Perceptual Analysis of Talking Avatar Head Movements: A Quantitative Perspective

Xiaohan Ma
Dept. of Computer Science
University of Houston
xiaohan@cs.uh.edu

Binh Huy Le
Dept. of Computer Science
University of Houston
bhle2@cs.uh.edu

Zhigang Deng
Dept. of Computer Science
University of Houston
zdeng@cs.uh.edu

ABSTRACT

Lifelike interface agents (e.g. talking avatars) have been increasingly used in human-computer interaction applications. In this work, we quantitatively analyze how human perception is affected by audio-head motion characteristics of talking avatars. Specifically, we quantify the correlation between perceptual user ratings (obtained via user study) and joint audio-head motion features as well as head motion patterns in the frequency-domain. Our quantitative analysis results clearly show that the correlation coefficient between the pitch of speech signals (but not the RMS energy of speech signals) and head motions is approximately linearly proportional to the perceptual user rating, and a larger proportion of high frequency signals in talking avatar head movements tends to degrade the user perception in terms of naturalness.

Author Keywords

Perceptual modeling, quantitative analysis, head motion, and audio-head motion features

ACM Classification Keywords

H.5.2 Information interfaces and presentation (e.g., HCI): User Interface—*Interaction styles*

General Terms

Human Factors

INTRODUCTION

Embodied interface agents (e.g., avatars) have been increasingly used in human-computer interfaces (HCI) [8]. In these HCI systems, perception of any human-like behaviors on avatars, such as facial expressions, hand gestures, lip movements, and head movements, can strongly influence user interaction experience. Among many human-like avatar behaviors [6], talking avatar head movement is considered as one of the crucial communication cues that can facilitate the social interaction between a human being and a humanoid avatar.

Previously, Wang *et al.* [6] studied human perception on head motion of a humanoid robot. They found that natural and appropriate robot head motions have a measurable positive effect on the social experience of users. Moreover, previous research studies [9, 5] show that head motion has a strong correlation with acoustic speech features. However, these studies were primarily focused on *qualitative* understanding of human perception on avatar head movements. The *quantitative* association between human perception and the audio-head motion characteristics of talking avatars remains to be uncovered, to the best of our knowledge.

Inspired by the above question, in this work we quantify how the characteristics of talking avatar head movements affect human perception (i.e. users' perceptual ratings). Specifically, it consists of the following steps: (1) Participants are asked to rate a number of constructed audio-head animation clips (as the visual stimuli); (2) joint features of speech signals and head motions are extracted from the audio-head animation clips; and (3) quantitative analysis is performed to study the association between the extracted features and the obtained subjective user ratings, including: a canonical correlation analysis based approach to model the correlation between joint audio-head motion features and subjective human ratings, and a frequency-domain analysis on the user-rated head movement patterns.

DATA ACQUISITION AND PROCESSING

The audio-head motion dataset used in this work was acquired from an actress using a VICON optical motion capture system. The actress was asked to speak a corpus (hundreds of sentences) while keeping her head movements as natural as possible. The original sampling frequency of the motion capture data was 120 Hz. The voice was recorded simultaneously using a close talking microphone at the sampling rate of 48 kHz. Based on the above acquired audio-motion data, three Euler angles of head rotation in each frame were computed using a Singular Value Decomposition (SVD) based technique [2]. Note that head translation is not the focus of this work. Meanwhile, we extracted the pitch (i.e. the fundamental frequency - F0) and RMS energy from the recorded acoustic speech signals. In this work, we used a Fourier analysis based cepstrum method to extract the F0 frequency from speech signals. The used window size of cepstrum sampling was 30 ms, and the output frame rate was set as 100 samples per second. RMS energy was extracted as the average squared intensity in a Hamming window (30 ms in our case). Finally, all the extracted Euler angles, pitches, and RMS energies were downsampled to the same 24 frames per second for our quantitative analysis.

CHI 2011, May 7–12, 2011, Vancouver, BC, Canada.

SUBJECTIVE EVALUATION

We generated 60 audio-head animation clips as the visual stimuli in our perceptual study. Specifically, based on 15 speech clips randomly selected from the recorded dataset, we used the following four different audio-driven head animation generation techniques to produce 60 audio-head animation clips (i.e., each for 15 clips): (1) playing back the original captured head motions (called "original data"), (2) the HMMs-based head motion synthesis algorithm [2] (called "HMMs"), (3) the mood-swings head motion synthesis framework [3] (called "Mood-Swings"), and (4) randomly generated head motions (called "randomly generated"). The main consideration of choosing four different head motion generation approaches is that if all the animations were generated by a single technique, their user evaluations might fall into a narrow range, which will directly affect the generality of our quantitative analysis.

To suppress the potential influences of other visual/animation factors on user perception, we further postprocessed the audio-head animation clips as follows. (1) We used a 3D face model without photorealistic texture in order to remove the potential influence of photorealistic facial texture. (2) We only kept lip-sync motions and removed the motions at other facial regions (e.g., eyebrows), and the eye gazes stayed still (looking straight ahead). (3) Lastly, we intentionally applied an image mosaic filter to the mouth area so that participants only can see obscure lip movements. Its main consideration is that we attempted to minimize or even remove the influence of lip-sync quality on the participants' perception on head motions. Note that a similar face masking methodology was used in previous face recognition research [7]. The left-top and left-bottom panels of Figure 1 show an example frame of our generated audio-head animation clips.

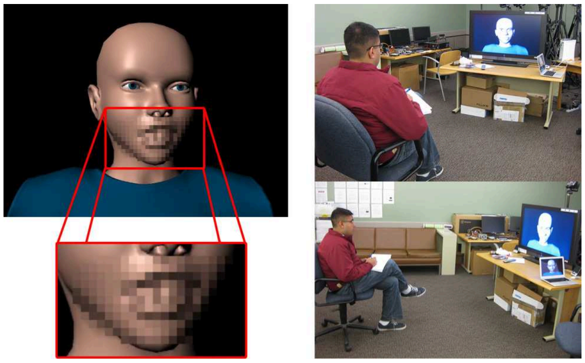

Figure 1. The left panels show an example frame of our generated audio-head animation clips. The right two panels show a snapshot of our user study.

We conducted a user study on the above 60 visual stimuli. A total of 18 student volunteers who were not informed the purpose of the study were invited to rate the naturalness of these clips individually. The demographics of the participants are: ages are from 23 to 28; 3 (16.67%) are female and 15 (83.33%) are male; and most of them are international students but all of them have been studying in the US for at least three years and can speak fluent English. It is noteworthy that although facial expression perception has been reported to be culture-dependent [10], it is still unclear whether a similar conclusion can be applied to the perception of head movements. The used rating scale was from 1 to 5 (1 represents "completely unnatural", and 5 represents "as natural as a real human's motion"). The participants viewed the clips in a 42 inches TV screen (approximately 3 meters away from them) and then wrote down their ratings (a snapshot is shown in Figure 1). For each participant, we first showed three example clips as test trials to allow him/her to get familiarized with the study procedure. The clip order for each participant was randomly generated to ensure the counter-balance of the experiment. Finally, we computed the average rating of each clip.

ANALYSIS OF RESULTS

For each audio-head animation clip, we obtained its average user rating, its 3D head rotation (Euler angles) sequence, and its speech pitch and RMS energy sequences. Our quantitative analysis is conducted in the following two ways. First, we analyze the quantitative correlation between the user rating and joint features of head motion and speech signals (in particular, the pitch and RMS energy). Second, we perform a frequency-domain analysis to study the association between the user ratings and head motion patterns.

Correlation between Speech-Head Motion Features and Human Perception

As reported in previous literature [5, 9], head motion has a strong correlation with the pitch (also known as the dominant frequency of repetition of speech signals) and the RMS energy of speech signals. Inspired by their work, we look into the quantitative association between human perception and joint audio-head motion feature. Specifically, we use Canonical Correlation Analysis (CCA) to measure the correlation between head motion and the two types of fundamental speech features: pitch and RMS energy.

For each used audio-head animation clip, we compute its CCA coefficient between its head rotation Euler angle sequence and its pitch sequence. Then, we plot the computed CCA coefficients versus their corresponding average user ratings (a total of 60) in Figure 2, where X axis represents the computed pitch and head motion CCA coefficients (i.e. between head motion Euler angles and pitch) and Y axis represents the average user ratings of these clips. As clearly shown in Figure 2, the highly rated animation clips (e.g., the audio-head animation clips generated by the original data approach) typically have high pitch and head motion CCA coefficients. Meanwhile, the lowly rated clips (e.g., those generated by the randomly generated approach) have low pitch and head motion CCA coefficients.

We also perform Pearson analysis to quantify the linear relationship between the computed pitch and head motion CCA coefficients and the subjective user ratings. The computed Pearson coefficient (in Figure 2), r = 0.731, which measurably proves the existence of a linear correlation between the pitch and head motion CCA coefficients and the perceptual user ratings.

The above validated tight coordination between avatar speech

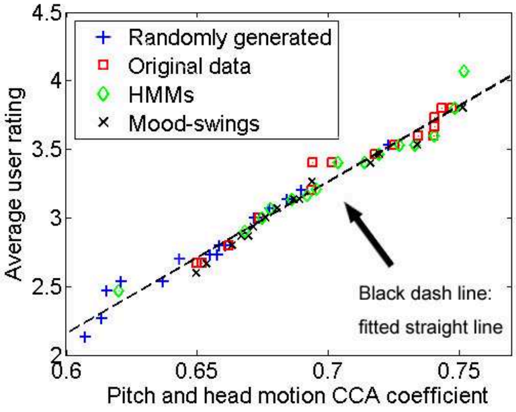

Figure 2. Plotting of the computed pitch and head motion CCA coefficients (X axis) and the obtained average user ratings (Y axis). The fitted line (the black dash line) shows the existence of an approximately linear correlation between the CCA coefficients and the user ratings.

and head motion implies the importance of precise timing in avatar head movement generation. Some previous head movement generation approaches (e.g. [1]) employ a delayed motion predictor for real-time applications. Our finding suggests that such delayed head movement generation strategy will potentially degrade human perception due to the importance of precise timing between head rotation and pitch, and for talking avatar head motion generation, designers need to heavily exploit vocal prosody timing (such as prosody-driven) in order to build a more perceptually believable head motion predictor. Also, our finding further gives a new and precise linear correlation between user ratings and the pitch and head motion CCA coefficients.

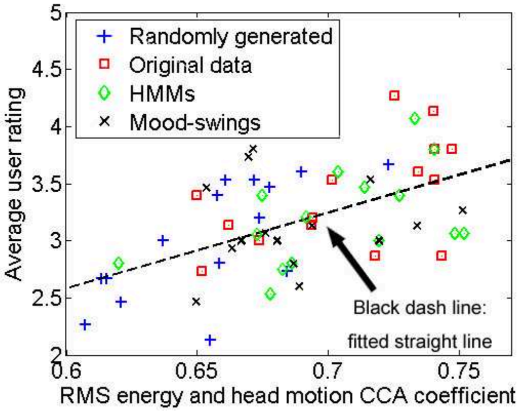

Figure 3. Plotting of the computed RMS energy and head motion CCA coefficients and the corresponding average user ratings. The fitted line (the black dash line) indicates that it is difficult to find a simple (e.g. linear) correlation between RMS-head motion CCA coefficients and the perceptual user ratings.

In a similar way, we also compute the CCA coefficients between head rotation Euler angles and the RMS energies of speech signals for all the 60 clips. Figure 3 plots the computed RMS energy and head motion CCA coefficients versus the obtained average user ratings. As shown in Figure 3, contrasting to the linear correlation between the user ratings and pitch-head motion CCA coefficients, we observe that such a linear correlation does not exist between the user ratings and RMS energy-head motion CCA coefficients. Our finding is contradicted with the previous results by Munhall et al. [5] that reported a strong correlation between RMS energy of speech signals and head movements helps to improve the auditory speech and non-verbal perception, and their evaluation experiment was limited to the case of a single sentence. Based on our finding, we argue that the quantitative relationship between RMS energy-head motion CCA correlation and human perception significantly varies, and it certainly does not fall into a simple correlation (e.g. linear), as in the case of the pitch and head motion CCA coefficients.

Frequency-Domain Analysis of Avatar Head Motion

As reported by Azuma and Bishop [1], human beings are more sensitive to the velocity and acceleration of head movements than static head poses. Thus, in this work, we use a frequency-domain analysis method to study the impacts of head motion spectrum patterns on human perception. Although frequency-domain analysis is not a new technique used in HCI community, previous frequency-domain analysis for head motion focused on analyzing the predicted results from linear head motion predictors [1], while our analysis intends to quantify the affects of frequency-domain head movements on human perception.

Our frequency-domain analysis of head motion (precisely, head rotation in our work) first separates 3D head motion Euler angles into three 1D signals inspired by the work of Azuma and Bishop [1]. This makes the analysis more accurate on each rotation axis. Then, we apply the Fourier and Z-transforms to the head rotation vector of each animation clip to derive its corresponding frequency spectrum. A classical Fast Fourier Transform (FFT) algorithm is used to compute the spectrum. The FFT point parameter is set as the smallest power of two that is greater than or equal to the absolute value of frame length of each animation clip. Figure 4 shows the frequency-domain analysis results of all the 60 clips, with an order from the highest user rating (4.2) to the lowest user rating (2.1). For a better illustration, in Figure 4, we only plot certain magnified part of head rotation spectrum with the squared magnitude less than 5 degree. Note that the degree unit (Y axis in Figure 4) is the squared magnitude of three Euler angles.

As shown in Figure 4, most of the highly rated audio-head animation clips have a low frequency pattern (i.e., most of the head motion frequencies are less than 12 Hz). We found that in the highly rated clips (i.e. the user rating is in the range of [3.5, 4.2]), 91.93% of their head motions are less than 10 Hz in the frequency domain, which suggests that natural head motions typically have low-frequency characteristics. By contrast, we also computed the frequency percentages of the lowly rated clips (i.e. the user rating is in the range of [2.1, 2.8]), and found that 19.39% of their head motions have are larger than 12 Hz in the frequency domain, which indicates that unnatural head motions usually enclose a significant portion of high-frequency movements (i.e. moving in a small range rapidly). For instance, in Figure 4, some of the lowly rated clips (<2.8 average user rat-

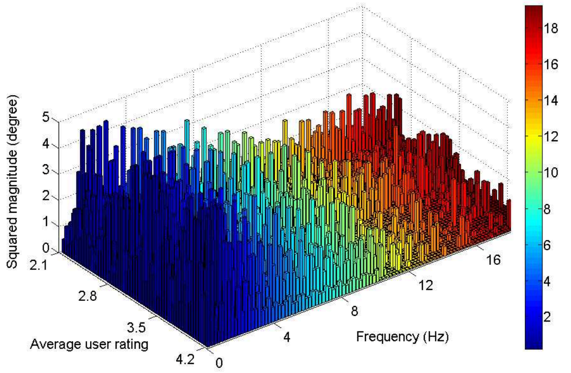

Figure 4. The frequency-domain analysis results of all the 60 clips. For a better illustration, we only plot certain magnified part of head motion spectrum with the squared magnitude less than 5 degree. The right color bar visualizes the head motion frequency range: from 0 Hz (blue) to 20 Hz(red).

ing) have a small range of head movements (<2.0 degree) with >14 Hz frequencies.

As pointed out by Grossman et al. [4], there is a critical instinctive mechanism that enables human beings to anchor their balance when the view is shifting, called Vestibulo Ocular Reflex (VOR). Therefore, during head rotations, if the maximum head velocity exceeds the range, the VOR mechanism will stabilize the motion spontaneously. Thus, our finding in this work is consistent with their results [4]. Furthermore, our quantitative analysis finding also suggests that 0∼12 Hz is the comfortable zone in which VOR enables human beings to maintain natural head movements. This leads to one important implication for avatar-based human-computer interfaces, that is, since humans tend to give a lower rating to head motions with a larger portion of high frequencies, synthetic talking avatar head motions should be smoothed or simply cropped as a post-processing step. In this way, more natural avatar head movements can be obtained.

DISCUSSION AND CONCLUSIONS

Our quantitative analysis results clearly show that the coupling between the pitch of speech signals and head motion has a strong correlation with human perception. Generally, a higher coupling coefficient between pitch and head motion leads to a higher user rating on the audio-head animation. Moreover, our frequency-domain analysis results show that perceived-natural head motions mainly consist of low-frequency components (e.g. <10 Hz). By contrast, a measurable portion of high-frequency signals (e.g. >12 Hz) are found in unnatural head motions.

Animation clips used in current study only enclose a single avatar. However, we believe that our current findings can be soundly applied and generalized to more realistic scenarios due to two main reasons. First, although we only acquired the head movements of one female subject, she was participating in a natural two-party conversation (with another auxiliary subject that was not motion captured) during the data capture procedure. Second, in the user study, participants sat in front of the synthetic avatar clips (about 3 meters away from the screen, refer to Figure 1), which was approximately a natural face-to-face, two-party, human-avatar conversation scenario. In the future, we plan to further extend the proposed methodology to systematically study head movement patterns in multi-party conversations.

ACKNOWLEDGMENTS

This work is supported in part by NSF IIS-0914965, Texas NHARP 003652-0058-2007, and grants from Google and Nokia. 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. R. Azuma and G. Bishop. A frequency-domain analysis of head-motion prediction. In *Proc. of SIGGRAPH'95*, pages 401–408, 1995.
2. C. Busso, Z. Deng, U. Neumann, and S. Narayanan. Natural head motion synthesis driven by acoustic prosody features. *Journal of Computer Animation and Virtual Worlds*, 16(3-4):283–290, July 2005.
3. E. Chuang and C. Bregler. Mood swings: expressive speech animation. *ACM Trans. Graph.*, 24:331–347, April 2005.
4. G. E. Grossman, R. J. Leigh, L. A. Abel, D. J. Lanska, and S. E. Thurston. Frequency and velocity of rotational head perturbations during locomotion. *Experimental Brain Research*, 70(3):470–476, 1988.
5. K. G. Munhall, J. A. Jones, D. E. Callan, T. Kuratate, and E. Vatikiotis-Basteson. Visual prosody and speech intelligibility: Head movement improves auditory speech perception. *Psychological Science*, 15:133–137, 2004.
6. E. Wang, C. Lignos, A. Vatsal, and B. Scassellati. Effects of head movement on perceptions of humanoid robot behavior. In *HRI'06: Proc. of ACM SIGCHI/SIGART Conf. on Human-robot interaction*, pages 180–185, 2006.
7. M. Williams, S. Moss, and J. Bradshaw. A unique look at face processing: the impact of masked faces on the processing of facial features. *Cognition*, 91(2):155–172, 2004.
8. N. Yee, J. N. Bailenson, and K. Rickertsen. A meta-analysis of the impact of the inclusion and realism of human-like faces on user experiences in interfaces. In *CHI'07*, pages 1–10, 2007.
9. H. C. Yehia, T. Kuratate, and E. Vatikiotis-Basteson. Linking facial animation, head motion and speech acoustics. *Journal of Phonetics*, 30:555–568, 2002.
10. C. Yun, Z. Deng, and M. Hiscock. Can local avatars satisfy a global audience? a case study of high-fidelity 3d facial avatar animation in subject identification and emotion perception by us and international groups. *ACM Computer In Entertainment*, 7(2):1–26, 2009.