Computational Paralinguistics in Everyday Environments

CHiME WORKSHOP

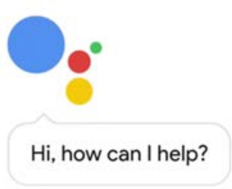
Björn W. Schuller

Imperial College London / UK
Machine Learning Group

University of Passau / Germany
Chair Complex & Intelligent Systems

audEERING GmbH / Germany

vs



Speech Recognition

1950 single speaker, digits

1970 1000 words

1980 several 1000 words

1990 trained dictation

2000 robust, million words

2010 everyday usage

Speaker Classification

1960 speaker identification

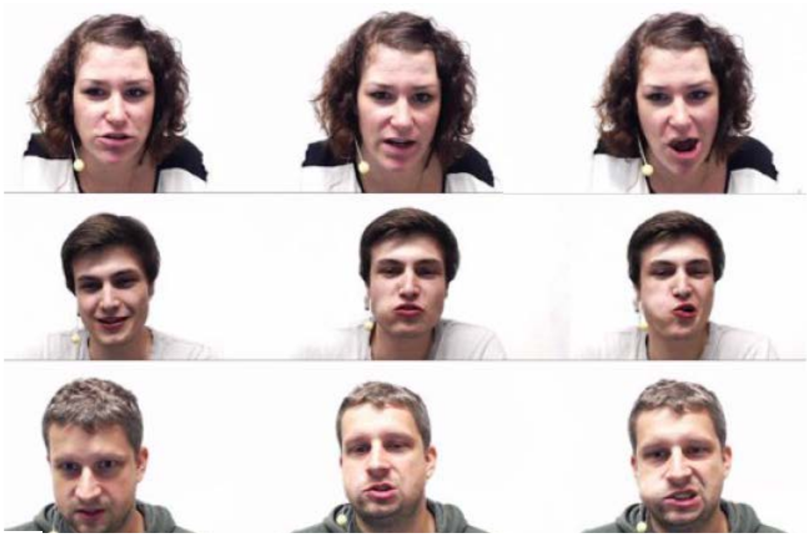
1997 single speaker, emotion

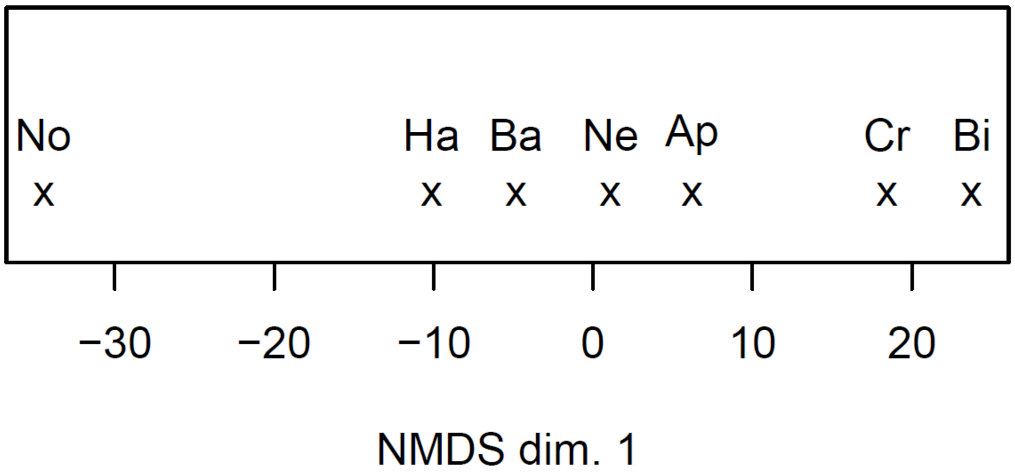
2010 few isolated states/traits

2020 everyday usage?

Paralinguistics.

- **Speech Under Eating & Food**
  30 subjects, 6 food types, +ASR features





R^2	
Crispness	.562

*"The Interspeech 2015 Computational Paralinguistics Challenge: Nativeness, Parkinson's & Eating Condition", **Interspeech**, 2015.*

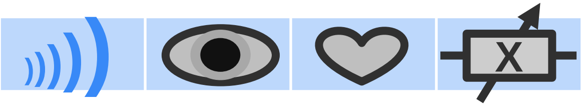
Paralings.

		# Classes	%UA/*AUC/+CC
2016	Deception	2	72.1
	Sincerity	[0,1]	65.4+
	Native Lang.	11	82.2
2015	Nativeness	[0,1]	43.3+
	Parkinson's	[0,100]	54.0+
	Eating	7	62.7
2014	Cognitive Load	3	61.6
	Physical Load	2	71.9
2013	Social Signals	2x2	92.7*
	Conflict	2	85.9
	Emotion	12	46.1
	Autism	4	69.4

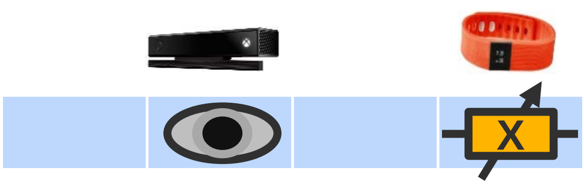
Paralings.

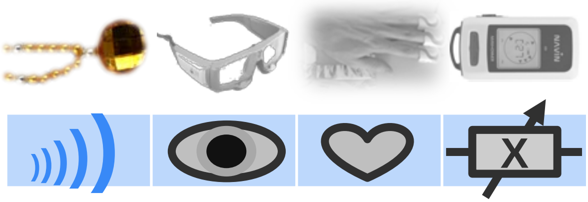
		# Classes	%UA/*AUC/+CC
2012	Personality	5x2	70.4
	Likability	2	68.7
	Intelligibility	2	76.8
2011	Intoxication	2	72.2
	Sleepiness	2	72.5
2010	Age	4	53.6
	Gender	3	85.7
	Interest	[-1,1]	42.8+
2009	Emotion	5	44.0
	Negativity	2	71.2

Paralings.



- **Pseudo Multimodality**



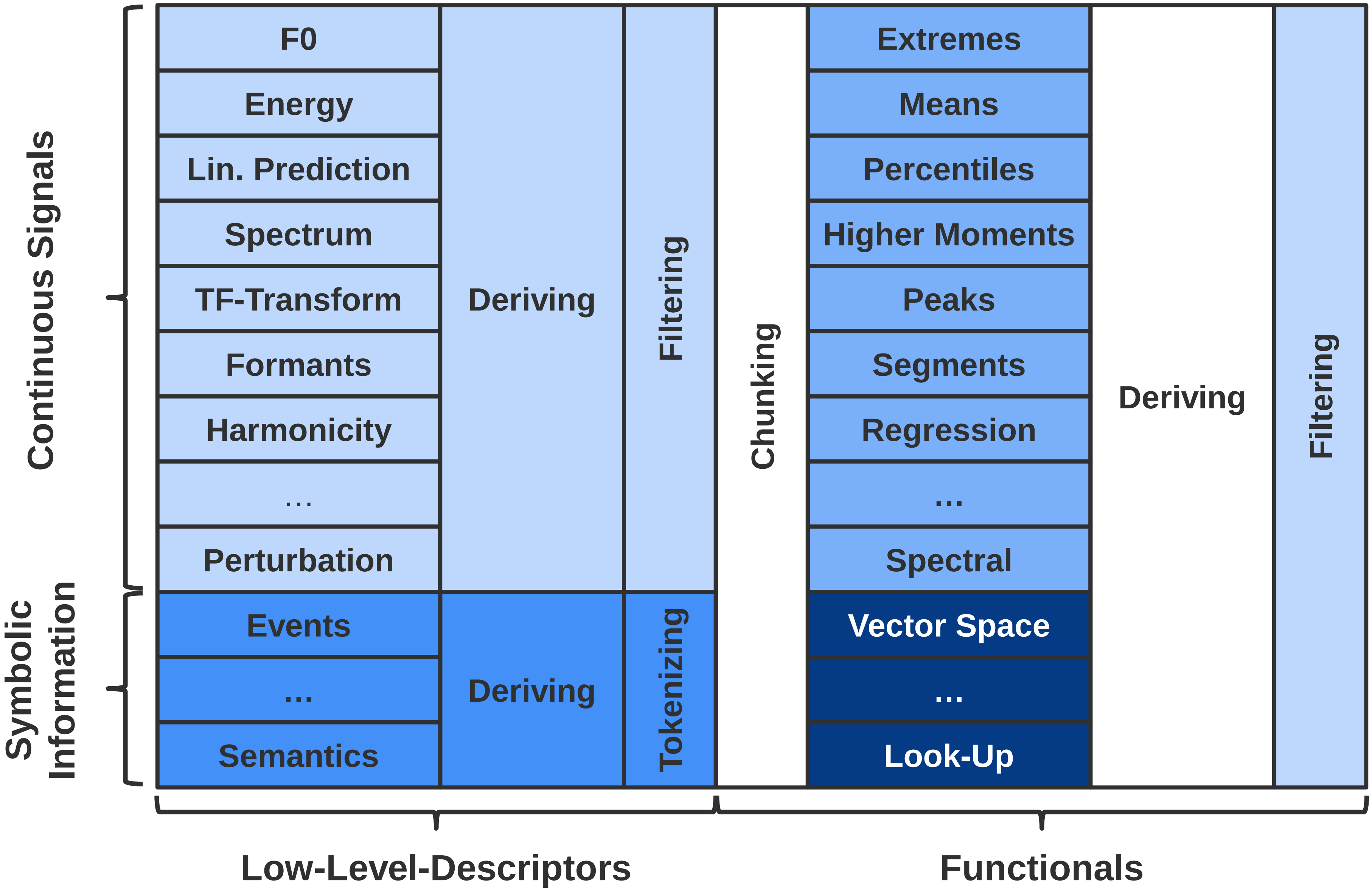


	*MAE $^+$CC %UA
Heart Rate	8.4*
Skin Conductance	.908$^+$
Facial Action Units	65.0
Eye-Contact	67.4

Acoustic Robustness.

Features

openSMILE:)



Feature Robustness

Pitch Detection

PDA in Time Domain

- Determination of 1. Partial
- Analysis of Time Signal
- Simplification of structure

PDA by Short Time Principle

- Correlation
- Analysis in Frequ. domain
- Maximum Likelihood

Feature Robustness

- **Pitch (FAU Aibo Corpus)**

  67.9% voiced frames, ~ 6% erroneous pitch (>10 % deviation)

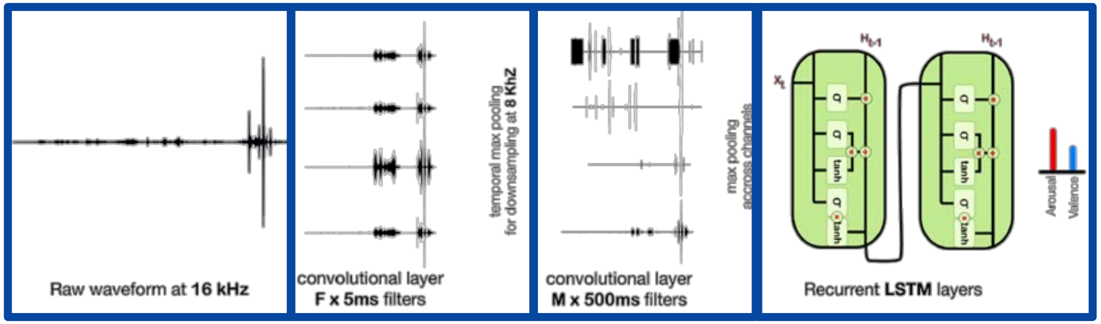
type	# frames	percent
identical	574 485	93.67
small errors	452	0.07
voiced errors	8 804	1.43
unvoiced errors	1 877	0.30
octave errors ↓	23 498	3.83
octave errors ↑	239	0.03
other gross errors	3 923	0.63

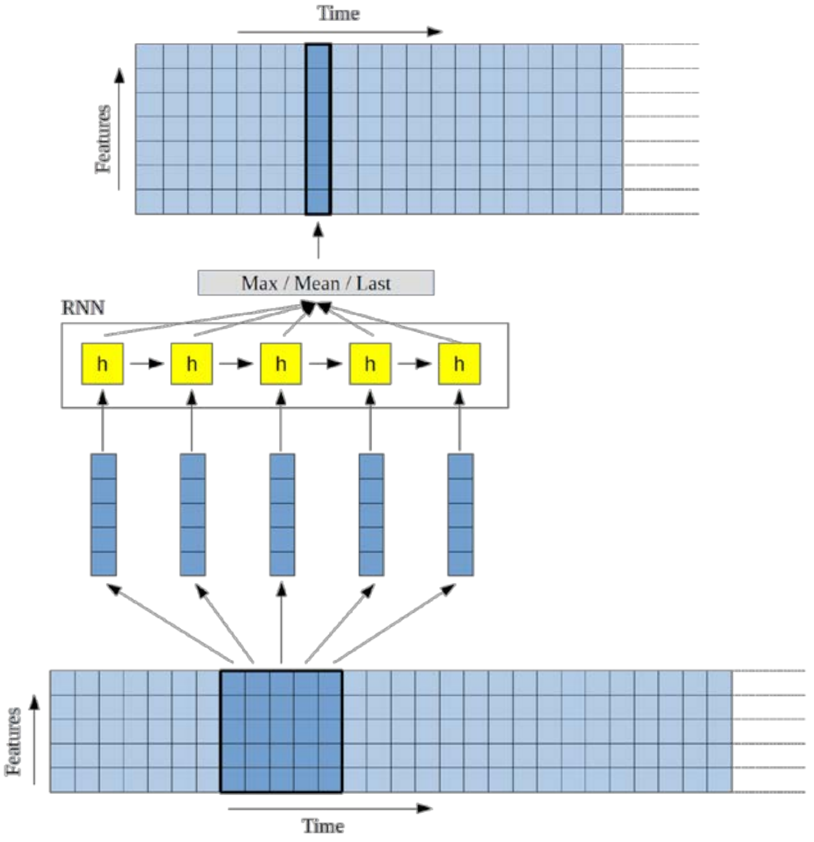
~2.0% loss in recognition accuracy (duration features less affected)

"The Impact of F0 Extraction Errors on the Classification of Prominence and Emotion", ICPhS, 2008.

End-2-End Learning

- **Convolutional RNNs**



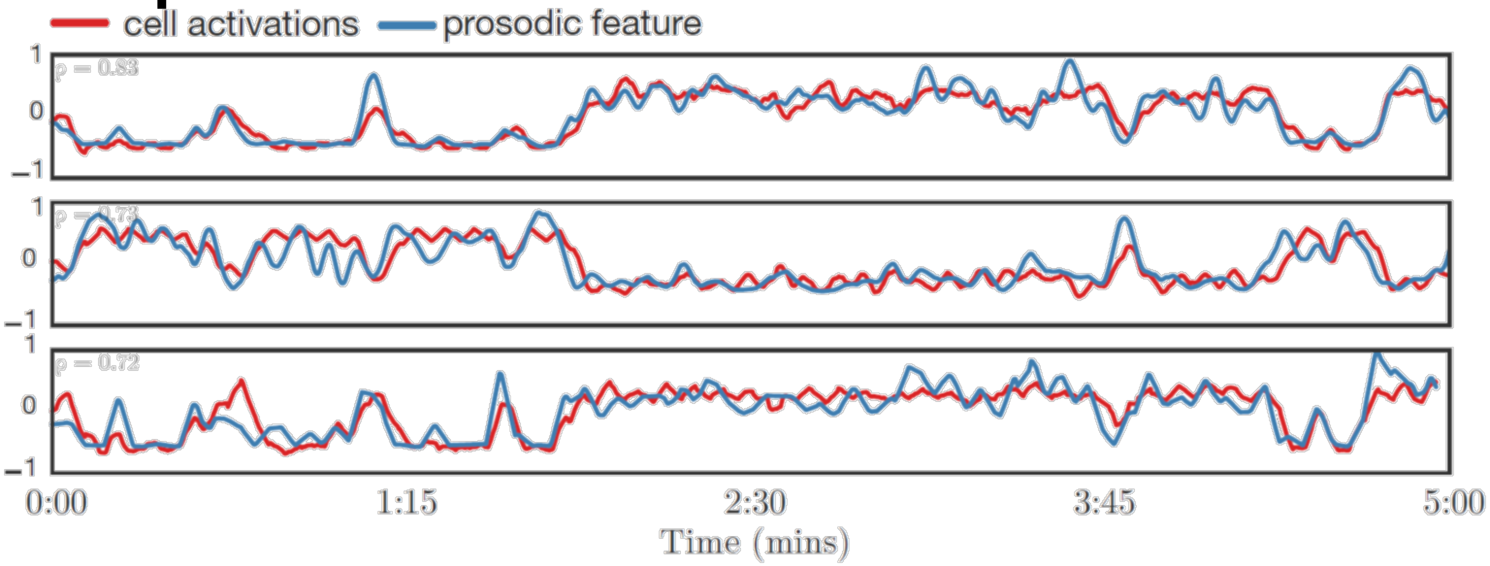


Arousal	CC
Baseline	.366
Deep CRNN	.686

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

End-2-End Learning

- **Example: AVEC 2016**



energy range (.77), loudness (.73), F0 mean (.71)

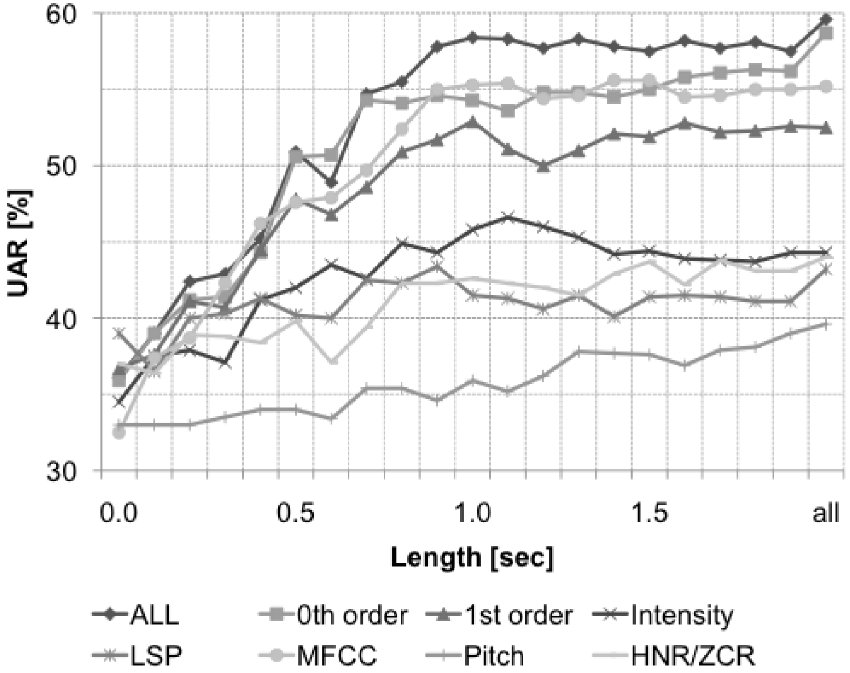
"Adieu Features? End-to-End Speech Emotion Recognition using a Deep Convolutional Recurrent Network", ICASSP, 2016.

Timing

- **Gating**

  Implications for feature normalization, on-set detection, etc.

  One second suffices?



(a) LLD: unweighted average recall.

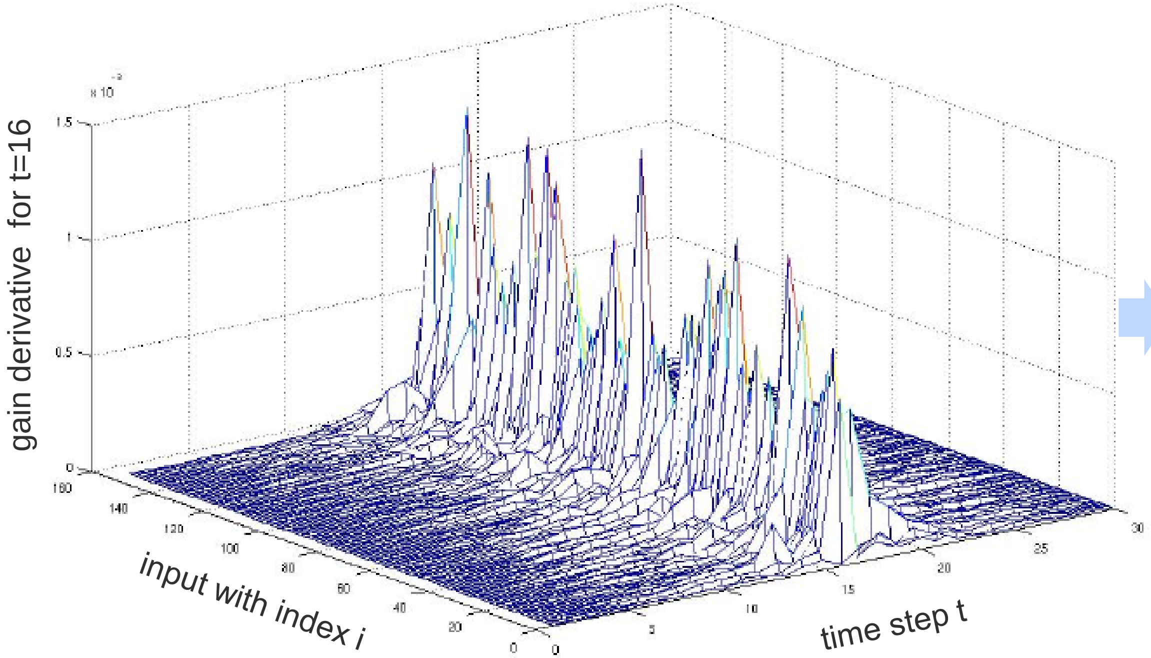
[Schuller;2010]

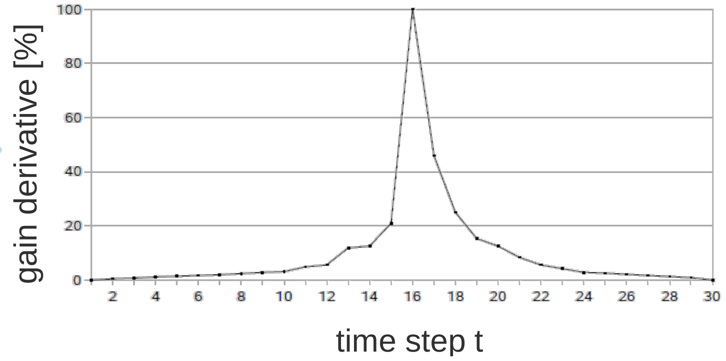
"Incremental Acoustic Valence Recognition: an Inter-Corpus Perspective on Features, Matching, and Performance in a Gating Paradigm", Interspeech, 2010.

Timing

- **Learning Temporal Context**

  LSTM: Sequential Jacobian





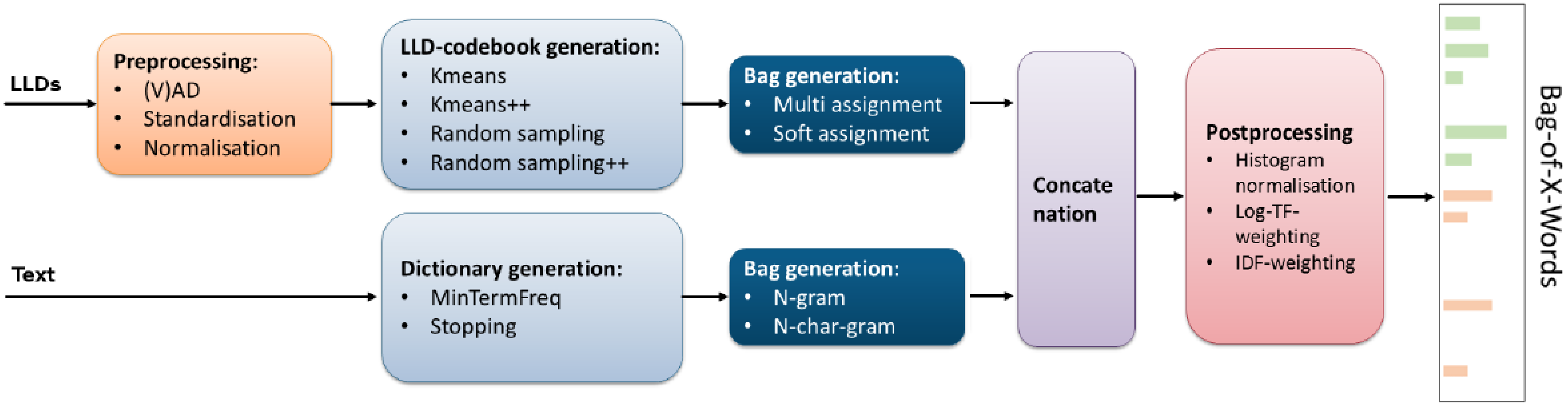
"Context-Sensitive Learning for Enhanced Audiovisual Emotion Classification", IEEE Transactions on Affective Computing, 3(2): 184-198, 2012.

Bag-of-Audio-Words

Split Vector Quantisation

+ Histrogram

openXBOW –|)→



Features

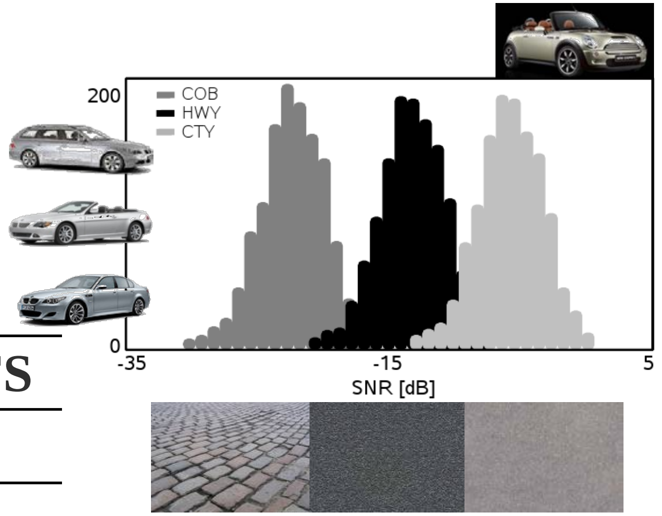
Comparison on the RECOLA (AVEC 2016) task

CCC Valid/Test	**Arousal**	**Valence**
Functionals	.790/.720	.459/.402
BLSTM-RNN	.800/.???	.398/.???
CNN (e2e)	.741/.686	.325/.261
BoAW	.793/.753	.550/.430
BoAW+Fctls	.799/.738	.521/.465

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

Acoustic Robustness

- **Additive Noise**

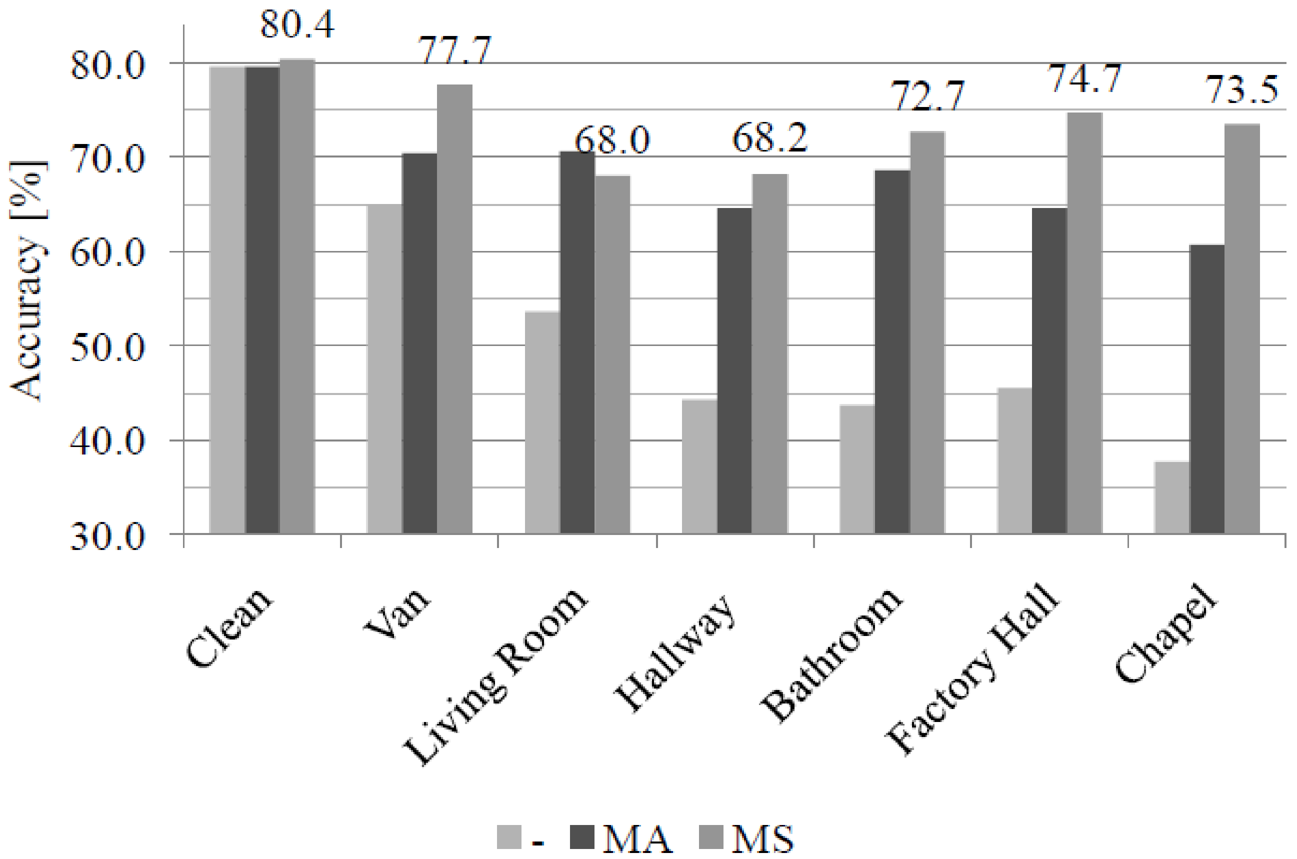


Accuracy [%]	-	NA	SA	NSA	NSA+FS
EMO-DB					
Clean Speech	74.9	-	79.6	-	**80.4**
Car Noise	60.5	72.1	75.1	76.3	**77.3**
Babble Noise	70.0	76.1	77.9	78.7	**80.5**
Babble+MINI	46.6	70.4	75.7	76.1	**79.5**
eNTERFACE					
Clean Speech	54.2	-	61.4	-	**62.8**
Car Noise	38.5	48.3	51.8	56.7	**59.7**
Babble Noise	42.1	53.2	54.2	61.0	**61.6**
Babble+MINI	30.6	49.8	46.2	55.8	**58.6**

“Emotion Recognition in the Noise Applying Large Acoustic Feature Sets”, Speech Prosody, 2006.

Acoustic Robustness

- **Reverberation**
  Matching to
  Acoustics (MA)
  Space (MS)



“Affective Speaker State Analysis in the Presence of Reverberation”, Int. Journal of Speech Technology, 2011.

Acoustic Robustness

- **NMF Features**

  Emotion Challenge Task

(a) Training with close-talk microphone (CT_{RM})

UAR [%]	C	CT_{RM}	RM	CTRV	Mean
IS	1.0	**67.62**	**60.51**	**53.06**	**60.40**
N30	1.0	65.48	52.36	50.23	56.02
$N31_I$	1.0	65.54	53.10	50.36	56.33
IS + N30	0.5	67.37	49.15	51.62	56.05
IS + $N31_I$	1.0	67.15	56.47	51.95	58.52

(b) Multicondition training ($CT_{RM} + RM + CTRV$)

UAR [%]	C	CT_{RM}	RM	CTRV	Mean
IS	0.01	**67.72**	59.52	66.06	64.43
N30	0.05	66.73	**67.55**	52.66	62.31
$N31_I$	0.2	65.81	64.61	63.32	64.58
IS + N30	0.005	67.64	62.64	**66.78**	**65.69**
IS + $N31_I$	0.005	67.07	61.85	65.92	64.95

(c) Training on room microphone (RM)

UAR [%]	C	CT_{RM}	RM	CTRV	Mean
IS	0.02	61.61	62.72	**62.10**	62.14
N30	0.2	53.57	65.61	54.87	58.02
$N31_I$	0.5	54.50	**66.54**	56.20	59.08
IS + N30	0.05	**65.13**	66.26	60.39	**63.93**
IS + $N31_I$	0.05	64.68	66.34	59.54	63.52

(d) Training on artificial reverberation (CTRV)

UAR [%]	C	CT_{RM}	RM	CTRV	Mean
IS	0.02	60.64	59.29	66.35	62.09
N30	0.05	60.73	**68.19**	62.72	**63.88**
$N31_I$	0.02	60.94	64.40	64.30	63.21
IS + N30	0.01	**61.70**	49.17	**66.68**	59.18
IS + $N31_I$	0.02	61.61	63.03	66.56	63.73

"Recognition of Non-Prototypical Emotions in Reverberated and Noisy Speech by Non-Negative Matrix Factorization", JASP, 2012.

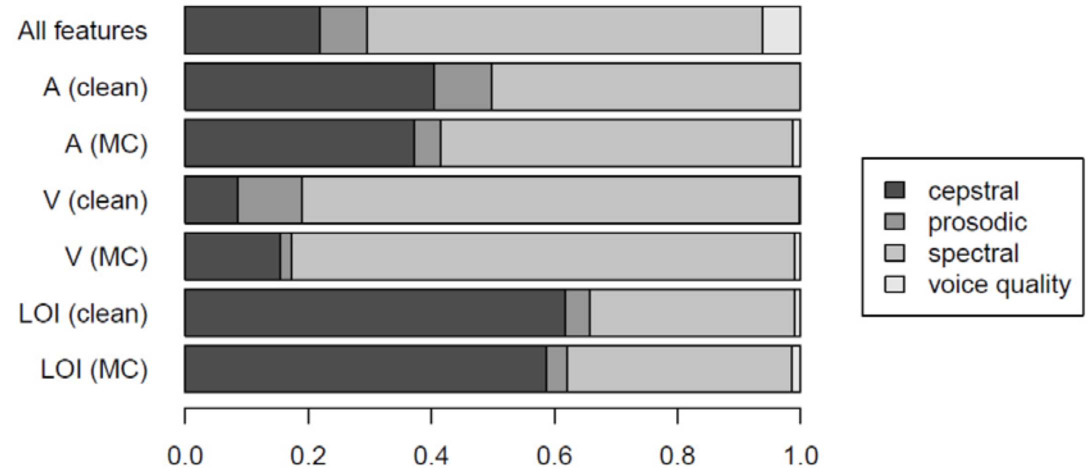
Acoustic Robustness

- **Multicondition**

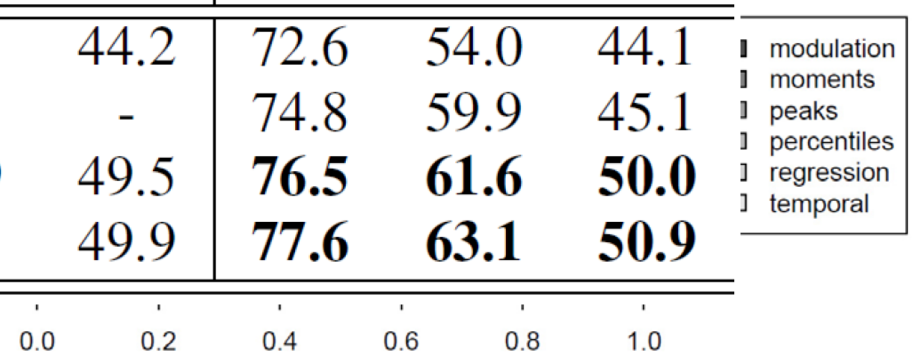
  Feature Selection

  +

  Training



[%] UAR	A	V	LOI	A	V	LOI
		all features			CC-FS	
clean	74.7	53.1	44.2	72.6	54.0	44.1
clean MC-FS	-	-	-	74.8	59.9	45.1
MCT mi.	74.4	56.9	49.5	**76.5**	**61.6**	**50.0**
MCT ma.	75.7	56.1	49.9	**77.6**	**63.1**	**50.9**

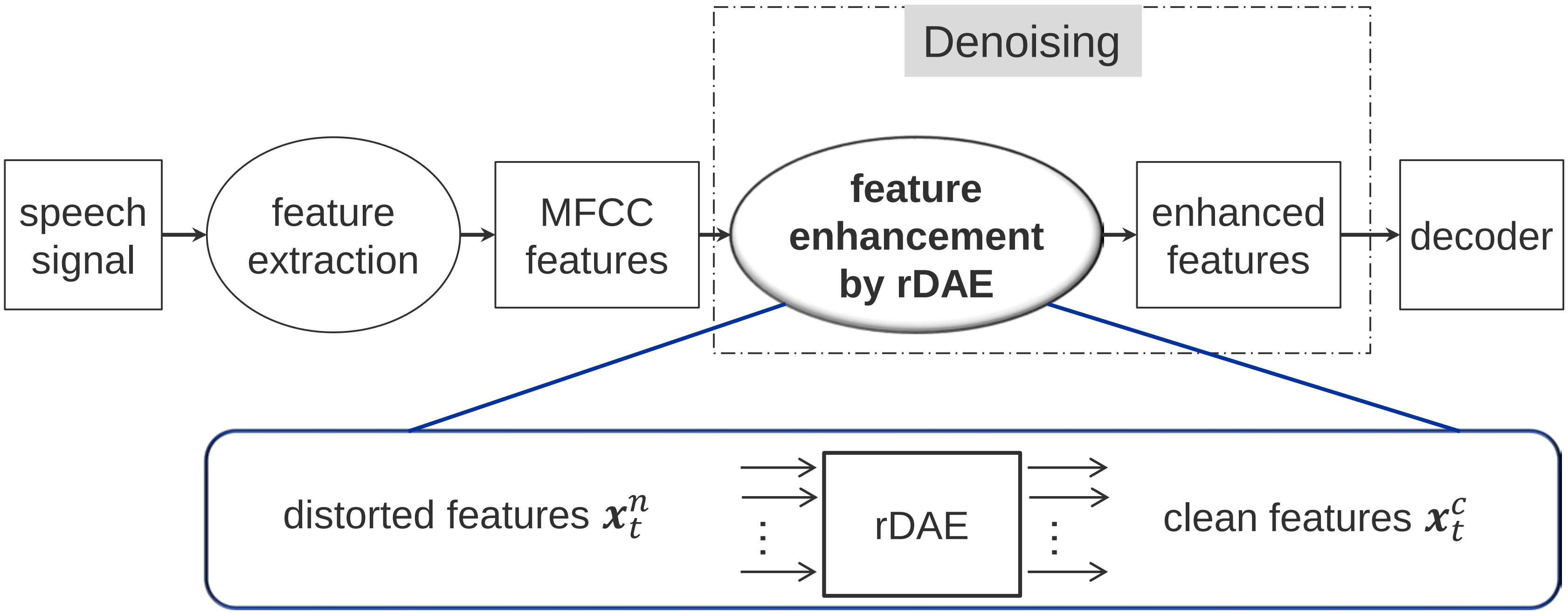


"Affect Recognition in Real-Life Acoustic Conditions – a New Perspective on Feature Selection", IEEE ACII, 2015.

Acoustic Robustness

- **Feature Enhancement**

  Recurrent Denoising Autoencoder

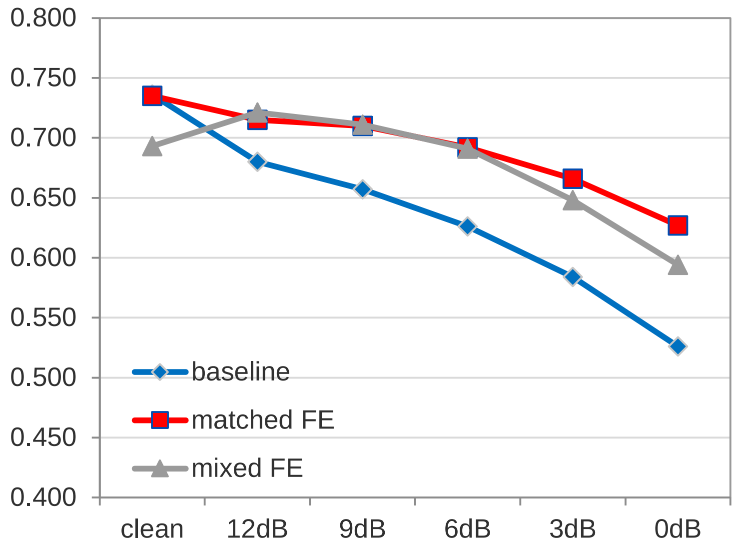


"Facing Realism in Spontaneous Emotion Recognition from Speech: Feature Enhancement by Autoencoder with LSTM Neural Networks", Interspeech, 2016.

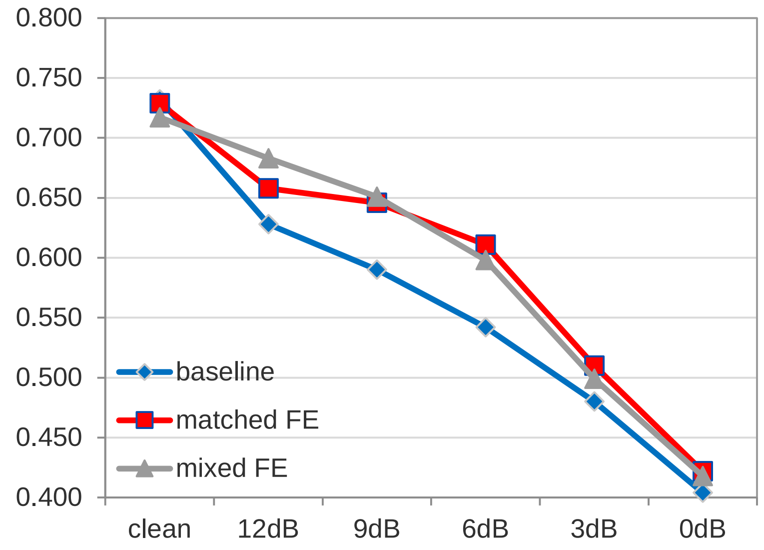
Acoustic Robustness

CHiME15 noise: **arousal**

validation



test

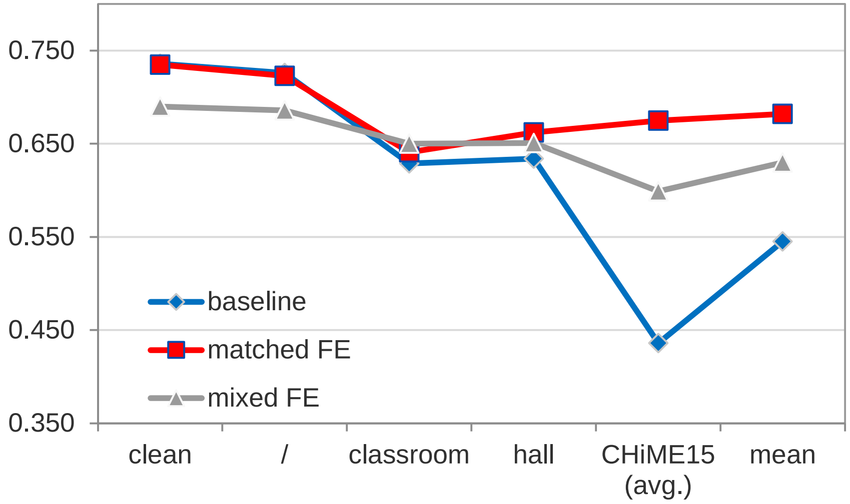


"Facing Realism in Spontaneous Emotion Recognition from Speech: Feature Enhancement by Autoencoder with LSTM Neural Networks", Interspeech, 2016.

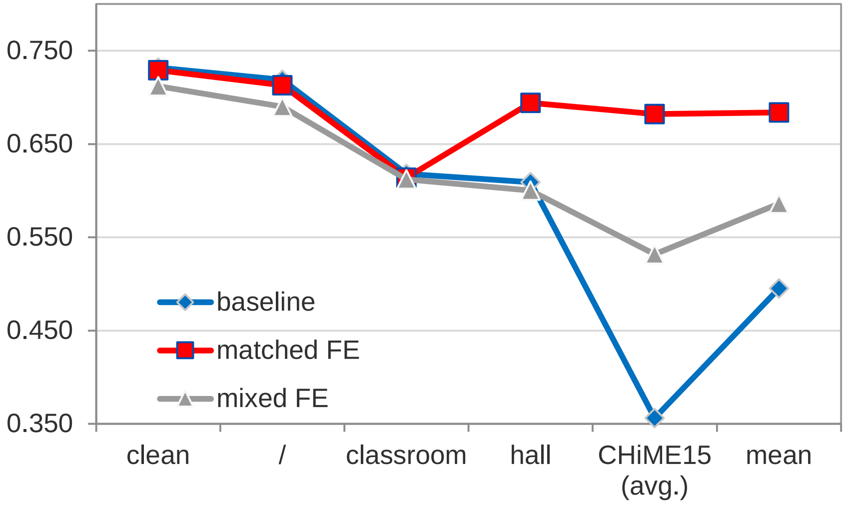
Acoustic Robustness

Smartphone noise: **arousal**

validation



test



“Facing Realism in Spontaneous Emotion Recognition from Speech: Feature Enhancement by Autoencoder with LSTM Neural Networks”, Interspeech, 2016.

Coding Robustness

- **Coding**

  Matched Learning

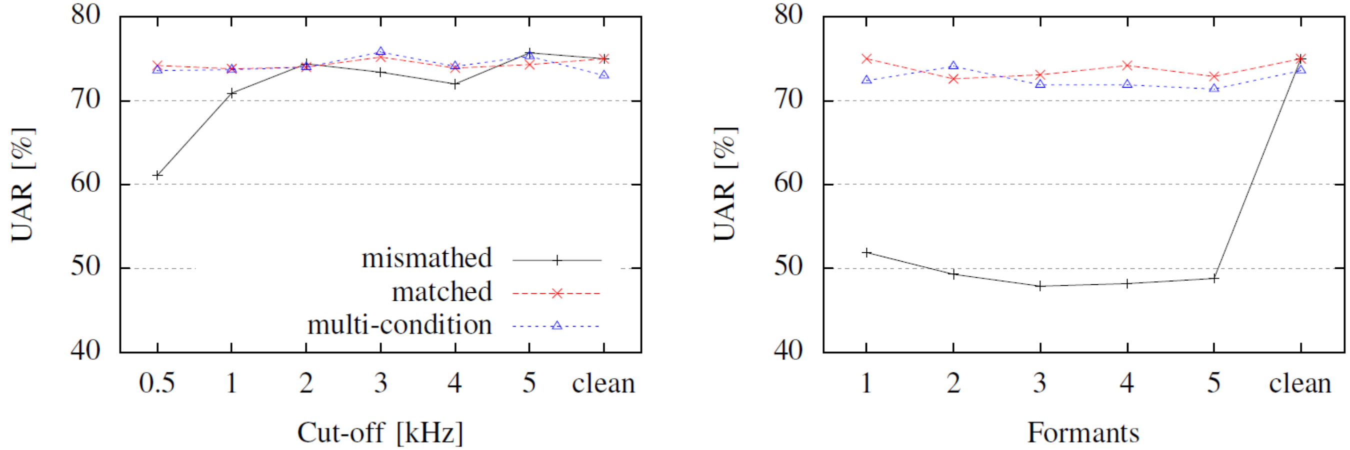
$codec_{[kbit/s]}$	GEMEP			CPSD		
	SSNR	IS	PESQ	SSNR	IS	PESQ
$G711_{64}$	29.4	0.10	4.3	34.8	0.01	4.4
$G726_{40}$	27.1	0.11	4.2	30.7	0.02	4.3
$G726_{32}$	24.6	0.14	4.0	26.7	0.04	4.2
$G726_{24}$	19.4	0.33	3.6	21.2	0.12	3.9
$G728_{16}$	15.6	0.21	4.0	16.2	0.21	4.0
GSM_{13}	10.2	0.40	3.4	11.6	0.40	3.4
$G7231_{6.3}$	-2.2	1.44	3.4	-2.5	1.58	3.2
$G7231_{5.3}$	-2.0	2.20	3.2	-2.4	2.24	3.1
$LPC10_{2.4}$	-3.1	29.08	2.4	-3.7	26.55	1.9
$codec2_{1.3}$	-2.8	3.57	2.1	-3.1	3.05	2.1

UAR[%]	GEMEP									CPSD						Avg.
	Arousal			**Valence**			**Emotion**			**Typicality**			**Diagnosis**			
	mi	*ma*	*mu*	*mi*	*ma*	*mu*	*mi*	*ma*	*mu*	*mi*	*ma*	*mu*	*mi*	*ma*	*mu*	
clean	75.0	75.0	74.2	61.6	61.6	62.5	40.9	40.9	40.5	90.7	90.7	89.8	67.1	67.1	63.6	66.7
PCM_{128}	76.5	75.2	74.6	63.1	62.3	62.5	34.3	36.8	40.4	87.8	89.4	89.7	59.7	63.3	64.7	65.4
$G711_{64}$	75.7	74.7	74.3	61.0	59.9	63.0	33.1	37.4	39.7	88.3	89.5	89.9	60.7	63.7	64.4	65.0
$G726_{40}$	75.7	74.5	74.2	61.2	59.7	62.5	33.7	37.6	38.4	88.4	89.5	89.9	60.0	63.0	63.8	64.8
$G726_{32}$	75.8	74.7	74.5	61.6	59.2	62.5	34.0	35.6	40.3	88.4	88.7	88.8	59.7	63.1	64.2	64.7
$G726_{24}$	74.5	74.2	73.4	58.6	58.5	61.2	26.0	33.9	36.7	87.7	89.7	89.3	60.3	60.6	61.2	63.1
$G728_{16}$	75.7	74.9	74.5	62.1	59.5	62.5	32.1	38.3	40.3	87.8	89.4	88.8	59.7	62.3	64.2	64.8
GSM_{13}	76.0	75.4	73.4	60.1	58.3	61.2	34.0	35.7	36.7	88.3	88.7	89.3	61.4	63.6	61.2	64.2
$G7231_{63}$	75.9	73.9	73.9	61.9	59.5	63.7	31.0	36.7	36.9	87.2	88.2	89.3	58.1	62.8	63.5	64.2
$G7231_{53}$	75.2	74.8	73.0	59.5	60.2	64.4	30.7	35.9	34.8	87.3	88.8	89.0	58.6	61.3	63.5	63.8
$LPC10_{24}$	71.2	75.7	73.0	61.9	64.6	63.5	27.8	34.1	31.0	72.1	85.7	81.9	44.2	64.1	57.2	60.5
$codec2_{13}$	73.8	73.9	73.9	57.9	60.1	62.3	25.6	35.3	36.3	85.5	88.5	88.8	53.7	62.3	56.2	62.3
Avg.	75.1	74.7	73.9	60.8	60.2	62.7	31.1	36.1	37.4	86.3	88.7	88.6	57.8	62.7	62.2	63.9

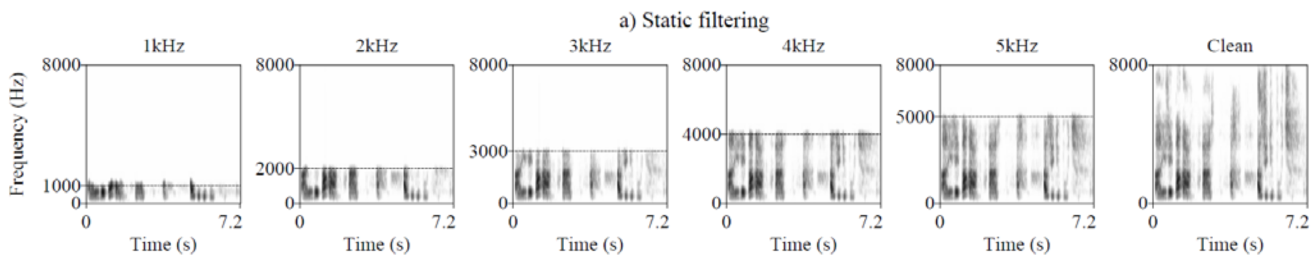
"The Effect of Narrow-band Transmission on Recognition of Paralinguistic Information From Human Vocalizations", IEEE Access, 2016.

Bandwidth Robustness

- **Channel**

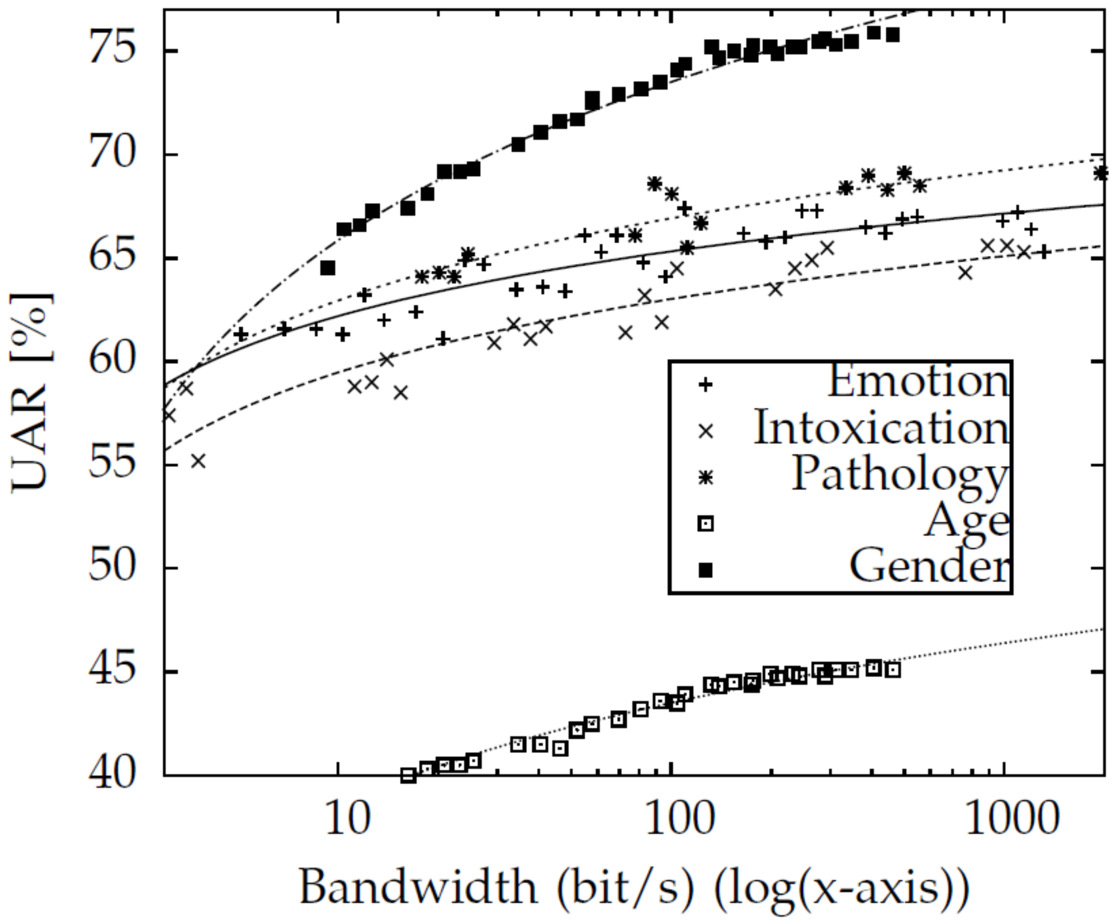


(a) Arousal



*"The Effect of Narrow-band Transmission on Recognition of Paralinguistic Information From Human Vocalizations", **IEEE Access**, 2016.*

Bandwidth Robustness



"Serious Gaming for Behavior Change – The State of Play," IEEE Pervasive Computing Magazine, 12: 48–55, 2013.

Linguistic Robustness.

Linguistic Robustness

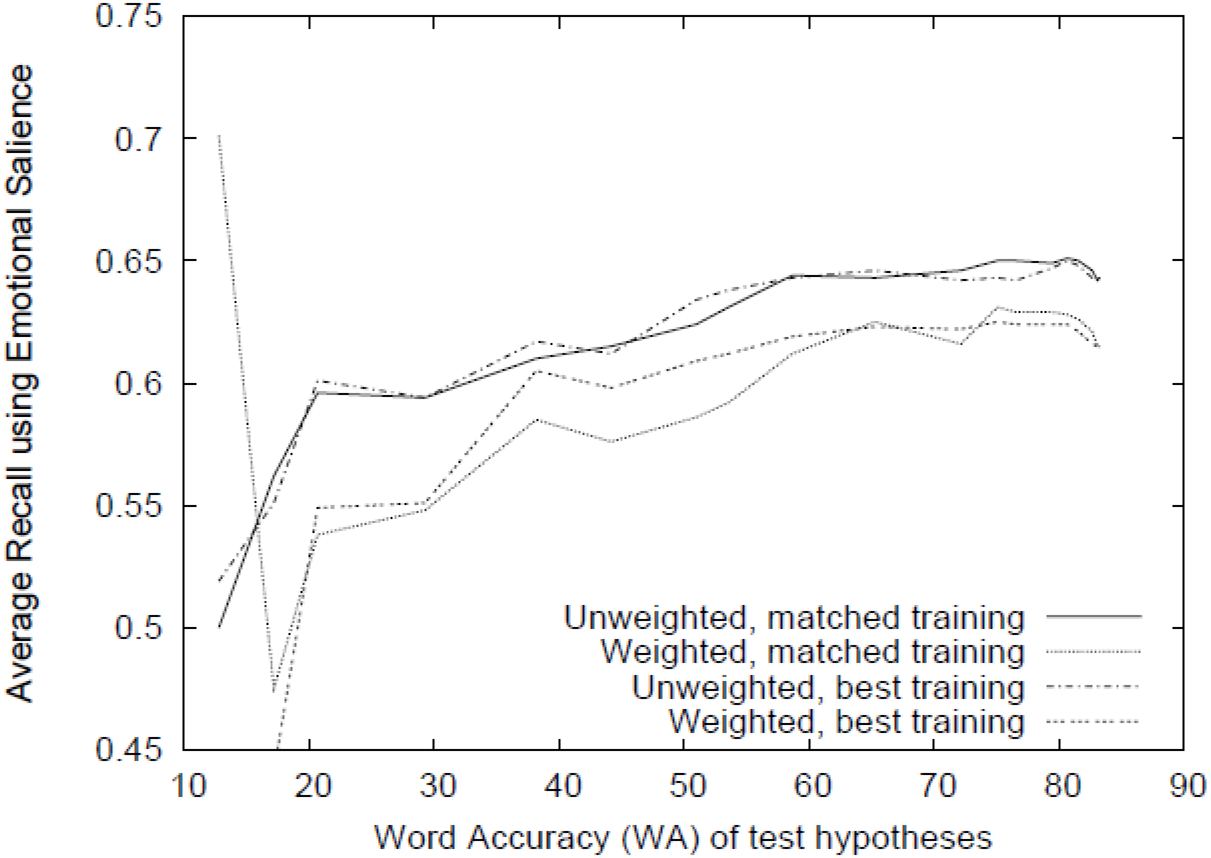
- **Spoken Content Matching**
  Examples (LOSO)

Model description	Acc. [%]	G 1	G 2	All
EMO-DB	matched	57.2	46.9	48.9
	mismatched	36.6	37.7	37.4
SUSAS	matched	64.6	60.3	60.7
	mismatched	52.4	54.4	55.2
AVIC	matched	79.7	57.8	60.9
	mismatched	49.2	51.3	50.1

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

Linguistic Robustness

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



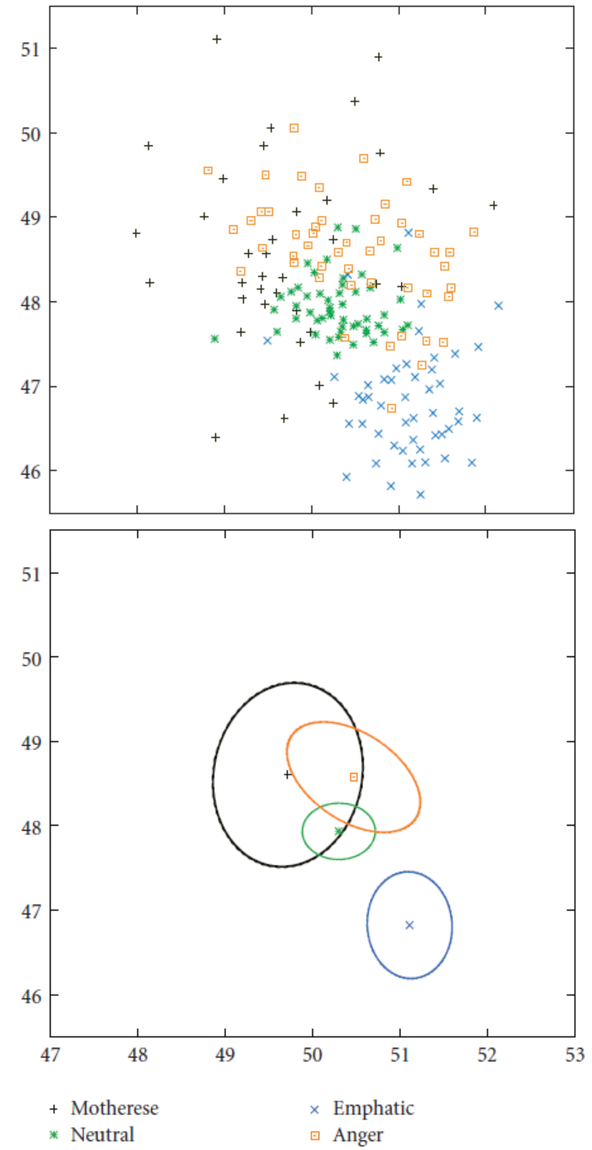
(INTERSPEECH 2010)

"On the Influence of Phonetic Content Variation for Acoustic Emotion Recognition," IEEE PIT, 2008.

Linguistic Robustness

- **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%



"Does Affect Affect Automatic Recognition of Children's Speech?," ACM WOCCI, 2008.

Linguistic Robustness

- **Training and Adapting Models**

  AM, LM, both

  Word accuracy
  Significance

(a) Scenario 1: "neutral versus emotional ASR engine"

	M	E	A
Baseline system	43.6	61.3	64.9
Adapted systems			
Acoustic models	43.1 ○ ○ ○ ○ ○	74.8 • • • • •	73.5 • • • • •
Linguistic models	49.3 • ○ ○ ○ ○	67.0 • • • • •	68.5 • • • • ○
Both	47.4 ○ ○ ○ ○ ○	76.5 • • • • •	75.3 • • • • •

(b) Scenario 2: "adaptation of neutral ASR engine"

	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," JASMP, 2010.

Multilingual: 2/3 Covered?

Language	% NS	Rank
Mandarin	14.40	1
Spanish	6.15	2
English	5.43	3
Hindi	4.70	4
Arabic	4.43	5
Portuguese	3.27	6
Bengali	3.11	7
Russian	2.33	8
Japanese	1.90	9
Punjabi	1.44	10
German	1.39	11
Malay/Indonesian	1.16	14
Telugu	1.15	15
Vietnamese	1.14	16
Korean	1.14	17
French	1.12	18
Marathi	1.10	19
Tamil	1.06	20
Urdu	0.99	21

Language	% NS	Rank
Persian	0.99	22
Turkish	0.95	23
Italian	0.90	24
Cantonese	0.89	25
Thai	0.85	26
Gujarati	0.74	27
Polish	0.61	30
Pashto	0.58	31
Burmese	0.50	38
Sindhi	0.39	47
Romanian	0.37	50
Dutch	0.32	57
Assamese	0.23	67
Hungarian	0.19	73
Greek	0.18	75
Czech	0.15	83
Swedish	0.13	91
Balochi	0.11	99

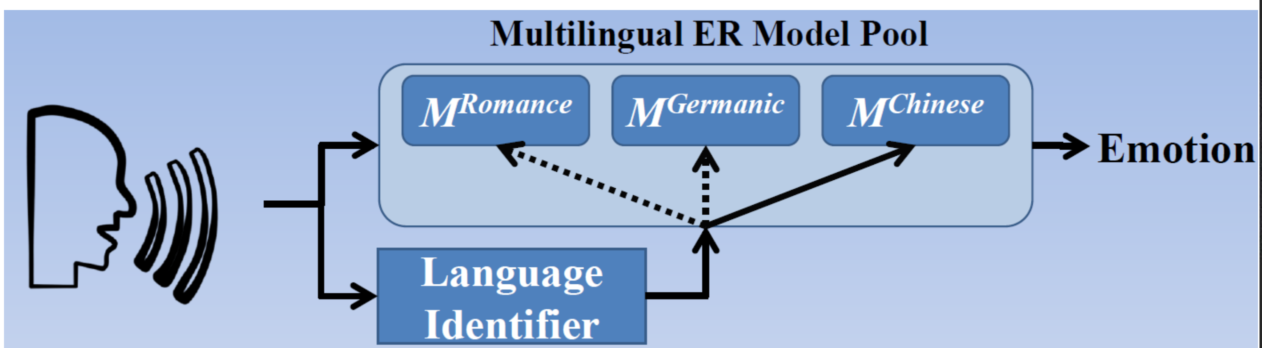
*“Cross-Language Acoustic Emotion Recognition – An Overview and Some Tendencies”,**ACII**, 2015.*

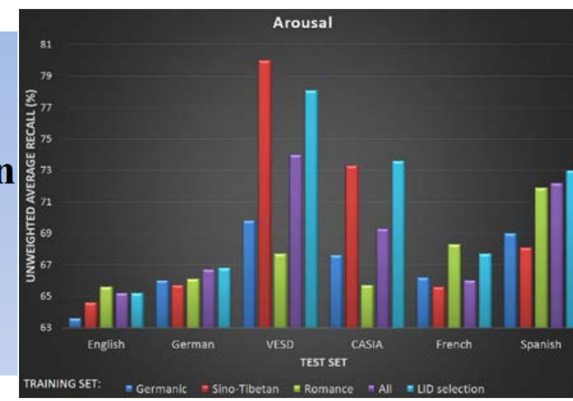
Linguistic Robustness

- **Cross-Language Acustics**

  Same language, within and across language family

% UA	same L	within LF	across LF
Arousal	94.0	66.3	62.7
Valence	81.7	61.9	54.6

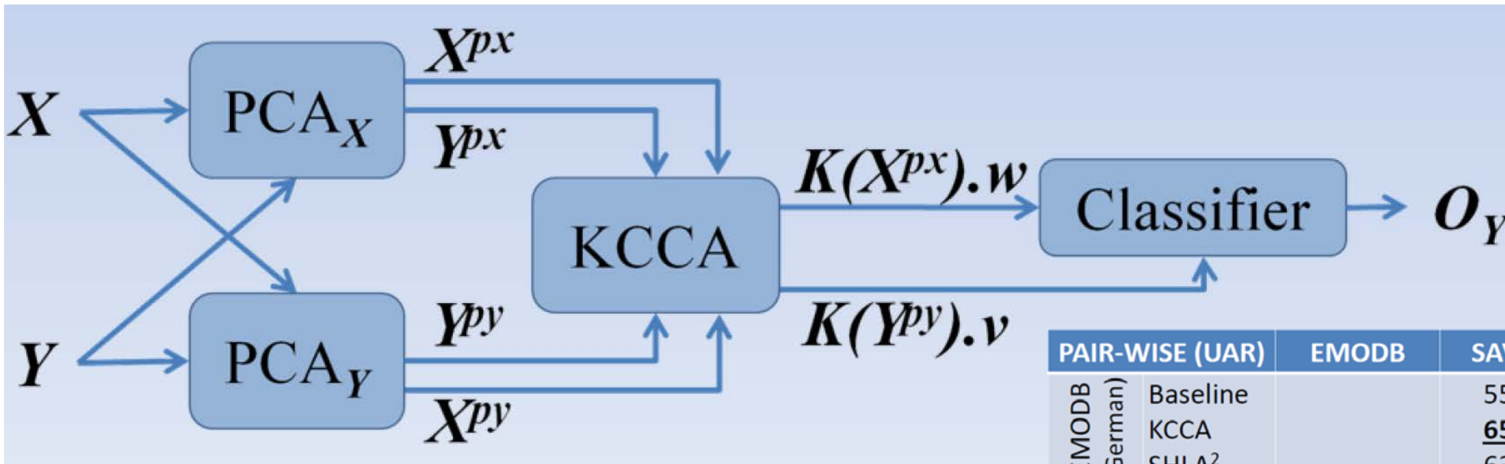




"Cross-Language Acoustic Emotion Recognition – An Overview and Some Tendencies",ACII, 2015.
"Enhancing Multilingual Recognition of Emotion in Speech by Language Identification", Interspeech, 2016.

Linguistic Robustness

- **Transfer Learning**



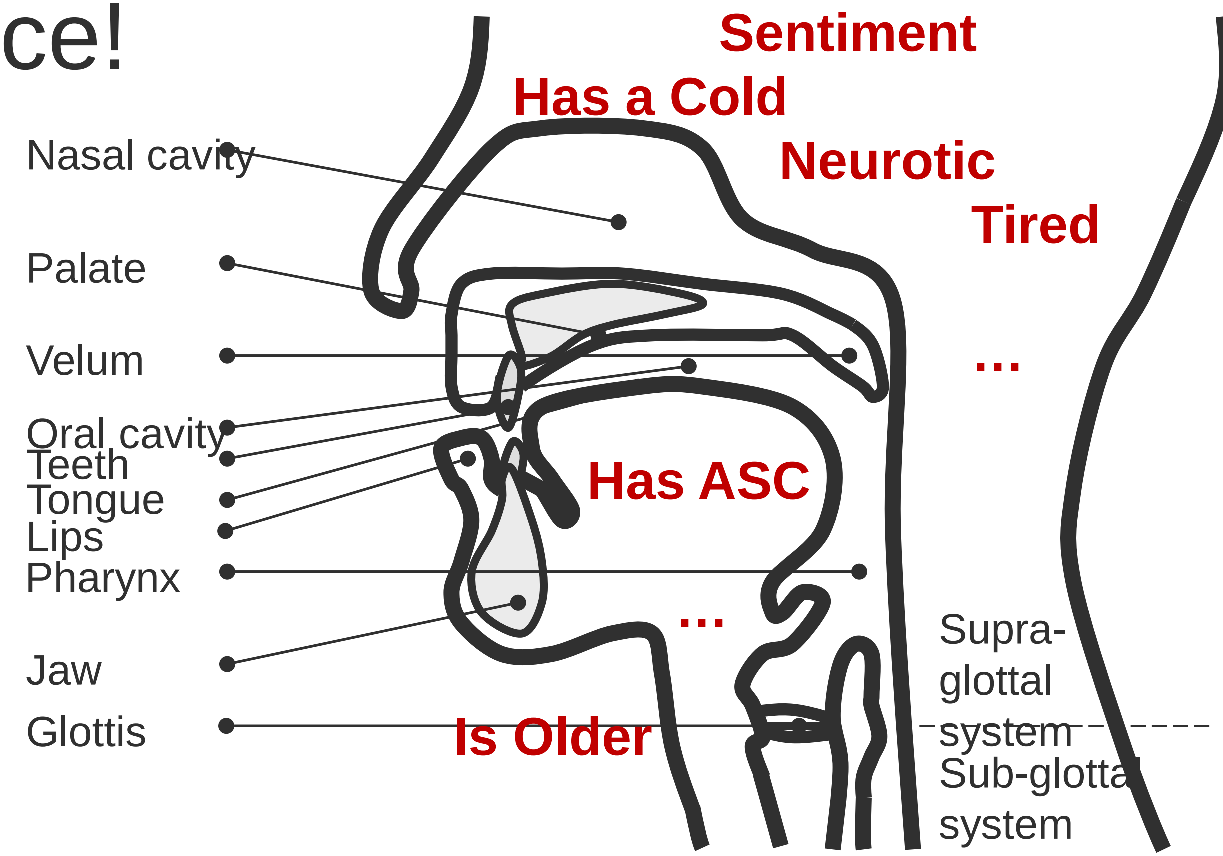
PAIR-WISE (UAR)		EMODB	SAVEE	EMOVO	Polish
EMODB (German)	Baseline		55.8	56.4	71.9
	KCCA		**65.2**	**57.9**	**72.8**
	SHLA[2]		63.7	56.8	59.7
SAVEE (English)	Baseline	62.5		54.6	65.0
	KCCA	**71.9**		**59.5**	69.4
	SHLA	67.7		59.0	**70.2**
EMOVO (Italian)	Baseline	58.0	51.2		55.3
	KCCA	62.9	**58.5**		**71.9**
	SHLA	**67.3**	58.2		64.3
Polish	Baseline	65.1	55.4	57.3	
	KCCA	70.9	**61.9**	56.1	
	SHLA	**71.1**	60.6	**58.3**	

*“Cross Lingual Speech Emotion Recognition using Canonical Correlation Analysis on Principal Component Subspace”,**ICASSP**, 2016.*

Paralinguistic Robustness.

Only One Voice!

- **Multiple-Targets**
  There is just one Vocal Production Mechanism…



% UA	Single	Multiple
Likability	59.1	(+A,G,Cl) **62.2**
Neuroticism	62.9	(+G,OCEA, Cl) **67.5**

Model Switching

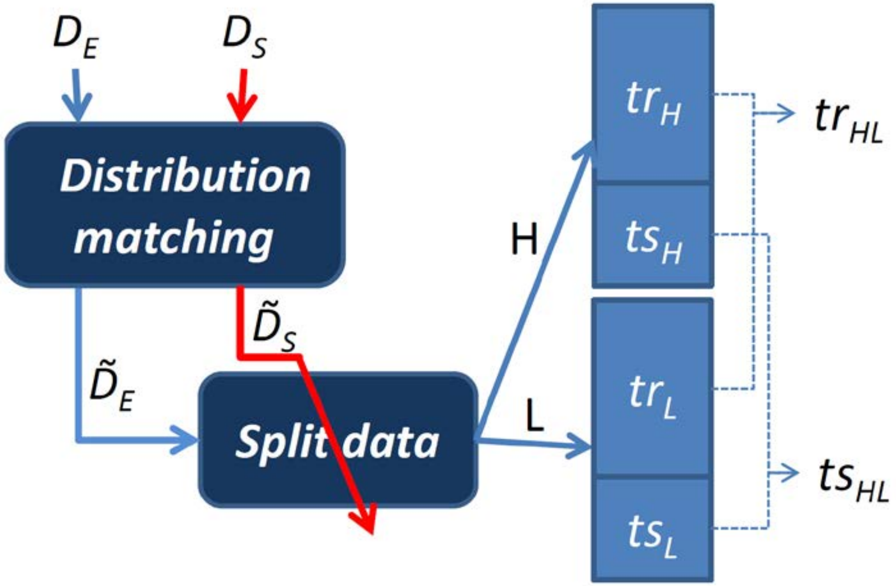
- **Model Selection**

  By: Age, Gender, Personality

  4 Emotional Speech Corpora

  AVIC, AEC

  eNTERFACE, SUSAS



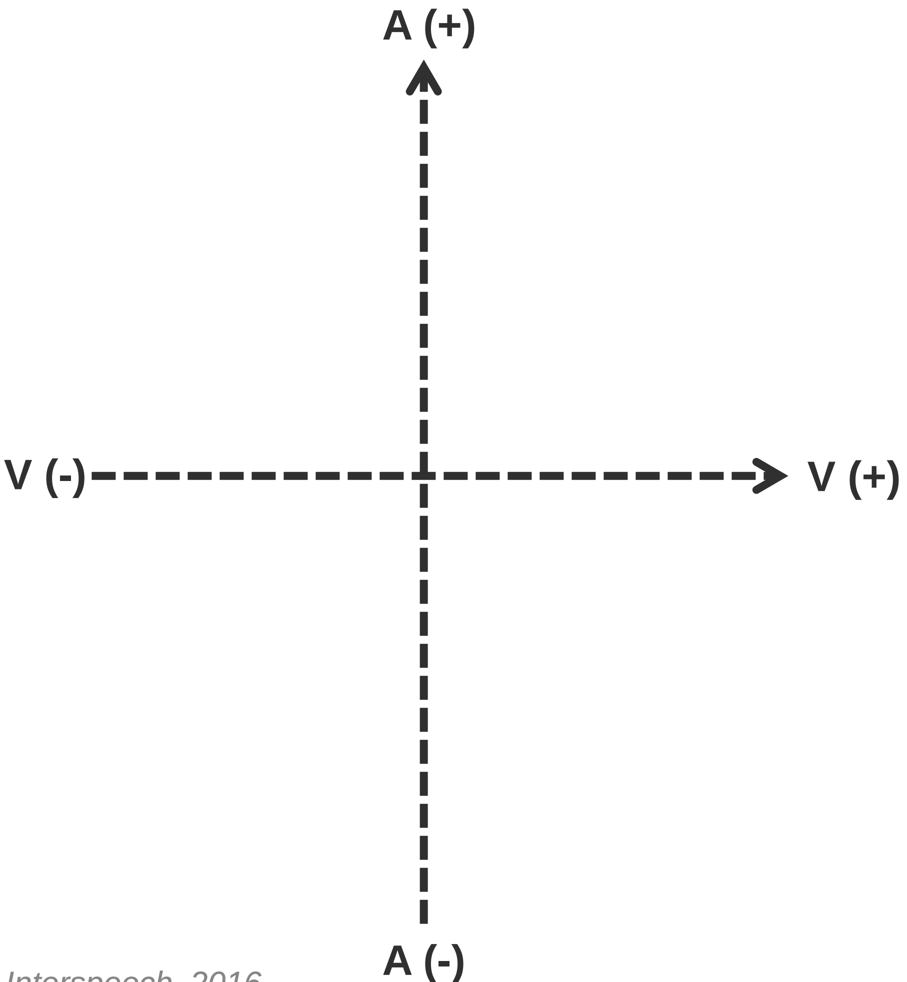
	Train	Test	O	C	E	A	N	Gender	Age
UAR	tr_{HL}	ts_{HL}	**73.66**	73.78	**73.33**	**73.66**	73.66	$(tr_{MF} \to ts_{MF})$ 73.38	$(tr_{AY} \to ts_{AY})$ 74.11
ΔUAR	tr_{HL}	ts_H	0.07^{ns}	-3.03	0.11^{ns}	-0.77	**0.75**	$(tr_{MF} \to ts_F)$ 1.63	$(tr_{AY} \to ts_A)$ -4.63
	tr_{HL}	ts_L	-1.83	**0.79**	-3.47	-1.21	-3.66	$(tr_{MF} \to ts_M)$ -2.78	$(tr_{AY} \to ts_Y)$ 2.47
ΔUAR	tr_H	ts_H	0.00^{ns}	-2.85	-0.02^{ns}	-0.24^{ns}	0.51^{ns}	$(tr_F \to ts_F)$ **2.47**	$(tr_A \to ts_A)$ -5.07
	tr_L	ts_L	-2.23	0.68	-2.61	-1.33	-3.34	$(tr_M \to ts_M)$ -3.15	$(tr_Y \to ts_Y)$ **2.66**
	tr_H	ts_L	-5.39	-4.61	-10.28	-8.84	-9.88	$(tr_F \to ts_M)$ -6.59	$(tr_A \to ts_Y)$ -8.10
	tr_L	ts_H	-7.19	-6.74	-6.78	-2.88	-4.21	$(tr_M \to ts_M)$ -5.65	$(tr_Y \to ts_A)$ -6.39
						(b)			
UAR	tr_{HL}	ts_{HL}	60.94	61.15	60.99	60.86	60.93	$(tr_{MF} \to ts_{MF})$ 63.16	$(tr_{AY} \to ts_{AY})$ 62.41
ΔUAR	Rule		-	1.04	-	-	0.40	0.18^{ns}	0.39

"The effect of personality trait, age, and gender on the performance of automatic speech emotion recognition", to appear.

Higher-level Features

ND and D speech from
Interspeech ComParE 2016

ND | **D**



“Is deception emotional? An emotion-driven predictive approach”, Interspeech, 2016.

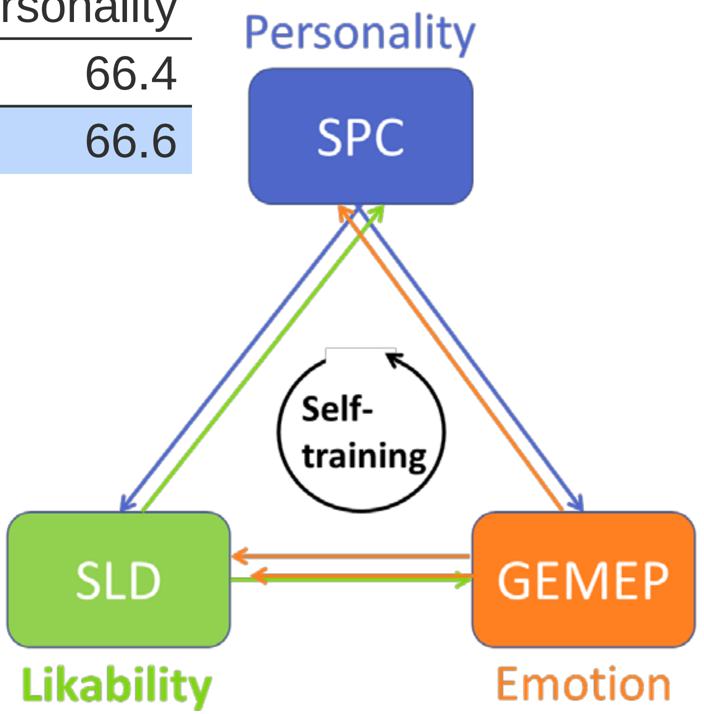
Holism: Vertical.

- **Cross-Task Self-Labelling**

% UA	Likability	Emotion	Personality
Baseline	57.2	68.9	66.4
Cross-Task Labelling	60.3	69.0	66.6

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

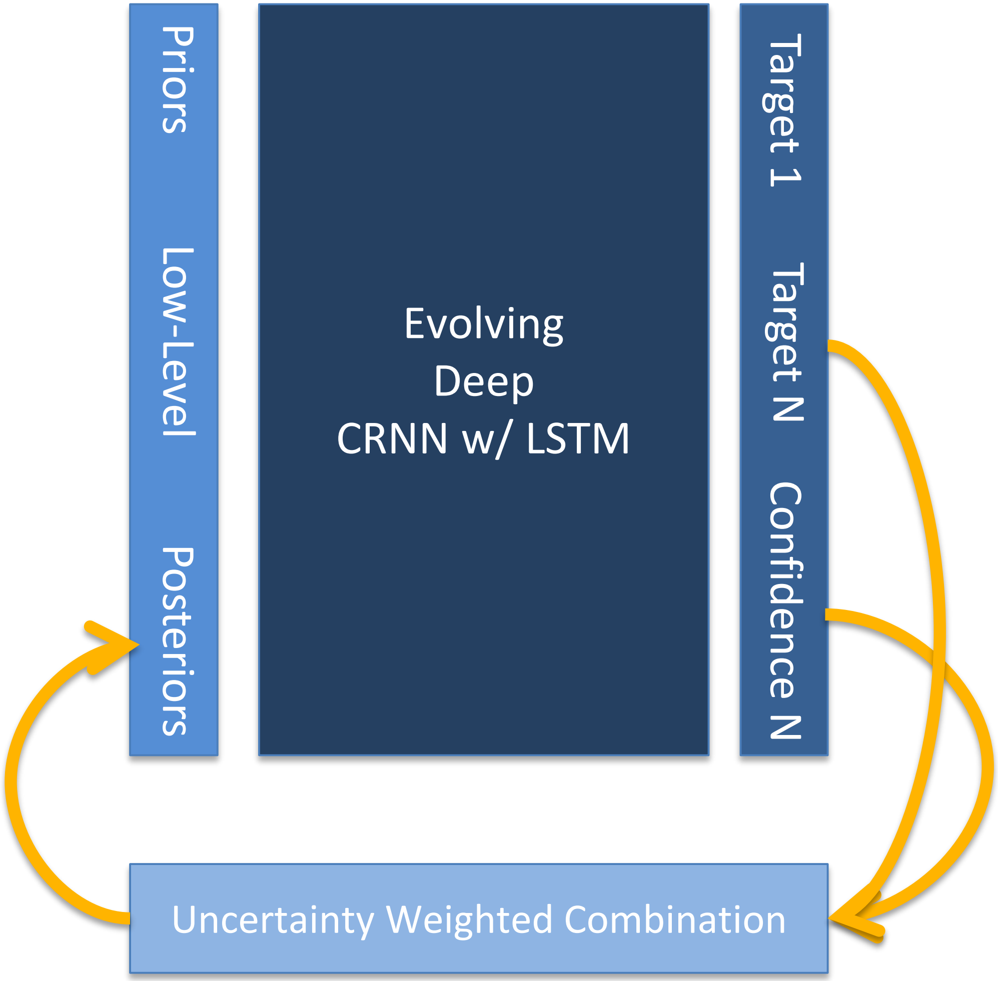
1. (Optional) Upsample training set $\mathcal{L}$ to even class distribution $\mathcal{L}_D$
2. Use $\mathcal{L}/\mathcal{L}_D$ to train classifier $\mathcal{H}$, then classify $\mathcal{U}$
3. Select a subset $\mathcal{N}_{st}$ that contains those instances predicted with the highest confidence values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

Holism: Next-Gen?

- **Evolutionary Learning**
- **Reinforced Learning**
- **Analysis/Synthesis Gap**



More Data: The answer to it all?

New Data

sewa
Automatic Sentiment Analysis in the Wild

- **In the Wild**

Cultural Background		Age Group		Years Known the Other Participant		Self-Reported Familiarity Rating	
British	66	**18~29**	203	**<1**	80	**Not Familiar**	9
				1	30		
German	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	13
Hungarian	70	**40~49**	25	**4**	37	**Somewhat Familiar**	35
				5~9	55		
Serbian	72	**50~59**	46	**10~14**	20	**Moderately Familiar**	114
				15~19	22		
Greek	56	**60+**	30	**20+**	75	**Extremely Familiar**	227
Chinese	70						

http://db.sewaproject.eu/

New Data

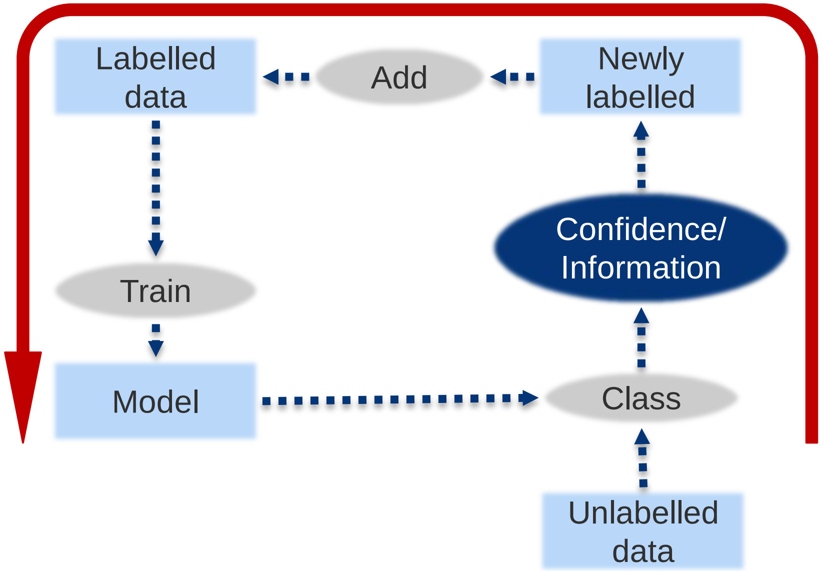
- **Graz Real-Life Affect in the Street & Supermarket (GRAS²)**

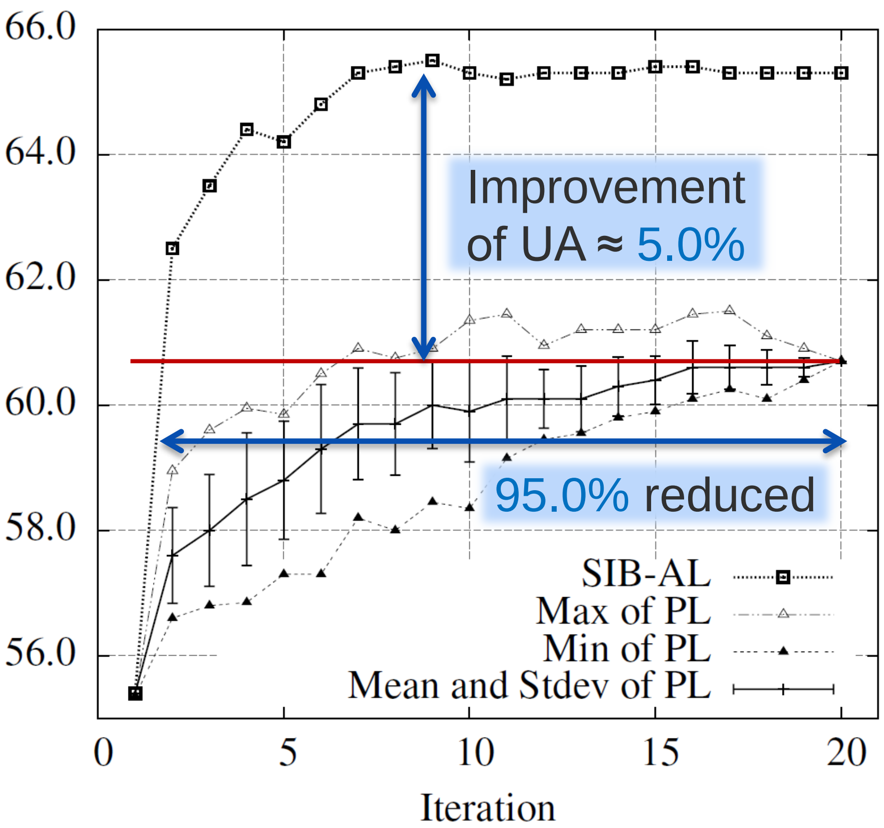
6 channel audio +
video + eyetracking +
EDA + temperature +
2x 3D motion

Ask for help
Gradually embarsassing:
denture adhisive
Anti-athlete’s foot cream

Efficient Labelling

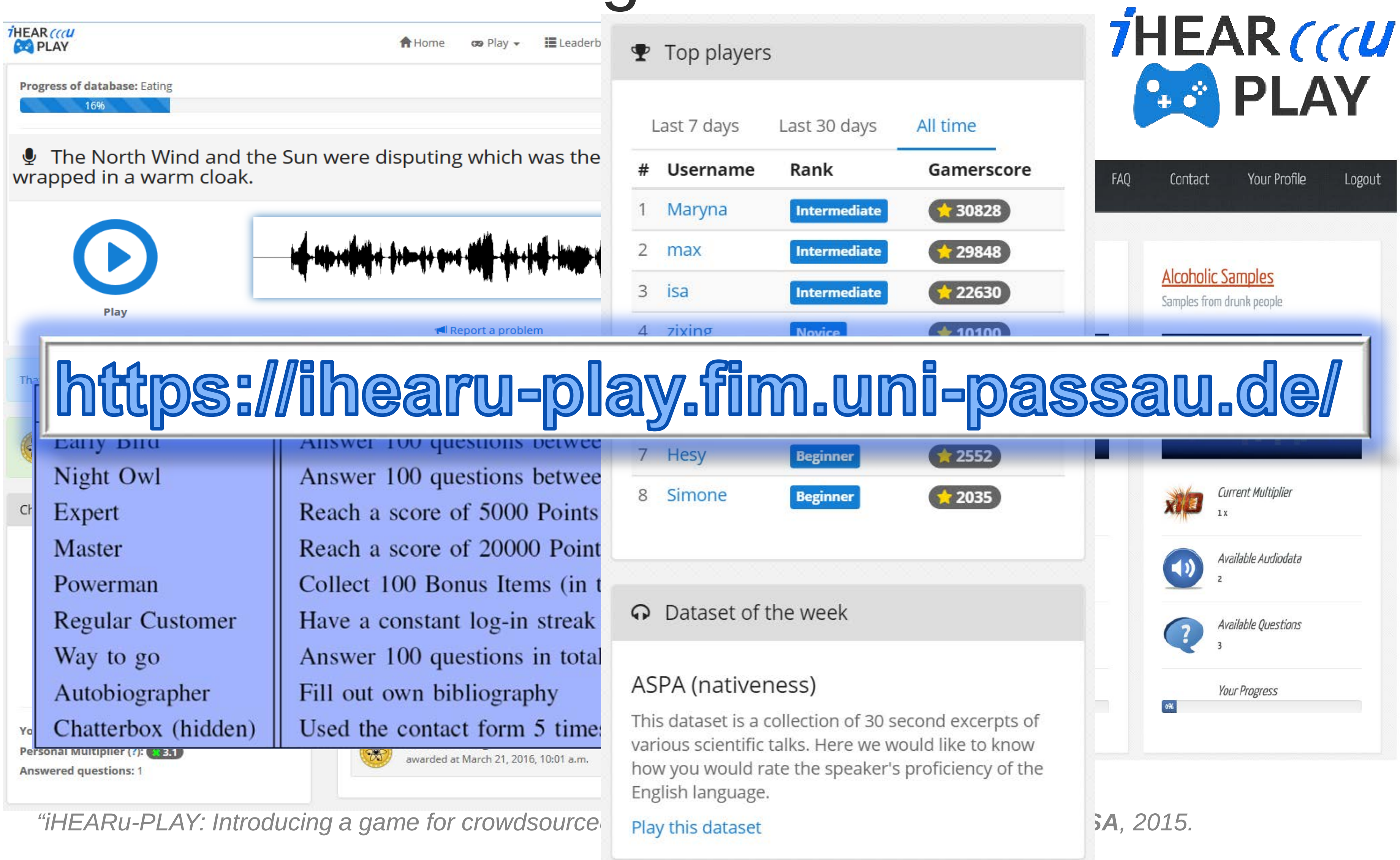
- **Cooperative Learning in aRMT**
  0) Transfer Learning
  1) Dynamic Active Learning
  2) Semi-Supervised Learning





"Cooperative Learning and its Application to ESR", ***IEEE Transactions ASLP***, 2015.

Efficient Labelling



https://ihearu-play.fim.uni-passau.de/

“iHEARu-PLAY: Introducing a game for crowdsource ... *SA, 2015.*

Vision.

#thankyou @CHiME ☺

Group Assessment

Cultural Robustness

Multiple Microphones, “Chips-Bag“,…?

Robust Gold Standard

Coupled ASR + CP?

Abstract

An increasingly long list of states and traits of speakers is being targeted for automatic recognition by computers including their age, emotion, health condition, or personality. However, hardly any of these have been encountered in “everyday” usage by the broad consumer mass up to now. This is certainly also owed to robustness issues, which shall be discussed here. Traditionally, these comprise speech enhancement, feature enhancement, feature space adaptation, or matched conditions training – mainly to cope with additive or convolutional noise. In addition, a number of further robustness issues mark this field of speech analysis, including interdependence of states and traits, potential subjectivity in the labels, phonetic content variation in the acoustic analysis, varying language and erroneous speech recognition in the linguistic analysis, and diversity of the cultural background of speakers. Finally, a number of hardly tackled issues remain such as the analysis of multiple speakers or in far field condition with multiple microphones. In the talk, an overview on these challenges and existing solutions is given. Then, required future research efforts will be named to help Computational Paralinguistics’ massive launch into the next generation dialogue systems and many other applications.