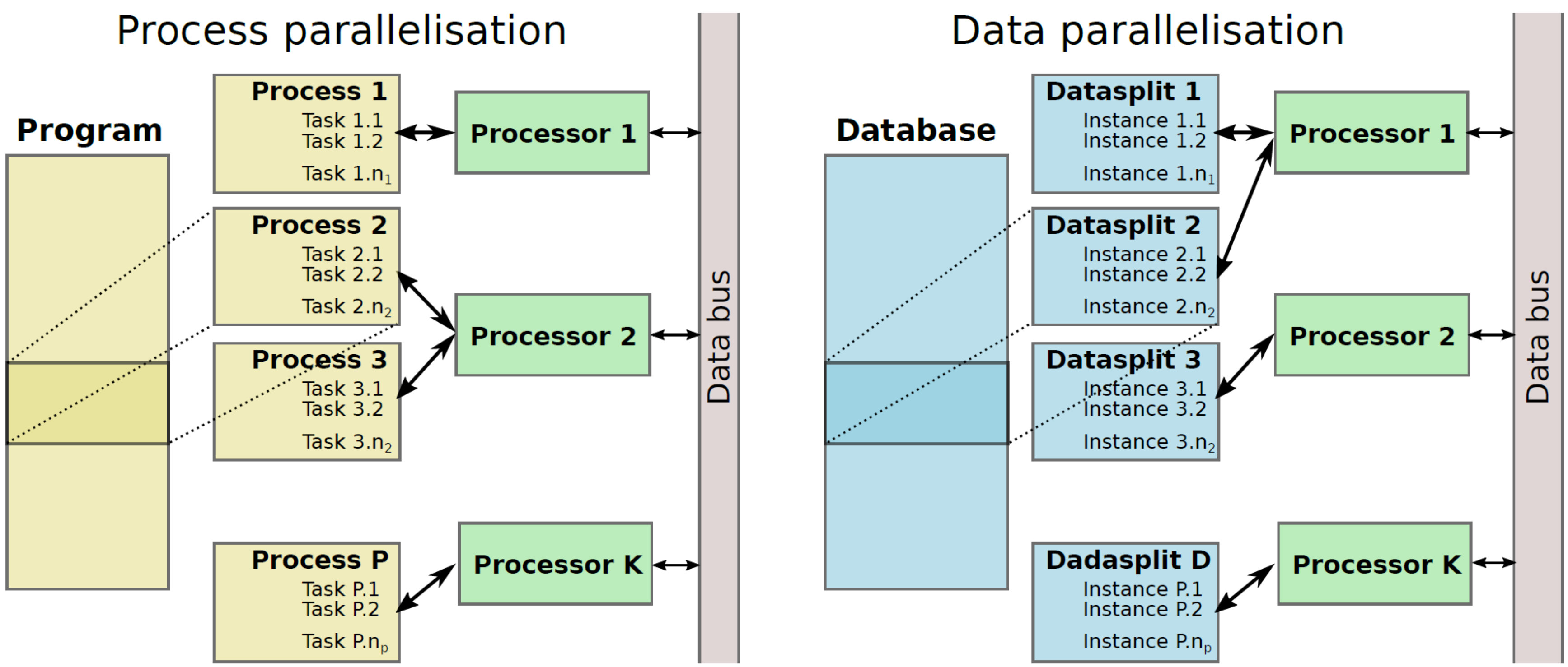
  skip compensation



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

Big Data: Velocity.

- **Parallelisation**



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

Big Data: Velocity.

RTF	CPU	GPU
Double	.522	.068
Single	.937	.033

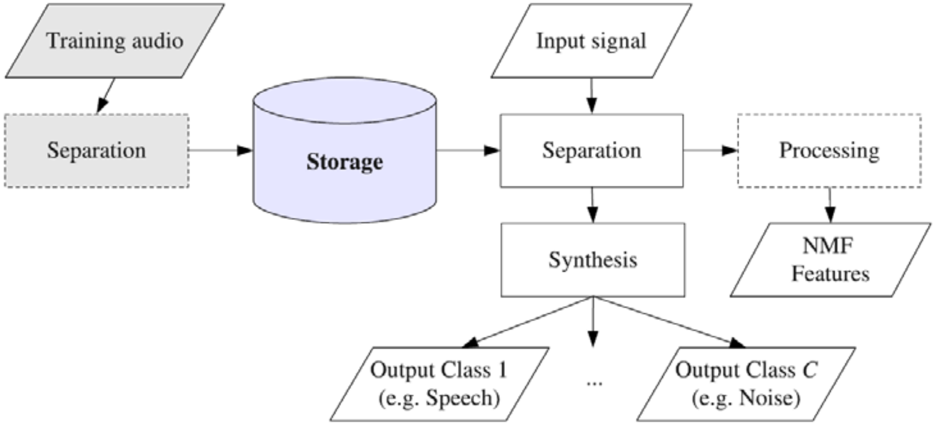
- **GPU-Preprocessing**

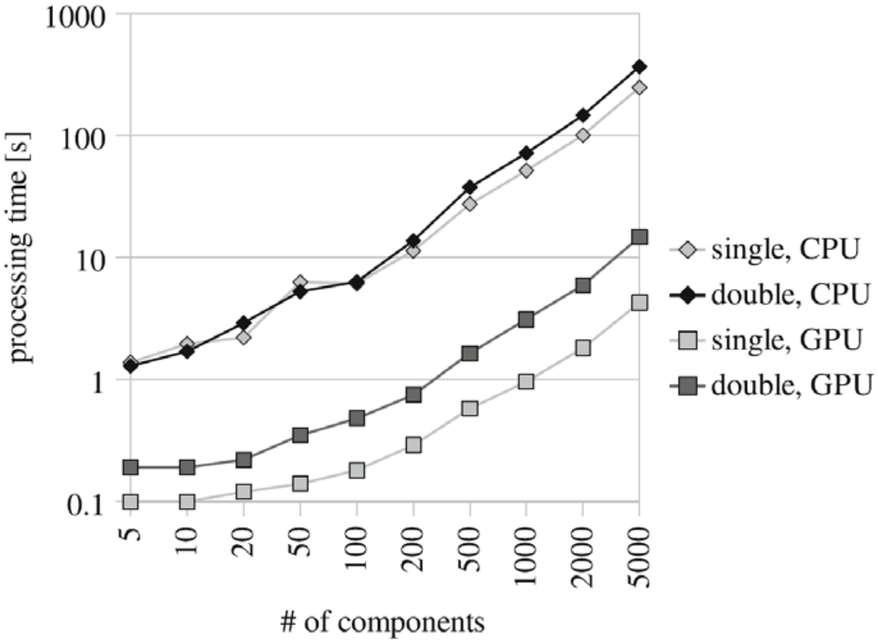
  Parallel NMF Source Separation

  500 x 1000 matrix

  KL divergence

  openBliSSART

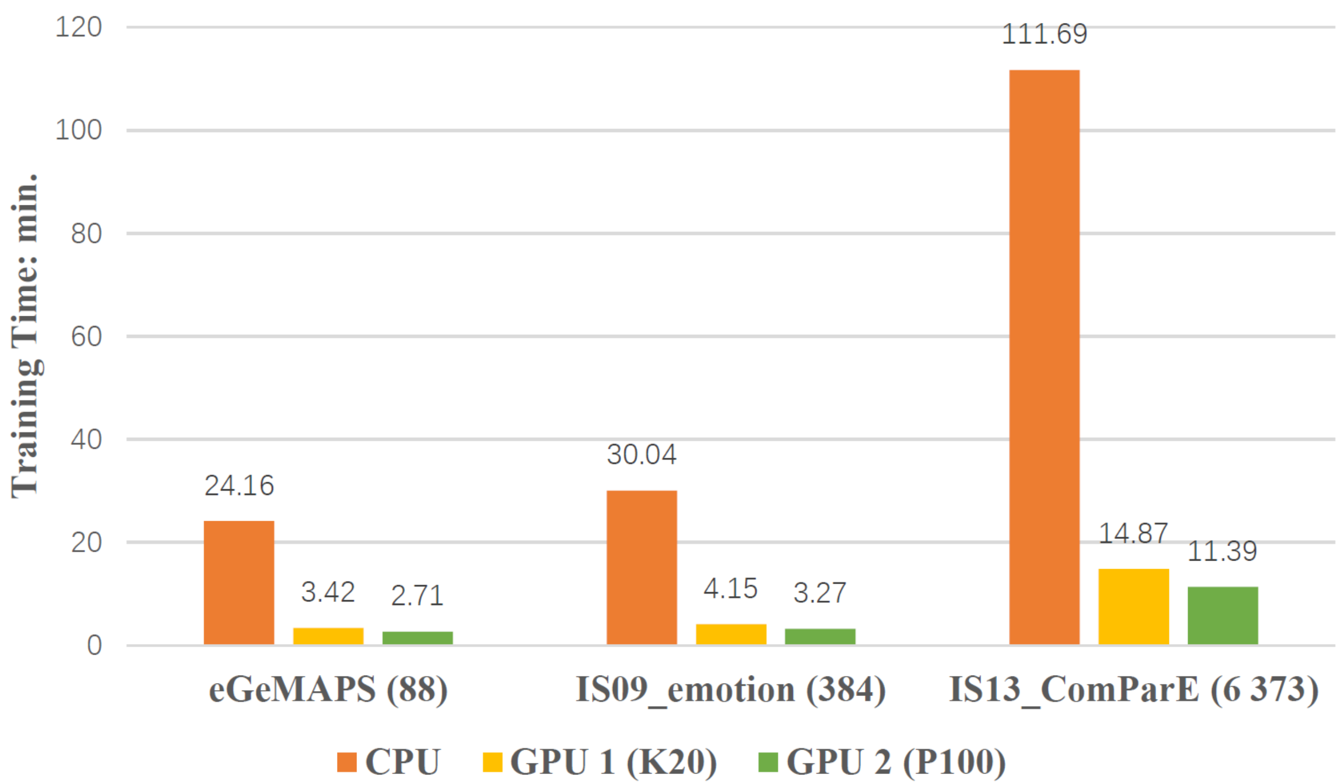




*"Optimization and Parallelization of Monaural Source Separation Algorithms in the openBliSSART Toolkit", **Journal of Signal Processing Systems**, Springer, 2012.*

Big Data: Velocity.

- **GPU feature extraction**



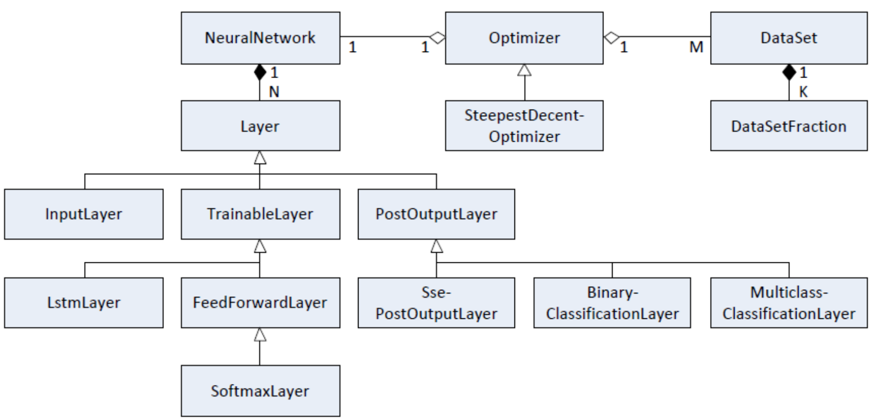
*"GPU-based Training of Autoencoders for Bird Sound Data Processing", **IEEE ICCE-TW**, 2017.*

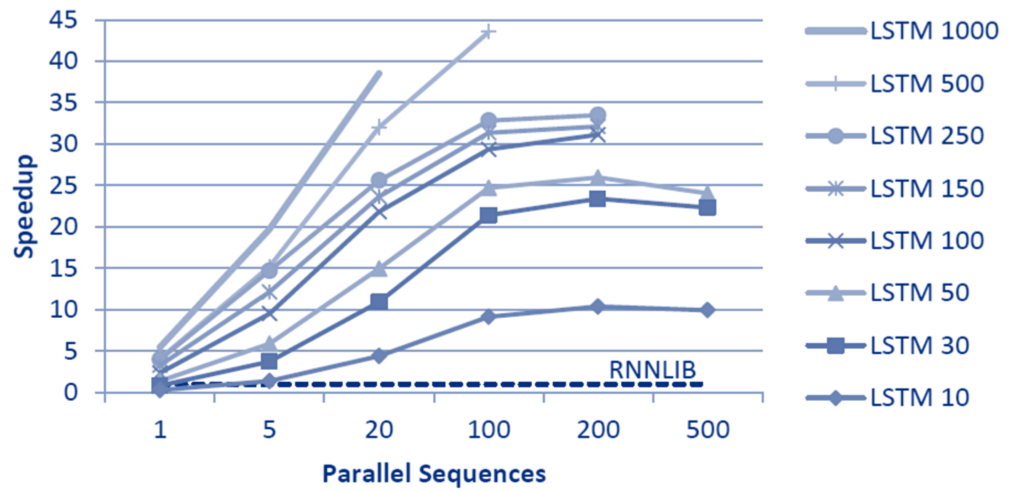
Big Data: Velocity.

CURRENNT

CHiME 2	RNNLIB	CURRENNT		
#seq.	1	1	10	200
speedup	(1)	2	13	22

- **GPU-Learning**
  10 – 1k LSTM cells,
  2k – 4Mio parameters
  GPGPU

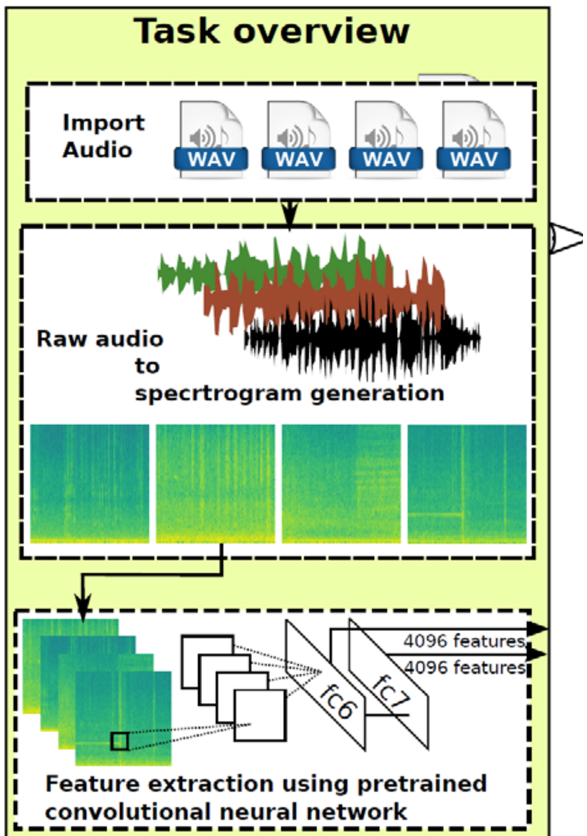


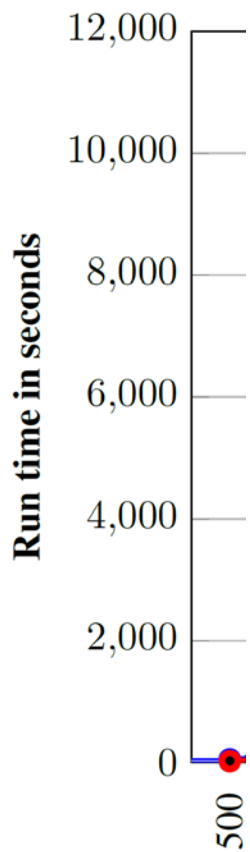


"*Introducing CURRENNT - the Munich Open-Source CUDA RecurREnt Neural Network Toolkit*",
Journal of Machine Learning Research, 2014.

Big Data: Velocity.

- **Parallel…**





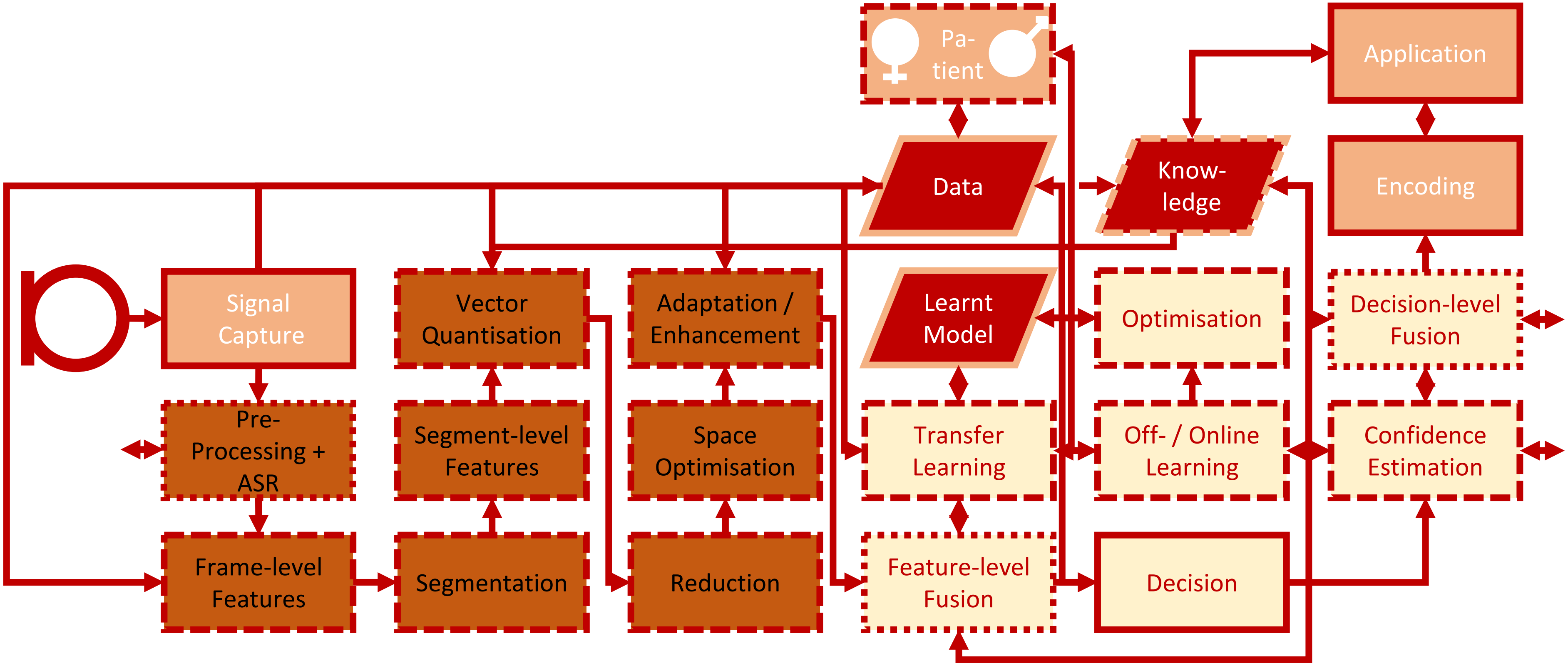
AlexNet	VGG19
Input = RGB image	
size: 227 × 227 pixels	size: 224 × 224 pixels
1×Convolution size: 11; ch: 96; stride: 4	2×Convolution size: 3; ch: 64; stride: 1
Maxpooling	
1×Convolution size: 5; ch: 256	2×Convolution size: 3; ch: 128
Maxpooling	
1×Convolution size: 3; ch: 384	4×Convolution size: 3; ch: 256
	Maxpooling
1×Convolution size: 3; ch: 384	4×Convolution size: 3; ch: 512
	Maxpooling
1×Convolution size: 3; ch: 256	4×Convolution size: 3; ch: 512
Maxpooling	
Fully connected *fc6* layer, 4 096 neurons	
Fully connected *fc7* layer, 4 096 neurons	
Fully connected 1 000 neurons	
Output = Probabilities for 1 000 object classes through soft-max	

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

Deep Learning.

Speech Analytics.

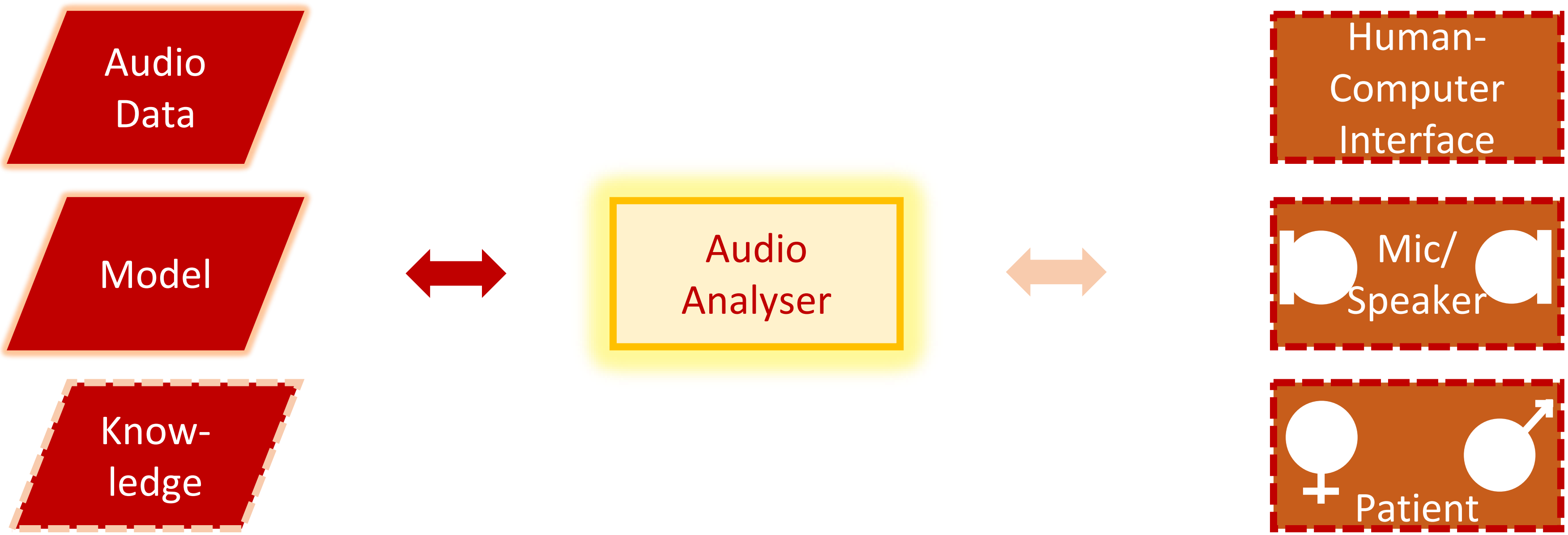
- **The "Traditional" Engine**



"Speech Emotion Recognition: 20 Years in a Nutshell, Benchmarks, and Ongoing Trends", Communications of the ACM, 2017.

Speech Analytics 2.0.

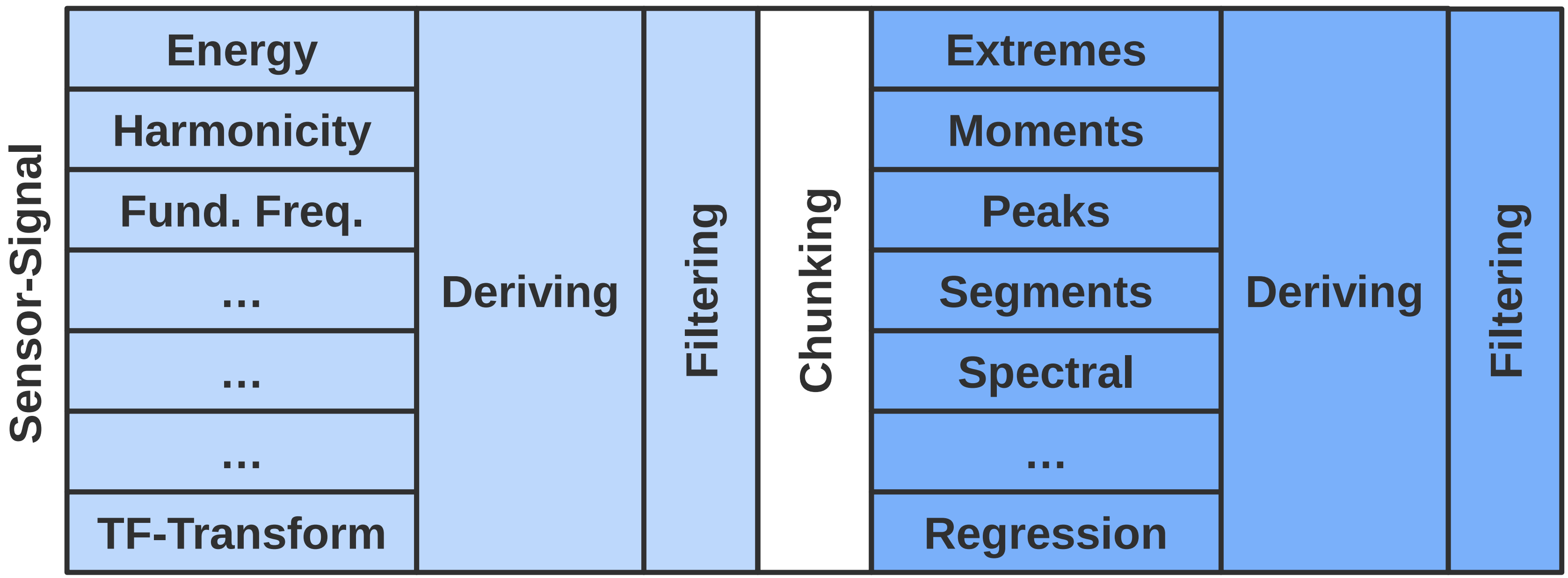
- **The "Modern" Engine?**



"Speech Emotion Recognition: 20 Years in a Nutshell, Benchmarks, and Ongoing Trends", Communications of the ACM, 2017.

Feature Extraction.

- **Brute-force**
  High-Dim. Space → Basis for selection
  Online update



"Recent Developments in openSMILE, the Open-Source Multimedia Feature Extractor", **ACM Multimedia**, 2013.

Functionals.

openSMILE:)

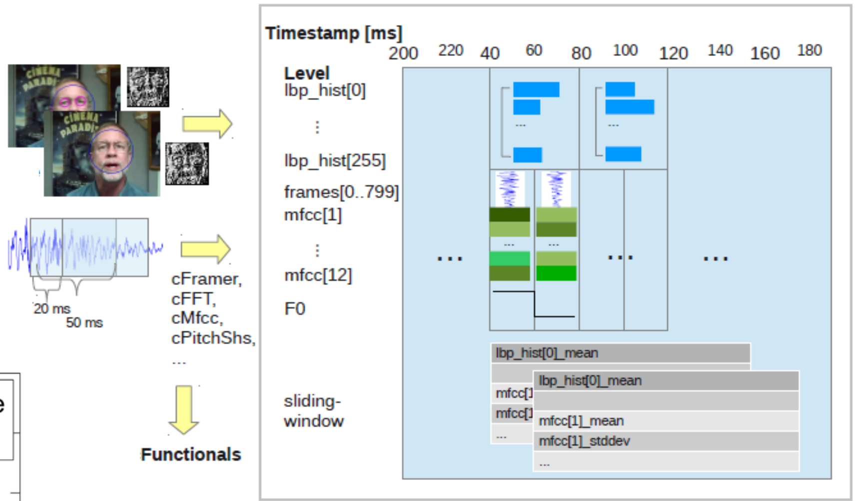
RTF (#feat)	Intel i7	HTC OneM9	Galaxy S3
.4k	.01	.06	.43
6.4k	.04	.23	.63

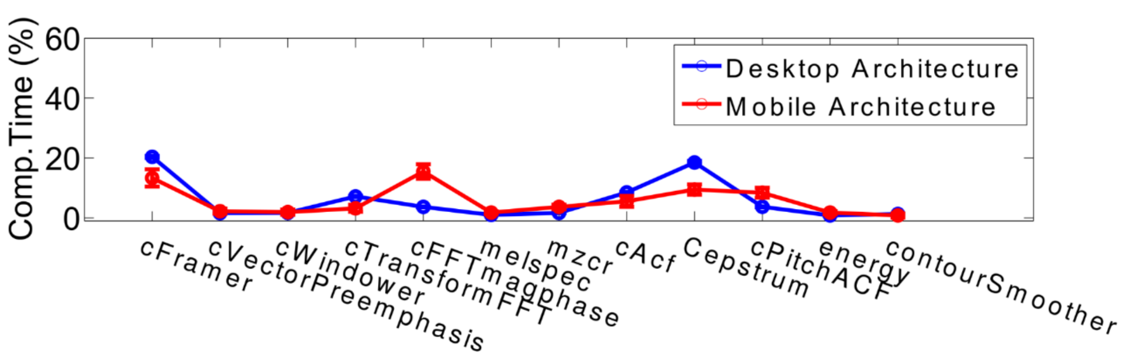
- **On-device Feature Extraction**

  Fast computation

  Cross-signal

  Energy/speed-aware selection





"Recent Developments in openSMILE, the Open-Source Multimedia Feature Extractor", **ACM Multimedia**, 2013.
(2nd place ACM MM Open Source Software Competition in 2010 and 2013, >1k citations for 3 papers)

Bag-of-”X”-Words.

LLDs over time ↓

Preprocessing:
- Normalisation of LLDs (online)

→

Codebook generation:
- K-means
- Random sampling

→

Vector quantisation:
- Single/Multi assignment

→

Postprocessing
- Histogram normalisation
- Log-TF-weighting
- IDF-weighting

↓

BoXW

openXBOW –|)→

https://github.com/openXBOW/openXBOW

Bag-of-"X"-Words.

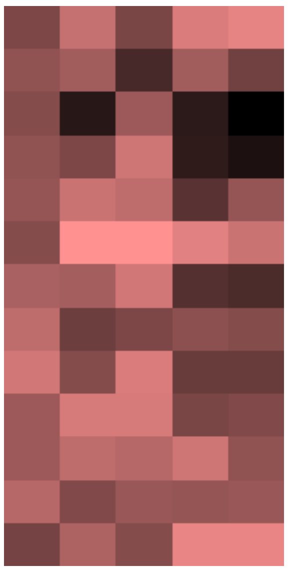
openXBOW –|)→

vector quantisation

1 2 3 4 5 …

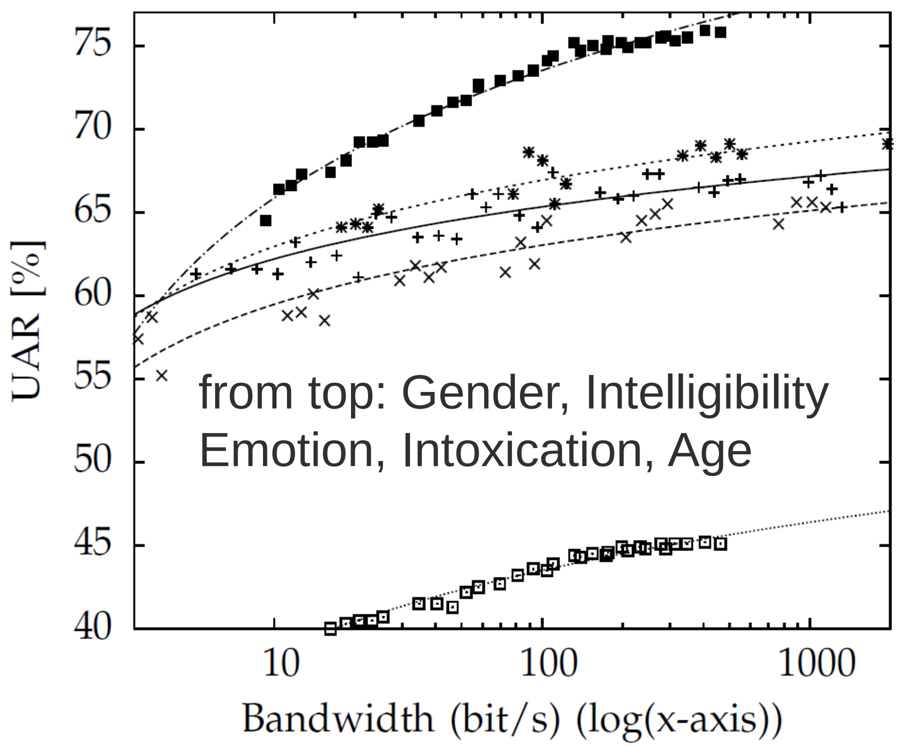
2 2 2 1 1 1 1 3 3 1 4 4 5 4

LLD x



LLD 1

dictionary



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

histogram

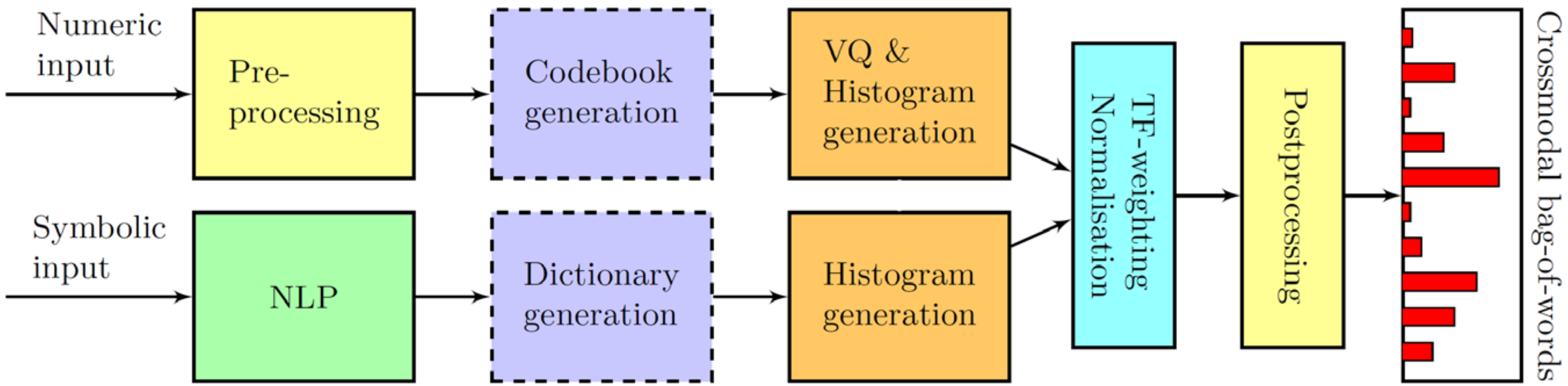
"openXBOW – Introducing the Passau Open-Source Crossmodal Bag-of-Words Toolkit", Journal of Machine Learning Research, 2017.

Bag-of-”X”-Words.

Automatic Sentiment Analysis in the Wild

- **Bags-of-X-Words**

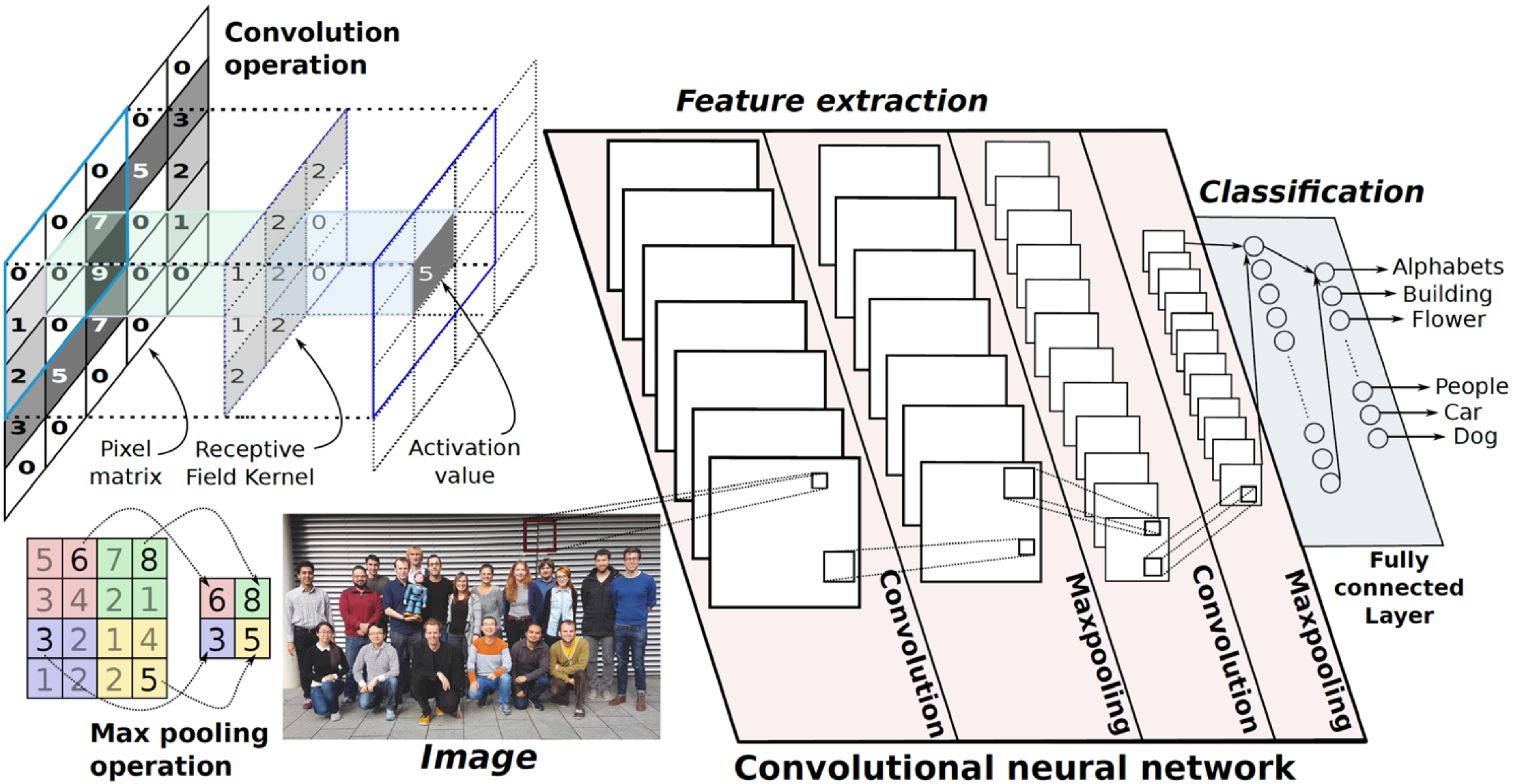
  SEWA, **openXBOW** –|)→



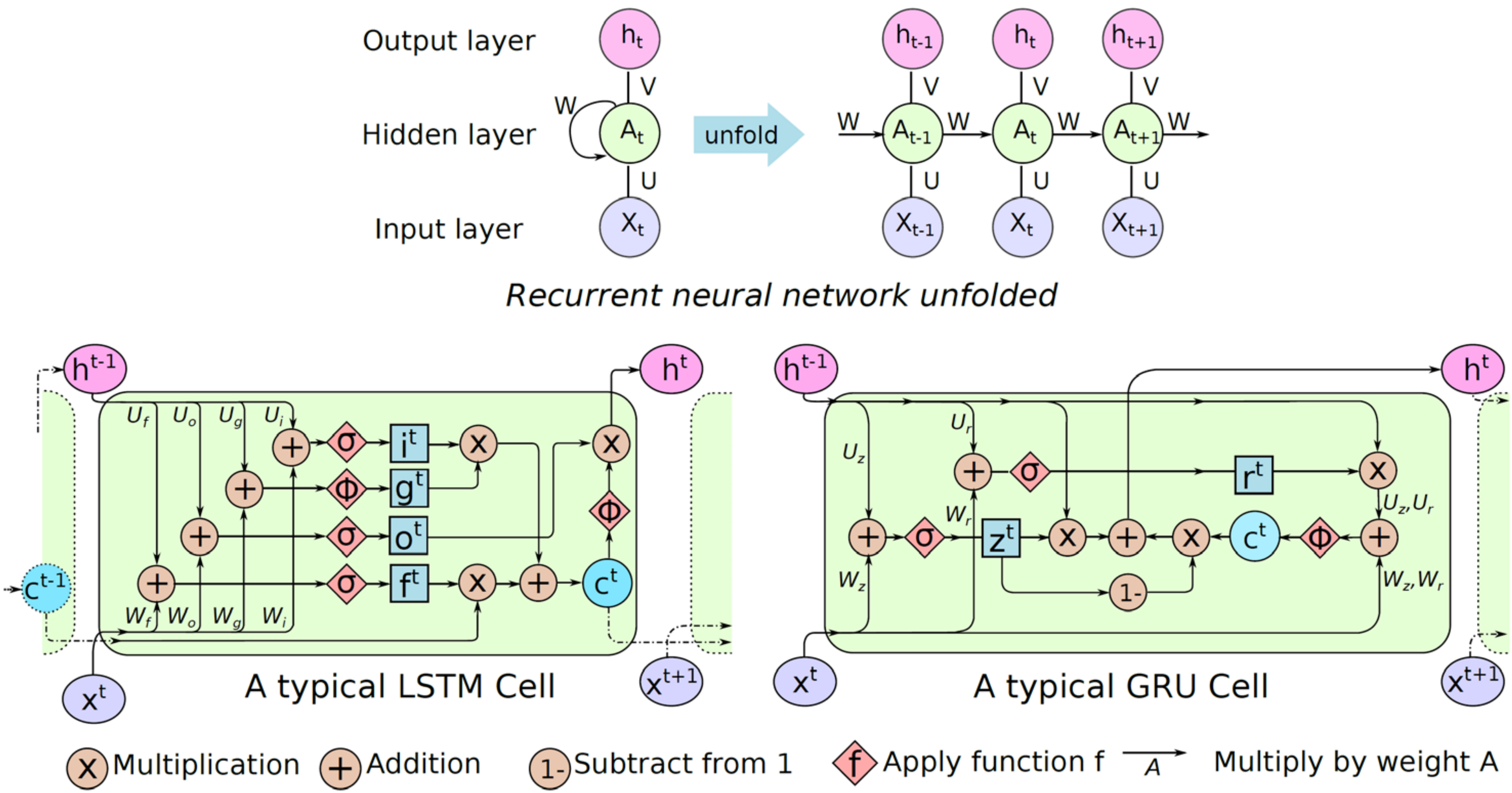
Modality	Arousal (test)	Valence (test)
Acoustic	.470	.426
Visual	.314	.344
Linguistic	.293	.320
Crossmodal (early fusion)	.432	.509
Crossmodal (late fusion)	.499	.523

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

Convolutional Neural Nets.



Deep Recurrent Nets.

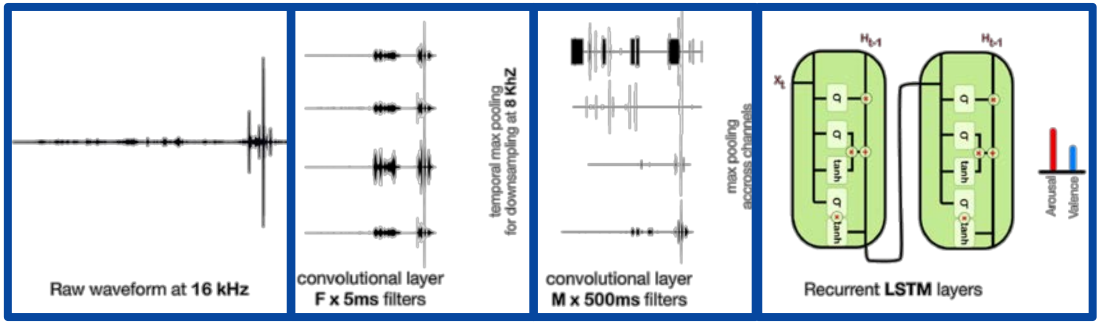


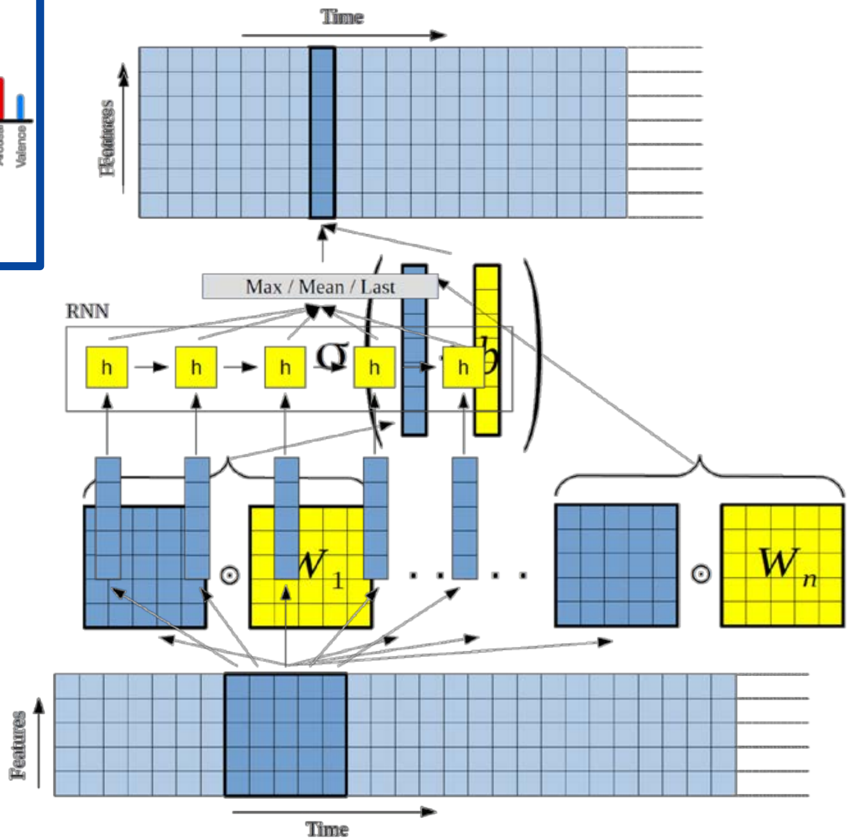
Recurrent neural network unfolded

End-to-End

Arousal	CCC
Baseline	.366
e2e	.686

- CNN + LSTM → CLSTM ?

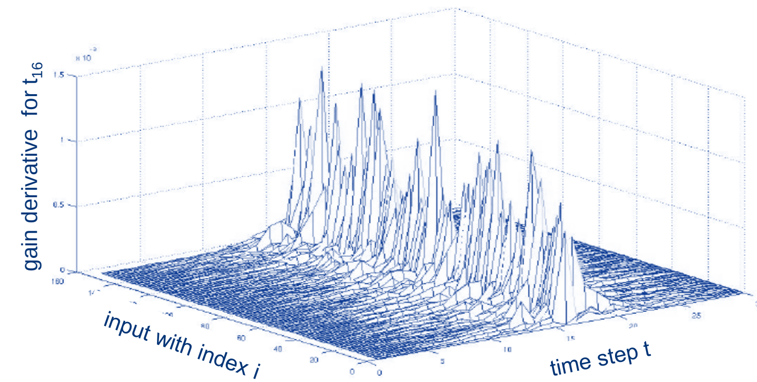


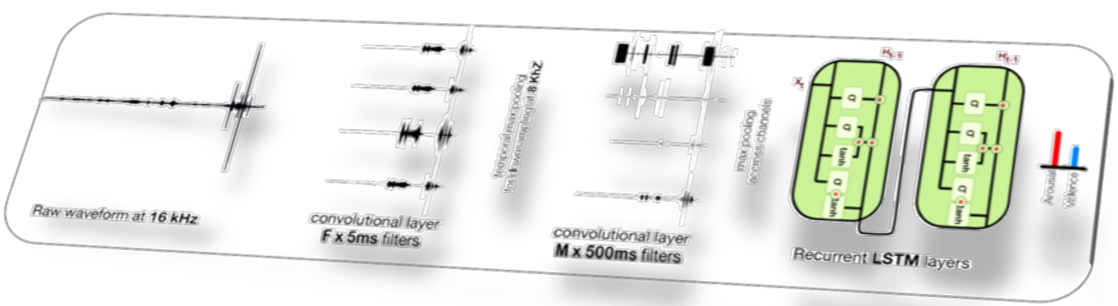


"Adieu Features? End-to-End Speech Emotion Recognition using a Deep Convolutional Recurrent Network", ICASSP, 2016.

End-to-End.

- **Black Box?**

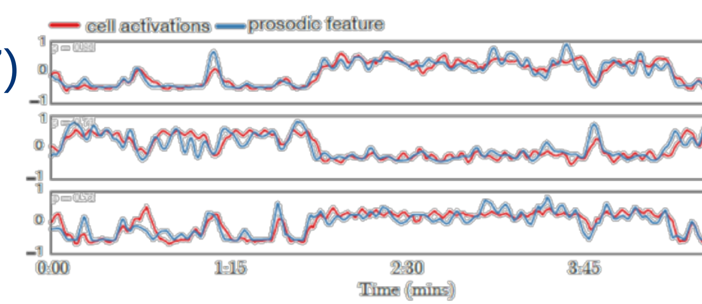




energy range (.77)

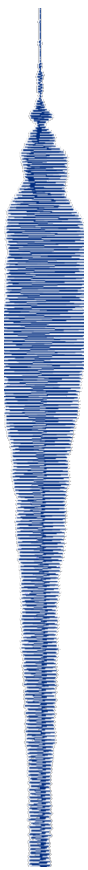
loudness (.73)

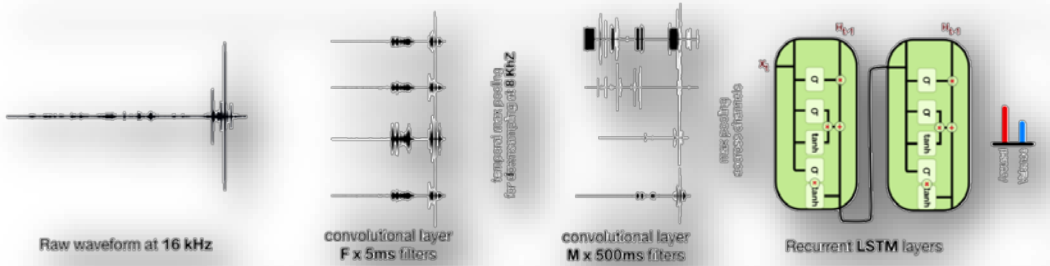
F0 mean (.71)

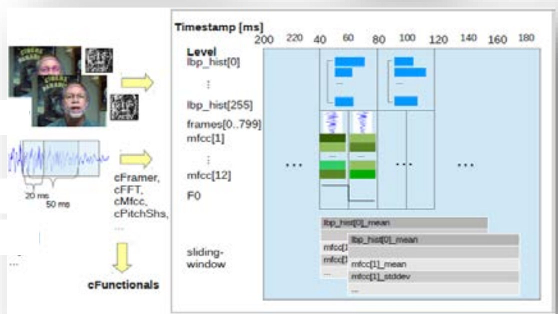


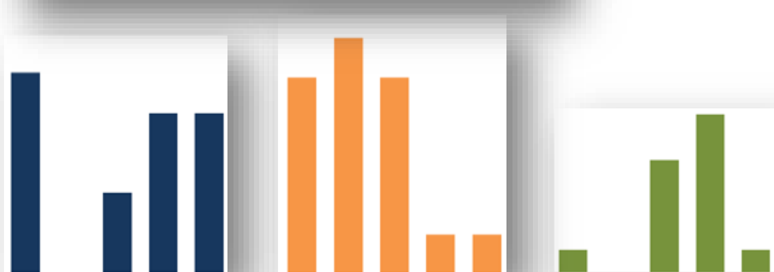
End-to-End.

- **e2e + functionals + BoAW?**









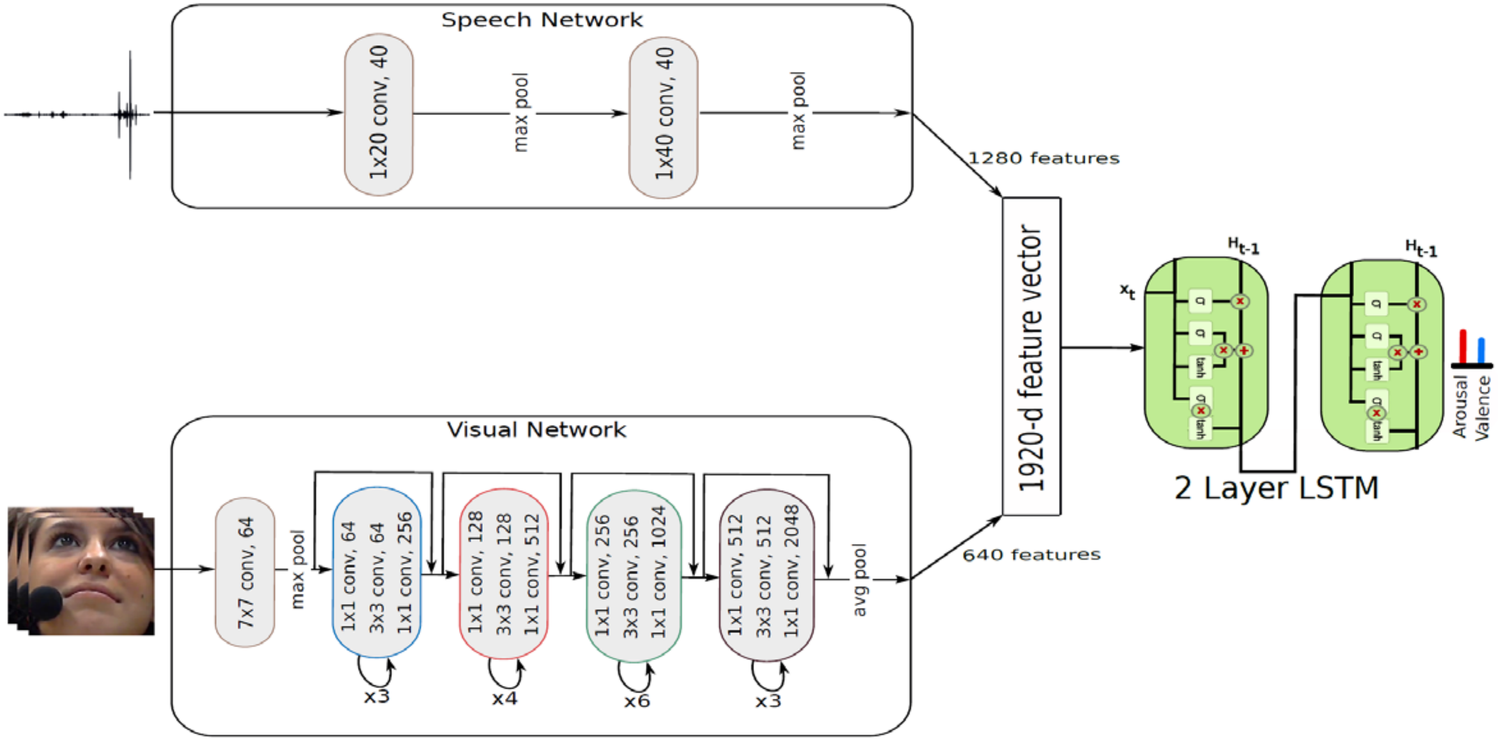
Speech under Cold	%UA
func	70.2
BoAW	69.7
e2e	60.0
func + BoAW	70.1
e2e + func	64.8
e2e + BoAW	62.5
all (conf.)	70.7
all (maj. vote)	71.0

"The INTERSPEECH 2017 Computational Paralinguistics Challenge: Addressee, Cold & Snoring", Interspeech, 2017.

End-to-End.

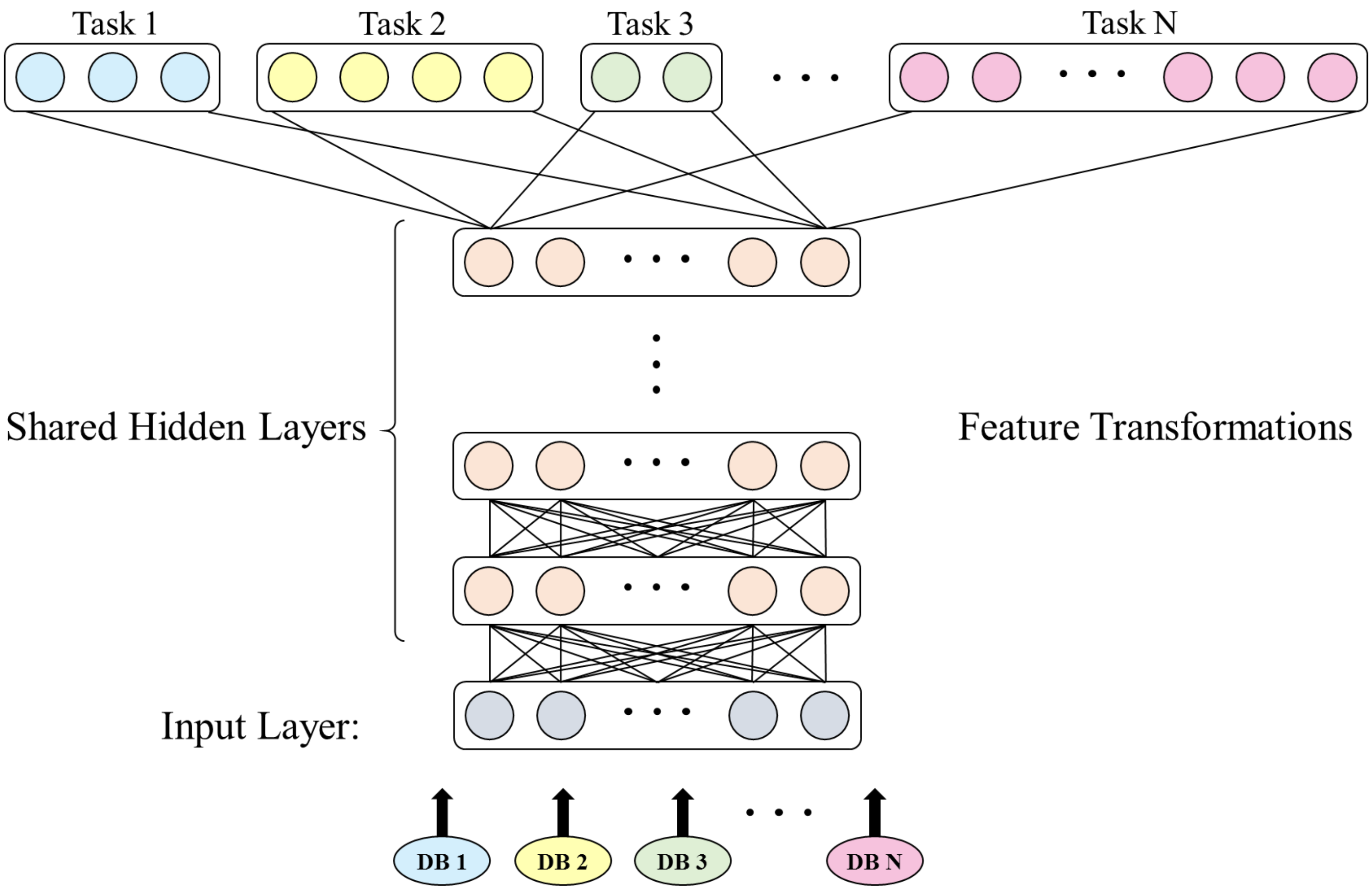
CCC	
Arousal	.770
Valence	.612

- **AVEC 2015/16 Task**



"End-to-End Multimodal Emotion Recognition using Deep Neural Networks", submitted / arxiv.org.

Multi-target.

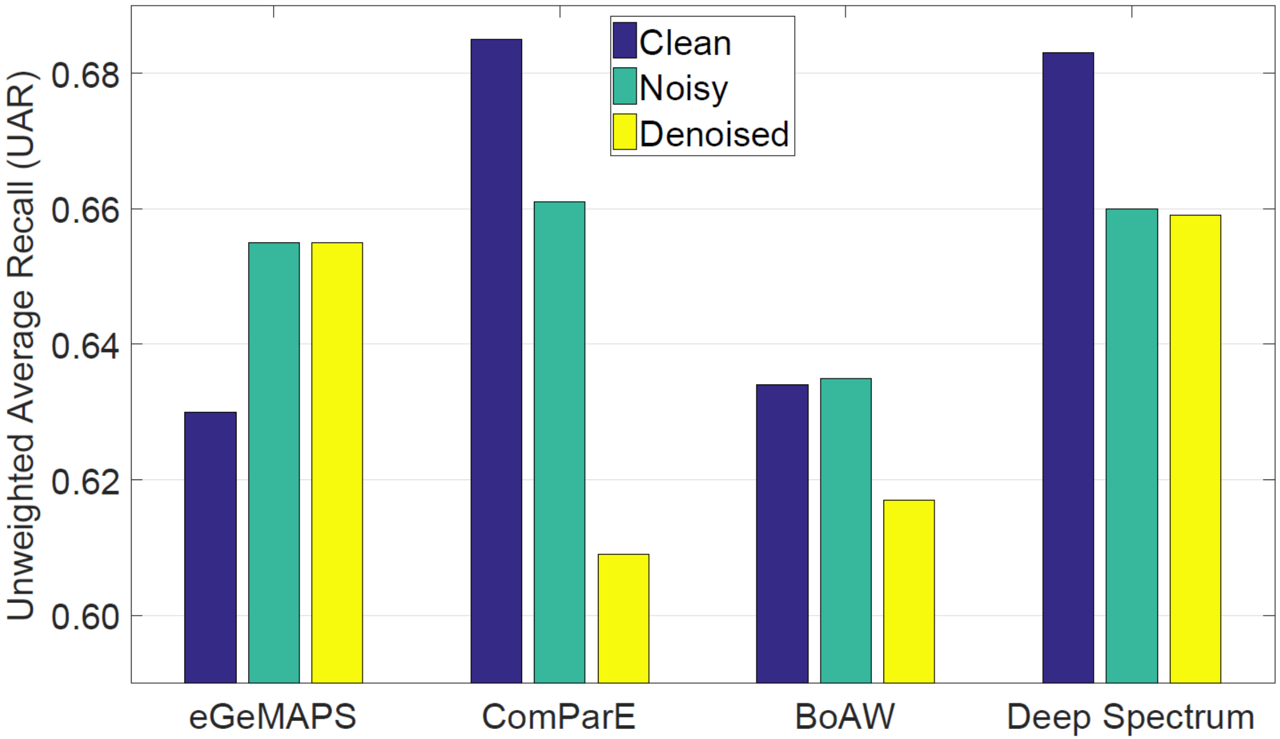


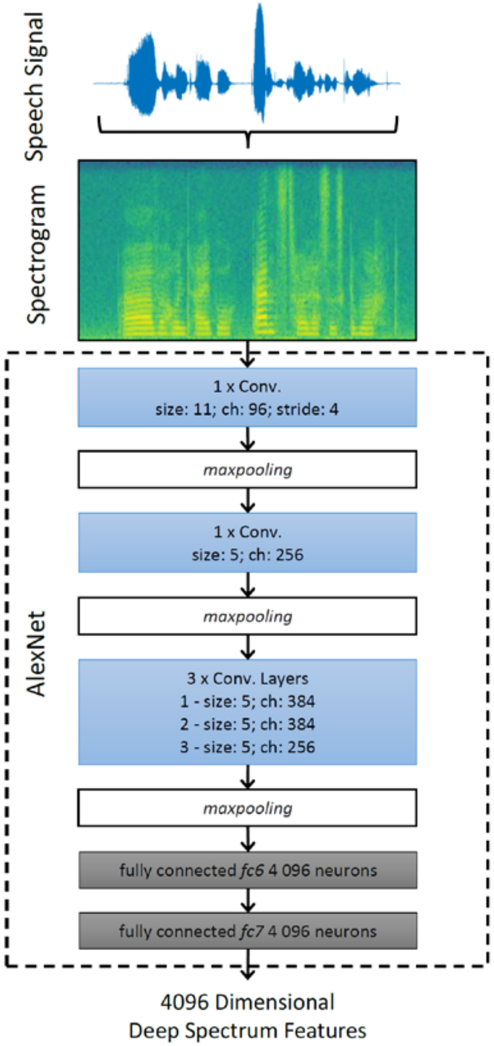
"Multi-task Deep Neural Network with Shared Hidden Layers: Breaking down the Wall between Emotion Representations", ICASSP, 2017.

Pre-Training.

- **Emotion with Image Nets**

  IS Emotion Challenge task – 2 classes



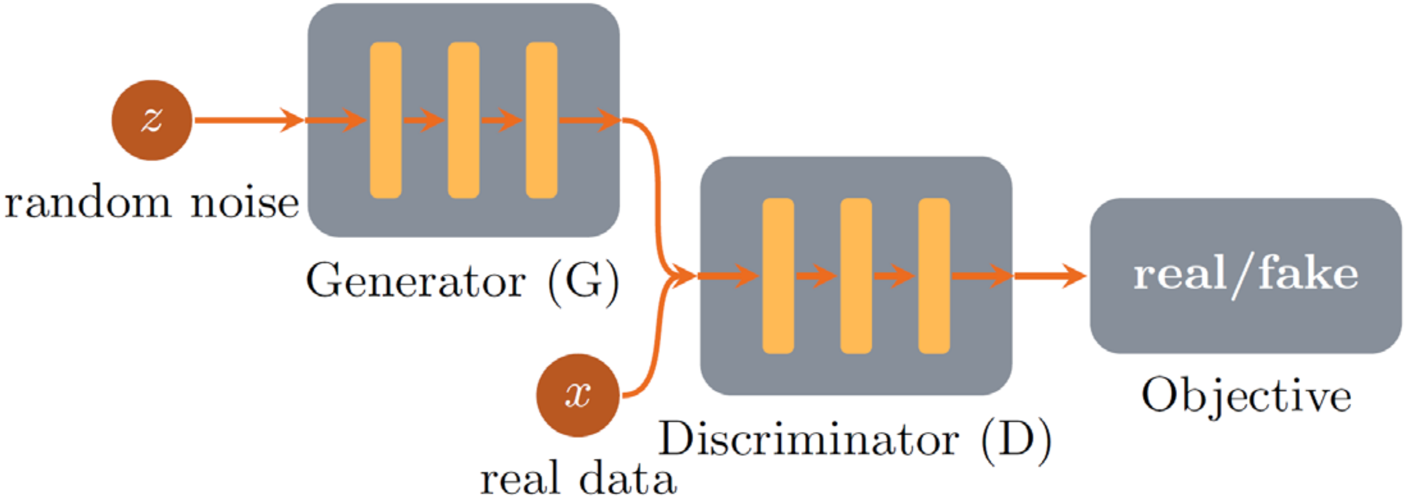


*"An Image-based Deep Spectrum Feature Representation for the Recognition of Emotional Speech", **ACM Multimedia**, 2017.*

Synthesis?

% UAR	
SVM	42.83
GAN	44.06

- **GANs**
  Generator & discriminator competing against each other in
  Zero-sum / Min-Max "game" framework



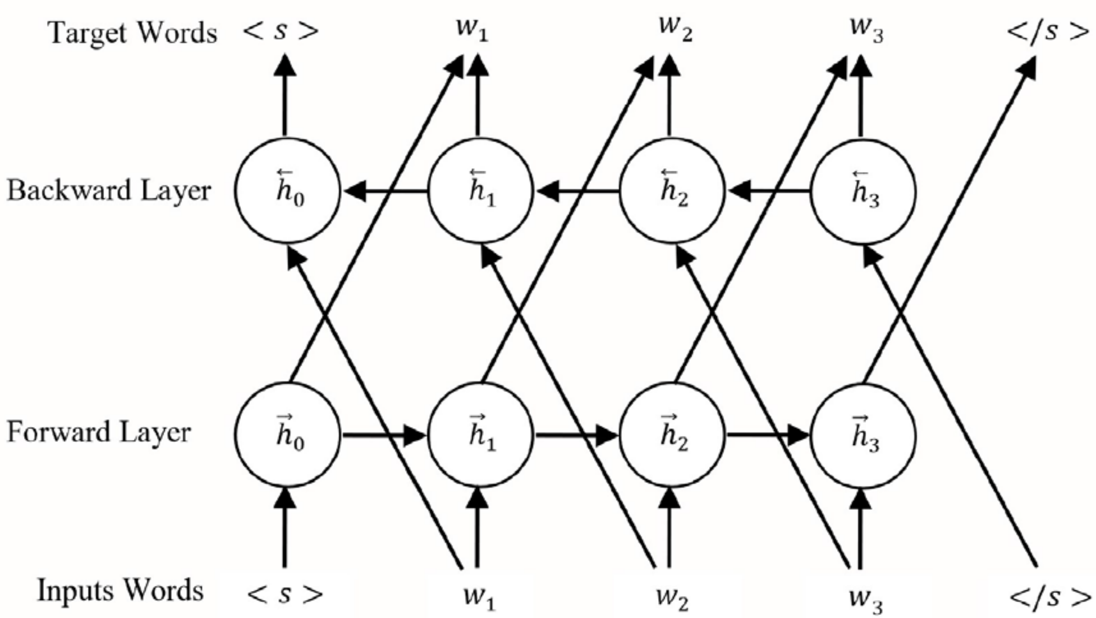
- **Example: Autism Diagnosis from Speech**
  CPESD database: 4 classes, children

"Speech-based Diagnosis of Autism Spectrum Condition by Generative Adversarial Network Representations", ACM Digital Health, 2017.

Deep Context Modelling.

contextual BLSTM (cBLSTM) LM

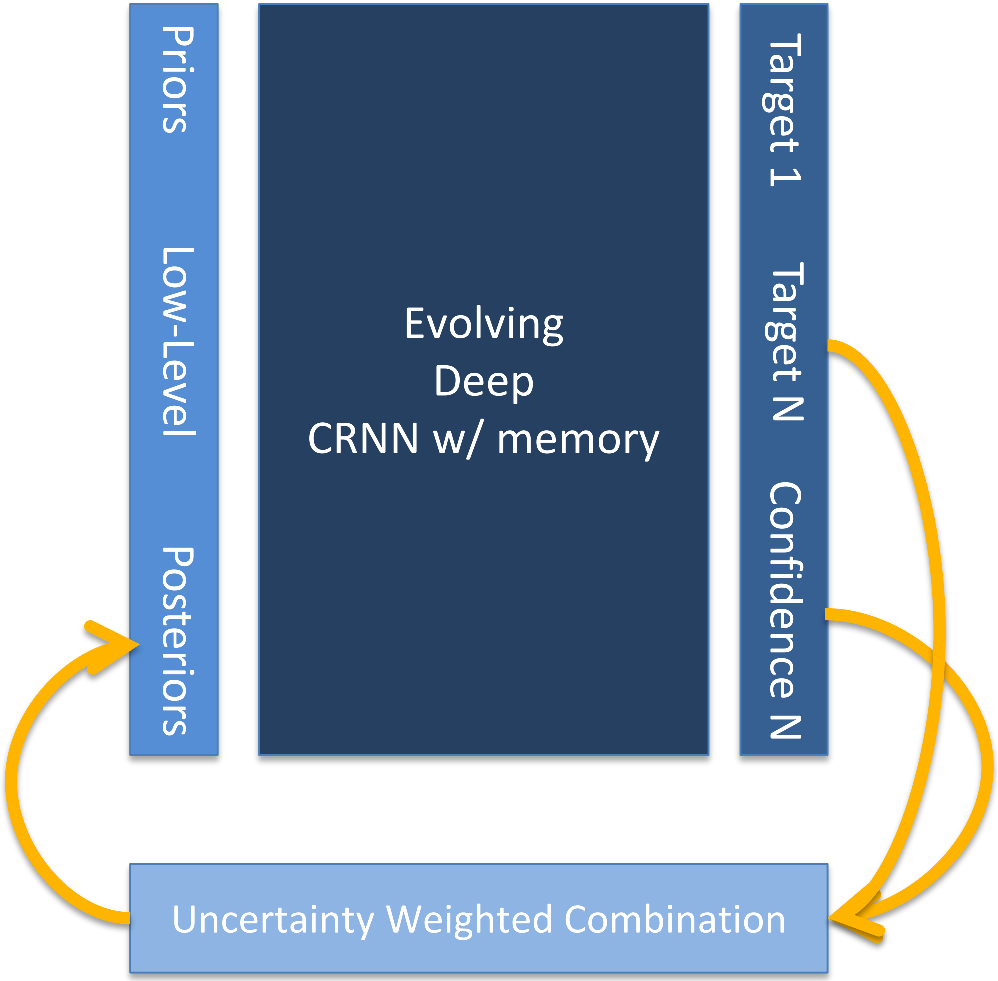
- exclude predicted word from conditional dependence
- cBLSTM: modified architecture, contextual dependence
- predict conditional probability $p(w_m | w_1^{m-1}, w_{m+1}^M)$
- CURRENNT toolkit `http://sourceforge.net/p/currennt`



*"Contextual BLSTM Language Models: A Generative Approach to Sentiment Analysis", **EACL**, 2017.*

Deep Embedding.

- **Seamless Holism**
- **Horizontal:**
  - Signal Enhancement
  - Feature Extraction
  - Feature Enhancement
  - Feature Transfer
  - Feature Alignment
  - Feature Selection (Bottleneck)
  - Classification / Regression
  - Language Modelling
- **Vertical:**
  - Multitarget
  - w/ Confidences (e.g., agreement)



Examples.

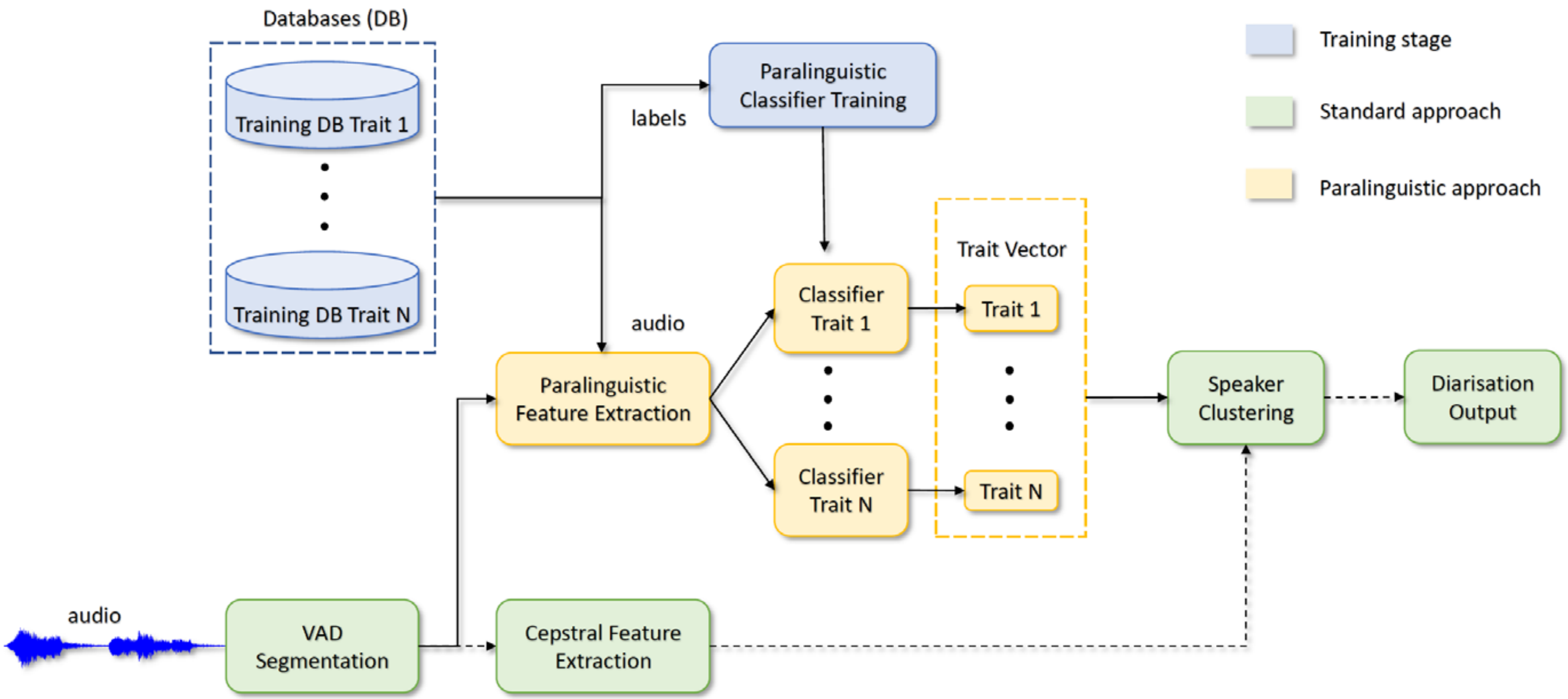
		# Classes	%UA/*AUC/$^+$CC
	Addressee	2	70.6
	Cold	2	72.0
	Snoring	4	70.5
	Deception	2	72.1
	Sincerity	[0,1]	65.4+
Personality	Native Lang.	11	82.2
Likability	Nativeness	[0,1]	43.3+
Intelligibility	Parkinson's	[0,100]	54.0$^+$
Intoxication	Eating	7	62.7
Sleepiness	Cognitive Load	3	61.6
Age	Physical Load	2	71.9
Gender	Social Signals	2x2	92.7*
Interest	Conflict	2	85.9
Emotion	Emotion	12	46.1
Negativity	Autism	4	69.4

Diarisation.

System	Miss	FA	sperr	DER
LIUM	6.3	20.1	39.0	65.4
sensAI	15.2	8.1	23.4	46.7
Paralings	6.3	20.4	38.0	64.7

- **Paralings for Diarisation**

  SEWA database



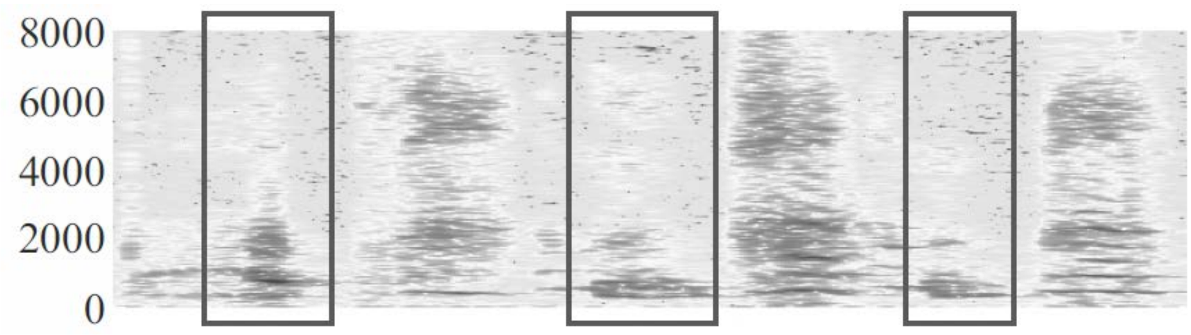
"A Paralinguistic Approach To Holistic Speaker Diarisation", ACM Multimedia, 2017.

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 Promising Novel Approach for the Earlier Identification of Rett Syndrome", Rett Syndrome Europe, 2016.
"Earlier Identification of Children with Autism Spectrum Disorder: An Automatic Vocalisation-based Approach", Interspeech, 2017.

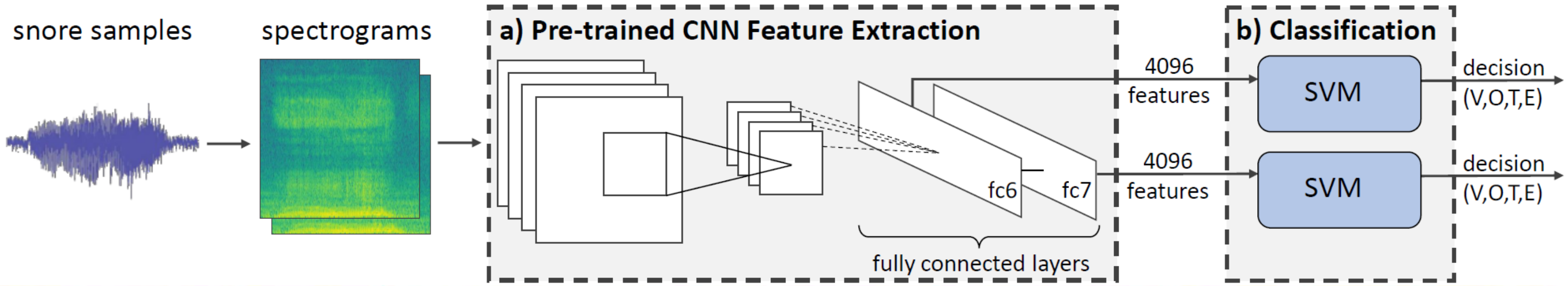
Products.

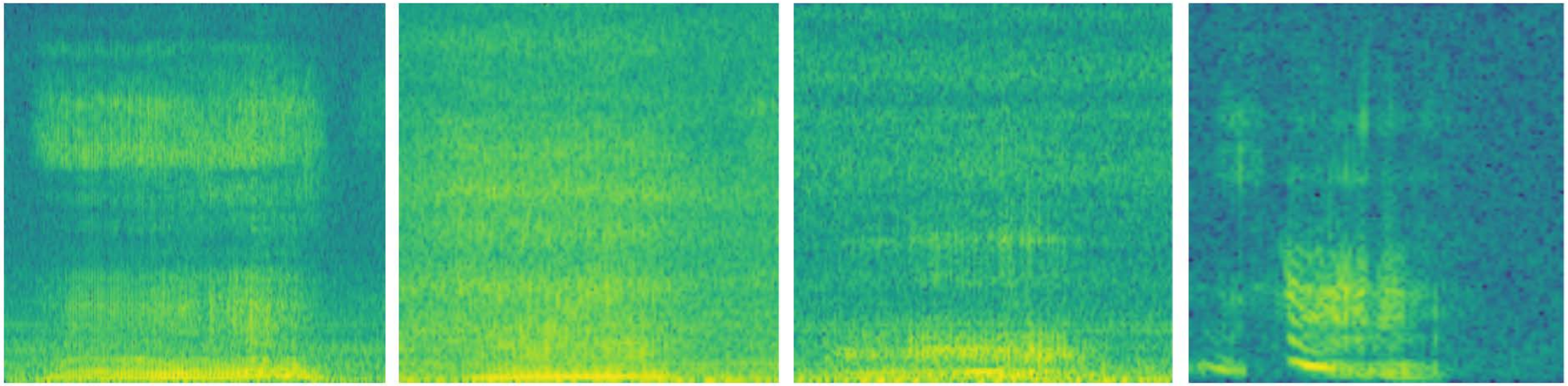
sensAi

Snoring.

	%UA
CNN+LSTM	40.3
Functionals	58.8
Deep Spec	67.0

VOTE classification
(site of vibration)





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

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

Animal Paralinguistics?

Recognition	%UA
Emotion	.42
Context	.40

- **Bark Context & Emotion**

Mudi, a Hungarian Herding Dog
226 Bark Sequences, 12 different dogs, 6 annotators
5 point likert scale per emotion → max emotion

Aggression. Despair. Fear. Fun. Happiness.

Alone. Ball. Fight. Food. Play. Stranger. Walk.

*"Classifying the Context and Emotions of Dog Barks: A Comparison of Acoustic Feature Representations", **SAS**, 2017.*

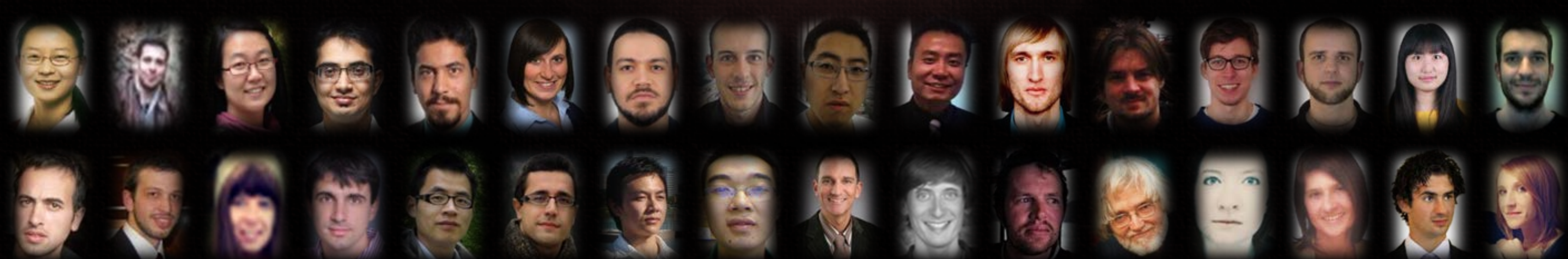
Vision.

Thank You.
Спасибо.

Conquering the Consumer Market.

Socio-Emotionally Intelligent Dialogs.

Super-human Speaker Analysis.



Imperial College London

audEERING
intelligent Audio Engineering

Abstract

With two years, one has roughly heard a thousand hours of speech - with ten years, around ten thousand. Similarly, an automatic speech recogniser's data hunger these days is often fed in these dimensions. In stark contrast, however, only few databases to train a speaker analysis system contain more than ten hours of speech. Yet, these systems are ideally expected to recognise the states and traits of speakers independent of the person, spoken content, language, cultural background, and acoustic disturbances at human parity or even super-human levels. While this is not reached at the time for many tasks such as speaker emotion recognition, deep learning - often described to lead to "dramatic improvements" - in combination with sufficient learning data satisfying the "deep data cravings" holds the promise to get us there. Luckily, every second, more than two hours of video are uploaded to the web and several hundreds of hours of audio and video communication in most languages of the world take place. If only a fraction of these data would be shared and labelled reliably, "x-ray"-alike automatic speaker analysis could be around the corner for next gen human-computer interaction, mobile health applications, and many further benefits to society.

In this light, first a solution towards utmost efficient exploitation of the "big" (unlabelled) data available is presented. Small-world modelling in combination with unsupervised learning help to rapidly identify potential target data of interest. Then, gamified dynamic cooperative crowdsourcing turn its labelling into an entertaining experience, while reducing the amount of required labels to a minimum by learning alongside the target task also the labellers' behaviour and reliability. Then, increasingly autonomous deep holistic end-to-end learning solutions are presented for the task at hand. Benchmarks are given from the 15 research challenges organised by the speaker over the years at Interspeech, ACM Multimedia, and related venues. The concluding discussion will contain some crystal ball gazing alongside practical hints not missing out on ethical aspects.