HEAD MOVEMENT SYNTHESIS BASED ON SEMANTIC AND PROSODIC FEATURES FOR A CHINESE EXPRESSIVE AVATAR

Shen Zhang[1, 2], Zhiyong Wu[2], Helen M. Meng[2] and Lianhong Cai[1]

[1]Department of Computer Science and Technology
Tsinghua University, 100084 Beijing, China
[2]Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong, HKSAR, China
zhangshen05@mails.tsinghua.edu.cn, john.zy.wu@gmail.com,
hmmeng@se.cuhk.edu.hk, clh-dcs@tsinghua.edu.cn

ABSTRACT

This paper proposes an approach for text-to-visual speech synthesis, where the synthetic head movements are rendered with an expressive talking avatar speaking Cantonese Chinese. The input text consists of descriptive information sourced from the Hong Kong tourism domain. The text is segmented into prosodic words (PW) and we adopt the PAD model [7] to describe the expressivity of a prosodic word based on its semantics. Within the PW, we consider two prosodic features relevant to head movement synthesis, namely, the stress and tone of the Chinese syllable. We designed and recorded an audiovisual speech corpus and analyzed the data to derive statistical correspondences between different (P,A) values for a Chinese prosodic word and head movement coordinates. These statistics help parameter selection in a sinusoidal movement model. Corpus analyses also enable us to locate "peak points" of head movements that are synchronized with prosodic features within a prosodic word. These help the design of three heuristics that control head movements within a prosodic word. Perceptual evaluation based on the expressive talking avatar shows that head movement synthesis can raise the MOS by 1.04 points on average, when compared to the baseline which only shows lip articulations without head movements.

Index Terms—visual prosody, PAD emotional model, talking head, text genres, expressivity

1. INTRODUCTION

We all move in various ways while we speak, including head movements, facial expressions, hand gestures etc, which contributes to expressivity beyond the speech signal. These visual movements are analogous to prosody in speech, to which some refer as "visual prosody" [1]. Visual prosody is important for visual speech synthesis [2], as it can make a talking head (avatar) more engaging, as well as convey additional non-verbal information. Previous research on visual prosody includes mostly qualitative analysis [1, 3] or correlations with acoustic features [4, 5].

This work explores the use of semantic and prosodic features for text-to-visual-speech (TTVS) synthesis. In particular, we focus on head movements. Our text corpus is sourced from the Hong Kong tourism domain and we focus on text that is descriptive in nature [6]. The textual input is segmented into prosodic words (PW) and we adopt the PAD model [7] to describe the expressivity of a prosodic word for head movement synthesis. Within the prosodic word, we consider two main prosodic features for head movement synthesis, namely, the stress and tone of the Chinese syllable. We designed and recorded an audiovisual speech corpus in Cantonese, a major dialect of Chinese. Analyses of this data enable us to derive head movement statistics (including *amplitude* and *average positions*) corresponding to different (P,A) combinations for a Chinese prosodic word. These are incorporated as parameters in a sinusoidal movement model. Corpus analysis also enables us to design heuristics to synchronize the synthetic head movements with the syllable stream.

In the following, we will focus on annotating the expressivity and prosodic features for textual input, capturing the visual correlates of semantic expressivity, finding the temporal relationship between prosodic features and head movements, as well as applying our approach to text-to-visual-speech synthesis.

2. TEXT GENRES AND VISUAL SPEECH CORPUS

The text corpus is selected from the Hong Kong tourism domain [6], which can be categorized into three different text genres, i.e. the descriptive genre, which presents description of scenic spots; the informative genre, which provides information or facts; as well as the procedural genre, which gives directions or guidance. The text for each scenic spot is organized into paragraphs.

The visual speech corpus is created by recording twelve paragraphs of text with all the three text genres covered. A female native Cantonese-speaker was invited for audio and video recording. The speaker was asked to sit in front of a teleprompter, keeping her eyes looking straight at the camera. The video frame size is 720x576 pixels and the face size occupies about 300x400 pixels. Total duration of the recording is about 10 minutes. Before recording, the speaker was required to read through the text and understand its meaning. During recording, the speaker was asked to speak expressively.

The scope of this study focuses on the descriptive text genre, because it contains many more descriptive words than the other genres and thus has a wide coverage on different expressivity. The corresponding visual speech corpus also exhibits more pronounced head movements. There are four paragraphs of descriptive text genre in our corpus, which consist of 94 Chinese prosodic words and 460 syllables in total.

3. TEXT ANNOTATION

The text paragraphs are segmented into prosodic words by a home-grown text analysis tool [6]. We devised a procedure to annotate the semantic expressivity for each prosodic word and prosodic features for each syllable, as follows:

The PAD emotional model proposed by Mehrabian [7] is adopted to annotate expressivity. The PAD model describes the emotional state along three nearly independent dimensions: "pleasure-displeasure" (P), "arousal-nonarousal" (A) and "dominance-submissiveness" (D). For text-to-expressive speech synthesis, we defined a set of principles to annotate the P and A values for each prosodic word [6], resulting in four different combinations, namely (P=0 and A=0), (P=0 and A=0.5), (P=1 and A=0.5), (P=1 and A=1). Our textual corpus for video speech recording is a subset of that in [6] as we focus only on the descriptive genre of text. Statistics are shown in Table 1. Table 2 gives an example of an annotated sentence.

Table 1. *Statistics of (P,A) annotation for prosodic words (PW) in the descriptive text sub-corpus*

(P,A)	(0,0)	(0,0.5)	(1,0.5)	(1,1)
#PW(total 94)	8	50	17	19
% of occurrence	8.5	53.2	18.1	20.2

Table 2. *Example of an annotated sentence about Victoria Peak. The tabulated Chinese prosodic words (PW) from left to right, may be translated as: "Victoria Peak", "is Hong Kong's", "most popular", "scenic spot", "climb up", "can overlook", "submontane", "row upon row of (skyscrapers)"*

PW	太平山頂	是	香港	最受歡迎的	名勝景點之一	登臨其間	可俯瞰	山下	鱗次櫛比的
(P,A)	(0,0.5)	(0,0)	(1,1)	(1,1)	(0,0)	(1,0.5)	(0,0)	(1,1)	

We choose the tone and stress as prosodic features of each syllable, since the head movements are found to be simple motion patterns which are repeatedly applied in synchrony with these prosodic features during speaking. The tone of syllable is provided by the text analyzer of our text-to-speech synthesis (TTS) engine [9]. The stress is much related with the semantic meaning of the text. In our corpus, the *superlative words* such as "最(most)", "极(super)", "很(very)", "非常(very)", etc are annotated as stressed words.

4. ANALYSIS ON VISUAL FEATURES

4.1. Visual Features Extraction

We use the OpenCV face detection tools [8] to extract the head movements in x-y plane for each video frame in our visual speech corpus. Head movements along the z-axis are found to be insignificant in these frames. The detection result is a rectangle indicating the face region. Head movements are calculated by the rectangle's displacement between current frame and the first frame of the video as illustrated in Figure 1. It should be noted that the vertical displacement is calculated by comparing the upper edges of the rectangles, because the lower edges are affected by mouth movements during speaking.

For each frame, we obtain the relative measurements of *vertical head movement* (ratio of vertical displacement to the rectangle height of the first frame) and *horizontal head movement* (ratio of horizontal displacement to the rectangle width of the first frame). The measured head movement results are shown in Figure 2.

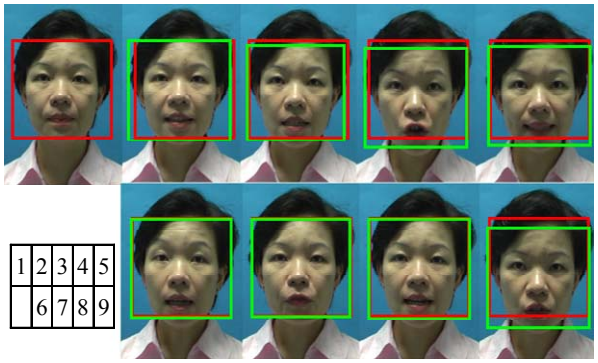

Figure 1. *Measuring head movements from the visual speech corpus by rectangle displacements. The first picture is the first frame of the video; the others correspond to the frames with the lowest vertical head positions in each prosodic word in Table 2.*

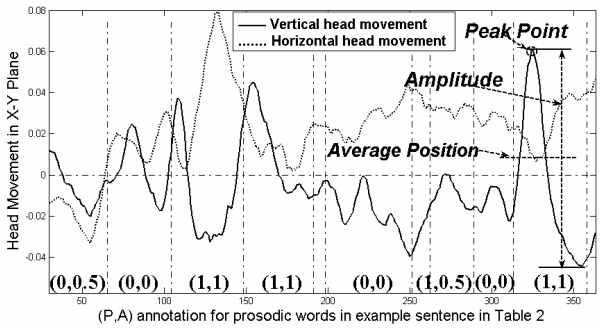

Figure 2. *Head movements in the x-y plane. The solid curve indicates downward positive displacement. The dotted curve indicates leftward positive displacement. The vertical dash-dot line indicates the boundaries of prosodic words. (P,A) values are marked below each prosodic word segment, based on the example sentence in Table 2.*

We studied the head movements for each prosodic word. The prosodic word is chosen as the basic unit for visual analysis, not only because it connects the semantic expressivity with speech elements, but also because we find repetitive motion patterns corresponding to prosodic words. Two main head motion patterns are observed: the head nod and the diagonal motion. The head nod is illustrated as the peak-valley pair for vertical head movements (i.e. the solid curve in Figure 2). The diagonal motion is the combination of vertical and horizontal head movements in the x-y plane which is usually person-dependent, and thus is not our concern.

Based on the above observation, we propose three measurements (as shown in Figure 2) to describe the head nod (i.e. vertical head movement) within prosodic word:

(i) *Amplitude*: the distance between highest and lowest position;
(ii) *Average Position*: the mean position of head;
(iii) *Peak Point*: the lowest position of head.

The amplitude and average position describe the head nod in prosodic word, and the peak point is used for the temporal synchronization.

4.2. Visual Correlates of Semantic Expressivity

The prosodic word bridges the semantic expressivity with visual features. For each prosodic word, we extracted the *amplitude* and

average position of the head nod. The prosodic words are classified into four sets according to their (P,A) values. For each set of prosodic words, we obtain the mean and standard derivation of the two visual features as shown in Table 3.

From Table 3, we find that for different (P,A) values, the amplitude of head nodding falls into different ranges. High values in pleasure (P) and arousal (A) usually lead to large nodding amplitudes. Statistics in Table 3 enable us to synthesize head movements with different amplitudes according to different (P,A) values. The synthesis methodology assumes that while people move their heads they maintain roughly the same overall head position.

Table 3. *Statistics of head nod Amplitude and Average Position for prosodic words (PW) with different (P,A) values*

(P,A)	PW num	Amplitude		Average Position	
		Mean	Std	Mean	Std
(0, 0)	8	0.039	0.014	-0.004	0.011
(0, 0.5)	50	0.058	0.025	-0.025	0.020
(1, 0.5)	17	0.059	0.028	-0.015	0.017
(1, 1)	19	0.081	0.037	-0.012	0.019

4.3. Head Movements within Prosodic Word

We also study the temporal relationship between head movements and prosodic features throughout the course of a prosodic word.

From the visual speech corpus, we find the speaker usually emphasizes one particular syllable a multi-syllabic prosodic word, which can be used as prime candidate for placing peak points of head positions. Figure 3 shows the prosodic features and peak points found in the example sentence in Table 2. Statistics of appearance of peak points on prosodic features (tone and stress) is shown in Table 4. About 63.8% of the peak points fall on syllables annotated with stress or falling tone. Another 16.0% of the peak points fall on the first syllable of the prosodic word. We also find that the speaker always emphasizes stressed syllables rather than syllables with falling tone. This observation can be used in the synthesis of head movements throughout a prosodic word.

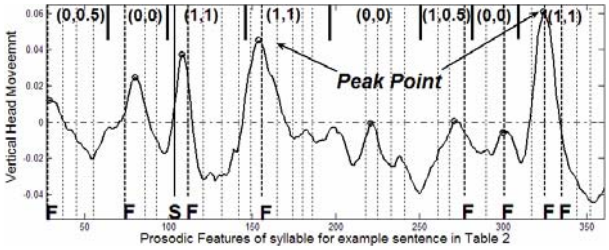

Figure 3. *Peak points and prosodic features found in the example sentence in Table 2. The bold solid vertical lines mark the stressed syllables (marked as S); the bold dash lines mark the syllables with falling tone (marked as F) and the dotted lines mark syllable boundaries. The (P,A) values for each prosodic word is marked on the top.*

Table 4. *Statistics of the correspondence between peak points and prosodic features of syllable*

Syllables in PW	Stressed syllable	Syllable with falling tone	First syllable	Others
Peak Points	20	40	15	19
% of occurrence	21.3	42.5	16.0	20.2

5. SYNTHESIS OF PROSODIC HEAD MOVEMENTS

Based on the analyses of visual correlates of semantic and prosodic features, we propose to use semantic expressivity in terms of (P,A) to modulate the head movements for each prosodic word, as well as the prosodic features (tone and stress) for temporal synchronization between head movements and the prosodic word (textual) sequence.

5.1. Head Movement Modulation for Prosodic Words

Head movement modulation is synthesized based on the (P,A) values of each prosodic word. The *sine* function is chosen to simulate the head nodding (equation 1):

$$y = Amp \times sin(\omega x + \phi) + Avg \quad (1)$$

where, *Amp* is the amplitude and *Avg* is the average position. The output *y* is the head position along y-axis. The modulation is implemented by adjusting the *Amp* and *Avg* parameters for each prosodic word. Four different value ranges to *Amp* are defined for different (P,A) combinations and then a specific *Amp* value for each prosodic word is generated randomly in the predefined range according to its (P,A) value as shown in Table 5. The mean value of average position is directly used as the *Avg* parameter.

Table 5. *Amp range and Avg value used in synthesis*

(P,A)	*Amp* Range	*Avg* Value
(0, 0)	[0.025, 0.053]	-0.004
(0, 0.5)	[0.033, 0.082]	-0.025
(1, 0.5)	[0.031, 0.087]	-0.015
(1, 1)	[0.043, 0.118]	-0.012

5.2. Intra-Prosodic-Word Rules for Synchronization

The temporal synchronization of head movements throughout a prosodic word and thus with the text input is also important for natural head movement synthesis. Missed synchronization will create an erratic effect to the observer. We set $\omega = 1$ in equation 1 to limit the *sine* curve to one period per prosodic word. For the phase part (ϕ) of *sine* function, we propose intra-prosodic-word rules based on the statistical result of temporal relationship between peak points and prosodic features of syllable in Table 4. The peak of *sine* function is taken as *peak point*, and exclusive intra-prosodic-word rules are defined as follows:

- *Rule1*: Place the peak points for head positions on the syllable with stress (if any);
- *Rule2*: Place the peak point for head positions on the syllable with falling tone (if any);
- *Rule3*: Place peak point for head positions on the first syllable of prosodic word.

6. PERCEPTUAL EVALUATION

We render the synthesized prosodic head movement with a three-dimensional Chinese avatar for text-to-audio-visual speech synthesis [9]. Figure 4 illustrates the synthesized output to the example sentence in Table 2.

We also conducted a set of experiments to evaluate the naturalness of synthetic head movements accompanying speech. Twenty text paragraphs (with about 20 prosodic words and 92 syllables per paragraph on average) are selected from the descriptive genre of information from the Hong Kong tourism domain. Head movements are synchronized with the synthetic visemes of recorded speech. We invited 18 native speakers of Cantonese as subjects in

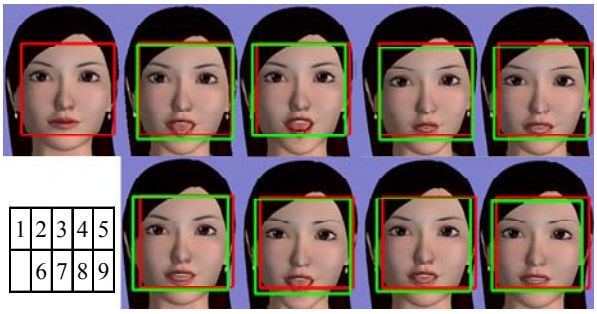

Figure 4. *Synthetic head movements for the example sentence in Table 2. The first frame indicates the neutral state; the others correspond to the frames with the lowest vertical head position in each prosodic word of the synthetic result. The rectangles help visualize the head movements.*

perceptual evaluation. Each subject was asked to listen to the recorded speech while watching three different sessions of synthetic visual speech for the same text paragraphs: (I) visual speech without any head movements; (II) visual speech with random head movements; (III) visual speech with prosodic head movements synthesized by our approach. The random head movements in (II) is synthesized using the equation 1, but without considering the semantic modulation and temporal synchronization. For each session, the subjects were asked to score the naturalness of head movements on a five level mean opinion score (MOS) scale: (5) expressive (4) natural (3) acceptable (2) unnatural (1) erratic. For each paragraph, the subjects were asked to give scores after comparing among all the three sessions.

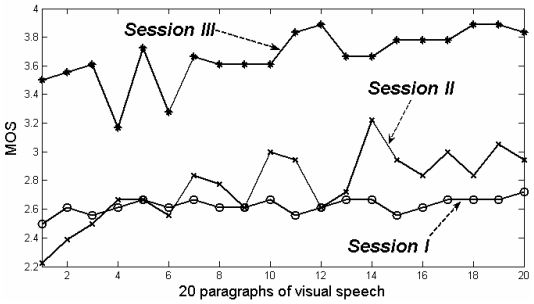

Figure 5 *MOS evaluation result of the synthetic visual speech*

The evaluation result is shown in Figure 5. The average MOS are 2.6 (session I), 2.8 (session II) and 3.7 (session III). A one-way ANOVA test revealed a significant main effect of head movement [$F(2,57)=192.97$, $p=0$]. *Post-hoc* analyses (Tukey HSD) showed that the synthetic result from session III is significantly better than result from session I and session II. As can be seen, the MOS for session I is the lowest, which suggests that synthetic lip-articulations without head movements make the avatar look stiff. The random head movements in session II give observers the impression that the avatar is alive. The head movements in session III seem coherent with the utterance semantics and obtains the highest score.

7. CONCLUSION AND FUTURE WORK

This paper proposes an approach for visual speech synthesis, namely, synthesizing head movements that are synchronized with the lip movements for a spoken utterance. Our approach adopts the PAD emotional model to describe the expressivity for each prosodic word and considers the tone and stress features within a prosodic word. Three visual features (*amplitude, average position* and *peak point*) are extracted from the video recordings to describe the head movements. We apply the *sine* function to simulate the vertical head movements within the prosodic word. The (P,A) values are used to modulate the head movements for each prosodic word. Intra-prosodic-word-rules are then applied to synchronize the head movements with the syllable sequence throughout the course of the prosodic word. Hence the head movements are synchronized with input text.

The proposed approach is rendered on a three-dimensional avatar. Perceptual evaluation shows that the synthesized visual speech improves the MOS by 1.04 points on a 5-point scale. Future work will focus on modeling the visual correlates of expressivity in a continuous PAD emotional space, as well as extending the approach to facial expression synthesis.

8. ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (NSFC) under grant No. 60433030 and the joint fund of NSFC-RGC (Research Grant Council of Hong Kong) under grant No. 60418012 and N-CUHK417/04. This work is affiliated with the Microsoft-CUHK Joint Laboratory for Human-centric Computing and Interface Technologies.

9. REFERENCE

[1] H.P. Graf, E. Cosatto, V. Strom, and F.J. Huang, "Visual Prosody: Facial Movements Accompanying Speech", in *Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 381-386, 2002.

[2] E. Casatto, J. Ostermann, H.P. Graf, and J. Schroeter, "Lifelike Talking Faces for Interactive Service", in *Proc of IEEE*, vol. 91, no. 9, 2003.

[3] K.G. Munhall, J.A. Jones, D.E. Callan, T. Kuratate, and E. Bateson, "Visual Prosody and Speech Intelligibility: Head Movement Improves Auditory Speech Perception", *Psychological Science*, vol. 15, no. 2, 2004.

[4] C. Busso, Z. Deng, U. Neumann, and S. Narayanan,"Natural Head Motion Synthesis Driven by Acoustic Prosodic Features", *Computer Animation and Virtual Worlds*, 2005.

[5] M. Costa, T. Chen, and F. Lavagetto, "Visual Prosody Analysis for Realistic Motion Synthesis of 3D Head Models", in *Proc. Int. Conf. on Augmented Virtual Environments and 3D Imaging*, pp. 343-346, 2001.

[6] H.W. Yang, H.M. Meng, and L.H. Cai, "Modeling the Acoustic Correlates of Expressive Elements in Text Genres for Expressive Text-to-Speech Synthesis", in *Proc. Int. Conf. on Spoken Language Processing*, pp.1806-1809, 2006.

[7] A. Mehrabian, "Pleasure-arousal-dominance: A General Framework for Describing and Measuring Individual Differences in Temperament", *Current Psychology: Developmental, Learning, Personality, Social,* vol.14, pp. 261-292, 1996.

[8] OpenCV Face Detection Tools
http://opencvlibrary.sourceforge.net/FaceDetection

[9] Z.Y. Wu, S. Zhang, L.H. Cai, and H.M. Meng, "Real-time Synthesis of Chinese Visual Speech and Facial Expressions using MPEG-4 FAP Features in a Three-dimensional Avatar", in *Proc. Int. Conf. on Spoken Language Processing*, pp.1802-1805, 2006.