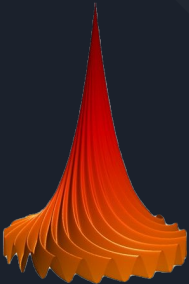




Head Motion Dataset captured Motion Capture (MoCap) Device Optitrack

- Valliappan CA
SPIRE Lab
Electrical Engineering Department
Indian Institute of Science



Optitrack (MoCap Device)

1. [Optitrack](#) is an industry leader providing various MoCap and 3D tracking products.
2. The device is capable of tracking the 3D positions of several reflective markers attached to a subject.
3. The reflective markers can be attached to custom positions on the face or according to some pre-specified Optitrack templates.
4. The device tracks the 3D coordinates that is the $\{x, y, z\}$ position of each marker at 120 frames per second.
5. The product specialized for facial motion capture consists of 7 Flex IR cameras system and the [Arena software](#).



Illustration of recorded data using Optitrack



- The left half in each video shows the actual recording of the subject and the right half of each video displays the corresponding motion of the markers in 3D.

Recorded data embedded on 3D face



- The left half of the video shows the actual video of the subject while recording and the right half shows the corresponding motion rendered on the 3D face.
- Autodesk 3ds max software is used to render the 3D coordinates on the realistic face.

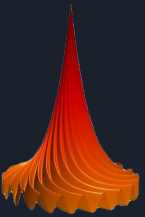


Head motion Dataset

Objective :

- To Develop a database that consisting of Audio Visual data using the 3D Head Motion.
- This technology has a wide variety of applications, from 3D animation for movies to gesture analysis.

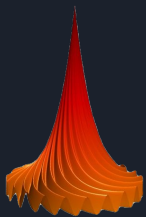
Database :


- The database consist of simultaneously recorded speech and head gestures from multiple subjects during both **storytelling** (spontaneous speech) and **poem recitation** (rhythmic speech).
 - The storytelling data consist of stories recited by subjects in their native language as well as English. Whereas the poem recitation was restricted to only English.
 - Subjects were uniformly distributed (2 male and 2 female) for each of the native languages {**Hindi, Bengali, Malayalam, Kannada, Tamil and Telugu**}.
 - The head gesture data for poem recitation comprises of 22 subjects (12 male and 10 female) and storytelling comprises of 24 subjects (12 male and 12 female).
- 



Research Works

- 1) **“Classification of story-telling and poem recitation using head gesture of the talker”**, accepted in International Conference on Signal Processing and Communications (SPCOM) 2018.
- 2) **“An information theoretic analysis of the temporal synchrony between head gestures and prosodic patterns in spontaneous speech”**, Proc. Interspeech 2017, 157-161.





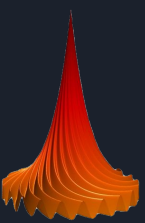
Classification of story-telling and poem recitation using head gesture of the talker


Abstract :

This work investigate the nature of head gestures in spontaneous speech during story-telling in comparison to that in poem recitation. It was hypothesize that head gestures during poem recitation would be more repetitive and structured compared to those in case of spontaneous speech. To quantify this, we proposed a measure called degree of repetition (DoR). We also perform a story-telling vs poem recitation classification experiment using deep neural network (DNN).

Results :

The classification experiment between storytelling and poem recitation using DNN demonstrates that the raw head gestures result in an average classification accuracy of 85.79% while the DoR results in an average accuracy 80.59%, indicating that the features learnt by DNN from raw head gestures are more discriminative than DoR features. While these accuracy are less than those (94.67%) obtained using acoustic feature such as MFCCs, raw head gestures and MFCCs together yield a higher average accuracy (98.62%), indicating that the head gestures are complementary to the acoustic features for the classification task.





An information theoretic analysis of the temporal synchrony between head gestures and prosodic patterns in spontaneous speech

Abstract :

The paper analyzes the temporal coordination between head gestures and prosodic patterns in spontaneous speech in a data-driven manner. For this study, head motion and speech data were used from 24 subjects while they tell a fixed set of five stories. The head motion, captured using a motion capture system, is converted to Euler angles and translations in X, Y and Z-directions to represent head gestures. Pitch and short-time energy in voiced segments are used to represent the prosodic patterns. To capture the statistical relationship between head gestures and prosodic patterns, mutual information (MI) is computed at various delays between the two using data from 24 subjects in six native languages.

Results :

The estimated MI, averaged across all subjects, is found to be maximum when the head gestures lag the prosodic patterns by 30msec. This is found to be true when subjects tell stories in English as well as in their native language. We observe a similar pattern in the root mean squared error of predicting head gestures from prosodic patterns using Gaussian mixture model.

