INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

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

Gaurav Fotedar, Prasanta Kumar Ghosh

Electrical Engineering, Indian Institute of Science, Bangalore 560012, India
gfotedar@gmail.com, prasantg@ee.iisc.ernet.in

Abstract

We analyze the temporal co-ordination between head gestures and prosodic patterns in spontaneous speech in a data-driven manner. For this study, we consider head motion and speech data 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. 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. These results indicate that there could be an asynchrony between head gestures and prosody during spontaneous speech where head gestures follow the corresponding prosodic patterns.

Index Terms: Head gestures, prosodic patterns, mutual information

1. Introduction

Head gestures among many other body gestures are naturally produced during speech and often convey information critical for face-to-face interaction. Studies [1], [2] report that head motion is important for auditory speech perception and can change the emotional perception of animations. The relationship between head gestures and corresponding speech is known to be complex [3]. Predicting the former from the latter has been an area of research for creating realistic avatars in natural human-computer interaction. In this work, we consider analyzing the temporal co-ordination between the two modalities. This could help in improving the accuracy of prediction of head gesture from speech.

To quantify the asynchrony between head gestures and speech, we use prosodic patterns to represent the speech acoustic. Prosodic patterns have been shown in literature to be tightly coupled with head gestures. For example, Kuratate et. al. [4] presented a system to estimate facial motion from speech and showed a high correlation between head motion and fundamental frequency F0. Yehia et al [5] developed a system to animate a talking head based on speech acoustics. They used F0 as a feature along with some constraints to estimate the head motion. Graf et al. [6] studied head and facial movements accompanying speech and concluded that despite differences from person to person, these movements are strongly correlated to the prosodic structure of text. Based on such studies many groups have proposed methods to synthesize head gestures based on prosody using models such as Hidden Markov Models(HMM) [7] [2] [8] [9], Multi-Stream HMM [10], Coupled HMM [11], Input-Output HMM[12], Dynamic Bayesian Networks(DBNs) [13] [14] and Neural Networks [15].

There have been a number of works investigating the temporal relation between speech and several body gestures including hand gestures [16] [17] [18] [19] [20] [21] [22] [23] [24] [25], lips, eyebrows [26] and other internal articulatory gestures [27]. Unlike these, there are relatively few studies that quantify the temporal coordination between speech and head motion. For example, Loehr [28] studied the rhythmic relationship between head, hands, eyeblinks and speech. He reported that each articulator produced pikes in complex synchrony with other articulators. Alexanderson et al. [29] [30] have studied the alignment of beat gestures such as head nods with syllables and reported that while there is considerable variation in fine temporal synchronization, syllables co-occurring with gestures generally have higher intensity, F0 and F0 range. Asor et al. [31] studied the alignment of head nods with respect to the prosodic structure in semi-spontaneous speech. They report that the timing of nod apexes and intensities are affected by stress, number of syllables and prosodic boundary positions. Paggio [32] studied time alignments of various kinds of head movements including nods, shakes, tilts, waggles with respect to words or phrases these movements are associated with. She reported that though there is a considerable variability in delay between the start of the gesture and corresponding speech (especially in case of waggles and shakes), there is a correlation between delay length and duration of associated speech.

All the above findings are based on datasets where specific head gestures and the associated events in speech or text are manually labelled. For example, Loehr [28] recorded four subjects in natural conversation with each other. Finally, four clips of range 20-60 seconds were manually annotated and used for analysis. Alexanderson et al. [29] [30] worked with a 20 minute dialogue motion capture data where head gestures were annotated in a semi-automatic way. Asor et al. [31] created Discourse Completion Tasks (DCT) which placed the participants in a hypothetical situation designed to elicit a declarative sentence expressing confirmation. Participants were required to use a pre-specified target word in their responses. The authors conducted their study with 155 instances of manually annotated head nods from this data. Paggio [32] used data from the Danish NOMCO corpus [33] which contains about an hour of video recordings of first-encounter dialogue interactions. Time stamps for head movements and associated speech segments are available with the database. While studies with such limited number of example events yield interesting results, time coordination between head gestures and prosodic patterns across large spontaneous dataset remains unclear. Hence, unlike example driven analysis of asynchrony between these two modalities,

http://dx.doi.org/10.21437/Interspeech.2017-999

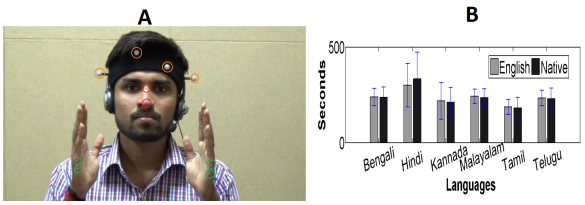


Figure 1: *Marker placements and story duration statistics*

we follow a data-driven approach. We quantify the statistical relationship between prosodic patterns and head gestures using an information theoretic measure, namely mutual information (MI). Instead of hand picking example speech segments and gesture types, MI is computed across all segments allowing us to investigate the temporal coordination in a more global sense.

We perform this information theoretic analysis with 15 hours of spontaneous speech where 24 subjects tell a fixed set of five stories in their own words. We study the asynchrony between head gestures and prosodic patterns when the stories are spoken in English as well as in speakers' native language. MI between head gestures and prosodic patterns is calculated at various delays. We find the MI, averaged across all subjects, to be maximum when head gestures lag the prosodic patterns by 30 msec. We have also trained Gaussian mixture model (GMM) based head gesture prediction from prosodic patterns at various delays and found that the root-mean squared error of the estimated head gestures follows a similar pattern across various delays.

2. Database

We collected data from 24 subjects comprising two males and two females each from six different Indian languages, namely, Bengali, Hindi, Kannada, Malayalam, Tamil, and Telugu. During recording, each subject had to tell a fixed set of five stories in his/her own words. The stories were chosen to be general and eventful so the subjects could articulate them easily without memorizing them word for word. The weblinks[1] of these stories were given to the subjects well in advance so they could read at their own pace and come prepared. Subjects were asked to rephrase every story in English as well as in their native language. Thus, we obtained ten recordings from each subject – five in English and five in their native language. During recording, subjects were not given any specific instructions. Each subject told the story in the manner they liked; this ensured that the head gestures during story-telling were natural.

Head motion was captured through the Optitrack motion capture system. Seven Optitrack IR cameras were connected to a PC through USB hubs to track reflective markers and give their 3D coordinates at 120 fps. Each subject wore a headband with four markers and two additional markers were placed on the nose as indicated by orange and red circles respectively in Figure 1. Sometimes the motion trajectories generated by the system had missing values in segments of duration ranging 10-40 frames. This was due to the markers getting occluded from the cameras either by the subject's hands while gesturing or in the process of a large head movement. Such gaps were filled using cubic interpolation in the Optitrack Arena software. The audio was recorded at 16kHz using a close-talk microphone. A

[1]http://www.worldstories.org.uk/stories/story/45-the-monkey, https://www.shortstories.co.in/birbals-journey-to-paradise/, http://mocomi.com/tenali-rama-and-the-three-dolls/, http://greece.mrdonn.org/greekgods/demeter.html, http://norse-mythology.org/tales/thor-the-transvestite/

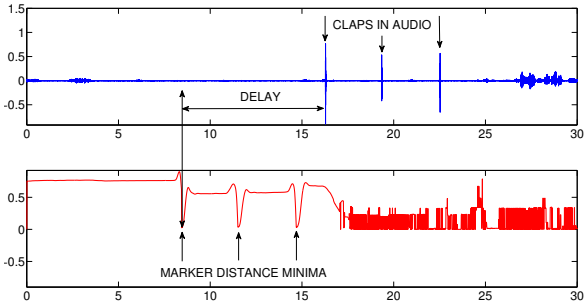


Figure 2: *Synchronization of the two modalities*

frontal face video was also captured through a Sony Handycam – model HDR-CX280E with a resolution of 1440×1080.

The Optitrack motion capture system used for recording the head gestures did not have the facility of simultaneous recording of speech. Hence, a separate laptop was used for audio recording. For synchronizing the audio and motion capture data streams, we utilized a clapping mechanism. Four optical markers were placed on the subject's hands (2 markers on each hand) as shown in Figure 1A using green circles. After both audio recorder and motion capture system started capturing data, the subject was asked to clap thrice with an interval of approximately 2 seconds before he/she started telling a story. The clapping sound is produced in a brief moment where the two hands touch each other i.e. when the distance between the two hands is minimum. Hence, the clapping sounds and the distance minima between left and right markers were used together to synchronize the two modalities. This is illustrated in Figure 2 where the top subplot shows the three clapping bursts while the respective three local minima in the distance between markers in left and right hands are shown in the bottom subplot. The delay between the two modalities was found by taking the difference between the time instant of the clapping burst and that of the corresponding local minimum in the distance.

Different subjects took different amounts of time to finish each story. Figure 1B shows the average duration of a story from speakers of each language. Interestingly, Tamil subjects took the least time to complete the story-telling task compared to subjects from other languages. This could be due to the speaking rate of the subjects as well as the manner in which they covered the events in the story. The total amount of data collected from 24 subjects is approximately 15 hours and the average duration of one story across all languages and subjects is 237 seconds (±83 seconds).

3. Representation for head gestures and prosodic patterns

We compute the features representing head gestures and prosodic patterns from the 3-dimensional position data of all Optitrack markers and the audio recordings respectively. Consider a 6×3 matrix M_i containing the X, Y and Z coordinates of six sensors (four headband and two nose markers) in the i-th frame. Suppose there are N frames in a recording. At first we compute an average position matrix $\bar{M} = \frac{1}{N}\sum_{i=1}^{N} M_i$. We consider the bottom nose marker as the center of rotation and compute the translation vector and rotation angles for the matrix M_i, $1 \leq i \leq N$. We translate M_i to $\bar{M}$ such that their centers coincide yielding a new matrix M_i'. The translation vector is denoted by $T^i = [T_x^i \;\; T_y^i \;\; T_z^i]$. Following translation, we use the singular value Decomposition (SVD) method as proposed

by Arun et al. [34] to find the rotation matrix that defines the Euler angles $\theta^i = [\theta^i_x \;\; \theta^i_y \;\; \theta^i_z]$. T^i and θ^i together are used to define a 6-dimensional feature vector representing the head gestures.

We compute pitch [35] and energy in a short-time window of duration 10msec with a shift of 10msec on the audio recordings. We do not consider the unvoiced frames as there is no pitch. Thus, in each frame of a voiced segment, combining pitch and short-time energy, we obtain a two-dimensional feature vector representing the prosodic patterns. This results in a two-dimensional prosodic feature sequence at 100Hz. However, the six-dimensional head gesture features are computed at 120Hz, the rate of the motion capture system. We resample the head gesture features at 100Hz to synchronize with the prosodic features. We discard frames which belong to a voiced segment of duration less than 0.5 sec. This results in 28771($\pm$13389) and 36350($\pm$15406) number of frames for each subject in English and native languages respectively.

4. Mutual information based analysis

We use mutual information (MI) [36] for quantifying the statistical relation between head gestures and prosodic patterns. MI indicates the statistical dependency between two random variables. Let $\mathbf{x}$ be a random vector representing the head gestures and $\mathbf{z}$ be a random vector representing the prosodic patterns. MI ($\mathcal{I}(\mathbf{z}, \mathbf{x})$) between $\mathbf{z}$ and $\mathbf{x}$ is calculated using their realizations in all voiced frames. We know:

$$\mathcal{I}(\mathbf{z}, \mathbf{x}) = \mathcal{H}(\mathbf{x}) - \mathcal{H}(\mathbf{x}|\mathbf{z}), \tag{1}$$

where $\mathcal{H}(\mathbf{x})$ is the entropy of $\mathbf{x}$ [36]. Thus, higher the MI, lower is the uncertainty of the head gestures ($\mathbf{x}$) given the prosodic patterns ($\mathbf{z}$), i.e., $\mathcal{H}(\mathbf{x}|\mathbf{z})$. Note that both $\mathbf{z}$ and $\mathbf{x}$ are continuous random variables. Thus, to compute MI, the probability density functions (PDF) of $\mathbf{z}$ and $\mathbf{x}$ need to be known. Since their PDFs are unknown, we quantize the space of $\mathbf{z}$ (denoted by $Q(\mathbf{z})$) and $\mathbf{x}$ (denoted by $Q(\mathbf{x})$) using the realizations of the head gestures and prosodic patterns with a finite number (K) of quantization bins using the K-means vector quantization [36, 37]. Then, we compute the MI by estimating the joint distribution of $\mathbf{z}$ and $\mathbf{x}$ in the finite alphabet space ($\mathbb{R}^{K\times K}$) using standard maximum likelihood criterion – frequency counts [37] and finally applying the definition of MI for discrete random variables as follows:

$$\mathcal{I}(Q(\mathbf{z}), Q(\mathbf{x})) = \sum_{z=1}^{K}\sum_{x=1}^{K} P(Q(\mathbf{z}) = z, Q(\mathbf{x}) = x) \times \log \frac{P(Q(\mathbf{z}) = z, Q(\mathbf{x}) = x)}{P(Q(\mathbf{z}) = z)P(Q(\mathbf{x}) = x)} \tag{2}$$

It can be shown that $\mathcal{I}(Q(\mathbf{z}), Q(\mathbf{x}))$ is a lower bound of the MI between $\mathbf{z}$ and $\mathbf{x}$. $\mathcal{I}(Q(\mathbf{z}), Q(\mathbf{x}))$ converges to the actual MI with more quantization bins (K). We have chosen K=64 since increasing K further does not change the results significantly.

5. Experiments and results

5.1. Experimental setup

We consider thirteen different delays δ (including the zero delay case) to analyze the temporal coordination between head gestures and prosodic patterns. Half of these delays are chosen where head gestures lead the prosodic patterns and for the remaining half the head gestures lag the prosodic patterns. The values of δ are -200, -150, -100, -70, -50, -30, 0, +30, +50, +70, +100, +150, +200 milliseconds. Let t^i_s and t^i_e denote the start and end time of the i^{th} voiced segment containing N_i frames. The pitch and short-time energy values in each frame between t^i_s and t^i_e give the $2\times N_i$ feature matrix P^i. For a given δ, the three dimensional angle vector (θ^i) and the three dimensional translation vector (T^i) between $t^i_s + \delta$ and $t^i_e + \delta$ give a $6\times N_i$ head gesture feature matrix H^i_δ.

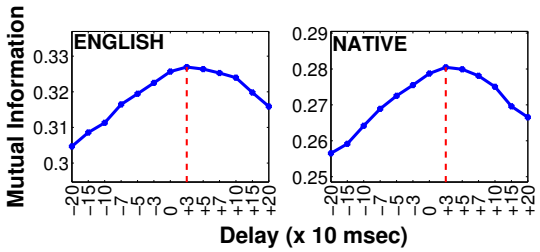


Figure 3: *MI averaged across all subjects for each delay when stories told in English and subjects' native languages are considered separately.*

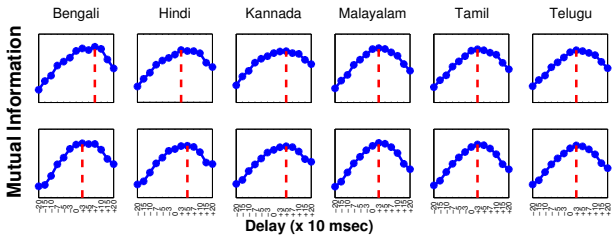
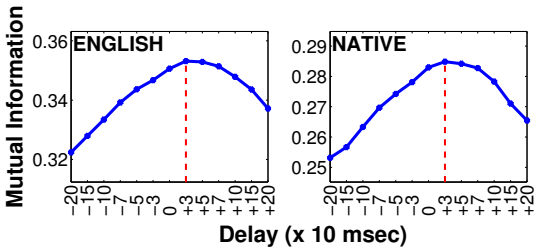


Figure 4: *MI averaged across stories in English (top row) and Native languages (bottom row) from subjects in each language separately.*

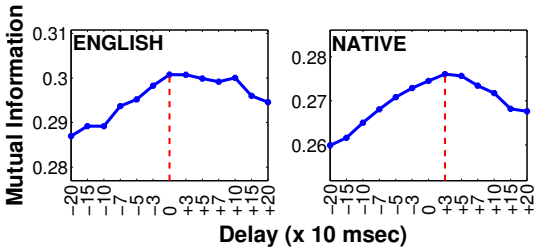
For each subject, concatenating all P^i from all five stories spoken in one language results in an overall prosodic feature matrix P. Similarly, for a given δ, concatenating all H^i_δ we obtain an overall head gesture feature matrix H_δ. This is done separately for both English and speaker's native language. Each of the two features in P is separately normalized such that it has zero mean and unit standard deviation. Similarly, six features in H_δ are also normalized separately. Following normalization, P and H_δ are used to estimate the MI.

MI estimation as outlined in Section 4 requires K-means vector quantization which depends on the initialization. To incorporate the variability in the initialization, we repeat the MI computation ten times for each delay. In each of these ten times, we randomly choose 90% of the data pairs from H_δ and P. The average of ten estimated MI values is reported for every delay separately for stories in English and subjects' native languages. MI values for different delays are reported after averaging across all 24 subjects. This is done to obtain an overall picture of the time coordination between head gestures and prosodic patterns. To investigate language and gender specific behaviors, we also report MI values for different delays by averaging across all subjects having the same native language as well as by averaging across subjects of the same gender.

To cross validate the findings on the temporal coordination using MI, we also develop a Gaussian mixture model (GMM) based head gesture prediction model from the prosodic patterns and examine the accuracy of prediction at different delays following the work by Toda et al.[38]. We consider a GMM with 16 mixture components. Since GMM parameters also vary depending on the initialization during training, we run the head



(a) *MI averaged across all female subjects*



(b) *MI averaged across all male subjects*

Figure 5: *Gender specific MI for different delays considering stories in English and Native languages separately*

gesture prediction experiment ten times in a subject-dependent manner. The average root mean squared error (RMSE) between the original and predicted head gestures for each delay is reported across these ten trials.

5.2. Results

Figure 3 shows the plots of the MI between head gestures and prosodic patterns at all delays considered in this study with the stories in English and native languages considered separately. The MI values in these plots are the average of the MI across all 24 subjects. It is clear that the highest average MI is obtained when the head gestures lag the prosodic patterns by 30msec (indicated by red dotted vertical line in Figure 3). This is true for stories when told in both English and subjects' native languages. This indicates that there is an asynchrony between the speech and the associated head gestures. It is also interesting to observe that the MI vs delay plot is asymmetric around the location of the highest MI for both English and native languages. This suggests that over a delay of 200msec between the head gestures and prosodic patterns, the prosodic patterns, when lead the head gestures, carry more information about the head gestures compared to when the prosodic patterns lag the head gestures for the same amount of delay. This could suggest that the head motion primarily follows what is spoken.

While Figure 3 shows the asynchrony between head gestures and prosodic patterns averaged across all subjects from six languages considered, we examine the temporal coordination between head gestures and prosodic patterns in a language specific manner. For this purpose, we average the MI across subjects who have the same native language. Figure 4 shows the MI vs delay plots for each of the six languages separately (one column for one language) when the subjects in the respective language tell the stories in English (top row in Figure 4) and in their native languages (bottom row in Figure 4). It is interesting to see that all twelve MI vs delay plots in Figure 4 are asymmetric similar to those in Figure 3. It is also interesting to observe that the highest average MI occurs when head gestures lag the prosodic patterns by either 30msec or 50msec or 70msec (indicated by red dashed line in Figure 4). A delay of 70msec is observed when MI is averaged over Bengali

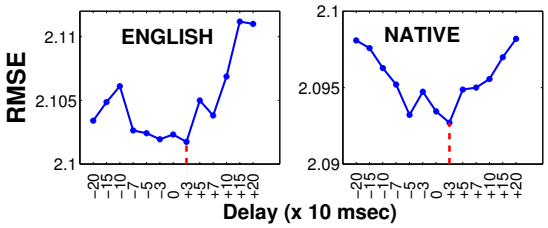


Figure 6: *RMSE of GMM based head gesture prediction for various delays*

subjects using stories spoken in English. Similarly, the delay of 50msec is observed when MI is averaged over Hindi subjects when stories were told in Hindi as well as for Kannada speakers when stories were told in both English as well as in Kannada. These results suggest that the asynchrony between head gestures and prosodic patterns is consistent across subjects from six different languages. We also examine the temporal coordination between head gestures and prosodic patterns in a gender specific manner. For this purpose, we average the MI across all female subjects considering stories in English and their native languages separately. Figure 5a shows the MI profile for various delays for female subjects. It is interesting that the MI profile appears similar to those in Figure 3 with the highest average MI appearing at +30msec (indicated by red dashed line) considering stories told in both English and native languages. Similarly, figure 5b shows the MI profile for various delays for male subjects. It is interesting to note that while the highest MI occurs at +30msec when stories spoken in subjects' native languages are considered, the highest MI occurs at 0msec when stories in English are considered although the MI at +30msec drops by only 0.03% compared to that at 0msec delay. Thus, the gender specific results also indicate that head gestures lag the prosodic patterns in spontaneous speech.

The RMSE of predicting head gestures from prosodic patterns using GMM at various delays averaged across all subjects is shown in Figure 6. We observe that minimum RMSE is obtained when the delay is +30msec in both English and speakers' native languages. This result is consistent with our observations from MI.

6. Conclusions

We conduct an information theoretic study of the temporal relationship between head gestures and prosodic patterns in speech. Head gestures along with spontaneous speech are captured from 24 subjects telling a fixed set of five stories in English and in their native languages. 3D Euler angles and translations are used to represent the head gesture while pitch and short-time energy are used to represent the prosodic patterns. MI is computed at various delays between head gestures and prosodic patterns for voiced segments. It is found that the MI averaged across all subjects is maximum when head gestures lag behind the prosodic patterns by 30 msec. This is found to be also true when subjects within each language as well as each gender are considered separately. The findings in this work are purely data-driven and obtained across different linguistic contexts and head movement types. It would be interesting to study the time asynchrony between these two modalities in a context-specific manner. This is part of our future work.

7. Acknowledgement

This work is supported by Department of Science and Technology (DST), Govt. of India.

8. References

[1] 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.

[2] 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.

[3] B. H. Le, X. Ma, and Z. Deng, "Live speech driven head-and-eye motion generators," *IEEE transactions on visualization and computer graphics*, vol. 18, no. 11, pp. 1902–1914, 2012.

[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, pp. 1279–1282.

[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] H. P. Graf, E. Cosatto, V. Strom, and F. J. Huang, "Visual prosody: Facial movements accompanying speech," in *Proceedings of Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, 2002, pp. 396–401.

[7] C. Busso, Z. Deng, U. Neumann, and S. Narayanan, "Natural head motion synthesis driven by acoustic prosodic features," *Computer Animation and Virtual Worlds*, vol. 16, no. 3-4, pp. 283–290, 2005.

[8] G. Hofer and H. Shimodaira, "Automatic head motion prediction from speech data," in *Proceedings INTERSPEECH*, 2007, pp. 722–725.

[9] Y. Ding, C. Pelachaud, and T. Artieres, "Modeling multimodal behaviors from speech prosody," in *International Workshop on Intelligent Virtual Agents*. Springer, 2013, pp. 217–228.

[10] M. E. Sargin, Y. Yemez, E. Erzin, and A. M. Tekalp, "Analysis of head gesture and prosody patterns for prosody-driven head-gesture animation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 8, pp. 1330–1345, 2008.

[11] A. Aly and A. Tapus, "Speech to head gesture mapping in multimodal human-robot interaction," in *Service Orientation in Holonic and Multi-Agent Manufacturing Control*. Springer, 2012, pp. 183–196.

[12] D. A. Braude, H. Shimodaira, and A. B. Youssef, "Template-warping based speech driven head motion synthesis." in *Proceedings INTERSPEECH*, 2013, pp. 2763–2767.

[13] S. Mariooryad and C. Busso, "Generating human-like behaviors using joint, speech-driven models for conversational agents," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 8, pp. 2329–2340, 2012.

[14] N. Sadoughi, Y. Liu, and C. Busso, "Speech-driven animation constrained by appropriate discourse functions," in *Proceedings of the 16th International Conference on Multimodal Interaction*. ACM, 2014, pp. 148–155.

[15] C. Ding, P. Zhu, L. Xie, D. Jiang, and Z.-H. Fu, "Speech-driven head motion synthesis using neural networks." in *Proceedings INTERSPEECH*, 2014, pp. 2303–2307.

[16] P. Morrel-Samuels and R. M. Krauss, "Word familiarity predicts temporal asynchrony of hand gestures and speech." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 18, no. 3, p. 615, 1992.

[17] L. Valbonesi, R. Ansari, D. McNeill, F. Quek, S. Duncan, K. E. McCullough, and R. Bryll, "Multimodal signal analysis of prosody and hand motion: Temporal correlation of speech and gestures," in *11th European Signal Processing Conference*. IEEE, 2002, pp. 1–4.

[18] L. Valbonesi, R. Ansari, D. McNeill, F. Quek, S. Duncan, K.-E. McCullough, and R. Bryll, "Analysis of speech and gestures: Gesture frequency during fluent and hesitant phases in speech," in *Proceedings of the Sixth Multi Conference on Systemics,Cybernetics and Informatics*, 2002.

[19] W.-M. Roth, "From action to discourse: The bridging function of gestures," *Cognitive Systems Research*, vol. 3, no. 3, pp. 535–554, 2002.

[20] D. P. Loehr, "Gesture and intonation," Ph.D. dissertation, Georgetown University, 2004.

[21] P. Treffner, M. Peter, and M. Kleidon, "Gestures and phases: The dynamics of speech-hand communication," *Ecological Psychology*, vol. 20, no. 1, pp. 32–64, 2008.

[22] K. Bergmann, V. Aksu, and S. Kopp, "The relation of speech and gestures: Temporal synchrony follows semantic synchrony," in *Proceedings of the 2nd Workshop on Gesture and Speech in Interaction (GeSpIn)*, 2011.

[23] T. Leonard and F. Cummins, "The temporal relation between beat gestures and speech," *Language and Cognitive Processes*, vol. 26, no. 10, pp. 1457–1471, 2011.

[24] N. Esteve-Gibert and P. Prieto, "Prosodic structure shapes the temporal realization of intonation and manual gesture movements," *Journal of Speech, Language, and Hearing Research*, vol. 56, no. 3, pp. 850–864, 2013.

[25] D. House, S. Alexanderson, and J. Beskow, "On the temporal domain of co-speech gestures: syllable, phrase or talk spurt?" *Working Papers 55. Linguistics Lund University*, p. 63, 2015.

[26] P. Keating, M. Baroni, S. Mattys, R. Scarborough, A. Alwan, E. Auer, and L. Bernstein, "Optical phonetics and visual perception of lexical and phrasal stress in english," in *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS)*, 2003, pp. 2071–2074.

[27] C. Qin and M. Á. Carreira-Perpiñán, "A comparison of acoustic features for articulatory inversion." in *Proceedings INTERSPEECH*, 2007, pp. 2469–2472.

[28] D. Loehr, "Aspects of rhythm in gesture and speech," *Gesture*, vol. 7, no. 2, pp. 179–214, 2007.

[29] S. Alexanderson, D. House, and J. Beskow, "Extracting and analyzing head movements accompanying spontaneous dialogue," in *Proceedings of Tilburg Gesture Research Meeting, Tilburg*, 2013.

[30] ——, "Aspects of co-occurring syllables and head nods in spontaneous dialogue." in *Proceedings of Auditory-Visual Speech Processing*, 2013, pp. 169–172.

[31] E. Asor, "The timing of head nods is constrained by prosodic structure," *e-Repositori, Pompeu Fabra University*, 2014.

[32] P. Paggio, "Coordination of head movements and speech in first encounter dialogues," in *Proceedings from the 3rd European Symposium on Multimodal Communication, Dublin, September 17-18*, no. 105. Linköping University Electronic Press, 2016, pp. 69–74.

[33] P. Paggio, J. Allwood, E. Ahlsén, K. Jokinen, and C. Navarretta, "The nomco multimodal nordic resource-goals and characteristics," in *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC´ 10) Valletta, Malta. May 19-21. European Language Resources Association (ELRA)*, 2010.

[34] 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.

[35] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis*, vol. 495, p. 518, 1995.

[36] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

[37] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification and scene analysis*. Wiley New York, 1973.

[38] T. Toda, A. W. Black, and K. Tokuda, "Acoustic-to-articulatory inversion mapping with Gaussian mixture model," in *Proceedings INTERSPEECH*, 2004, pp. 1129–1132.