AUTOMATIC IDENTIFICATION OF SPEAKERS USING HEAD GESTURES IN A NARRATION

Sanjeev Kadagathur Vadiraj, Achuth Rao M V, Prasanta Kumar Ghosh

SPIRE LAB
Electrical Engineering,
Indian Institute of Science (IISc), Bangalore, India

Lecture Sesssion: WE2.L3.5: Multimodal Processing of Language
Wednesday, 6 May, 2020, 11:30-13:30

Plan

1. Introduction
2. Data
3. Proposed approach
4. Results
5. Conclusion

Importance of Head Gestures

- Specific information - Nod and Shake [1]
- Intensity of speech [2]
- Better Comprehension in speech or dialogue [3]
- Convey emotional state [4]

[1] Tanya Stivers, "Stance, alignment, and affiliation during story telling: When nodding is a token of affiliation,"Research on language and social interaction, vol. 41, no. 1, pp. 31–57,2008.

[2] Paul Ekman and Wallace V Friesen, "Head and body cues in the judgment of emotion: A reformulation,"Perceptual and motor skills, vol. 24, no. 3 PT 1, pp. 711–724, 1967.

[3] Kevin G Munhall, Jeffery A Jones, Daniel E Callan, Takaaki Kuratate, and Eric Vatikiotis-Bateson, "Visual prosody and speech intelligibility: Head movement improves auditory speech perception,"Psychological science, vol. 15, no. 2, pp.133–137, 2004.

[4] Hatice Gunes and Maja Pantic, "Dimensional emotion prediction from spontaneous head gestures for interaction with sensitive artificial listeners," in International conference on intelligent virtual agents. Springer, 2010, pp. 371–377

Uniqueness of head gestures

- Personality traits and temperament [5]
- Music induced movement [6]
- Hill and Johnston experiment [7]
  - 4 subjects, 4 recordings each
  - 4 buckets - Score from 0 to 48 - (Number of actors X Number of Recordings X Other recordings in same bin)
  - Score of 25 on average $\sim$ Accuracy of 0.5208

[5] Anne Campbell and J Philippe Rushton, "Bodily communication and personality,"British Journal of Social and Clinical Psychology, vol. 17, no. 1, pp. 31–36, 1978.

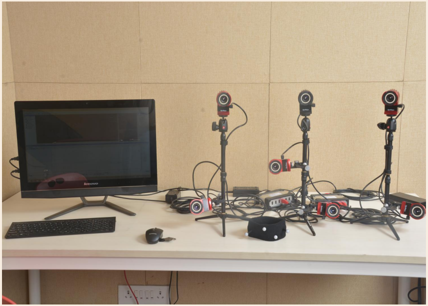
[6] Geoff Luck, Suvi Saarikallio, and Petri Toiviainen, "Personality traits correlate with characteristics of music-induced movement," in ESCOM 2009: 7th Triennial Conference of European Society for the Cognitive Sciences of Music, 2009.

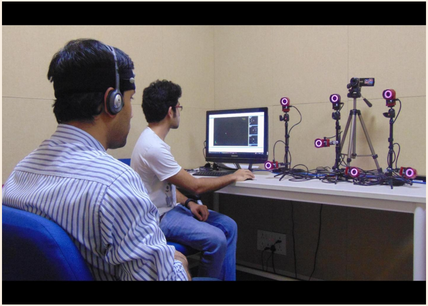
[7] Harold Hill and Alan Johnston, "Categorizing sex and identity from the biological motion of faces," Current biology, vol. 11, no. 11, pp.880–885, 2001.

Applications of Head Gestures

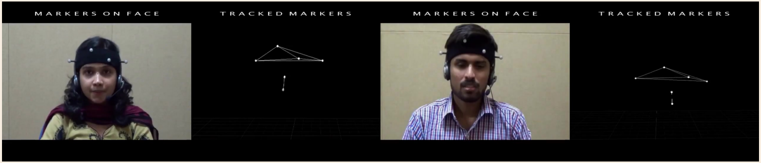
- Infer emotional state of the speaker
- Conversational robots interact better
- Forensics - Blurry or masked videos

Recording Setup

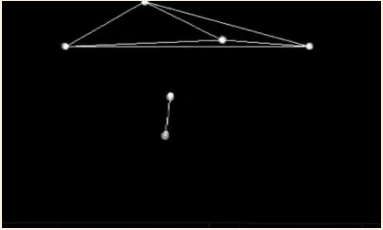


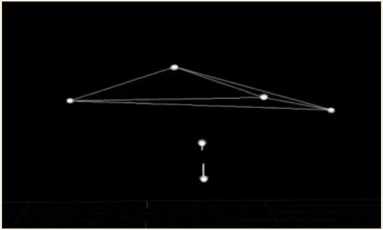


Head Motion


MARKERS ON FACE
TRACKED MARKERS
MARKERS ON FACE
TRACKED MARKERS

Head Motion + Audio





Plan

1 Introduction

2 Data

3 Proposed approach

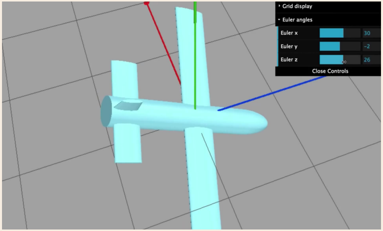
4 Results

5 Conclusion

Euler Angles

Euler angles are the three angles used to describe the orientation of a rigid body with respect to a fixed coordinate system

[8]



Grid display

Euler angles

Euler x 30

Euler y -2

Euler z 26

[8] Gaurav Fotedar and Prasanta Kumar Ghosh, "An information theoretic analysis of the temporal synchrony between head gestures and prosodic patterns in spontaneous speech.," in INTERSPEECH,2017, pp. 157–161.

Data description

Story	E1	N1	E2	N2	E3	N3	E4	N4	E5	N5
Mean	235	232	204	200	231	246	245	250	267	262
Std	67	64	82	87	76	83	71	90	74	113
Min	102	142	79	78	112	144	120	139	123	125
Max	410	38	508	511	507	542	438	552	479	668

Table: Duration of recordings in the dataset (s)

- Speakers give varied level of detail while narrating a story
- Mean normalization of Euler angles

Plan

1 Introduction

2 Data

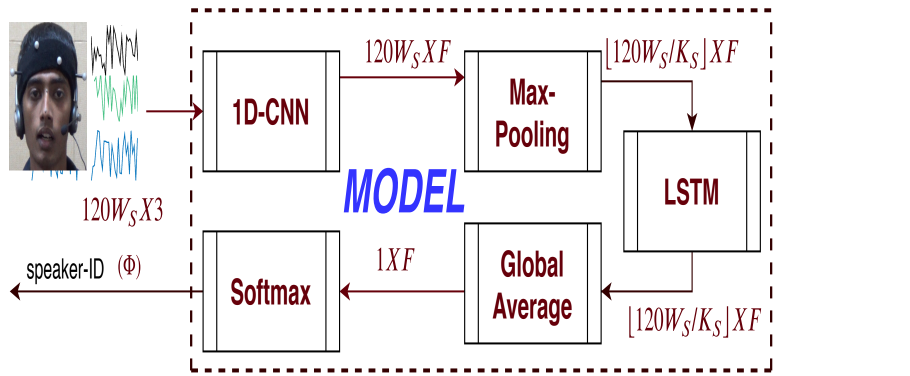
3 Proposed approach

4 Results

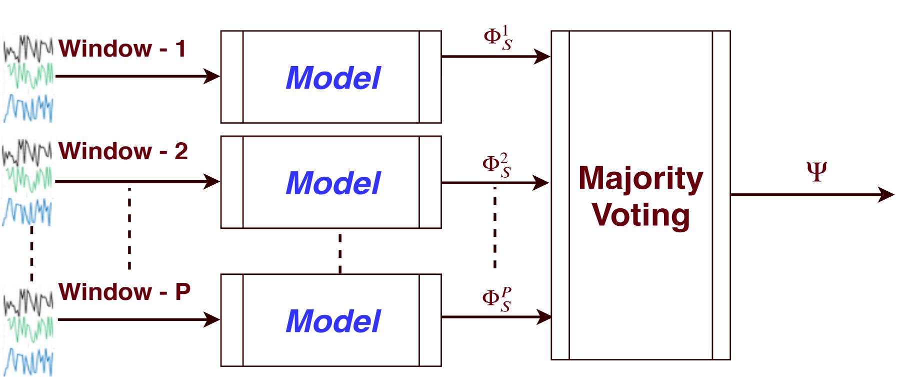
5 Conclusion

Architecture

W_S : *Duration*, 120 : *Sampling rate*, F : *CNN Filters*, K_S : *Max pooling size*, $Optimizer$: Adam, $Loss$: Categorical cross entropy, $Batch$: 10



Voting


Window - 1
Window - 2
Window - P
Model
Model
Model
Φ_S^1
Φ_S^2
Φ_S^P
Majority Voting
Ψ

Plan

1 Introduction

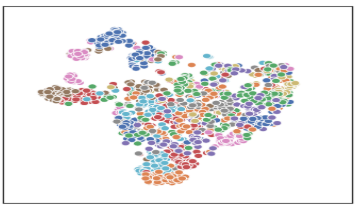
2 Data

3 Proposed approach

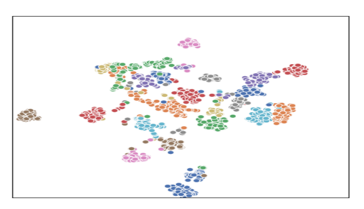
4 Results

5 Conclusion

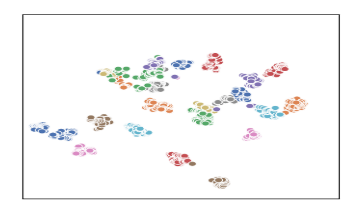
t-SNE



(a) Head Gestures — 5s

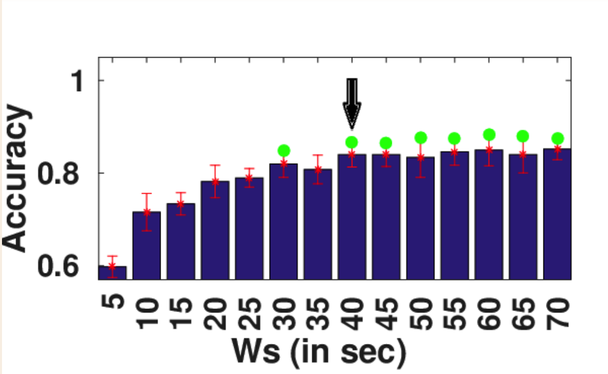


(b) Head Gestures — 20s

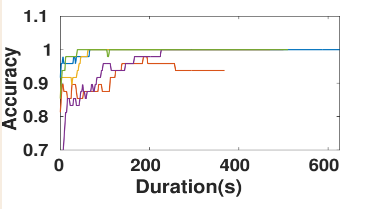


(c) Head Gestures — 40s

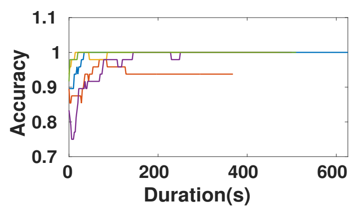
Speaker identification performance



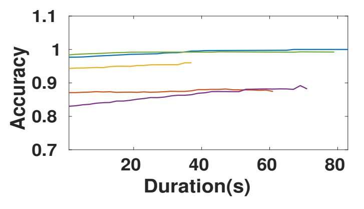
Location Specific and Location Independent Analysis



Location Specific Analysis — Beginning



(b) Location Specific Analysis — End



(c) Location Independent Analysis

Speech Vs Head Gestures

Head gestures	Fold	0	1	2	3	4	Avg
	ValAcc	0.83	0.81	0.85	0.88	0.81	0.836
	TestAcc	0.96	0.86	0.94	0.83	0.96	0.91
Audio	ValAcc	0.93	0.97	0.98	0.98	0.99	0.970
	TestAcc	0.93	0.99	0.99	0.99	0.99	0.978

Table: Fold wise speaker identification accuracy using head gestures over 40s duration and audio over 3.2s duration.

Plan

1 Introduction

2 Data

3 Proposed approach

4 Results

5 Conclusion

Key Takeaways

- Speaker specific information is encoded in natural head gestures - Average accuracy of 83.6% across windows of 40s duration
- Longer sequences are better at identifying

Future Works

- Run the model on well-established datasets like IEMOCAP
- Collect more data for native languages and see if there are any language-specific patterns in head gestures
- Analyze the correlation between head gestures and speech

References

- Harold Hill and Alan Johnston, “Categorizing sex and identity from the biological motion of faces,” Current biology, vol. 11, no. 11, pp. 880–885, 2001.
- Gaurav Fotedar and Prasanta Kumar Ghosh, “An information theoretic analysis of the temporal synchrony between head ges- tures and prosodic patterns in spontaneous speech.,” in INTER- SPEECH, 2017, pp. 157–161.
- CA Valliappan, Anurag Das, and Prasanta Kumar Ghosh, “Classification of story-telling and poem recitation using head gesture of the talker,” in 2018 International Conference on Sig- nal Processing and Communications (SPCOM). IEEE, 2018, pp. 36–40.

Acknowledgement

We thank the Department of Science and Technology, Govt. of India for their support in this work.

THANK YOU

Have Questions/Suggestions?
Write to us at spirelab.ee@iisc.ac.in