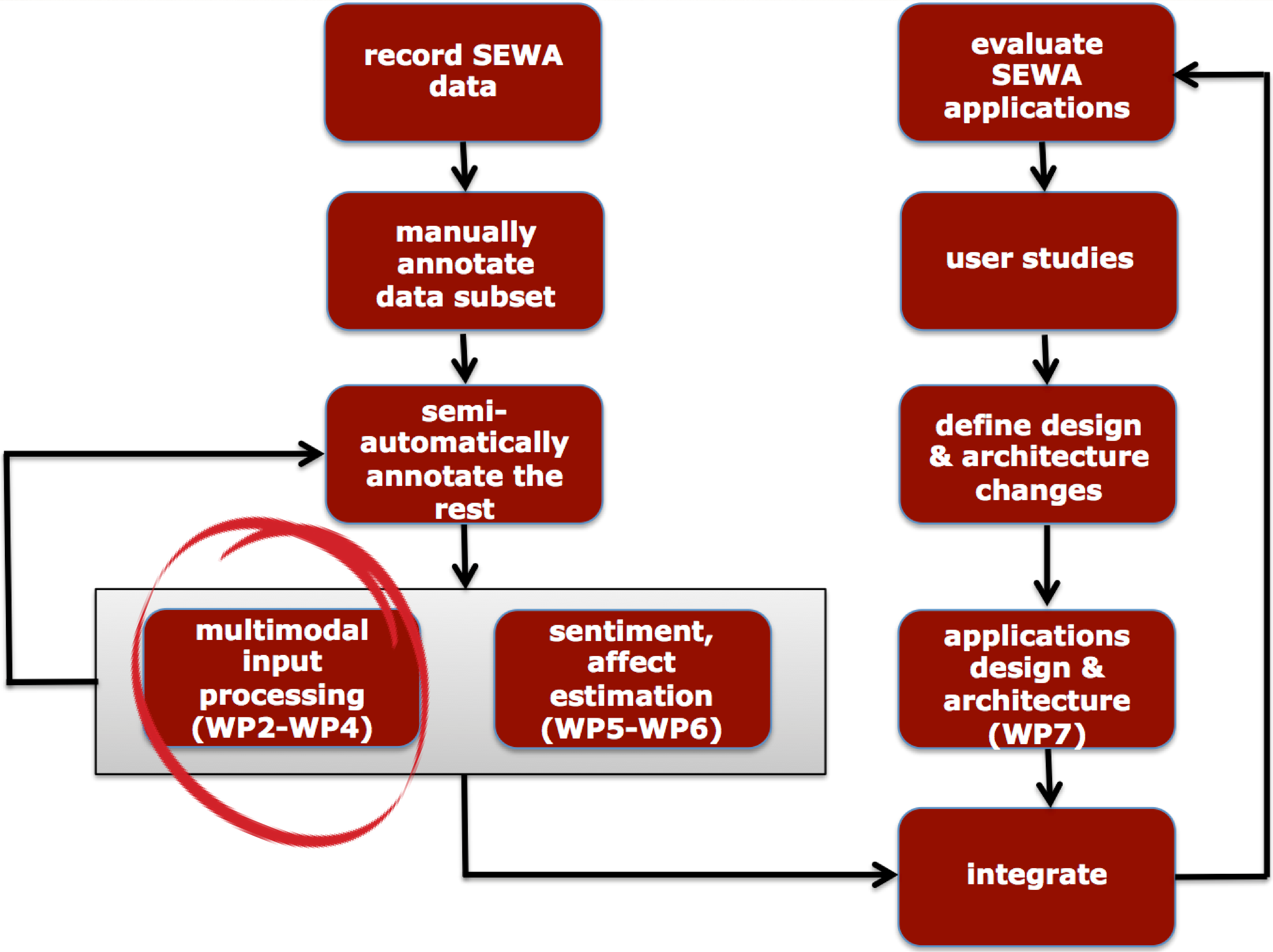
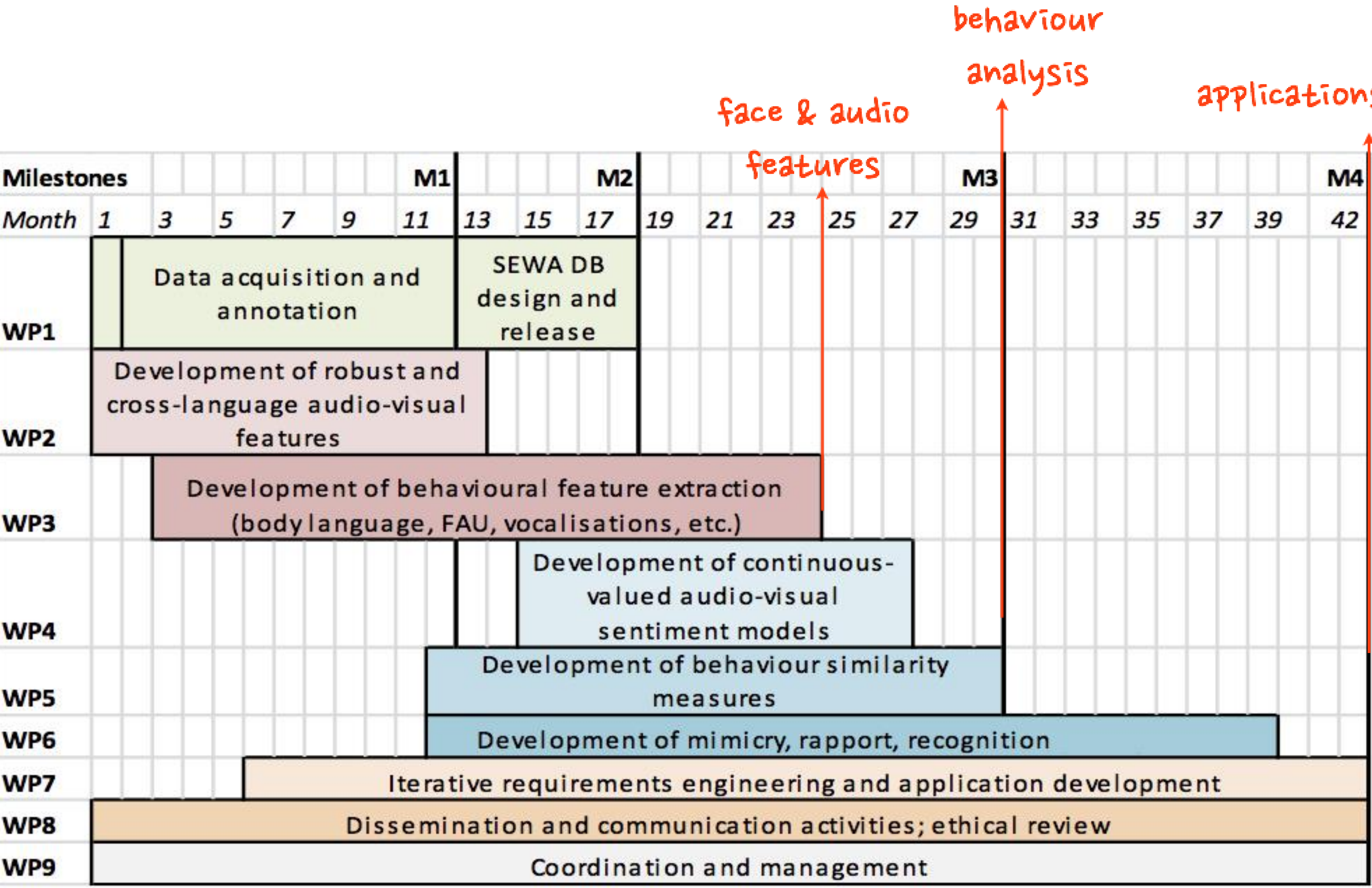
WP2:
Low-level Feature Extraction

Björn Schuller
Jie Shen

sewa
Automatic Sentiment Analysis in the Wild


sewa
Automatic Sentiment Analysis in the Wild
record SEWA data
manually annotate data subset
semi-automatically annotate the rest
multimodal input processing (WP2-WP4)
sentiment, affect estimation (WP5-WP6)
evaluate SEWA applications
user studies
define design & architecture changes
applications design & architecture (WP7)
integrate


behaviour analysis
face & audio features
applications
Milestones
M1
M2
M3
M4
Month
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
37
39
42
WP1
Data acquisition and annotation
SEWA DB design and release
WP2
Development of robust and cross-language audio-visual features
WP3
Development of behavioural feature extraction (body language, FAU, vocalisations, etc.)
WP4
Development of continuous-valued audio-visual sentiment models
WP5
Development of behaviour similarity measures
WP6
Development of mimicry, rapport, recognition
WP7
Iterative requirements engineering and application development
WP8
Dissemination and communication activities; ethical review
WP9
Coordination and management

Low-level feature extraction

- Process audio-visual input
  (e.g. facial expressions, vocalisations and casual speech)
  - Real-life conditions
  - Multiple languages
- Obtain:
  - Acoustic features (Passau)
  - Visual features (ICL)
- Requirements:
  - Independence of **language**, user facial/vocal **characteristics**
  - Environmental **robustness** (e.g. equipment, background noise)
- Enables detection of **sentiment**, **affect** and **intentions**

Objectives

- ❖ Task 2.1: Environmentally robust acoustic features
- ❖ Task 2.2: Environmentally robust visual features

  → Robust visual feature extractor (D2.2, February 2016, M13)
- ❖ Task 2.3: Cross-lingual language-related features

sewa
Automatic Sentiment Analysis in the Wild

Facial Landmark Tracking

- Goal: to accurately track facial landmarks in SEWA applications.
- Further requirements:
  - Reliability.
  - High processing speed.

Imperial College London

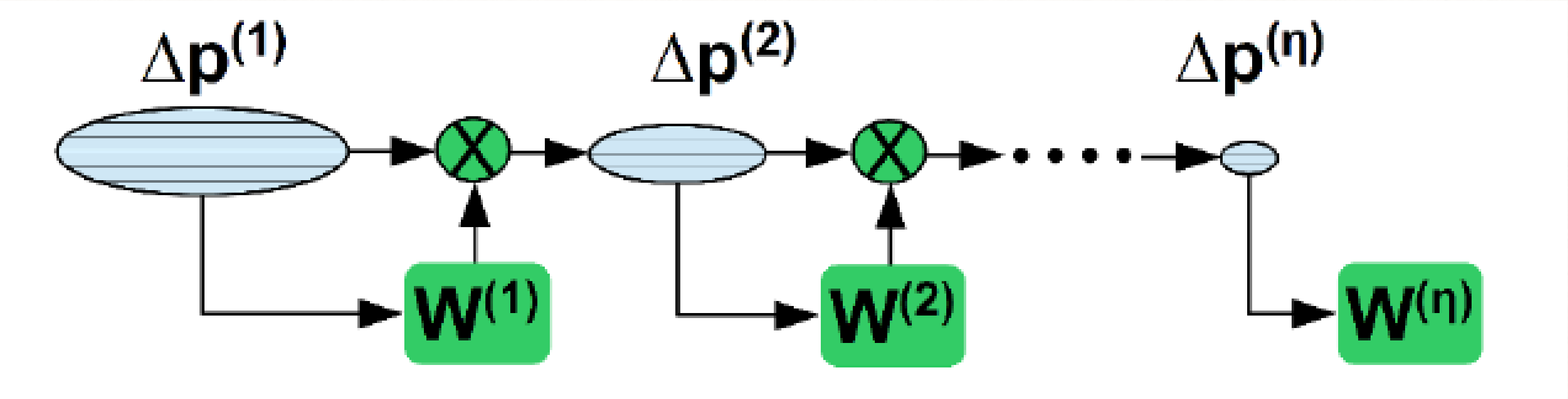
Incremental Face Alignment

- Given new unseen examples, automatically update the existing fitting models.
- Challenges:
  - How to update the model efficiently?
  - How to incorporate new training data?

Imperial College London

Cascade Linear Regression (CLR)

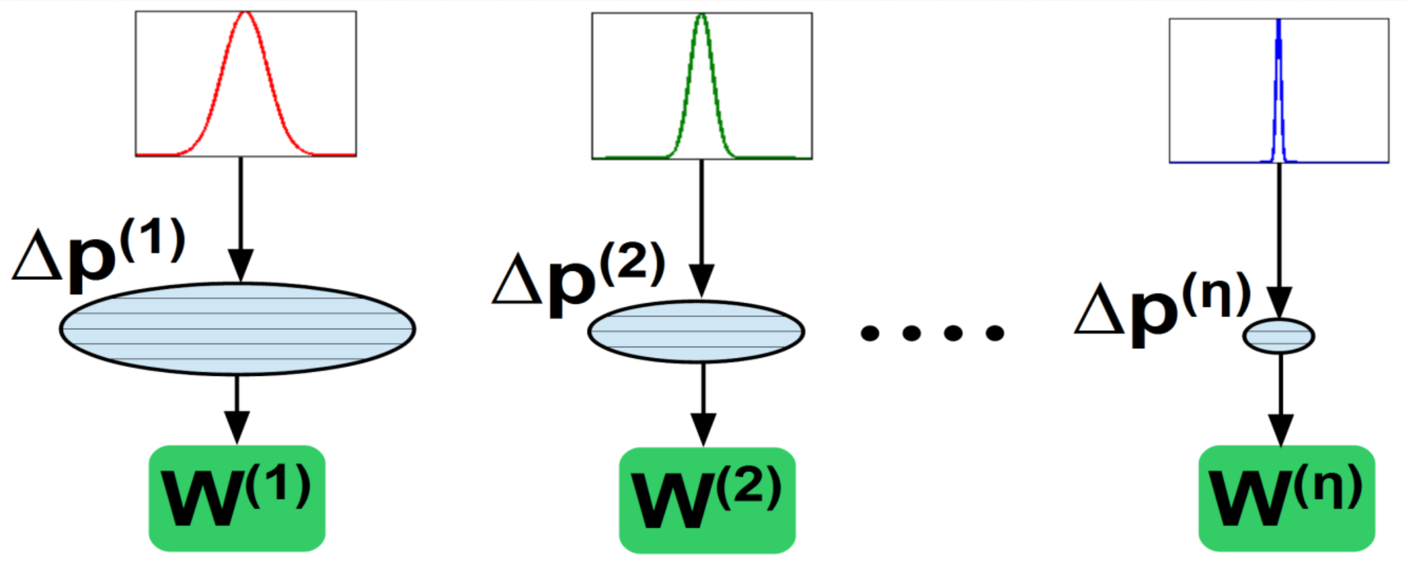
- Generate perturbed shapes within a predefined range.
- Compute HOG features around each landmark point.
- Find a function that can map the features to the displacement between the ground truth and perturbed shapes, using CLR:



Imperial College London

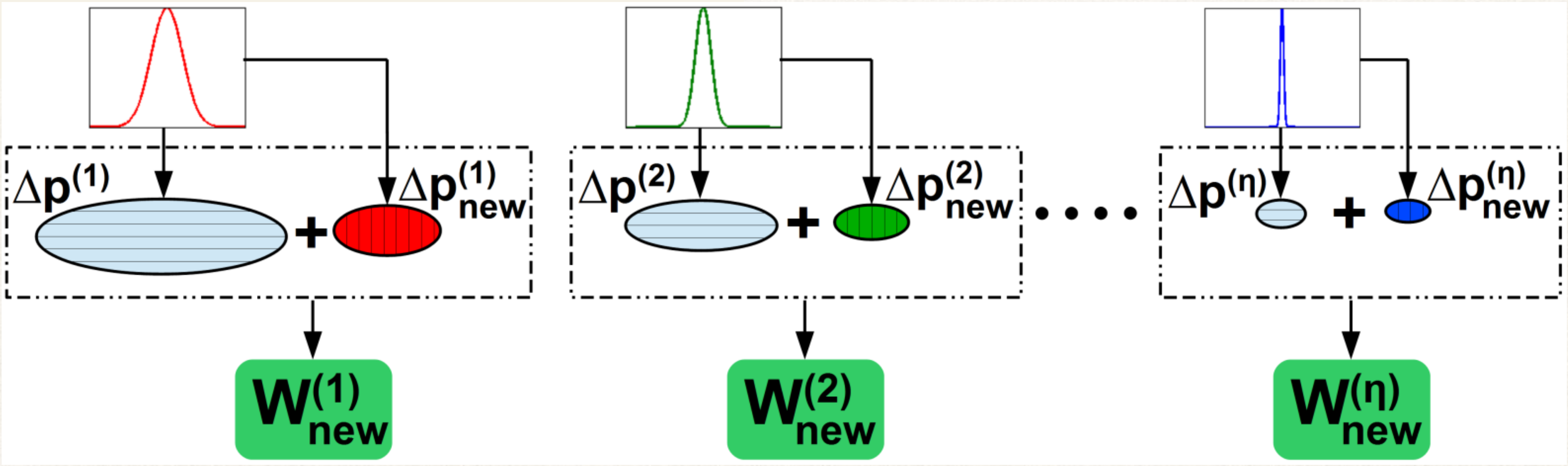
Parallel-CLR (Par-CLR)

- Learning the cascade of regression is by nature a Monte-Carlo procedure.
  - Collect the statistics for the shape parameters at each level.
  - Draw the perturbations from the distribution to train the regressors in parallel.



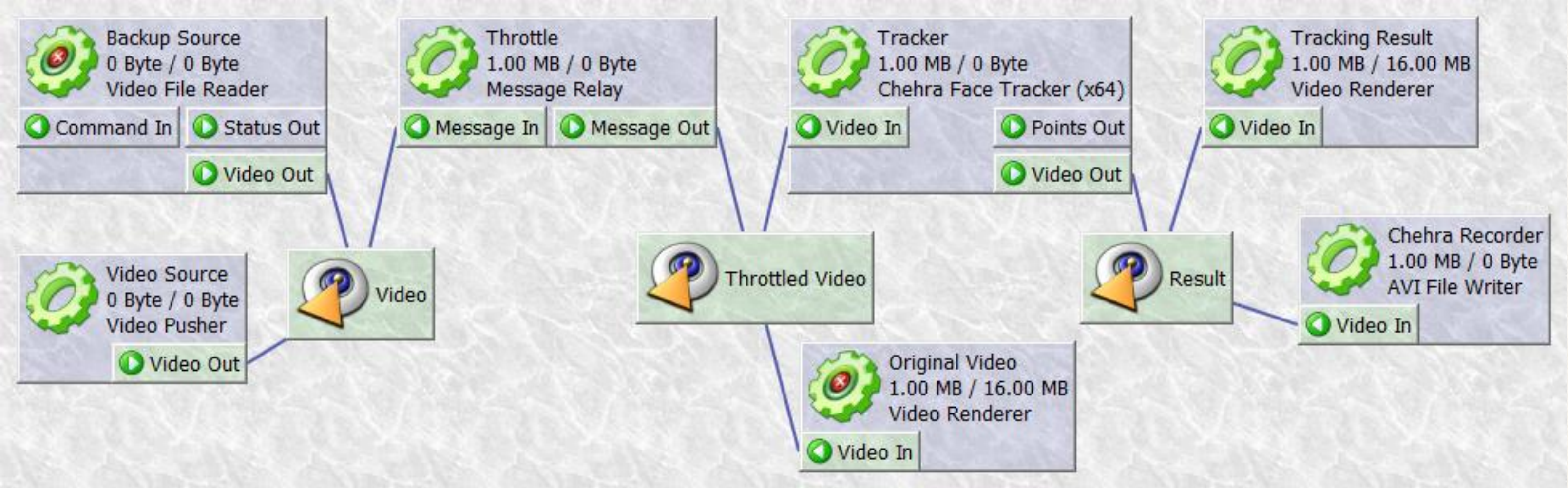
Incremental Par-CLR (iPar-CLR)

- Uses incremental linear least squares solution to perform the updates.
- Allows for all the level of the cascade to be updated with new examples independently in parallel.

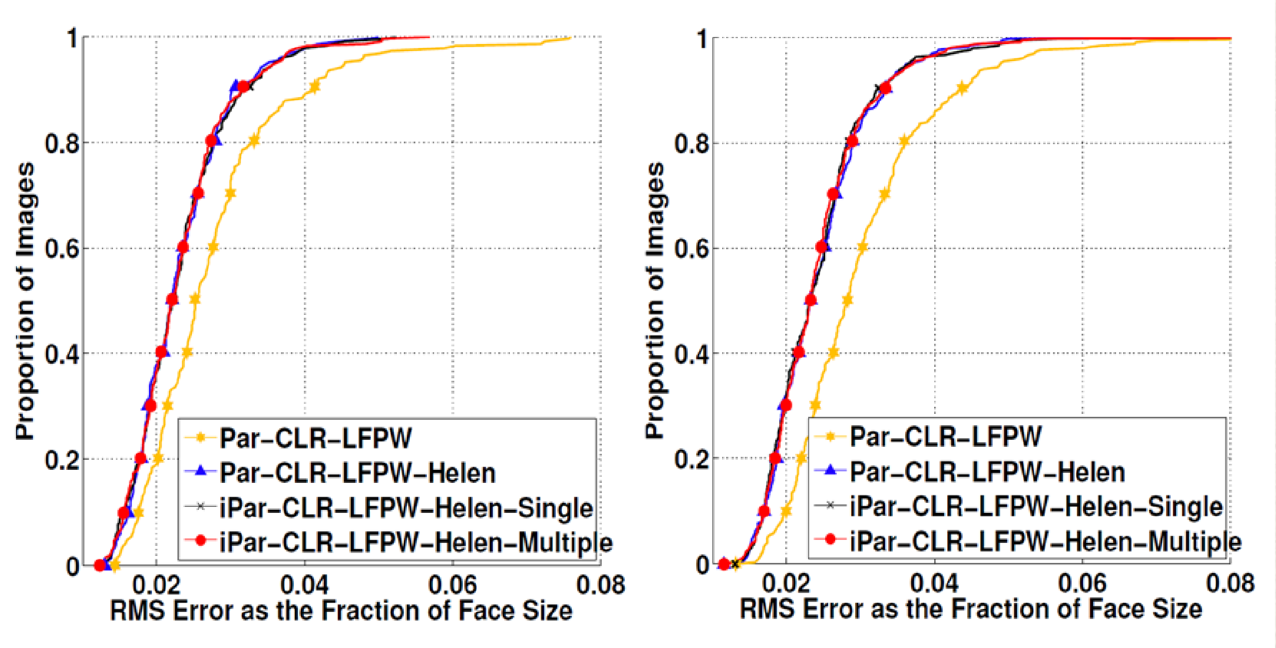


Software Implementation

- iPar-CLR method is implemented into the Chehra tracker.
  - Use daemon process for crash recovery.
  - Can track 8 streams at 50 fps in parallel.
  - Now integrated into the SEWA back-end server.



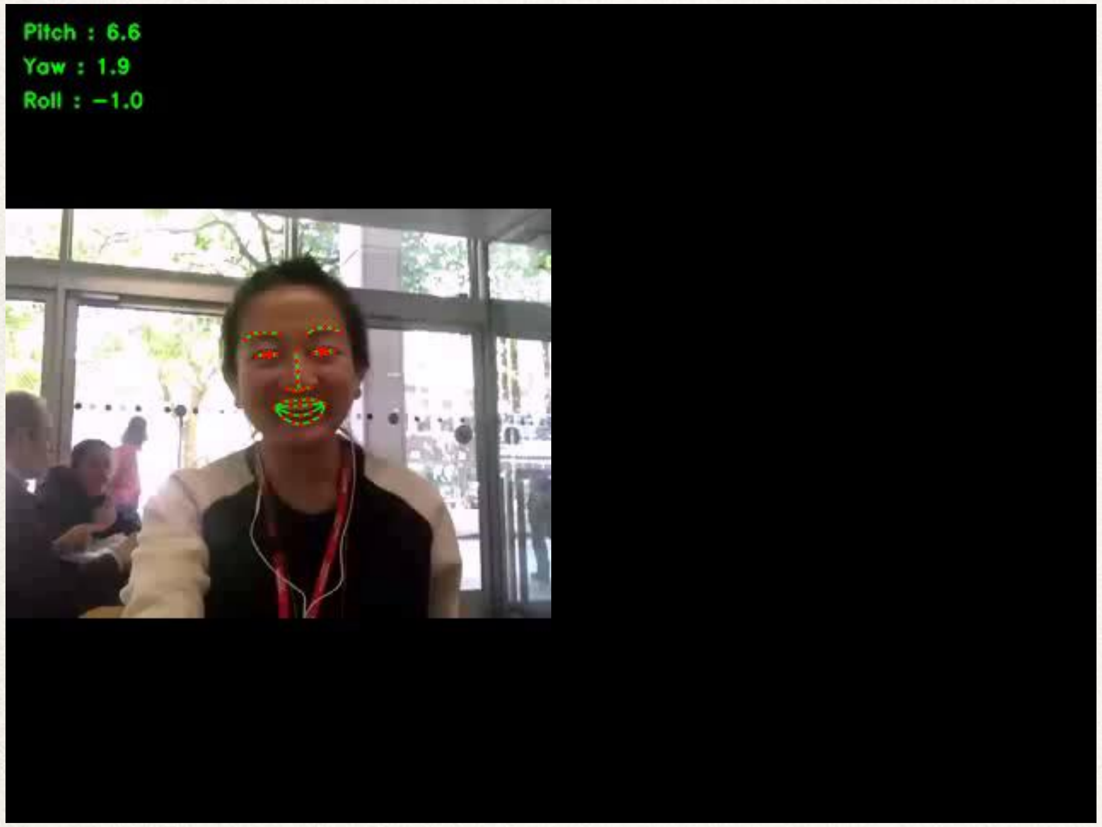
Result on LPFW and Helen Data

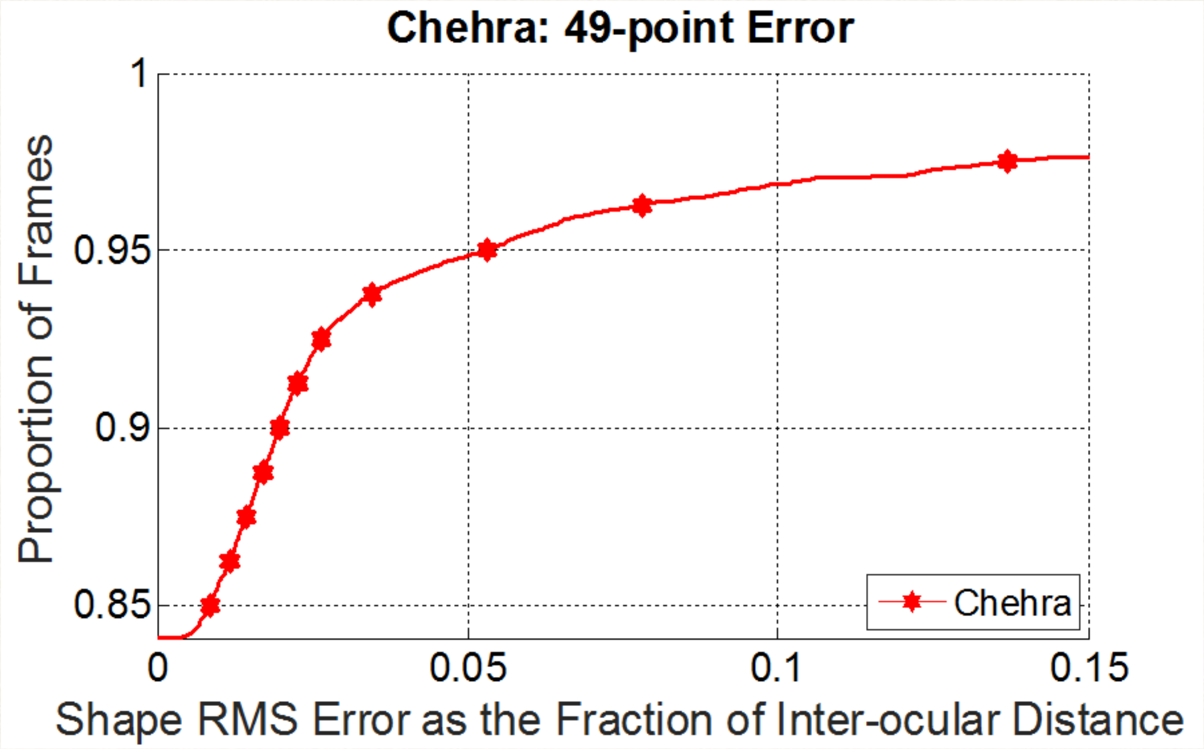


(a) LFPW Test Set
(b) Helen Test Set

Imperial College London

Result on SEWA Data





Imperial College London

More Result on SEWA Data

Environmentally robust visual features

- Facial landmark tracking ✓

Objectives

- ❖ Task 2.1: Environmentally robust acoustic features

  → Improved acoustic feature extractor (D2.1, October 2015, M9)

- ❖ Task 2.2: Environmentally robust visual features

- ❖ Task 2.3: Cross-lingual language-related features

Environmentally robust acoustic features

1. **Selection of features** that are correlated with target labels in noisy data

❖ State-of-the-art acoustic emotion recognition feature sets

❖ Bag-of-audio-words (BoAW) representations (generated, e.g., by Vector Quantisation or Deep Semi-NMF)

2. **Feature enhancement** by deep de-noising auto-encoders such as LSTM-RNN

❖ On raw spectral features (as in previous studies on ASR)

❖ Learning of non-linear distortions in
   (a) Emotion-related features, e.g., low-level descriptor contours
   (b) BoAW representations

State-of-the-art feature sets

Selection of noise robust features:

- 132 features, including **prosody**, **voice quality**, **auditory spectrum**, **spectral** / **cepstral** and **deltas**
- Data: RECOLA
- Noise: “Smartphone”

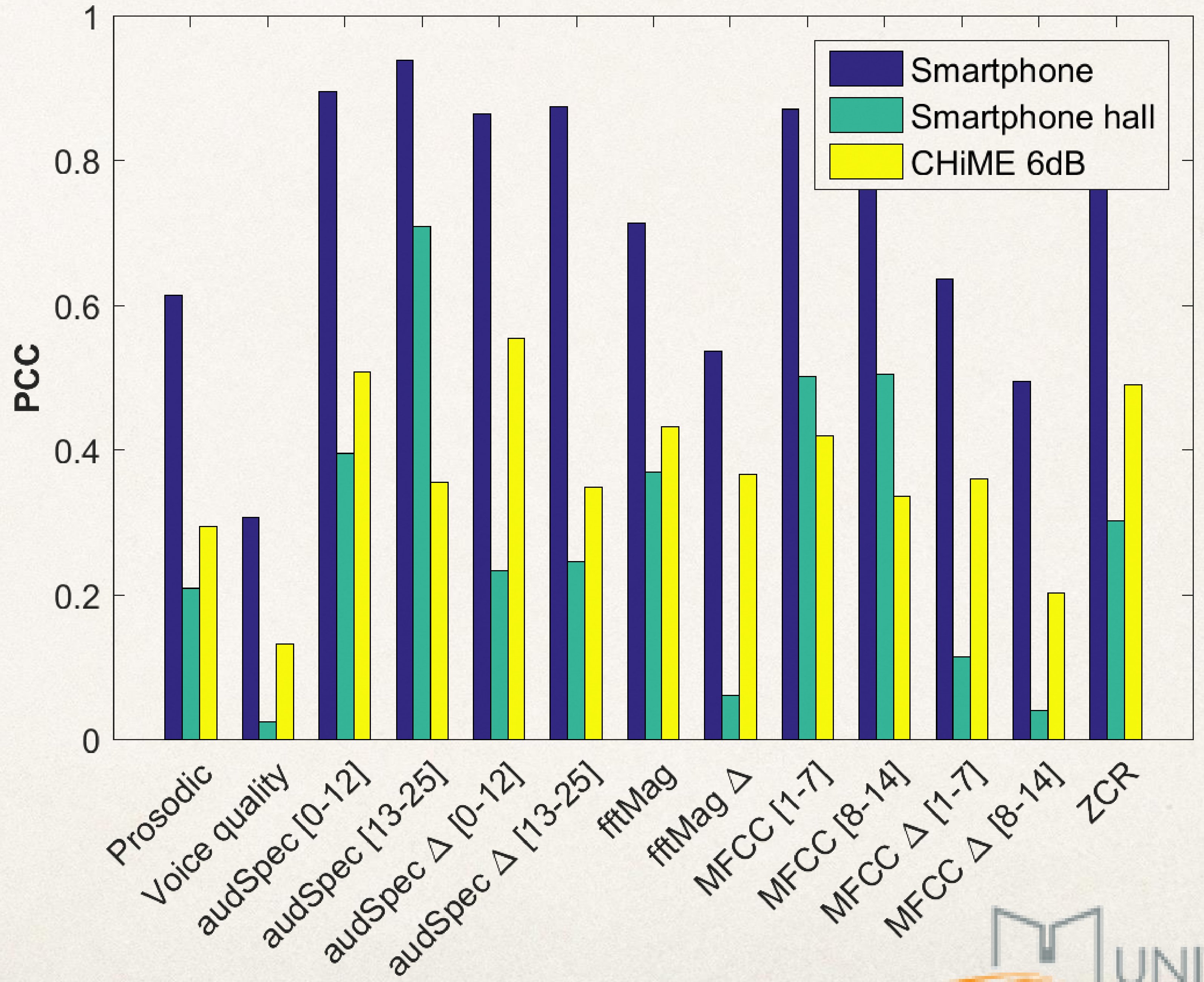
1. convolutive, IR from Google Nexus one
2. + reverberation (convolutive)
3. + CHiME noise (additive, 6 dB SNR)

State-of-the-art feature sets

Correlation:
LLDs clean vs
3 noise types

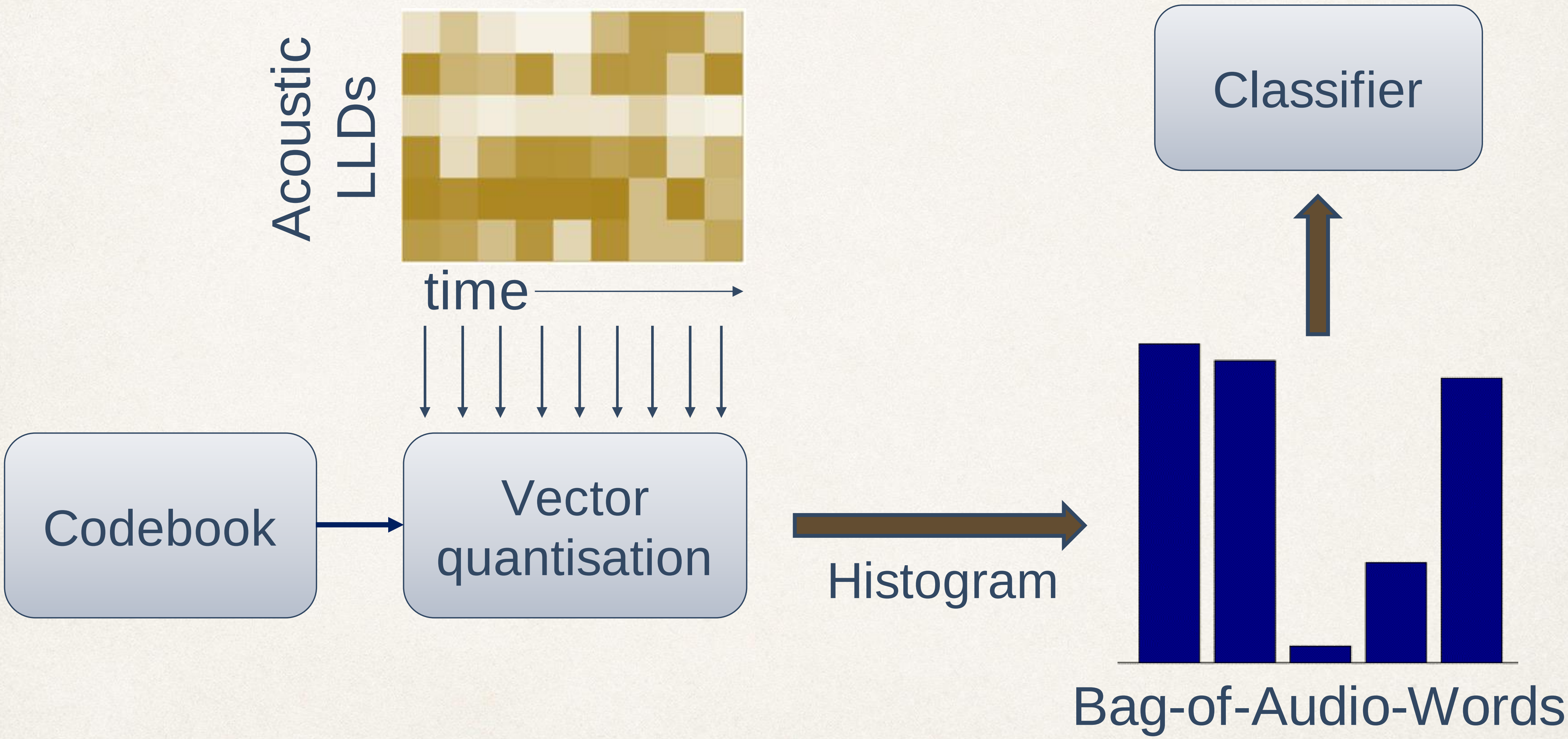
++:
prosody, spectral

--:
Voice quality



Male speaker from RECOLA

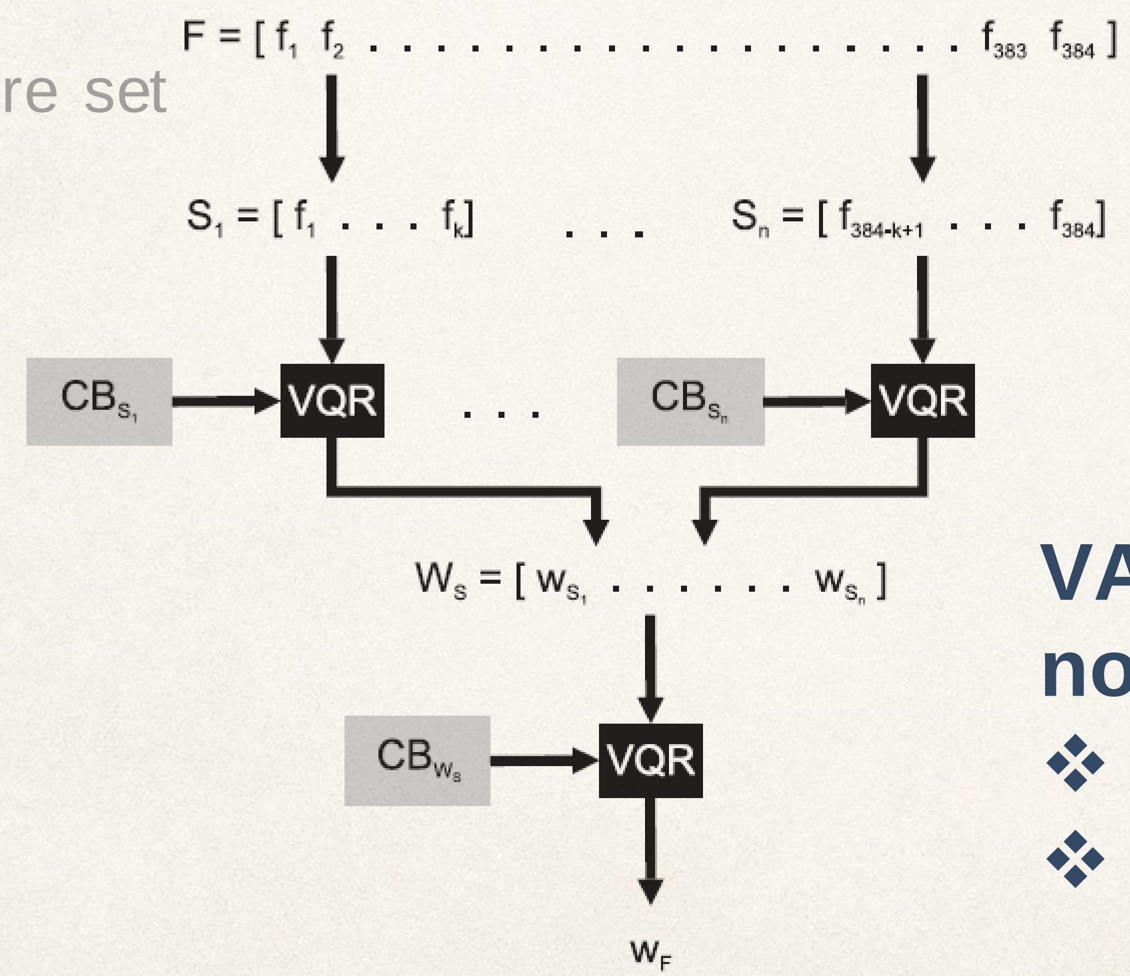
Bag-of-Audio-Words

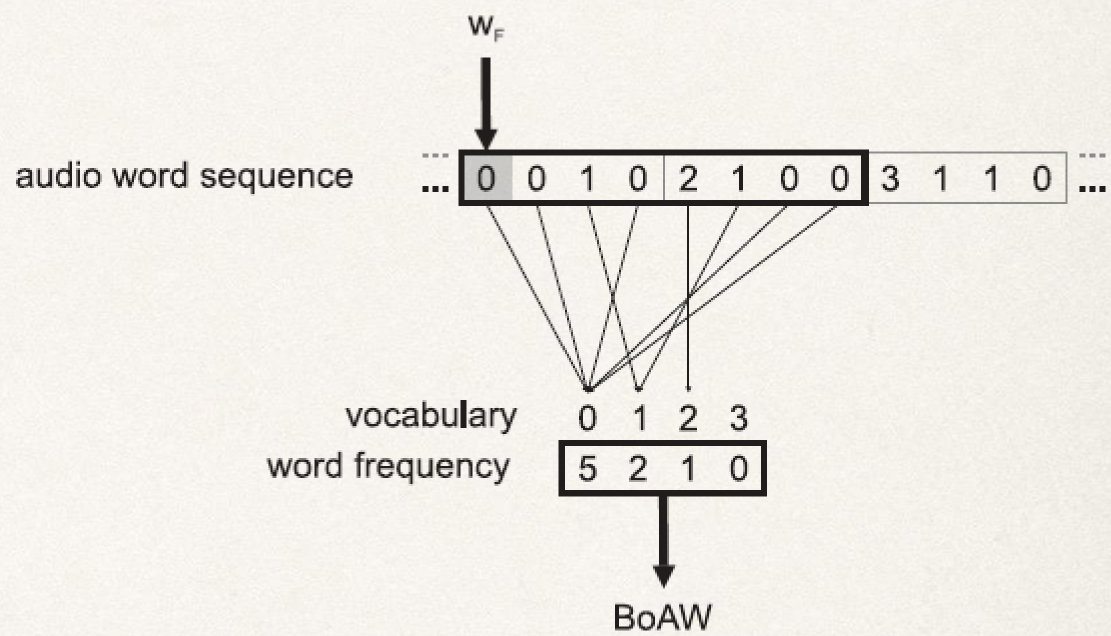


Bag-of-Audio-Words

Split vector quantisation (SVQ)

IS09
Feature set





VAM Corpus (negative vs. nonnegative emotions)

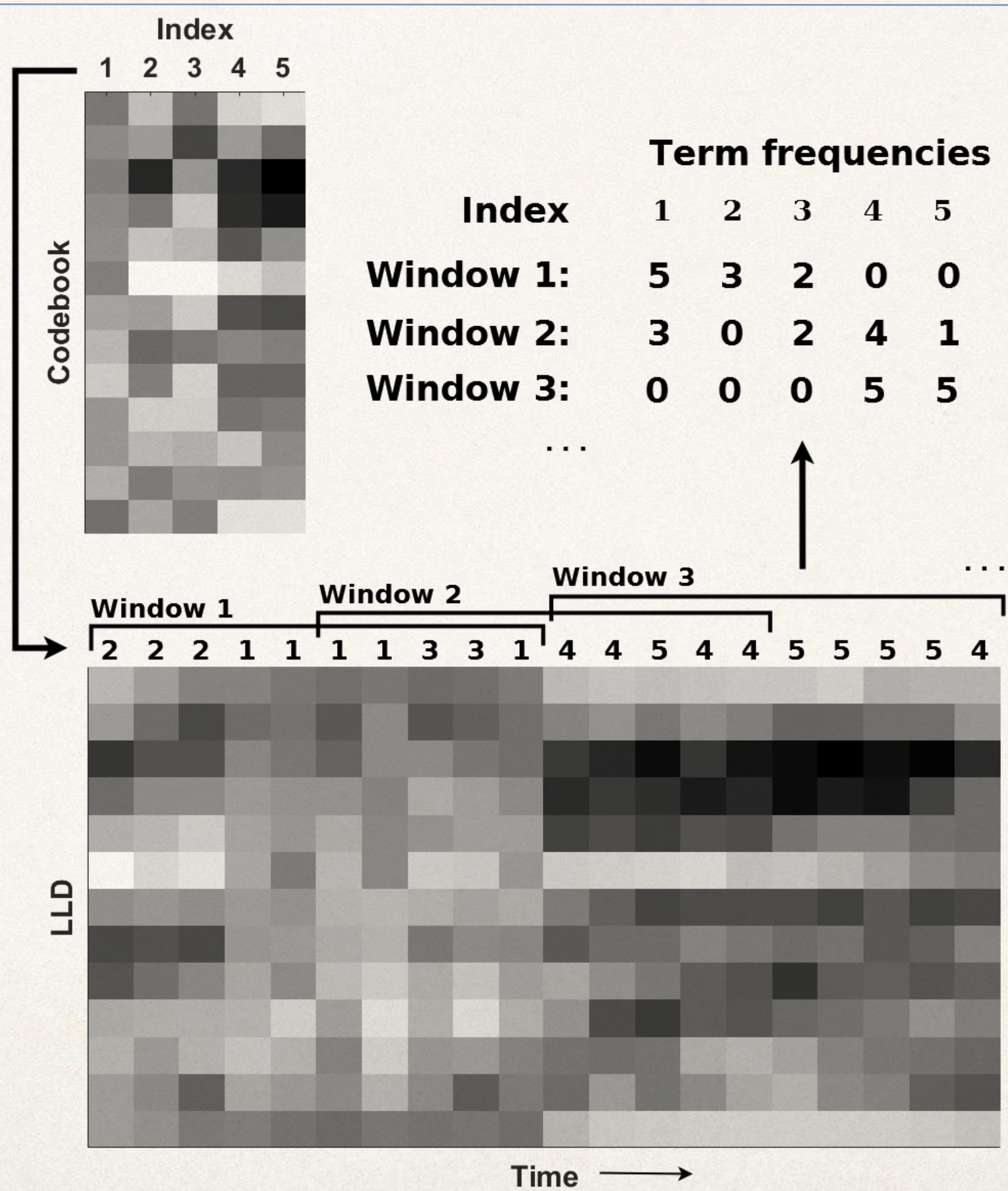
- Raw features (IS09): **54.3 % (UA)**
- BoAW with SVQ (IS09): **64.2 % (UA)**

"Detection of Negative Emotions in Speech Signals Using Bags-of-Audio-Words", ACII, 2015

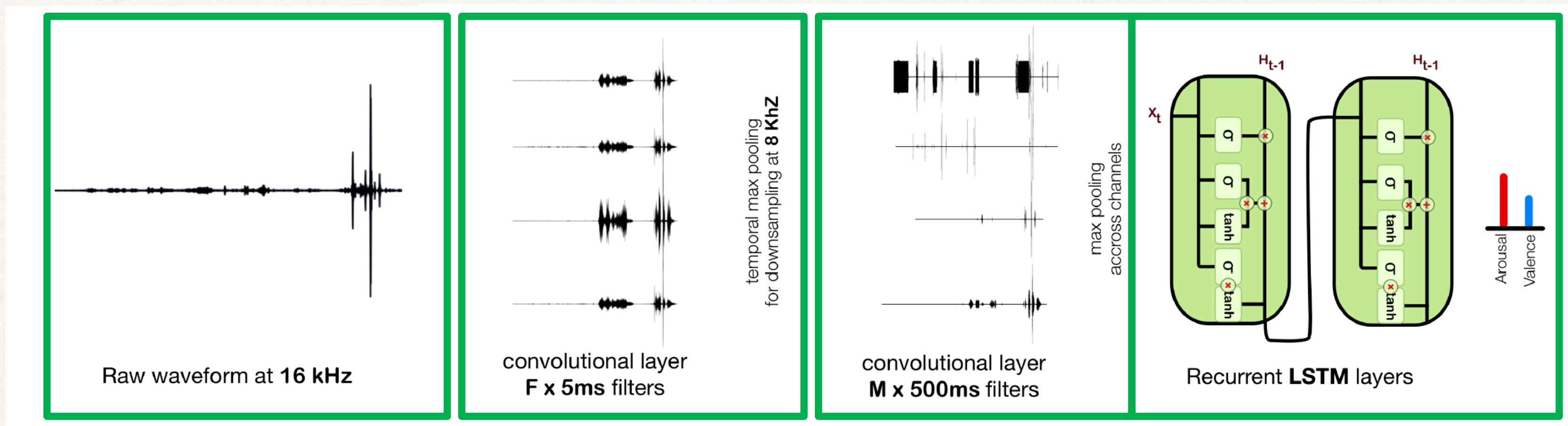
Bag-of-Audio-Words

Time-continuous emotion recognition with BoAW

❖ LLDs:
- MFCC(1-12)
- log-energy



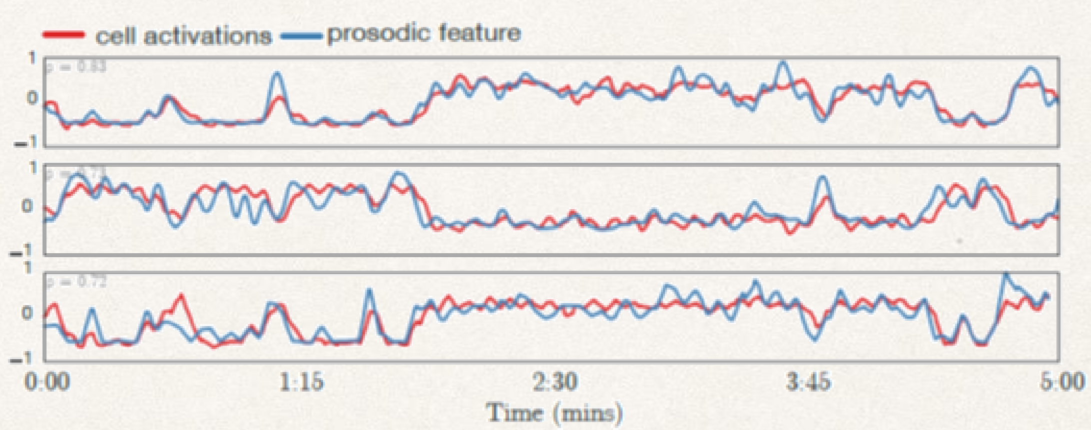
End-2-End Learning



RMS energy range (.81)

loudness (.73)

F0 mean (.72)



First Deep Learning from the *raw signal* in Affective Computing

"Adieu features? End-To-End Speech Emotion Recognition using a Deep Convolutional Recurrent Network", ICASSP, 2016 (Winner SPS StTrGr)

Emotion recognition using BoAW

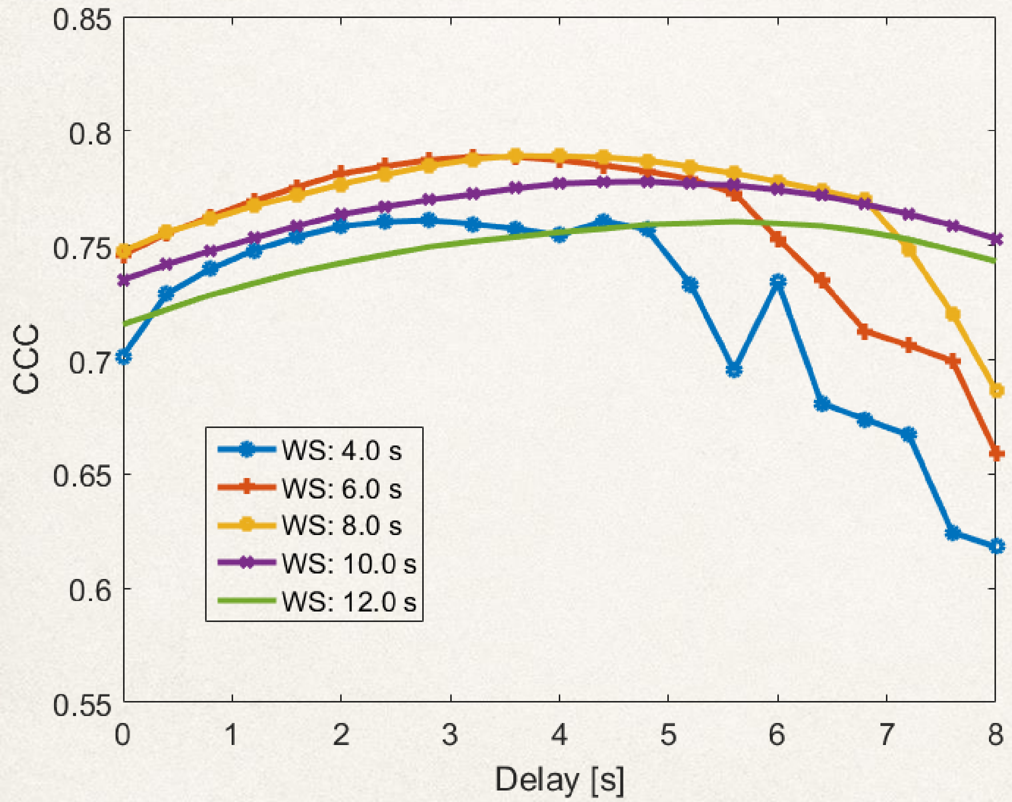
❖ **RECOLA**:
- Dyadic conversation in French
- 46 subjects x 5 min = 230 min
- 6 annotators

Model	CCC	
	Arousal	Valence
	Test	Test
BoAW	<u>.753</u>	.430
BoAW+functionals	.738	<u>.465</u>
Raw signal (CNN+BLSTM)	.686	.261
Baseline AVEC 15 / 16	.382 / .648	.187 / .375

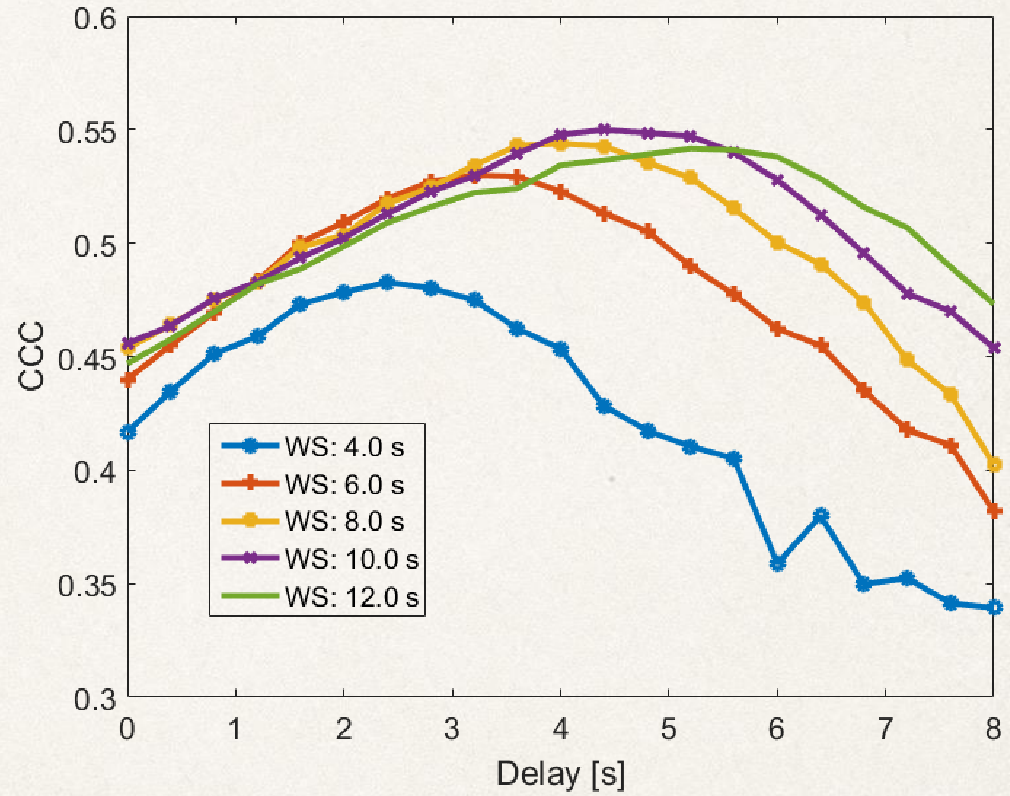
SUBMITTED: "At the Border of Acoustics and Linguistics: Bag-of-Audio Words for the Recognition of Emotions in Speech", Interspeech, 2016

Emotion recognition using BoAW

- Optimisation of **delay** (between shown emotion & gold standard) and **window size**



Arousal



Valence

UNIVERSITÄT PASSAU

Deep Semi-NMF

- ❖ Representation of acoustic features similar to Bag-of-Audio-Words
- ❖ Deep Semi-NMF model learns a hierarchical structure of features

- ❖ Experiments:
  - Berlin Emo-DB
  - Acoustic features: eGeMAPs (88 selected LLDs with functionals)

- ❖ Results:
  - eGeMAPs: 78.2 % (UA)
  - eGeMAPs w/ Deep Semi-NMF: 82.5 % (UA)

Feature enhancement

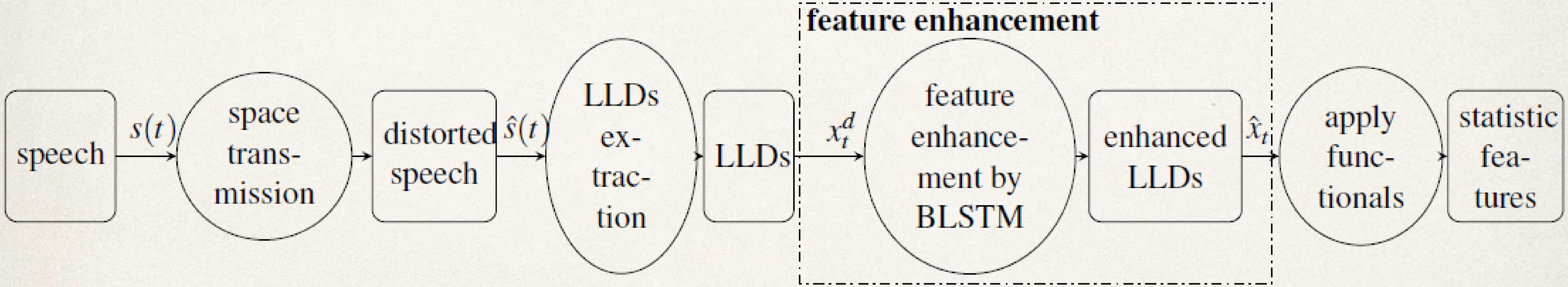
To train de-noising auto-encoders, **stereo** data (noisy recordings with corresponding time-aligned clean recordings) are required.

1. Data generated **artificially**, simulating various room reverberation parameters and additive ambient noise ✓

2. Artificial data **augmented by real-life data** by means of semi-supervised learning ✓

Feature enhancement

- ❖ Acoustic features corrupted by noise (recordings `in the wild`)
- ❖ Denoising autoencoders: remove distortions from features



Feature enhancement

❖ **Results:**

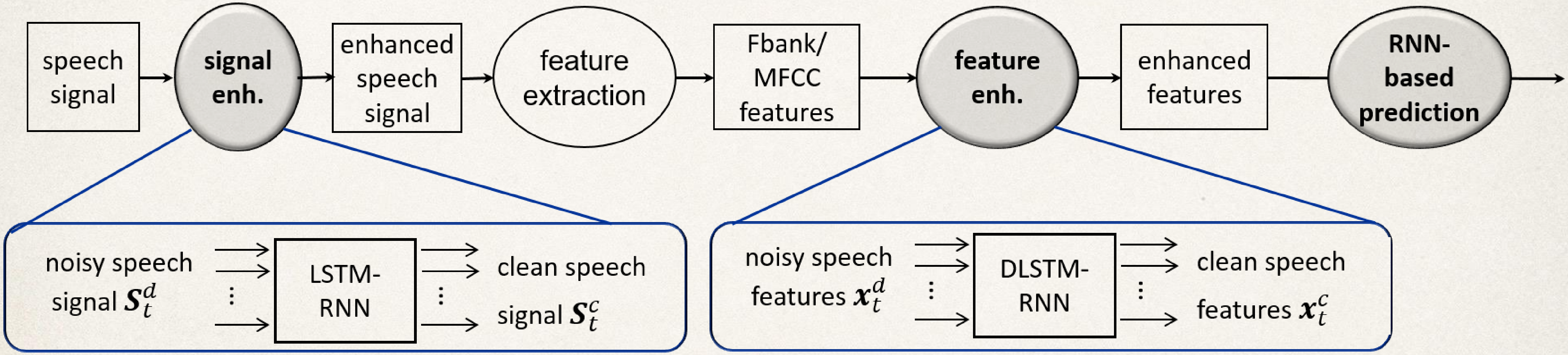
- Database: RECOLA
- Task: Arousal
- Baseline: LSTM
- Noise: CHiME noise w/ SNRs (12 dB → 0 dB)

		clean	**12 dB**	**9 dB**	**6 dB**	**3 dB**	**0 dB**
No feature enhancement	**CCC**	<u>.661</u>	.556	.526	.472	.420	.329
Feature enhancement	**CCC**	.467	<u>.648</u>	<u>.631</u>	<u>.612</u>	<u>.521</u>	<u>.368</u>

Feature enhancement

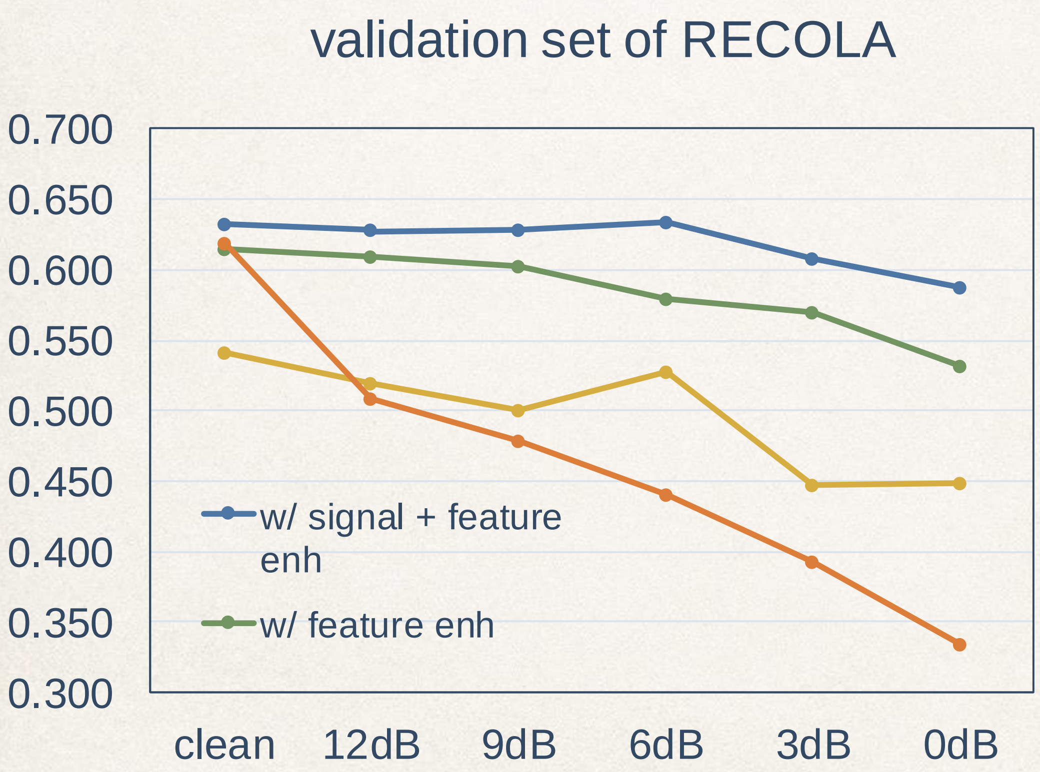
Enhancement of the raw speech signal

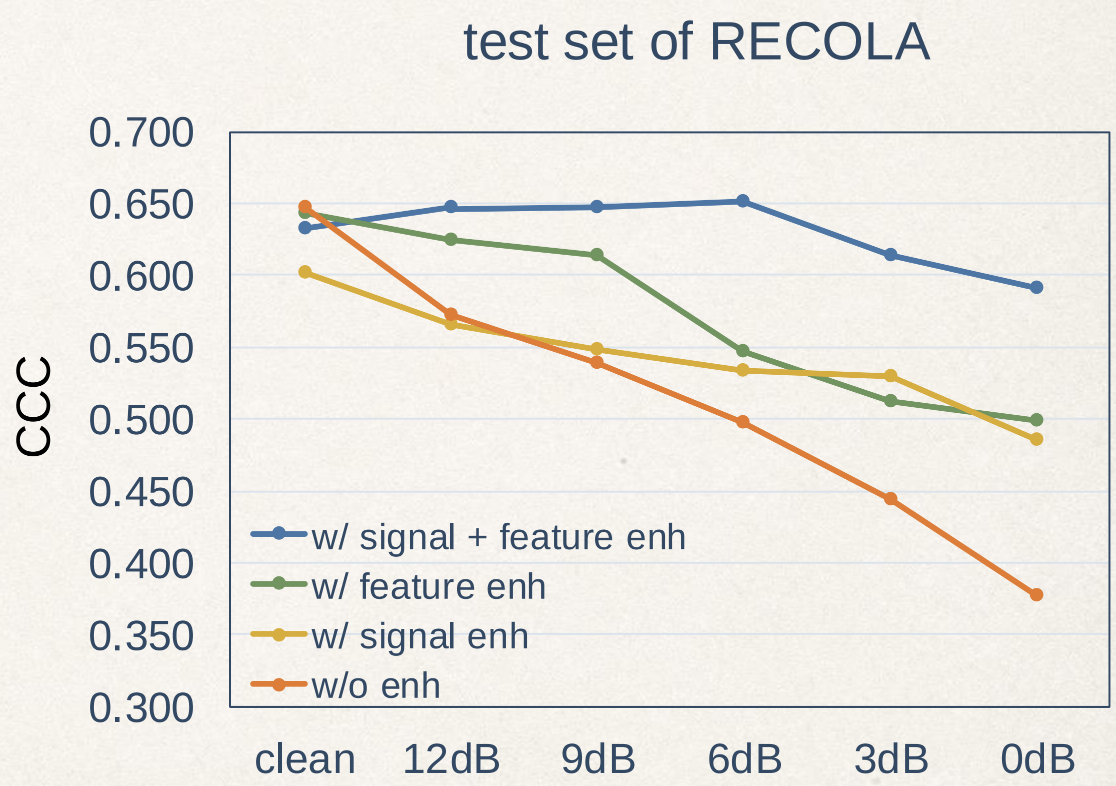
Deep LSTM-RNN



Feature enhancement

Speech Enhancement by Deep LSTM-RNN for Continuous Emotion Regression





Environmentally robust acoustic features

1. **Selection of features** that are correlated with target labels in noisy data
   - State-of-the-art acoustic emotion recognition feature sets
   - Bag-of-audio-words (BoAW) representations (generated, e.g., by Vector Quantisation or Deep Semi-NMF)

2. **Feature enhancement** by deep de-noising auto-encoders such as LSTM-RNN
   - On raw spectral features
   - Learning of non-linear distortions in
     (a) Emotion-related features
     (b) BoAW representations

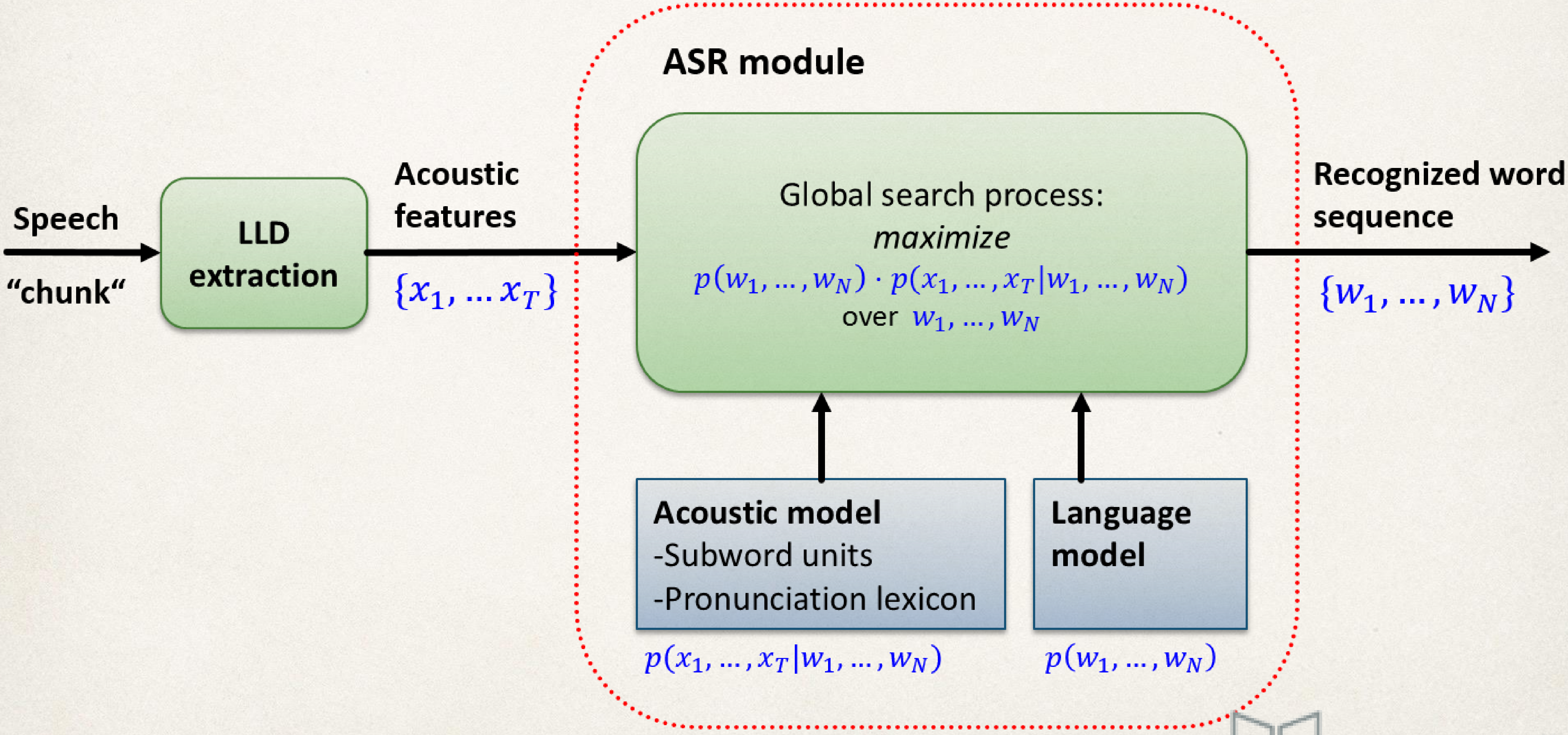
Objectives

- Task 2.1: Environmentally robust acoustic features
- Task 2.2: Environmentally robust visual features
- Task 2.3: Cross-lingual language-related features

  → Improved acoustic-linguistic feature extractor

  (D2.3, February 2016, M13)

Automatic speech recognition



Automatic speech recognition

- Based on Kaldi toolkit
- Features: MFCCs + Δ + ΔΔ

- AM: Context-dependent *triphone models* trained by hybrid DNN-HMM
- LM: *Kneser-Ney smoothed backoff 4-gram LM*

- Training: *LibriSpeech* (1000 hours of audiobooks, 2.3k speakers)

- Pre-trained LM, trained on 14.5k books taken from *Project Gutenberg*

Automatic speech recognition

❖ Results on **LibriSpeech** corpus:

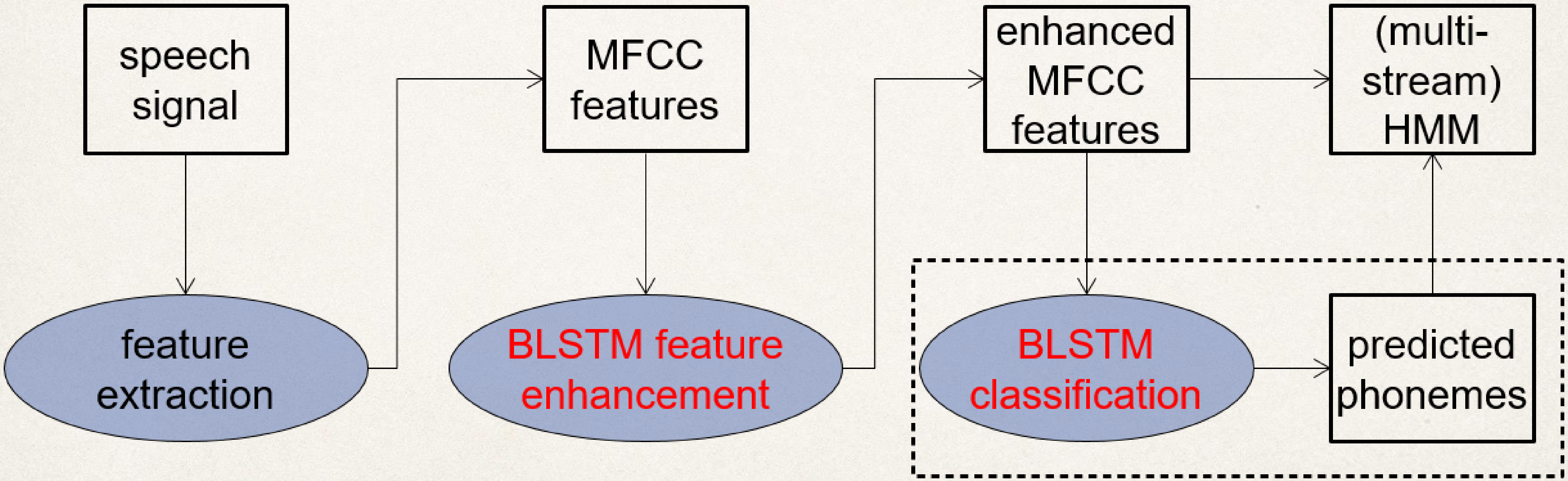
Data set	**WER (%)**	
	Panayotov, ICASSP 2015	**Uni Passau**
Test clean	5.51	<u>5.30</u>
Test other	13.97	<u>13.68</u>

Panayotov et al.: LibriSpeech: An ASR Corpus Based on Public Domain Audio Books, ICASSP, 2015

❖ Training corpora in-domain: Buckeye, COSINE

Feature enhancement for ASR

- AM Enhancing for noisy/reverberated speech recognition
- Feature enhancement (FE) + multi-stream (MS) by BLSTM-RNN



Feature enhancement for ASR

Experimental results

- Buckeye corpus (spontaneous)
- train/dev/test = 20.7 h / 2.6 h / 2.4 h
- vocab size = 9.1k words
- CHiME noise
- BLSTM-RNN: 3 hidden layers
  Features: MFCC 1-12 + log-energy

	SNR [dB]							
WER [%]	**-6**	**-3**	**0**	**3**	**6**	**9**	**Avg.**	**Clean**
Clean	78.8	76.9	74.6	72.2	69.2	65.5	72.9	49.0
Noisy	74.8	72.6	69.9	68.4	65.8	63.0	69.1	56.2
Noisy + FE	**67.5**	**65.6**	**62.8**	**61.4**	**59.1**	**56.9**	**62.2**	**55.6**

Feature enhancement for ASR

Experimental results

- WSJ0 corpus
- Reverberated by Aachen IR database
- Training w/ reverberation (w/o stairway)

	Tested on					
WER [%]	**Stairway 1_90**	**Stairway 2_90**	**Stairway 3_45**	**Stairway 3_90**	**Stairway 1_135**	**Avg.**
Baseline	40.6	70.0	93.3	86.5	89.5	76.0
+ FE	19.6	30.0	63.0	38.5	51.5	40.5
+ re-training	21.5	28.5	47.1	32.4	38.7	33.6
Reverb. Train	19.4	30.1	56.7	43.2	51.9	40.3
+ FE	**18.5**	**24.6**	**42.5**	**29.4**	**36.1**	**30.2**

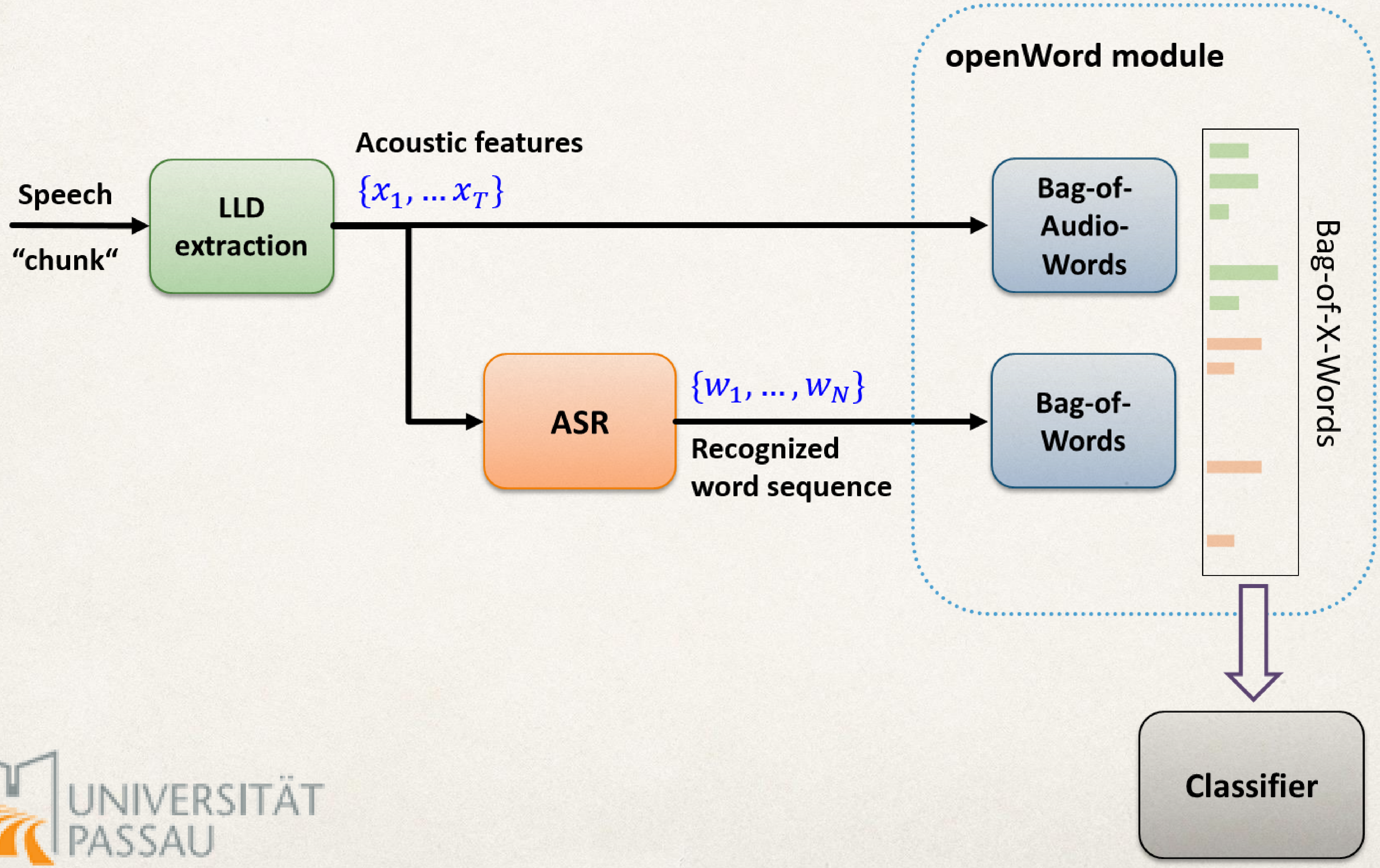
Feature enhancement for ASR

Experimental results

- WSJ0 corpus
- Track 2 of CHiME 2013

	SNR [dB]						
WER [%]	**-6**	**-3**	**0**	**3**	**6**	**9**	**Avg.**
Baseline	70.4	63.1	58.4	51.1	45.3	41.7	55.0
FE	62.0	54.6	50.1	44.7	40.3	37.0	48.2
FE + re-training	56.9	50.3	45.1	39.3	34.6	31.8	43.0
MS	58.6	50.1	43.9	37.1	32.7	28.3	41.8
FE+re-training+MS	**56.1**	**48.3**	**40.5**	**35.9**	**31.1**	**27.7**	**39.9**

Acoustic-linguistic features



openXBOW – Bag-of-X-Words tool

- Implemented in Java
- Fast and flexible
- Multiple input/output formats: ARFF, CSV, Libsvm
- JUnit tests
- Open source: GitHub repository

openXBOW – Bag-of-X-Words tool

- Generates single feature vector of acoustic, visual & textual features

- Preprocessing: standardisation, normalisation, VAD
- Windowing
- Supervised codebook generation
- Split vector quantisation
- Multiple assignments
- Soft vector quantisation
- Term-frequency/inverse document-frequency weighting
- N-grams, stopping
- Histogram normalisation

openXBOW – Bag-of-X-Words tool

LLDs

Preprocessing:
- (V)AD
- Standardisation
- Normalisation

LLD-codebook generation:
- Kmeans
- Kmeans++
- Random sampling
- Random sampling++

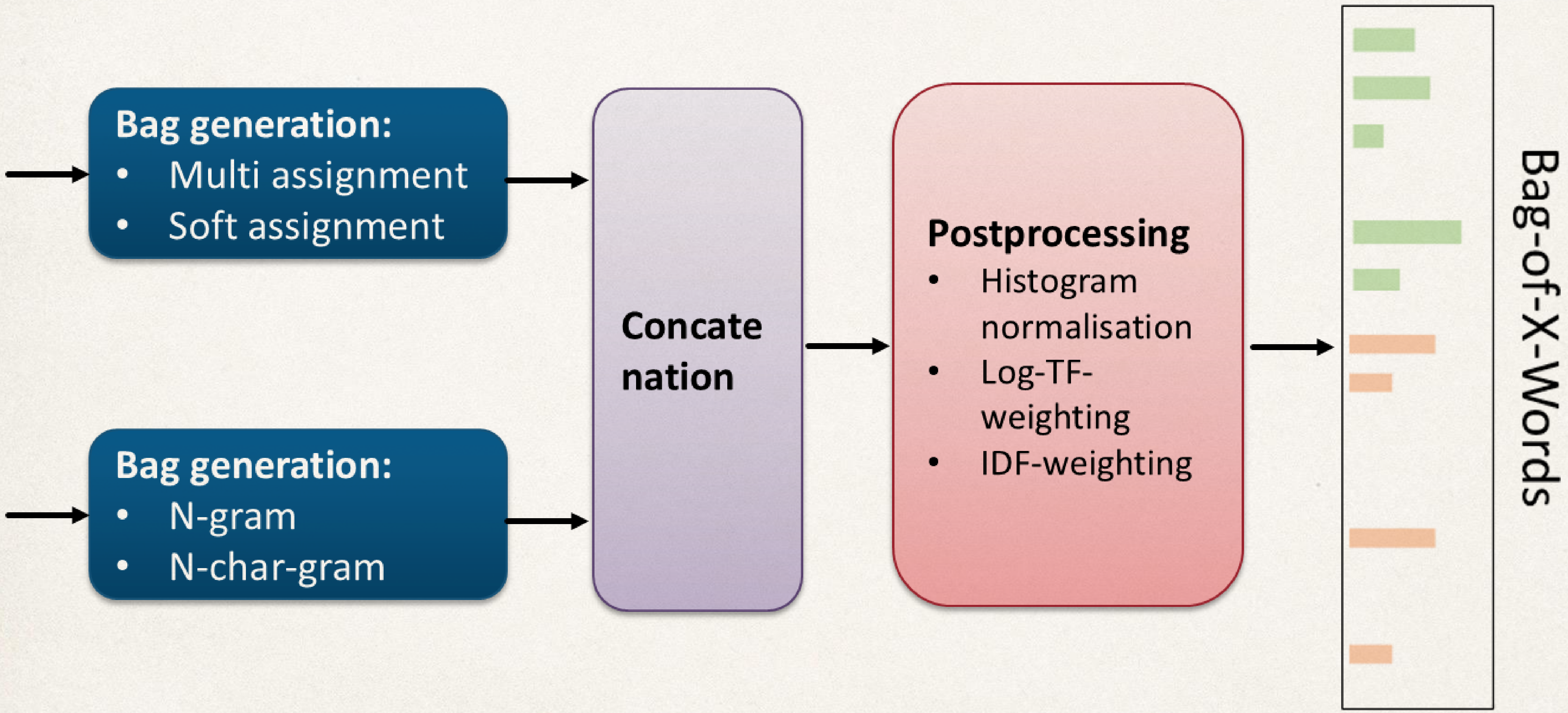
Bag generation:
- Multi assignment
- Soft assignment

Transcriptions (text)

Dictionary generation:
- MinTermFreq
- Stopping

Bag generation:
- N-gram
- N-char-gram

openXBOW – Bag-of-X-Words tool



Natural language processing with openXBOW

- **Gender recognition** on SEWA (from transcriptions)
  2-grams, log-IDF weighting, Naïve Bayes (10-fold CV):

  - British: 72.7 % (UA)
  - German: 75.6 % (UA)

- **Cross-language gender recognition**
  multilingual dictionaries (10-fold CV):

  - British → German: 62.1 % (UA)
  - German → British: 59.1 % (UA)

Natural language processing with openXBOW

- **Sentiment analysis:**
  Thinknook database
  1.5 Mio **tweets**, +/- sentiment

- WA: 75.8 % (UA: 74.8 %)
- WA: 75.0 % is state-of-the-art by *Thinknook*

"@MariaLKanellis U know what I was thinking about? What
u sang at Otiz, was it one of your secret recordings?
Loved it anyway... Jay" → positive

Acoustic landmarks

- Overcome the problem of language dependence in ASR
- Extract acoustic landmarks from *f0 / energy contours*
- Find significant changes in *speech production* or *perception*
- More *robust* to noise and acoustic variations due to emotional encoding

1. **Voiced/unvoiced segments:**
   Based on continuity of the *f0 contour*

2. **Pseudo-vowels:** Unsupervised detection of vocalic nuclei

3. **P-center:** Rythmic prominence of speech

Acoustic landmarks

- Landmarks constitute a language independent dictionary
  → BoW features are generated

- **Results on the SEMAINE corpus:**

Dimension	Partition	UA (%)
Arousal	Development	59.6
	Test	60.2
Valence	Development	56.7
	Test	55.6

Acoustic-linguistic features

Retrieve features related to linguistic content, largely **language** and **context-independent**

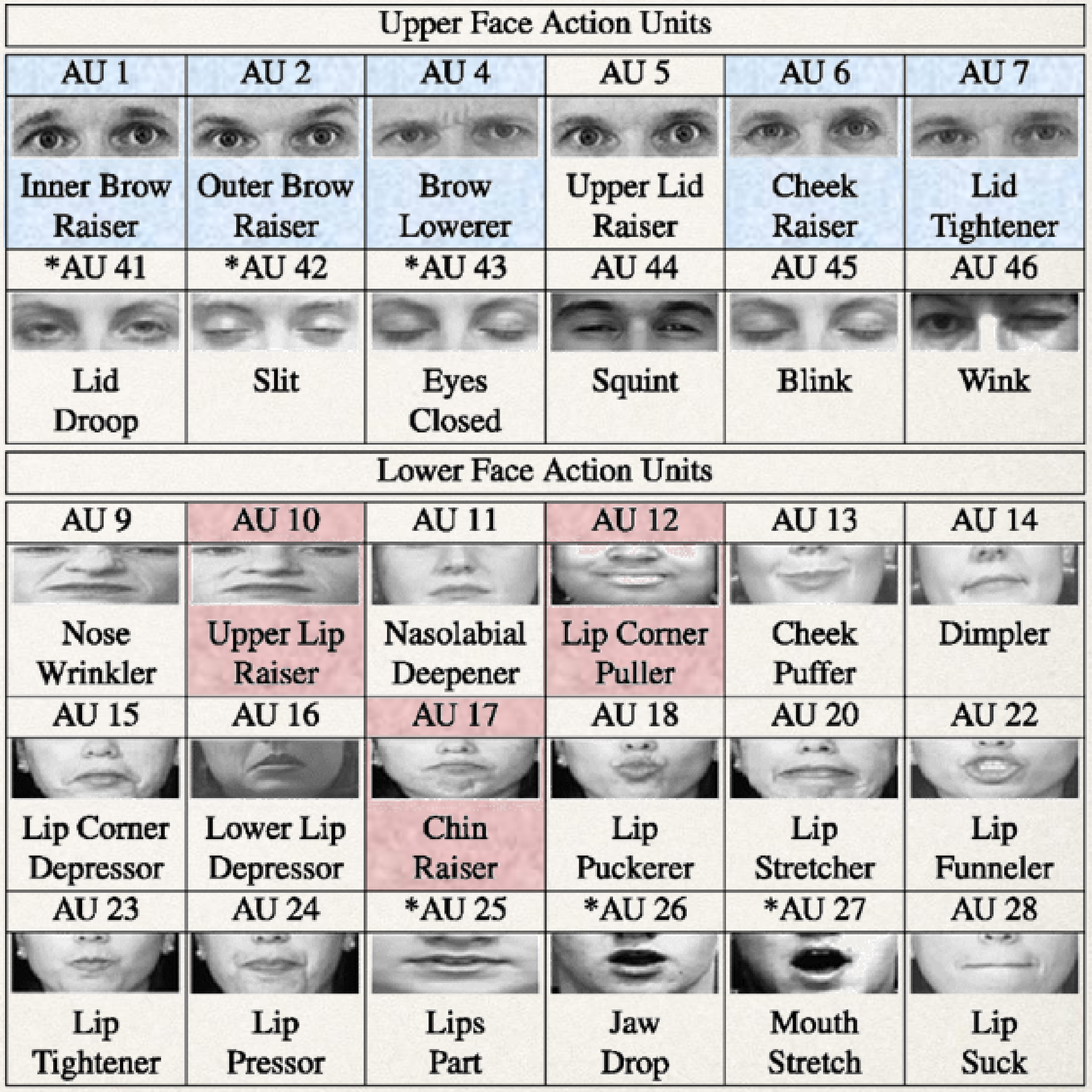
- Multi-lingual dictionaries for BoAW generated from fully automatic **syllabification** of unlabelled multi-lingual speech data ✓
- Generation of language-independent bag-of-words (BoW) type representations by ASR, natural language processing and machine translation systems: stemming, dictionary lookup and/or machine translation ✓
- Linguistic Inquiry and Word Count (LIWC) features can be generated from ASR outputs in twelve different languages

Not considered as it is not open source

Further work (selected)

Face Reading from Speech – Predicting Facial Action Units from Audio Cues
Interspeech 2015

UA [%]	Mean
SVM	57.3
Deep NN	<u>65.0</u>



Further work (selected)

Cross Lingual Speech Emotion Recognition Using Canonical Correlation Analysis on Principal Component Subspace
IEEE ICASSP 2016

Cross-Language Acoustic Emotion Recognition: An Overview and Some Tendencies
IEEE/AAAC ACII 2015

Enhanced Semi-supervised Learning for Multimodal Emotion Recognition
IEEE ICASSP 2016

Continuous Estimation of Emotions in Speech by Dynamic Cooperative Speaker Models
IEEE Transactions on Affective Computing

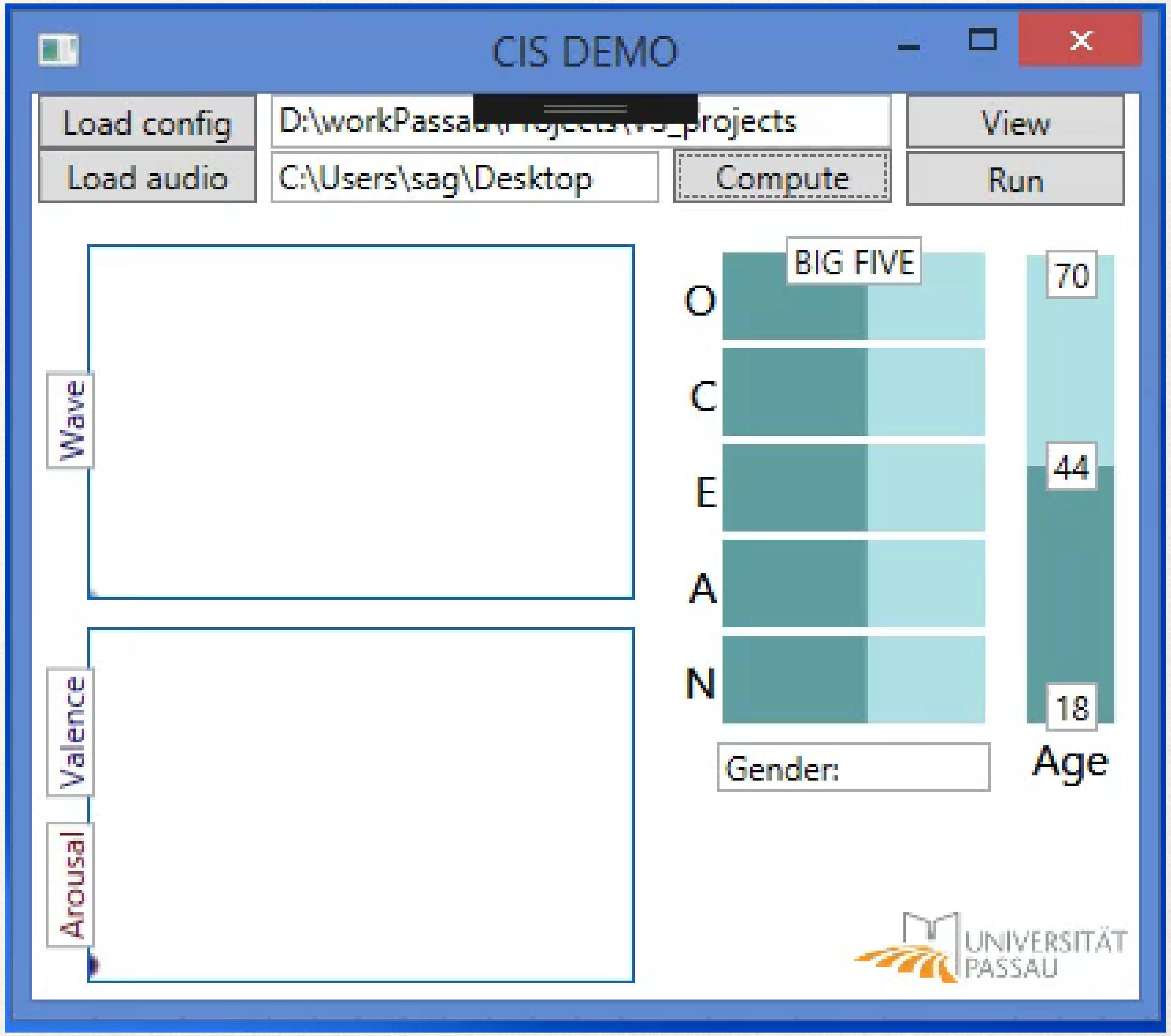
Further work (selected)

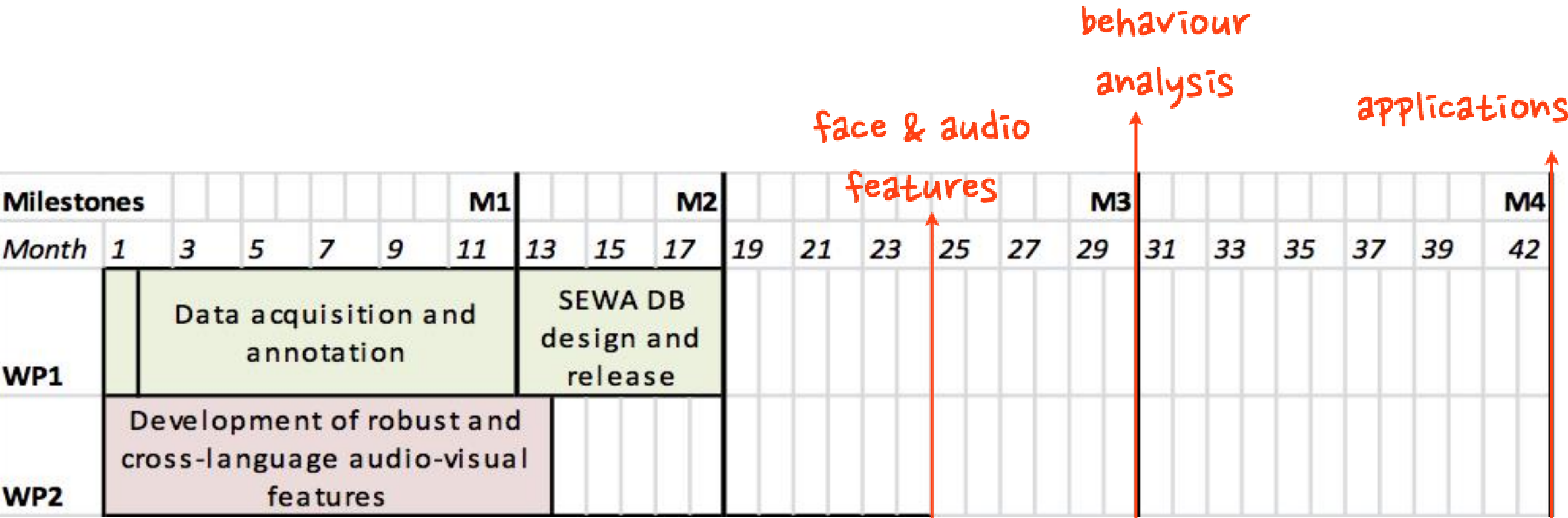
AVEC 2015 – The First Affect Recognition Challenge Bridging Across Audio, Video, and Physiological Data
ACM Multimedia 2015

The ICL-TUM-PASSAU Approach for the MediaEval 2015 “Affective Impact of Movies” Task
MediaEval 2015
(Winning team (1./2./3. arousal/valence/violence – 22 registered teams))
Video: CNN of 1000 objects to detect (ILSVRC 2013)

Demo

Clinton & Trump





→ Improved acoustic feature extractor (D2.1, Oct 15, M9) ✓

→ Robust visual feature extractor (D2.2, Feb 16, M13) ✓

→ Improved acoustic-linguistic feature extr. (D2.3, Feb 16, M13) ✓

WP2:
Low-level Feature Extraction

Björn Schuller

sewa

Automatic Sentiment Analysis in the Wild