Non-linear Dynamics Characterization from Wavelet Packet Transform for Automatic Recognition of Emotional Speech

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

[1] Faculty of Engineering, Universidad de Antioquia UdeA
[2] Pattern Recognition Lab, Friedrich Alexander Universität, Erlangen-Nürnberg

Outline

1. **Introduction**
2. Methodology
3. Databases
4. Results
5. Conclusion

UNIVERSIDAD DE ANTIOQUIA
1803

jcamilo.vasquez@udea.edu.co

1. Introduction

Recognition of emotion in speech:

- ✓ Call centers
- ✓ Emergency services
- ✓ Psychologic therapy
- ✓ Intelligent vehicles
- ✓ Video games





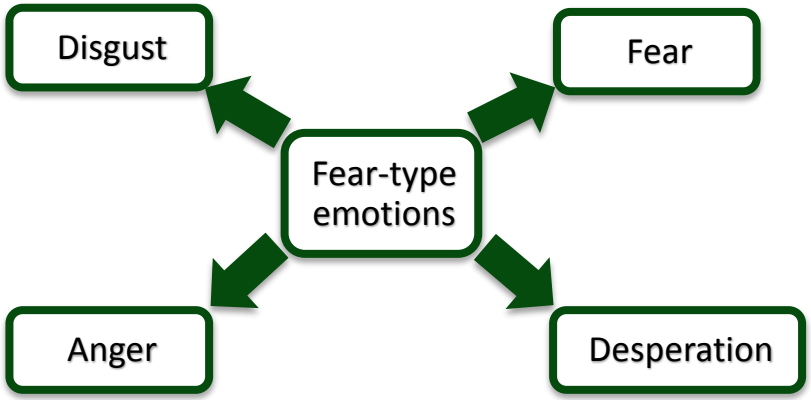


UNIVERSIDAD DE ANTIOQUIA
1803

jcamilo.vasquez@udea.edu.co

1. Introduction

- The interest has been focused on detection of fear-type emotions which appear in situations where the human integrity is at risk.

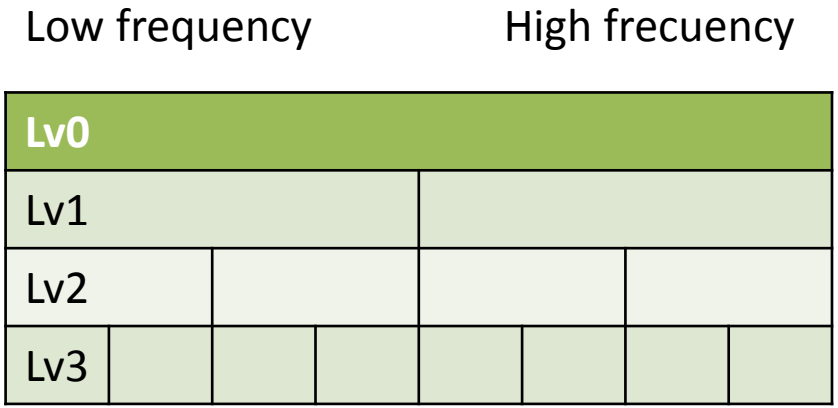


UNIVERSIDAD DE ANTIOQUIA
1803

jcamilo.vasquez@udea.edu.co

1. Introduction

- WPT provides a time-frequency multi-resolution analysis. NLD measures are estimated in each decomposed band.



Wavelet Packet Transform (WPT)

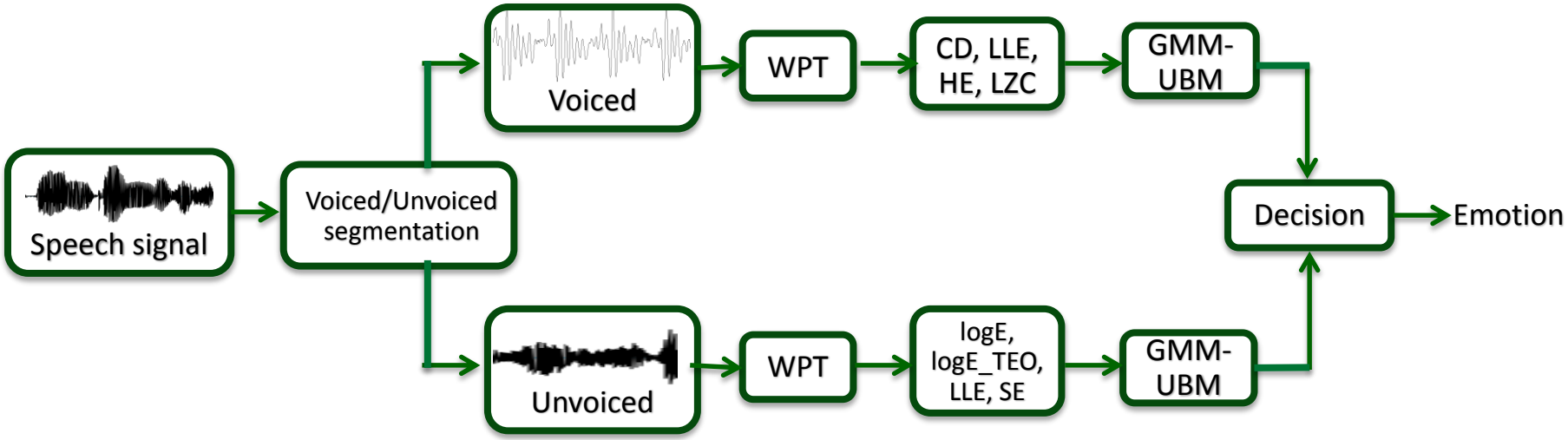
Outline

1. Introduction
2. **Methodology**
3. Databases
4. Results
5. Conclusion

UNIVERSIDAD DE ANTIOQUIA
1803

jcamilo.vasquez@udea.edu.co

2. Methodology



UNIVERSIDAD DE ANTIOQUIA
1803

jcamilo.vasquez@udea.edu.co

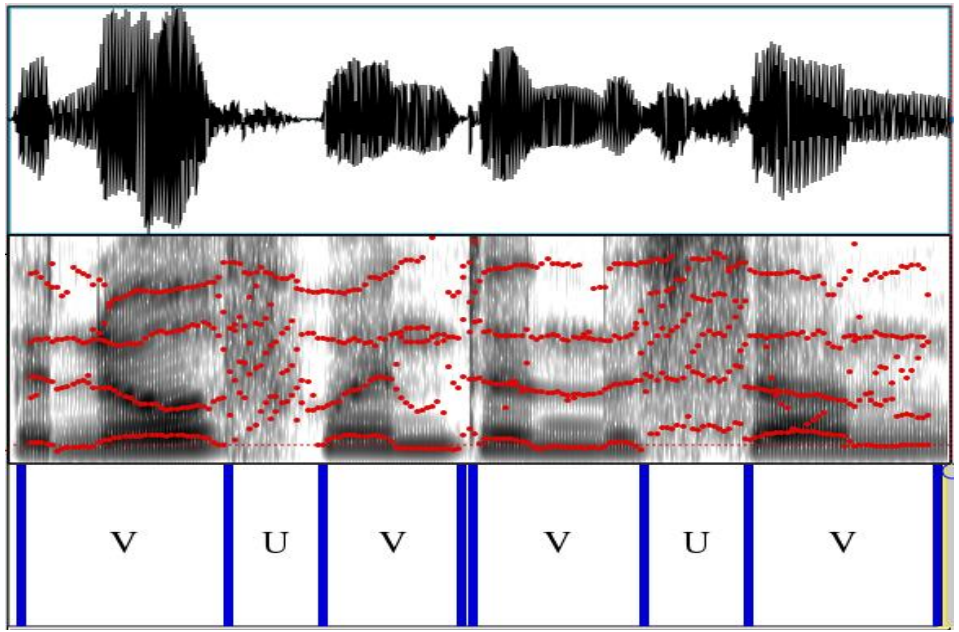
2. Methodology

Segmentation

Two types of sound:

- ✓ Voiced
- ✓ Unvoiced

Both kind of segments are processed independently



UNIVERSIDAD DE ANTIOQUIA
1803

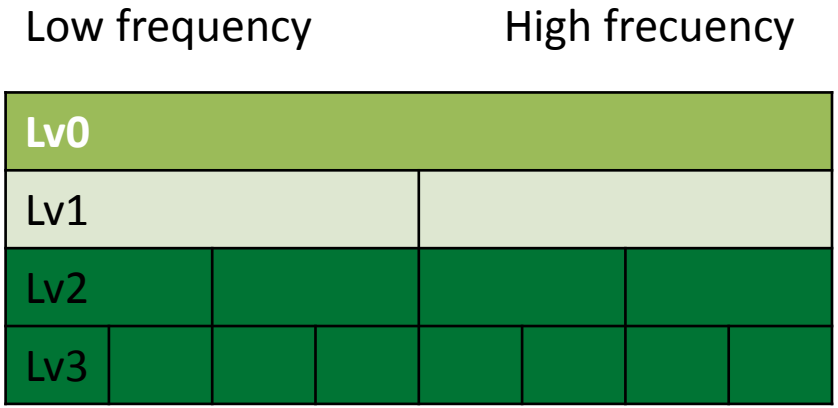
jcamilo.vasquez@udea.edu.co

2. Methodology

Wavelet Packet Transform

Features are estimated on each band:

- ✓ Log-Energy
- ✓ Teager Energy Operator (TEO)
- ✓ Entropies (Shannon, log-Energy)
- ✓ NLD (CD, LLE, HE, LZC)



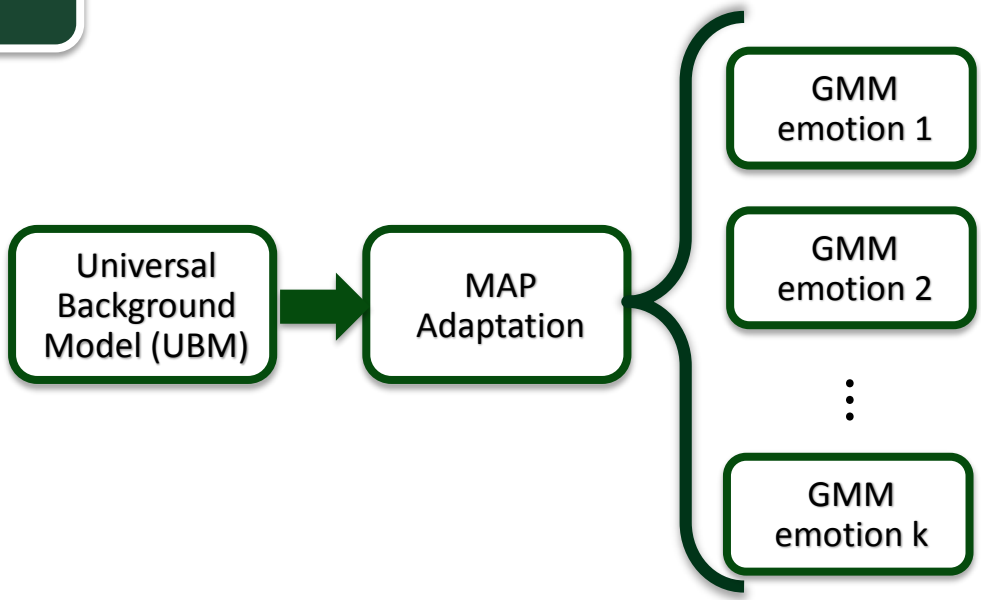
Wavelet Packet Transform (WPT)

UNIVERSIDAD DE ANTIOQUIA
1803

jcamilo.vasquez@udea.edu.co

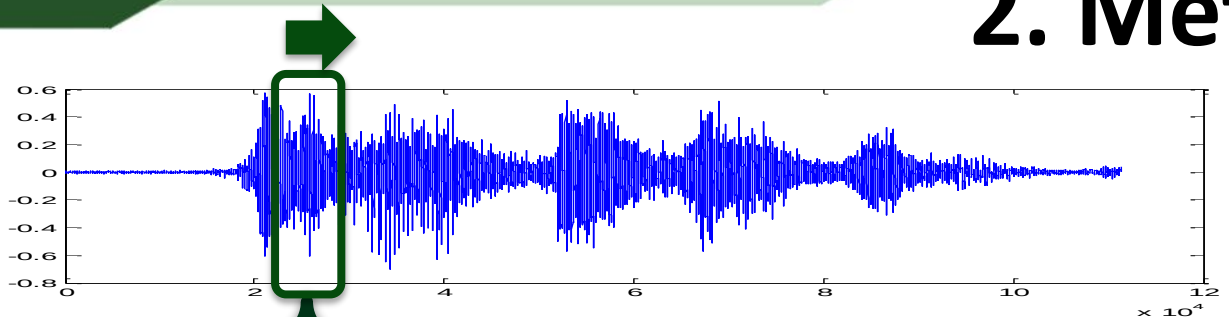
2. Methodology

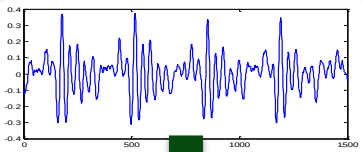
GMM-UBM



UNIVERSIDAD DE ANTIOQUIA
1803

2. Methodology





1764 samples 50% overlapping

Feature estimation

clasification

$P(X,\Theta)_k$ $P(X,\Theta)_{k+1}$ ... $P(X,\Theta)_n$

$$LL(X,\Theta) = \sum_{k=1}^{n} \log(P(X,\Theta)_k)$$

UNIVERSIDAD DE ANTIOQUIA
1803

jcamilo.vasquez@udea.edu.co

2. Methodology

Final Decision

Two different GMM were created for classification task, which are based on:

1. Voiced segments
2. Unvoiced segments

Then are combined in a second classification stage according to

$$P(\text{Score fusion}) = \alpha * P(\text{GMM Voiced}) + (1 - \alpha) * P(\text{GMM Unvoiced})$$

UNIVERSIDAD DE ANTIOQUIA
1803

jcamilo.vasquez@udea.edu.co

Outline

1. Introduction
2. Methodology
3. **Databases**
4. Results
5. Conclusion

UNIVERSIDAD DE ANTIOQUIA
1803

jcamilo.vasquez@udea.edu.co

3. Databases

Database	Num recordings	Num speakers	Sample frequency	Emotions recognized
GVEESS	224	12	44100	Anger Disgust Fear Desperation
Berlin	534	10	16000	Anger Disgust Fear
eNTERFACE05	1317	44	44100	Anger Disgust Fear

UNIVERSIDAD DE ANTIOQUIA
1803

Outline

1. Introduction
2. Methodology
3. Databases
4. **Results**
5. Conclusion

UNIVERSIDAD DE ANTIOQUIA
1803

jcamilo.vasquez@udea.edu.co

3. Results

Voiced Segments

Features	GVEESS Accuracy	Berlin Accuracy	eNTERFACE Accuracy
DC	57.1±14.6	62.7±13.9	47.6±3.8
LLE	68.0±16.2	67.6±8.1	52.1±4.9
HE	68.1±28.0	67.6±8.1	52.0±4.9
LZC	**82.0±11.3**	78.3±9.9	**54.0±7.3**
Comb	65.0±21.2	**79.0±10.0**	51.1±8.0

UNIVERSIDAD DE ANTIOQUIA
1803

jcamilo.vasquez@udea.edu.co

3. Results

Unvoiced Segments

Features	GVEESS Accuracy	Berlin Accuracy	eNTERFACE Accuracy
LogEnergy	93.4±9.8	64.7±11.1	46.9±4.4
LogEnergy TEO	93.1±8.8	60.8±8.1	54.2±4.9
SE	93.4±9.8	71.0±12.7	53.7±5.8
LEE	92.3±10.3	**77.2±10.9**	57.0±4.1
Comb.	**99.0±2.5**	69.1±16.0	**63.1±15.7**

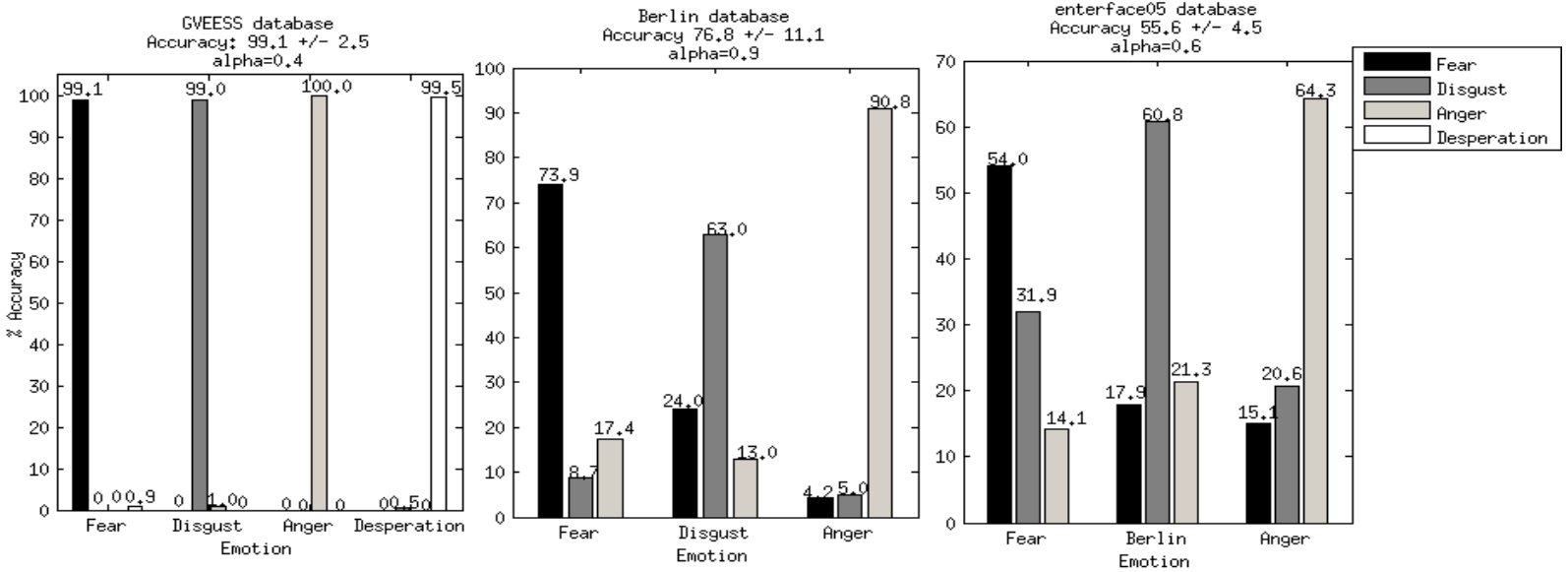
UNIVERSIDAD DE ANTIOQUIA
1803

jcamilo.vasquez@udea.edu.co

3. Results

Combination of probabilities

LZC Voiced and Comb. Unvoiced



UNIVERSIDAD DE ANTIOQUIA
1803

jcamilo.vasquez@udea.edu.co

Outline

1. Introduction
2. Methodology
3. Databases
4. **Results**
5. Conclusion

UNIVERSIDAD DE ANTIOQUIA
1803

jcamilo.vasquez@udea.edu.co

4. Conclusion

1. A new set of features based on NLD measures calculated from WPT are extracted from speech signals to perform the automatic recognition of fear-type emotions. The voiced and unvoiced segments of each recording are characterized separately.

2. The results indicate that LZC evaluated from wavelet decomposition in voiced segments provides a good representation of emotional speech.

3. Features derived from energy and entropy calculated from unvoiced segments are suitable to characterize emotional speech.

UNIVERSIDAD DE ANTIOQUIA
1803

jcamilo.vasquez@udea.edu.co

4. Conclusion

4. The evaluation of proposed features could be used as complement of classical features for emotion recognition from speech.

5. The proposed features must be evaluated in speech recordings in non-controlled noise conditions, and the wavelet transform in superior levels of decomposition must be addressed in future work in order to consider more resolution in frequency domain.

UNIVERSIDAD DE ANTIOQUIA
1803

jcamilo.vasquez@udea.edu.co

Thanks!

UNIVERSIDAD DE ANTIOQUIA
1803

jcamilo.vasquez@udea.edu.co