High-level Features for Multimodal Deception Detection in Videos

Rodrigo Rill-García[*1]
Hugo Jair Escalante[1,2], Luis Villaseñor-Pineda[1],
Verónica Reyes-Meza[3]

*rodrigo.rill@{inaoep.mx, hotmail.com}
1. INAOE, Mexico
2. CINVESTAV, Mexico
3. UAT, Mexico

Motivation

- An “optimal” decision can be ***harmful*** if it is based on inaccurate (or wrong) data
- Purposely spreading ***inaccurate/wrong*** information is a way to ***mislead people***
  - Doing so for personal gain is the definition of ***deceiving***

Problem Description

- Deception detection is a ***hard task*** for humans
  - Untrained people have an average accuracy ~54% [1]

- Research supports that there is a ***difference*** in the way ***liars*** communicate in contrast with ***truth tellers***
  - Furthermore, such difference ***can be pointed out using Machine Learning***

Problem Description (2)

- There are many available sources of ***cues of deception interpretable*** by humans
  - Eye movements
  - Facial expressions
  - Voice
  - Speech
  - Etc.

- Recent research suggests ***multimodal analysis*** can ***improve the performance*** of analyzing different modalities separately

Objective

To develop a ***multimodal*** information ***fusion method***, inspired by classifier ***ensemble techniques***, for ***deception detection in videos*** using ***high-level features***

Related Work

- ***"Detecting deceptive behavior via integration of discriminative features from multiple modalities"*** [2]
  - Physiological features, thermal videos and transcriptions
  - Early fusion
  - Fused non-invasive features surpassed physiological ones
- ***"Deception detection using real-life trial data"*** [3]
  - Videos (image) and transcriptions
  - Early fusion
  - Best performance with fused features
- ***"Deception detection in videos"*** [4]
  - Videos (image and audio) and trasncriptions
  - Late fusion
  - Best performance with fused features
- ***"Toward End-to-End Deception Detection in Videos"*** [5]
  - Videos (image and audio)
  - Early fusion
  - Best performance with fused features

***No focus on multimodal fusion strategies**

Datasets

Database	*Court Trial*	*Abortion/Friend Spanish*
Deceptive/Truthful	61/60	22/21
Subjects	60	12

Table 1. Summary of the databases used.

Figure 1. Examples of Spanish videos.

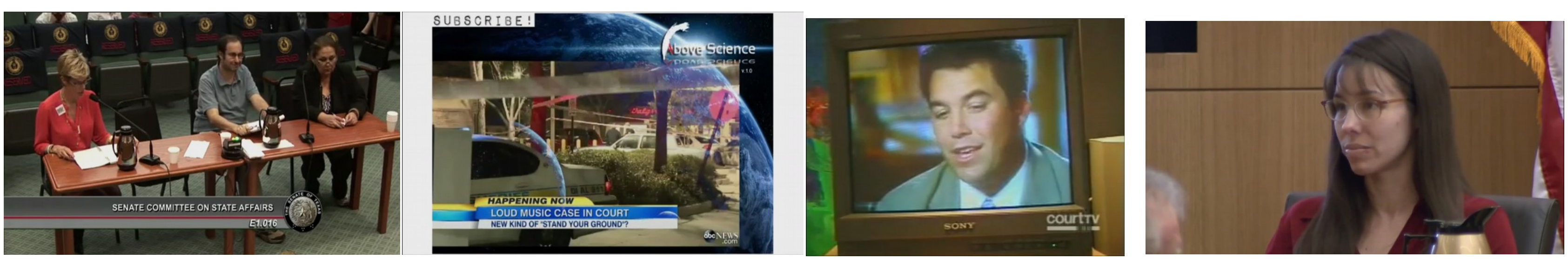

Figure 2. Examples of court videos [3].

Feature Extraction

Modality	Visual	Accoustic	Textual
	* per frame	** per frame	*** per video
Views	AU Int	Voice	Char 1-grams
	AU Pres	Glottal Flow	Char 2-grams
	Eye LM	MCEP	Char 3-grams
	Facial LM	HMPDM	Char 4-grams
	Gaze	HMPDD	POS 1-grams
	Head		POS 2-grams
			POS 3-grams
			POS 1-grams
			BoW
			LIWC
			Syntax Info

Figure 3. The different views extracted for each of the 3 proposed modalities.

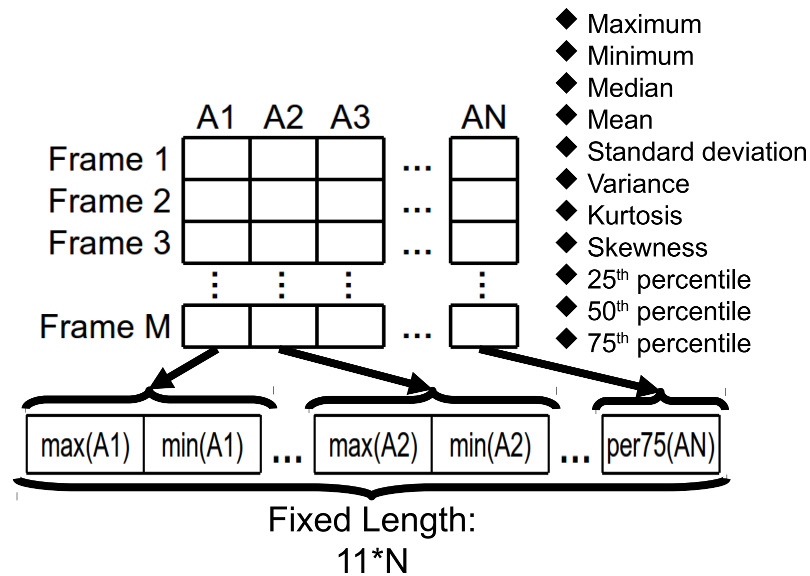

Figure 4. Creation of a fixed size vector from a number-of-frames-dependent matrix.

* OpenFace
** COVAREP
*** IBM Watson ASR, Google SyntaxNet, Python NLTK

Experimental settings

- ***N feature sets*** (views) are extracted per video
  - Textual modality is not extracted for Spanish
    - Lack of a Mexican Spanish ASR system
- Metric: ***AUC ROC*** of the **Deceptive** class
  - ***10-folds cross-validation***
    - ***No subject seen in training*** is contained ***in the validation*** set

Single views

- Court (Sklearn, LinearSVC)

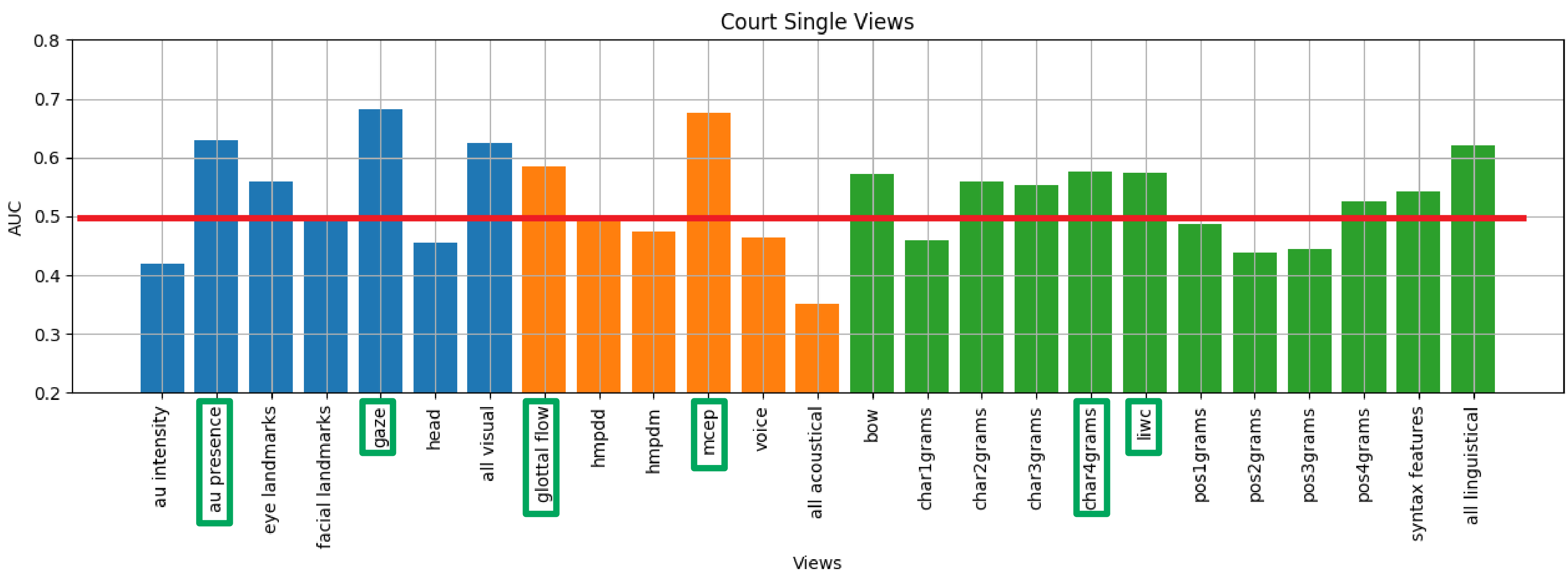

Figure 5. Results for single views/modalities in the court database.

Single Views (2)

- Spanish (Sklearn, SVC: kernel=poly, C=0.01)

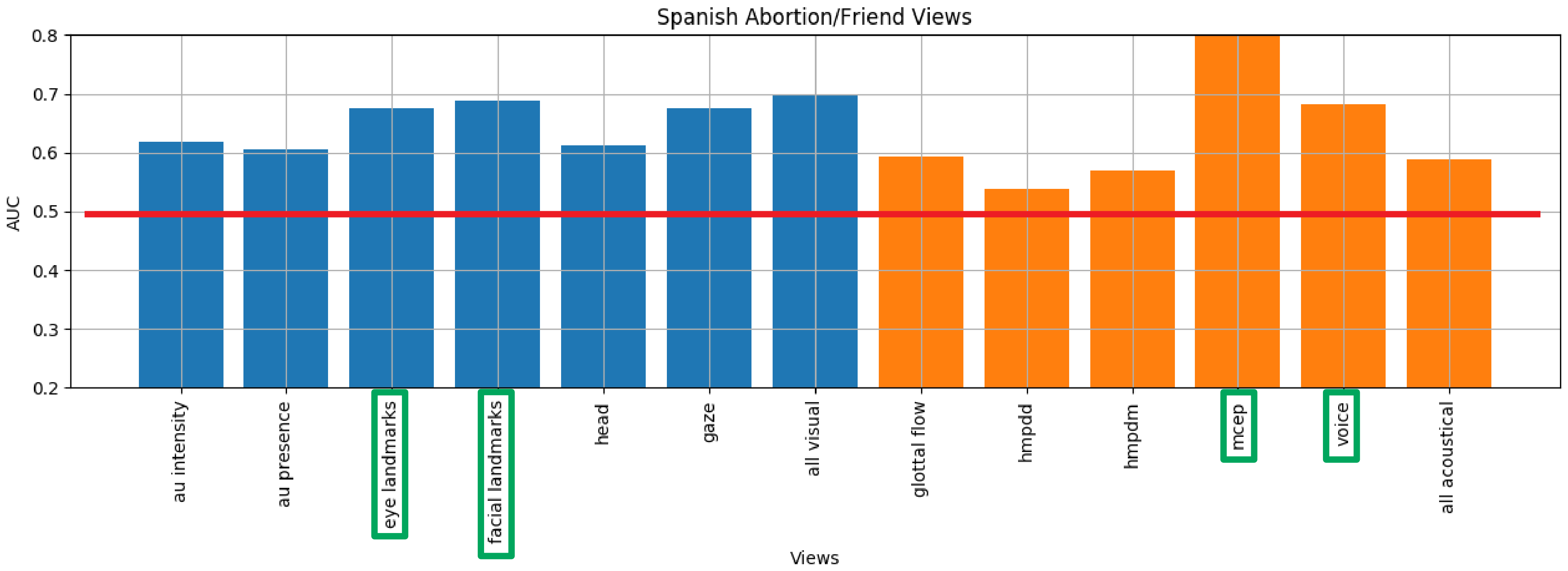

Figure 6. **Results for single views/modalities in the Spanish database.**

Complementarity

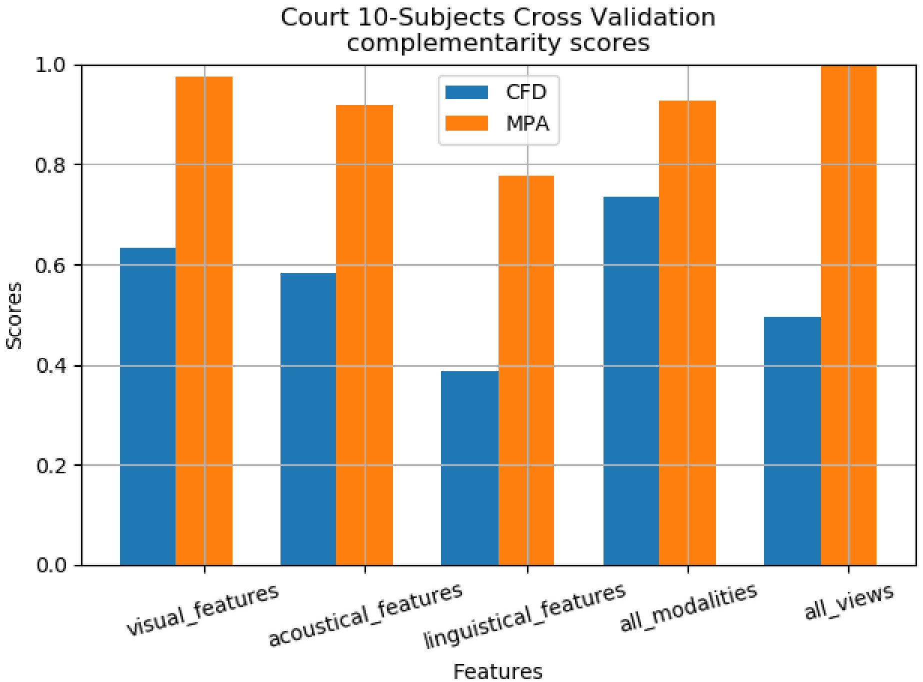

Figure 7. Complementarity measures for the court database.

There is diversity in the errors committed by each view

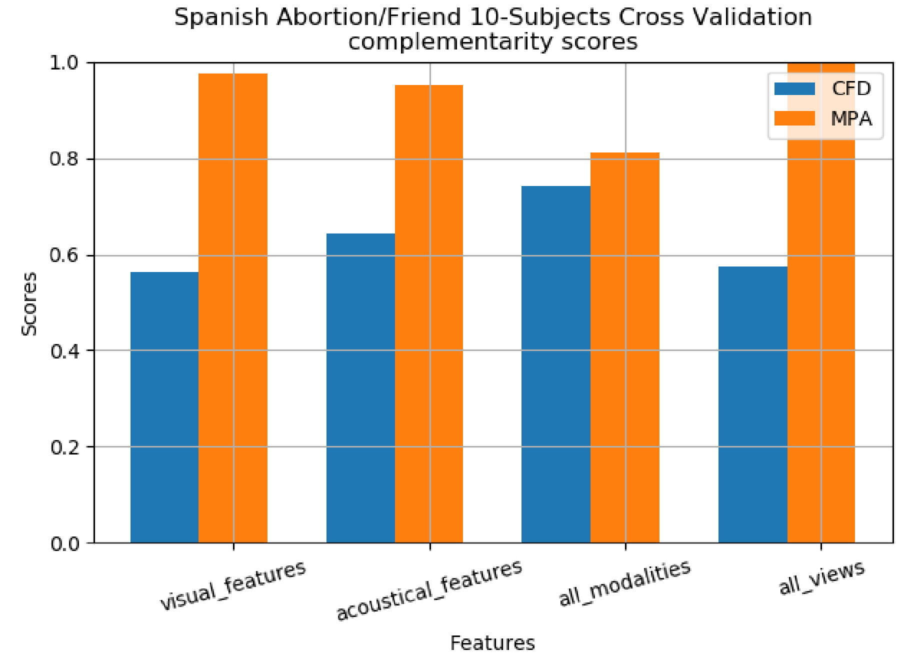

Figure 8. Complementarity measures for the Spanish database.

The correct predictions from different views predict the whole datasets

Proposed Methods (2)

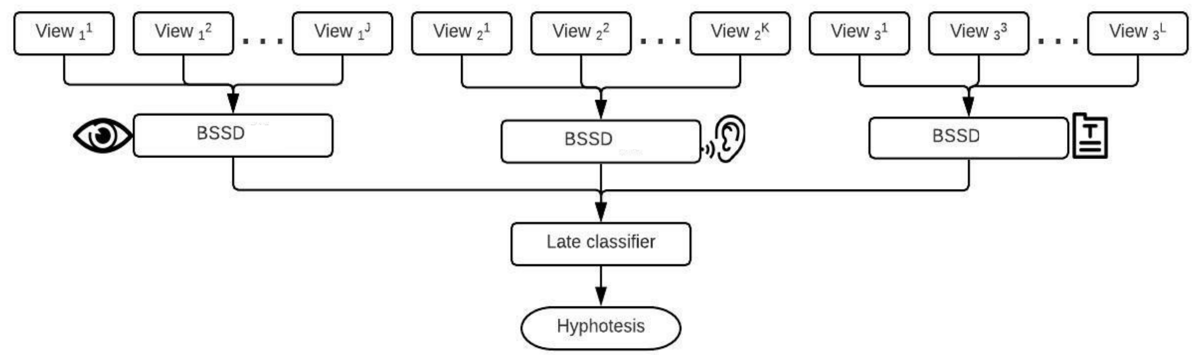

Figure 9. **Block diagram of *Hierarchical* Boosting with Shared Sampling Distribution.**

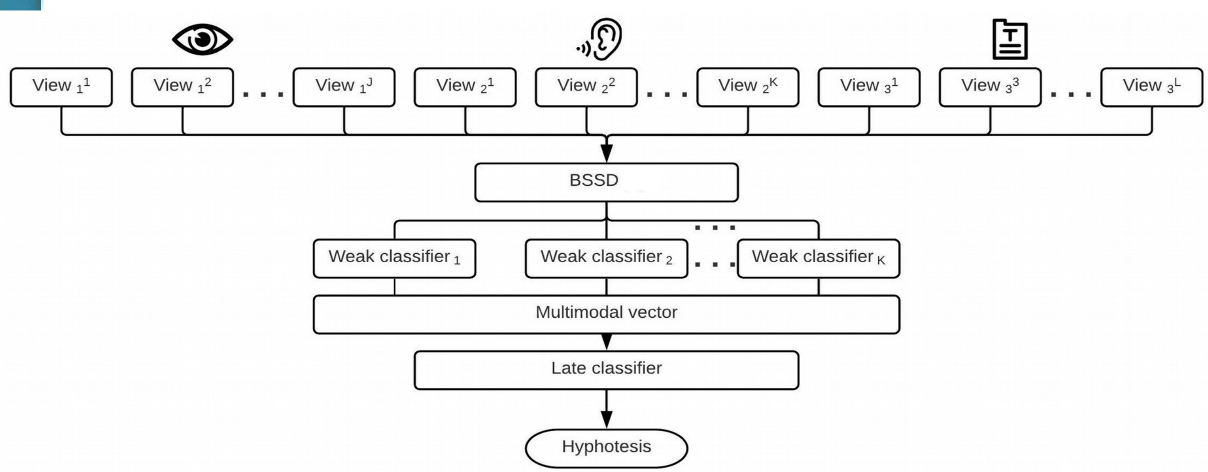

Figure 10. **Block diagram of *Stacked* Boosting with Shared Sampling Distribution.**

Algorithm 1: Boosting With Shared Sampling Distribution (BSSD) [5]

1. Input: $z_0^j = \{x_i^j, y_i)\}_{i=1}^n$, $j = 1, \cdots, M$.
2. Initialization: $W_1 = \{w_1(i) = \frac{1}{n}\}_{i=1}^n$.
3. For $k = 1$ to k_{max}
   - (a) Sample z_k^j from z_0^j using the distribution W_k.
   - (b) Compute hypothesis h_k^j from z_k^j for each view j.
   - (c) Calculate error ϵ_k^j: $\epsilon_k^j = P_{i \sim W_k}[h_k^j(x_i^j) \neq y_i]$
   - (d) If for each view: $\{\epsilon_k^j\}_{j=1}^M \leq 0.5$, select h_k^* corresponding to $\epsilon_k^* = \min_j\{\epsilon_k^j\}$.
   - (e) Calculate $\alpha_k^* = \frac{1}{2} ln(\frac{1-\epsilon_k^*}{\epsilon_k^*})$.
   - (f) Update $w_{k+1}(i) = \frac{w_k(i)}{Z_k^*} \times e^{-h_k^*(x_i^*)y_i\alpha_k^*}$, where Z_k^* is a normalizing factor.
4. Output: $F(x) = \sum_{k=1}^{k_{max}} \alpha_k^* h_k^*(x^*)$.
5. Final hypothesis: $H(x) = sign(F(x))$.

Fusion Results Court

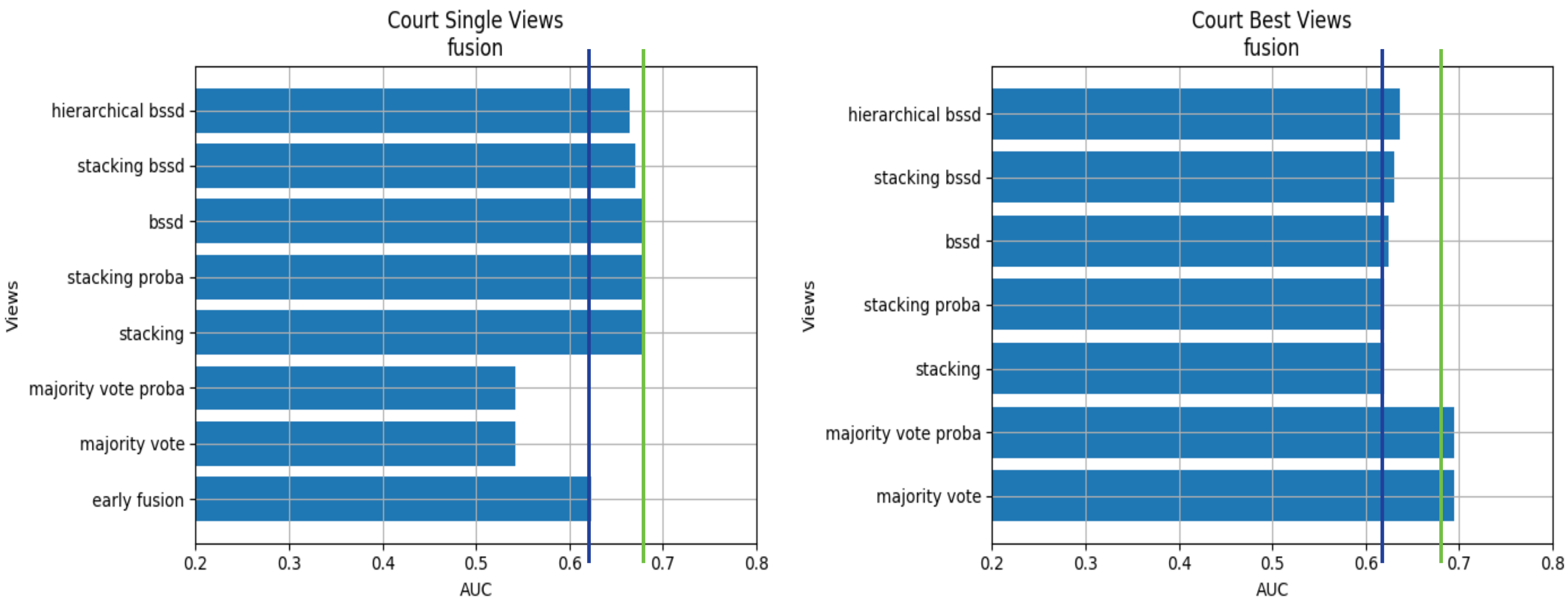

Figure 11. **Results of fusion methods using all the views (left) and the best two views per modality (right) from the court database.**

***Best view*: Gaze direction (0.683)**
***Early fusion*: 0.623**

Fusion Results Spanish

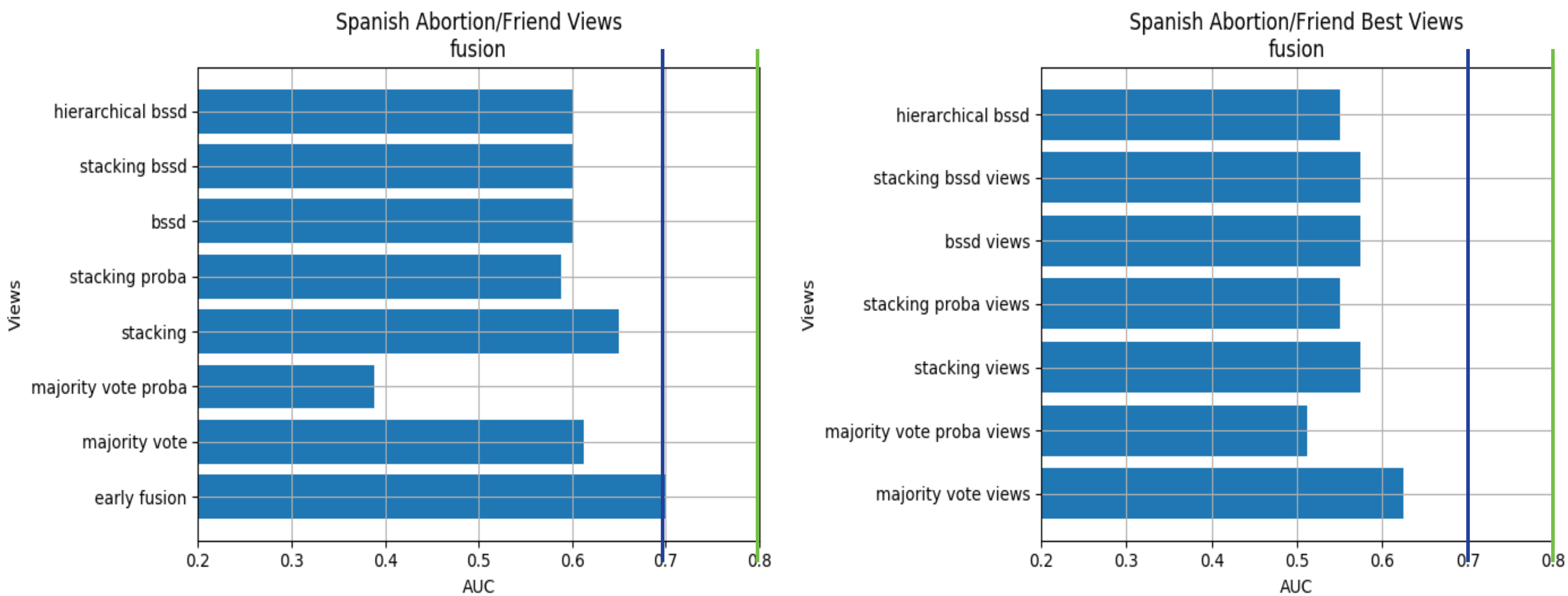

Figure 12. **Results of fusion methods using all the views (left) and the best two views per modality (right) from the Spanish database.**

Best view**: MCPE (0.856)**

Early fusion**: 0.700**

Conclusions

- Despite language, context and topic differences, there are ***views useful for deception detection*** in ***both datasets***
  - ***Action units, eye landmarks, gaze direction*** (visual)
  - ***MCEP, glottal flow*** (acoustical)
- ***Fundamental frequency and voiced/unvoiced intervals*** seem useful to ***detect deception on uninterrupted speech***
- Complementarity analysis suggest it is ***useful to fuse*** features to improve performance
  - Fusion is ***not trivial***
  - ***Alternatives to concatenating*** the multimodal features can improve the performance of a simple early fusion

Future work

- To explore ***LSTM*** networks for temporal analysis of features
- To use ***boosting methods with tuned hyperparameters*** per view
- To study pure ***NN*** approaches preserving high-level features
- To expand the ***Spanish dataset***

References

1. Bond Jr, Charles F and Bella M DePaulo (2006). “Accuracy of deception judgments”. In: Personality and social psychology Review 10.3, pp. 214–234.
2. Abouelenien, Mohamed, Verónica Pérez-Rosas, Rada Mihalcea, et al. (2017). “Detecting deceptive behavior via integration of discriminative features from multiple modalities”. In: IEEE Transactions on Information Forensics and Security 12.5, pp. 1042–1055.
3. Pérez-Rosas, Verónica et al. (2015). “Deception detection using real-life trial data”. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. ACM, pp. 59–66.
4. Wu, Zhe et al. (2018). “Deception detection in videos”. In: Thirty-Second AAAI Conference on Artificial Intelligence.
5. Barbu, Costin, Jing Peng, and Guna Seetharaman (2010). “Boosting information fusion”. In: 2010 13th International Conference on Information Fusion. IEEE, pp. 1–8.