AUTOMATIC IDENTIFICATION OF SPEAKERS FROM HEAD GESTURES IN A NARRATION

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

Department of Electrical Engineering, Indian Institute of Science, Bangalore - 560012, India.

ABSTRACT

In this work, we focus on quantifying speaker identity information encoded in the head gestures of speakers, while they narrate a story. We hypothesize that the head gestures over a long duration have speaker-specific patterns. To establish this, we consider a classification problem to identify speakers from head gestures. We represent every head orientation as a triplet of Euler angles and a sequence of head orientations as head gestures. We use a database having recordings from 24 speakers where the head movements are recorded using a motion capture device, with each subject narrating ten stories. We get the best speaker identification accuracy of 0.836 using head gestures over a duration of 40 seconds. Further, the accuracy increases by combining decisions from multiple 40 second windows when a recording is available with duration more than the window length. We achieve an average accuracy of 0.9875 on our database when the entire recording is used. Analysis of the speaker identification performance over 40 second windows across a recording reveals that the speaker-identity information is more prevalent in some parts of a story than others.

Index Terms— Euler angles, Speaker recognition, CNN, LSTM, Head gestures.

1. INTRODUCTION

Patterns in the movement of the head are referred to as head gestures. It has been well established that head gestures play a vital role in face-to-face interactions by conveying several para-linguistic cues [1]. While head gestures are involuntary in general, in certain cases, they carry specific information [2]. Nod and shake are two well-known head gestures where the former generally conveys an approval and the latter a disapproval by the subject. Head motions have been shown to convey a sense of intensity with which the subject speaks, in addition to the semantic information [3][4]. Sargin et al. [5] have shown that there exists a co-occurring pattern between speech and head gestures [6]. Research has shown that speech accompanied by proper head gestures improves the listener's comprehension [7][8]. People also form impressions about one another based on the degree and nature of the head motion they perceive during an interaction [9]. Not only are head gestures prevalent during a dialogue between multiple speakers, but they also significant when a single speaker is giving either a discourse, a narration, or a monologue. Head gestures are influenced by the emotional state of the subject in addition to the content being spoken [10][11].

Significant research has gone into inferring para-linguistic cues from the head motion of human subjects, and also synthesizing realistic head motion. Research has been done to infer emotions from the head motion and also emotional states from videos [12][1]. Head gestures have also been used to differentiate a poem recitation from a story narration [13]. Conversational robots use the para-linguistic information conveyed by head gestures during a conversation to interact better [14]. Head gestures have also been synthesized using prosody variation with varying levels of objective and subjective metrics [15][16][17][18]. Semantic variation descriptors like pleasure, arousal, and dominance give informative cues while synthesizing head gestures [19].

Among all the information conveyed by head gestures, in this work, we focus on the identity of a speaker. Understanding how the identity of a speaker is encoded in head gestures can aid in identifying the speaker in noisy videos where the speaker's face may not be visible. It can also be used to embed personal identity while synthesizing subject dependent head gestures.

Identifying speakers from head gestures has important applications in forensics where a face in a video appears blurry or masked. Several works in the past indicate that personal identity and head gestures could be related. Multiple experiments have shown that there exists a correlation between personality traits and non-verbal cues. Campbell and Rushton [20], with their extensive research, have established many interesting relationships between personality and non-verbal communication. Extroverts tend to talk more than introverts, which, in turn, affects their head gestures. Unlike introverts, they tend to look less at the listeners and also make fewer head gestures. Non-verbal communication gives reasonable cues about personality and temperament [20]. Luck et al. conducted experiments and established a correlation between personality traits and movements induced in the subjects by music [21]. Hill and Johnston [22] designed an interesting subjective experiment to determine whether human head gestures encode any speaker identity information. The experiment animated neutral heads (avatars) with the head motion of real human actors recorded using a motion capture device. Multiple recordings were animated for every human actor. Another set of human evaluators were asked to group the animations into groups belonging to the same human actor. Based on a custom evaluation metric, Hill and Johnston concluded that humans could discriminate speakers using head gestures with statistical significance. The results in this experiment approximately translate to an accuracy of 0.5208 in a four speakers setting, with three test recordings from each speaker. The experiments conducted by Hill and Johnston [22], however, has a limitation of being subjective, and, hence, challenging to automate and scale. Unlike the subjective experiments, we, in this work, develop algorithms to identify the speaker from head gestures, which is both automated and scalable.

The average duration of animations shown to human evaluators in the experiment by Hill and Johnston [22] was 7.2s. This indicates that the head gestures spanning several seconds could encode speaker identity. Our work analyses head gestures of different durations to identify subject-specific patterns that might be encoded. For this purpose, we use Convolutional Neural Networks (CNN) to capture local head motion patterns and Long Short Term Memory (LSTM) to capture long term subject-specific cues in those patterns.

Among many durations considered in this work, on a dataset of 24 speakers, each narrating ten stories, we get the best speaker identification performance for head gestures over a duration of 40s. Combining decisions from multiple windows, we achieve an average accuracy of 0.9875 for the 240 recordings in our database. We use the same CNN-LSTM architecture to predict the speaker-identity from the audio. Interestingly, with the same architecture, a much smaller audio signal of 1.2s duration yields an accuracy similar to that of 40s of head gestures.

2. DATASET

For all the experiments in this work, we use the database introduced by Fotedar et al. [23] for our experiments. The dataset consists of 24 subjects in the age range of 20 to 38 years, from 6 different native languages - Kannada, Tamil, Telugu, Malayalam, Bangla, and Hindi. For each language, we have four subjects, and each subject narrates five different stories. Though the set of five stories is common for all the subjects, every subject narrates the story in his/her own words. Each subject narrates a story once in English and once in his/her native language. In total, we have ten recordings per subject, five in English and five in respective native languages across 24 subjects, resulting in 240 recordings. All these 240 recordings together amount to 6.6 hours of data, details of which are summarized in Table 1.

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 1: Story-wise summary of duration (sec) of recordings. English-i and Native-i refer to the i^{th} story narrated in English and native language, respectively.

It should be noted from Table 1 that the duration of a story varies greatly from one speaker to another. As the subjects narrate a story impromptu, some of the subjects narrate in great detail, resulting in longer recordings while others summarize the same story in a shorter duration.

Each of the 240 recordings in the database has parallel audio, and head motion data. High precision coordinates of 6 different points from the head captured using a motion capture device constitute the head motion data. The motion capture device captures X, Y, and Z coordinates of the points at a rate of 120 frames per second. More information regarding the entire procedure of recording can found in the work by Fotedar et al. [23].

While analysing gestures of a head, the head is assumed to be a rigid body. Any orientation of a rigid body can be captured in terms of rotation angles along any three axes which span the 3-D space — Euler Angles. From the coordinates captured by the motion capture device, we calculate Euler angles for our experiments according to the procedure adopted by Fotedar et al. [23]. Three calculated Euler angles $[\theta_x^i, \theta_y^i, \theta_z^i]$ serve as a representation at every frame in our database. Every component of this triplet is mean removed to ensure that there is no bias in the triplets. In other words, a triplet of Euler angles represents the orientation of the head of a subject in every frame.

We use the audio data in the database to compare the subject identity encoded in audio and head gestures. Kaldi [24] is used to remove silent regions in every audio as silent regions do not carry any speaker-specific information. 13 Mel frequency cepstral coefficients (MFCCs) are calculated with a frame size of 0.025s and a frame shift of 0.0083s. This results in 13 audio features in every frame at 120 frames per second, identical to the rate of head gestures.

3. METHODOLOGY

Experiments by Hill and Johnston [22] indicate that the head gestures spanning several seconds contain subject-identity information. However, the optimal duration for obtaining the best speaker identification performance needs to be determined. To determine the best duration, we divide every recording of head gestures into windows $\{w_S^j\}$ of different duration, where w_S^j denotes the j^{th} window of duration W_S seconds. Each recording with a duration of R_S seconds is divided into multiple windows using a shift of S_S seconds. The resultant windows are denoted by $H_S = \left\{ w_S^i \mid 1 \leq i \leq M, M = \left\lfloor \frac{R_S - W_S}{S_S} \right\rfloor + 1 \right\}$.

Speaker specific head gestures may have any duration and occur at arbitrary instants over an entire recording. To address these scenarios, we propose an architecture shown in Fig. 1(a) for speaker identification.

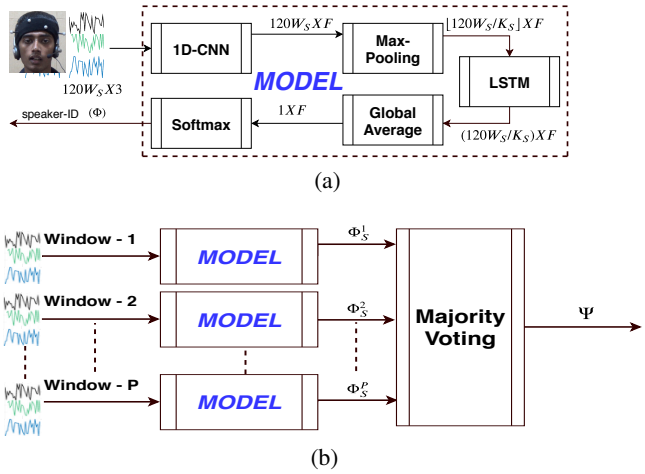
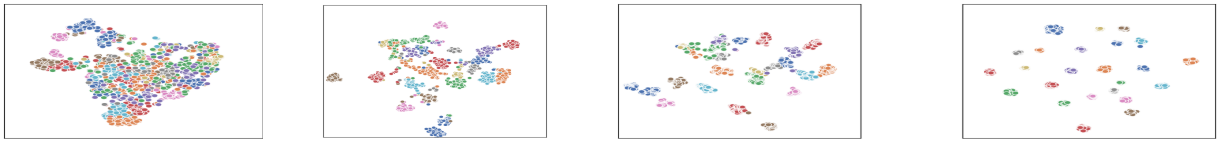


Fig. 1: (a) Network architecture for head gesture based speaker identification (b) Scheme for speaker identification using head gestures from multiple windows

In general, CNN applies the same filter throughout its input, thus extracting features with no importance to the location. In our architecture, we use 1D CNN to extract the local features present anywhere in the input head gestures. The max-pooling layer is used to find a single representation for a small kernel spanning K_S seconds. In turn, max-pooling reduces the number of input timestamps for LSTM by a factor of K_S. LSTM captures the temporal dependency in the sequence of head gestures. LSTM is followed by a global average layer to ensure that the individual *signature* of a subject is captured irrespective of the location of that *signature* in the input. The output of the global average is then fed to a softmax layer, which outputs the speaker-ID corresponding to the speaker with the highest probability for the given input head gestures.

While predicting speaker-ID, varying degree of speaker identity specific information might be present at different time locations of the input head gestures. To address this problem and utilise the information present in a recording longer than the window size used while training, we combine the speaker-IDs from multiple windows using election mechanism illustrated in Fig.1(b). Corresponding to every window in the input head gestures w_S^i, the model predicts a speaker-ID, denoted by (ϕ_S^j). We choose a subset of windows,



(a) Head Gestures — 5s (b) Head Gestures — 20s (c) Head Gestures — 40s (d) Audio — 5s

Fig. 2: t-SNE plots - Visualization of speaker-wise clusters

$E_S \subseteq H_S$ to participate in the election mechanism to get the final prediction for the input head gestures. Windows in E_S, $w_S^j \epsilon E_S$, are fed separately into the model to get speaker-ID predictions for those windows $\phi_S = \{\phi_S^j\}$. The most frequent speaker-ID in ϕ_S is used as the final prediction (Ψ_{E_S}) for the input head gestures. $\Psi_{E_S} = \arg\max_k \left\{ \sum_{i=1}^{|E_S|} [\phi_S^i = k] \middle| 0 \leq k \leq N-1 \right\}$, where N denotes the number of speakers, [] denotes the Iverson bracket [25].

4. EXPERIMENTS AND RESULTS

4.1. Experimental Setup

We experiment with different window sizes to determine the duration of head gestures that gives the best speaker identification accuracy. As the least duration of a story in our dataset is 77.68s (Table 1), we vary the window size (W_S) from 5s to 70s in steps of 5s. As the sampling rate of the head gesture data is 120Hz, a window of W_S second has $120W_S$ samples. Considering three Euler angles, this results in an input dimension of $120W_S \times 3$ for the proposed model (Fig. 1(a)).

Experiments are conducted in a five-fold cross-validation setup. While preparing the folds, we ensure that out of the ten recordings belonging to every speaker, six recordings are in the training set, two in the validation, and the remaining two in the test set in a round robin fashion. However, it should be noted that no hyper-parameter tuning is done on the validation set in our experiments. Each individual recording is mean removed to ensure that there is no bias in the recording. We also ensure that each recording is present in the test set exactly once.

We also examine the speaker identification performance based on the recorded speech using the same architecture, as in Fig. 1(a), to ensure reliability of our architecture as an identity classifier. For this purpose, MFCCs are used as features. Since it is known that a relatively smaller duration of audio is sufficient for speaker identification, we vary the window duration from 0.4s to 4.8s in steps of 0.4s. In addition to these window durations, we also experiment with 5s audio duration, which is identical to the smallest head gesture window duration considered.

Experiments are run in Keras environment with tensorflow backend. Adam is used as the optimizer and categorical cross-entropy as the loss function. All the experiments are run for 200 epochs with 20 epochs patience for improvement in the categorical accuracy.

Speaker identification experiments are carried out using both single and multiple windows. While for the former, CNN-LSTM architecture (as shown in Fig. 1(a)) is used, for the latter an election mechanism (as shown in Fig. 1(b)) is used in unison with the same CNN-LSTM architecture. The details of these experimental setups are described below.

Single Window: The goal of the single window based speaker identification is to examine how accurately head gestures within a window can be used to identify a speaker. By varying the window size, we also examine head gestures over multiple durations to determine the window duration that provides maximal information about speaker identity. Single window-based speaker identification accuracy is also used to examine parts of the story where head gestures provide more speaker-specific information relative to the rest of the story.

Multiple Window: Multiple window analysis is done when the duration of recording exceeds the window duration used in training. In multiple window analysis, speaker identity (Ψ_S) is predicted by combining the decisions of P contiguous windows of size W_S, which effectively considers head gestures of duration $W_S + (P-1) \times S_S$. In our work, we generate windows with a shift of 1s while analysing the performance of our model (i.e., $S_S = 1$).

We investigate whether head gestures at a particular location of a story provide more speaker-specific cues than elsewhere. For this purpose, we perform *Location Specific Analysis (LSA)* and *Location Independent Analysis (LIA)*.

Location Specific Analysis (LSA): We compare the multiple-window based speaker identification accuracy for segments of stories in the *beginning*, the *middle* and the *end* of the story. This is done to check for concentration of user-specific information in different regions of a recording. The segment duration is varied by varying P. If the segment duration is more than the length of the story for a chosen P, the entire story is used for speaker identification task. In the case of *middle*, the segment is centered exactly at the middle of the story.

Location Independent Analysis (LIA): Unlike Location Specific Analysis, which always considers windows in either the *beginning*, the *middle*, or the *end* of a recording, this analysis considers all possible P-contiguous windows for predicting the speaker identity. As a result, it gives a more robust, location independent metric than LSA by capturing the performance of our model irrespective of location in a recording. There could be multiple segments of contiguous P windows within a recording. As a result, the number of segments for any P depends on the duration of recordings. *Accuracy* is used as an evaluation metric.

4.2. Results and Discussion

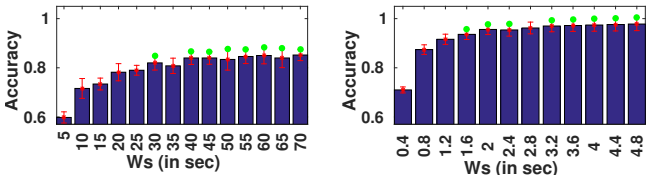
Fig. 3(a) shows the speaker identification accuracy across five folds (bar height shows the average, and errorbar shows the standard deviation (SD)) when the W_S is varied. We get the best average accuracy of 85.2% with the 70s window. We find the windows which have accuracies that are not statistically different from this window using the Wilcoxon test. These windows are indicated with green solid circles in Fig. 3(a). We consider the smallest window after which every window's performance is similar to the performance of the 70s window. We declare this smallest window with performance similar to that of the best performing window as the optimal window. From the figure, it is clear that 40s is the duration of the optimal window.

The results of similar experiments on audio data are illustrated in Fig. 3(b). We get the best performance using the 4.8s window, with an average accuracy of 0.978. We observe that accuracies using all windows longer than 3.2s (including) duration are statistically similar to that using the 4.8s window, based on the Wilcoxon test. Thus, we choose 3.2s as the duration of the optimal window for audio data

Head gestures	Fold	0	1	2	3	4	Avg
	ValAcc	.83	.81	.85	.88	.81	.836
	TestAcc	.96	.86	.94	.83	.96	.91
Audio	ValAcc	.93	.97	.98	.98	.99	.970
	TestAcc	.93	.99	.99	.99	.99	.978

Table 2: Fold wise speaker identification accuracy using head gestures over 40s duration and audio over 3.2s duration.

in our experiments. Table 2 shows the accuracies using the head gestures and audio from the optimal window duration. We notice that the audio performs consistently better than head gestures at identifying speakers in all the folds, even though the experiments with audio receive significantly shorter duration of data. It is interesting to note that an audio of duration 1.2s results in an accuracy identical to that using head gestures of duration 40s using the proposed CNN-LSTM architecture. This suggests that the head gestures over a duration of more than thirty times that of audio are needed to identify speakers with similar accuracies.



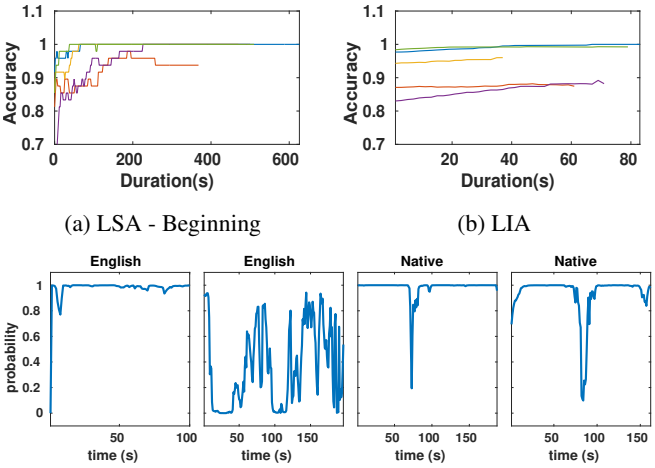
(a) Head Gestures (b) Audio

Fig. 3: Speaker identification performance across five folds on validation set by varying the window size using (a) head gestures and (b) audio

When the first 1/3-rd of the story is used for speaker identification using the multiple-window based approach, we obtain an average speaker identification accuracy of 0.942 across all five folds. When this is repeated for middle 1/3-rd and last 1/3-rd parts of the story, we obtain a speaker identification accuracy of 0.933 and 0.970, respectively. Based on the Wilcoxon test, the performances in the beginning, the middle, and the end 1/3-rd of the story were found to be not statistically significantly different. We further examine how the speaker identification performance changes when the length of the segment in a story from the beginning is increased. Fig. 4(a) shows that the accuracy increases with an increase in segment duration for all five folds separately. Accuracy with increasing segment duration is plotted in different colors for different folds. It is clear from the figure that, in general, the accuracy increases with increasing segment length, although it differs from one fold to another.

To understand how the speaker identification performance varies for every 1s in a story, we compute the probability at a location by finding the percentage of the number of 40s windows overlapping with that time location that correctly identifies the speaker. Fig. 4(c) shows such a probability profile for four randomly chosen stories from four speakers. It is clear that not all locations in a story have equal probability of identifying the speaker correctly. In fact, the locations where the probability is high, change, depending on the story. This could be related to the content of the story at that location, the way the speaker narrates the story, and also the extent to which head gestures at those locations capture speaker-specific details.

Fig. 4(b) shows the speaker identification accuracy from LIA for all the five folds separately with the increasing duration of the story segment. The accuracy increases with increasing segment duration. This suggests that the majority voting, as defined in Fig. 1(b) benefits more with the availability of more windows over longer segments. This, in turn, indicates that the best speaker identification performance is achieved using the entire story. In fact, when the entire story is used, we achieve an accuracy of 0.9875.



(a) LSA - Beginning (b) LIA

(c) Speaker identification probability at every one second based on the prediction from 40s window

Fig. 4: Illustration of location specific and location independent speaker identification performance

The 2-D plots in Fig. 2 is the t-SNE plot of the output from the Global-Average layer in the network (Fig. 1(a)) used in the experiments in this work. The four plots are the outputs for 5s, 20s, 40s head gestures, and 5s audio features (MFCC), respectively. t-SNE is a non-linear dimensionality reduction algorithm [26], which helps in visualizing higher dimensional objects in 2-D or 3-D.

From the plot, we observe that most of the data points belonging to a particular speaker — same color — form a cluster, while the data points belonging to a different user are a bit apart. Such discrimination is more clearly visible with 5s audio data and 40s head gesture data. Whereas for 5s, and 20s head gesture data, some of the clusters are not well separated from the rest of the clusters, resulting in a performance drop in our model. However, there is more overlap among speakers with 5s head gestures compared to that with 20s head gestures. Though the speaker-specific information in head gestures increases with the duration, the information contained in audio is much more discriminative, even with considerably lesser duration.

5. CONCLUSION

In this work, we conducted experiments to identify speakers using head gestures during a story narration and achieved an average accuracy of 0.836 with an input window of 40s. We also observe that the speaker specific head gestures are not equally prevalent in all parts of the story. While synthesizing head gestures based on cues from other channels like prosody, this identity specific patterns should be retained for the gestures to appear realistic and desirable to the person interacting with the avatar. A neural network can be used as an adversary while synthesizing head gestures. We plan to run our experiments on diverse and larger datasets in our future work. We also intend to collect more data for every language and analyse language-specific cues in head gestures.

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

6. REFERENCES

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

[2] Tanya 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.

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

[4] 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.

[5] Mehmet Emre Sargin, Oya Aran, Alexey Karpov, Ferda Ofli, Yelena Yasinnik, Stephen Wilson, Engin Erzin, Yücel Yemez, and A Murat Tekalp, "Combined gesture-speech analysis and speech driven gesture synthesis," in *2006 IEEE International Conference on Multimedia and Expo*. IEEE, 2006, pp. 893–896.

[6] Takaaki Kuratate, Kevin G Munhall, Philip E Rubin, Eric Vatikiotis-Bateson, and Hani Yehia, "Audio-visual synthesis of talking faces from speech production correlates," in *Sixth European Conference on Speech Communication and Technology*, 1999.

[7] 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.

[8] Katja Grauwinkel, Britta Dewitt, and Sascha Fagel, "Visual information and redundancy conveyed by internal articulator dynamics in synthetic audiovisual speech," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[9] Markus Koppensteiner and Karl Grammer, "Motion patterns in political speech and their influence on personality ratings," *Journal of Research in Personality*, vol. 44, no. 3, pp. 374–379, 2010.

[10] 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.

[11] Minghao Yang, Jinlin Jiang, Jianhua Tao, Kaihui Mu, and Hao Li, "Emotional head motion predicting from prosodic and linguistic features," *Multimedia Tools and Applications*, vol. 75, no. 9, pp. 5125–5146, 2016.

[12] Carlos Busso, Zhigang Deng, Michael Grimm, Ulrich Neumann, and Shrikanth 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.

[13] 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 Signal Processing and Communications (SPCOM)*. IEEE, 2018, pp. 36–40.

[14] Shinya Fujie, Yasuhi Ejiri, Kei Nakajima, Yosuke Matsusaka, and Tetsunori Kobayashi, "A conversation robot using head gesture recognition as para-linguistic information," in *RO-MAN 2004. 13th IEEE International Workshop on Robot and Human Interactive Communication (IEEE Catalog No. 04TH8759)*. IEEE, 2004, pp. 159–164.

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

[16] Chuang Ding, Lei Xie, and Pengcheng Zhu, "Head motion synthesis from speech using deep neural networks," *Multimedia Tools and Applications*, vol. 74, no. 22, pp. 9871–9888, 2015.

[17] Najmeh Sadoughi and Carlos Busso, "Novel realizations of speech-driven head movements with generative adversarial networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6169–6173.

[18] Najmeh Sadoughi and Carlos Busso, "Head motion generation with synthetic speech: A data driven approach.," in *INTERSPEECH*, 2016, pp. 52–56.

[19] Jia Jia, Zhiyong Wu, Shen Zhang, Helen M Meng, and Lianhong Cai, "Head and facial gestures synthesis using pad model for an expressive talking avatar," *Multimedia Tools and Applications*, vol. 73, no. 1, pp. 439–461, 2014.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011, number CONF.

[25] Donald E Knuth, "Two notes on notation," *The American Mathematical Monthly*, vol. 99, no. 5, pp. 403–422, 1992.

[26] Laurens van der Maaten and Geoffrey Hinton, "Visualizing data using t-SNE," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.