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

C.A.Valliappan, Anurag Das, Prasanta Kumar Ghosh
Department of Electrical Engineering, Indian Institute of Science Bangalore, 560012, India.
Email :{valliappanc, anuragd, prasantg}@iisc.ac.in

***Abstract*—In this work, we investigate the nature of head gestures in spontaneous speech during story-telling in comparison to that in poem recitation. We 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). For the classification, both DoR as well as context dependent raw head gesture data are used. Analysis and experiments are performed using a database of 24 subjects each telling five stories and a different set of 10 subjects each reciting 20 poems, three times each, thus having data of comparable durations for story telling and poem recitation. Analysis of head gestures using DoR reveals that the DoR, on average, is higher during poem recitation compared to that during story-telling. A four-fold classification experiment between story-telling and poem recitation using DNN demonstrates that the raw head gestures result in an average classification accuracy of 85.79% and an average F-score of 89.05% while the DoR results in an average accuracy and F-score of 80.59% and 82.30% respectively indicating that the features learnt by DNN from raw head gestures are more discriminative than DoR features. While these accuracy and F-score are less than those (94.67% & 95.60%) obtained using acoustic feature such as Mel frequency cepstral coefficients (MFCCs), raw head gestures and MFCCs together yield a higher average accuracy (98.62%) and F-score (98.92%), indicating that the head gestures are complementary to the acoustic features for the classification task.**

I. Introduction

Head gestures accompanying speech occur naturally during face-to-face communication. While speech carries linguistic information from talker to listener, talker often use head gestures as well as hand gestures in communication to convey several para-linguistic factors, e.g., agreement/disagreement, critical for face-to-face interaction [1]. Use of such gestures, in turn, makes the interaction engaging and helps the listener to understand talker's message better. Perceptually, head gestures have been shown to improve the understanding of speech [2]. Head gestures have also been shown to be closely related to speech. For example, distinct head gestures like side-to-side nodding correlates expressions of inclusivity and uncertainity [3]. Head gestures have also been shown to be correlated to the fundamental frequency F0 [4]. F0 based features have also been used to synthesize head gestures [5] [6] [7].

Head gestures not only occur in face-to-face interaction, but also during monologue such as giving lecture, story telling [8], poem recitation. Several factors in speech could determine the naturally occurring head gestures while speaking. These factors include the content of what is being spoken, the emotional state of the talker, modes of speaking, and response from the listeners [9]. There have been a number of works in the literature to determine speaker's emotional and mental states as well as gender based on head gestures. For example, head gesture parameters have been shown to discriminate between different emotions as reported by Busso et al. [10]. Similarly, head gestures have been used to infer complex mental states of the speaker from a video stream [1]. Head gestures have also been shown to identify speakers and their gender [11]. However, the degree to which head gestures may differ across various modes of speech has not been investigated well. We, in this work, focus on the head gestures during spontaneous speech in story telling and rhythmic speech in poem recitation. A better understanding of the head gesture style in different modes of speech could help develop models for head gesture synthesis for different styles of speech. This could also be useful to identify modes of speech in a human-computer communication, which could further be used for determining the speaking style and associated head gesture of the computer agent or avatar.

Spontaneous speech, particularly in story telling contains both intentional and unintentional pauses as well as hesitation, repetition of phrases [12][13][14]. Story telling also involves exaggeration and stressing at appropriate instants as well as change of intonation to make the story lively. All these aspects in story telling modulate the head motion in different ways at various points in time during story telling. On the other hand, poem recitation requires the subject to remember the poem and recite in a manner that makes the recitation sounds well. This, in turn, requires giving pauses, stress and having prosodic patterns that match well with the poem. Thus the variation in the long term acoustic events is expected to be more structured in poem recitation than in story telling. Poems are also known to be primarily rhythmic [15]. Thus the speech pattern, particularly the stress and intonation, also repeats following the rhythm in poem. Thus, the head gesture patterns during recitation could be influenced by the structured and rhythmic nature in poem. We, in this work, quantify the head gesture patterns to analyze the degree to which they could be different during story telling and poem recitation. We also perform a head gesture based classification to quantify how representative the head gestures are for story telling and poem

978-1-5386-3821-7/18/$31.00 © 2018 IEEE

recitation.

We propose a measure, called degree of repetition (DoR) to capture the rhythmic nature in the head gesture, if any. This is computed over a long term analysis window. Interestingly, DoR was found to be more in the head gesture during poem recitation compared to that during story telling suggesting a structured and rhythmic nature of head motion during poem recitation. To examine the discriminative power of the head gestures to classify a given recording to story telling or poem recitation, we perform classification experiments with DoR as well as context dependent raw head gesture data recorded using motion capture system from 24 subjects each telling an identical set of five stories and a different set of 10 subjects each reciting 20 poems, each repeated three times. The classification experiments are done in four fold with no overlap between training and test subjects in every fold. The classification accuracy averaged across all folds is found to be 80.59% and 85.69% when DoR and raw head gesture values are used for classification task respectively. However, this is found to be lower than the classification accuracy of 94.67% obtained by acoustic feature, namely Mel frequency cepstral coefficients (MFCCs). Interestingly, when MFCC and raw head gestures are combined the classification accuracy increases to 98.62% suggesting that head gestures, while inferior to MFCC for discriminating story telling and poem recitation, provide cues complementary to MFCC for the classification task.

II. Dataset

We use simultaneously recorded speech and head gestures from multiple subjects during both story telling and poem recitation. The head gesture data for poem recitation comprises of 10 subjects (6 male and 4 female) where two subjects are chosen from each of the following native languages, Hindi, Bengali, Gujarati, Kannada, Tamil. Each speaker recites a set of 20 English poems with every poem being repeated three times. Thus, each speaker recites a set of 60 English poems. After finishing recitation of a poem, each subject is asked to take sufficient break before reciting the next poem. The poems were given to the subjects well in advance so that the subjects could memorize them well. During recitation the subjects were not allowed to look at the poems. We selected 20 elementary level poems for the recording so that either the subjects are familiar with them or it is easy to memorize them. The entire recording of poem recitation by 10 subjects corresponds to 4.36 hours. Table I shows the poems with their respective average ($\pm$standard deviation (SD)) duration in seconds.

For recording head gestures during story-telling, we use a database of 24 subjects used in the work by Fotedar et al. [16] with four subjects (two male, two female) from each of the following native languages namely, Hindi, Bengali, Kannada, Malayalam, Tamil and Telugu. Each subject tells a fixed set of five stories in English as well as their native language. However, in this work, we use stories spoken in English only, that correspond to $\sim$7 hours. The chosen stories were narrated by the subjects in their own words without exact memorization.

Poem	Duration
Baa Baa Black Sheep	14.75($\pm$2.16)
Twinkle Twinkle Little Star	18.96 ($\pm$3.94)
Johnny Johnny Yes Papa	10.35 ($\pm$1.6)
Humpty Dumpty	13.78 ($\pm$3.57)
Jack and Jill	24.22 ($\pm$6.68)
London Bridge	40.93 ($\pm$11.56)
Ding Dong Bell	22.07 ($\pm$2.42)
Old MacDonald	69.04 ($\pm$18.56)
Five Little Ducks	39.32 ($\pm$11.15)
Itsy Bitsy Spider	18.04 ($\pm$4.23)
Mary had a Little Lamb	59.88 ($\pm$15.65)
Pat-a-Cake	16.21 ($\pm$10.59)
Rain Rain Go Away	52.85 ($\pm$10.74)
Rub-a-dub-dub	12.71 ($\pm$1.81)
Little Tom Tucker	21.84 ($\pm$19.6)
One Two Buckle My Shoe	23.04 ($\pm$4.2)
Pease Porridge	15.58 ($\pm$2.11)
Little Miss Muffet	17.61 ($\pm$13.59)
Cock a Doodle Do	10.89 ($\pm$1.82)
Little Bo Peep	23.36 ($\pm$4.18)

TABLE I
THE POEMS AND THEIR RESPECTIVE AVERAGE DURATION (WITH SD IN BRACKET) IN SECONDS

To capture the head gesture of the subjects, the Optitrack motion capture system [17] comprising seven IR cameras, was used. The cameras were used to keep track of the markers attached to the subject and capture their positions at a sampling rate of 120 Hz. Four of these markers were present on the headband attached to the forehead of the subject being recorded. In addition to this, two more markers were placed on a subject's nose and two more on each of the hands. The audio was recorded using a close-talking microphone at 16 kHz using Praat [18] software as it is not possible to record audio using Optitrack system. A Sony Handycam was also used to capture the frontal face video of the subjects[1]. To synchronize the audio and head gesture recording, a clapping based scheme is used as described in the work by Fotedar et al. [16].

To calculate the head gesture features, we use the position data from six head motion markers: two nose markers and four head markers. Each marker provides values in X,Y and Z coordinates and hence, for six markers we get a total of 18 values at each sampling instant. Considering a 6×3 matrix P_i for the position data at the i-th frame and N as the total number of frames, we first obtain the average of all frames $\bar{T}=\frac{1}{N}\sum_{i=1}^{N} P_i$. Considering the bottom nose marker as the origin, we calculate the translation and rotation vectors for the position data in all frames. We translate the matrix P_i to $\bar{T}$, which provides the translation vector $T(i)=\big[T_x(i)\ T_y(i)\ T_z(i)\big]$ at the i-th frame. After translation we use a singular value decomposition (SVD) based technique as proposed by Arun et al. [19] to obtain the rotation matrix representing the Eular Angles $R(i)=\big[R_x(i)\ R_y(i)\ R_z(i)\big]$.

III. Head gesture analysis using degree of repetition

We hypothesize that the head gestures during poem recitation are more structured and rhythmic in nature. For example Fig. 1 compares the head gesture trajectories for a duration of

[1] https://spire.ee.iisc.ac.in/spire/mocap.php

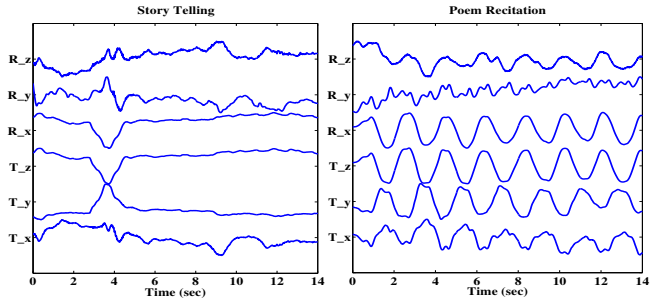

Fig. 1. Head motion trajectory of a subject while story telling and poem recitation

14 seconds when a subject was telling a story and reciting a poem. It is clear that the head gesture during poem recitation is more rhythmic in nature. In order to capture the rhythmic nature of the head gesture, we propose a measure called degree of repetition (DoR) for given head gesture contour. Let us consider a head gesture feature trajectory $x(i), 1 \leq i \leq L$, where $x \in \{T_x, T_y, T_z, R_x, R_y, R_z\}$ and L is the length of the trajectory in terms of the number of samples in a recording. To calculate DoR, we first compute the auto-correlation sequence of $x(i)$ as follows:

$$R_x(m) = \sum_i x(i)x(i+m), \quad -L \leq m \leq L, \tag{1}$$

where m denotes the index for lag in the auto-correlation sequence. It is known that $R_x(m) \leq R_x(0),\ |m| > 0$ and $R_x(-m) = R_x(m)$. If there is repetition in the head gesture due to rhythmic nature of speech, we expect high valued peak in the auto-correlation at the respective lag. For this purpose, we find out peaks in $R_x(m),\ 1 \leq m \leq L$. Let the highest peak occur at a lag of m_h. Then, the DoR is defined as

$$\text{DoR} = \frac{R_x(m_h) - \min_m R_x(m)}{R_x(0) - \min_m R_x(m)} \tag{2}$$

Thus, 0≤DoR≤1 as $R_x(m) \leq R_x(0), \forall m$. DoR quantifies to what extent the largest peak $R_x(m_h)$ is close to $R_x(0)$. An exact repetition of the head gesture would result in a large value of $R_x(m_h)$ and, hence, the ratio as defined in eqn. (2) will be close to 1. If there is no repetition, the largest peak $R_x(m_h)$ will be small compared to $R_x(0)$.

We compute DoR for $T_x, T_y, T_z, R_x, R_y, R_z$ separately for head gestures recorded during story telling as well as poem recitation. L=600 (corresponding to 5 seconds of head gesture data) with a shift of 120 samples is used for this purpose. Figure 2 compares the histogram of DoR computed using head gesture recordings from all subjects in story telling as well poem recitation. The comparison is done separately for $T_x, T_y, T_z, R_x, R_y, R_z$. The average along with SD is also indicated using dashed vertical lines for the histograms corresponding to story telling (in red colour) and poem recitation (in blue colour). It is clear from the figure that the average DoR for poem recitation is more than that for story telling. This is true for all six head gesture features. This, in turn, suggests that the head gestures during poem recitation are more repetitive than that during story telling. This could be due to rhythmic nature of the poem recitation. From Figure 2, it is interesting to observe that many of the histograms for poem recitation are bimodal in nature unlike unimodal histogram in case of story telling. For example, the histogram for T_z and R_x has a peak around 0.75 in addition to a peak around 0.5 DoR. This suggests that in several analysis windows, the head gesture for poem recitation is highly repetitive unlike that for story telling. We can also observe a large count near DoR=0 in the histogram. These correspond to analysis windows where head gestures do not repeat itself at all. Interestingly, the histogram value at DoR=0 is higher for story telling than that for poem recitation. This suggests that more segments of head gestures in story telling do not have rhythmic pattern at all compared to those in poem recitation.

The histograms in Figure 2 is generated using all twenty poems and five stories' head gesture recordings. We examine how the DoR varies across different poems and stories in our dataset. For this purpose, we compute DoR, averaged across $T_x, T_y, T_z, R_x, R_y, R_z$ as well as all subjects. These average DoR values with the respective SD values are shown in Figure 3 for twenty poems and five stories. It is clear from the figure that each of the five stories, on average, has lower DoR compared to that of every poem. This confirms our hypothesis that head gestures are less structured and rhythmic in story telling compared to those in poem recitation. Among twenty poems, the highest average DoR of 0.52 is obtained for the poem# 12- Pat-a-Cake followed by an average DoR of 0.51 for poem# 4- Humpty Dumpty. The lowest average DoR of 0.45 is found for poem# 3- Johnny Johnny Yes Papa. The highest average DoR of 0.32 among all five stories is obtained for the fifth story.

IV. Head Gesture Based Classification

In order to quantify how much information head gestures provide for story telling and poem recitation, head gesture based binary classification experiments are carried out using various representations derived from head gestures. The classification experiments are carried out in four folds. For each fold, ten subjects (among 24) are randomly picked from the story telling database and used for training. Stories from the remaining 14 subjects are used as the test set. This results in a total of 70 test stories (14 subjects $\times$ 5 stories per subject). Similarly, for each fold, eight subjects (among 10) are randomly picked from the poem recitation database for training the classifier. Remaining two subjects are used for testing which, in turn, results in a total of 120 test poems (2 subjects $\times$ 60 poems per subject).

As temporal pattern in head gesture could be indicative of the mode of speech associated with the head gesture, we consider an analysis window of five seconds (with a shift of one second) and extract various features from it for classification. Considering head gesture data during story telling from 10 training subjects, we obtain an average of 10724 frames for training across four folds. Similarly, poem recitation data from eight training subjects results in an average of 9827 training frames across four folds. Thus, it is clear that the choice of 10 subjects in story telling and 8 subjects in poem recitation result in a balanced number of training data

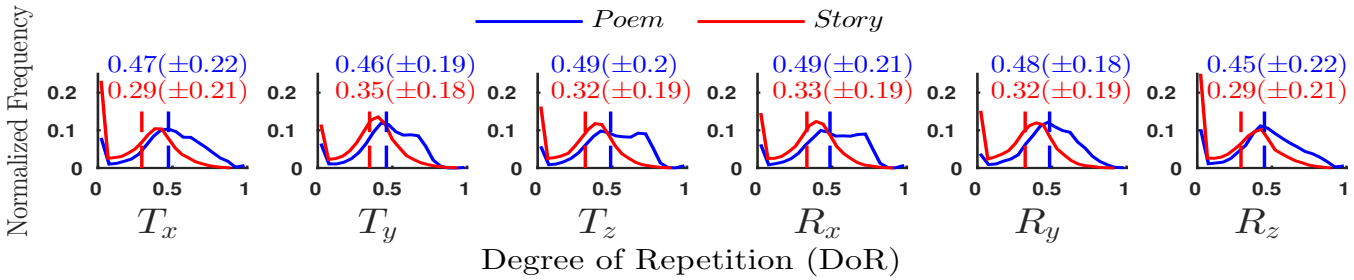

Fig. 2. Comparison between normalized histograms of DoR of head gestures during story telling and poem recitation separately for T_x, T_y, T_z, R_x, R_y, R_z. The average values (with SD in bracket) are indicated on top of the histograms.

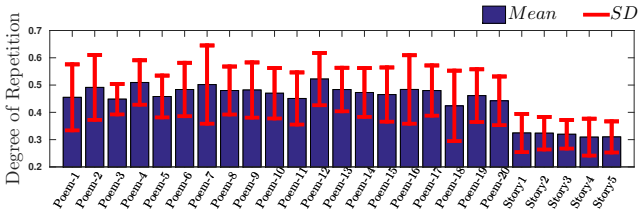

Fig. 3. Illustration of average DoR (with SD as errorbar) for each poem and story.

points for two classes. For a given test recording, each frame is classified into story telling or poem recitation class. The final decision for the test recording is obtained by majority voting of the estimated class labels across all frames.

In every analysis frame, different representative head gesture features are computed. For this purpose, we use six dimensional DoR feature computed for $T_x, T_y, T_z, R_x, R_y, R_z$ as defined in eqn. (2). This is denoted by DoR_6. We also experiment with DoR corresponding to the rotation angles only, R_x, R_y, R_z. This results in a three dimensional DoR feature, denoted by DoR_3.

In addition to the proposed DoR features, we directly use the head gesture feature trajectory, corresponding to Eular Angles only, i.e., R_x, R_y, R_z. For this purpose, we resample the trajectories of R_x, R_y, R_z from 120 Hz to 100 Hz. For a five second analysis window, this results in 1500 feature values (5 second $\times$ 100 samples per second $\times$ 3 trajectories). Velocity and acceleration are computed for each of these three feature trajectories and are used in addition to the 1500 static feature values. This results in a 4500 dimensional head gesture representation, denoted by HG_{4500}.

We compare the discriminative power of the head gesture with that of acoustic feature derived from the speech during story telling and poem recitation. For this purpose, we use 39 dimensional MFCC feature (with velocity and acceleration coefficients) computed for every 25 msec with a shift of 10 msec. This yields MFCC sequence at 100Hz, identical to the rate of the resampled trajectories, R_x, R_y, R_z. MFCCs were computed using Kaldi toolkit [20]. This results in 500 39-dimensional MFCC vectors over the five second long analysis window, which, in turn, results in a 19500-dimensional (=39$\times$500) feature vector denoted by $MFCC_{19500}$.

To examine the complementary characteristics of the head gesture and acoustic features, we also perform classification experiments by combining HG_{4500} and $MFCC_{19500}$. This results in a 24000-dimensional feature vector denoted by $COMB_{24000}$.

To perform the classification experiments using HG_{4500}, $MFCC_{19500}$ and $COMB_{24000}$, we use a four layer Deep Neural Network (DNN) and train it with 100 epochs. We also use early stopping with a patience of 7 epochs. The number of nodes in the layers is set as $[2700, 1300, 700, 300]$. We choose the number of neurons in each hidden layer in such a way that it is approximately half of the number of neurons of the previous layer. We hypothesize such an architecture could learn parsimonious representation for classification. In each hidden layer we perform batch normalization followed by dropout. Batch normalization is used to make sure that the inputs to a layer have zero mean and unit variance. However, Batch normalization operation may not always yield zero mean and the obtained mean may instead vary according to the requirement of the next layer [21]. We use ReLu-Rectified linear unit as the activation function in every layer except the last one. ReLu activation is used due to the faster convergence of ReLu compared to other activation functions. We use softmax activation in the last layer to obtain the scores for the classification. Each layer has a dropout of 0.2 in order to avoid over fitting [22]. For classification experiments using DoR_3 and DoR_6, we use a neural network with one hidden layer with four nodes. Similar to the previous network, ReLu is used as the activation function in the first layer with softmax action at the output layer. Accuracy and F-score are used to evaluate the classification performance. These are calculated using true positive (TP), true negative (TN), false positive (FP), false negative (FN), precision (P) and recall (R) as follows: F-score=$2RP/(R+P)$, accuracy=$(TP+TN)/(TP+TN+FP+FN)$, where $P = TP/(TP+FP)$ and $R = TP/(TP+FN)$.

V. Results and Discussion

Feature	Fold1	Fold2	Fold3	Fold4	Average
DoR_3	80.00	74.72	72.13	85.39	78.06
	81.95	76.53	73.85	86.60	79.73
DoR_6	81.62	77.47	77.05	86.21	80.59
	83.50	79.40	79.41	86.87	82.30
HG_{4500}	83.51	83.54	91.80	84.32	85.79
	88.12	88.18	93.39	86.51	89.05
$MFCC_{19500}$	90.29	94.53	97.01	96.86	94.67
	93.29	94.53	97.21	97.36	95.60
$COMB_{24000}$	98.42	99.47	99.30	97.29	98.62
	98.73	99.58	99.45	97.90	98.92

TABLE II
The Accuracy (in blue) and F-score in percentage, in each fold

Table II shows the accuracy (in blue) and F-score of classification using different features for each fold as well as averaged

across all folds. DoR_6 results in an classification accuracy of 80.59%. This suggests that the extent to which head gestures repeat could be used as a cue to classify a test clip to story telling or poem recitation with an accuracy of 80.59%. Interestingly, when DoR_3 is used, the accuracy turns out to be 78.06%. Thus, when the DoR of translation components are not used, the classification accuracy drops by only 2.53% suggesting that the DoR of the rotation components carry the primary cues in distinguishing two classes.

Based on this observation, we consider the HG_{4500} features based on rotation components only. HG_{4500} results in a classification accuracy of 85.79%, which is more than that obtained by DoR_3 which suggests that, for discriminating two classes, the DNN classifier learns features which are not captured by the DoR measure. Interestingly, classification accuracy for each fold reveals that HG_{4500} results in an improved classification accuracy over DoR_3 for all folds except Fold4.

$MFCC_{19500}$ results in a classification accuracy higher than that using HG_{4500} for every fold and, hence, also on average (94.67%). This suggests that acoustics have more discriminatory power than head gesture for story telling vs poem recitation classification. However, when these features are combined ($COMB_{24000}$), the classification accuracy becomes 98.62%, which is better than that using each of these features alone. This suggests that the head gestures and MFCCs provide complementary cues and, hence, improve the classification performance when combined. It could be that the repetition of the head movement, although an important cue for classification, may not be directly present in the MFCC feature sequence. We experiment with varying duration of the test clip to examine the minimum required data to achieve a target classification performance. Fig. 4 shows both frame level accuracy and F-score using $MFCC_{19500}$, HG_{4500}, and $COMB_{24000}$ when the duration of the test clip is varied from 5 seconds to 150 seconds. It is clear that with 40 seconds of test clip the classification performance saturates and does not increase significantly further.

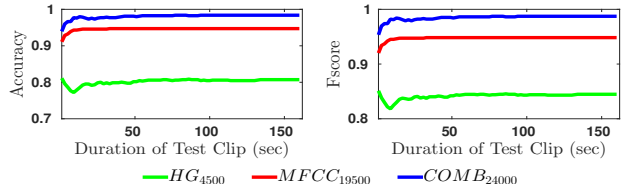

Fig. 4. Average accuracy and F-score values using MFCC ($MFCC_{19500}$), Head Gesture (HG_{4500}) and their combination ($COMB_{24000}$) for varying duration of test clip

VI. CONCLUSION

Head gestures for story telling and poem recitation are analyzed using a proposed measure called degree of repetition, which reveals that the head gestures are more rhythmic in poem recitation compared to that in story telling. Classification experiments with head gestures show that they perform worse compared to acoustic features. However, they are found to carry complementary cues and, thus, improve the classification accuracy when combined. While head gesture during poem recitation show rhythmic nature, it is not clear to what extent it is related to the rhythm of the poem. Understanding that and developing model of head gesture synthesis during poem recitation are parts of our future work.

REFERENCES

[1] R. El Kaliouby and P. Robinson, "Real-time inference of complex mental states from facial expressions and head gestures," in *Real-time vision for human-computer interaction*. Springer, 2005, pp. 181–200.

[2] K. G. Munhall, J. A. Jones, D. E. Callan, T. Kuratate, and E. Vatikiotis-Bateson, "Visual prosody and speech intelligibility: Head movement improves auditory speech perception," *Psychological science*, vol. 15, no. 2, pp. 133–137, 2004.

[3] E. Z. McClave, "Linguistic functions of head movements in the context of speech," *Journal of pragmatics*, vol. 32, no. 7, pp. 855–878, 2000.

[4] T. Kuratate, K. G. Munhall, P. Rubin, E. Vatikiotis-Bateson, and H. Yehia, "Audio-visual synthesis of talking faces from speech production correlates." in *EuroSpeech*, 1999.

[5] H. C. Yehia, T. Kuratate, and E. Vatikiotis-Bateson, "Linking facial animation, head motion and speech acoustics," *Journal of Phonetics*, vol. 30, no. 3, pp. 555–568, 2002.

[6] M. E. Sargin, E. Erzin, Y. Yemez, A. M. Tekalp, A. T. Erdem, C. Erdem, and M. Ozkan, "Prosody-driven head-gesture animation," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 2. IEEE, 2007, pp. II–677.

[7] M. Yang, J. Jiang, J. Tao, K. Mu, and H. Li, "Emotional head motion predicting from prosodic and linguistic features," *Multimedia Tools and Applications*, vol. 75, no. 9, pp. 5125–5146, 2016.

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

[9] H. Gunes and M. Pantic, "Dimensional emotion prediction from spontaneous head gestures for interaction with sensitive artificial listeners," in *Intelligent virtual agents*. Springer, 2010, pp. 371–377.

[10] C. Busso, Z. Deng, M. Grimm, U. Neumann, and S. Narayanan, "Rigid head motion in expressive speech animation: Analysis and synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1075–1086, 2007.

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

[12] S. R. Rochester, "The significance of pauses in spontaneous speech," *Journal of Psycholinguistic Research*, vol. 2, no. 1, pp. 51–81, 1973.

[13] D. O'Shaughnessy, "Recognition of hesitations in spontaneous speech," in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, vol. 1. IEEE, 1992, pp. 521–524.

[14] M. Goto, K. Itou, and S. Hayamizu, "A real-time filled pause detection system for spontaneous speech recognition," in *Sixth European Conference on Speech Communication and Technology*, 1999.

[15] P. Suppes, "Rhythm and meaning in poetry," *Midwest Studies in Philosophy*, vol. 33, no. 1, pp. 159–166, 2009.

[16] G. Fotedar and P. K. Ghosh, "An information theoretic analysis of the temporal synchrony between head gestures and prosodic patterns in spontaneous speech," *Proc. Interspeech 2017*, pp. 157–161, 2017.

[17] N. Point, "Optitrack," *Natural Point, Inc.,[Online]. Available: http://www. naturalpoint. com/optitrack/.[Accessed 22 2 2014]*, 2011.

[18] P. Boersma and D. Weenink, "Praat, software for speech analysis and synthesis," 2005.

[19] K. S. Arun, T. S. Huang, and S. D. Blostein, "Least-squares fitting of two 3-d point sets," *IEEE Transactions on pattern analysis and machine intelligence*, no. 5, pp. 698–700, 1987.

[20] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[21] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, 2015, pp. 448–456.

[22] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting." *Journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.