Wavelet-Based Time-Frequency Representations for Automatic Recognition of Emotions from Speech

J. C. Vásquez-Correa[1,2*], **T. Arias-Vergara**[1], J. R. Orozco-Arroyave[1,2], J. F. Vargas-Bonilla[1], E. Nöth[2]

[1]Department of Electronics and Telecommunication Engineering, University of Antioquia UdeA.
[2]Pattern recognition Lab. Friedrich Alexander Universität. Erlangen-Nürnberg.

**jcamilo.vasquez@udea.edu.co*

UNIVERSIDAD DE ANTIOQUIA 1803

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG
TECHNISCHE FAKULTÄT

Outline

Introduction: Emotions

Introduction: Emotion recognition

Recognition of emotion from speech:

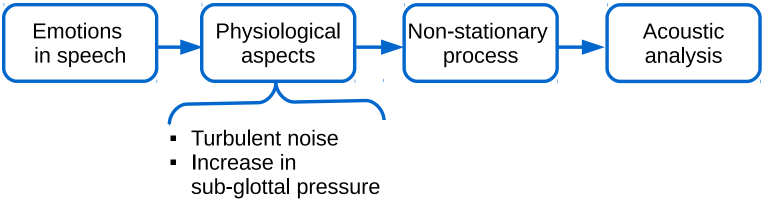
- Call centers
- Emergency services
- Depression Treatment
- Intelligent vehicles
- Public surveillance





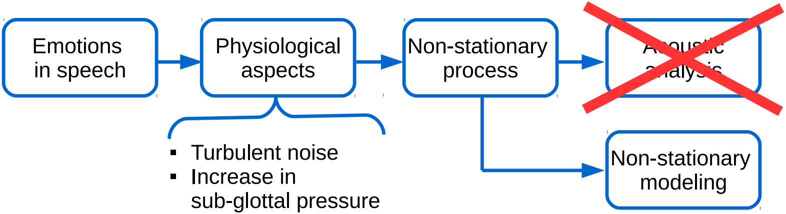



Introduction: Non-stationary analysis



Introduction: Non-stationary analysis

UNIVERSIDAD DE ANTIOQUIA



- Time–Frequency Analysis

  Wavelet Transform
  Wigner–Ville distribution
  Modulation Spectra

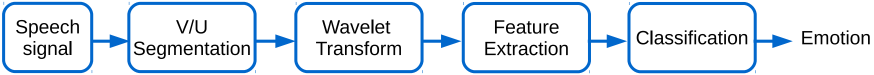
Introduction: Proposal

Features based on the energy content of three Wavelet–based TF representations for the classification of emotions from speech.

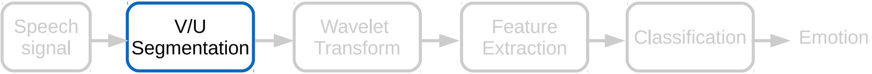
- Continuous Wavelet transform
- Bionic Wavelet transform
- Synchro–squeezing Wavelet transform

Methodology

UNIVERSIDAD DE ANTIOQUIA

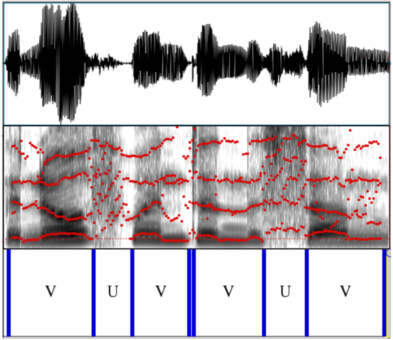


Methodology: segmentation

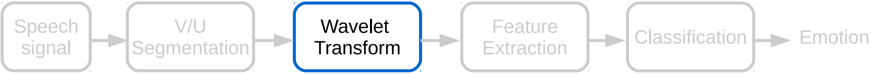


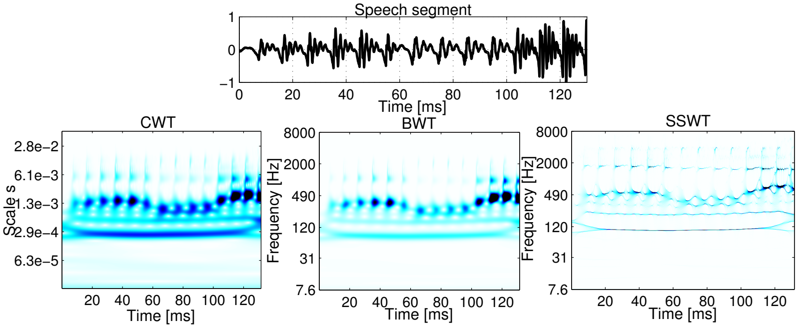
Two types of sounds:

- Voiced
- Unvoiced



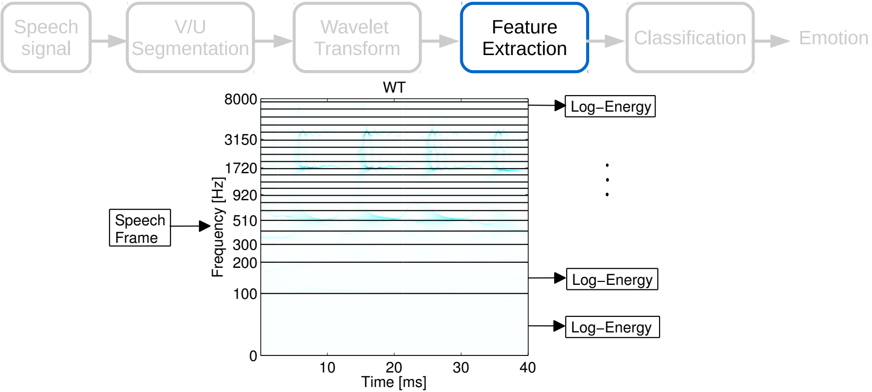
Methodology: Wavelet Transforms





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

Methodology: feature extraction



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

Methodology: feature extraction

UNIVERSIDAD DE ANTIOQUIA

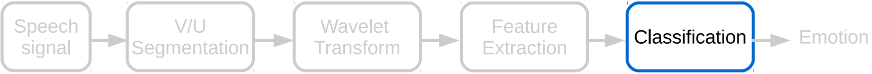
Speech signal → V/U Segmentation → Wavelet Transform → **Feature Extraction** → Classification → Emotion

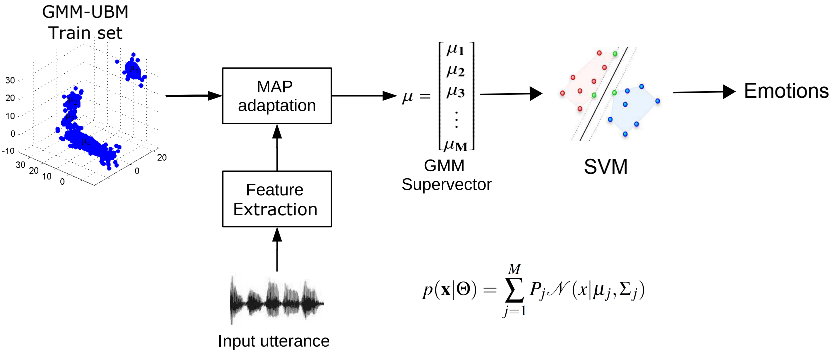
Descriptors (16 $\times$ 2)	**statistic functions (12)**
ZCR	mean
RMS Energy	standard deviation
F_0	kurtosis, skewness
HNR	max, min, relative position, range
MFCC 1-12	slope, offset, MSE linear regression
Δs	

Table: Features implemented using openEAR[1]

[1] 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.

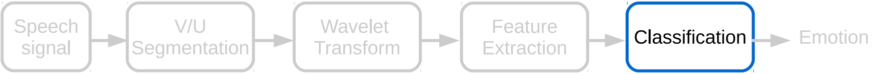
Methodology: classification





$$p(\mathbf{x}|\Theta) = \sum_{j=1}^{M} P_j \mathcal{N}(x|\mu_j, \Sigma_j)$$

Methodology: classification



- 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.

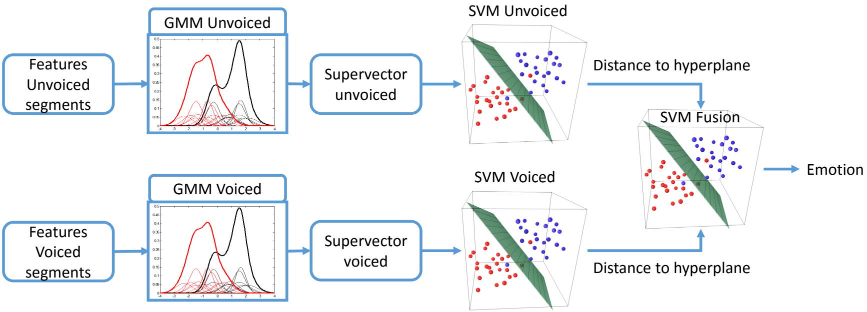
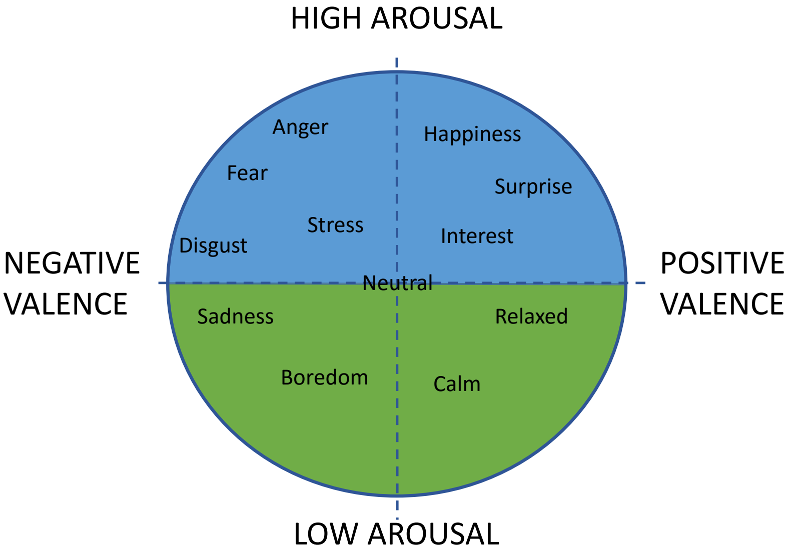


Table: Databases used in this study

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


HIGH AROUSAL
Anger
Happiness
Fear
Surprise
Stress
Interest
Disgust
NEGATIVE VALENCE
POSITIVE VALENCE
Neutral
Sadness
Relaxed
Boredom
Calm
LOW AROUSAL

UNIVERSIDAD DE ANTIOQUIA

Table: Detection of high vs. low arousal emotions. V: voiced, U: unvoiced.

Features	Segm.	Berlin	SAVEE	enterface05	IEMOCAP
CWT	V	96 ± 6	83 ± 9	81 ± 2	74 ± 4
	U	89 ± 9	80 ± 8	80 ± 1	75 ± 3
	Fusion	93 ± 8	87 ± 7	81 ± 3	76 ± 3
BWT	V	96 ± 6	82 ± 8	82 ± 2	74 ± 4
	U	90 ± 9	80 ± 7	80 ± 2	75 ± 3
	Fusion	94 ± 7	85 ± 7	82 ± 2	76 ± 4
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
OpenEAR	-	97 ± 3	83 ± 9	81 ± 2	76 ± 4

Table: Detection of high vs. low arousal emotions. V: voiced, U: unvoiced.

Features	Segm.	Berlin	SAVEE	enterface05	IEMOCAP
CWT	V	96 ± 6	83 ± 9	81 ± 2	74 ± 4
	U	89 ± 9	80 ± 8	80 ± 1	75 ± 3
	Fusion	93 ± 8	87 ± 7	81 ± 3	76 ± 3
BWT	V	96 ± 6	82 ± 8	82 ± 2	74 ± 4
	U	90 ± 9	80 ± 7	80 ± 2	75 ± 3
	Fusion	94 ± 7	85 ± 7	82 ± 2	76 ± 4
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
OpenEAR	-	97 ± 3	83 ± 9	81 ± 2	76 ± 4

Table: Detection of high vs. low arousal emotions. V: voiced, U: unvoiced.

Features	Segm.	Berlin	SAVEE	enterface05	IEMOCAP
CWT	V	96 ± 6	83 ± 9	81 ± 2	74 ± 4
	U	89 ± 9	80 ± 8	80 ± 1	75 ± 3
	Fusion	93 ± 8	87 ± 7	81 ± 3	76 ± 3
BWT	V	96 ± 6	82 ± 8	82 ± 2	74 ± 4
	U	90 ± 9	80 ± 7	80 ± 2	75 ± 3
	Fusion	94 ± 7	85 ± 7	82 ± 2	76 ± 4
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
OpenEAR	-	97 ± 3	83 ± 9	81 ± 2	76 ± 4

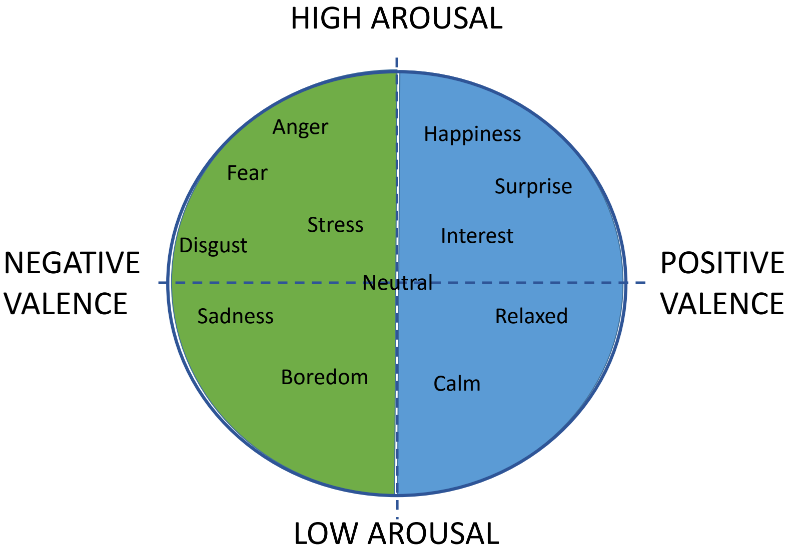

HIGH AROUSAL
Anger
Fear
Stress
Disgust
Happiness
Surprise
Interest
NEGATIVE VALENCE
POSITIVE VALENCE
Neutral
Sadness
Boredom
Relaxed
Calm
LOW AROUSAL

Table: Detection of positive vs. negative valence emotions. V: voiced, U: unvoiced.

Features	Segm.	Berlin	SAVEE	enterface05	IEMOCAP
CWT	V	80 ± 4	64 ± 5	75 ± 2	55 ± 4
	U	76 ± 5	64 ± 3	73 ± 3	58 ± 2
	Fusion	78 ± 4	67 ± 4	74 ± 2	58 ± 5
BWT	V	80 ± 4	64 ± 6	74 ± 2	55 ± 4
	U	76 ± 7	64 ± 5	74 ± 3	58 ± 2
	Fusion	78 ± 6	65 ± 6	74 ± 4	58 ± 3
SSWT	V	82 ± 5	64 ± 5	76 ± 3	56 ± 4
	U	77 ± 6	63 ± 3	74 ± 3	58 ± 2
	Fusion	79 ± 4	65 ± 5	74 ± 4	60 ± 3
OpenEAR	-	87 ± 2	72 ± 6	81 ± 4	59 ± 3

Table: Detection of positive vs. negative valence emotions. V: voiced, U: unvoiced.

Features	Segm.	Berlin	SAVEE	enterface05	IEMOCAP
CWT	V	80 ± 4	64 ± 5	75 ± 2	55 ± 4
	U	76 ± 5	64 ± 3	73 ± 3	58 ± 2
	Fusion	78 ± 4	67 ± 4	74 ± 2	58 ± 5
BWT	V	80 ± 4	64 ± 6	74 ± 2	55 ± 4
	U	76 ± 7	64 ± 5	74 ± 3	58 ± 2
	Fusion	78 ± 6	65 ± 6	74 ± 4	58 ± 3
SSWT	V	82 ± 5	64 ± 5	76 ± 3	56 ± 4
	U	77 ± 6	63 ± 3	74 ± 3	58 ± 2
	Fusion	79 ± 4	65 ± 5	74 ± 4	60 ± 3
OpenEAR	-	87 ± 2	72 ± 6	81 ± 4	59 ± 3

UNIVERSIDAD DE ANTIOQUIA

Table: Detection of positive vs. negative valence emotions. V: voiced, U: unvoiced.

Features	Segm.	Berlin	SAVEE	enterface05	IEMOCAP
CWT	V	80 ± 4	64 ± 5	75 ± 2	55 ± 4
	U	76 ± 5	64 ± 3	73 ± 3	58 ± 2
	Fusion	78 ± 4	67 ± 4	74 ± 2	58 ± 5
BWT	V	80 ± 4	64 ± 6	74 ± 2	55 ± 4
	U	76 ± 7	64 ± 5	74 ± 3	58 ± 2
	Fusion	78 ± 6	65 ± 6	74 ± 4	58 ± 3
SSWT	V	82 ± 5	64 ± 5	76 ± 3	56 ± 4
	U	77 ± 6	63 ± 3	74 ± 3	58 ± 2
	Fusion	79 ± 4	65 ± 5	74 ± 4	60 ± 3
OpenEAR	-	87 ± 2	72 ± 6	81 ± 4	59 ± 3

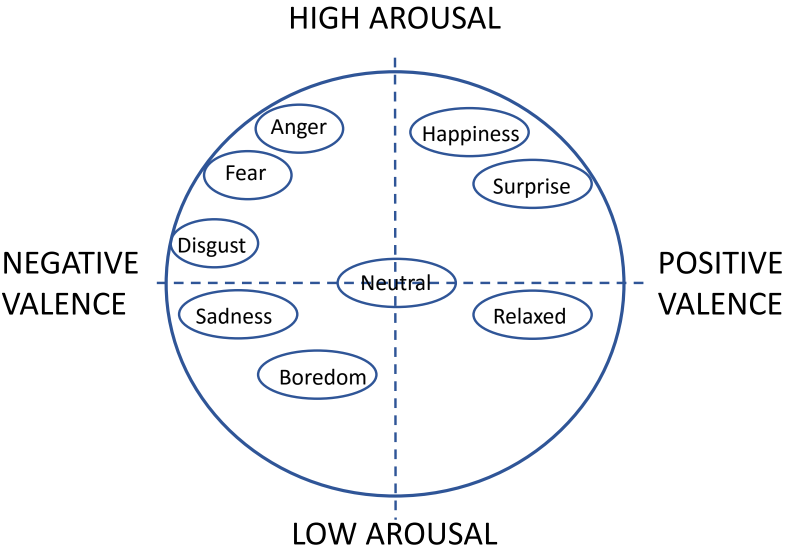

HIGH AROUSAL
Anger
Fear
Happiness
Surprise
Disgust
NEGATIVE VALENCE
Neutral
POSITIVE VALENCE
Sadness
Relaxed
Boredom
LOW AROUSAL

Table: Classification of multiple emotions. V: voiced, U: unvoiced.

Features	**Segm.**	**Berlin**	**SAVEE**	**enterface-05**	**IEMOCAP**
CWT	V	61 ± 8	41 ± 13	48 ± 5	47 ± 6
	U	55 ± 7	39 ± 6	46 ± 4	51 ± 4
	Fusion	67 ± 7	44 ± 9	51 ± 6	56 ± 5
BWT	V	64 ± 9	41 ± 15	48 ± 4	47 ± 5
	U	56 ± 7	40 ± 4	45 ± 4	51 ± 4
	Fusion	66 ± 7	47 ± 10	50 ± 4	55 ± 6
SSWT	V	64 ± 8	43 ± 11	48 ± 4	49 ± 5
	U	55 ± 8	40 ± 6	46 ± 4	52 ± 3
	Fusion	69 ± 8	45 ± 12	49 ± 6	58 ± 4
OpenEAR	-	80 ± 8	49 ± 17	63 ± 7	57 ± 3

Table: Classification of multiple emotions. V: voiced, U: unvoiced.

Features	Segm.	Berlin	SAVEE	enterface-05	IEMOCAP
CWT	V	61 ± 8	41 ± 13	48 ± 5	47 ± 6
	U	55 ± 7	39 ± 6	46 ± 4	51 ± 4
	Fusion	67 ± 7	44 ± 9	51 ± 6	56 ± 5
BWT	V	64 ± 9	41 ± 15	48 ± 4	47 ± 5
	U	56 ± 7	40 ± 4	45 ± 4	51 ± 4
	Fusion	66 ± 7	47 ± 10	50 ± 4	55 ± 6
SSWT	V	64 ± 8	43 ± 11	48 ± 4	49 ± 5
	U	55 ± 8	40 ± 6	46 ± 4	52 ± 3
	Fusion	69 ± 8	45 ± 12	49 ± 6	58 ± 4
OpenEAR	-	80 ± 8	49 ± 17	63 ± 7	57 ± 3

Conclusion

UNIVERSIDAD DE ANTIOQUIA

- This study evaluates different wavelet based TF representations to model emotional speech (CWT, BWT, SSWT).
- When comparing these three TF–based transformations, SSWT provides better results.
- In most of the cases the highest UARs are obtained with the features extracted from voiced segments.
- The fusion scheme shows to be useful to combine the information provided by both kinds of segments.
- The results with the proposed approach are better than those obtained with openEAR when classifying high vs. low arousal emotions.
- Further experiments shall be performed considering other descriptors extracted from the TF representations to improve the results in other classification tasks.

Conclusion

UNIVERSIDAD DE ANTIOQUIA

- This study evaluates different wavelet based TF representations to model emotional speech (CWT, BWT, SSWT).
- When comparing these three TF–based transformations, SSWT provides better results.
- In most of the cases the highest UARs are obtained with the features extracted from voiced segments.
- The fusion scheme shows to be useful to combine the information provided by both kinds of segments.
- The results with the proposed approach are better than those obtained with openEAR when classifying high vs. low arousal emotions.
- Further experiments shall be performed considering other descriptors extracted from the TF representations to improve the results in other classification tasks.

- This study evaluates different wavelet based TF representations to model emotional speech (CWT, BWT, SSWT).
- When comparing these three TF–based transformations, SSWT provides better results.
- In most of the cases the highest UARs are obtained with the features extracted from voiced segments.
- The fusion scheme shows to be useful to combine the information provided by both kinds of segments.
- The results with the proposed approach are better than those obtained with openEAR when classifying high vs. low arousal emotions.
- Further experiments shall be performed considering other descriptors extracted from the TF representations to improve the results in other classification tasks.

Conclusion

- This study evaluates different wavelet based TF representations to model emotional speech (CWT, BWT, SSWT).
- When comparing these three TF–based transformations, SSWT provides better results.
- In most of the cases the highest UARs are obtained with the features extracted from voiced segments.
- The fusion scheme shows to be useful to combine the information provided by both kinds of segments.
- The results with the proposed approach are better than those obtained with openEAR when classifying high vs. low arousal emotions.
- Further experiments shall be performed considering other descriptors extracted from the TF representations to improve the results in other classification tasks.

Conclusion

- This study evaluates different wavelet based TF representations to model emotional speech (CWT, BWT, SSWT).
- When comparing these three TF–based transformations, SSWT provides better results.
- In most of the cases the highest UARs are obtained with the features extracted from voiced segments.
- The fusion scheme shows to be useful to combine the information provided by both kinds of segments.
- The results with the proposed approach are better than those obtained with openEAR when classifying high vs. low arousal emotions.
- Further experiments shall be performed considering other descriptors extracted from the TF representations to improve the results in other classification tasks.

Conclusion

- This study evaluates different wavelet based TF representations to model emotional speech (CWT, BWT, SSWT).
- When comparing these three TF–based transformations, SSWT provides better results.
- In most of the cases the highest UARs are obtained with the features extracted from voiced segments.
- The fusion scheme shows to be useful to combine the information provided by both kinds of segments.
- The results with the proposed approach are better than those obtained with openEAR when classifying high vs. low arousal emotions.
- Further experiments shall be performed considering other descriptors extracted from the TF representations to improve the results in other classification tasks.

Thanks!

jcamilo.vasquez@udea.edu.co

Wavelet-Based Time-Frequency Representations for Automatic Recognition of Emotions from Speech

J. C. Vásquez-Correa[1,2*], **T. Arias-Vergara**[1],
J. R. Orozco-Arroyave[1,2], J. F. Vargas-Bonilla[1], E. Nöth[2]

[1]Department of Electronics and Telecommunication Engineering,
University of Antioquia UdeA.
[2]Pattern recognition Lab. Friedrich Alexander Universität. Erlangen-Nürnberg.

**jcamilo.vasquez@udea.edu.co*

UNIVERSIDAD DE ANTIOQUIA 1803

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG
TECHNISCHE FAKULTÄT