Analysis of Emotional Effect on Speech-Body Gesture Interplay

Zhaojun Yang and Shrikanth Narayanan

Signal Analysis and Interpretation Lab (SAIL), University of Southern California, Los Angeles, CA

zhaojuny@usc.edu, shri@sipi.usc.edu

Abstract

In interpersonal interactions, speech and body gesture channels are internally coordinated towards conveying communicative intentions. The speech-gesture relationship is influenced by the internal emotion state underlying the communication. In this paper, we focus on uncovering the emotional effect on the interrelation between speech and body gestures. We investigate acoustic features describing speech prosody (pitch and energy) and vocal tract configuration (MFCCs), as well as three types of body gestures, viz., head motion, lower and upper body motions. We employ mutual information to measure the coordination between the two communicative channels, and analyze the quantified speech-gesture link with respect to distinct levels of emotion attributes, i.e., activation and valence. The results reveal that the speech-gesture coupling is generally tighter for low-level activation and high-level valence, compared to high-level activation and low-level valence. We further propose a framework for modeling the dynamics of speech-gesture interaction. Experimental studies suggest that such quantified coupling representations can well discriminate different levels of activation and valence, reinforcing that emotions are encoded in the dynamics of the multimodal link. We also verify that the structures of the coupling representations are emotion-dependent using subspace-based analysis.

Index Terms: emotion attributes, body gesture, speech prosody, speech-gesture interplay, mutual information

10.21437/Interspeech.2014-437

1. Introduction

Body gesture is an important nonverbal behavior in interpersonal communication. The expression of body gesture often spontaneously accompanies speech production. Such verbal and nonverbal behavior forms, which are usually modulated by the internal emotion state, are coherently linked to an integrated communication system towards signaling a desired message [1] [2]. Understanding the interplay of speech and body gesture as a function of the underlying emotions can facilitate research on multimodal emotion recognition, emotional gesture synthesis driven by speech, as well as human-machine interaction.

Speech and gesture are coordinated towards conveying communication intentions and many research efforts have been devoted to study such connection. Bernardis and Gentilucci found that the voice spectra were enhanced by gestures when speech and gesture were emitted simultaneously [3]. Kelly *et al.* examined the neural correlates between speech and hand gesture comprehension and suggested a possible integration of hand gesture and speech at the early and late stages of language processing [4]. Based on the analyses of speech-gesture correlation, much progress has been made on speech-driven gesture synthesis. A rule-based conversational agent has been developed in [5] by synthesizing coordinated facial expressions and hand gesture with speech intonation. Likewise, a framework for full body language synthesis in real-time from speech prosody has been proposed in [6].

The work was supported in part by NSF and DARPA.

As one of the major elements that control and influence the multimodal communicative channels, emotion has been widely studied in terms of its relation to speech and gesture. The emotional fingerprint in the acoustic measurements of speech prosody and spectral energy distribution has been explored [7] [8]. Evidence has shown that both speech prosody and short-time spectral measurements carry emotional content. Such audio features have been commonly used by the affective computing community, e.g., [9] [10]. Besides speech, gesture is another essential communicative channel which encodes emotion information. Castellano *et al.* used information from facial expressions and body gestures to discriminate eight categorical emotions [11]. Metallinou *et al.* similarly applied body language features for dynamically tracking changes in continuous emotion over an interaction [12]. Recently, we have also attempted to model attitude-related dynamics of hand gesture based on data-driven gesture primitives [13]. Despite these efforts focusing on the link of emotions and a given expressed communication modality, relatively few studies have analyzed the emotional influence on the joint relationship between speech and gesture. An HMM framework was proposed in [14] to capture the emotion-related dependency of speech and head motion. Busso *et al.* made an initial attempt at quantizing the emotional effect on the linear mapping between speech and facial gestures using data of a single subject, and found a strong speech-gesture correlation depending on the emotional content of the utterance [15]. Nevertheless, they also pointed out that a linear mapping is not sufficient to capture the speech-gesture relation especially for some facial gestures and the vocal tract features which are associated with more complex structures.

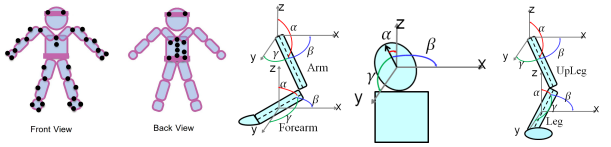
In this work, we aim at quantitatively modeling the more general dynamic coupling between speech and body gestures at utterance level and uncovering the emotional modulation of such speech-gesture dependency. To this end, we investigate acoustic features describing speech prosody (pitch and energy) and vocal tract configuration (MFCCs), as well as three types of body gestures, viz., head motion, lower and upper body motions. The speech-gesture coupling is analyzed with respect to distinct levels of emotion attributes, i.e., activation and valence. This work is based on the USC CreativeIT database which contains improvised dyadic interactions performed by 16 actors [16]. We first examine the activeness of body gestures, which is quantified as a measure of angular velocities, as a function of different activation/valence levels, and find that there exists a significant inter-emotion difference of body gesture activeness. We further employ mutual information to measure the coordination of each speech-gesture pair (e.g., prosody-head motion), and analyze the quantified speech-gesture link with respect to distinct levels of activation or valence. Analysis results reveal that the speech-gesture coordination depends on the emotions, i.e., the coupling is generally tighter for low-level activation and high-level valence, compared to high-level activation and low-level valence. In addition, a stronger MFCCs-gesture relationship is observed compared to the one between prosody and gestures, suggesting a closer interaction between the speech ar-

ticulatory process and gesture production. Motivated by these analyses, we propose a framework for quantitatively modeling the speech-gesture coupling to further inspect how the dynamics of their dependency are affected by emotions. More specifically, we quantize both speech and gesture feature vectors at each frame into discrete representations through unsupervised clustering. The speech-gesture coupling at utterance level is then quantified in a straightforward manner by computing the transition probabilities between speech and gesture. Experimental studies show that such coupling representations can well discriminate distinct levels of activation or valence, reinforcing that emotions are encoded in the dynamics of the multimodal link. We also verify that the quantified speech-gesture coupling representations are located in a low-dimensional subspace using principal component analysis (PCA), and the corresponding subspace structures are emotion-dependent. These results also shed light on emotion-dependent multimodal modeling.

2. Data Description

We use the USC CreativeIT database in this work, which is a multimodal database of dyadic theatrical improvisations [16]. The interactions performed by the pairs of actors are either improvisations of scenes from theatrical plays or theatrical exercises where actors repeat sentences to express interaction goals featuring specific emotions. The interactions were guided by a theater expert (professor/director), and were performed following the Active Analysis improvisation technique pioneered by Stanislavsky [17]. According to this technique, interactions are goal-driven; actors have predefined goals, e.g., to comfort or to avoid, which can elicit natural realization of emotions as well as expressive speech and body language behavior.

This database contains detailed full body Motion Capture (MoCap) data of the two interacting participants during a dyadic interaction, and audio data obtained through close-up microphones at 48 kHZ. A Vicon motion capture system with 12 cameras was used to capture the (x, y, z) positions of the 45 markers of each actor at 60fps, as shown in Figure 1(a). There are 50 interactions in total performed by 16 actors (9 female).



(a) (b) (c) (d)

Figure 1: (a) The positions of the Motion Capture markers; (b) – (d) Euler angles of the hand, head and leg joints.

2.1. Feature Extraction

After capturing the motion data, we manually mapped the 3D locations of markers to the angles of different human body joints using MotionBuilder [18]. The mapped angles will be used as body gesture features. The joint angles are preferred instead of 3D coordinates to describe gestures, because they are more suitable for animation purposes [6] [19] and subject-dependent gesture characteristics (e.g., the arm length) have been removed through the mapping process. In this work, we focus on three types of body gestures: lower body motion (legs), upper body motion (hands) and head motion. Figure 1(b) – (d) illustrates the Euler angles of the hand (arm and forearm), head and leg (upleg and leg) joints in the x, y and z directions. The head motion is described by the three angles (3D) shown in Figure 1(c); the upper body gesture is represented by the angles of both right and left hands (12D); and the lower body gesture is featured by the angles of both right and left legs (12D).

We extracted acoustic features of pitch and the rms energy representing the speech prosody, as well as 12 Mel Frequency Cepstral Coefficients (MFCCs) describing the vocal tract configuration, using the Praat speech processing software [20]. These features were extracted every 16.67ms (60fps) with an analysis window length of 30ms, in order to match with the MoCap frame rate. The pitch features were smoothed and interpolated over the unvoiced/silence regions. We further augment the acoustic features with their $1st$ and $2nd$ derivatives to incorporate the temporal dynamics. In this way, we have 6D prosody and 36D MFCC feature vectors. All the audio features are z-score normalized in a subject-dependent manner.

2.2. Emotion Annotation

To preserve the continuous flow of body gestures during the improvisation, we annotated time-continuous dimensional emotion attributes for each actor throughout the interaction, i.e., activation (excited vs calm) and valence (positive vs negative). Annotators used the Feeltrace instrument [21] to time-continuously indicate the attribute value from -1 to 1 for each actor while watching the video recording. The emotional state of each actor was annotated by three or four annotators. For the detailed annotation process, we refer readers to [22].

We first examine the inter-rater agreement for the continuous emotion annotations. As described in [12], we define the agreement as the linear correlation between two annotators. For each actor recording, we compute the correlation between every pair of annotators and only keep the annotator pairs with correlations greater than 0.5. Our work is based on sentences/utterances in an interaction. Each actor recording is partitioned into utterances according to speech regions. For each selected annotator, we take the average emotional annotation over an utterance, and map the value into high-level $[0.5, 1]$ and low-level $[-1, -0.5]$ activation or valence. To better capture potential differences between the extreme emotional expressions, the ambiguous values in the middle level $[-0.5, 0.5]$ are not analyzed in this work, but we intend to include this level in future work. The inter-rater agreement for the categorical labels of utterances is 0.85 for activation and 0.87 for valence. The final emotional attribute level of each utterance is decided by majority voting. This process results in 444 utterances from all subjects for activation (195 low-level and 249 high-level), and 611 for valence (280 low-level and 331 high-level).

3. Emotional Modulation

3.1. Emotional Modulation of Body Gestures

In this section, we investigate how body gestures are influenced by emotions during speech. For this purpose, we define a measure δ describing the activeness of the body gesture of M joints at the utterance level,

$$\delta = \frac{1}{N}\sum_{t=1}^{N}\frac{1}{M}\sum_{i=1}^{M}\omega(i)_t^2, \quad (1)$$

where N is the number of frames in the utterance, $\omega(i)_t$ is the angular velocity of the i-th joint $(\alpha(i)_t, \beta(i)_t, \gamma(i)_t)$ at frame t, i.e., $\omega(i)_t = \sqrt{\Delta\alpha(i)_t^2 + \Delta\beta(i)_t^2 + \Delta\gamma(i)_t^2}$, and Δ means the $1st$ order derivative of the corresponding angle.

Table 1 shows the average gesture activeness amongst utterances within a specific emotional level. We also perform t-tests to examine whether the activeness difference between distinct emotional levels is statistically significant. The p-values for the comparison tests are also presented in Table 1. We can observe that the activeness difference between the two activation levels is significant, i.e., body gestures are generally more active for excitement (high activation) compared to that for calmness

(low activation). In addition, head motion is less active than the other two motions for the high-level activation. The significant activeness difference between the low and high levels of valence can also be observed for the head and lower body motions. These results underscore the encoding of emotional cues in body gestures during speech communication.

Table 1: Average activeness of body gestures during speech, and statistical significance of activeness difference between emotion levels (Act: Activation, Val: Valence).

Mean activeness δ				
Body gesture	**Low Act**	**High Act**	**Low Val**	**High Val**
Upper	0.51	2.33	1.47	1.29
Lower	0.83	2.03	1.23	1.59
Head	0.68	1.59	1.22	0.95
Statistical significance (p-values)				
Body gesture	**Low-High Act**		**Low-High Val**	
Upper	0.000		0.127	
Lower	0.000		0.013	
Head	0.000		0.009	

3.2. Emotional Modulation of Speech-Gesture Coupling

In the previous section, we explored the emotional fingerprint in the single communicative channel of body gestures as reflected in activeness. Since speech and gesture are known to interact with one another and such interaction is influenced by the internal emotion state, this section focuses on understanding how the speech-gesture interplay is affected by underlying emotions. We use mutual information to measure the strength of the speech-gesture dependency at the utterance level. Compared to the correlation coefficient which captures a linear relationship between random variables, mutual information can model more general inter-variable connection.

Given an utterance, we denote the speech feature vector at frame t as $\mathbf{x}_t \in R^n$, and the gesture feature vector as $\mathbf{y}_t \in R^m$. We model the joint probability distribution of $\mathbf{x}_t$ and $\mathbf{y}_t$ as a Gaussian distribution: $P(\mathbf{x}_t, \mathbf{y}_t) = \mathcal{N}(\boldsymbol{\mu}^{(xy)}, \boldsymbol{\Sigma}^{(xy)})$, where $\boldsymbol{\mu}^{(xy)}$ is the mean vector and $\boldsymbol{\Sigma}^{(xy)}$ is the covariance matrix. The speech-gesture relationship in terms of mutual information at the utterance level is:

$$I(X,Y) = H(X) + H(Y) - H(X,Y) \quad (2)$$

$$= \frac{1}{2} log \frac{|\boldsymbol{\Sigma}^{(x)}| \cdot |\boldsymbol{\Sigma}^{(y)}|}{|\boldsymbol{\Sigma}^{(xy)}|}. \quad (3)$$

It is noteworthy that the mutual information I ($I \geq 0$) depends on the dimensionality $(m+n)$ of the joint feature vector. To compare the speech-gesture coupling across different speech-gesture pairs, we use the normalized version by dividing I with $(m+n)$, i.e., $\bar{I} = \frac{I}{m+n}$.

Table 2 presents the average mutual information for each speech-gesture pair (e.g., prosody-head motion) across utterances from a specific emotional level. Similar to Section 3.1, we perform comparison tests to examine whether the speech-gesture interplay in terms of mutual information is emotion-dependent. The statistical significance is shown in Table 2.

As can be observed, the normalized mutual information between speech and body gestures is much greater than 0, implying certain level of speech-gesture dependency. Moreover, the coupling of speech and the lower/upper body motions is tighter, compared to the connection between speech and head motion. This may result from the lower level of expressiveness and activeness of head motion as analyzed in Section 3.1. It is also interesting to observe that the link between MFCCs and body gestures, especially the upper and lower body gestures, is much stronger than the prosody-gesture link. Since MFCCs represent the vocal tract configuration which is related to the articulatory movements, this result indicates that the body gesture production is more closely coordinated with the speech articulatory process. Another interesting observation is the significant inter-emotion difference in speech-gesture coupling. Specifically, the speech-gesture interrelation is stronger for the low-level activation utterances than that with the high-level activation speech. Similarly, a tighter speech-gesture coupling is observed for the high-level valence utterances. These results are in concordance with the analysis of the interrelation between speech and facial gestures in [15], suggesting that the strength of speech-gesture coupling depends on emotions.

Table 2: Average mutual information of speech-gesture pairs at utterance level, and statistical significance of the speech-gesture coupling difference between emotion levels.

Average normalized mutual information				
Speech-Gesture pair	**Low-Act**	**High-Act**	**Low-Val**	**High-Val**
Prosody-Upper	0.162	0.142	0.149	0.162
Prosody-Lower	0.158	0.140	0.146	0.156
Prosody-Head	0.112	0.095	0.101	0.107
MFCCs-Upper	0.292	0.223	0.257	0.286
MFCCs-Lower	0.288	0.223	0.255	0.283
MFCCs-Head	0.127	0.091	0.108	0.126
Statistical significance (p-values)				
Speech-Gesture pair	**Low-High Act**		**Low-High Val**	
Prosody-Upper	0.003		0.028	
Prosody-Lower	0.006		0.027	
Prosody-Head	0.010		0.170	
MFCCs-Upper	0.000		0.001	
MFCCs-Lower	0.000		0.001	
MFCCs-Head	0.000		0.000	

4. Speech-Gesture Interplay Modeling

In section 3.2, we analyzed the speech-gesture coupling strength with respect to emotions in a holistic manner. Herein, we aim at modeling the speech-gesture interplay and decoding emotion from the quantified link to further inspect how the dynamics of their mutual dependency are affected by emotions.

4.1. Framework For Speech-Gesture Modeling

Inspired by the literature [23] which has reported the tight temporal co-occurrence between speech and gesture, we propose a transition model to capture the speech-gesture dependency. We first employ the clustering approach of k-means to separately group speech and gesture feature vectors of all utterances into K_s and K_g clusters. In this way, each speech or gesture feature vector at frame t is described as the discrete cluster ID C_t, where $C_t \in \{1, 2, \cdots K\}$ with $K = K_s$ for speech signals and $K = K_g$ for gesture signals. According to this quantization, each utterance is represented by a sequence of speech cluster IDs and a sequence of gesture cluster IDs.

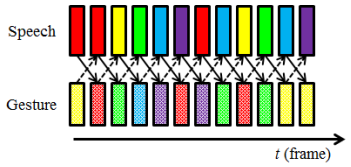


Figure 2: The illustration of speech-gesture coupling modeling.

Based on the sequences of cluster IDs, the speech-gesture coupling of each utterance is explicitly modeled as the transition probability $P(g_{t+1}|s_t)$ from the speech signal s_t to the gesture signal g_{t+1}, and vice versa, i.e., $P(s_{t+1}|g_t)$. The modeling process is illustrated in Figure 2. Hence, the temporal dynamics of speech-gesture dependency of an utterance are quantified by the transition probabilities of $P(g_{t+1}|s_t)$ and $P(s_{t+1}|g_t)$, i.e., a $2K_sK_g \times 1$ feature vector. This framework corresponds to the hidden state layers of coupled hidden Markov model (CHMM) which is popular in learning interaction dynamics [24]. A similar approach has been applied to model the temporal coordina-

tion of two interaction partners in [13].

4.2. Empirical Studies and Results

In order to validate the emotional modulation of speech-gesture coupling dynamics, we use the quantified coupling representations (the $2K_sK_g \times 1$ transition probability vector) in Section 4.1 as features for emotion recognition, i.e., classifying an utterance as high or low level of activation or valence. In addition, we investigate the widely applied linear mapping from speech to gesture [15] [25], i.e., $\mathbf{y} = \mathbf{T}\mathbf{x} + \mathbf{m}$, where $\mathbf{y}$ is the gesture vector and $\mathbf{x}$ is the speech vector (see Section 3.2). The mapping parameters $\mathbf{T}$, an $m \times n$ matrix, is obtained for each utterance by affine minimum mean square error estimation (AMMSE) [15]. Since the linear relation of speech and gesture is mainly captured by $\mathbf{T}$, we reshape it into an $mn \times 1$ vector $\mathbf{t}$ and utilize $\mathbf{t}$ as baseline features for emotion recognition. In the experiment, we employ the SVM classifier and the leave-one-subject-out scheme. In each fold, the emotional utterances are divided into training and testing data according to subjects, and the optimal cluster numbers K_s and K_g ranging from 5 to 30 are determined using cross-validation on the training set.

Table 3 presents the recognition results with the transition probabilities and with the linear mapping parameters $\mathbf{t}$ for each speech-gesture pair. In general, the coupling dynamics represented by the transition probabilities exhibit much higher discriminative power for distinguishing different levels of emotion attributes, compared to the linear mappings, indicating that the transition model in Section 4.1 can better capture the general dynamic interaction between speech and gesture. For example, the recognition accuracy for activation is 62.8% using the linear mapping of prosody and lower body motion, and has been improved to 90.8% using our proposed coupling dynamics. Specifically for the transition modeling, the coupling of lower body motion and speech especially shows the superiority of emotion discrimination over other speech-gesture pairs.

The effectiveness of decoding (recognizing) emotion from our quantified speech-gesture relationship reinforces that the dynamics of the multimodal link are controlled and modulated by the emotional content of an utterance. These results also underscore possibilities for multimodal emotion recognition and gesture synthesis by emotional speech.

Table 3: Recognition accuracy ($\%$) for discriminating low and high levels of activation and valence.

Method	Prosody			MFCC		
	Upper	Lower	Head	Upper	Lower	Head
Activation (Chance: 56.1)						
Linear	66.4	62.8	58.8	67.8	61.7	62.6
Transition	87.6	90.8	86.7	88.3	88.1	86.9
Valence (Chance: 54.2)						
Linear	59.1	58.9	56.8	56.5	57.8	56.6
Transition	67.4	79.7	73.5	73.5	81.3	74.6

4.3. Analysis of Speech-Gesture Coupling Structure

As described in Section 4.1, the transition probabilities capture the temporal dynamics of speech-gesture interaction. Hence, analysis of the structures of such quantified coupling representations can provide further insights about the emotional effect on speech-gesture coupling dynamics.

Similar to [15], we use PCA to analyze the complexity of coupling structure. If the coupling representations (transition probabilities) are located in a linear low-dimensional subspace, PCA is capable of finding the subspace by selecting the eigenvectors of the covariance matrix of the coupling data. The selected eigenvectors correspond to the highest eigenvalues which explain the most variance of the data. For each speech-gesture pair, we compute the coupling representations using the corresponding optimal cluster numbers K_s and K_g which are obtained through cross-validation in Section 4.2. We perform PCA upon the coupling representations of emotion-specific utterances (from a specific emotional level) to obtain an emotion-dependent subspace, and also for utterances from both low and high emotional levels to find the emotion-independent subspace. Note that emotion-independent utterances are obtained by randomly sampling equal utterances from low and high emotional levels, such that the emotion-independent utterance number is comparable to the emotion-specific ones.

Table 4 presents the fraction of eigenvectors which explain 90% or more of total variance of the coupling representations for each speech-gesture pair. As can be observed, for both activation and valence, the fraction of eigenvectors needed for spanning emotion-independent subspace is higher than the percentage for emotion-dependent subspace. This result demonstrates that the structure complexity of speech-gesture coupling increases as the emotion variability grows, supporting the observation that in addition to the coupling strength, the emotion content also affects the corresponding structures. Moreover, we can observe that a much higher fraction (greater than 0.5) of eigenvectors is needed for the valence-related subspace of the coupling representations between prosody and upper body motion. This observation is consistent with the result in Section 4.2 that a relatively lower accuracy for discriminating positive and negative emotions is obtained when using the corresponding transition probabilities. We could infer that the coupling structures between prosody and the upper body motion are more difficult to model in terms of valence.

Table 4: Fraction of eigenvectors explaining 90% or more of total variance of the speech-gesture coupling representations.

Gesture	Low-Act	High-Act	All-Act	Low-Val	High-Val	All-Val
Prosody						
Upper	0.21	0.23	0.32	0.60	0.64	0.68
Lower	0.16	0.19	0.24	0.20	0.22	0.26
Head	0.16	0.19	0.23	0.26	0.22	0.28
MFCCs						
Upper	0.26	0.27	0.33	0.40	0.41	0.46
Lower	0.20	0.21	0.24	0.18	0.23	0.26
Head	0.22	0.23	0.27	0.25	0.20	0.27

5. Conclusion and Future Work

In this paper, we studied how the relationship between speech and body gestures is affected by the emotional state. Overall, the analysis results revealed that the emotion content of an utterance modulates the corresponding speech-gesture coupling. The interrelation between the multimodal channels measured by mutual information is generally stronger for low-level activation and high-level valence, compared to the high-level activation and low-level valence. We further proposed a framework for modeling the dynamics of speech-gesture dependency. Experimental studies showed that such quantified coupling representations can well discriminate different levels of activation and valence, reinforcing that emotions are encoded in the coupling dynamics. We also verify that the structures of the coupling representations are emotion-dependent using PCA.

These results provide important implications for emotion-dependent multimodal modeling. For example, the tight coupling between speech and lower body motion, as well as their significant inter-emotion difference, suggest the possibility of synthesizing emotional lower body gesture driven by speech. Moreover, the effectiveness of our proposed coupling representations for discriminating distinct emotions indicates the usefulness of temporal dynamic models, such as CHMM, for speech-gesture modeling. In future work, it would also be interesting to study inter-subject and inter-gender variabilities regarding gesture activeness and speech-gesture interplay.

6. References

[1] D. McNeill, *Hand and mind: What gestures reveal about thought*. University of Chicago Press, 1992.

[2] F. Quek, D. McNeill, R. Bryll, S. Duncan, X.-F. Ma, C. Kirbas, K. E. McCullough, and R. Ansari, "Multimodal human discourse: gesture and speech," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 9, no. 3, pp. 171–193, 2002.

[3] P. Bernardis and M. Gentilucci, "Speech and gesture share the same communication system," *Neuropsychologia*, vol. 44, no. 2, pp. 178–190, 2006.

[4] S. D. Kelly, C. Kravitz, and M. Hopkins, "Neural correlates of bimodal speech and gesture comprehension," *Brain and language*, vol. 89, no. 1, pp. 253–260, 2004.

[5] J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Becket, B. Douville, S. Prevost, and M. Stone, "Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents," in *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*. ACM, 1994, pp. 413–420.

[6] S. Levine, C. Theobalt, and V. Koltun, "Real-time prosody-driven synthesis of body language," in *ACM Transactions on Graphics*, vol. 28, no. 5. ACM, 2009, p. 172.

[7] S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, Z. Deng, S. Lee, S. Narayanan, and C. Busso, "An acoustic study of emotions expressed in speech." in *INTERSPEECH*, 2004.

[8] P. N. Juslin and K. R. Scherer, "Vocal expression of affect," *The new handbook of methods in nonverbal behavior research*, pp. 65–135, 2005.

[9] C.-M. Lee, S. Narayanan, and R. Pieraccini, "Recognition of negative emotions from the speech signal," in *Proc. of ASRU*, 2001, pp. 240–243.

[10] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *Signal Processing Magazine, IEEE*, vol. 18, no. 1, pp. 32–80, 2001.

[11] G. Castellano, L. Kessous, and G. Caridakis, "Emotion recognition through multiple modalities: face, body gesture, speech," in *Affect and emotion in human-computer interaction*. Springer, 2008, pp. 92–103.

[12] A. Metallinou, A. Katsamanis, and S. Narayanan, "Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information," *Image and Vision Computing, Special Issue on Continuous Affect Analysis*, 2012.

[13] Z. Yang, A. Metallinou, E. Erzin, and S. Narayanan, "Analysis of interaction attitudes using data-driven hand gesture phrases," in *Proc. of ICASSP*, 2014.

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

[15] C. Busso and S. Narayanan, "Interrelation between speech and facial gestures in emotional utterances: a single subject study," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2331–2347, 2007.

[16] A. Metallinou, C.-C. Lee, C. Busso, S. Carnicke, and S. Narayanan, "The USC CreativeIT database: A multimodal database of theatrical improvisation," in *Proc. of Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality (MMC)*, 2010.

[17] S. M. Carnicke, *Stanislavsky in Focus: An Acting Master for the Twenty-First Century*. Routledge, UK, 2008.

[18] I. Guide, "Autodesk®," 2008.

[19] M. Sargin, Y. Yemez, E. Erzin, and A. 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.

[20] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, no. 9/10, pp. 341–345, 2002.

[21] R. Cowie, E. Douglas-Cowie, S. Savvidou*, E. McMahon, M. Sawey, and M. Schröder, "'feeltrace': An instrument for recording perceived emotion in real time," in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.

[22] A. Metallinou and S. Narayanan, "Annotation and processing of continuous emotional attributes: Challenges and opportunities," in *Automatic Face and Gesture Recognition (FG), IEEE International Conference and Workshops on*, 2013, pp. 1–8.

[23] J. M. Iverson and E. Thelen, "Hand, mouth and brain. the dynamic emergence of speech and gesture," *Journal of Consciousness Studies*, vol. 6, no. 11-12, pp. 11–12, 1999.

[24] A. V. Nefian, L. Liang, X. Pi, L. Xiaoxiang, C. Mao, and K. Murphy, "A coupled hmm for audio-visual speech recognition," in *Acoustics, Speech, and Signal Processing (ICASSP), IEEE International Conference on*, vol. 2, 2002, pp. II–2013.

[25] 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.