KOÇ UNIVERSITY

Agreement and Disagreement Classification of Dyadic Interactions Using Vocal and Gestural Cues

Hossein Khaki, Elif Bozkurt, Engin Erzin
Multimedia, Vision and Graphics Lab (MVGL)
Department of Electrical and Electronics Engineering

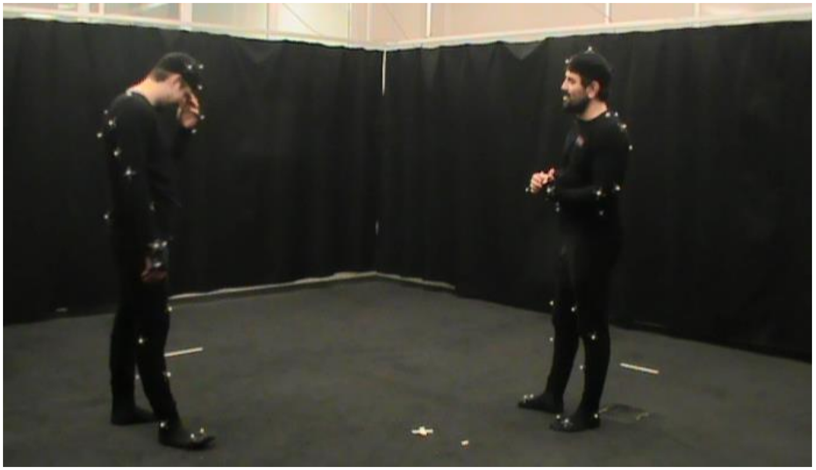
41st IEEE International Conference on Acoustics, Speech and Signal Processing
ICASSP 2016
20-25 March 2016 – Shanghai, China

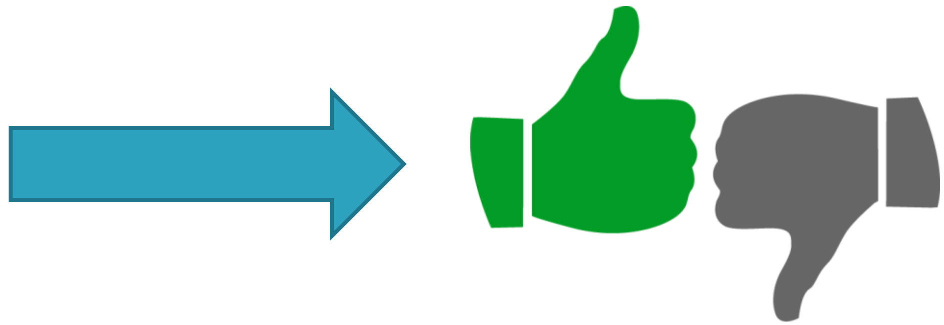
Koç University
Istanbul, Turkey

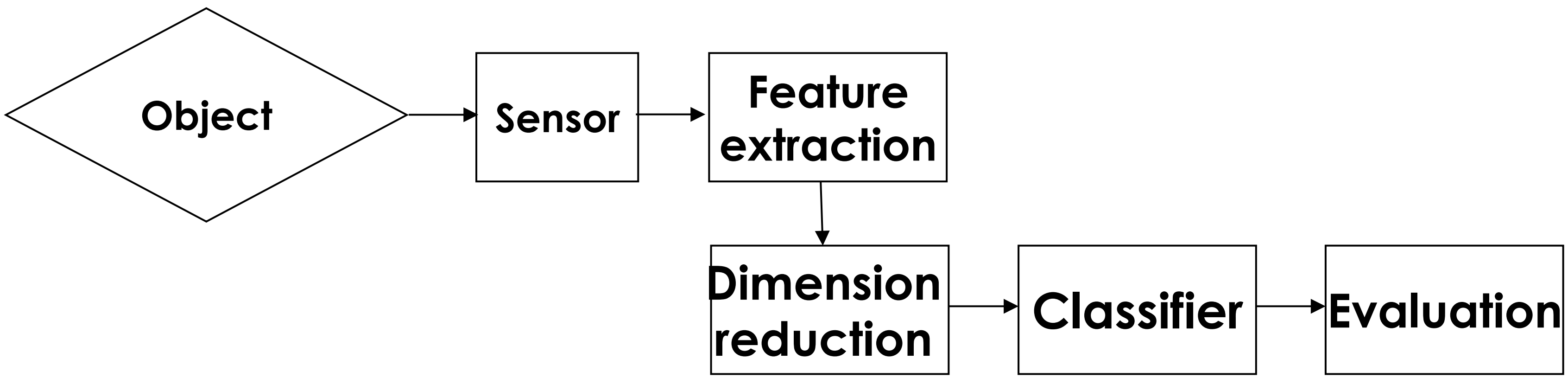
Outline

- Problem Definition
- JESTKOD database
- Agreement/Disagreement Classification
- Experimental Evaluations
- Conclusions

Problem Definition





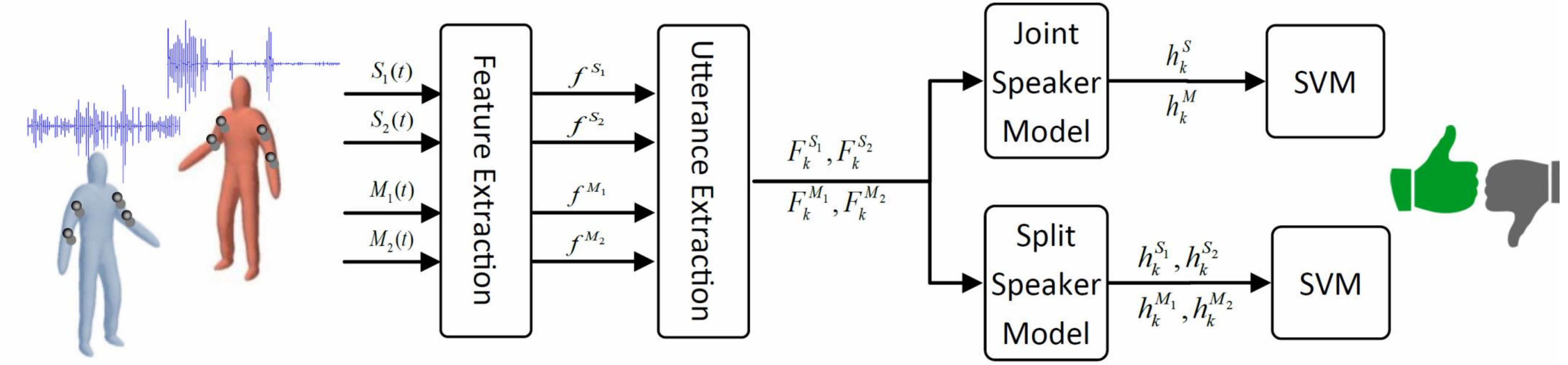


JESTKOD database

- A natural and affective dyadic interactions
- Equipment:
  - A high-definition video recorder
  - Full body motion capture system with 120 fps
  - Individual audio recorders
- 5 sessions, totally 66 agree and 79 disagree clips
- In each clips: 2 participants, around 2~4 minutes
- Totally 10 participants
  - 4 female/6 male, ages: 20 - 25
- Language: Turkish
- Annotation (Not used in this paper)
  - Activation
  - Valence
  - Dominance

Pair #	Topics in the JESTKOD database			
	Agreement scenario	Num. clips	Disagreement scenario	Num. clips
1	Cinema, World cuisine, Holiday resorts, TV series	13	Football, Maths, Game consoles, PC Games	13
2	Football, World cuisine, Music, Cinema, Literature	13	Geography, Holiday resorts, PC Games, Theatre, Dance	16
3	Cinema, Sports, PC Games, Music, World cuisine	11	Cinema History, TV series, Animals, Education	17
4	World cuisine, Holiday resorts, Science-fiction, History, Theatre, Cities	16	Football, Cinema PC Games, TV series, Literature, Physics	17
5	Cinema, Languages, PC Games, Cities, Game consoles	13	Cinema, Sports, Holiday resorts, Nutrition, Musicals	16
Total		66		79

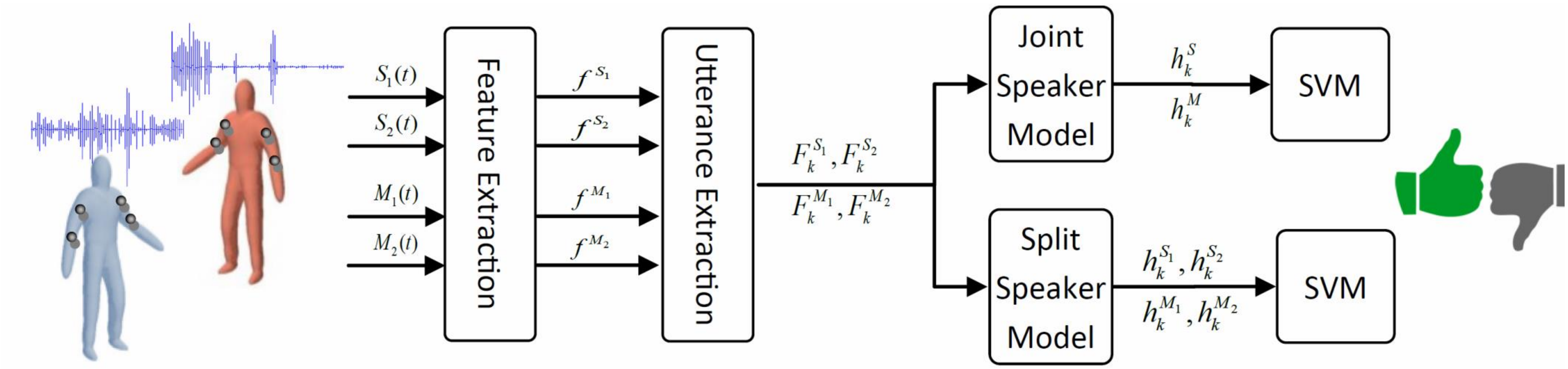
Agreement/Disagreement Classification



- A two-class dyadic interaction type (DIT) estimation problem
- **Input:** speech and motion modalities of two participants
- **Feature Extraction:**
  - Speech: 20 ms win with 10 ms frame shifts $\Rightarrow f^{S_i}$: $39D = 13MFCCs + \Delta + \Delta\Delta$
  - Motion: f^{M_i}: $24D = (\varphi, \theta, \psi)$ of the arm & forearm joints with their derivatives

$i = 1,2.$
Index of two participants.

Agreement/Disagreement Classification



- **Utterance Extraction:** collect frame level feature vectors over the temporal duration of the utterance and construct matrices of features
  - **Speech**: only vocal frames, $\mathrm{F}_{\mathrm{k}}^{\mathrm{S_i}} = \left[f_1^{\mathrm{S_i}}, \dots, f_{N_S}^{S_i}\right]$
  - **Motion**: All frames, $\mathrm{F}_{\mathrm{k}}^{\mathrm{M_i}} = \left[f_1^{\mathrm{M_i}}, \dots, f_{N_S}^{M_i}\right]$

i = 1,2.
Index of two participants.

Agreement/Disagreement Classification (Cont.)

$$\begin{bmatrix} f_{11} & \cdots & f_{1n} \\ \vdots & \ddots & \vdots \\ f_{m1} & \cdots & f_{mn} \end{bmatrix} \rightarrow \boxed{\text{Feature Summarizer}} \rightarrow [h_1 \quad \ldots \quad h_r]$$

matrices of features: F

Summarized vector: $\underline{h}$

- Two Feature Summarization techniques
  - Using statistical functions followed by PCA [1]
    - mean, standard deviation, median, minimum, maximum, range, skewness, kurtosis, the lower and upper quantiles and the interquantile range.
  - Using i-vector representation in total variability space (TVS) [2]
    - GMM models followed by Factor Analysis

[1]- A. Metallinou, A. Katsamanis, and S. Narayanan, “Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information,” Image and Vision Computing, vol. 31, no. 2, pp. 137– 152, 2013.

[2]- H. Khaki and E. Erzin, “Continuous emotion tracking using total variability space,” in Sixteenth Annual Con. of the International Speech Communication Association, 2015.

Agreement/Disagreement Classification (Cont.)

- Dyadic modeling:
  - Joint Speaker Model (JSM)

$$\begin{bmatrix} \mathrm{F}_{\mathrm{k}}^{S/M_1} \\ \mathrm{F}_{\mathrm{k}}^{S/M_2} \end{bmatrix} \longrightarrow \boxed{\text{Feature Summarizer}} \longrightarrow \underline{h}_k^{S/M}$$

  - Split Speaker Model (SSM)

$$\begin{bmatrix} \mathrm{F}_{\mathrm{k}}^{S/M_1} & \mathrm{F}_{\mathrm{k}}^{S/M_2} \end{bmatrix} \longrightarrow \boxed{\text{Feature Summarizer}} \longrightarrow \begin{bmatrix} \underline{h}_k^{S/M_1} & \underline{h}_k^{S/M_2} \end{bmatrix}$$

- Support Vector Machine

	Speech	Motion	Multimodal
JSM	$SVM(h^S)$	$SVM(h^M)$	$SVM(h^S, h^M)$
SSM	$SVM(h^{S_1}, h^{S_2})$	$SVM(h^{M_1}, h^{M_2})$	$SVM(h^{S_1}, h^{S_2}, h^{M_1}, h^{M_2})$

* SVM(h): A notation to describe an SVM classifier using feature vector h.

Experimental Evaluations (parameters)

- **Training and testing strategy:** Leave-one-clip-out
- **Feature Summarizer:**
  - **statistical functions:** Adjust the PCA output dimension to preserve 90% of the total variance
  - **i-vector:** 128 GMM for TVS and 30 dimensional i-vector.
- **SVM:** Linear kernel from LibSVM package.
- **Performance metric:** The average of classification accuracy
- **Chance level recognition rate:** 49.99%
- **Two levels of evaluation:**
  - **Clip level:** decision over a whole clip
  - **Utterance level:** decision over a couple of seconds of a clip

Experimental Evaluations (clip level)

- Unimodal and multimodal classification accuracy for clip level DIT estimation
  - **Lowest accuracy**: Motion
  - i-vector inappropriate for motion compare to statistical functions.

Method	Accuracy
JSM: i-vector(Motion)	55.74%
JSM: i-vector(Speech)	99.18%
JSM: i-vector(Speech+Motion)	98.36%
SSM: i-vector(Motion)	57.38%
SSM: i-vector(Speech)	85.25%
SSM: i-vector(Speech+Motion)	86.89%
JSM: statistics(Motion)	82.79%
JSM: statistics(Speech)	83.61%
JSM: statistics(Speech+Motion)	86.07%
SSM: statistics(Motion)	79.51%
SSM: statistics(Speech)	89.34%
SSM: statistics(Speech+Motion)	90.16%

Experimental Evaluations (clip level)

- Unimodal and multimodal classification accuracy for clip level DIT estimation
  - **Lowest accuracy**: Motion
  - i-vector inappropriate for motion compare to statistical functions.
  - Speech modality outperforms motion modality
  - Low performance:
    - SSM + i-vector
    - JSM + Statistical functions
  - High performance:
    - JSM + i-vector
    - SSM + Statistical functions

Method	Accuracy
JSM: i-vector(Motion)	55.74%
JSM: i-vector(Speech)	99.18%
JSM: i-vector(Speech+Motion)	98.36%
SSM: i-vector(Motion)	57.38%
SSM: i-vector(Speech)	85.25%
SSM: i-vector(Speech+Motion)	86.89%
JSM: statistics(Motion)	82.79%
JSM: statistics(Speech)	83.61%
JSM: statistics(Speech+Motion)	86.07%
SSM: statistics(Motion)	79.51%
SSM: statistics(Speech)	89.34%
SSM: statistics(Speech+Motion)	90.16%

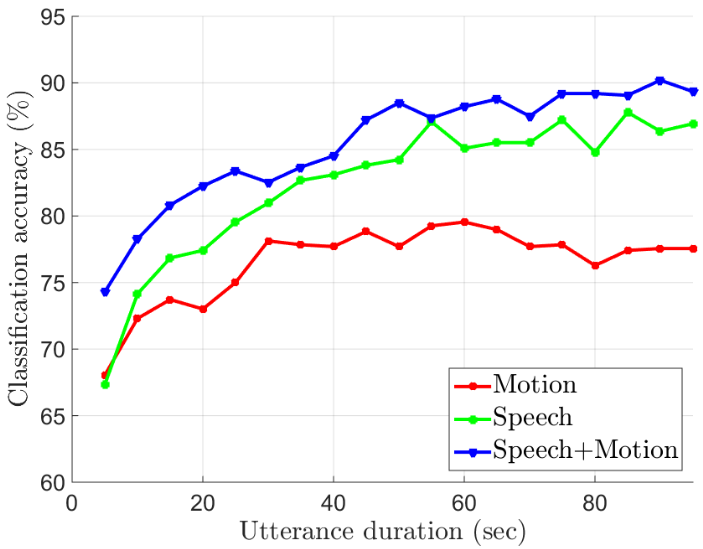
Experimental Evaluations (clip level)

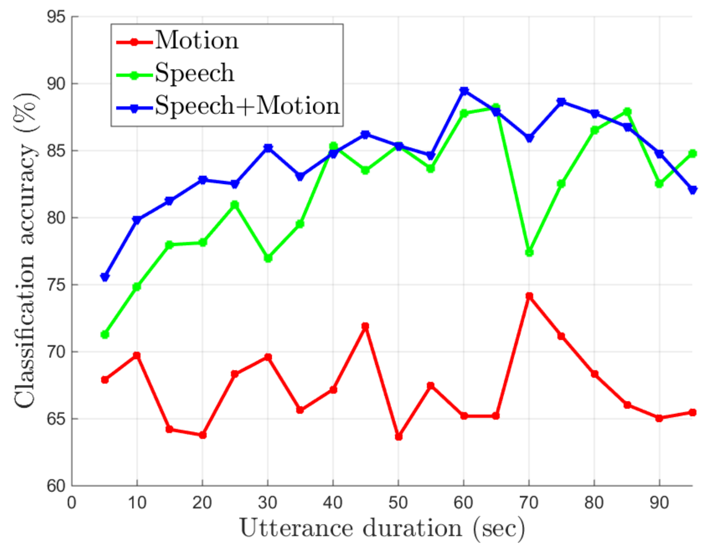
- Unimodal and multimodal classification accuracy for clip level DIT estimation
  - **Lowest accuracy**: Motion
  - i-vector inappropriate for motion compare to statistical functions.
  - Speech modality outperforms motion modality
  - **Highest accuracy**: The multimodal scenarios except JSM + i-vector!
  - Low performance:
    - SSM + i-vector
    - JSM + Statistical functions
  - High performance:
    - JSM + i-vector
    - SSM + Statistical functions

Method	Accuracy
JSM: i-vector(Motion)	55.74%
JSM: i-vector(Speech)	99.18%
JSM: i-vector(Speech+Motion)	98.36%
SSM: i-vector(Motion)	57.38%
SSM: i-vector(Speech)	85.25%
SSM: i-vector(Speech+Motion)	86.89%
JSM: statistics(Motion)	82.79%
JSM: statistics(Speech)	83.61%
JSM: statistics(Speech+Motion)	86.07%
SSM: statistics(Motion)	79.51%
SSM: statistics(Speech)	89.34%
SSM: statistics(Speech+Motion)	90.16%

Experimental Evaluations (utterance level)

- DIT estimation for overlapping utterances:
  - SSM with statistical functions
  - JSM with i-vector





✓ Multimodal has the highest performance for short utterances
✓Duration >15 sec ➔ Multimodal accuracy > 80%
✓Speech and Multimodal have similar curves.
✓Motion is not reliable with JSM+i-vector

*The duration is the total time of dyadic interaction, including silent and speech segments.

Conclusion

- **JESTKOD as A natural and affective dyadic interactions**
  - JESTKOD: A multimodal database of speech, motion capture and video recordings of affective dyadic interactions
- **Early results on the two-class dyadic interaction type detection**
  - Joint and split speaker model to estimate the dyadic interaction type
  - Accuracy of speech features > Accuracy of motion features
  - The multimodal has the highest accuracy over the short utterances.
- **Future works:**
  - Studding the relationship between the AVD and DIT
  - Using JESTKOD as a rich database for emotion recognition and synthesis

Thanks.

!?QUESTIONS?!

For more questions, please, contact to mail: **hkhaki13@ku.edu.tr**

This work is supported by TÜBİTAK under Grant Number 113E102.

i-vector Extraction

First a GMM models the data distribution:

$$P(\mathcal{D}) = \sum_{i=1}^{M} \omega_i \mathcal{N}\left(\mathcal{D}; \underline{\mu_i}, \boldsymbol{\Sigma_i}\right)$$

- $\mathcal{D}$: The speech feature space
- ω_i, $\underline{\mu_\mathrm{i}}$, and Σ_i: The weight, mean vector, and covariance matrix of the i'th Gaussian mixture
- M: The total number of mixtures

Then Factor Analysis reduces the dimension:

$$\mu = m + Tw,$$

- $\mu = \left[\underline{\mu_1}^T, \underline{\mu_2}^T, \dots, \underline{\mu_M}^T\right]^T$: The super-vector
- m: The Universal Background Model (UBM),
- T: The TVS basis,
- w: The reduced feature known as i-vector