2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)

Perceptual Enhancement of Emotional Mocap Head Motion: An Experimental Study

Yu Ding
Univeristy of Houston
Houston, TX, USA
yding7@uh.edu

Lei Shi
Univeristy of Houston
Houston, TX, USA
lshi16@uh.edu

Zhigang Deng
Univeristy of Houston
Houston, TX, USA
zdeng4@uh.edu

Abstract—**Motion capture (mocap) systems have been widely used to collect various human behavior data. Despite existing numerous research efforts on mocap motion processing and understanding, to the best of our knowledge, to date few works have been dedicated to the investigation into whether the mocap human behavior data can be further enhanced to improve its perception. In this work, we investigate whether and how it is feasible to consistently manipulate mocap emotional head motion to enhance its perceived expressiveness. Our study relies on a mocap audiovisual dataset acquired in a laboratory setting. Participants are invited to view the animation clips of a virtual talking character displaying the original mocap head motion or manipulated head motion, and then to rate their perceived expressiveness. Statistical analysis of the rated perceptions shows that humans are sensitive to the mean of head pitch rotation (called up-down rotation) in an utterance and that the expressiveness of emotion could be improved by adjusting the mean of head pitch rotation in an utterance.**

1. Introduction

While the explicit channel (linguistic) of human communication has been widely studied through speech signal analysis and synthesis [1], the implicit one (paralinguistic) that is responsible of transmitting a complex set of signals encoding the speaker's emotional state and other hidden meaning beyond the spoken content [2], still presents many unexplored aspects. The signals through the implicit channel are able to augment the comfort level and to improve the effectiveness of human-computer interaction. Recently, the consumer market of human-computer interaction has been rapidly growing. Hence, it has become more and more important to understand face-to-face human communication.

To better investigate nonverbal behaviors in the implicit channel, mocap human behavior data has been increasingly used to carry out various research studies. For example, a large number of research efforts have been attempted to learn the relationships between nonverbal behaviors and speech signals and then to utilize the learned relationships to automatically produce human-like animations for virtual characters uttering even with emotions, including facial expressions [3] [4] [5] [6] [7] [8], head movements [9] [10] [11] [12] [13] [14] [15] [7] [16] [17] [18] [19] [20] [21] [22] [23], and hand gestures [24] [25] [26] [27] [28]. In these works, often the mocap human movement data is by default considered as the gold standard. However, the quality of mocap human movement data essentially depends on the qualities of the mocap performer including expressiveness, naturalness, etc.

In this work, we investigate whether and how we can consistently manipulate mocap emotional head motion for the purpose of perceptual enhancement. Specifically, we use an audiovisual dataset consisting of four typical emotions including neutral, sadness, happiness, and anger. This study is carried out by adjusting the mean of head pitch rotation in an utterance. The averaged value ranges from -15° (towards down) to 15° (towards up). The head motion data with and without adjusted values is used to animate a human-like virtual character [29]. Then, participants are invited to rate the level of perceived emotion when viewing the head movements of the virtual talking character. Our experimental results show certain consistent manipulations of mocap emotional head motion can measurably improve its perceived expressiveness.

Our study consists of three inter-related user studies: an *experimental* user study, a *validation* user study, and an *exploratory* user study. The experimental user study is conducted to infer the optimal mean of head pitch rotation in an utterance for four emotions (neutral, sadness, happiness, and anger), respectively. Then the validation user study is carried out to validate those optimal values refined from the experimental user study. The audiovisual data used in the experimental and the validation user studies is captured from the same subject. Finally, the exploratory user study is conducted to explore whether the obtained optimal values can be generalized, to a certain extent, to other individuals.

2. Experimental User Study

An audiovisual dataset consisting of the simultaneously captured head movements and speech of a subject is used in the experimental user study. It contains 440 audiovisual sequences (110 sentences × 4 types of emotion) from a professional actress who was instructed to utter 110 sentences

978-1-5386-0563-9/17/$31.00 ©2017 IEEE

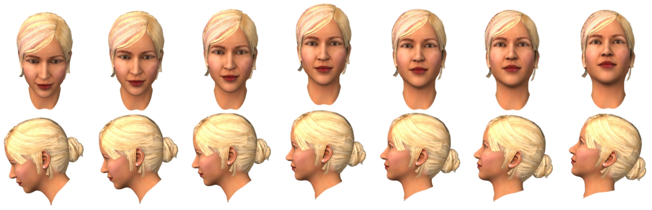


Figure 1: Reference image (front view and side view) of head pitch rotation ranging from $-15°$, $-10°$, $-5°$, $0°$, $5°$, $10°$, and $15°$.

with four common emotions including neutral, sadness, happiness, and anger, respectively. Each audiovisual utterance is comprised of speech audio signal and 3-dimensional head rotation angles. The utterance length ranges from 1.11s to 10.96s (Mean = 3.32; Std = 1.67).

During the dataset recording, the actress did not receive any specific instructions about how to move her head during speaking. She wore a headband with four attached markers. Three MoCap cameras were set up to track the 3-dimensional position of each marker at 120 frames per second. While speaking, a microphone was used to record her speech signals at 48 kHz. In the data processing, the recorded 3-dimensional positions of the head markers were used to compute the corresponding 3-dimensional rotation angles (i.e., pitch, roll and yaw) of head movement.

As the first step, our work is dedicated to looking into the mean of head pitch rotation denoted as p_{am}, which refers to as the static feature of head motion. We investigate p_{am} from $-15°$ to $15°$ with a step size of $5°$, where $0°$ denotes face straight. In total, p_{am} has seven candidates visualized in Figure 1. To figure out the impact of p_{am} on perception, twelve audiovisual sequences were selected from the dataset. They consist of three randomly selected sentences uttered four times with the four emotions.

In the experimental user study, the mean of head pitch rotation is first calculated from the head motions of the twelve utterances and then substituted with p_{am}. Note that no other operation is performed on head pitch rotation, and no operation is performed on head roll and yaw rotation values as well as speech audio signals.

The experimental user study is designed to investigate the differences among the seven conditions of p_{am}. In the user study, a virtual character [29] is used to display the 3-dimensional head rotation data under the seven conditions. So, for each utterance, seven animation clips are made under the seven conditions but with the same original speech audio. In total, 84 animation clips (3 sentences × 4 emotions × 7 conditions) were created.

In each animation clip, the virtual character displays head animation and simultaneously utters the corresponding speech audio. To avoid unnecessary visual distractions on participants, the facial expression region (including the eye region) and the mouth region are intentionally masked with mosaic. Figure 3 shows the masked virtual character.

In the user study, 25 participants were recruited and asked to view all the 84 animation clips with the aid of an online webpage interface. Each webpage encloses one animation clip that conveys a specific emotion. After viewing each animation clip as many times as the participant wants, he/she is asked to rate the level of the specific emotion using a 5-points Likert scale.

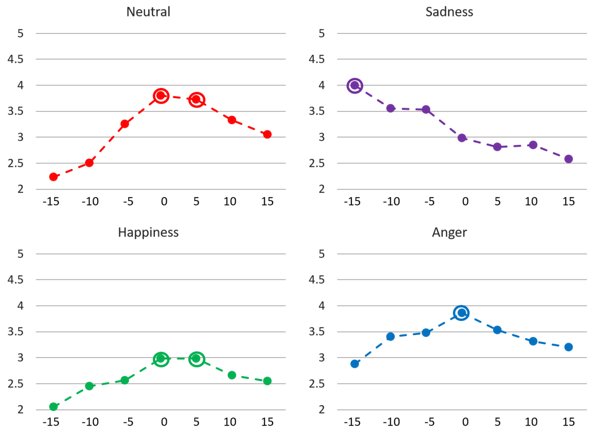


Figure 2: Perception of emotion rated by participants. For each emotion, the point(s) with the highest value(s) is marked with circles. Particularly, no significant difference is statistically found between the two circled points for neutral and happiness. More details about statistically significant differences can be found in Table 1.

Figure 2 shows the averaged scores rated by the participants for the four emotions. We are interested in p_{am} with the highest score(s) only. To learn about it, a two-tailed t-test is applied to seven scores under each condition of p_{am}. The t-test allows us to pick out p_{am} with the highest score(s) by verifying statistical difference between the averaged scores. The results of statistical difference are shown in Table 1.

For **neutral** and **happiness**, the two highest scores are observed. They are $0°$ and $5°$. No statistical difference is found between $0°$ and $5°$. The scores of $0°$ and $5°$ are statistically higher than the others. This means that the expressiveness of neutral or happiness can be enhanced most by p_{am} valued at $0°$ or $5°$. For **sadness/anger**, $-15°/0°$ is observed as the value of p_{am} with the highest score. The score of $-15°/0°$ is statistically higher than the others. This means that the expressiveness of sadness/anger can be enhanced most by p_{am} valued at $-15°/0°$. The highest scores above are marked by circles in Figure 2.

Considering that the above results are refined from only three sentences uttered with four different emotions, another user study is conducted to further validate these values of p_{am} with the highest score(s) by comparing them with the p_{am} of MoCap data (the original data), which is described below.

3. Validation User Study

To validate the values of P_{am} with the highest score(s), we formulate the hypotheses below:

TABLE 1: Statistical analysis results through a two-tailed t-test on perception of neutral (Neu), sadness (sad), happiness (hap) and anger (ang). - marks no significant difference. * marks p<.05 and ** marks p<.01. This table reports t-values of the statistical results about the significant differences between the condition(s) of p_{am} with the highest value and the others. The condition(s) of p_{am} with the highest value is marked with circle(s) in Figure 2.

Neu	$-15°$	$-10°$	$-5°$	$0°$	$5°$	$10°$	$15°$
$0°$	8.23 **	7.16 **	2.86 **		0.33 -	2.33 *	3.85 **
$5°$	7.70 **	6.62 **	2.46 *	-0.33 -		1.96 *	3.43 **

Sad	$-15°$	$-10°$	$-5°$	$0°$	$5°$	$10°$	$15°$
$-15°$		2.47 *	2.63 **	5.09 **	6.68 **	6.72 **	7.87 **

Hap	$-15°$	$-10°$	$-5°$	$0°$	$5°$	$10°$	$15°$
$0°$	5.11 **	2.76 **	2.16 *		0 -	1.70 -	2.11 *
$5°$	5.40 **	2.89 **	2.27 *	0 -		2.29 *	2.20 *

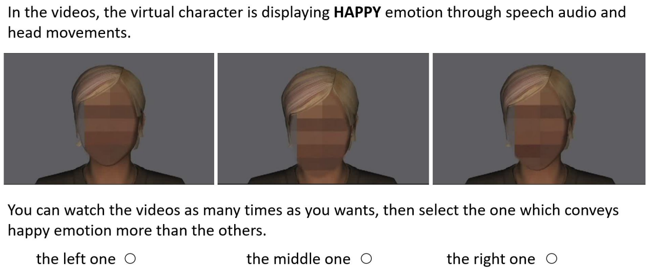
Ang	$-15°$	$-10°$	$-5°$	$0°$	$5°$	$10°$	$15°$
$0°$	5.23 **	2.70 **	2.23 *		2.00 *	3.12 **	3.6 **

- **H-Neu/H-Hap**: a virtual character animated by p_{am} at $0°$ or $5°$ is perceived as more neutral/happier than that animated by the original emotional mocap head motion.
- **H-Sad/H-Ang**: a virtual character animated by p_{am} at $-15°$/$0°$ is perceived as sadder/angrier than that animated by the original emotional mocap head motion.

To test the above hypotheses, we compare the original emotional mocap head motion with its variation(s) through a virtual character. The variation(s) refers to the head motion whose average pitch rotation angle is substituted with the above values of p_{am} (see the above hypotheses). These hypotheses are validated if the virtual character displaying the variation(s) is perceived as more neutral, sadder, happier, or angrier than that displaying the original data.

The validation user study is conducted through an online website to compare the virtual character animated with the original data and with its variation(s), respectively. The used virtual character is the same as the one in the experimental user study. 28 audiovisual sequences are randomly selected from the same dataset that is employed in the experimental user study. They are comprised of 7 sentences uttered with the four emotions and do not contain any of those utterances used in the experimental user study.

Each comparison set contains several (2 or 3 in this study) animation clips with the same speech audio, displaying the original head motion and its variation(s), respectively. Particularly, each comparison set contains three animation clips of virtual character for neutral and happiness and two clips for sadness and anger. The animation clips in a comparison set are randomly arranged into a row in a webpage. The participant can freely view the animation clips in one webpage as many times as he/she wants. The participant is explicitly informed of the specific emotion conveyed through the original audiovisual data. Then he/she is invited to cast a vote for the animation clip best conveying the specified emotion among those in one webpage. Figure 3 shows a snapshot of the validation user study webpage.



In the videos, the virtual character is displaying **HAPPY** emotion through speech audio and head movements.

You can watch the videos as many times as you wants, then select the one which conveys happy emotion more than the others.

the left one ○ the middle one ○ the right one ○

Figure 3: A snapshot of the webpage used in the validation user study. Three candidate animation clips of the virtual character are used for neutral/happiness and two ones are used for anger/sadness.

Each comparison set was judged and voted a total of 40 times by 40 participants. We counted the number of the received votes for each animation clip that was picked as the best one by the participants. Considering three (or two) candidate animation clips for neutral and happiness (or sadness and anger), the formulated hypotheses can be validated if the number of votes for picking the animation clips with the variation is statistically higher than 13.3 (13.3=40 $\times\frac{1}{3}$ refers to the baseline votes) for neutral and happiness and 20 (20=40 $\times\frac{1}{2}$ refers to the baseline votes) for sadness and anger.

To test the statistical differences of the recorded votes from the baseline votes, the 95% confidence interval for each emotion is estimated by calculating the following equation:

$$p \pm Z_{.95}\sqrt{\frac{p(1-p)}{N}} \pm \frac{0.5}{N} \tag{1}$$

where p is the percentage of the votes for picking the animation clips animated with the original data or the variation(s); $Z_{.95}$ is 1.96, representing the 95% confidence interval; and N is 40, the total number of the participants. To correct for the fact that the discrete distribution used in our work is approximated with a continuous distribution, $\frac{0.5}{N}$ is subtracted from the low limit and added to the upper limit of the interval [30]. Figure 4 illustrates the 95% confidence intervals for each emotion. From this figure, the 95% confidence intervals are higher than the baseline proportions ($66.6\% = 40 \times \frac{2}{3}$ for neutral and happiness and $50\% = 40 \times \frac{1}{2}$ for sadness and anger). These observations validate our hypotheses H-Neu/H-Hap and H-Sad/H-Ang.

To better understand whether the variation data improves the perception of emotion for each sentence, Figure 5 shows the number of votes for the animation clips animated with the original mocap data and with the variation data, for each sentence. In terms of *sadness* and *anger*, the variation data is voted for more than the baseline votes in all the utterances.

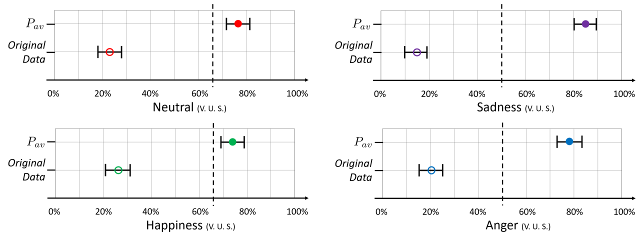


Figure 4: The 95% confidence intervals in the validation user study (V.U.S). This figure demonstrates the 95% confidence intervals for each emotion. The dash lines highlight the baseline proportion. The hollow and solid circles mark the intervals of selecting the animation clips with the original data and with its variation(s), respectively.

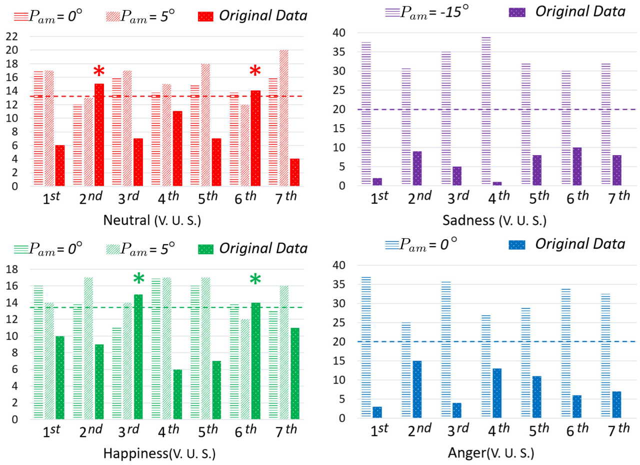


Figure 5: Votes for selecting the animation clips animated with the original mocap head motion data and with the variation data in the validation user study (V.U.S). The horizonal dash lines highlight the baseline votes which is 13.3 for neutral and happiness (one out of three), and 20 for sadness and anger (one out of two). $*$ denotes the animation clip animated with the original mocap head motion data selected more than the baseline votes.

In terms of *neutral* and *happiness*, the variation data is voted for more than the baseline votes in five utterances, and fewer than those in the other two utterances (marked with $*$ in Figure 5).

To explain the observations of the $*$ marked utterances in Figure 5, we look into the number of votes for selecting the animation clip animated with the original mocap data, V_h, and the difference in the pitch rotation value between the original data and the variation data, D_{h-v}. D_{h-v} is 0 when the averaged pitch rotation values in the original data and the variation data are the same. Figure 6 illustrates V_h along with D_{h-v} in an increasing order. As can be seen in this figure, D_{h-v} is smaller than 4° for the utterances marked with $*$ in Figure 5. In particular, neutral and happiness, D_{h-v} for the utterances with $*$ is smaller than that for the others. It suggests that, when the original head motion data is approximate to the variation data (that is, D_{h-v} is close to 0), their values of V_h would be close to each other and about the baseline votes. It explains why the original data is selected slightly more than the baseline votes in those utterances with $*$. The observations in Figures 4 and 6 are in line with the validated hypotheses.

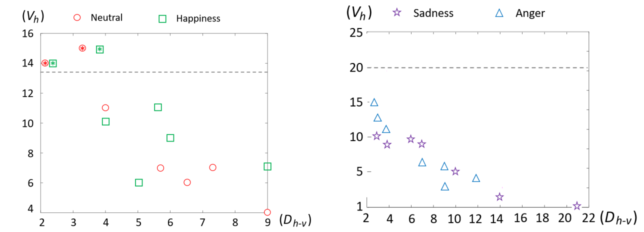


Figure 6: The relationship between V_h and D_{h-v}. V_h refers to the votes for the animation clip animated with the original head motion data; D_{h-v} refers to the difference in the average pitch rotation value between the original data and the variation data. $*$ marks the animation clips animated with the original data that have more votes than the baseline votes, which is 13.3 for neutral and happiness and 20 for sadness and anger.

Our validation user study confirms the findings of the optimal values of P_{am} from the experimental user study. The two user studies rely on the same dataset where the audiovisual utterances are collected from the same subject. To further generalize these findings to other individuals, we conducted another user study based on another dataset (i.e., the recorded audiovisual data of other subjects).

4. Exploratory User Study

An exploratory user study is conducted to investigate whether the above findings can be reasonably generalized to other individuals. The user study relies on another audio-visual dataset [31] recording the audiovisual data from eight female and six male participants. The participants are invited to utter twelve sentences two times, one with the neutral state and the other with an emotional state. While uttering with the emotional state, they are instructed to spontaneously utter the sentences. In other words, the utterance recorded in the dataset [31] may convey hybrid emotions (e.g. mixing of happy and surprise, or mixing of anger and sad). This differs from the dataset used in the experimental and the validation user studies. Since those findings are inferred from the audiovisual data of a female actress, our exploratory user study employs female data only in [31].

We choose twenty-eight audiovisual sequences of the female participants in the dataset [31]: seven sequences for neutral, sadness, happiness, and anger, respectively. As done in the validation user study, a few candidate animation clips, which display either the original mocap head motion data or the variation data, are produced for each selected utterance. 40 participants are invited to select the best one among the candidate animation clips in one comparison set. The other

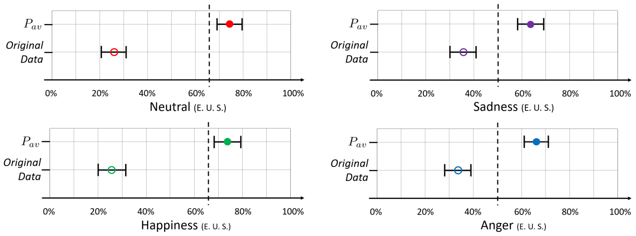


Figure 7: The 95% confidence intervals in the exploratory user study. This figure illustrates the 95% confidence interval for each emotion. The dash vertical lines denote the reference percentages. The hollow and solid circles mark the intervals of selecting the animation clips animated with the original MoCap data and with the manipulated pitch rotation, respectively.

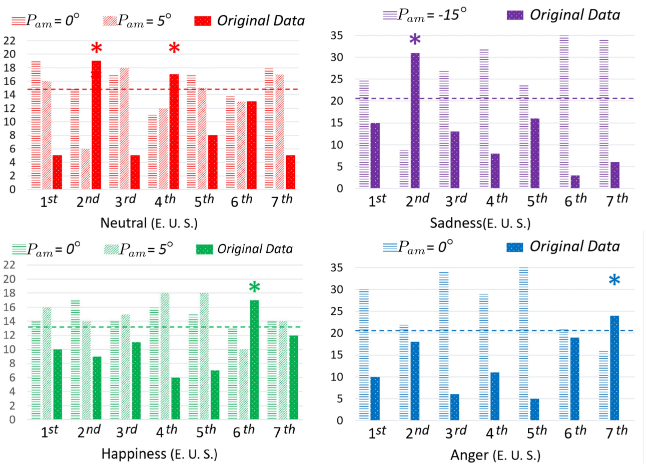


Figure 8: The numbers of user votes for the clips animated with the original mocap data and with the variation data in the exploratory user study (E.U.S). The dash lines denote the baseline votes. $*$ marks the clips animated with the original mocap data that have more votes than the baseline votes.

procedure details of the exploratory user study are the same as those of the validation user study.

The results of the exploratory user study are illustrated in Figures 7 and 8. As can be observed in Figure 7, the 95% confidence intervals are clearly higher than the baseline percentages for all the four emotions. This observation is in line with the hypotheses verified in the validation user study.

Figure 8 provides more detailed analysis of the user votes for each utterance. The results show that more utterances are perceived more neutral, sadder, happier, and angrier with the variation data than with the original mocap data, including 5 out of 7 in the neutral case and 6 out of 7 in the cases of sadness, happiness, and anger. $*$ marks the animation clips displaying the original mocap data that have more votes than the baseline votes. Different from the validation user study, we fail to use the relationships between V_h and D_{h-v} to well explain the $*$ cases in the exploratory user study (refer to Figures 5 and 6).

5. Discussion and Conclusion

While the previous work [32] has suggested that the mean of head pitch rotation is related to the discourse function (for example, a conversational agent may look up for questions), our work shows that the mean of head pitch rotation is helpful to enhance the expressiveness of emotion. Our user study results together show that humans are sensitive to the static head pitch rotation and are able to decode the emotion information from the static head pitch rotation. Also, our results show that the expressiveness of emotion of virtual characters could be improved from mocap head motion by properly manipulating head pitch rotation including towards straight or sightly up for neutral and happiness, towards straight for anger, and towards down for sadness. Particularly, our main findings are summarized below:

- *Static Motion*: The mean of head pitch rotation has impact on its expressiveness.
- *Perception*: Humans are sensitive to the mean of head pitch rotation in terms of emotion perception; and the perceptual sensitivity is no more than 5°.
- *Expressiveness*: Emotional head motions could be adjusted with consistent manipulation to enhance the expressiveness.
- *Neutral* and *Happiness*: The expressiveness of neutral and happiness can be enhanced by adjusting the average head pitch rotation straight (e.g. 0°) or slightly up (e.g. 5°).
- *Sadness*: The expressiveness of sadness can be enhanced by adjusting the mean of head pitch rotation towards down (e.g. $-15°$).
- *Anger*: The expressiveness of anger can be enhanced by adjusting the mean of head pitch rotation towards straight (e.g. 0°).

The exploratory user study results show that the votes for the variation data are higher than the baseline votes for the four emotions when taking into account all the seven chosen utterances. It appears that the consistent manipulations of head pitch rotation are not strictly validated for other individuals. This may be explained by the differences between the two used datasets. The consistent manipulations are inferred from a professional actress who is instructed to utter sentences with one typical emotion (could be neutral, sadness, anger, or happiness) each time. The professional actress is probably more skilled at encoding emotion information into the audiovisual signals than non-professional participants. It is reasonable to assume that the audiovisual data from this actress is able to fully convey one emotion of four. Since the actress is not asked to spontaneously utter the sentences, she may probably pay much attention to how to distinguish the four emotions from each other. This means that she probably conveys the specified emotion fully in an utterance while skillfully excluding other emotions in the utterance. On the other hand, in the dataset used in the exploratory user study, non-professional subjects are invited to utter sentences spontaneously, which suggests the audio-visual data may convey the mixing of multiple emotions in an utterance. This could explain certain differences between the exploratory user study results and the validation user study results.

Our work provides the early experimental evidence that certain consistent manipulations of the averaged head pitch rotation could improve the emotional perception of some subjects, but it has not been strictly validated with a general dataset. Therefore, currently we cannot generalize this finding to emotional mocap head motion of any subjects.

Acknowledgments

This research is supported in part by NSF IIS-1524782. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the agencies.

References

[1] K. Munhall, J. A. Jones, D. E. Callan, T. Kuratate, and E. Vatikiotis-Bateson, “Visual prosody and speech intelligibility,” *Psychological Science*, vol. 15, no. 2, pp. 133–137, 2004.

[2] R. Banse and K. R. Scherer, “Acoustic profiles in vocal emotion expression.” *Journal of personality and social psychology*, vol. 70 3, pp. 614–36, 1996.

[3] M. Brand, “Voice puppetry,” in *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, ser. SIGGRAPH ’99, 1999, pp. 21–28.

[4] Y. Cao, W. C. Tien, P. Faloutsos, and F. Pighin, “Expressive speech-driven facial animation,” *ACM Trans. Graph.*, vol. 24, no. 4, pp. 1283–1302, 2005.

[5] Z. Deng, J. P. Lewis, and U. Neumann, “Automated eye motion using texture synthesis,” *IEEE Computer Graphics and Applications*, vol. 25, no. 2, pp. 24–30, 2005.

[6] J. Xue, J. Borgstrom, J. Jiang, L. E. Bernstein, and A. Alwan, “Acoustically-driven talking face synthesis using dynamic bayesian networks,” in *IEEE International Conference on Multimedia and Expo*, 2006, pp. 1165–1168.

[7] Y. Ding, M. Radenen, T. Artieres, and C. Pelachaud, “Speech-driven eyebrow motion synthesis with contextual markovian models,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 3756–3760.

[8] Y. Ding, K. Prepin, J. Huang, C. Pelachaud, and T. Artières, “Laughter animation synthesis,” in *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, 2014, pp. 773–780.

[9] 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, July 2005.

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

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

[12] X. Ma and Z. Deng, “Natural eye motion synthesis by modeling gaze-head coupling,” in *2009 IEEE Virtual Reality Conference*, 2009, pp. 143–150.

[13] 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, pp. 1902–1914, 2012.

[14] 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, October 2012.

[15] Y. Ding, C. Pelachaud, and T. Artieres, “Modeling multimodal behaviors from speech prosody,” in *International Conference on Intelligent Virtual Agents*. Springer Berlin Heidelberg, 2013, pp. 217–228.

[16] A. B. Youssef, H. Shimodaira, and D. A. Braude, “Articulatory features for speech-driven head motion synthesis,” in *INTERSPEECH*, 2013, pp. 2758–2762.

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

[18] H. Van Welbergen, Y. Ding, K. Sattler, C. Pelachaud, and S. Kopp, “Real-time visual prosody for interactive virtual agents,” in *International Conference on Intelligent Virtual Agents*, 2015, pp. 139–151.

[19] Y. Ding, J. Huang, N. Fourati, T. Artieres, and C. Pelachaud, “Upper body animation synthesis for a laughing character,” in *International Conference on Intelligent Virtual Agents*, 2014, pp. 164–173.

[20] A. B. Youssef, H. Shimodaira, and D. A. Braude, “Speech driven talking head from estimated articulatory features,” in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014*, 2014, pp. 4573–4577.

[21] C. Ding, L. Xie, and P. Zhu, “Head motion synthesis from speech using deep neural networks,” *Multimedia Tools and Applications*, vol. 74, no. 22, pp. 9871–9888, 2015.

[22] N. Sadoughi and C. Busso, “Head motion generation with synthetic speech: a data driven approach,” in *Interspeech 2016*, September 2016, pp. 52–56.

[23] K. Haag and H. Shimodaira, “Bidirectional LSTM networks employing stacked bottleneck features for expressive speech-driven head motion synthesis,” in *International Conference on Intelligent Virtual Agents*, 2016, pp. 198–207.

[24] C.-C. Chiu and S. Marsella, “How to train your avatar: A data driven approach to gesture generation,” in *International Conference on Intelligent Virtual Agents*, 2011, pp. 127–140.

[25] S. Levine, C. Theobalt, and V. Koltun, “Real-time prosody-driven synthesis of body language,” *ACM Trans. Graph.*, vol. 28, no. 5, pp. 172:1–172:10, 2009.

[26] S. Levine, P. Krähenbühl, S. Thrun, and V. Koltun, “Gesture controllers,” in *ACM Trans. Graph.*, vol. 29, no. 4, 2010, p. 124.

[27] C.-C. Chiu and S. Marsella, “Gesture generation with low-dimensional embeddings,” in *International Conference on Autonomous Agents and Multi-agent Systems*, 2014, pp. 781–788.

[28] C.-C. Chiu, L.-P. Morency, and S. Marsella, “Predicting co-verbal gestures: a deep and temporal modeling approach,” in *International Conference on Intelligent Virtual Agents*, 2015, pp. 152–166.

[29] A. Hartholt, D. Traum, S. C. Marsella, A. Shapiro, G. Stratou, A. Leuski, L.-P. Morency, and J. Gratch, “All Together Now: Introducing the Virtual Human Toolkit,” in *International Conference on Intelligent Virtual Agents*, 2013, pp. 368–381.

[30] “http://onlinestatbook.com/2/estimation/proportion_ci.html.”

[31] G. Fanelli, J. Gall, H. Romsdorfer, T. Weise, and L. V. Gool, “A 3-d audio-visual corpus of affective communication,” *IEEE Transactions on Multimedia*, vol. 12, no. 6, pp. 591 – 598, October 2010.

[32] 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 &amp; spoken intonation for multiple conversational agents,” in *Annual Conference on Comp. Graph. and Inter. Tech.*, 1994, pp. 413–420.