Emotion Recognition in Speech under Environmental Noise Conditions using Wavelet Decomposition

J.C. Vásquez-Correa[1], N. García[1], J.R. Orozco-Arroyave[1,2], J.D. Arias-Londoño[1], J.F. Vargas-Bonilla[1], Elmar Nöth[2]

[1]Faculty of Engineering, University of Antioquia UdeA, Medellín, Colombia.
[2]Pattern Recognition Lab., Friedrich-Alexander-Universität, Erlangen-Nürnberg, Germany

jesus.vargas@udea.edu.co

UNIVERSIDAD DE ANTIOQUIA 1803

SISTEMIC Sistemas Embebidos e Inteligencia Computacional

Introduction: Emotion recognition

Recognition of emotion in speech:

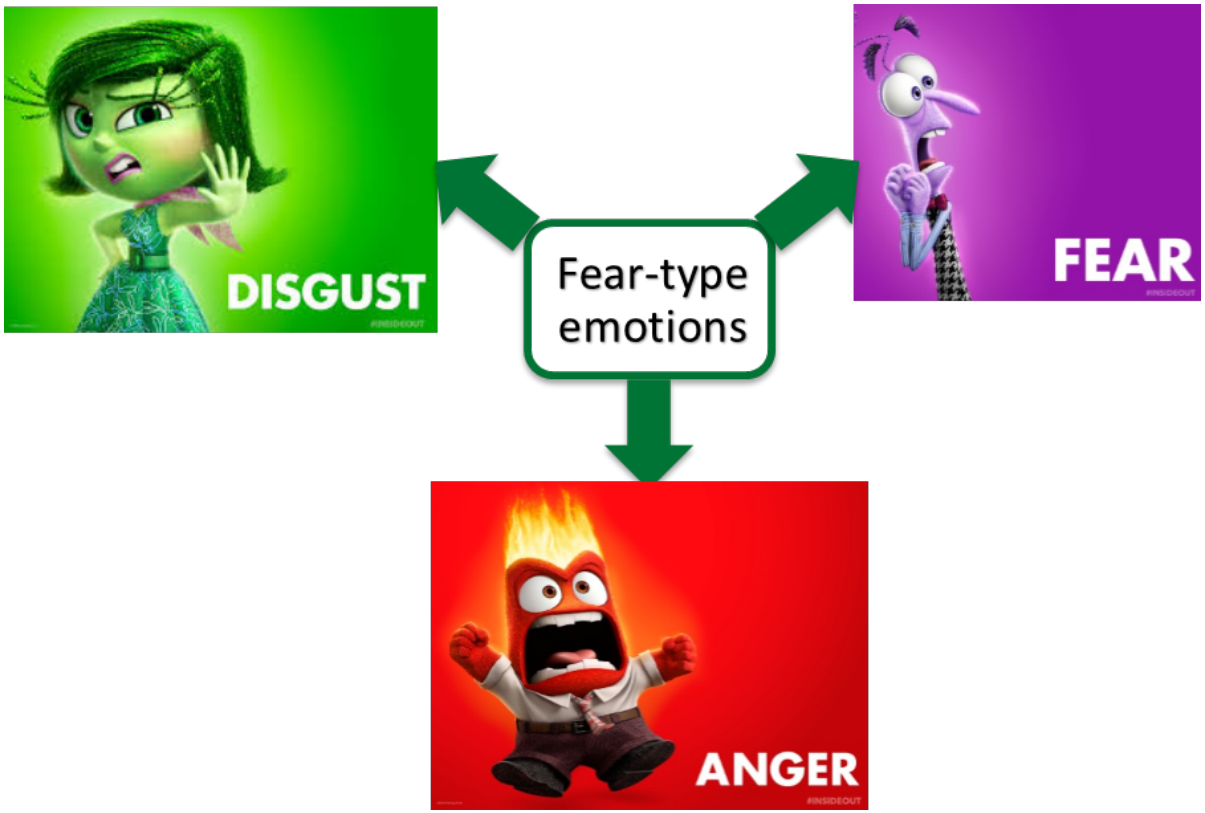
- Call centers
- Emergency services
- Psychologic therapy
- Intelligent vehicles
- Public surveillance







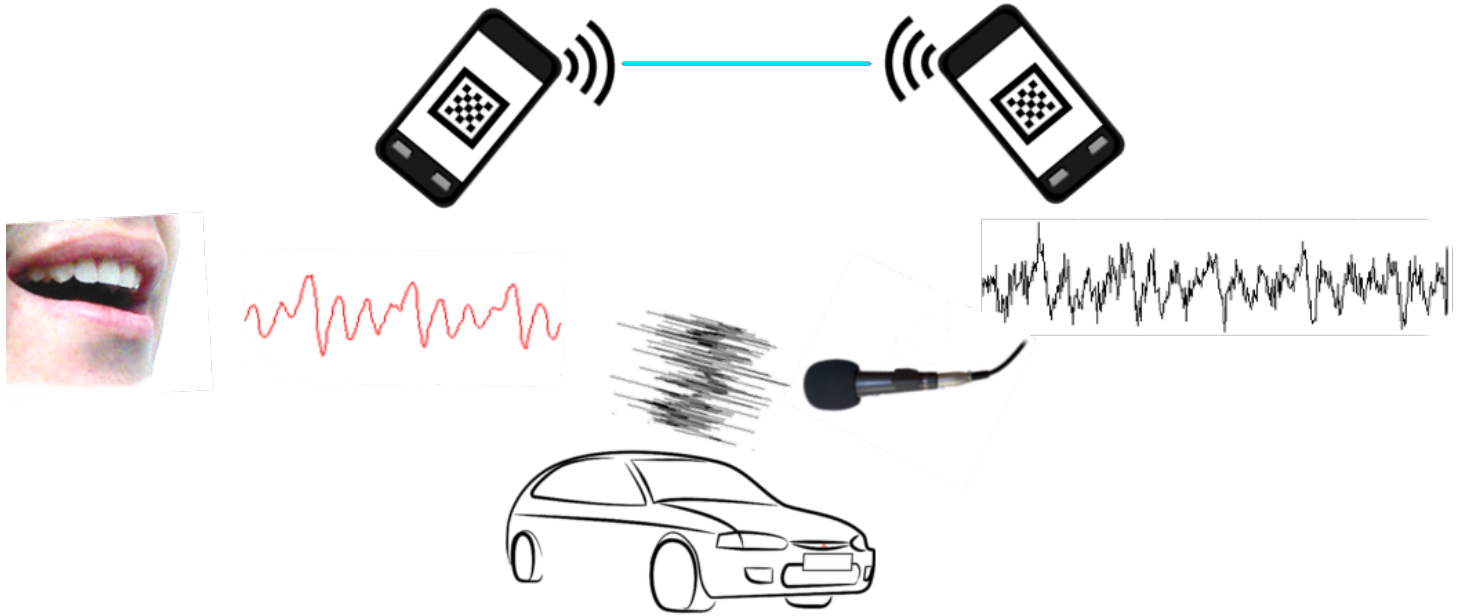

Introduction: Fear-type emotions



Introduction: Challenges

UNIVERSIDAD DE ANTIOQUIA

- Naturalness of databases (Acted, Natural, Evoked)
- Large set of features
- Acoustic conditions (Telephone, Background noise)



Introduction: Previous Work (2-class)

UNIVERSIDAD DE ANTIOQUIA

- Emotion recognition under AWGN noise
- Emotion recognition under GSM and wired-line telephone channel

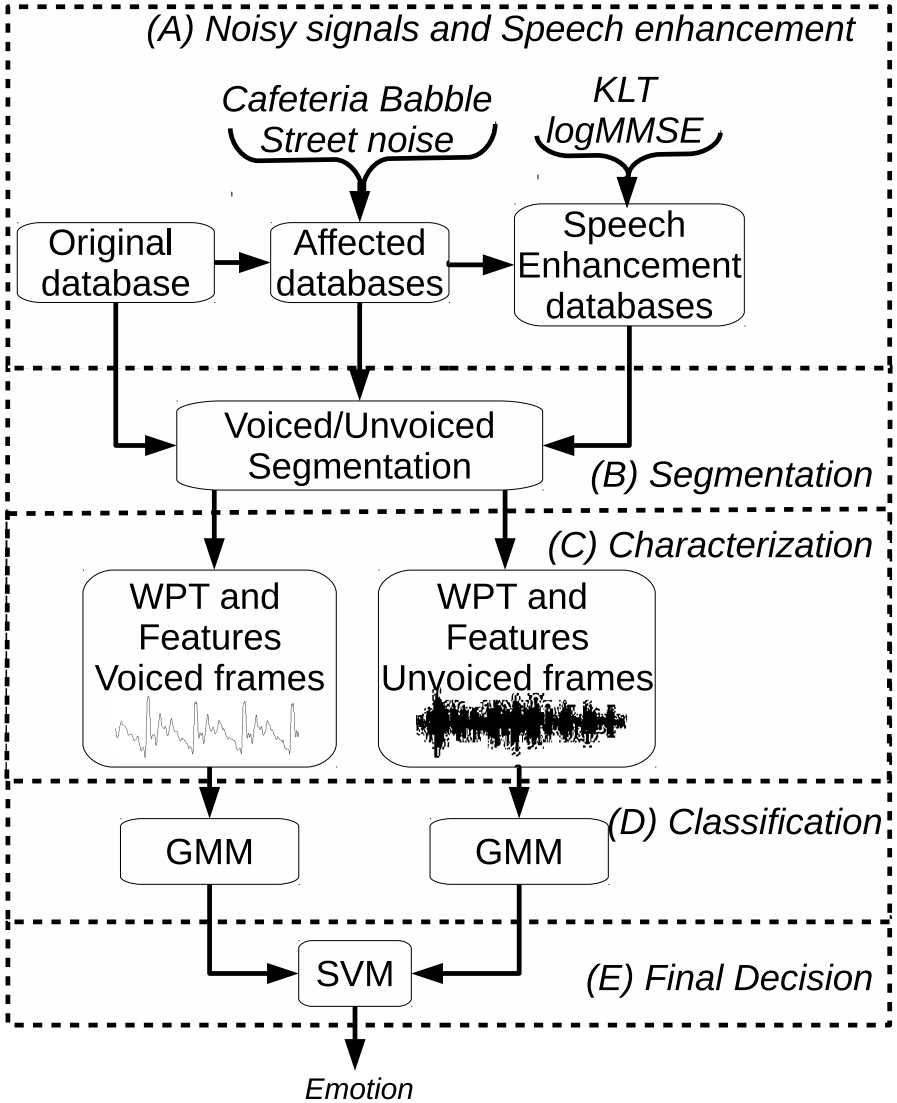
Condition	**Original**	**Affected**	**KLT**	**logMMSE**
AWGN SNR=3dB	76.9%	71.3%	78.1%	74.7%
AWGN SNR=10dB	76.9%	74.7%	80.1%	76.7%
GSM channel	76.9%	77.8%	62,9%	70.6%
wired-line	76.9%	65.2%	59.0%	75.1%

Table: Emotion recognition Berlin database

Methodology

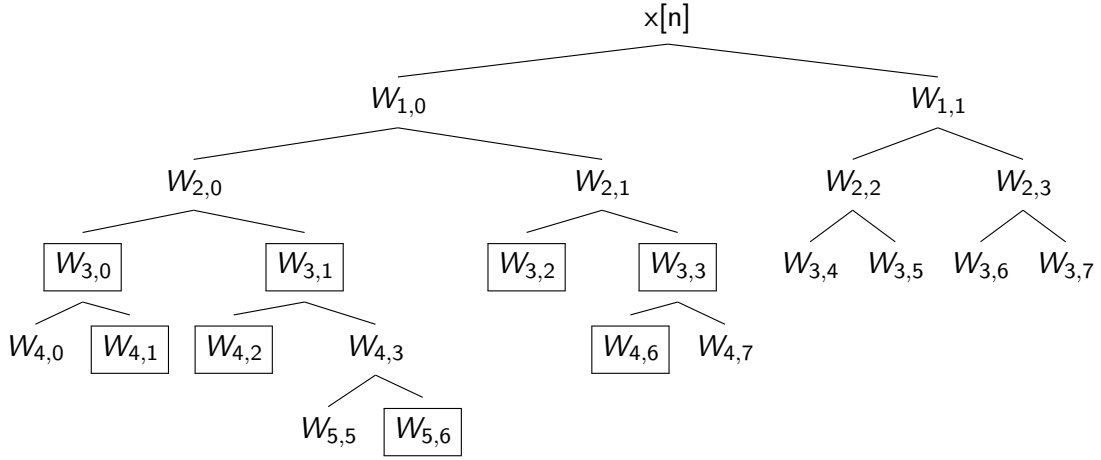
A new characterization approach based on wavelet packet transform for recognition of emotions in speech evaluated in non-controlled noise conditions.

- Log-energy
- Log-energy entropy
- MFCC
- Lempel-Ziv complexity

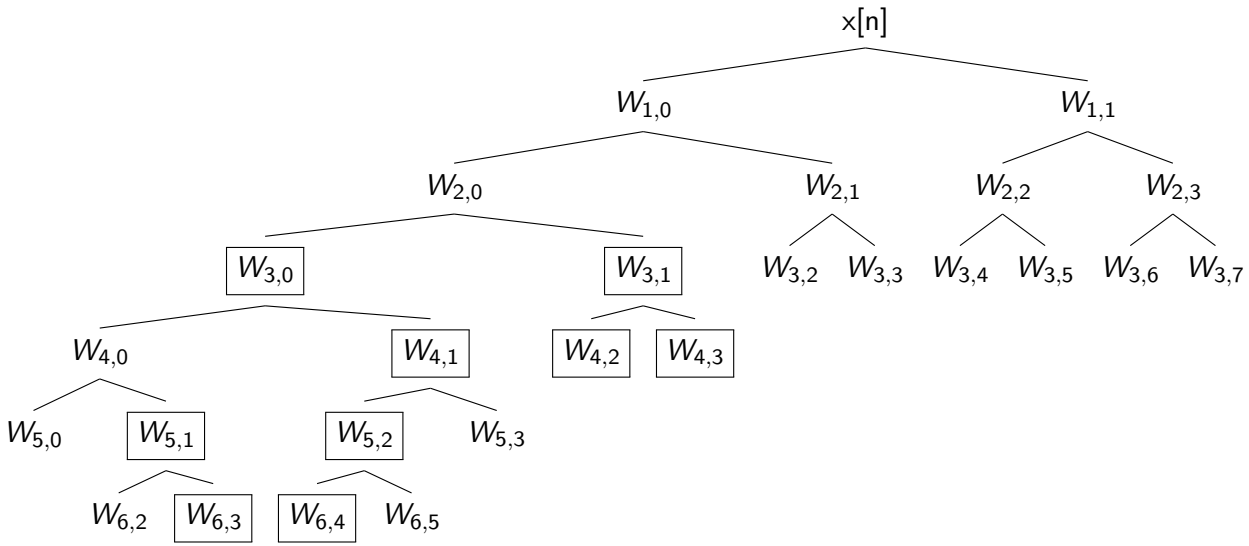


Methodology: Characterization

Wavelet decomposition Voiced segments



Wavelet decomposition Unvoiced segments



Databases

UNIVERSIDAD DE ANTIOQUIA

database	# recordings	Speakers	Fs (Hz)	Naturalness	Emotions
Berlin	534	12	16000	Acted	Hot anger Boredorm Disgust Anxiety/Fear Happiness Sadness Neutral
Enterface05 (Audio-Video)	1317	44	44100	Evoked	Hot anger Happiness Disgust Anxiety/Fear Sadness Surprise

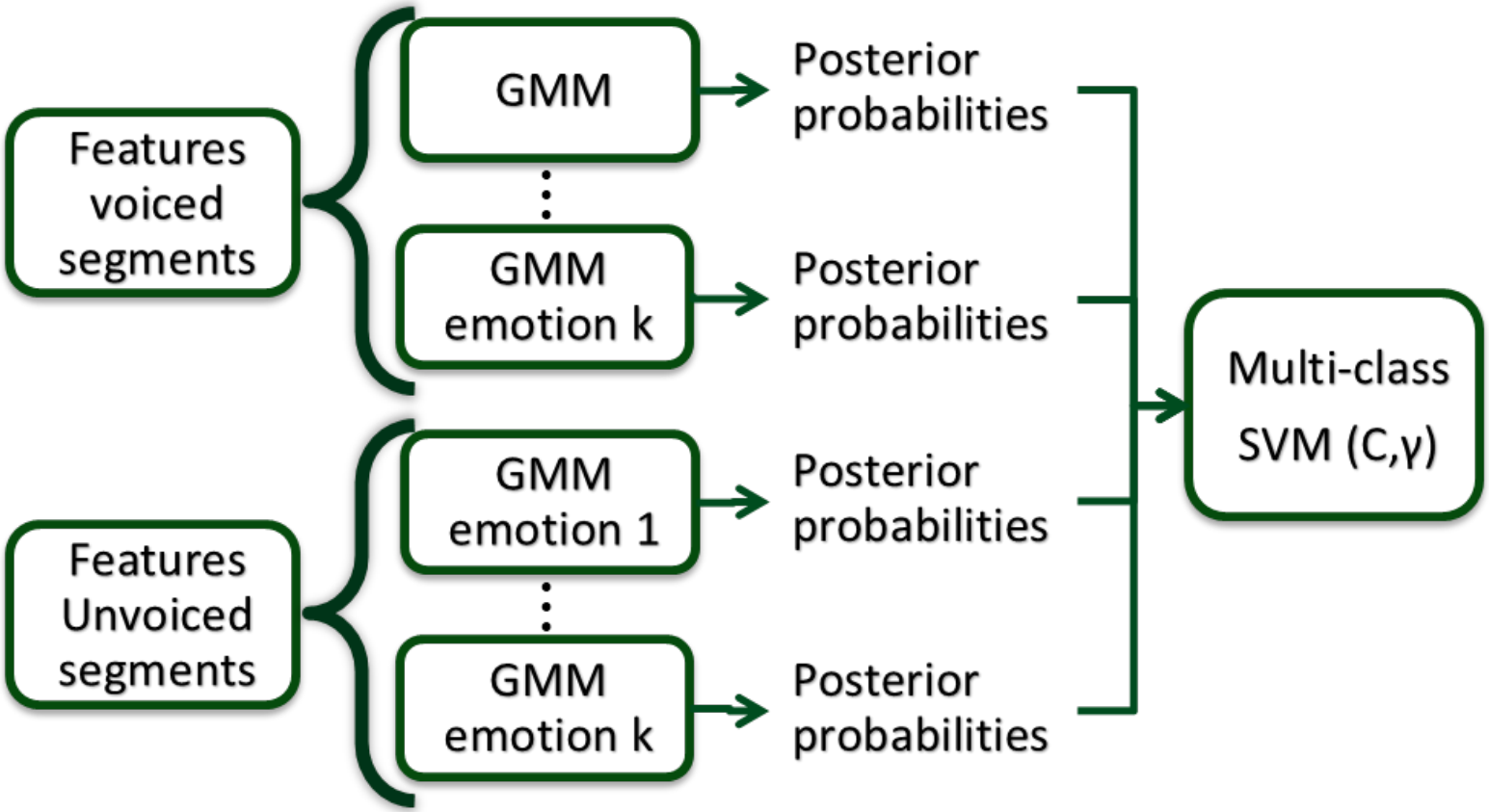
Experiments

UNIVERSIDAD DE ANTIOQUIA

Experiment	**Berlin DB**	**enterface05 DB**
Multi-class	Anger Disgust Fear Neutral	Anger Disgust Fear
2-class	(Anger, disgust, fear) vs Neutral	(Anger, disgust, fear, sadness) vs (Happiness, Surprise)

Table: Experiments performed

Methodology: Classification



Results: Original signals

UNIVERSIDAD DE ANTIOQUIA

Segments	feat.	Class. task	Berlin DB	enterface05 DB
Voiced	120	multi-class	80.0 ± 11.6	57.7 ± 6.8
		2-class	89.9 ± 7.8	65.1 ± 4.6
Unvoiced	120	multi-class	62.5 ± 5.0	55.4 ± 6.8
		2-class	82.5 ± 8.6	64.6 ± 6.0
Fusion		multi-class	74.7 ± 11.9	61.6 ± 4.5
		2-class	94.6 ± 5.1	69.2 ± 1.5
all signal openEAR [Eyben2012]	384	multi-class	84.3 ± 6.6	66.6 ± 4.2
		2-class	94.9 ± 4.1	68.6 ± 4.8

Table: Accuracy for original non-affected speech signals

Results: Original signals

UNIVERSIDAD DE ANTIOQUIA

Segments	feat.	Class. task	Berlin DB	enterface05 DB
Voiced	120	multi-class	80.0 ± 11.6	57.7 ± 6.8
		2-class	89.9 ± 7.8	65.1 ± 4.6
Unvoiced	120	multi-class	62.5 ± 5.0	55.4 ± 6.8
		2-class	82.5 ± 8.6	64.6 ± 6.0
Fusion		multi-class	74.7 ± 11.9	61.6 ± 4.5
		2-class	94.6 ± 5.1	69.2 ± 1.5
all signal openEAR [Eyben2012]	384	multi-class	84.3 ± 6.6	66.6 ± 4.2
		2-class	94.9 ± 4.1	68.6 ± 4.8

Table: Accuracy for original non-affected speech signals

Results: Original signals

UNIVERSIDAD DE ANTIOQUIA

Segments	feat.	Class. task	Berlin DB	enterface05 DB
Voiced	120	multi-class	80.0 ± 11.6	57.7 ± 6.8
		2-class	89.9 ± 7.8	65.1 ± 4.6
Unvoiced	120	multi-class	62.5 ± 5.0	55.4 ± 6.8
		2-class	82.5 ± 8.6	64.6 ± 6.0
Fusion		multi-class	74.7 ± 11.9	61.6 ± 4.5
		2-class	94.6 ± 5.1	69.2 ± 1.5
all signal openEAR [Eyben2012]	384	multi-class	84.3 ± 6.6	66.6 ± 4.2
		2-class	94.9 ± 4.1	68.6 ± 4.8

Table: Accuracy for original non-affected speech signals.
Previous Work: 76.9%

Results: Original signals

UNIVERSIDAD DE ANTIOQUIA

Segments	feat.	Class. task	Berlin DB	enterface05 DB
Voiced	120	multi-class	80.0 ± 11.6	57.7 ± 6.8
		2-class	89.9 ± 7.8	65.1 ± 4.6
Unvoiced	120	multi-class	62.5 ± 5.0	55.4 ± 6.8
		2-class	82.5 ± 8.6	64.6 ± 6.0
Fusion		multi-class	74.7 ± 11.9	61.6 ± 4.5
		2-class	94.6 ± 5.1	69.2 ± 1.5
all signal openEAR [Eyben2012]	384	multi-class	84.3 ± 6.6	66.6 ± 4.2
		2-class	94.9 ± 4.1	68.6 ± 4.8

Table: Accuracy for original non-affected speech signals

Results: Original signals

UNIVERSIDAD DE ANTIOQUIA

Segments	feat.	Class. task	Berlin DB	enterface05 DB
Voiced	120	multi-class	80.0 ± 11.6	57.7 ± 6.8
		2-class	89.9 ± 7.8	65.1 ± 4.6
Unvoiced	120	multi-class	62.5 ± 5.0	55.4 ± 6.8
		2-class	82.5 ± 8.6	64.6 ± 6.0
Fusion		multi-class	74.7 ± 11.9	61.6 ± 4.5
		2-class	94.6 ± 5.1	69.2 ± 1.5
all signal openEAR [Eyben2012]	384	multi-class	84.3 ± 6.6	66.6 ± 4.2
		2-class	94.9 ± 4.1	68.6 ± 4.8

Table: Accuracy for original non-affected speech signals

Experiments: Environments

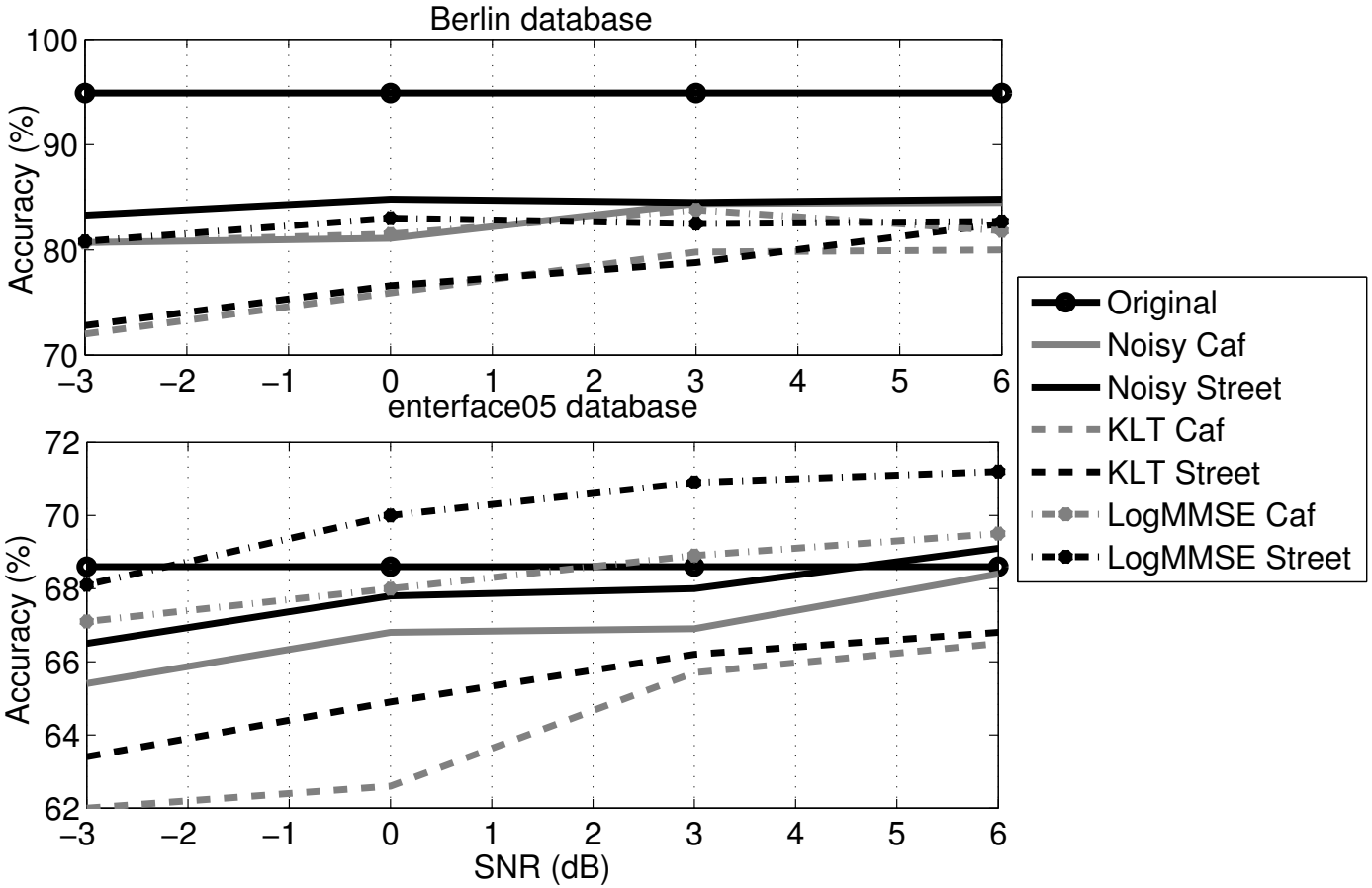
UNIVERSIDAD DE ANTIOQUIA

- Original non-affected speech signals
- Cafeteria babble noise
- Street noise

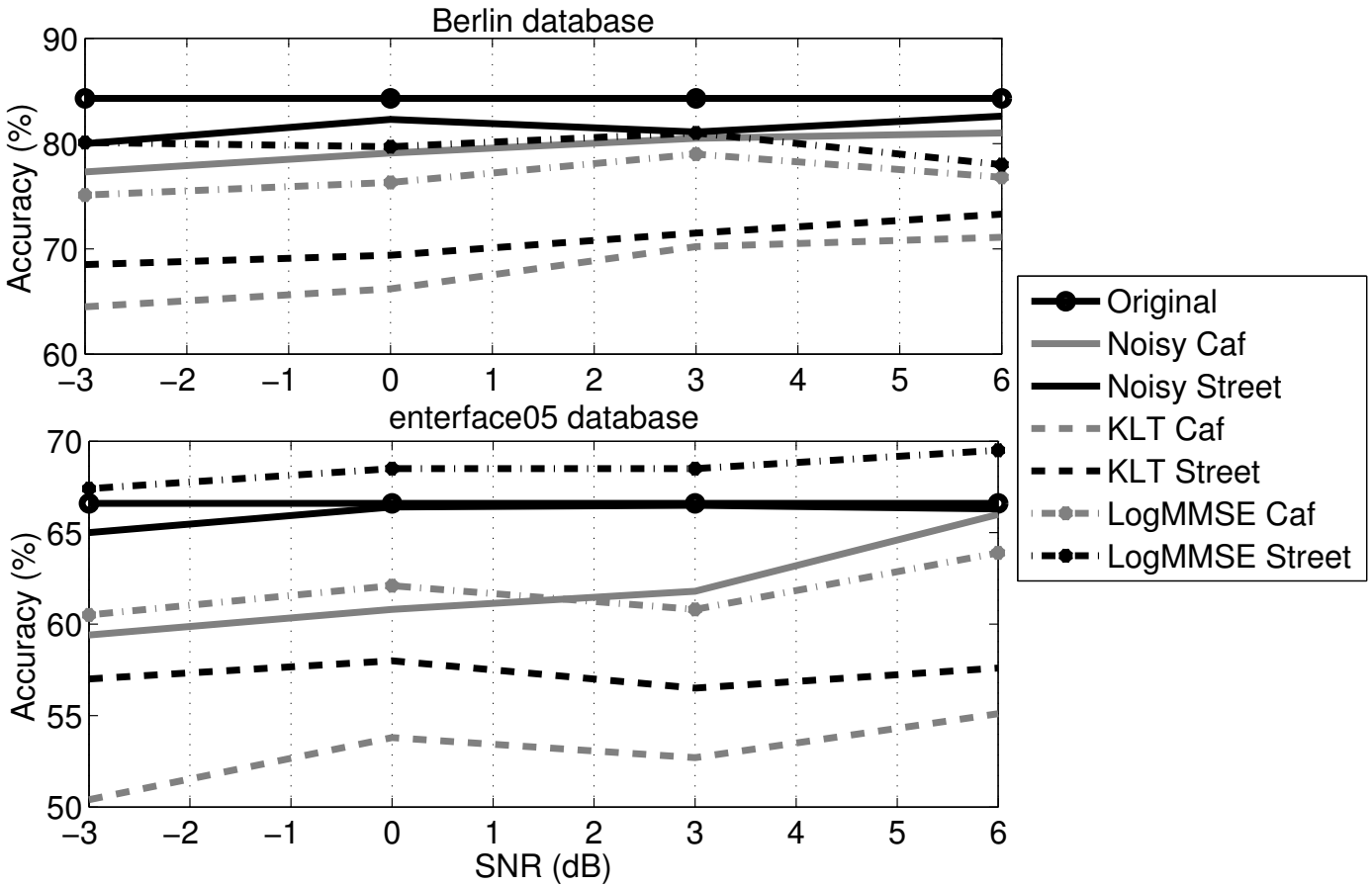
- KLT algorithm
- LogMMSE algorithm

SNR evaluated ranges from -3dB to 6dB

Results: Affected signals, 2-class (OpenEAR)



Results: Affected signals, M-class (OpenEAR)



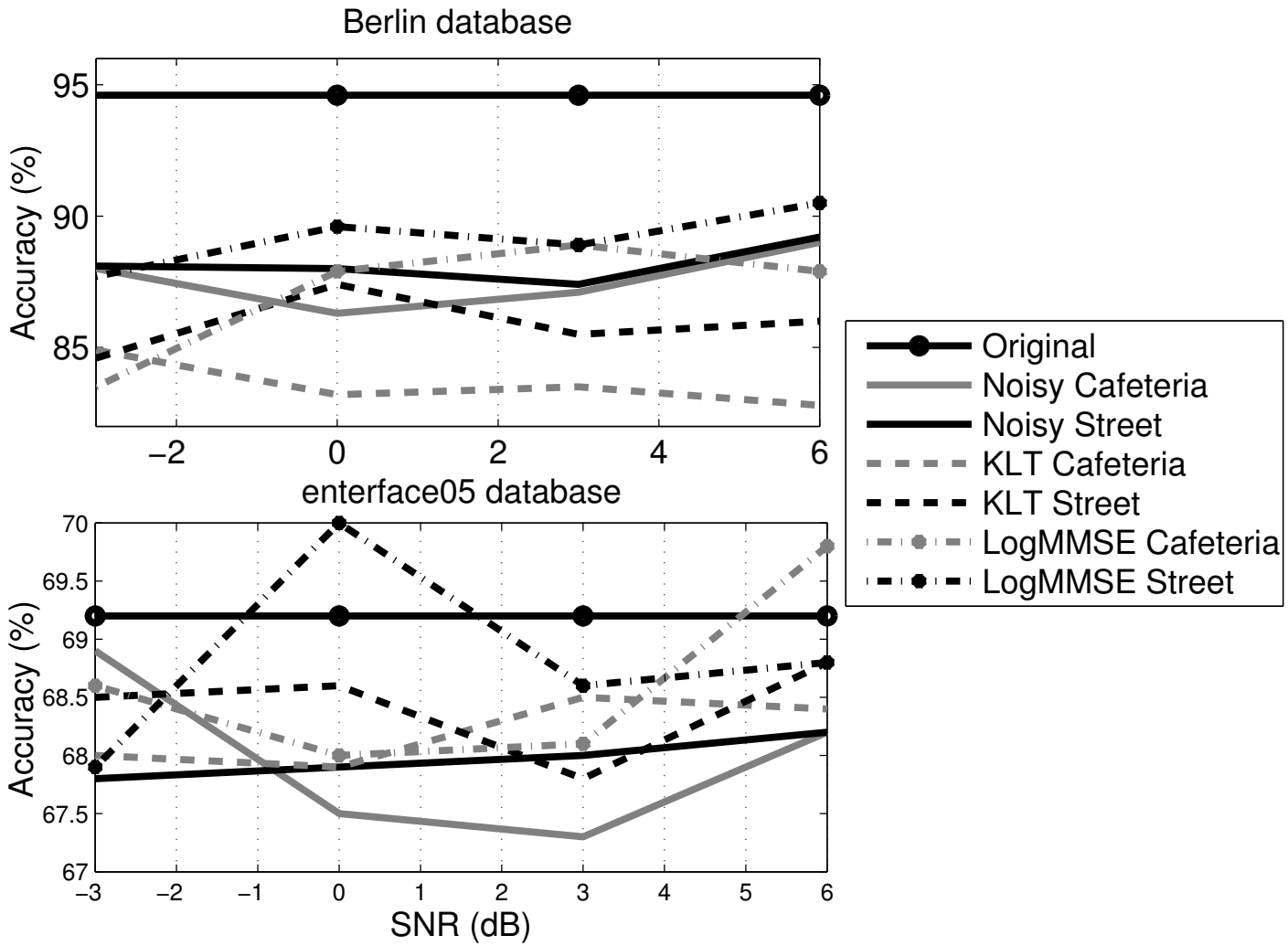
Databases

UNIVERSIDAD DE ANTIOQUIA

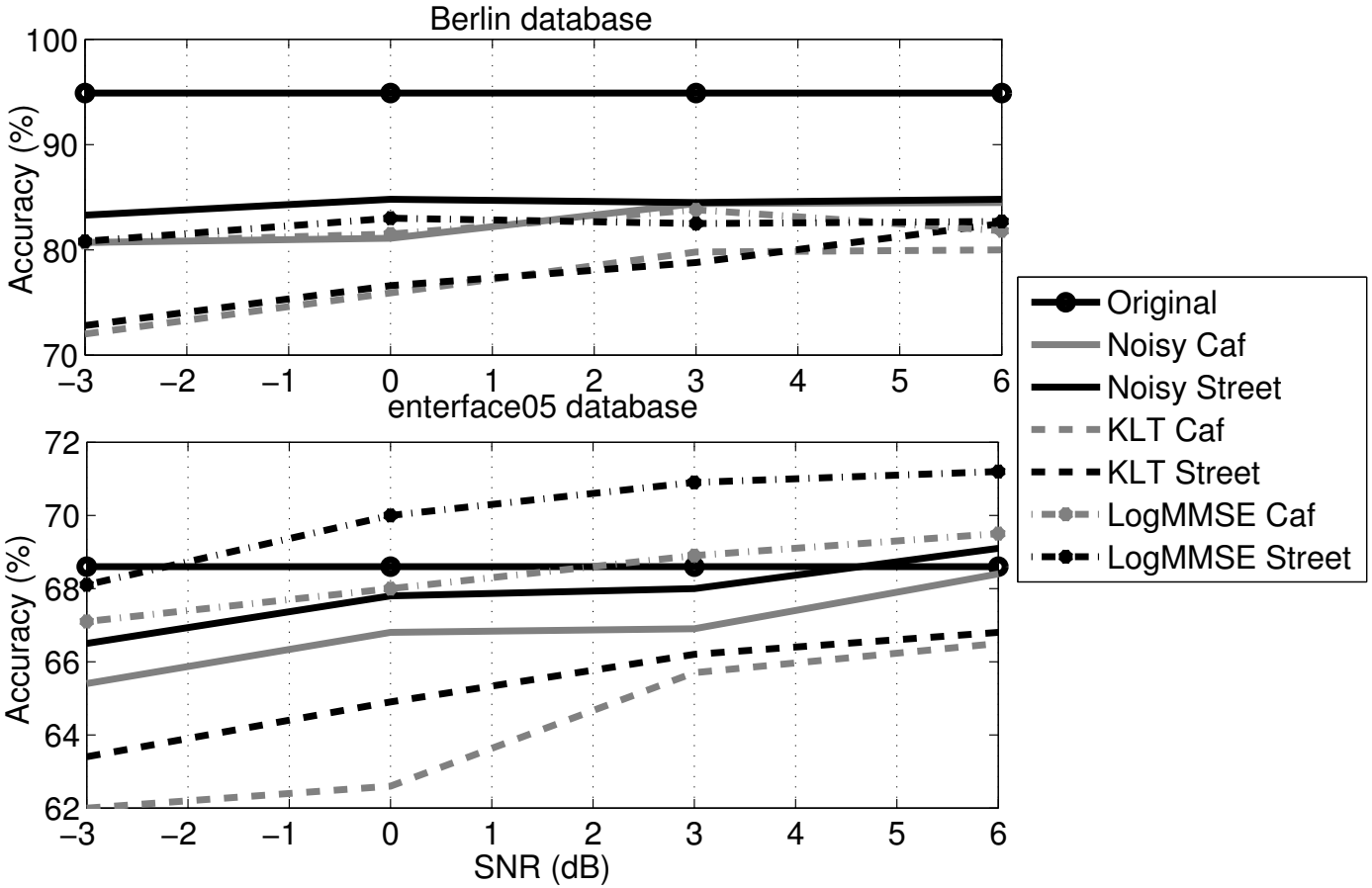
database	# recordings	Speakers	Fs (Hz)	Naturalness
Berlin	534	12	16000	Acted
Enterface05 (Audio-Video)	1317	44	44100	Evoked

Segments	Classif task	enterface05 logMMSE	Difference
openEAR	multi-class	66.9 ± 4.2	$+0.3$
	2-class	68.8 ± 3.1	$+0.2$

Results: Affected signals, 2class (WPT)



Results: Affected signals, 2-class (OpenEAR)



Conclusion I

1. A different scheme for feature extraction based on WPT is presented, it highlights the low frequency zone from the speech signal. Its performance it is acceptable for the 2-class problem when compared with a well established scheme as OpenEAR.

2. The use of WPT in low frequency bands must be evaluated more deeply in order to improve performance for Multi-class problem.

3. Other features calculated from the wavelet decompositions must be considered, specially for unvoiced segments.

Conclusion II

4. New methodology seems to be more robust against non-controlled conditions. Although logMMSE algorithm outperforms KLT, performance for Speech Enhancement is not good enough. The affectation produced by the cafeteria babble noise is more critical than the produced by the street noise.

5. Evaluation of non-additive environmental noise must be addressed in the future.

Thanks!
Q?

jesus.vargas@udea.edu.co

Thanks!
Q?

jesus.vargas@udea.edu.co