Emotion Recognition from Speech with Acoustic, Non-Linear and Wavelet-based Features Extracted in Different Acoustic Conditions

Juan Camilo Vásquez Correa

Advisor: PhD. Jesús Francisco Vargas Bonilla.

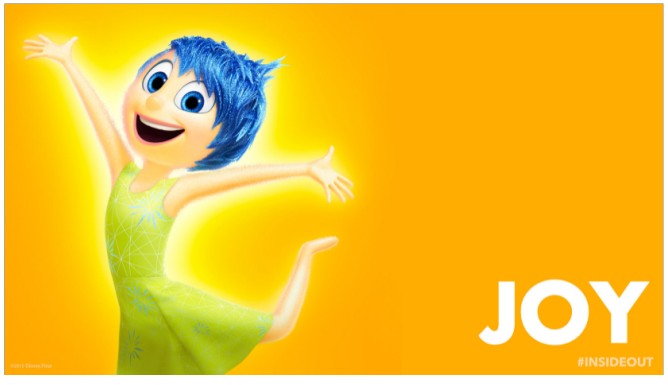
Department of Electronics and Telecommunication Engineering,
University of Antioquia UdeA.

jcamilo.vasquez@udea.edu.co

Outline

Outline

Introduction: Emotions



Introduction: Emotion recognition

Recognition of emotion from speech:

- Call centers
- Emergency services
- Depression Treatment
- Intelligent vehicles
- Public surveillance








Introduction: Emotions and speech

Table: Relationship between specch and emotions[1]

Feature	Anger	Happiness	Fear	Disgust	Sadness
Speech Rate	Faster or faster	Slightly slower	Much faster	Very much faster	Slightly slower
F_0	Very much higher	Much higher	Very much higher	Very much lower	Slightly lower
F_0 Range	Much wider	Much wider	Much wider	Slightly wider	Slightly narrower
ΔF_0	Abrupt on stressed	smooth, upward inflections	Normal	Wide, downward inflections	Downward inflections
Energy content	Higher	Higher	Normal	Lower	Lower
Voice Quality	Breathy chest blaring tone	Breathy	Irregular chest tone	Grumble voicing	Resonant
Articulation	Tense	Normal	Precise	Normal	Slurring

[1] R. Cowie et al. "Emotion recognition in human-computer interaction". In: *IEEE Signal Processing Magazine* 18.1 (2001), pp. 32–80.

Introduction: Emotions and speech

Table: Relationship between specch and emotions[2]

Feature	Anger	Happiness	Fear	Disgust	Sadness
Speech Rate	Faster or faster	Slightly slower	Much faster	Very much faster	Slightly slower
F_0	Very much higher	Much higher	Very much higher	Very much lower	Slightly lower
F_0 Range	Much wider	Much wider	Much wider	Slightly wider	Slightly narrower
ΔF_0	Abrupt on stressed	smooth, upward inflections	Normal	Wide, downward inflections	Downward inflections
Energy content	Higher	Higher	Normal	Lower	Lower
Voice Quality	Breathy chest blaring tone	Breathy	Irregular chest tone	Grumble voicing	Resonant
Articulation	Tense	Normal	Precise	Normal	Slurring

[2] R. Cowie et al. "Emotion recognition in human-computer interaction". In: *IEEE Signal Processing Magazine* 18.1 (2001), pp. 32–80.

Introduction: Emotions and speech

Table: Relationship between specch and emotions[3]

Feature	Anger	Happiness	Fear	Disgust	Sadness
Speech Rate	Faster or faster	Slightly slower	Much faster	Very much faster	Slightly slower
F_0	Very much higher	Much higher	Very much higher	Very much lower	Slightly lower
F_0 Range	Much wider	Much wider	Much wider	Slightly wider	Slightly narrower
ΔF_0	Abrupt on stressed	smooth, upward inflections	Normal	Wide, downward inflections	Downward inflections
Energy content	Higher	Higher	Normal	Lower	Lower
Voice Quality	Breathy chest blaring tone	Breathy	Irregular chest tone	Grumble voicing	Resonant
Articulation	Tense	Normal	Precise	Normal	Slurring

[3] R. Cowie et al. "Emotion recognition in human-computer interaction". In: *IEEE Signal Processing Magazine* 18.1 (2001), pp. 32–80.

Introduction: Emotions and speech

Table: Relationship between specch and emotions[4]

Feature	Anger	Happiness	Fear	Disgust	Sadness
Speech Rate	Faster or faster	Slightly slower	Much faster	Very much faster	Slightly slower
F_0	Very much higher	Much higher	Very much higher	Very much lower	Slightly lower
F_0 Range	Much wider	Much wider	Much wider	Slightly wider	Slightly narrower
ΔF_0	Abrupt on stressed	smooth, upward inflections	Normal	Wide, downward inflections	Downward inflections
Energy content	Higher	Higher	Normal	Lower	Lower
Voice Quality	Breathy chest blaring tone	Breathy	Irregular chest tone	Grumble voicing	Resonant
Articulation	Tense	Normal	Precise	Normal	Slurring

[4] R. Cowie et al. "Emotion recognition in human-computer interaction". In: *IEEE Signal Processing Magazine* 18.1 (2001), pp. 32–80.

Introduction: State of the art

- Three kinds of databases
  1. Acted databases: (Burkhardt et al. 2005; Busso, Bulut, et al. 2008; Haq and Jackson 2010; Bänziger, Mortillaro, and K. R. Scherer 2012)
  2. Evoked databases: (Martin et al. 2006; McKeown et al. 2012)
  3. Natural databases: (Steidl 2009; Ringeval et al. 2013)

Introduction: State of the art

Table: Feature extraction

Acoustic Analysis	(Busso, S. Lee, and Narayanan 2009; B. Schuller, Steidl, and Batliner 2009; B. Schuller, Batliner, et al. 2011; Sethu, Ambikairajah, and Epps 2013; Mariooryad and Busso 2014; F. Eyben, K. Scherer, et al. 2015)
Non–linear dynamics	(Alam et al. 2013; Henríquez et al. 2014; Zao, Cavalcante, and Coelho 2014)
Wavelets	(Huang et al. 2014; Zao, Cavalcante, and Coelho 2014)
Deep learning	(Degaonkar and Apte 2013; Li et al. 2013; Kim, H. Lee, and Provost 2013; Xia et al. 2014)

Introduction: State of the art

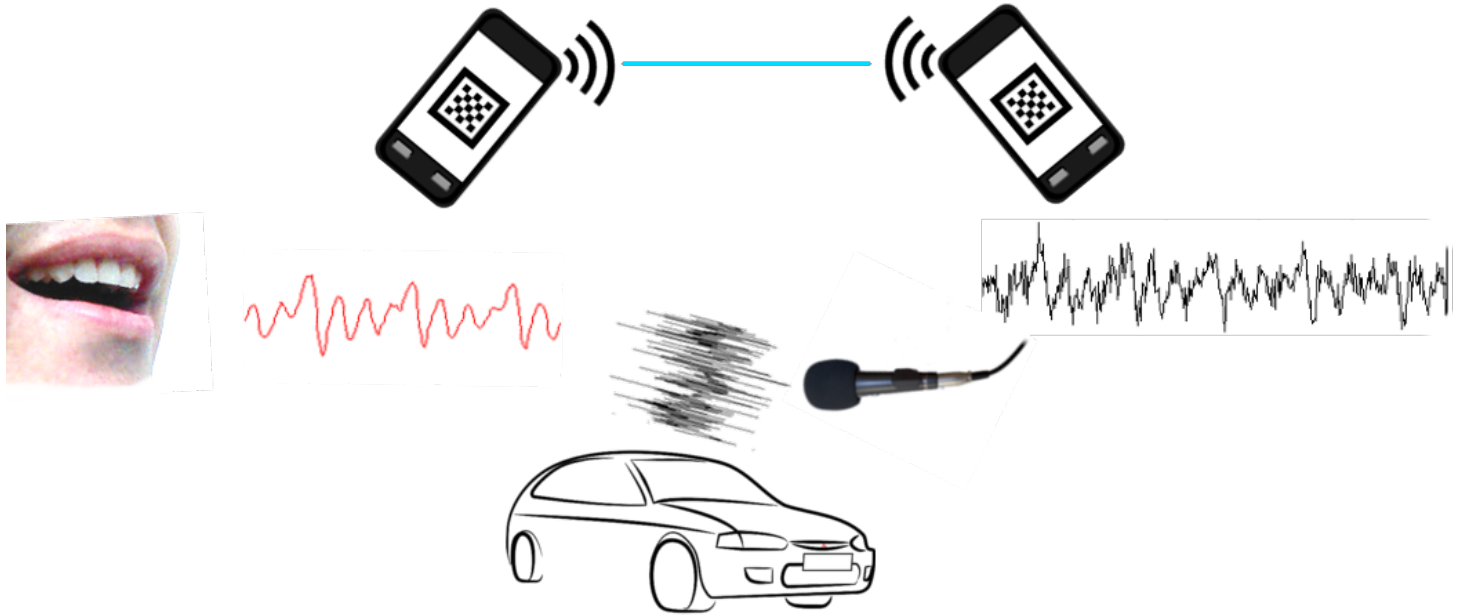
Table: Non–controlled acoustic conditions

Background noise	(B. Schuller, Rigoll, et al. 2007; Tawari and Trivedi 2010)
Telephone channels	(Pohjalainen and Alku 2013; Pohjalainen and Alku 2014)

Outline

Challenges

- Type of databases (Acted, Evoked, Natural)
- Feature extraction techniques
- Acoustic conditions (Telephone, Background noise)



Objectives: General objective

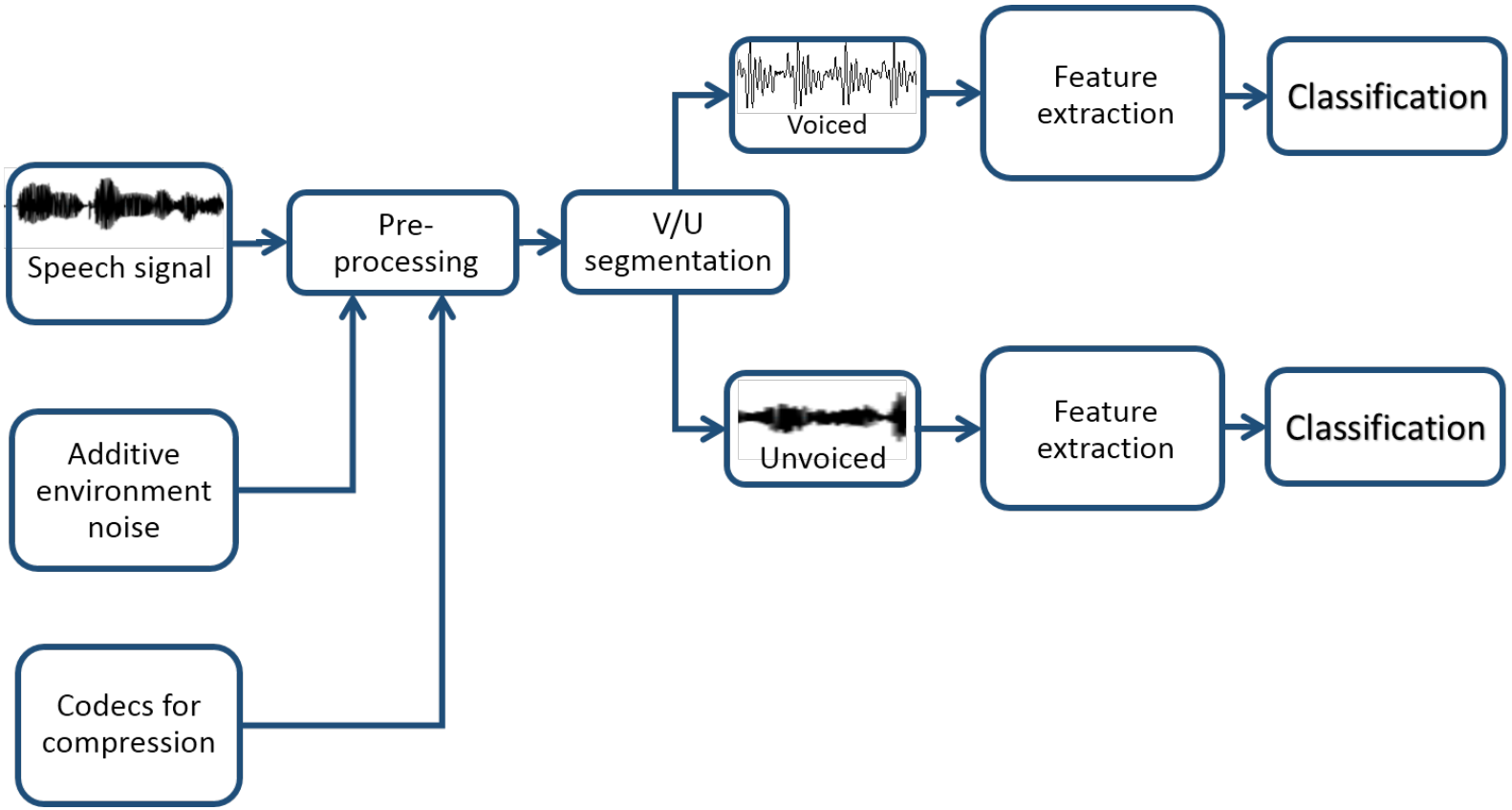
- Propose a methodology to recognize emotions from speech signals in non–controlled acoustic conditions, using several acoustics and non–linear features.

Objectives: Specific objectives

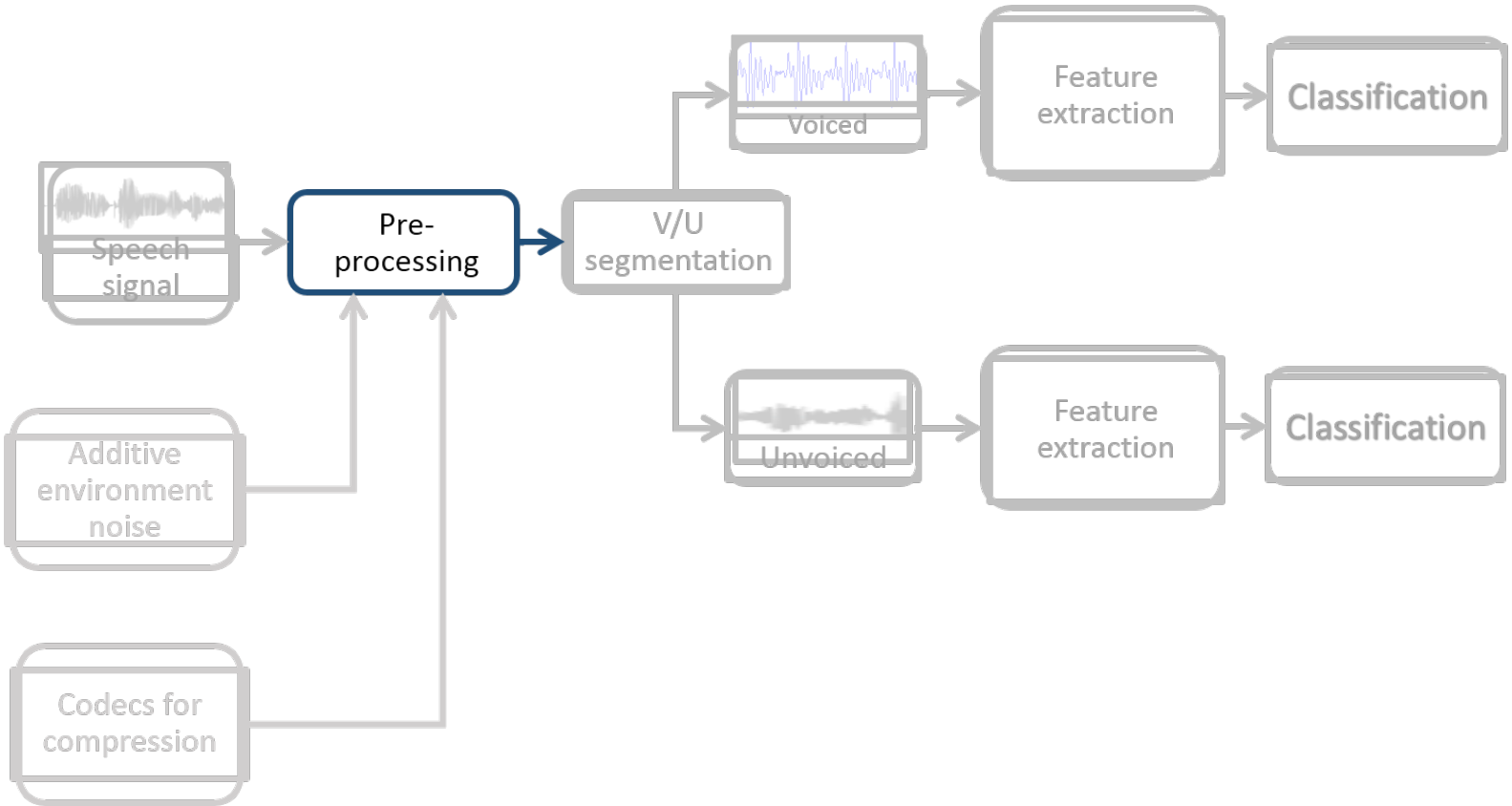
1. Analyze the representation capability of different feature extraction techniques based on acoustic and non–linear analysis with the aim of recognize emotions from speech.
2. Evaluate the effect of different non–controlled acoustic conditions for the recognition of emotions from speech.
3. Evaluate the performance of speech enhancement methods to improve the emotion recognition in non–controlled acoustic conditions.

Outline

Methodology



Methodology: Pre-processing

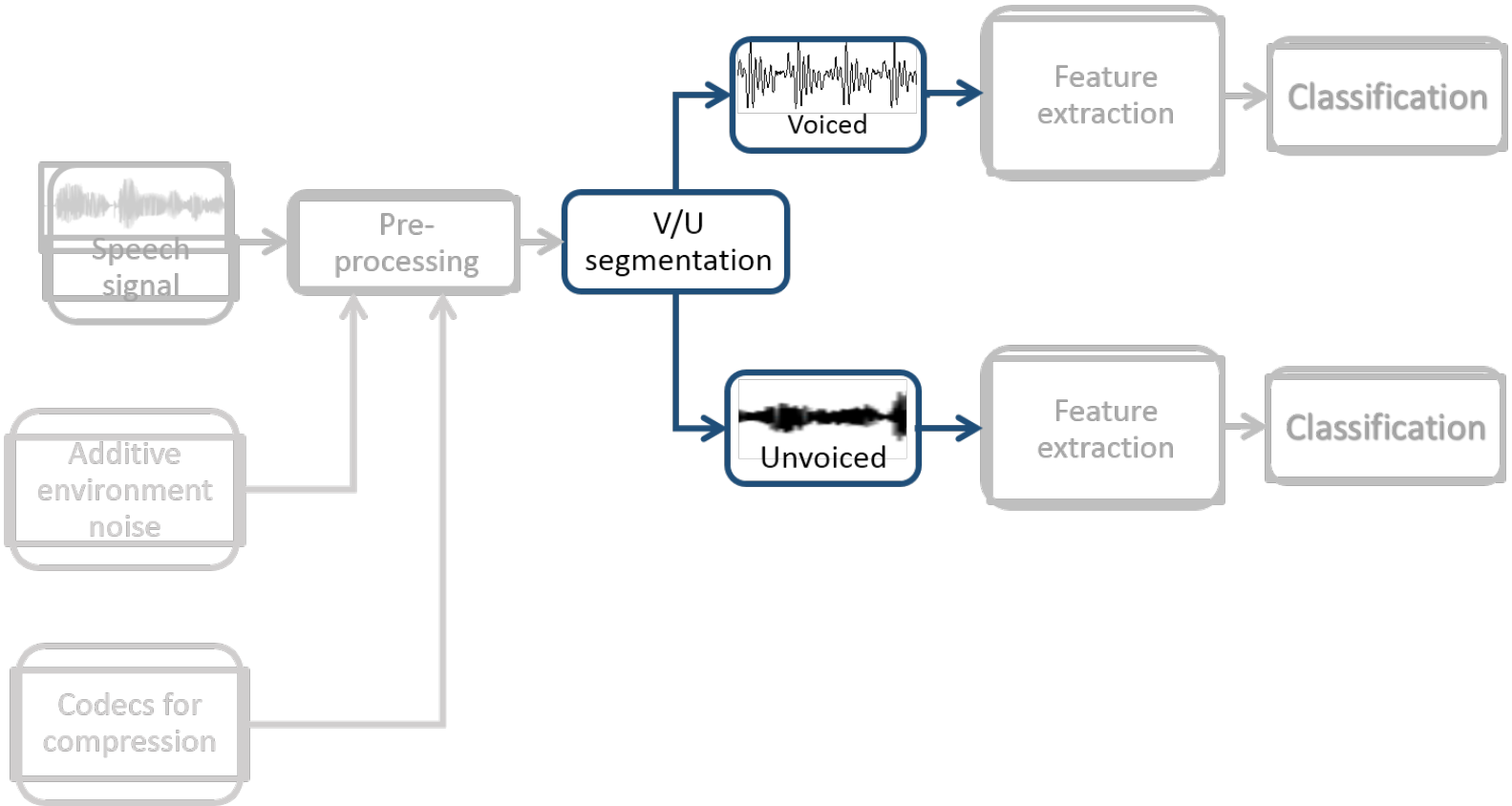


Methodology: Pre–processing

1. Cepstral mean subtraction
2. Amplitude Normalization
3. Speech Enhancement
   - KLT
   - logMMSE

KLT: Karhunen-Loeve Transform.
logMMSE: logarithmic minimum mean square error

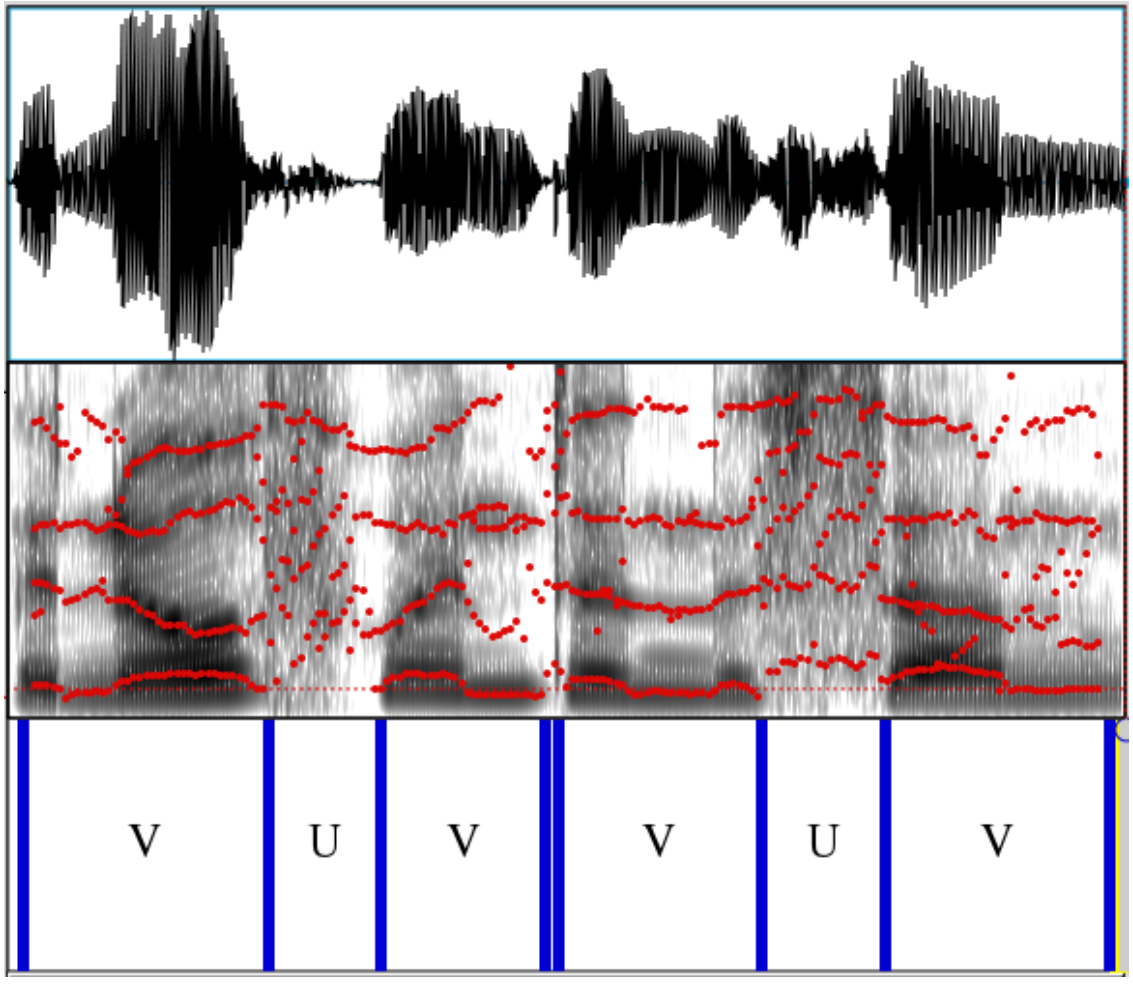
Methodology: Segmentation



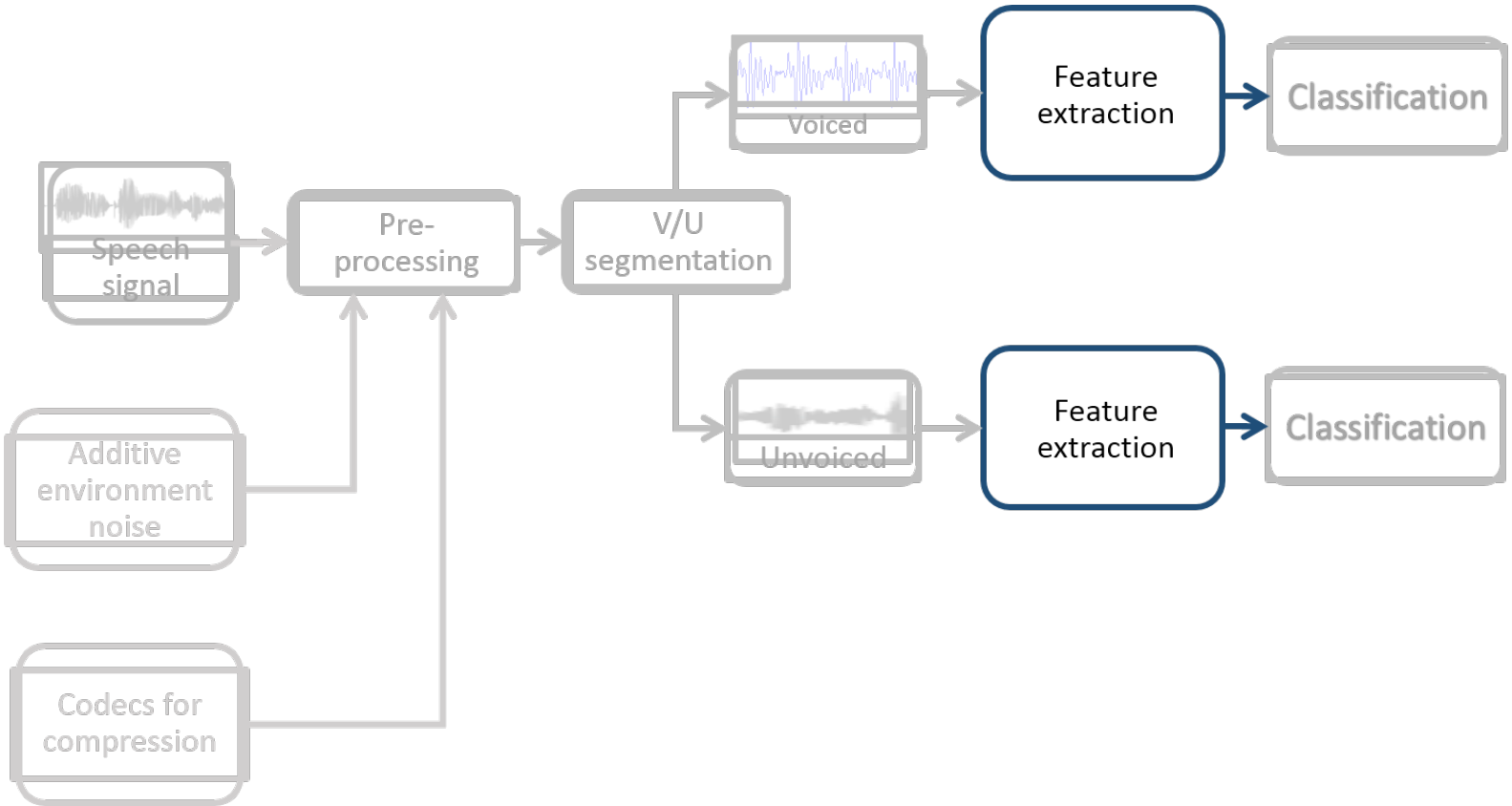
Methodology: Segmentation

Two types of sounds:

- Voiced
- Unvoiced



Methodology: Feature extraction



Methodology: Feature extraction

1. Acoustic Analysis: OpenSmile, prosody
2. Non-Linear Dynamics
3. TARMA models
4. Wavelet Packet Transform (WPT)
5. Time-frequency representations

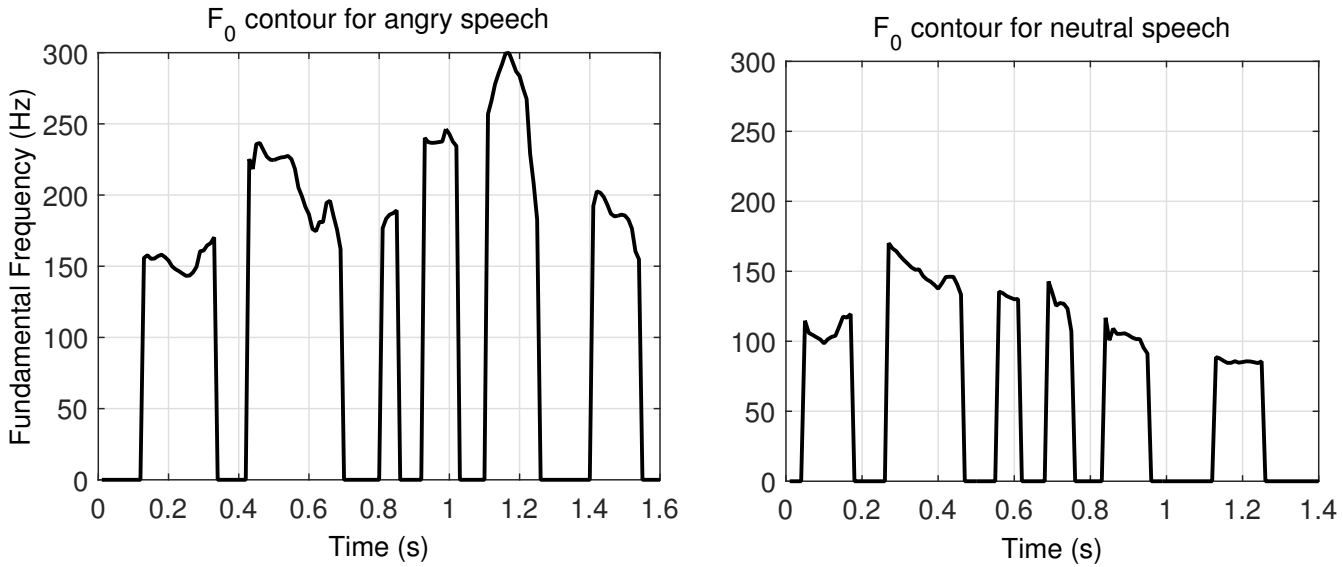
Methodology: Feature extraction 1, OpenSmile :)

Descriptors (16 $\times$ 2)	**statistic functions (12)**
ZCR	mean
RMS Energy	standard deviation
F_0	kurtosis, skewness
HNR	max, min, relative position, range
MFCC 1-12	slope, offset, MSE linear regression
Δs	

Table: Features implemented using OpenSmile[5]

[5] Florian Eyben, Martin Wöllmer, and Björn Schuller. "OpenSmile: the munich versatile and fast open-source audio feature extractor". In: *18th ACM international conference on Multimedia*. ACM. 2010, pp. 1459–1462.

Methodology: Feature extraction 1, Prosody



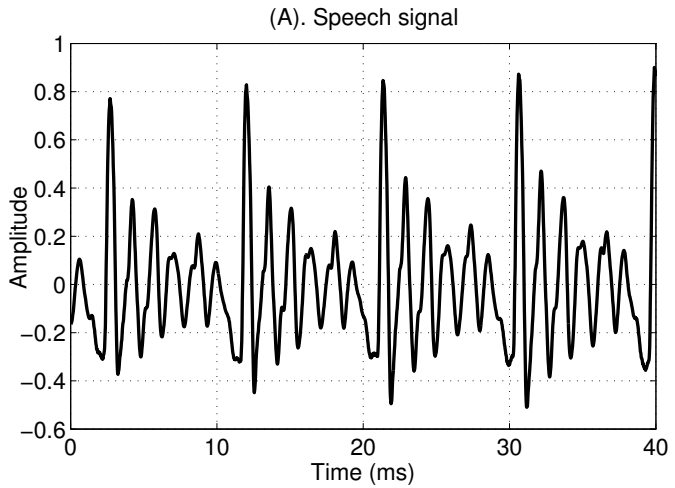
Descriptors	**statistic functions**
Duration	$\mathrm{sil}/(\mathrm{v}+\mathrm{u})$, v/u, $\mathrm{u}/(\mathrm{v}+\mathrm{u})$, $\mathrm{v}/(\mathrm{v}+\mathrm{u})$, v/sil, u/sil
F_0, ΔF_0	mean, max, min, range, std, skewness, kurtosis, median
Energy, ΔEnergy	mean, max, min, range, std, skewness, kurtosis, median

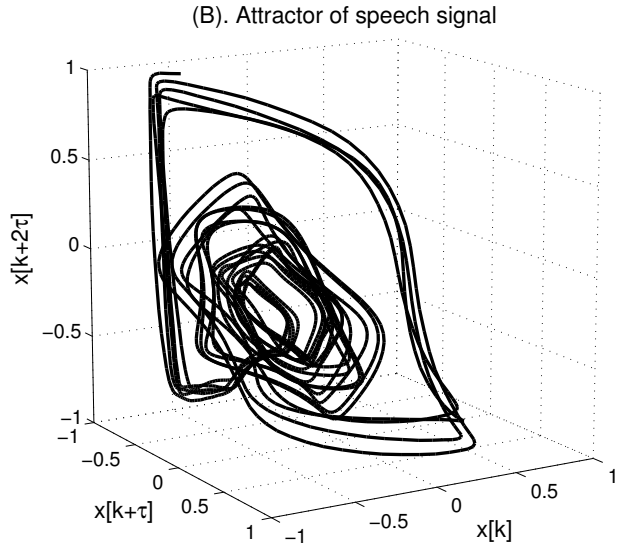
Table: Features implemented derived from prosody

Methodology: Feature extraction 2, Non-linear Dynamics

Features extracted

- Correlation Dimension
- Largest Lyapunov exponent
- Hurst exponent
- Lempel-Ziv complexity
- Shannon entropy
- Log-energy entropy

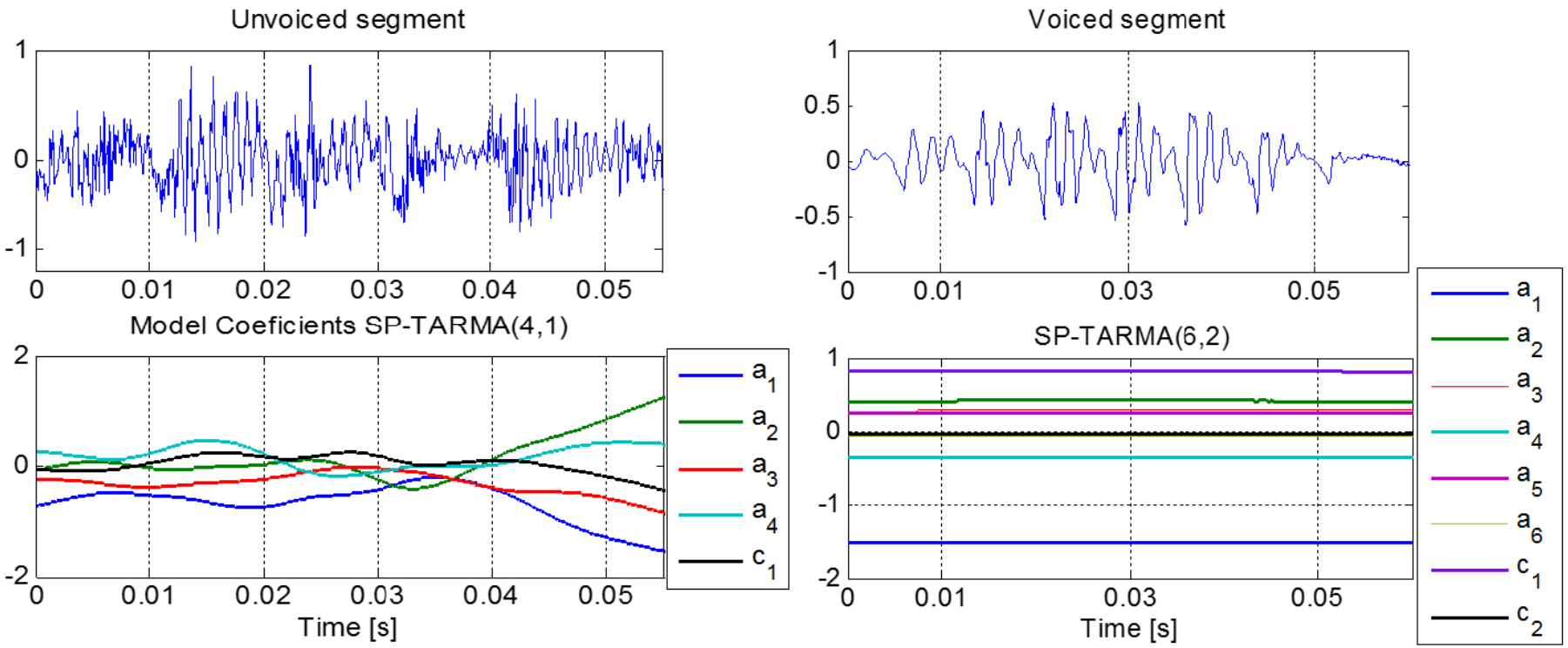




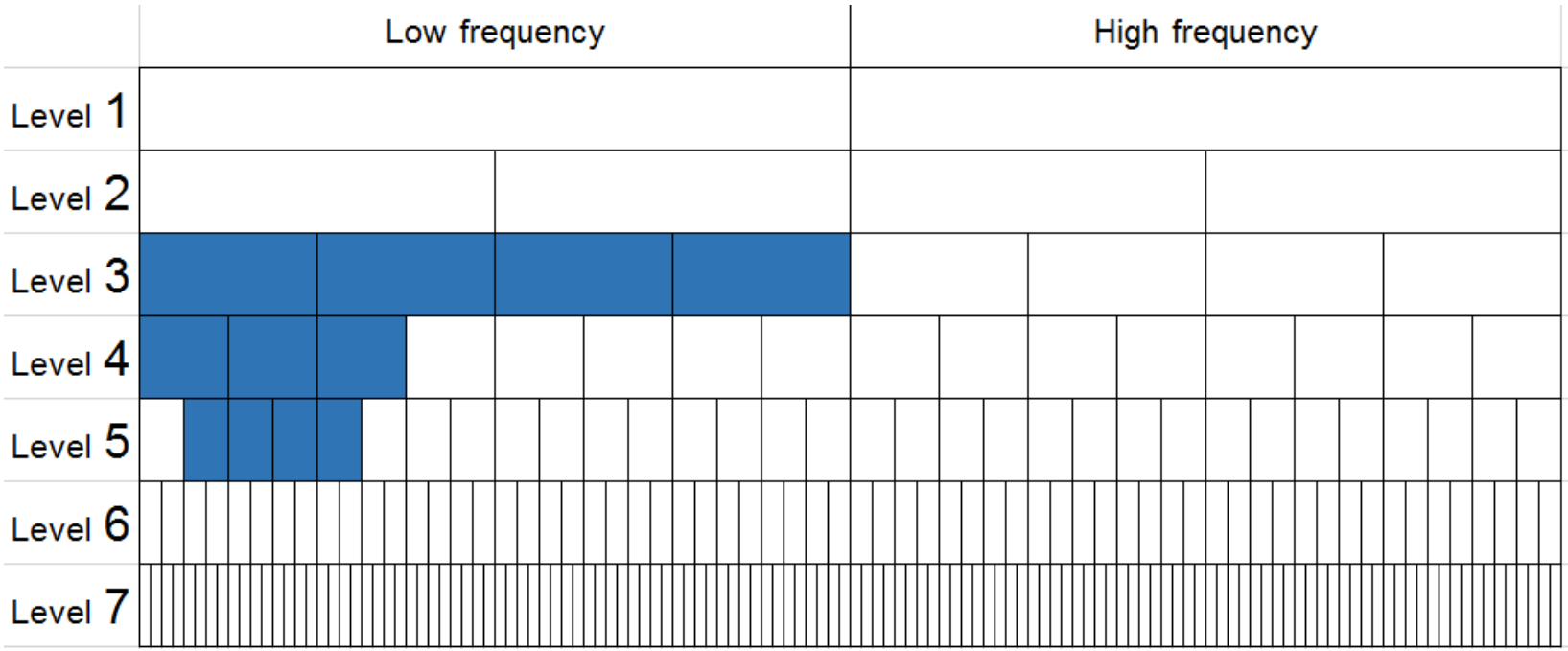
Methodology: Feature extraction 3, TARMA models

Parametric method to analyze non–stationary processes

$$x[t] + \underbrace{\sum_{i=1}^{n_a} a_i[t] \cdot x[t-i]}_{AR\ part} = e[t] + \underbrace{\sum_{i=1}^{n_c} c_i[t] \cdot e[t-i]}_{MA\ part}$$



Methodology: Feature extraction 4, Wavelet packet transform

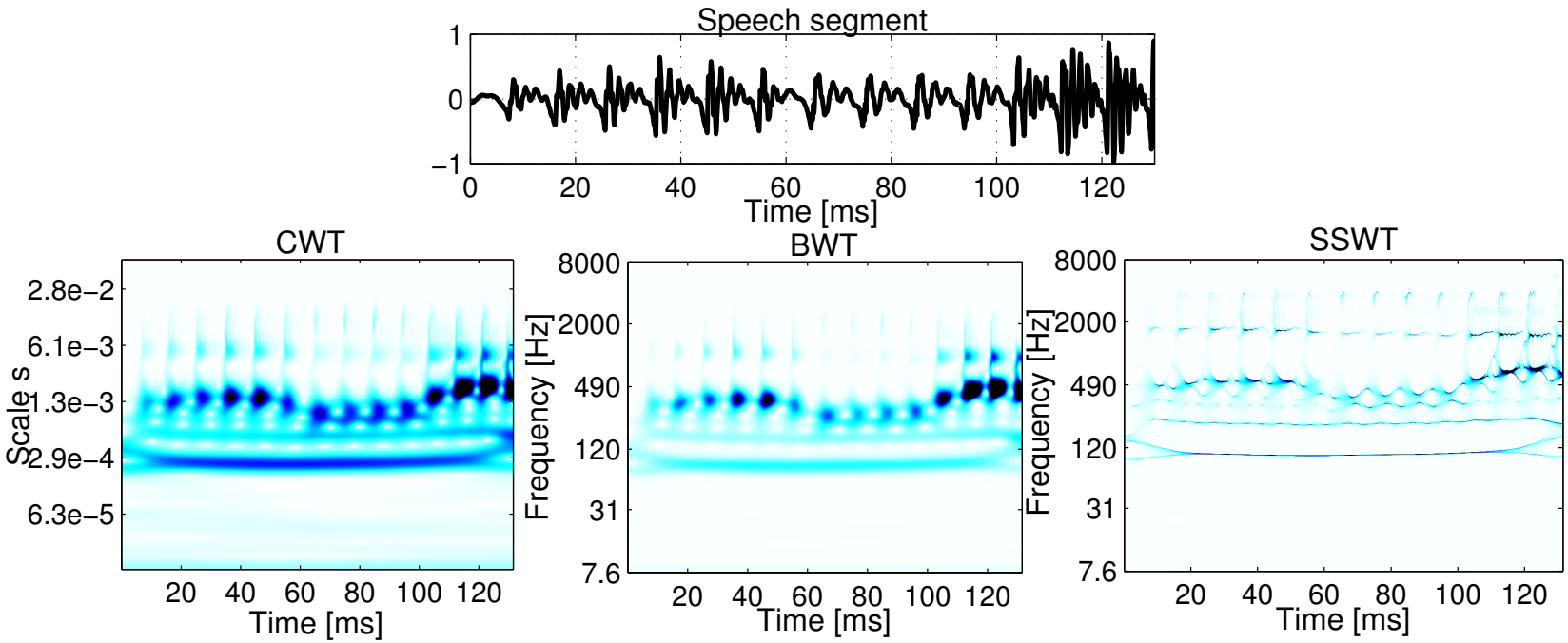


Methodology: Feature extraction 4, Wavelet packet transform

Features extracted from wavelet packets:

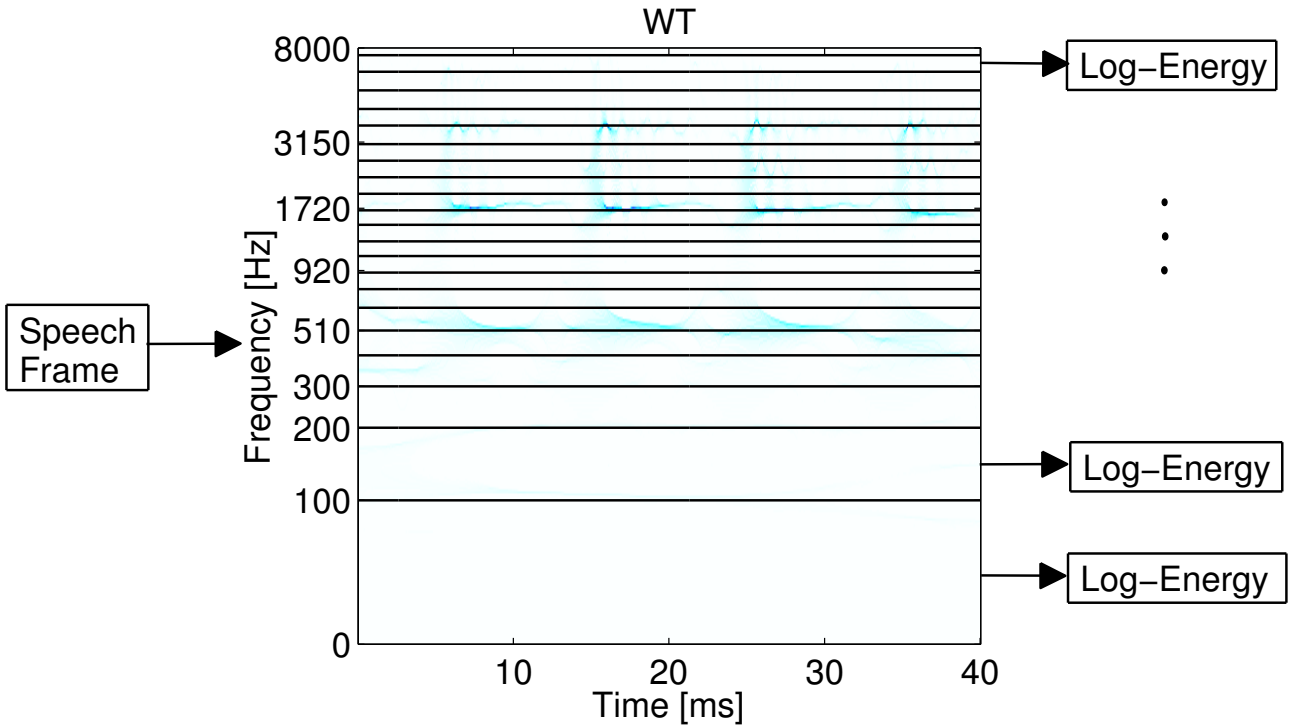
- Log-energy
- Shannon entropy
- Non-linear dynamics measures
- Statistic functionals

Methodology: Feature extraction 5, Time-frequency representations



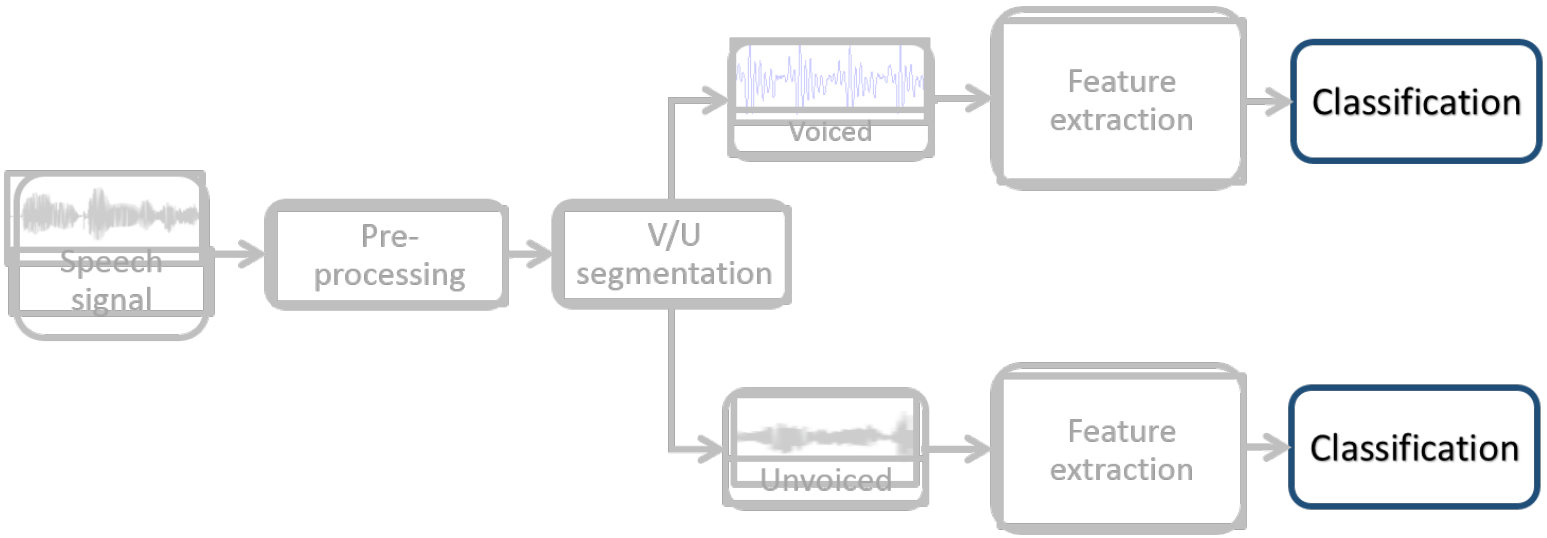
CWT: continuous wavelet transform
BWT: bionic wavelet transform
SSWT: synchro-squeezed wavelet transform

Methodology: Feature extraction 5, Time-frequency representations

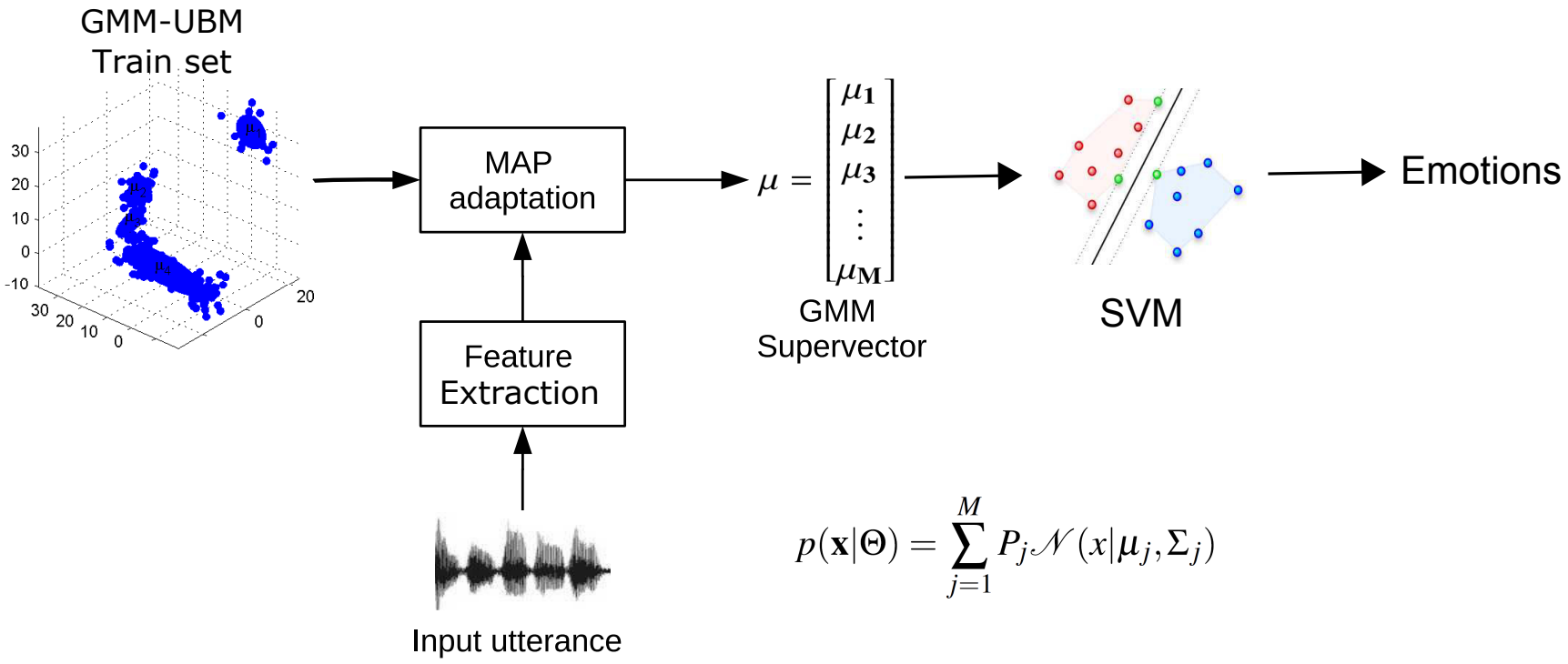


$$E[i] = log \left| \frac{1}{N} \sum_{f_i} \sum_{u_k}^{N} \left| \mathrm{WT}_{(u_k, f_i)} \right|^2 \right| \qquad (1)$$

Methodology: Classification

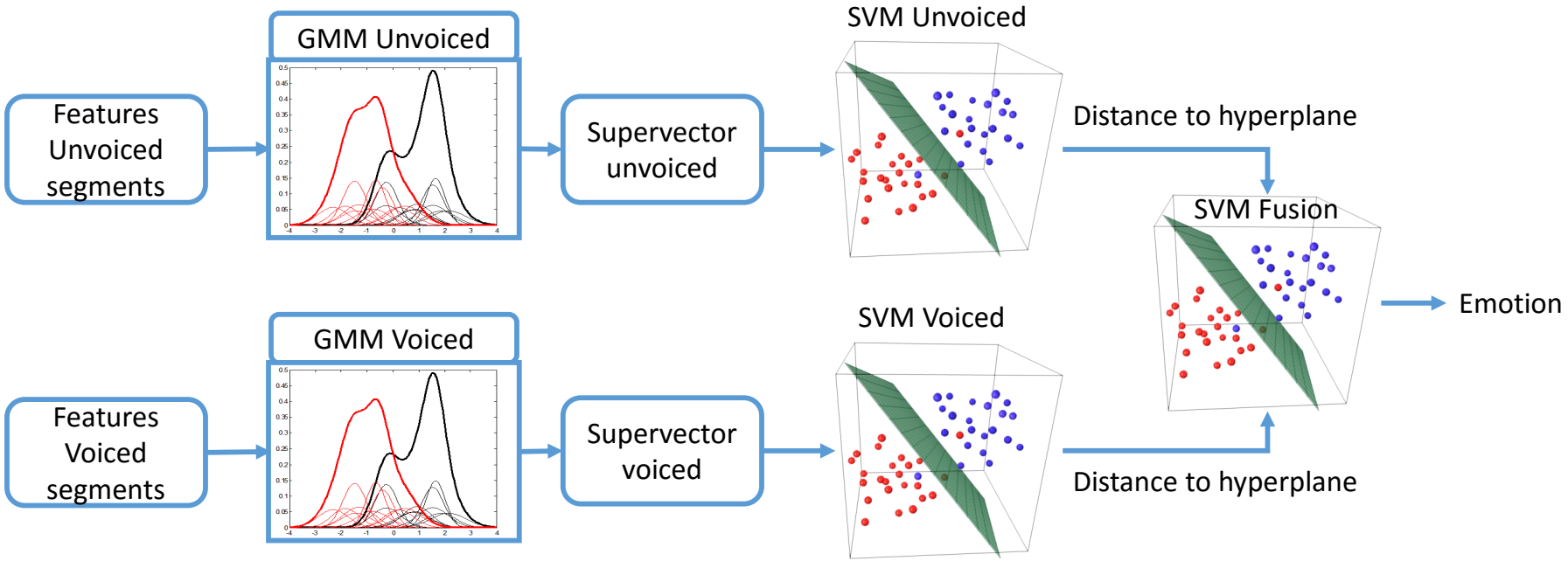


Methodology: Classification



Methodology: Classification

- Supervectors are extracted both for voiced and unvoiced segments based features.
- The scores of the SVM are fused and used as new features for a second SVM.
- Leave one speaker out cross validation is performed.
- UAR as performance measure.



Outline

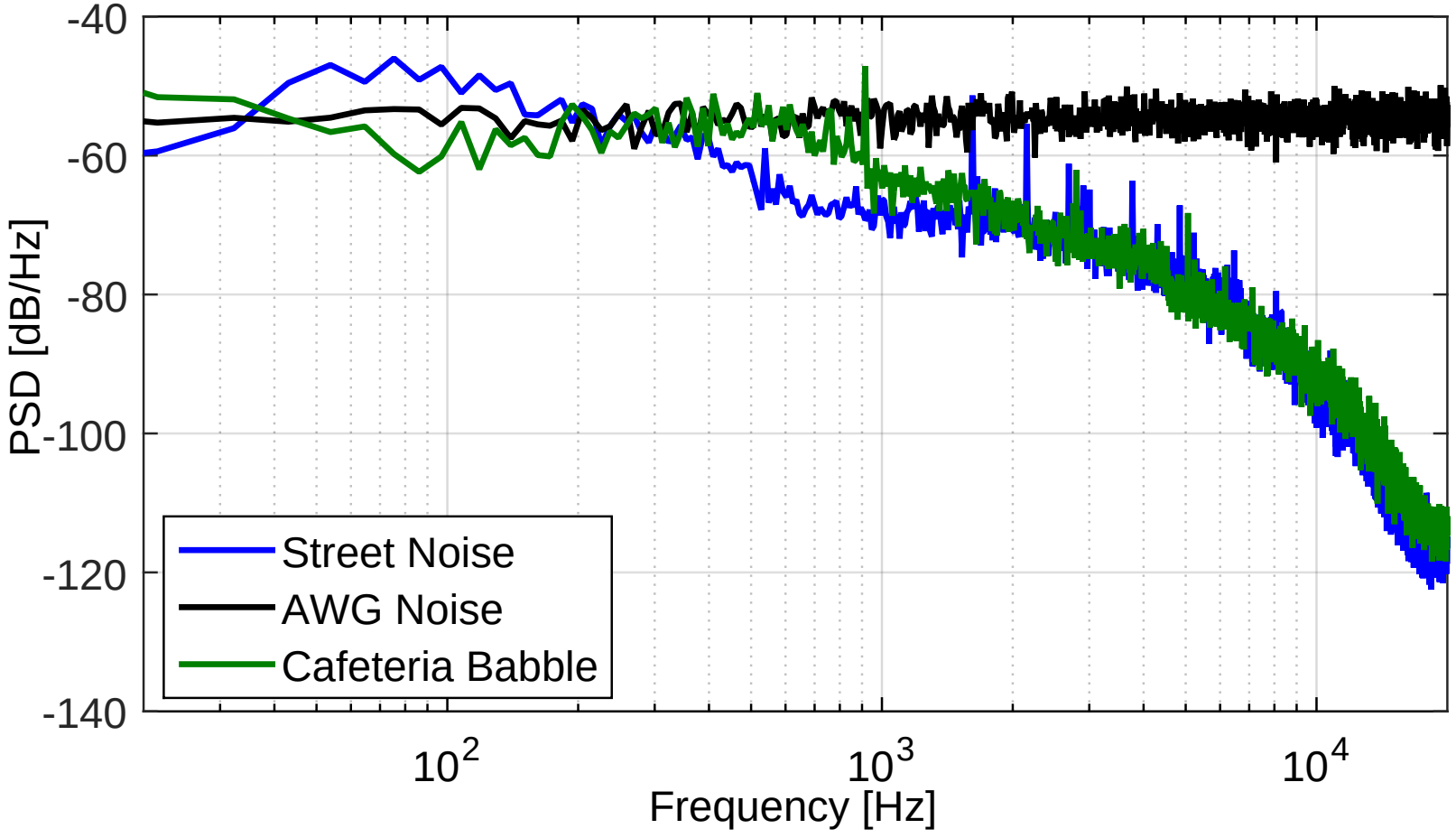
Experiments: Databases

Table: Databases used in this study

Database	# Rec.	# Speak.	Fs (Hz)	Type	Emotions
Berlin	534	10	16000	Acted	Fear, Disgust Happiness, Neutral Boredom, Sadness Anger
IEMOCAP	10039	10	16000	Acted	Fear, Disgust Happiness, Anger Surprise, Excitation Frustration, Sadness Neutral
SAVEE	480	4	44100	Acted	Anger, Happiness Disgust, Fear, Neutral Sadness, Surprise
enterface05	1317	44	44100	Evoked	Fear, Disgust Happiness, Anger Surprise, Sadness
FAU-Aibo	18216	57	16000	Natural	Anger, Emphatic Neutral, Positive Rest

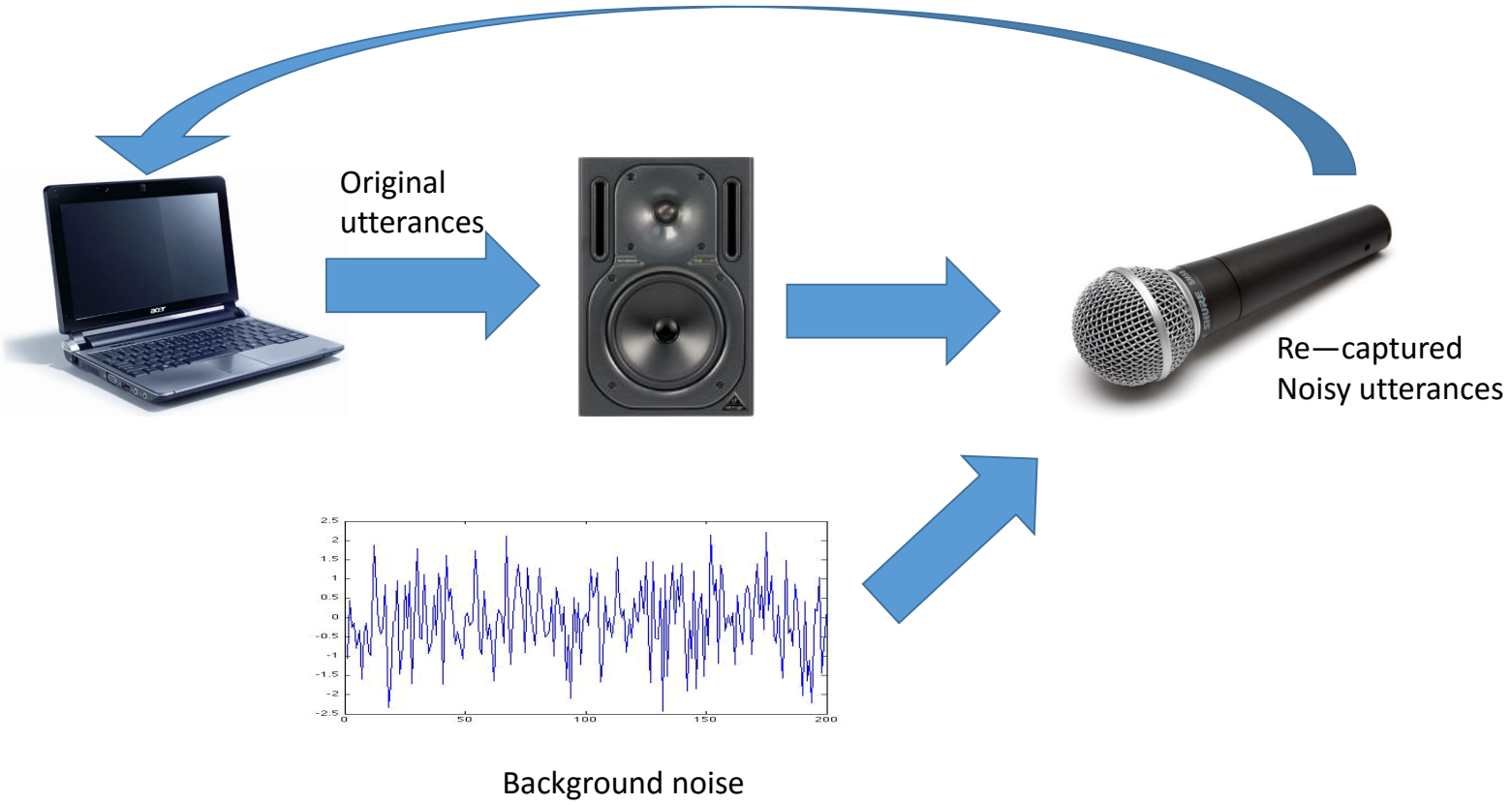
Experiments: additive environment noise

1. Cafeteria babble
2. Street noise



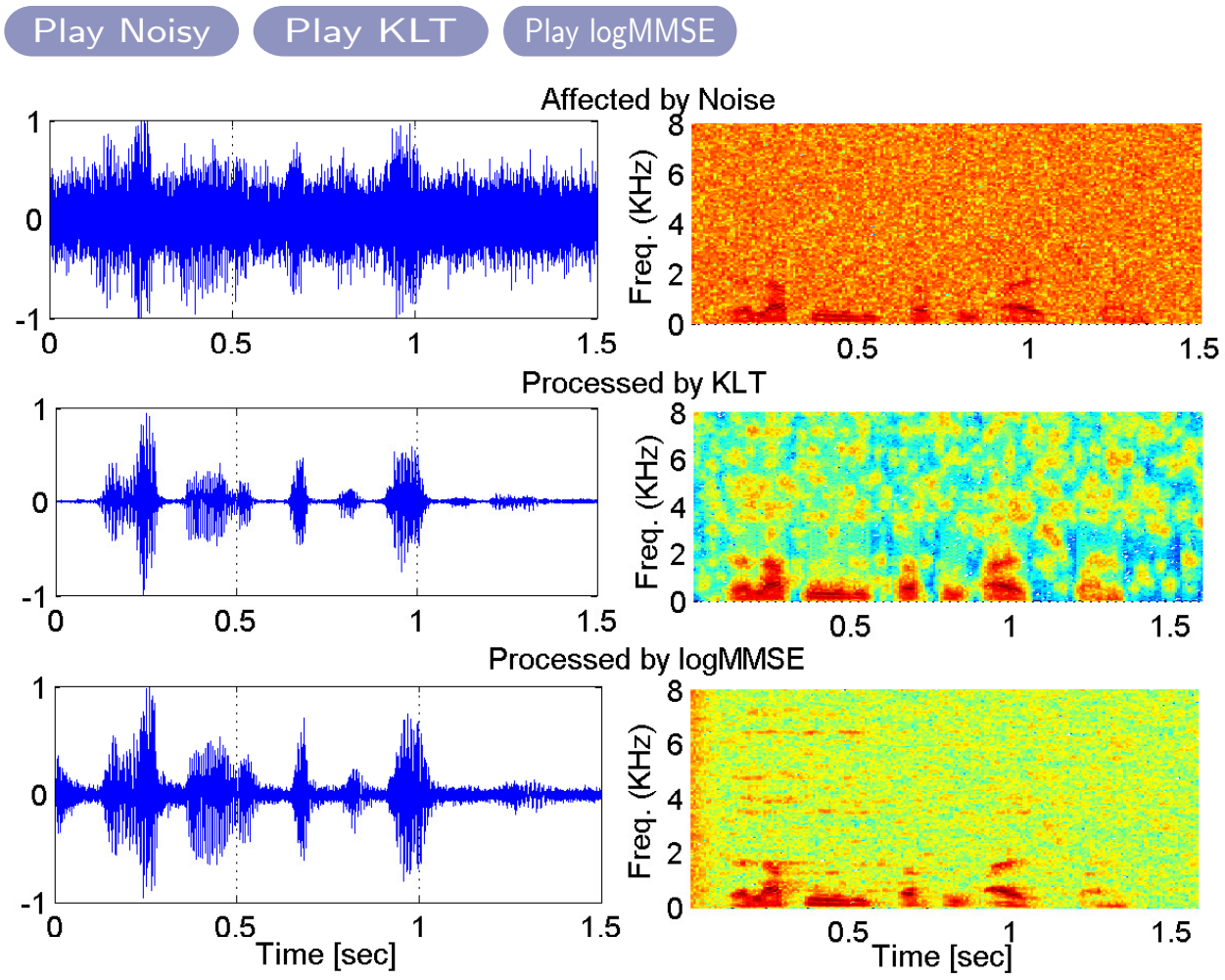
Experiments: non–additive environment noise

1. Office babble
2. Street noise



Experiments: Speech Enhancement

- KLT (Hu and Loizou 2003)
- LogMMSE (Ephraim and Malah 1985)

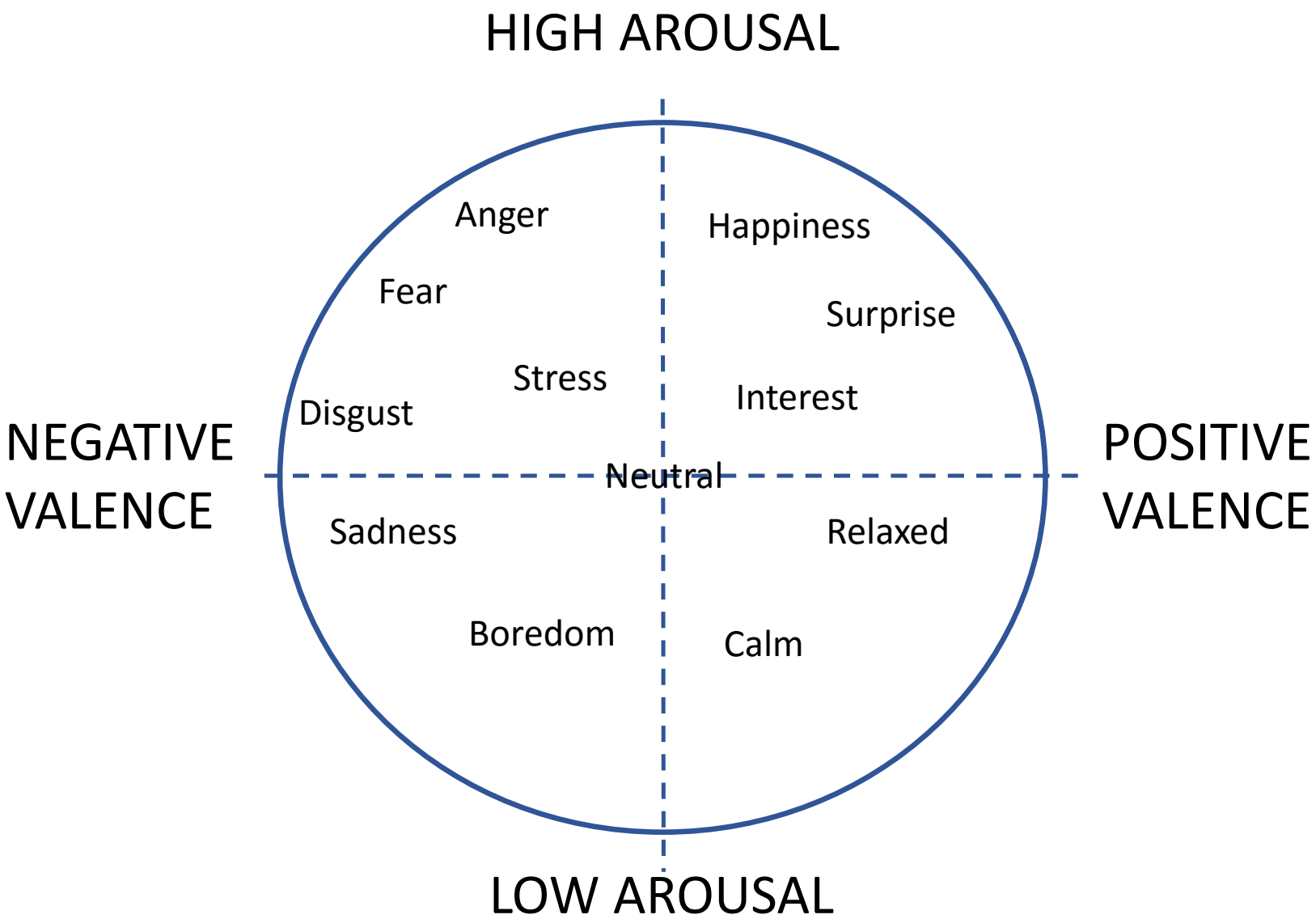


KLT: Karhunen-Loeve Transform.
logMMSE: logarithmic minimum mean square error

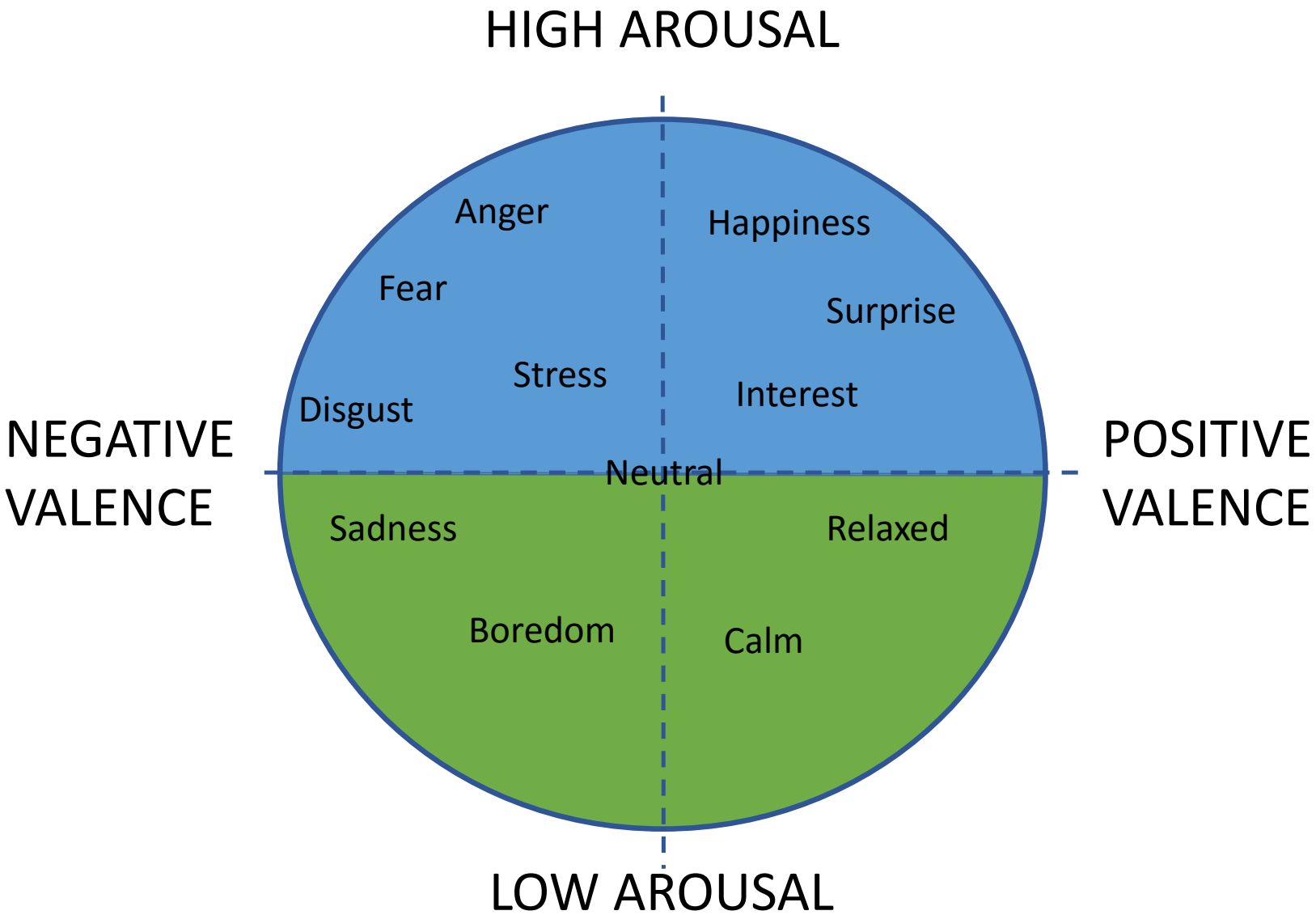
Experiments: codecs

- G.722: LAN VoIP
- G.726: International trunks
- AMR-NB: mobile phone networks
- GSM-FR: mobile phone networks
- AMR-WB: modern mobile networks
- SILK: Skype
- Opus: WebRTC (Google, Facebook)

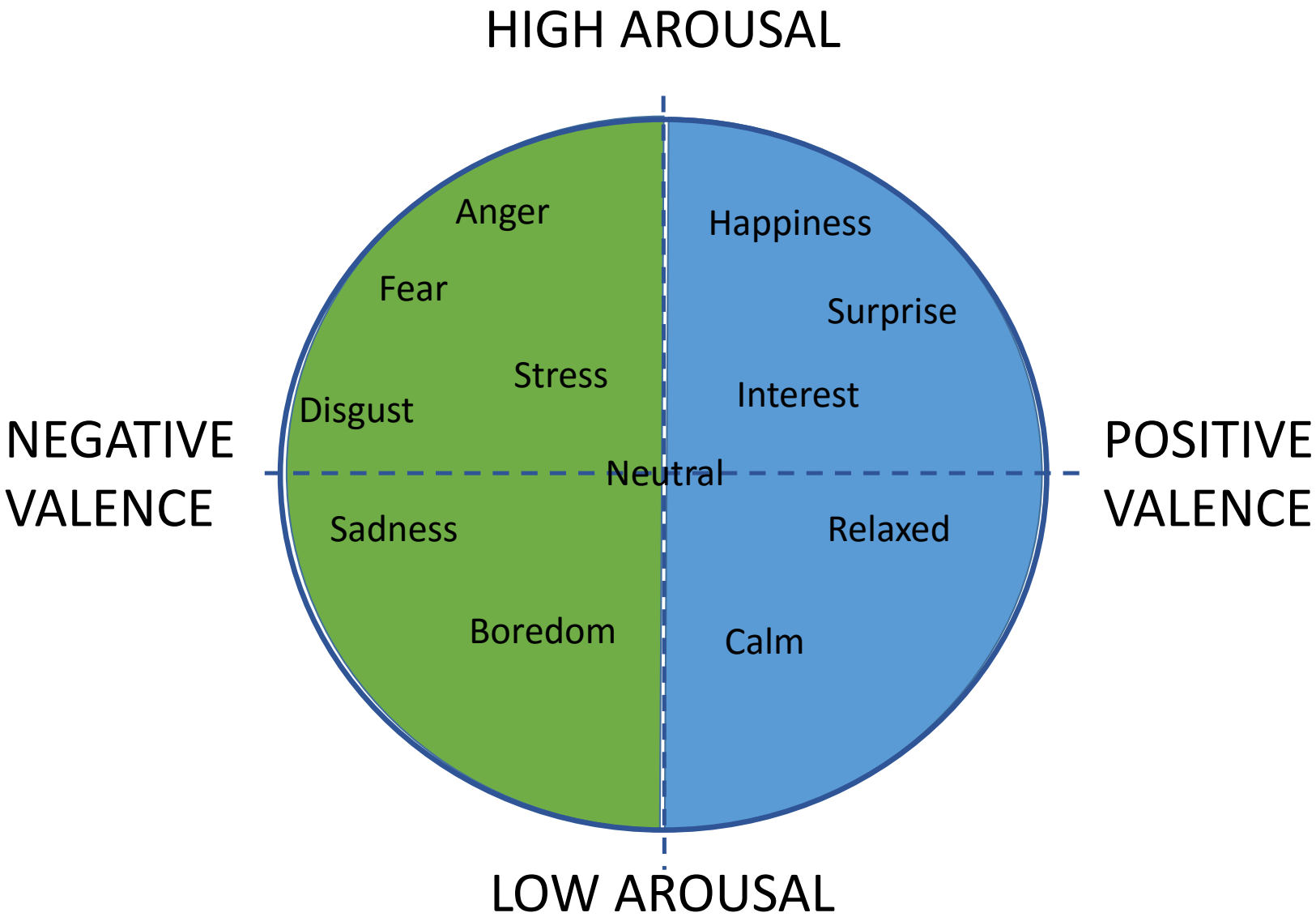
Experiments



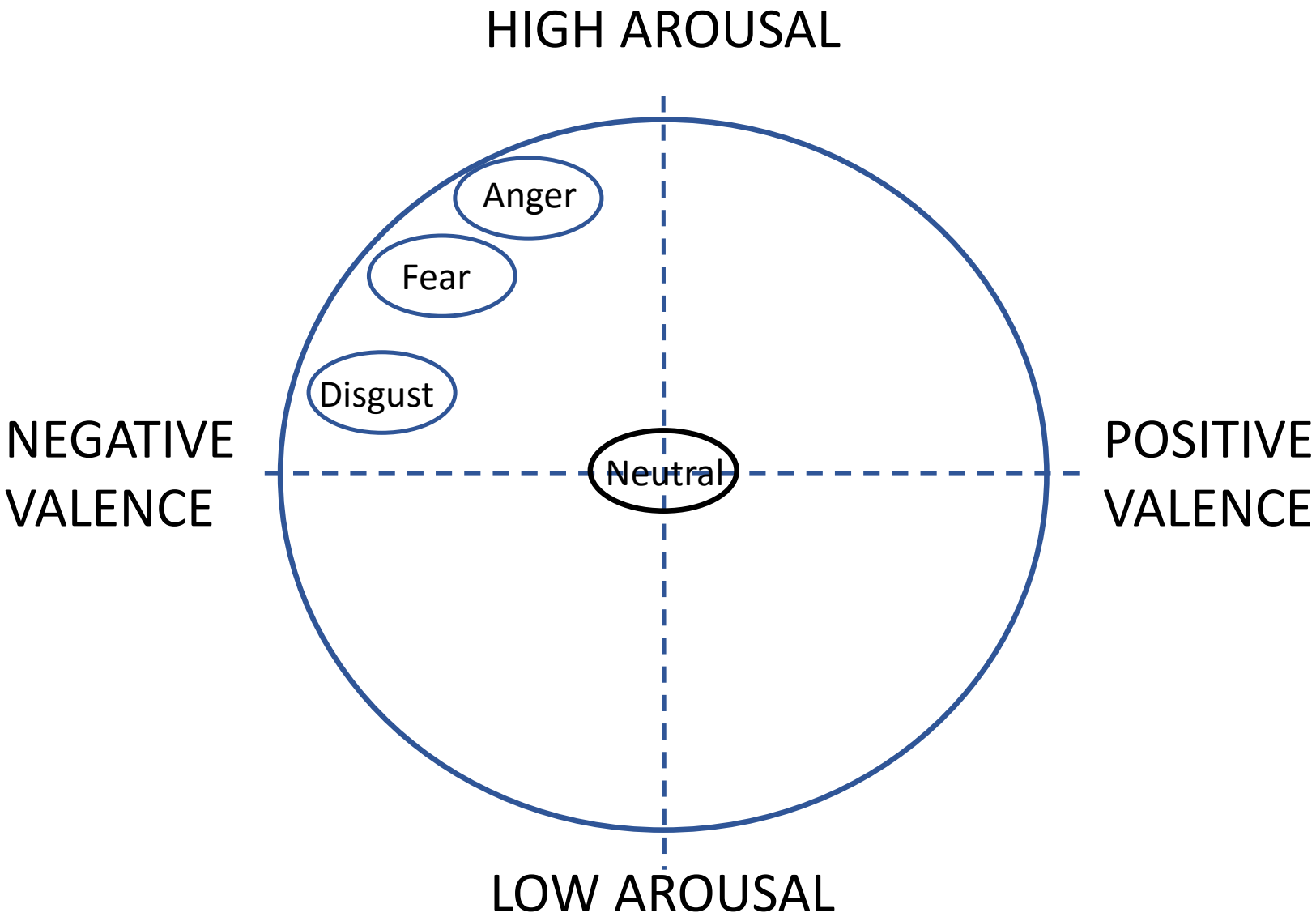
Experiments: High vs. Low Arousal detection



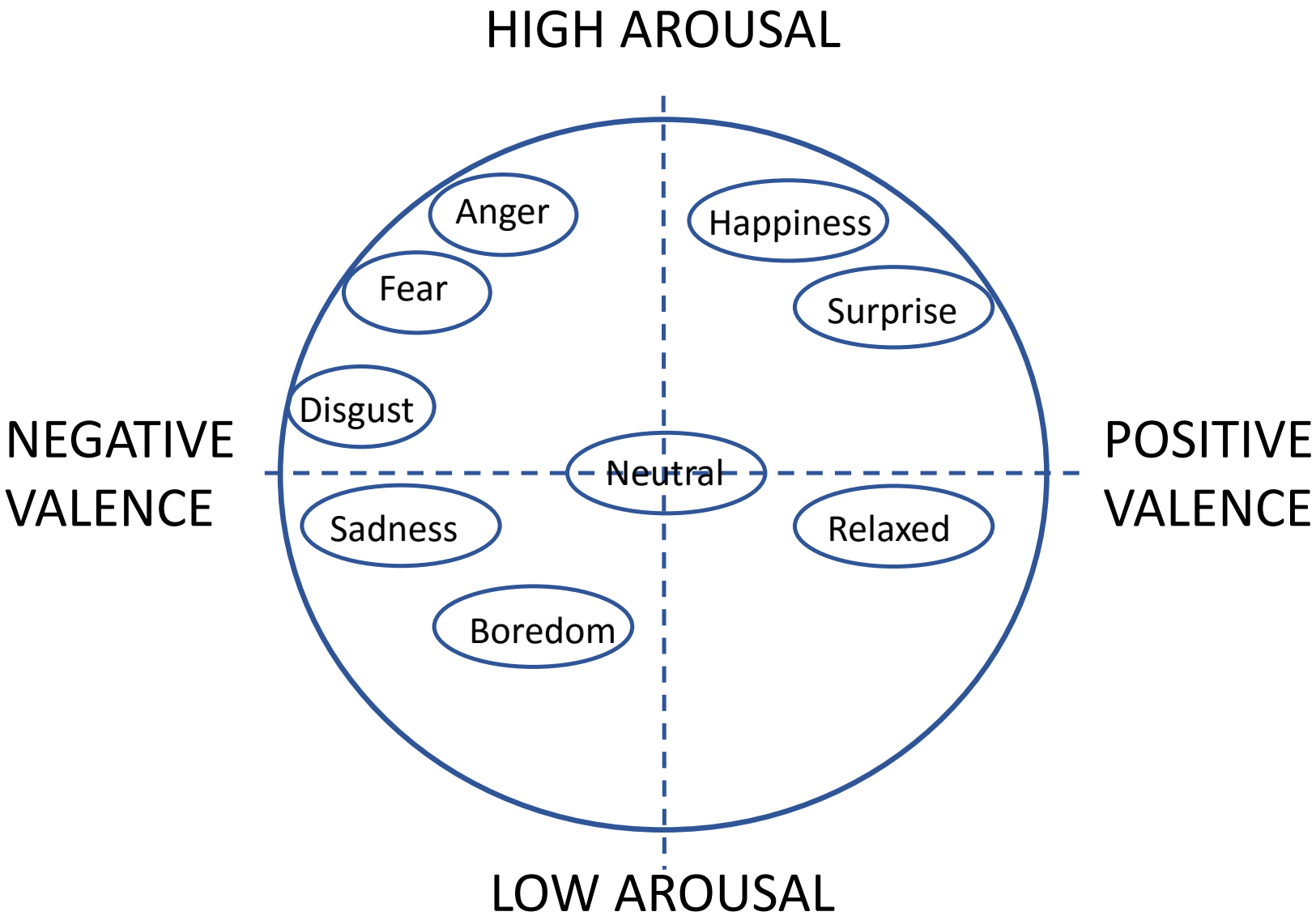
Experiments: Positive vs. Negative Valence detection



Experiments: classification of fear-type emotions



Experiments: classification of multiple emotions



Outline

Results: Original recordings

Table: Performance (%) in Detection of High vs. Low arousal emotions

Features	Segm.	Berlin	SAVEE	enterface05	IEMOCAP
OpenSmile	-	97 ± 3	83 ± 9	81 ± 2	76 ± 4
Prosody	-	92 ± 4	87 ± 5	76 ± 5	72 ± 4
Acoustic+NLD	V	97 ± 4	78 ± 10	80 ± 2	72 ± 6
	U	82 ± 8	81 ± 7	78 ± 2	75 ± 3
	Fusion	93 ± 6	83 ± 6	80 ± 2	72 ± 4
TARMA	U	86 ± 6	71 ± 3	79 ± 1	64 ± 3
WPT	V	96 ± 4	89 ± 6	81 ± 5	75 ± 4
	U	82 ± 6	82 ± 10	78 ± 1	73 ± 6
	Fusion	93 ± 5	87 ± 4	79 ± 2	75 ± 3
SSWT	V	96 ± 6	84 ± 8	81 ± 2	76 ± 5
	U	89 ± 8	80 ± 7	80 ± 1	76 ± 3
	Fusion	95 ± 6	82 ± 6	80 ± 3	77 ± 4

Results: Original recordings

Table: Performance (%) in Detection of High vs. Low arousal emotions

Features	Segm.	Berlin	SAVEE	enterface05	IEMOCAP
OpenSmile	-	97 ± 3	83 ± 9	81 ± 2	76 ± 4
Prosody	-	92 ± 4	87 ± 5	76 ± 5	72 ± 4
Acoustic+NLD	V	97 ± 4	78 ± 10	80 ± 2	72 ± 6
	U	82 ± 8	81 ± 7	78 ± 2	75 ± 3
	Fusion	93 ± 6	83 ± 6	80 ± 2	72 ± 4
TARMA	U	86 ± 6	71 ± 3	79 ± 1	64 ± 3
WPT	V	96 ± 4	89 ± 6	81 ± 5	75 ± 4
	U	82 ± 6	82 ± 10	78 ± 1	73 ± 6
	Fusion	93 ± 5	87 ± 4	79 ± 2	75 ± 3
SSWT	V	96 ± 6	84 ± 8	81 ± 2	76 ± 5
	U	89 ± 8	80 ± 7	80 ± 1	76 ± 3
	Fusion	95 ± 6	82 ± 6	80 ± 3	77 ± 4

Results: Original recordings

Table: Performance (%) in Detection of High vs. Low arousal emotions

Features	Segm.	Berlin	SAVEE	enterface05	IEMOCAP
OpenSmile	-	97 ± 3	83 ± 9	81 ± 2	76 ± 4
Prosody	-	92 ± 4	87 ± 5	76 ± 5	72 ± 4
Acoustic+NLD	V	97 ± 4	78 ± 10	80 ± 2	72 ± 6
	U	82 ± 8	81 ± 7	78 ± 2	75 ± 3
	Fusion	93 ± 6	83 ± 6	80 ± 2	72 ± 4
TARMA	U	86 ± 6	71 ± 3	79 ± 1	64 ± 3
WPT	V	96 ± 4	89 ± 6	81 ± 5	75 ± 4
	U	82 ± 6	82 ± 10	78 ± 1	73 ± 6
	Fusion	93 ± 5	87 ± 4	79 ± 2	75 ± 3
SSWT	V	96 ± 6	84 ± 8	81 ± 2	76 ± 5
	U	89 ± 8	80 ± 7	80 ± 1	76 ± 3
	Fusion	95 ± 6	82 ± 6	80 ± 3	77 ± 4

Results: Original recordings

Table: Performance (%) in Detection of High vs. Low arousal emotions

Features	Segm.	Berlin	SAVEE	enterface05	IEMOCAP
OpenSmile	-	97 ± 3	83 ± 9	81 ± 2	76 ± 4
Prosody	-	92 ± 4	87 ± 5	76 ± 5	72 ± 4
Acoustic+NLD	V	97 ± 4	78 ± 10	80 ± 2	72 ± 6
	U	82 ± 8	81 ± 7	78 ± 2	75 ± 3
	Fusion	93 ± 6	83 ± 6	80 ± 2	72 ± 4
TARMA	U	86 ± 6	71 ± 3	79 ± 1	64 ± 3
WPT	V	96 ± 4	89 ± 6	81 ± 5	75 ± 4
	U	82 ± 6	82 ± 10	78 ± 1	73 ± 6
	Fusion	93 ± 5	87 ± 4	79 ± 2	75 ± 3
SSWT	V	96 ± 6	84 ± 8	81 ± 2	76 ± 5
	U	89 ± 8	80 ± 7	80 ± 1	76 ± 3
	Fusion	95 ± 6	82 ± 6	80 ± 3	77 ± 4

Results: Original recordings

Table: Performance (%) in Detection of High vs. Low arousal emotions

Features	Segm.	Berlin	SAVEE	enterface05	IEMOCAP
OpenSmile	-	97 ± 3	83 ± 9	81 ± 2	76 ± 4
Prosody	-	92 ± 4	87 ± 5	76 ± 5	72 ± 4
Acoustic+NLD	V	97 ± 4	78 ± 10	80 ± 2	72 ± 6
	U	82 ± 8	81 ± 7	78 ± 2	75 ± 3
	Fusion	93 ± 6	83 ± 6	80 ± 2	72 ± 4
TARMA	U	86 ± 6	71 ± 3	79 ± 1	64 ± 3
WPT	V	96 ± 4	89 ± 6	81 ± 5	75 ± 4
	U	82 ± 6	82 ± 10	78 ± 1	73 ± 6
	Fusion	93 ± 5	87 ± 4	79 ± 2	75 ± 3
SSWT	V	96 ± 6	84 ± 8	81 ± 2	76 ± 5
	U	89 ± 8	80 ± 7	80 ± 1	76 ± 3
	Fusion	95 ± 6	82 ± 6	80 ± 3	77 ± 4

Results: Original recordings

Table: Performance (%) in Detection of Positive vs. Negative valence emotions

Features	Segm.	Berlin	SAVEE	enterface05	FAU-Aibo	IEMOCAP
OpenSmile	-	87 ± 2	72 ± 6	81 ± 4	62	59 ± 3
Prosody	-	81 ± 6	68 ± 7	66 ± 6	63	58 ± 2
Acoustic+NLD	V	83 ± 6	67 ± 4	75 ± 2	70	57 ± 3
	U	74 ± 5	63 ± 4	71 ± 2	63	54 ± 3
	Fusion	80 ± 6	67 ± 5	74 ± 5	69	60 ± 3
TARMA	U	74 ± 6	60 ± 3	69 ± 1	56	59 ± 3
WPT	V	81 ± 3	71 ± 10	76 ± 3	68	57 ± 3
	U	75 ± 5	65 ± 4	73 ± 2	65	56 ± 6
	Fusion	76 ± 5	70 ± 8	73 ± 4	68	59 ± 2
SSWT	V	82 ± 5	64 ± 5	76 ± 3	70	56 ± 4
	U	77 ± 6	63 ± 3	74 ± 3	61	58 ± 2
	Fusion	79 ± 4	65 ± 5	74 ± 4	69	60 ± 3

Results: Original recordings

Table: Performance (%) in Detection of Positive vs. Negative valence emotions

Features	Segm.	Berlin	SAVEE	enterface05	FAU-Aibo	IEMOCAP
OpenSmile	-	87 ± 2	72 ± 6	81 ± 4	62	59 ± 3
Prosody	-	81 ± 6	68 ± 7	66 ± 6	63	58 ± 2
Acoustic+NLD	V	83 ± 6	67 ± 4	75 ± 2	70	57 ± 3
	U	74 ± 5	63 ± 4	71 ± 2	63	54 ± 3
	Fusion	80 ± 6	67 ± 5	74 ± 5	69	60 ± 3
TARMA	U	74 ± 6	60 ± 3	69 ± 1	56	59 ± 3
WPT	V	81 ± 3	71 ± 10	76 ± 3	68	57 ± 3
	U	75 ± 5	65 ± 4	73 ± 2	65	56 ± 6
	Fusion	76 ± 5	70 ± 8	73 ± 4	68	59 ± 2
SSWT	V	82 ± 5	64 ± 5	76 ± 3	70	56 ± 4
	U	77 ± 6	63 ± 3	74 ± 3	61	58 ± 2
	Fusion	79 ± 4	65 ± 5	74 ± 4	69	60 ± 3

Results: Original recordings

Table: Performance (%) in Detection of Positive vs. Negative valence emotions

Features	Segm.	Berlin	SAVEE	enterface05	FAU-Aibo	IEMOCAP
OpenSmile	-	87 ± 2	72 ± 6	81 ± 4	62	59 ± 3
Prosody	-	81 ± 6	68 ± 7	66 ± 6	63	58 ± 2
Acoustic+NLD	V	83 ± 6	67 ± 4	75 ± 2	70	57 ± 3
	U	74 ± 5	63 ± 4	71 ± 2	63	54 ± 3
	Fusion	80 ± 6	67 ± 5	74 ± 5	69	60 ± 3
TARMA	U	74 ± 6	60 ± 3	69 ± 1	56	59 ± 3
WPT	V	81 ± 3	71 ± 10	76 ± 3	68	57 ± 3
	U	75 ± 5	65 ± 4	73 ± 2	65	56 ± 6
	Fusion	76 ± 5	70 ± 8	73 ± 4	68	59 ± 2
SSWT	V	82 ± 5	64 ± 5	76 ± 3	70	56 ± 4
	U	77 ± 6	63 ± 3	74 ± 3	61	58 ± 2
	Fusion	79 ± 4	65 ± 5	74 ± 4	69	60 ± 3

Results: Original recordings

Table: Performance (%) in Detection of Positive vs. Negative valence emotions

Features	Segm.	Berlin	SAVEE	enterface05	FAU-Aibo	IEMOCAP
OpenSmile	-	87 ± 2	72 ± 6	81 ± 4	62	59 ± 3
Prosody	-	81 ± 6	68 ± 7	66 ± 6	63	58 ± 2
Acoustic+NLD	V	83 ± 6	67 ± 4	75 ± 2	70	57 ± 3
	U	74 ± 5	63 ± 4	71 ± 2	63	54 ± 3
	Fusion	80 ± 6	67 ± 5	74 ± 5	69	60 ± 3
TARMA	U	74 ± 6	60 ± 3	69 ± 1	56	59 ± 3
WPT	V	81 ± 3	71 ± 10	76 ± 3	68	57 ± 3
	U	75 ± 5	65 ± 4	73 ± 2	65	56 ± 6
	Fusion	76 ± 5	70 ± 8	73 ± 4	68	59 ± 2
SSWT	V	82 ± 5	64 ± 5	76 ± 3	70	56 ± 4
	U	77 ± 6	63 ± 3	74 ± 3	61	58 ± 2
	Fusion	79 ± 4	65 ± 5	74 ± 4	69	60 ± 3

Results: Original recordings

Table: Performance (%) in Classification of fear–type emotions

Features	Segm.	Berlin (4)	enterface05 (3)	SAVEE (4)
OpenSmile	-	91 ± 5	65 ± 18	78 ± 6
Prosody	-	76 ± 7	70 ± 16	53 ± 4
Acoustic+NLD	V	88 ± 10	59 ± 14	70 ± 6
	U	69 ± 9	54 ± 8	57 ± 6
	Fusion	83 ± 10	65 ± 14	67 ± 6
TARMA	U	67 ± 7	62 ± 5	54 ± 5
WPT	V	84 ± 6	71 ± 14	71 ± 5
	U	69 ± 27	60 ± 15	65 ± 4
	Fusion	83 ± 7	72 ± 12	71 ± 9
SSWT	V	88 ± 7	62 ± 13	70 ± 6
	U	80 ± 6	56 ± 7	69 ± 4
	Fusion	90 ± 6	69 ± 9	74 ± 6

Results: Original recordings

Table: Performance (%) in Classification of fear–type emotions

Features	Segm.	Berlin (4)	enterface05 (3)	SAVEE (4)
OpenSmile	-	91 ± 5	65 ± 18	78 ± 6
Prosody	-	76 ± 7	70 ± 16	53 ± 4
Acoustic+NLD	V	88 ± 10	59 ± 14	70 ± 6
	U	69 ± 9	54 ± 8	57 ± 6
	Fusion	83 ± 10	65 ± 14	67 ± 6
TARMA	U	67 ± 7	62 ± 5	54 ± 5
WPT	V	84 ± 6	71 ± 14	71 ± 5
	U	69 ± 27	60 ± 15	65 ± 4
	Fusion	83 ± 7	72 ± 12	71 ± 9
SSWT	V	88 ± 7	62 ± 13	70 ± 6
	U	80 ± 6	56 ± 7	69 ± 4
	Fusion	90 ± 6	69 ± 9	74 ± 6

Results: Original recordings

Table: Performance (%) in Classification of fear–type emotions

Features	Segm.	Berlin (4)	enterface05 (3)	SAVEE (4)
OpenSmile	-	91 ± 5	65 ± 18	78 ± 6
Prosody	-	76 ± 7	70 ± 16	53 ± 4
Acoustic+NLD	V	88 ± 10	59 ± 14	70 ± 6
	U	69 ± 9	54 ± 8	57 ± 6
	Fusion	83 ± 10	65 ± 14	67 ± 6
TARMA	U	67 ± 7	62 ± 5	54 ± 5
WPT	V	84 ± 6	71 ± 14	71 ± 5
	U	69 ± 27	60 ± 15	65 ± 4
	Fusion	83 ± 7	72 ± 12	71 ± 9
SSWT	V	88 ± 7	62 ± 13	70 ± 6
	U	80 ± 6	56 ± 7	69 ± 4
	Fusion	90 ± 6	69 ± 9	74 ± 6

Results: Original recordings

Table: Performance (%) in Classification of fear–type emotions

Features	Segm.	Berlin (4)	enterface05 (3)	SAVEE (4)
OpenSmile	-	91 ± 5	65 ± 18	78 ± 6
Prosody	-	76 ± 7	70 ± 16	53 ± 4
Acoustic+NLD	V	88 ± 10	59 ± 14	70 ± 6
	U	69 ± 9	54 ± 8	57 ± 6
	Fusion	83 ± 10	65 ± 14	67 ± 6
TARMA	U	67 ± 7	62 ± 5	54 ± 5
WPT	V	84 ± 6	71 ± 14	71 ± 5
	U	69 ± 27	60 ± 15	65 ± 4
	Fusion	83 ± 7	72 ± 12	71 ± 9
SSWT	V	88 ± 7	62 ± 13	70 ± 6
	U	80 ± 6	56 ± 7	69 ± 4
	Fusion	90 ± 6	69 ± 9	74 ± 6

Results: Original recordings

Table: Classification of multiple emotions

Features	Segm.	Berlin (7)	SAVEE (7)	enterface (6)	FAU-Aibo (5)	IEMOCAP (4)
OpenSmile	-	80 ± 8	49 ± 18	63 ± 7	33	57 ± 3
Prosody	-	65 ± 7	48 ± 12	32 ± 4	37	51 ± 5
Acoustic+NLD	V	69 ± 10	42 ± 12	49 ± 4	39	50 ± 7
	U	43 ± 6	35 ± 7	34 ± 3	29	52 ± 4
	Fusion	63 ± 11	43 ± 9	48 ± 5	34	56 ± 3
TARMA	U	46 ± 6	34 ± 4	33 ± 3	23	43 ± 3
WPT	V	65 ± 4	50 ± 13	49 ± 3	38	56 ± 2
	U	49 ± 19	42 ± 12	39 ± 4	29	50 ± 9
	Fusion	66 ± 5	52 ± 14	49 ± 6	39	57 ± 4
SSWT	V	64 ± 8	43 ± 11	48 ± 4	33	49 ± 5
	U	55 ± 8	40 ± 6	46 ± 4	22	52 ± 3
	Fusion	69 ± 8	45 ± 12	49 ± 6	31	58 ± 4

Results: Original recordings

Table: Classification of multiple emotions

Features	Segm.	Berlin (7)	SAVEE (7)	enterface (6)	FAU-Aibo (5)	IEMOCAP (4)
OpenSmile	-	80 ± 8	49 ± 18	63 ± 7	33	57 ± 3
Prosody	-	65 ± 7	48 ± 12	32 ± 4	37	51 ± 5
Acoustic+NLD	V	69 ± 10	42 ± 12	49 ± 4	39	50 ± 7
	U	43 ± 6	35 ± 7	34 ± 3	29	52 ± 4
	Fusion	63 ± 11	43 ± 9	48 ± 5	34	56 ± 3
TARMA	U	46 ± 6	34 ± 4	33 ± 3	23	43 ± 3
WPT	V	65 ± 4	50 ± 13	49 ± 3	38	56 ± 2
	U	49 ± 19	42 ± 12	39 ± 4	29	50 ± 9
	Fusion	66 ± 5	52 ± 14	49 ± 6	39	57 ± 4
SSWT	V	64 ± 8	43 ± 11	48 ± 4	33	49 ± 5
	U	55 ± 8	40 ± 6	46 ± 4	22	52 ± 3
	Fusion	69 ± 8	45 ± 12	49 ± 6	31	58 ± 4

Results: Original recordings

Table: Classification of multiple emotions

Features	Segm.	Berlin (7)	SAVEE (7)	enterface (6)	FAU-Aibo (5)	IEMOCAP (4)
OpenSmile	-	80 ± 8	49 ± 18	63 ± 7	33	57 ± 3
Prosody	-	65 ± 7	48 ± 12	32 ± 4	37	51 ± 5
Acoustic+NLD	V	69 ± 10	42 ± 12	49 ± 4	39	50 ± 7
	U	43 ± 6	35 ± 7	34 ± 3	29	52 ± 4
	Fusion	63 ± 11	43 ± 9	48 ± 5	34	56 ± 3
TARMA	U	46 ± 6	34 ± 4	33 ± 3	23	43 ± 3
WPT	V	65 ± 4	50 ± 13	49 ± 3	38	56 ± 2
	U	49 ± 19	42 ± 12	39 ± 4	29	50 ± 9
	Fusion	66 ± 5	52 ± 14	49 ± 6	39	57 ± 4
SSWT	V	64 ± 8	43 ± 11	48 ± 4	33	49 ± 5
	U	55 ± 8	40 ± 6	46 ± 4	22	52 ± 3
	Fusion	69 ± 8	45 ± 12	49 ± 6	31	58 ± 4

Results: Original recordings

Table: Classification of multiple emotions

Features	Segm.	Berlin (7)	SAVEE (7)	enterface (6)	FAU-Aibo (5)	IEMOCAP (4)
OpenSmile	-	80 ± 8	49 ± 18	63 ± 7	33	57 ± 3
Prosody	-	65 ± 7	48 ± 12	32 ± 4	37	51 ± 5
Acoustic+NLD	V	69 ± 10	42 ± 12	49 ± 4	39	50 ± 7
	U	43 ± 6	35 ± 7	34 ± 3	29	52 ± 4
	Fusion	63 ± 11	43 ± 9	48 ± 5	34	56 ± 3
TARMA	U	46 ± 6	34 ± 4	33 ± 3	23	43 ± 3
WPT	V	65 ± 4	50 ± 13	49 ± 3	38	56 ± 2
	U	49 ± 19	42 ± 12	39 ± 4	29	50 ± 9
	Fusion	66 ± 5	52 ± 14	49 ± 6	39	57 ± 4
SSWT	V	64 ± 8	43 ± 11	48 ± 4	33	49 ± 5
	U	55 ± 8	40 ± 6	46 ± 4	22	52 ± 3
	Fusion	69 ± 8	45 ± 12	49 ± 6	31	58 ± 4

Results: Original recordings, Summary

Table: Summary of results for original recordings

Source	# Feat.	Arousal	Valence	All	Fear–type
Berlin database					
openSMILE	384	97	87	80	91
Acoustic+NLD	76	97	83	69	88
WPT	128	96	81	66	84
SSWT	88	96	82	69	90
enterface05 database					
openSMILE	384	81	81	63	65
Acoustic+NLD	76	80	75	49	65
WPT	128	80	76	49	72
SSWT	88	81	76	48	69
IEMOCAP database					
openSMILE	384	76	59	57	-
Acoustic+NLD	76	75	60	56	-
WPT	128	75	59	57	-
SSWT	88	77	60	58	-
FAU-Aibo database					
openSMILE	384	-	62	32	-
Acoustic+NLD	76	-	69	39	-
WPT	128	-	68	38	-
SSWT	88	-	70	33	-

Results: Original recordings, Summary

Table: Summary of results for original recordings

Source	# Feat.	Arousal	Valence	All	Fear–type
Berlin database					
openSMILE	384	97	87	80	91
Acoustic+NLD	76	97	83	69	88
WPT	128	96	81	66	84
SSWT	88	96	82	69	90
enterface05 database					
openSMILE	384	81	81	63	65
Acoustic+NLD	76	80	75	49	65
WPT	128	80	76	49	72
SSWT	88	81	76	48	69
IEMOCAP database					
openSMILE	384	76	59	57	-
Acoustic+NLD	76	75	60	56	-
WPT	128	75	59	57	-
SSWT	88	77	60	58	-
FAU-Aibo database					
openSMILE	384	-	62	32	-
Acoustic+NLD	76	-	69	39	-
WPT	128	-	68	38	-
SSWT	88	-	70	33	-

Results: Original recordings, Summary

Table: Summary of results for original recordings

Source	# Feat.	Arousal	Valence	All	Fear–type
Berlin database					
openSMILE	384	97	87	80	91
Acoustic+NLD	76	97	83	69	88
WPT	128	96	81	66	84
SSWT	88	96	82	69	90
enterface05 database					
openSMILE	384	81	81	63	65
Acoustic+NLD	76	80	75	49	65
WPT	128	80	76	49	72
SSWT	88	81	76	48	69
IEMOCAP database					
openSMILE	384	76	59	57	-
Acoustic+NLD	76	75	60	56	-
WPT	128	75	59	57	-
SSWT	88	77	60	58	-
FAU-Aibo database					
openSMILE	384	-	62	32	-
Acoustic+NLD	76	-	69	39	-
WPT	128	-	68	38	-
SSWT	88	-	70	33	-

Results: Original recordings, Summary

Table: Summary of results for original recordings

Source	# Feat.	Arousal	Valence	All	Fear–type
Berlin database					
openSMILE	384	97	87	80	91
Acoustic+NLD	76	97	83	69	88
WPT	128	96	81	66	84
SSWT	88	96	82	69	90
enterface05 database					
openSMILE	384	81	81	63	65
Acoustic+NLD	76	80	75	49	65
WPT	128	80	76	49	72
SSWT	88	81	76	48	69
IEMOCAP database					
openSMILE	384	76	59	57	-
Acoustic+NLD	76	75	60	56	-
WPT	128	75	59	57	-
SSWT	88	77	60	58	-
FAU-Aibo database					
openSMILE	384	-	62	32	-
Acoustic+NLD	76	-	69	39	-
WPT	128	-	68	38	-
SSWT	88	-	70	33	-

Results: Original recordings, Summary

Table: Summary of results for original recordings

Source	# Feat.	Arousal	Valence	All	Fear–type
Berlin database					
openSMILE	384	97	87	80	91
Acoustic+NLD	76	97	83	69	88
WPT	128	96	81	66	84
SSWT	88	96	82	69	90
enterface05 database					
openSMILE	384	81	81	63	65
Acoustic+NLD	76	80	75	49	65
WPT	128	80	76	49	72
SSWT	88	81	76	48	69
IEMOCAP database					
openSMILE	384	76	59	57	-
Acoustic+NLD	76	75	60	56	-
WPT	128	75	59	57	-
SSWT	88	77	60	58	-
FAU-Aibo database					
openSMILE	384	-	62	32	-
Acoustic+NLD	76	-	69	39	-
WPT	128	-	68	38	-
SSWT	88	-	70	33	-

Results: Additive noise

Table: High vs. Low Arousal

	OpenSMILE			SSWT		
Original	97			96		
	0 dB	3 dB	6 dB	0 dB	3 dB	6 dB
Street noise	96	97	96	92	93	93
Cafeteria noise	96	97	96	93	94	94
KLT Street	92	96	95	88	91	92
KLT Cafeteria	92	96	95	90	90	93
logMMSE Street	96	95	96	93	93	95
logMMSE Cafeteria	96	95	96	94	94	95

Results: Additive noise

Table: High vs. Low Arousal

	OpenSMILE			SSWT		
Original	97			96		
	0 dB	3 dB	6 dB	0 dB	3 dB	6 dB
Street noise	96	97	96	92	93	93
Cafeteria noise	96	97	96	93	94	94
KLT Street	92	96	95	88	91	92
KLT Cafeteria	92	96	95	90	90	93
logMMSE Street	96	95	96	93	93	95
logMMSE Cafeteria	96	95	96	94	94	95

Results: Additive noise

Table: High vs. Low Arousal

	OpenSMILE			SSWT		
Original	97			96		
	0 dB	3 dB	6 dB	0 dB	3 dB	6 dB
Street noise	96	97	96	92	93	93
Cafeteria noise	96	97	96	93	94	94
KLT Street	92	96	95	88	91	92
KLT Cafeteria	92	96	95	90	90	93
logMMSE Street	96	95	96	93	93	95
logMMSE Cafeteria	96	95	96	94	94	95

Results: Additive noise

Table: Positive vs. Negative Valence

	OpenSMILE			SSWT		
Original	87			82		
	0 dB	3 dB	6 dB	0 dB	3 dB	6 dB
Street noise	86	87	87	76	78	80
Cafeteria noise	83	82	82	75	78	78
KLT Street	80	82	80	77	78	79
KLT Cafeteria	77	79	79	75	75	78
logMMSE Street	85	85	86	77	81	78
logMMSE Cafeteria	79	83	83	74	76	78

Results: Additive noise

Table: Positive vs. Negative Valence

	OpenSMILE			SSWT		
Original	87			82		
	0 dB	3 dB	6 dB	0 dB	3 dB	6 dB
Street noise	86	87	87	76	78	80
Cafeteria noise	83	82	82	75	78	78
KLT Street	80	82	80	77	78	79
KLT Cafeteria	77	79	79	75	75	78
logMMSE Street	85	85	86	77	81	78
logMMSE Cafeteria	79	83	83	74	76	78

Results: Additive noise

Table: Positive vs. Negative Valence

	OpenSMILE			SSWT		
Original	87			82		
	0 dB	3 dB	6 dB	0 dB	3 dB	6 dB
Street noise	86	87	87	76	78	80
Cafeteria noise	83	82	82	75	78	78
KLT Street	80	82	80	77	78	79
KLT Cafeteria	77	79	79	75	75	78
logMMSE Street	85	85	86	77	81	78
logMMSE Cafeteria	79	83	83	74	76	78

Results: Additive noise

Table: Positive vs. Negative Valence

	OpenSMILE			SSWT		
Original	87			82		
	0 dB	3 dB	6 dB	0 dB	3 dB	6 dB
Street noise	86	87	87	76	78	80
Cafeteria noise	83	82	82	75	78	78
KLT Street	80	82	80	77	78	79
KLT Cafeteria	77	79	79	75	75	78
logMMSE Street	85	85	86	77	81	78
logMMSE Cafeteria	79	83	83	74	76	78

Results: Additive noise

Table: Fear–type emotions

	OpenSMILE			SSWT		
Original	91			89		
	0 dB	3 dB	6 dB	0 dB	3 dB	6 dB
Street noise	85	90	89	78	81	80
Cafeteria noise	85	86	89	77	81	84
KLT Street	80	81	81	75	78	79
KLT Cafeteria	76	80	79	73	75	75
logMMSE Street	86	88	87	81	82	85
logMMSE Cafeteria	83	83	86	78	79	81

Results: Additive noise

Table: Fear–type emotions

	OpenSMILE			SSWT		
Original	91			89		
	0 dB	3 dB	6 dB	0 dB	3 dB	6 dB
Street noise	85	90	89	78	81	80
Cafeteria noise	85	86	89	77	81	84
KLT Street	80	81	81	75	78	79
KLT Cafeteria	76	80	79	73	75	75
logMMSE Street	86	88	87	81	82	85
logMMSE Cafeteria	83	83	86	78	79	81

Results: Additive noise

Table: Fear–type emotions

	OpenSMILE			SSWT		
Original	91			89		
	0 dB	3 dB	6 dB	0 dB	3 dB	6 dB
Street noise	85	90	89	78	81	80
Cafeteria noise	85	86	89	77	81	84
KLT Street	80	81	81	75	78	79
KLT Cafeteria	76	80	79	73	75	75
logMMSE Street	86	88	87	81	82	85
logMMSE Cafeteria	83	83	86	78	79	81

Results: Additive noise

Table: Fear–type emotions

	OpenSMILE			SSWT		
Original	91			89		
	0 dB	3 dB	6 dB	0 dB	3 dB	6 dB
Street noise	85	90	89	78	81	80
Cafeteria noise	85	86	89	77	81	84
KLT Street	80	81	81	75	78	79
KLT Cafeteria	76	80	79	73	75	75
logMMSE Street	86	88	87	81	82	85
logMMSE Cafeteria	83	83	86	78	79	81

Results: Additive noise

Table: Fear–type emotions

	OpenSMILE			SSWT		
Original	91			89		
	0 dB	3 dB	6 dB	0 dB	3 dB	6 dB
Street noise	85	90	89	78	81	80
Cafeteria noise	85	86	89	77	81	84
KLT Street	80	81	81	75	78	79
KLT Cafeteria	76	80	79	73	75	75
logMMSE Street	86	88	87	81	82	85
logMMSE Cafeteria	83	83	86	78	79	81

Results: Additive noise

Table: Multiple emotions

	OpenSMILE			SSWT		
Original	80			64		
	0 dB	3 dB	6 dB	0 dB	3 dB	6 dB
Street noise	74	77	77	57	57	59
Cafeteria noise	65	69	71	56	58	62
KLT Street	63	67	66	56	58	59
KLT Cafeteria	54	63	61	51	55	56
logMMSE Street	73	73	75	59	59	64
logMMSE Cafeteria	62	68	69	55	54	58

Results: Additive noise

Table: Multiple emotions

	OpenSMILE			SSWT		
Original	80			64		
	0 dB	3 dB	6 dB	0 dB	3 dB	6 dB
Street noise	74	77	77	57	57	59
Cafeteria noise	65	69	71	56	58	62
KLT Street	63	67	66	56	58	59
KLT Cafeteria	54	63	61	51	55	56
logMMSE Street	73	73	75	59	59	64
logMMSE Cafeteria	62	68	69	55	54	58

Results: Additive noise

Table: Multiple emotions

	OpenSMILE			SSWT		
Original	80			64		
	0 dB	3 dB	6 dB	0 dB	3 dB	6 dB
Street noise	74	77	77	57	57	59
Cafeteria noise	65	69	71	56	58	62
KLT Street	63	67	66	56	58	59
KLT Cafeteria	54	63	61	51	55	56
logMMSE Street	73	73	75	59	59	64
logMMSE Cafeteria	62	68	69	55	54	58

Results: Additive noise

Table: Multiple emotions

	OpenSMILE			SSWT		
Original	80			64		
	0 dB	3 dB	6 dB	0 dB	3 dB	6 dB
Street noise	74	77	77	57	57	59
Cafeteria noise	65	69	71	56	58	62
KLT Street	63	67	66	56	58	59
KLT Cafeteria	54	63	61	51	55	56
logMMSE Street	73	73	75	59	59	64
logMMSE Cafeteria	62	68	69	55	54	58

Results: Non–additive noise

Table: Results for Berlin DB re-captured in noisy environments

	High–Low arousal **SSWT**	Pos.–Neg. valence **SSWT**	Fear-Type **SSWT**	All **SSWT**
Original	96 ± 6	82 ± 5	88 ± 7	64 ± 8
Street Noise	96 ± 6	82 ± 5	86 ± 9	63 ± 4
Office Noise	97 ± 4	81 ± 6	87 ± 9	65 ± 8
KLT Street	96 ± 6	82 ± 5	86 ± 9	64 ± 7
KLT Office	97 ± 5	82 ± 3	85 ± 8	64 ± 8
logMMSE Street	96 ± 5	81 ± 7	83 ± 10	60 ± 6
logMMSE Office	96 ± 3	82 ± 6	84 ± 6	62 ± 6

Results: Non–additive noise

Table: Results for Berlin DB re-captured in noisy environments

	High–Low arousal **SSWT**	Pos.–Neg. valence **SSWT**	Fear-Type **SSWT**	All **SSWT**
Original	96 ± 6	82 ± 5	88 ± 7	64 ± 8
Street Noise	96 ± 6	82 ± 5	86 ± 9	63 ± 4
Office Noise	97 ± 4	81 ± 6	87 ± 9	65 ± 8
KLT Street	96 ± 6	82 ± 5	86 ± 9	64 ± 7
KLT Office	97 ± 5	82 ± 3	85 ± 8	64 ± 8
logMMSE Street	96 ± 5	81 ± 7	83 ± 10	60 ± 6
logMMSE Office	96 ± 3	82 ± 6	84 ± 6	62 ± 6

Results: Non–additive noise

Table: Results for Berlin DB re-captured in noisy environments

	High–Low arousal **SSWT**	Pos.–Neg. valence **SSWT**	Fear-Type **SSWT**	All **SSWT**
Original	96 ± 6	82 ± 5	88 ± 7	64 ± 8
Street Noise	96 ± 6	82 ± 5	86 ± 9	63 ± 4
Office Noise	97 ± 4	81 ± 6	87 ± 9	65 ± 8
KLT Street	96 ± 6	82 ± 5	86 ± 9	64 ± 7
KLT Office	97 ± 5	82 ± 3	85 ± 8	64 ± 8
logMMSE Street	96 ± 5	81 ± 7	83 ± 10	60 ± 6
logMMSE Office	96 ± 3	82 ± 6	84 ± 6	62 ± 6

Results: Non–additive noise

Table: Results for Berlin DB re-captured in noisy environments

	High–Low arousal **SSWT**	Pos.–Neg. valence **SSWT**	Fear-Type **SSWT**	All **SSWT**
Original	96 ± 6	82 ± 5	88 ± 7	64 ± 8
Street Noise	96 ± 6	82 ± 5	86 ± 9	63 ± 4
Office Noise	97 ± 4	81 ± 6	87 ± 9	65 ± 8
KLT Street	96 ± 6	82 ± 5	86 ± 9	64 ± 7
KLT Office	97 ± 5	82 ± 3	85 ± 8	64 ± 8
logMMSE Street	96 ± 5	81 ± 7	83 ± 10	60 ± 6
logMMSE Office	96 ± 3	82 ± 6	84 ± 6	62 ± 6

Results: Audio codecs

Table: Results for Berlin DB audio codecs

Codec	bit-rate	High–Low arousal **SSWT**	Pos.–Neg. valence **SSWT**	Fear-Type **SSWT**	All **SSWT**
Original	256	96 ± 6	82 ± 5	88 ± 7	64 ± 8
Down-sampled	128	95 ± 4	82 ± 6	85 ± 6	65 ± 7
AMR-NB	4.75	93 ± 4	81 ± 6	83 ± 8	63 ± 6
AMR-NB	7.95	95 ± 5	82 ± 5	84 ± 6	63 ± 5
GSM	12.2	94 ± 5	82 ± 6	82 ± 6	64 ± 7
AMR-WB	6.6	96 ± 4	82 ± 5	87 ± 7	61 ± 6
AMR-WB	23.85	96 ± 5	81 ± 5	85 ± 8	65 ± 10
G.722	64	96 ± 6	82 ± 4	87 ± 6	67 ± 8
G.726	16	94 ± 5	82 ± 5	84 ± 6	62 ± 7
SILK	64*	96 ± 6	82 ± 5	87 ± 7	63 ± 7
Opus	25*	96 ± 5	83 ± 5	87 ± 6	65 ± 6

Results: Audio codecs

Table: Results for Berlin DB audio codecs

Codec	bit-rate	High–Low arousal **SSWT**	Pos.–Neg. valence **SSWT**	Fear-Type **SSWT**	All **SSWT**
Original	256	96 ± 6	82 ± 5	88 ± 7	64 ± 8
Down-sampled	128	95 ± 4	82 ± 6	85 ± 6	65 ± 7
AMR-NB	4.75	93 ± 4	81 ± 6	83 ± 8	63 ± 6
AMR-NB	7.95	95 ± 5	82 ± 5	84 ± 6	63 ± 5
GSM	12.2	94 ± 5	82 ± 6	82 ± 6	64 ± 7
AMR-WB	6.6	96 ± 4	82 ± 5	87 ± 7	61 ± 6
AMR-WB	23.85	96 ± 5	81 ± 5	85 ± 8	65 ± 10
G.722	64	96 ± 6	82 ± 4	87 ± 6	67 ± 8
G.726	16	94 ± 5	82 ± 5	84 ± 6	62 ± 7
SILK	64*	96 ± 6	82 ± 5	87 ± 7	63 ± 7
Opus	25*	96 ± 5	83 ± 5	87 ± 6	65 ± 6

Outline

Conclusion I

- Features derived from acoustic, non–linear, and wavelet analysis were computed to characterize emotions from speech.
- The effect of different non–controlled acoustic conditions was tested.
- All feature sets are more suitable to recognize high vs. low arousal rather than positive vs. negative valence.
- Strong need to define new features more useful to classify emotions similar in arousal and different in valence.
- Further studies might be performed to improve the results for the recognition of multiple emotions.

Conclusion II

- Better results are obtained with features extracted from voiced segments relative to the obtained with features from unvoiced.
- The logMMSE technique seems to be useful to improve the results in some of the non–controlled acoustic conditions, while KLT has a negative impact in the system's performance.
- The effect of non–additive noise is not high, speech enhancement methods are not able to improve the results.
- The audio codecs do not have a high impact in the results, specially in detection of arousal and valence .
- Mobile telephone codecs decrease the results .
- Further studies might be performed to manage the effect of the mobile channels.

Academic Results I

- **J. C. Vásquez-Correa**, J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. Vargas-Bonilla, and E. Nöth. Non-linear dynamics characterization from wavelet packet transform for automatic recognition of emotional speech". *Smart Innovation, Systems and Technologies*, 48 pp. 199–207, 2016.

- **J. C. Vásquez-Correa**, J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. Vargas-Bonilla, L. D. Avendaño, and E. Nöth. "Time dependent ARMA for automatic recognition of fear-type emotions in speech". *Lecture Notes in Artificial Intelligence*, 9302, pp. 110–118, 2015.

- **J. C. Vásquez-Correa**, T. Arias-Vergara, J. R. Orozco-Arroyave, J. F. Vargas-Bonilla, J. D. Arias-Londoño and E. Nöth. "Automatic Detection of Parkinson's Disease from Continuous Speech Recorded in Non-Controlled Noise Conditions". *16th Anual conference of the international speech and communication association (INTERSPEECH)*, Dresden, 2015.

- **J. C. Vásquez-Correa**, N. García, J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. Vargas-Bonilla, and E. Nöth. "Emotion recognition from speech under environmental noise conditions using wavelet decomposition". *49th IEEE International Carnahan Conference on Security Technology (ICCST)*, Taipei, 2015.

Academic Results II

- N. García, **J. C. Vásquez-Correa**, J.F. Vargas-Bonilla, J.R. Orozco-Arroyave, J.D. Arias-Londoño. "Automatic Emotion Recognition in Compressed Speech Using Acoustic and Non-Linear Features". *20th Symposium of Image, Signal Processing, and Artificial Vision (STSIVA)*, Bogotá, 2015.

- **J. C. Vásquez-Correa**, N. García, J. F. Vargas-Bonilla, J. R. Orozco-Arroyave, J. D. Arias-Londoño, and O. L. Quintero-Montoya. "Evaluation of wavelet measures on automatic detection of emotion in noisy and telephony speech signals". *48th IEEE International Carnahan Conference on Security Technology (ICCST)*, Rome, 2014.

- **J. C. Vásquez-Correa**, J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. Vargas-Bonilla and E. Nöth. "New Computer Aided Device for Real Time Analysis of Speech of People with Parkinson's Disease". *Revista Facultad de Ingeniería Universidad de Antioquia*, N. 72 pp. 87-103, 2014.

- N. García, **J. C. Vásquez-Correa**, J.F. Vargas-Bonilla, J.R. Orozco-Arroyave, J.D. Arias-Londoño. "Evaluation of the effects of speech enhancement algorithms on the detection of fundamental frequency of speech". *19th Symposium of Image, Signal Processing, and Artificial Vision (STSIVA)*, Armenia, 2014.

Academic Results III

- Research Internship Pattern Recognition Lab. Friedrich–Alexander–Universität, Erlangen–Nürnberg, Germany.
  `https://www5.cs.fau.de/`.

- Research Internship Telefónica Research, Barcelona, Spain.
  `http://www.tid.es/`.

References I

Alam, M. J. et al. "Amplitude modulation features for emotion recognition from speech." In: *Annual conference of the international speech and communication association (INTERSPEECH)*. 2013, pp. 2420–2424.

Attabi, Y. and P. Dumouchel. "Anchor models for emotion recognition from speech". In: *IEEE Transactions on Affective Computing* 4.3 (2013), pp. 280–290.

Bänziger, T., M. Mortillaro, and K. R. Scherer. "Introducing the Geneva multimodal expression corpus for experimental research on emotion perception". In: *Emotion* 12.5 (2012), p. 1161.

Burkhardt, F. et al. "A database of German emotional speech". In: *Anual conferenece of the international speech and communication association (INTERSPEECH)* (2005), pp. 1517–1520.

Busso, C., M. Bulut, et al. "IEMOCAP: Interactive emotional dyadic motion capture database". In: *Language resources and evaluation* 42.4 (2008), pp. 335–359.

Busso, C., S. Lee, and S. Narayanan. "Analysis of emotionally salient aspects of fundamental frequency for emotion detection". In: *IEEE Transactions on Audio, Speech, and Language Processing* 17.4 (2009), pp. 582–596.

Cowie, R. et al. "Emotion recognition in human-computer interaction". In: *IEEE Signal Processing Magazine* 18.1 (2001), pp. 32–80.

Degaonkar, V. N and S. D. Apte. "Emotion modeling from speech signal based on wavelet packet transform". In: *International Journal of Speech Technology* 16.1 (2013), pp. 1–5.

References II

Deng, J. et al. "Autoencoder-based Unsupervised Domain Adaptation for Speech Emotion Recognition". In: *IEEE Signal Processing Letters* 21.9 (2014), pp. 1068–1072.

Ephraim, Y. and D. Malah. "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator". In: *IEEE Transactions on Acoustics, Speech and Signal Processing* 33.2 (1985), pp. 443–445.

Eyben, F., A. Batliner, and B. Schuller. "Towards a standard set of acoustic features for the processing of emotion in speech". In: *Proceedings of Meetings on Acoustics*. Vol. 9. 1. 2010, pp. 1–12.

Eyben, F., K. Scherer, et al. "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing". In: *IEEE Transactions on Affective Computing* (2015).

Eyben, Florian, Martin Wöllmer, and Björn Schuller. "OpenSmile: the munich versatile and fast open-source audio feature extractor". In: *18th ACM international conference on Multimedia*. ACM. 2010, pp. 1459–1462.

Haq, S. and P. J. B. Jackson. "Multimodal emotion recognition". In: *Machine audition: principles, algorithms and systems, IGI Global, Hershey* (2010), pp. 398–423.

Henríquez, P. et al. "Nonlinear dynamics characterization of emotional speech". In: *Neurocomputing* 132 (2014), pp. 126–135.

Hu, Y. and P. C. Loizou. "A generalized subspace approach for enhancing speech corrupted by colored noise". In: *IEEE Transactions on Speech and Audio Processing* 11.4 (2003), pp. 334–341.

References III

Huang, Y. et al. "Speech Emotion Recognition Based on Coiflet Wavelet Packet Cepstral Coefficients". In: *Pattern Recognition*. 2014, pp. 436–443.

Kim, Y., H. Lee, and E. M. Provost. "Deep learning for robust feature generation in audiovisual emotion recognition". In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. 2013, pp. 3687–3691.

Lee, C. C. et al. "Emotion recognition using a hierarchical binary decision tree approach". In: *Speech Communication* 53.9 (2011), pp. 1162–1171.

Li, L. et al. "Hybrid Deep Neural Network-Hidden Markov Model (DNN-HMM) Based Speech Emotion Recognition". In: *Humaine Association Conference on Affective Computing and Intelligent Interaction*. 2013, pp. 312–317.

Mariooryad, S. and C. Busso. "Compensating for speaker or lexical variabilities in speech for emotion recognition". In: *Speech Communication* 57 (2014),

Martin, O. et al. "The eNTERFACE'05 Audio-Visual Emotion Database". In: *Proceedings of International Conference on Data Engineering Workshops*. 2006, pp. 8–15.

McKeown, G. et al. "The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent". In: *IEEE Transactions on Affective Computing* 3.1 (2012), pp. 5–17.

Pohjalainen, J. and P. Alku. "Automatic detection of anger in telephone speech with robust auto-regressive modulation filtering". In: *Proceedigns of International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*. 2013, pp. 7537–7541.

References IV

Pohjalainen, J. and P. Alku. "Multi-scale modulation filtering in automatic detection of emotions in telephone speech". In: *Proceedings of International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*. 2014, pp. 980–984.

Ringeval, F. et al. "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions". In: *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*. 2013, pp. 1–8.

Schuller, B., A. Batliner, et al. "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge". In: *Speech Communication* 53.9 (2011), pp. 1062–1087.

Schuller, B., G. Rigoll, et al. "Effects of in-car noise-conditions on the recognition of emotion within speech". In: *Fortschritte der Akustik* 33.1 (2007), pp. 305–306.

Schuller, B., S. Steidl, and A. Batliner. "The INTERSPEECH 2009 emotion challenge". In: *Anual conferenece of the international speech and communication association (INTERSPEECH)*. 2009, pp. 312–315.

Sethu, V., E. Ambikairajah, and J. Epps. "On the use of speech parameter contours for emotion recognition". In: *EURASIP Journal on Audio, Speech, and Music Processing* 2013.1 (2013), pp. 1–14.

Steidl, S. *Automatic classification of emotion related user states in spontaneous children's speech*. University of Erlangen-Nuremberg Germany, 2009.

Stuhlsatz, A. et al. "Deep neural networks for acoustic emotion recognition: raising the benchmarks". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2011, pp. 5688–5691.

References V

Tawari, A. and M. Trivedi. "Speech emotion analysis in noisy real-world environment". In: *International Conference on Pattern Recognition*. 2010, pp. 4605–4608.

Xia, R. et al. "Modeling gender information for emotion recognition using Denoising autoencoder". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2014, pp. 990–994.

Zao, L., D. Cavalcante, and R. Coelho. "Time-frequency feature and AMS-GMM mask for acoustic emotion classification". In: *IEEE Signal Processing Letters* 21.5 (2014), pp. 620–624.

Zheng, W. et al. "A novel speech emotion recognition method via incomplete sparse least square regression". In: *IEEE Signal Processing Letters* 21.5 (2014), pp. 569–572.

Questions

Thanks!

jcamilo.vasquez@udea.edu.co

Appendix I

Table: Comparison of the results obtained with the state of the art

Source	# Feat.	Arousal	Valence	All
Berlin database				
F. Eyben, Batliner, and B. Schuller 2010	384	96.0%	80.0%	80.0%
Stuhlsatz et al. 2011	6552	97.4%	87.5%	81.9%
Li et al. 2013	42	-	-	77.9%
Zao, Cavalcante, and Coelho 2014	12 per frame	-	-	68.1%
F. Eyben, K. Scherer, et al. 2015	88	97.8%	86.7%	86.0%
openSMILE	384	97.3%	87.2%	80.4%
Acoustic+NLD	19 per frame	96.9%	82.9%	69.2%
WPT	128 per frame	95.7%	81.2%	66.1%
SSWT	22 per frame	95.8%	81.7%	69.3%
enterface05 database				
F. Eyben, Batliner, and B. Schuller 2010	384	76.0%	65.0%	68.0%
Stuhlsatz et al. 2011	6552	80.8%	79.7%	61.1%
Zheng et al. 2014	1582	-	-	69.3%
Li et al. 2013	42	-	-	53.9%
openSMILE	384	81.0%	81.4%	63.2%
Acoustic+NLD	19 per frame	80.2%	74.9%	49.0%
WPT	128 per frame	79.7%	75.9%	49.2%
SSWT	22 per frame	81.1%	75.6%	48.0%

Appendix II

Table: Comparison of the results obtained with the state of the art

Source	# Feat.	Arousal	Valence	All
IEMOCAP database				
C. C. Lee et al. 2011	384	-	-	56.3%
Mariooryad and Busso 2014	513	-	-	56.7%
Xia et al. 2014	1584	-	-	63.1%
openSMILE	384	75.5%	59.0%	57.2%
Acoustic+NLD	37 per frame	75.1%	59.5%	56.4%
WPT	128 per frame	75.4%	59.1%	57.1%
SSWT	22 per frame	77.2%	59.5%	58.2%
FAU-Aibo database				
C. C. Lee et al. 2011	384	-	-	39.9%
Deng et al. 2014	384	-	64.2%	-
Attabi and Dumouchel 2013	1584	-	-	44.2%
F. Eyben, K. Scherer, et al. 2015	88	-	76.5%	43.1%
openSMILE	384	-	62.0%	32.5%
Acoustic+NLD	19 per frame	-	69.6%	38.9%
WPT	128 per frame	-	68.2%	38.0%
SSWT	22 per frame	-	70.3%	32.6%

Apendix III

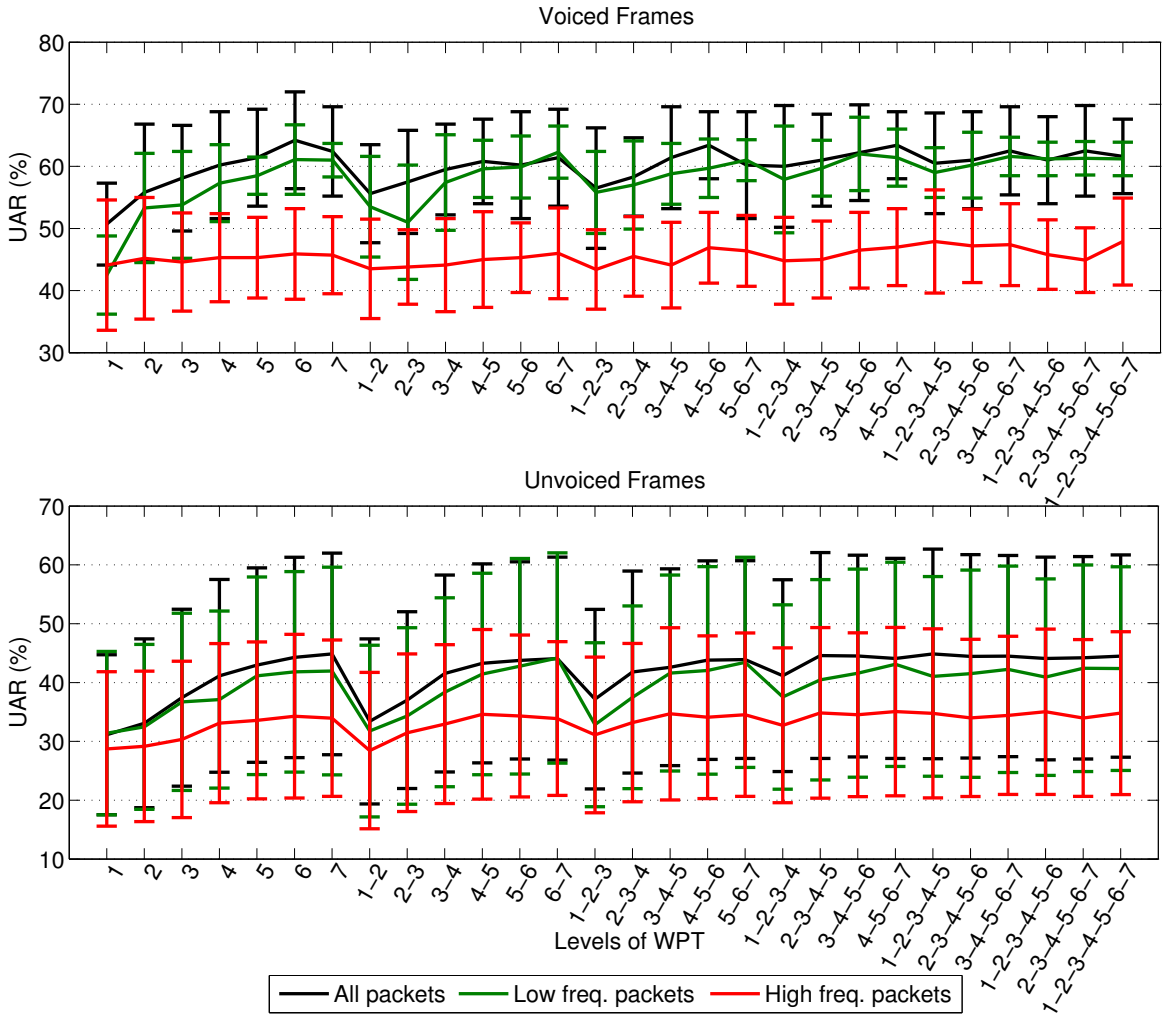
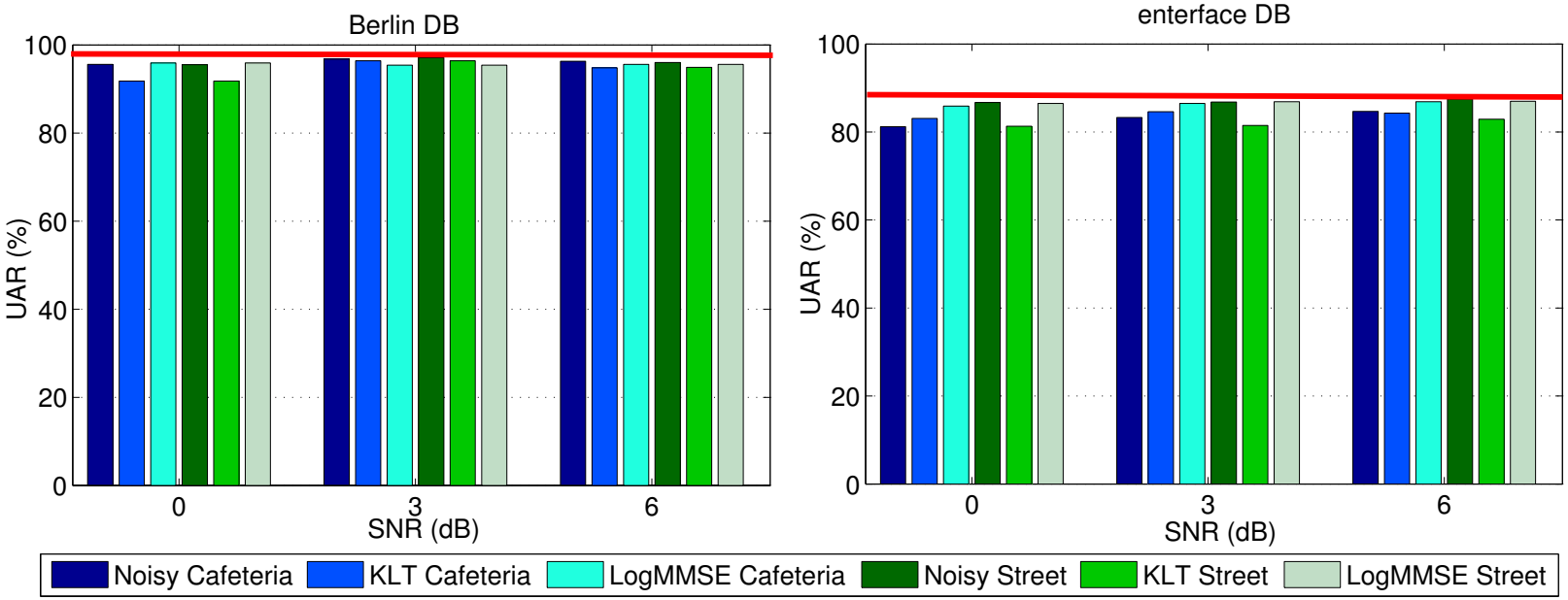


Figure: Frequency band selection in WPT

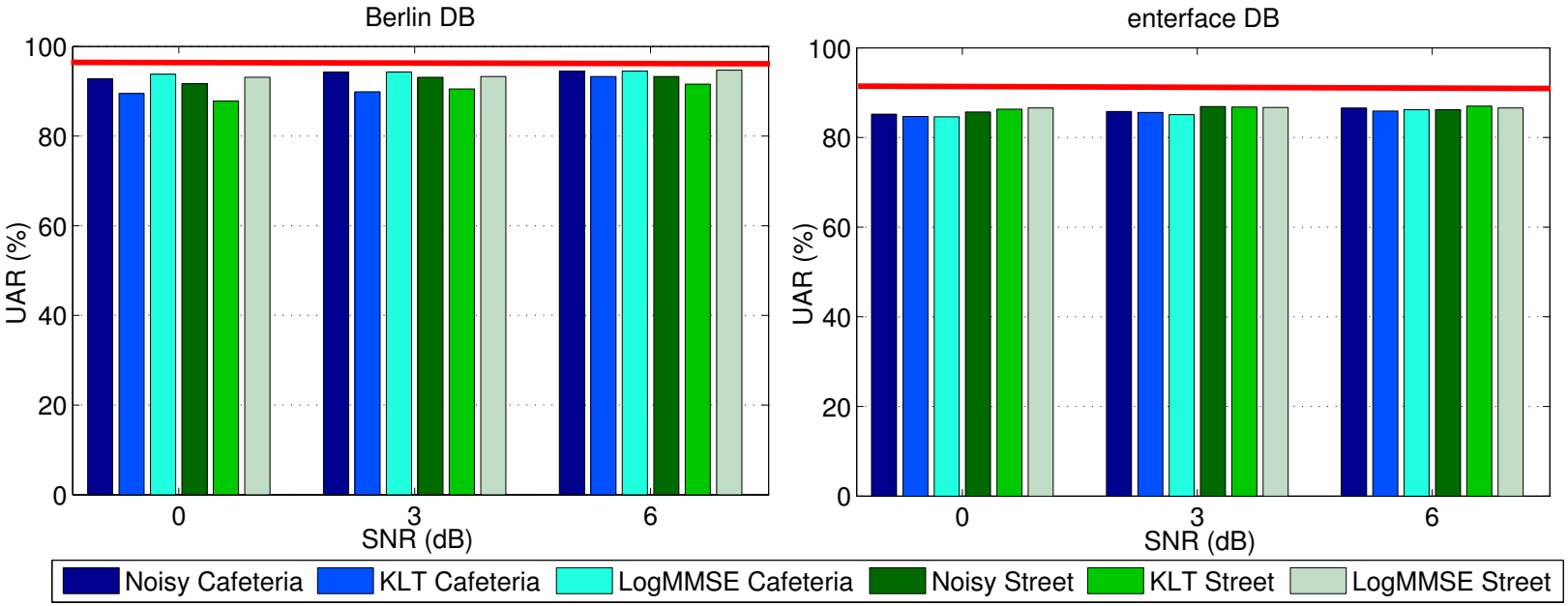
Results: Noisy recordings, High vs. Low arousal detection: OpenSMILE

Red line: results for Original signals.



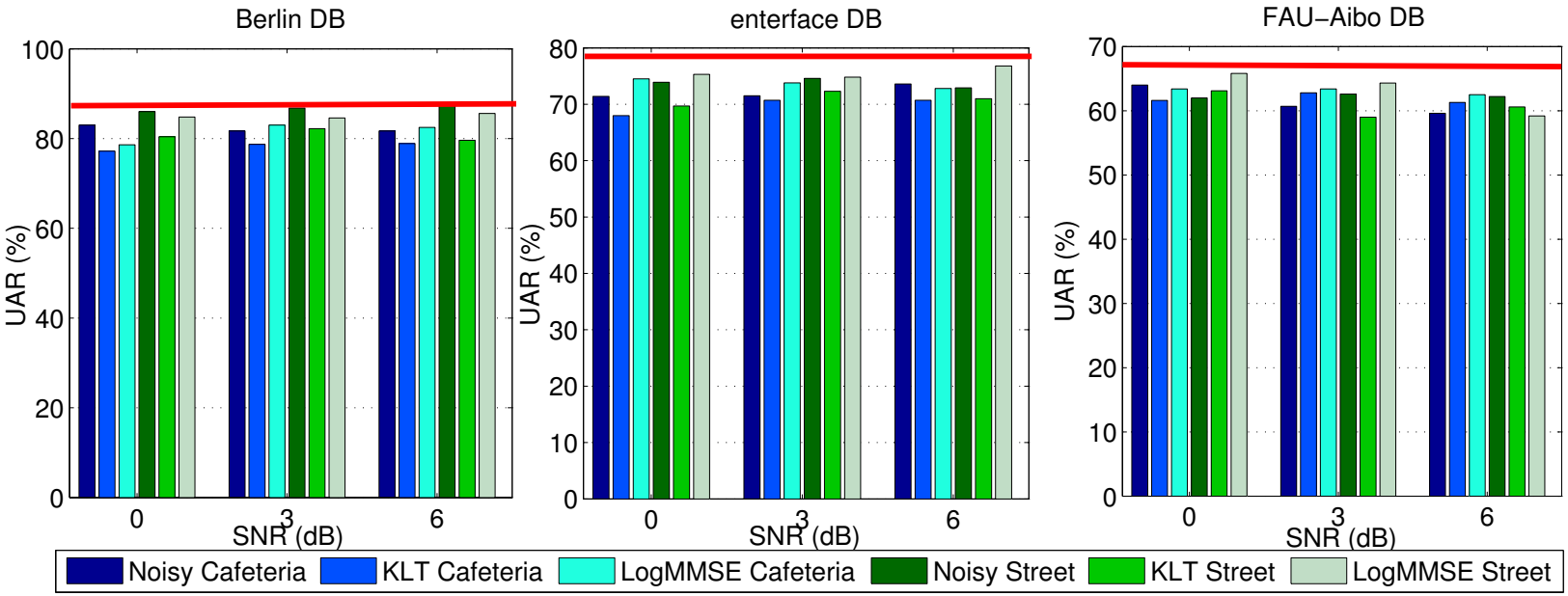
Results: Noisy recordings, High vs. Low arousal detection: SSWT

Red line: results for Original signals.



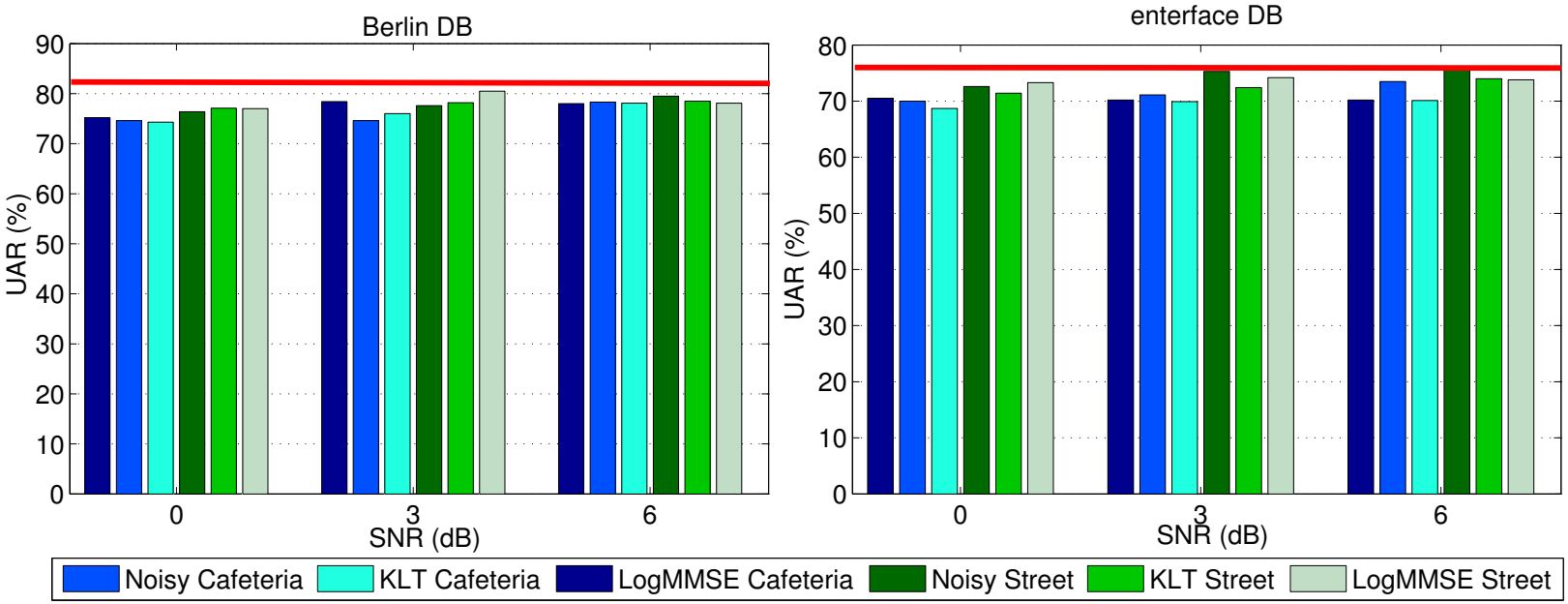
Results: Noisy recordings, positive vs. negative valence detection: OpenSMILE

Red line: results for Original signals.



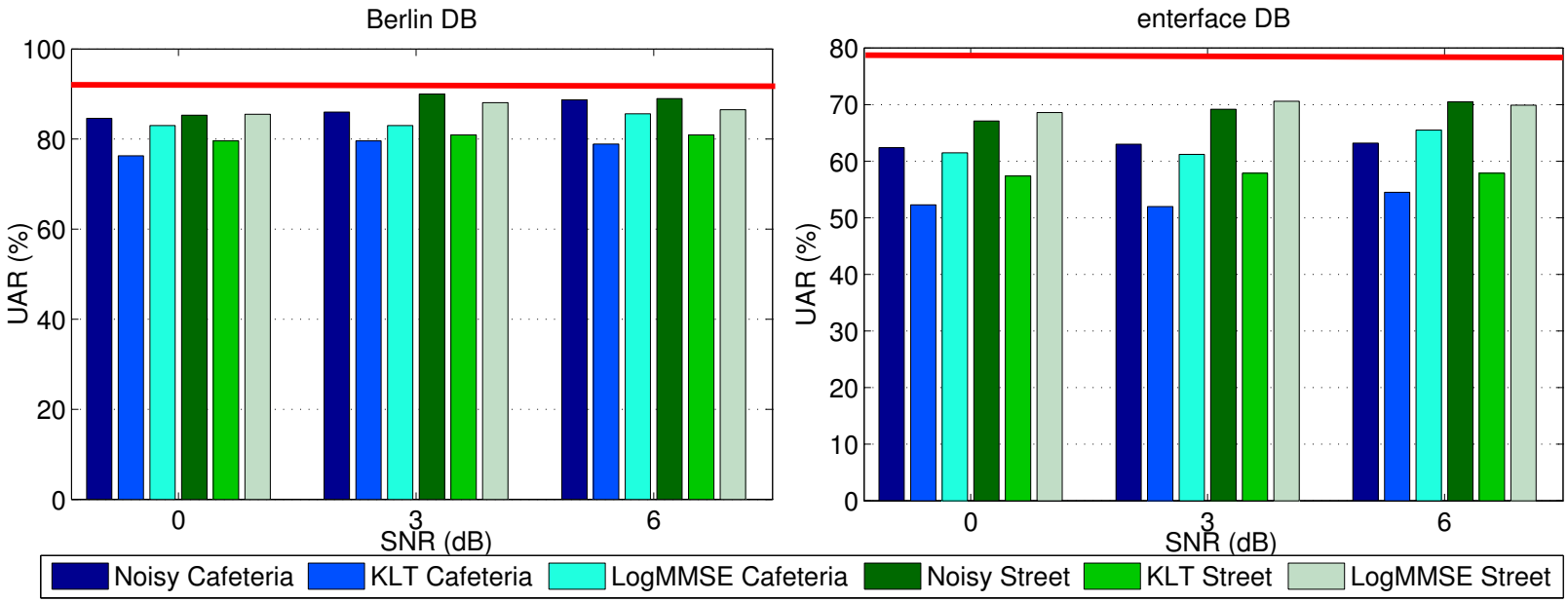
Results: Noisy recordings, positive vs. negative valence detection: SSWT

Red line: results for Original signals.



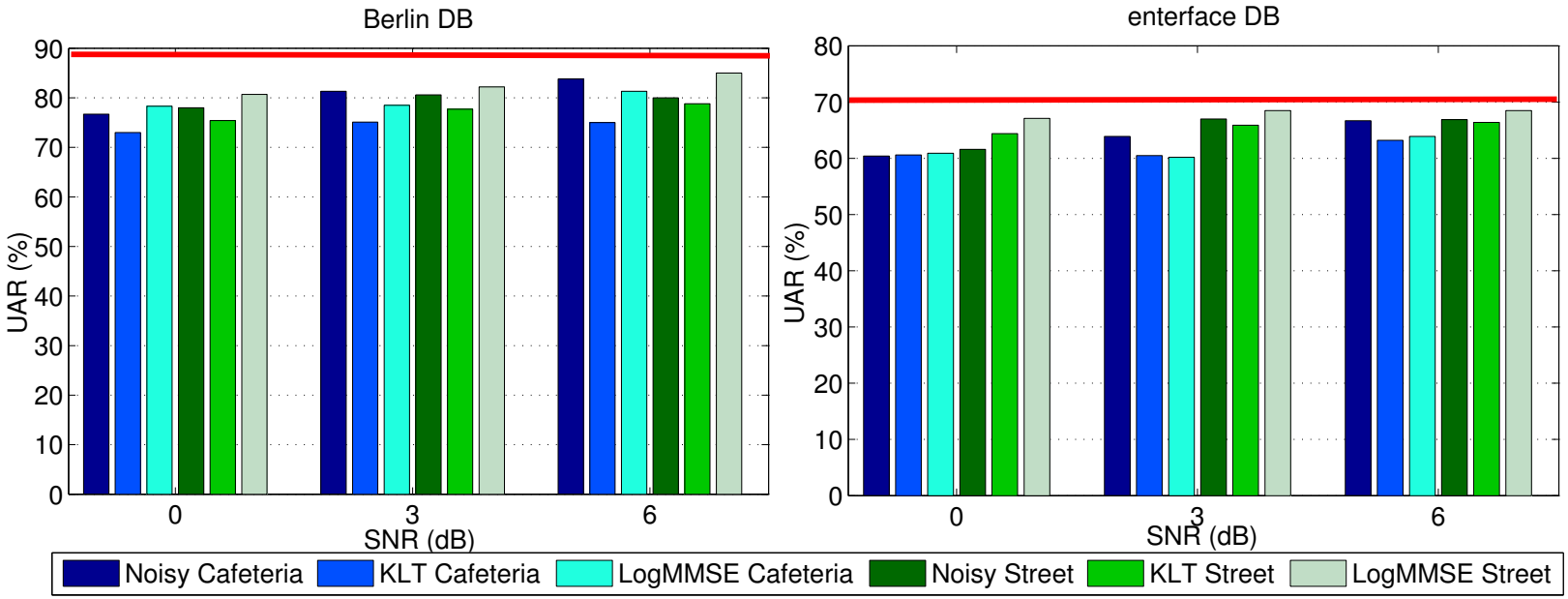
Results: Noisy recordings, Fear–type emotion: OpenSMILE

Red line: results for Original signals.



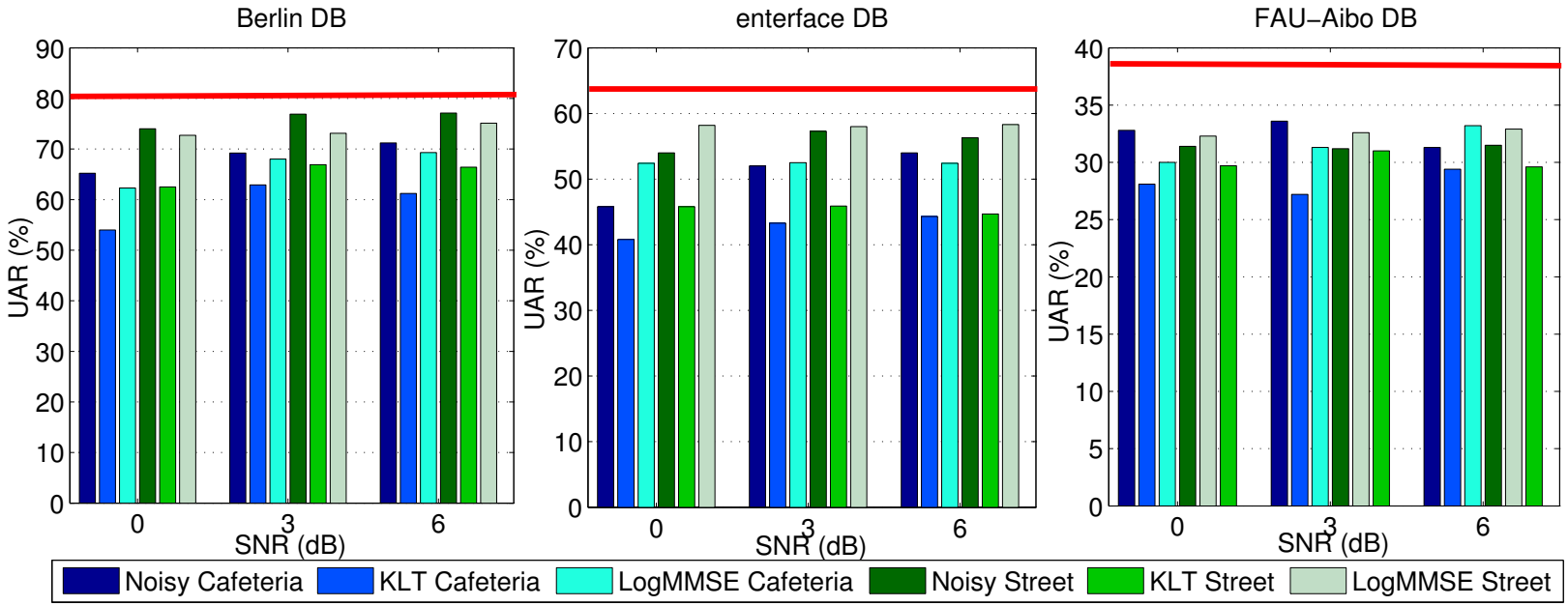
Results: Noisy recordings, Fear–type emotion: SSWT

Red line: results for Original signals.



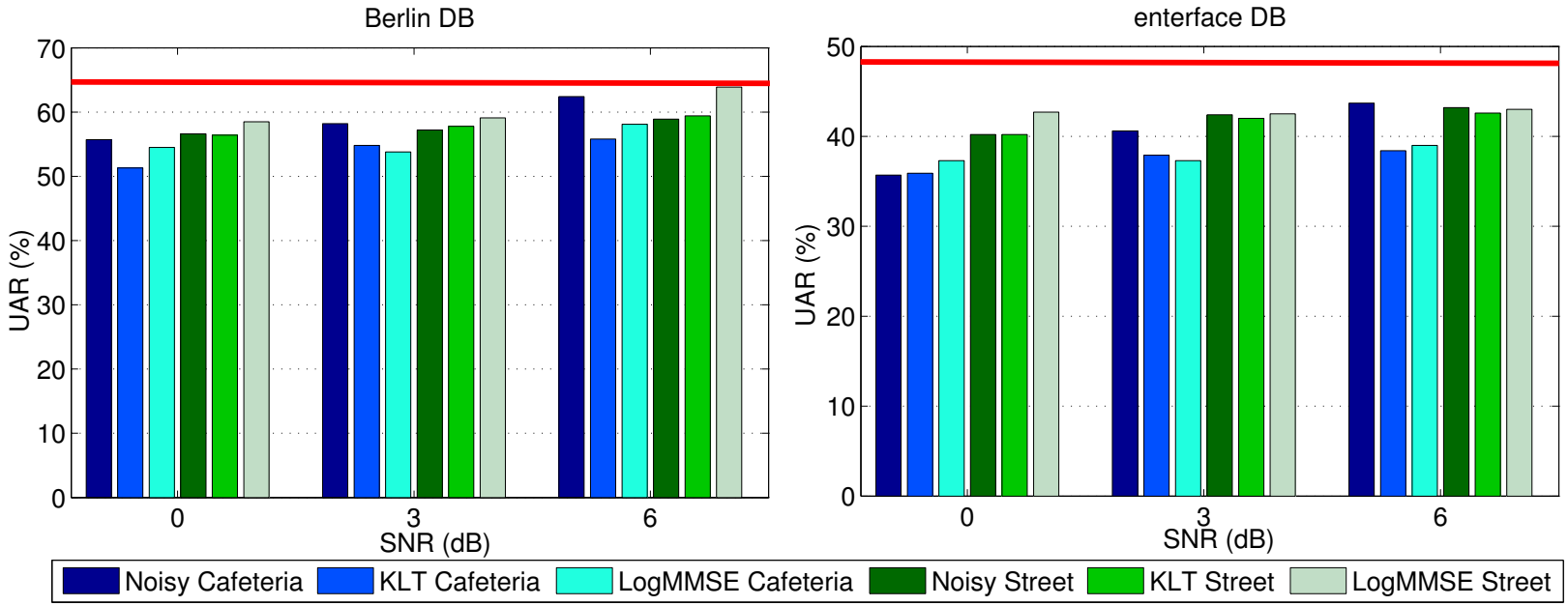
Results: Noisy recordings, recognition of multiple emotions: OpenSMILE

Red line: results for Original signals.



Results: Noisy recordings, recognition of multiple emotions: SSWT

Red line: results for Original signals.



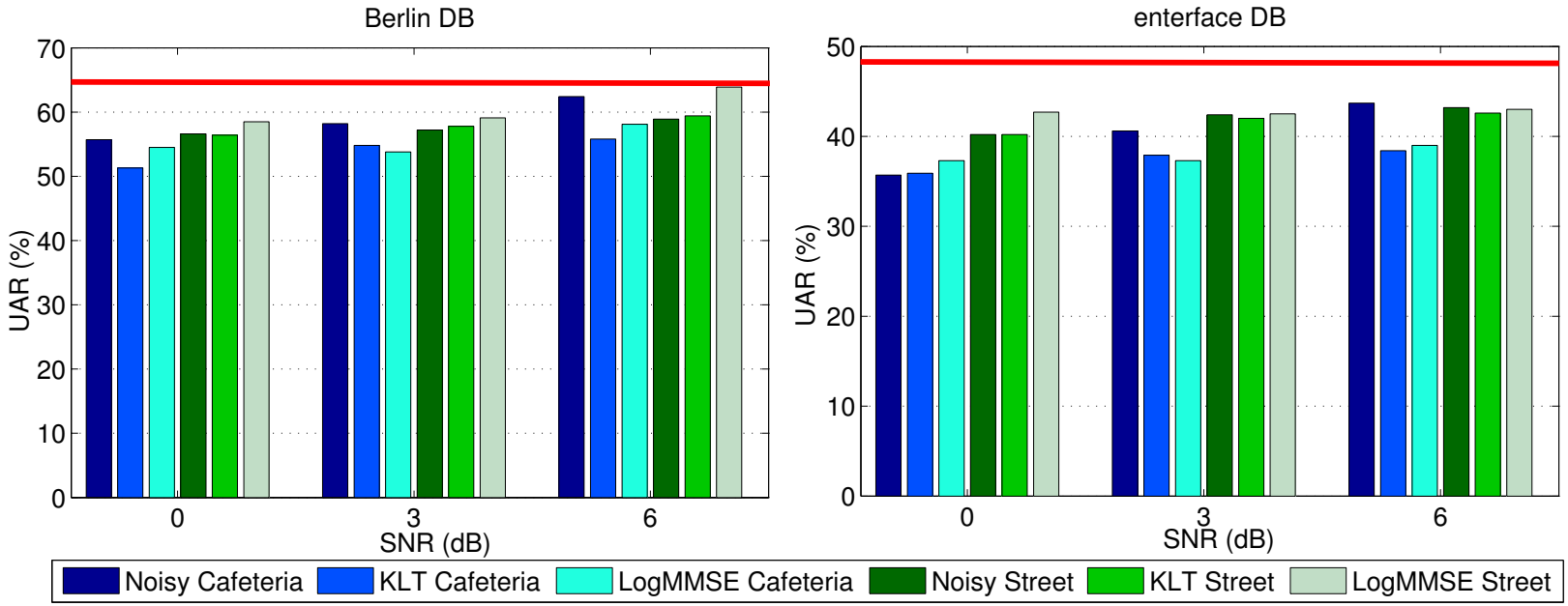
Results: Audio codecs

Table: Results for Berlin DB audio codecs

		High–Low arousal		Pos.–Neg. valence		Fear-Type		All	
Codec	bit-rate	**openSMILE**	**SSWT**	**openSMILE**	**SSWT**	**openSMILE**	**SSWT**	**openSMILE**	**SSWT**
Original	256	97 ± 3	96 ± 6	87 ± 2	82 ± 5	91 ± 5	88 ± 7	80 ± 8	64 ± 8
Down-sampled	128	95 ± 4	95 ± 4	83 ± 5	82 ± 6	82 ± 6	85 ± 6	74 ± 6	65 ± 7
AMR-NB	4.75	96 ± 4	93 ± 4	83 ± 7	81 ± 6	83 ± 6	83 ± 8	75 ± 6	63 ± 6
AMR-NB	7.95	95 ± 4	95 ± 5	81 ± 5	82 ± 5	85 ± 6	84 ± 6	74 ± 5	63 ± 5
GSM	12.2	97 ± 4	94 ± 5	81 ± 5	82 ± 6	82 ± 5	82 ± 6	74 ± 6	64 ± 7
AMR-WB	6.6	97 ± 5	96 ± 4	86 ± 4	82 ± 5	84 ± 8	87 ± 7	79 ± 7	61 ± 6
AMR-WB	23.85	97 ± 6	96 ± 5	87 ± 5	81 ± 5	87 ± 7	85 ± 8	78 ± 6	65 ± 10
G.722	64	97 ± 6	96 ± 6	82 ± 4	82 ± 4	88 ± 4	87 ± 6	80 ± 7	67 ± 8
G.726	16	97 ± 3	94 ± 5	82 ± 5	82 ± 5	82 ± 5	84 ± 6	76 ± 8	62 ± 7
SILK	64*	97 ± 7	96 ± 6	84 ± 3	82 ± 5	88 ± 7	87 ± 7	77 ± 6	63 ± 7
Opus	25*	99 ± 3	96 ± 5	85 ± 5	83 ± 5	90 ± 6	87 ± 6	77 ± 5	65 ± 6

Results: Noisy recordings, recognition of multiple emotions: SSWT

Red line: results for Original signals.



Results: Non–additive noise

Table: Results for Berlin DB re-captured in noisy environments

	High–Low arousal		Pos.–Neg. valence		Fear-Type		All	
Recordings	**openSMILE**	**SSWT**	**openSMILE**	**SSWT**	**openSMILE**	**SSWT**	**openSMILE**	**SSWT**
Street Noise	95.8 ± 6.1	95.5 ± 6.2	86.0 ± 2.5	82.3 ± 4.5	91.6 ± 5.3	85.6 ± 9.1	78.1 ± 4.4	63.3 ± 4.0
Office Noise	96.4 ± 4.0	96.7 ± 4.0	88.0 ± 3.3	81.3 ± 5.7	91.3 ± 3.7	87.2 ± 9.2	76.5 ± 6.3	65.4 ± 7.5
KLT Street	95.1 ± 5.0	96.4 ± 5.6	85.3 ± 4.2	81.8 ± 5.3	85.4 ± 6.4	85.5 ± 9.1	75.5 ± 7.9	63.7 ± 6.9
KLT Office	95.6 ± 5.1	96.7 ± 4.5	86.5 ± 4.0	81.6 ± 3.3	89.9 ± 5.3	85.2 ± 7.9	73.7 ± 7.0	63.5 ± 7.7
logMMSE Street	95.7 ± 4.0	96.2 ± 4.6	86.3 ± 5.0	81.3 ± 6.7	86.9 ± 5.7	82.8 ± 10.3	75.8 ± 6.0	59.9 ± 6.2
logMMSE Office	96.1 ± 3.7	96.0 ± 3.3	82.6 ± 4.2	81.8 ± 5.7	87.1 ± 7.4	84.3 ± 6.0	74.9 ± 5.7	61.9 ± 6.4
Original	97.3 ± 3.0	95.8 ± 5.5	87.2 ± 2.4	81.7 ± 4.6	91.4 ± 5.0	88.3 ± 7.0	80.4 ± 8.0	64.0 ± 8.0