Fusing Affective Dimensions and Audio-Visual Features from Segmented Video for Depression Recognition

Humberto Pérez, Hugo Jair Escalante, Luis Villaseñor, Manuel Montes, David Pinto, Verónica Reyes

Laboratorio de Tecnologías del Lenguaje
Coordinación de Ciencias Computacionales
Instituto Nacional de Astrofísica, Óptica y Electrónica

Outline

- Introduction
- *Challenges* of the depression recognition *challenge*
- Proposed approach
- Experimental results
- Discussion

Depression and mental disorders

- Depression affects a large portion of world population (350 million in 2012, WHO)

- The leading cause of disability in the world

- ITs could offer support for therapists:
  - Massive / Online / anytime monitoring of patients
  - Identification of patients suffering depression
  - Support tools to quantify the progress of the disease
  - Large scale studies
  - …

AVEC '14: Problem settting

- To learn a model to predict the degree of depression (BDI-II) of patients by analyzing clips (video+audio) in which patients *interact* (one -way) with a computer

Challenges of the AVE challenge

- Tiny training data set
- Raw video and audio (recorded with a webcam)
- Imbalanced “categories”
- Predictive variable was BDI-II
- Clips were not necessarily recorded when the patient is expressing the corresponding BDI
- Wide variety of subjects
- For some clips no word was pronounced
- …

In spite of these challening conditions, the potential impact of DR systems is huge and, therefore, it is worth approaching it

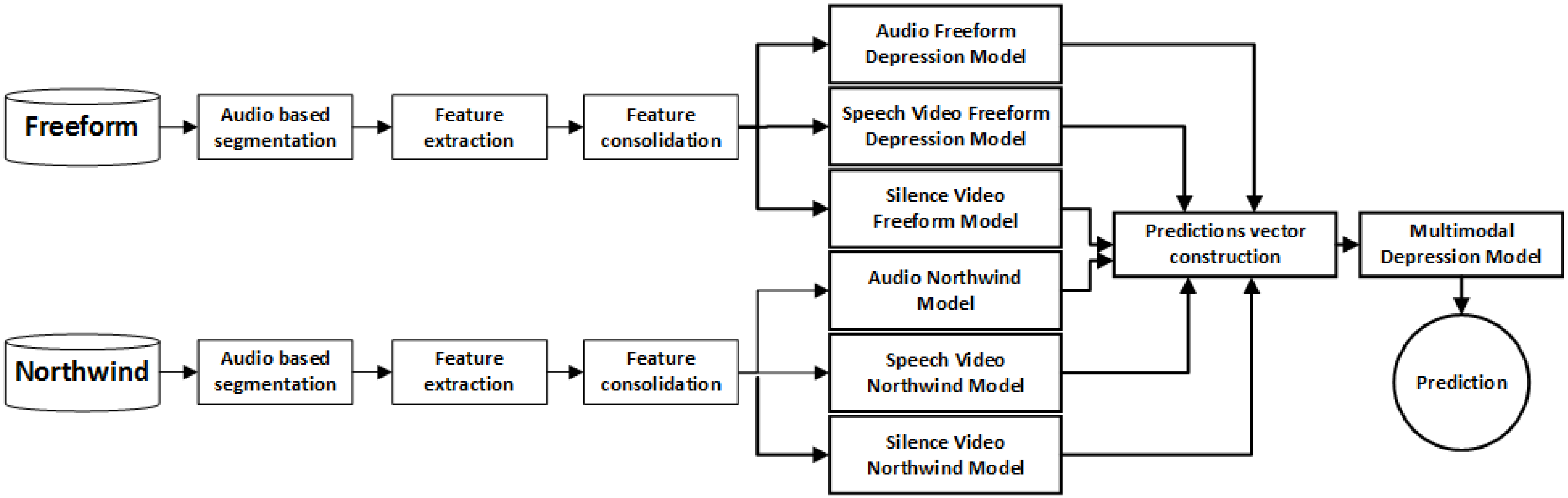
Our solution to AVEC2014's DRC

- We approach the problem as one of regression, with two novel components:
  - Clip segmentation, and the use of segment-level features
  - Using affective dimensions as features
- Further
  - Exploiting multimodal information
  - Exploring two segment aggregation strategies

Our solution to AVEC2014's DRC

- Working hypotheses (research questions):
  - How strongly correlated are the affective dimensions to the depression indicator?
  - What is the appropriate segment size to estimate more accurately valence, arousal and dominance?
  - Is worth combining multimodal information?, how?

Our solution to AVEC2014’s DRC



Audio-based semgentation

- Motivation:
  - Local modeling of affective and audiovisual information
  - Affect is expressed intensively in short episodes, emotions can change rapidly
- Clips are segmented into sound and silence intervals (PRAAT),
  - segments of [0.5-2] seconds long
- Voice-segment identification (syllable detection and classifier)

Affective dimensions as features

- Can affective dimensions be good predictive variables for depression recognition?

- Affective dimensions we computed on a segment-level basis (we took the average across a segment)
  - **Training and development:** use te ground truth dimensions
  - **Test:** use predictions from a regression model

Affective dimensions as features

- Initial evidence (ground-truth AD):

Primitive	Northwind	Freeform
Arousal	-0.45	-0.32
Dominance	-0.44	-0.20
Valence	-0.46	-0.46
Average	-0.45	-0.32

Pearson correlation coefficient BDI-II –vs. Affective dimensions (training data)

Primitive	A	D	V
A	1	0.64	0.58
D	0.64	1	0.58
V	0.58	0.58	1

Pearson correlation coefficient among affective dimensions

Affective dimensions as features

- Realistic scenario: obtaining AD values for test samples
  - We used a regression model (SVR) at the segment level, trained with baseline audio features
  - Comparison of two segmentation methods

Task	Arousal	Dominance	Valence
Provided VAD Segmentation			
Freeform	0.5060	0.4764	0.5045
Northwind	0.6312	0.5565	0.2858
Proposed Segmentation			
Freeform	0.6477	0.6680	0.3771
Northwind	0.4532	0.6430	0.5781

Affective dimensions as features

- Affective attributes were combined with additional features derived from the audio signal:
  - Averaged speech rate along clip (number of detected syllables/segment duration).
  - Number of silence intervals greater than 10 seconds and less than 20 seconds.
  - Total time, in seconds, of silence intervals greater than 10 seconds and less than 20 seconds.
  - Number of silence intervals greater than 20 seconds
  - Total time, in seconds, of silence intervals greater than 20 seconds
  - Percentage of total voice time classified as neutral
  - Percentage of total voice time classified as happiness
  - Total duration of voice intervals

Each clip represented by the average values of attributes across segments

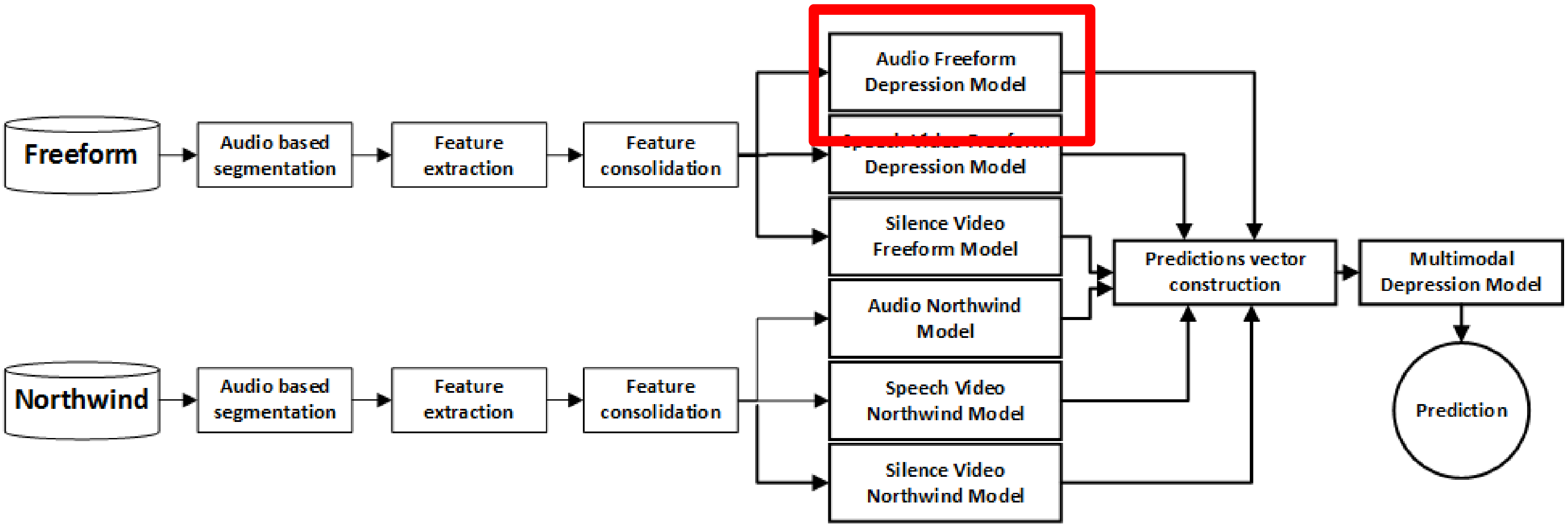
Visual features

- We consider raw motion/velocity attributes

- Face and eyes were detected (Viola & Jones) in segments we characterized segments as follows:
  - Difference of initial and final positions of face/eyes
  - Average, maximum, minimum, coordinates of face/eyes during the clip
  - Average velocity of face/eyes (x/y axis)
  - Motion history image / static history image

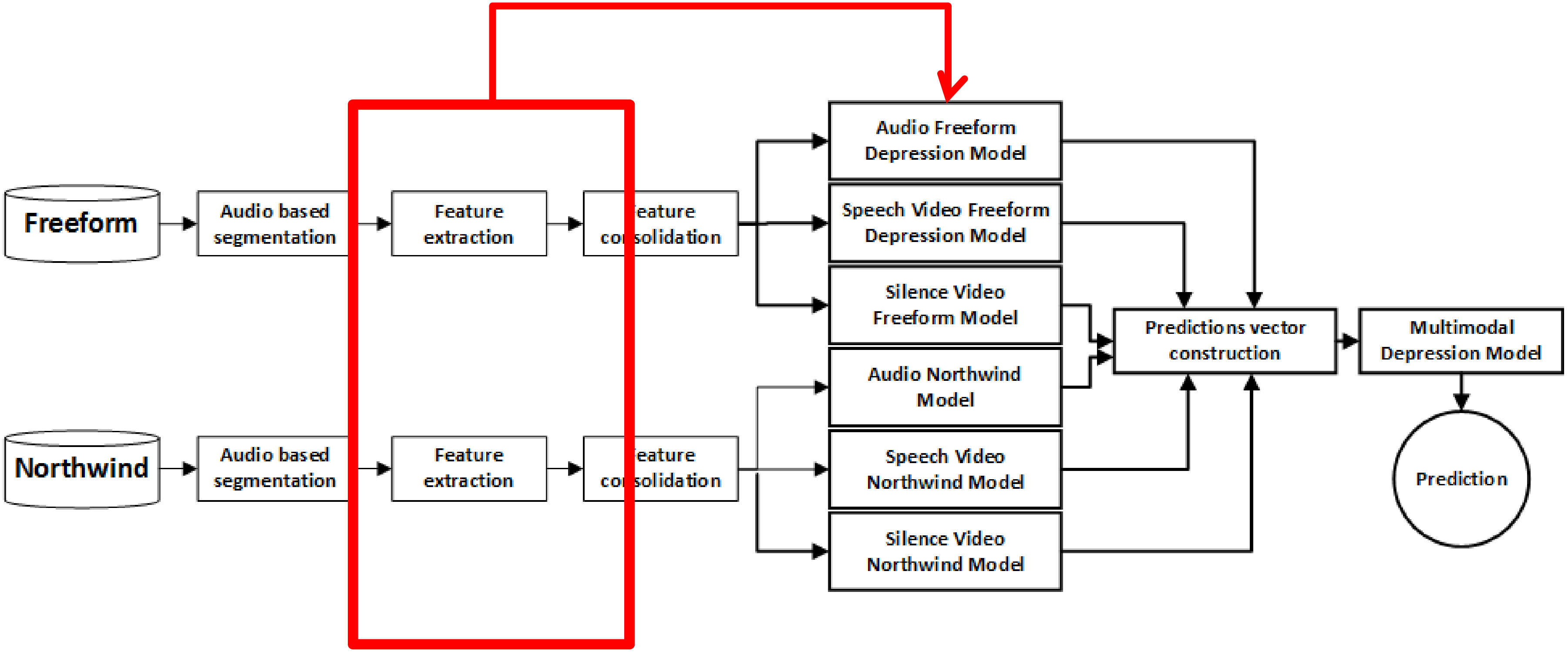
Visual features were extracted from both: voice and silence segments

Our solution to AVEC2014's DRC



- Two variants to our proposal:
  - Best individual model
  - Majority voting

Our solution to AVEC2014's DRC



- Two variants to our proposal:
  - Best individual model
  - Majority voting

Experimental study

- Depression recognition performance using AD only (training-development)

Task	Correlation	MAE	RMS
Freeform	0.4583	8.2976	11.2962
Northwind	0.5224	7.906	10.9192
Both	0.5224	7.906	10.9192

Experimental study

- Best individual model (training-development)

Modality	Correlation	MAE	RMS
North Wind			
Audio	0.4811	8.902	10.6195
Video Voice	0.3156	9.4721	11.51
Video Silence	0.4573	9.6723	11.18
Audio+Video*	0.6026	7.7969	9.7873
Free Form			
Audio	**0.6864**	**7.4895**	**8.9676**
Video Voice	0.1146	8.64	10.4754
Video Silence	0.0614	8.7861	10.2169
Audio+Video*	0.6534	7.4723	9.0336

Audio+Video (*) means that audio features were combined with both Video Voice (VViddeo) and Video Silence (SVideo).

Experimental study

- Majority voting (training-development)

Modality	Correlation	MAE	RMS
North Wind			
Audio	0.43804	8.7660	10.800
Video Voice	0.16385	9.7447	11.832
Video Silence	0.38159	9.7692	11.419
Audio+Video	0.4678	9.1763	10.5641
Free Form			
Audio	0.34598	10.146	13.447
Video Voice	0.23876	8.6591	10.714
Video Silence	**0.32435**	**8.4634**	**9.8414**
Audio+Video	0.3759	9.1512	11.0124

Experimental study

- Meta model (training-development) :

Modality	Correlation	MAE	RMS
Feature consolidation			
Audio+Video	0.7261	6.7862	8.3058
Majority vote approach			
Audio+Video	0.5209	7.9641	10.1376

- Meta model (test) :

Modality	MAE	RMS
Direct Prediction		
Audio Freeform	9.3539	11.9165
Meta-classifier		
Audio+VVideo+SVideo	**8.9910**	**10.8239**

Conlusions?

- Using AD as features is a promising and fruitful approach for depression recognition, although results were somewhat disapointing

- The best individual model (audio-based) resulted very competitive as well

- The meta-model approach proved to be effective to (slightly) boost performance

- The clip segmentation method performed better than the baseline model

Thank you

Laboratorio de Tecnologías del Lenguaje
Coordinación de Ciencias Computacionales
Instituto Nacional de Astrofísica, Óptica y Electrónica

BENEMERITA UNIVERSIDAD AUTONOMA DE PUEBLA

UPAEP

INAOE