# Low-level Characterization of Expressive Head Motion through Frequency Domain Analysis

Yu Ding*, Lei Shi*, and Zhigang Deng+, *Senior Member, IEEE*

**Abstract**—For the purpose of understanding how head motions contribute to the perception of emotion in an utterance, we aim to examine the perception of emotion based on Fourier transform-based static and dynamic features of head motion. Our work is to conduct intra-related objective analysis and perceptual experiments on the link between the perception of emotion and the static/dynamic features. The objective analysis outcome shows that the static and dynamic features are effective in characterizing and recognizing emotions. The perceptual experiments enable us to collect human perception of emotion through head motion. The collected perceptual data shows that humans are unable to reliably perceive emotion from head motion alone but reveals that humans are sensitive to the static feature (in reference to the averaged up-down rotation angle) and the dynamic features (which reflect the fluidity and speed of movement). It also indicates that humans perceive emotion carried in head motion and the naturalness of head motion in two different channels. Our work contributes to the understanding and the characterization of head motion in expressive speech through low-level descriptions of motion features, instead of commonly used high-level motion style (e.g. head nods, shakes, tilts, and raises).

**Index Terms**—Human Behavior, Head Motion, Expression, Emotion, Conversation, Frequency-domain, Discrete Fourier Transform.

## 1 Introduction

A large amount of research efforts have been devoted to the understanding of expression and the perception of emotion through nonverbal behaviors, including facial expression, head motion, and hand gesture. Most works rely on stylized motions to examine the emotional information carried in nonverbal behaviors. For example, facial expression is often regarded as the coordination of a few basic motions called Action Units (AUs), each of which describes the contraction of a basic muscular unit [1]. For instance, sadness can be displayed through the combination of AU1 (Inner Brow Raiser), AU4 (Brow Lowerer), AU15 (Lip Corner Depressor), and AU23 (Lip Tightener). In fact, numerous works have been developed to recognize AUs to automatically detect emotion. Head backward may transmit anger, surprise, and fear while head forward may link with sadness [2]. Arms' rising or stretching out to the front may be used to express joy [2]. The above behavioral descriptions rely on a predefined set of stylized or basic motions.

Recently, with the increasing availability of high-quality mocap data, various subtle and complex motion trajectories are captured. This may result in that a set of stylized motions may be inadequate to describe the diversity of mocap data and to characterize the subtlety of mocap behavioral data. On the other hand, some existing studies show that human behavioral expressions linking to the same emotion category share generally a core set of action-independent features [3] [4]. For example, the speed of knocking motion predominantly affects the intensity of activation in the perceived affect [5]. Sadness is displayed through slow body movements [6]. The work of [7] studies the characterization of body expressions of 8 emotions across 7 daily actions (e.g., lifting and throwing an object with one hand). The body motion is characterized by 8 body cues including power, fluidity, speech of movement, etc. In addition to the above works on emotional body expression, another work of [8] aims at portraying personality for virtual character by defining low-level parameters to effectively characterize behavioral dynamics including space (indirect vs. direct), weight (light vs. strong), time (sustained vs. sudden) and flow (free vs. bound). The works mentioned above made efforts on exploring action-independent features. Hence, there is a clear need of creating *an action-independent model of how affect is expressed rather than building a recognition system for each type of action* [9], which relies on motion characterization instead of action detection.

To the best of our knowledge, few previous efforts have been focused on the characterization of emotional head motion. Most existing works employ the temporal 3-dimensional rotation angles to characterize head motion. Moreover, the work of [10] describes head motion in an utterance with the means, the standard deviations, the ranges, the maximums and the minimums of 3-dimensional rotation angles. However, these proposed features may not be intuitively linked with the perception of emotion and could not further explain how head motions affect the perception of emotion.

Our work aims to explore the low-level characterization of head motion in expressive speech and to further understand meaningful head motion cues linked to the perception of emotion. The characterized cues are not limited to any specific high-level types of head motion (e.g. head nodding and shakes). They can be used to describe any head motion

- * *The two co-first authors contributed equally to this work.* + *Correspondence E-mail: zdeng4@uh.edu.*
- *Y. Ding, L. Shi and Z. Deng are with the Department of Computer Science at the University of Houston, Houston, Texas, USA.*

*Manuscript received xxxx xx, 20xx; revised xxxx xx, 20xx.*

in a low-level. Indeed, in addition to head nodding and shakes, other categories of head motions are also helpful to transmit emotional information [11].

As the first step, our work employs the dynamic features estimated through Fourier transform as well as the static feature called the direct current component (see Section 3) to characterize head motions and to study their link with emotion perception. This work relies on a mocap dataset recording head motion data and the simultaneous expressive speech, which allows us to quantitatively analyze head motion in utterances. Our study also relies on a human-like virtual character driven by human motion data with or without manipulation, which allows us to characterize and quantify emotion perception through head motion.

To achieve the above goal our study is dedicated to the objective analysis and the perceptual understanding of the static and dynamic features of head motion in expressive speech. In the objective analysis, machine learning algorithms are employed to recognize emotion type from the static and dynamic features of head motion. The recognition accuracies suggest that the static and dynamic features provide a promising characterization of expressive head motion and that they may carry meaningful emotion information. The perceptual experiments are to explore perceptual aspects of emotional information carried in the static and dynamic features.

The main contributions of our study consist in an exploration of the action-independent characterization of head motion in expressive speech and an experimentally-grounded explanation of how humans perceive and decode emotions embodied in head motions.

## 2 Background

Previously, a few works have been dedicated to the automatic generation of head motion based on expressive speech [10] [12]. Although the produced head motions are able to transmit appropriate emotions, the use of 3-dimensional rotation angles in their works cannot allow us to understand how emotional information is encoded into head motions and how head motions are perceived by humans. In the previous works, head motion is usually featured by a temporal sequence of three head rotation angles at each time frame. While three head rotation angles describe the spatial position of the head at each time frame, humans may not be able to intuitively correlate a static spatial position of the head with emotion in an utterance.

In the above works [10] [12], a short utterance lasting for several seconds is labeled by one emotion state. This suggests that the emotion may be stable within a short utterance. To find out the correlation between head motion and the emotion state in a period of a few seconds, the temporal head rotation angles at each time frame may be inappropriate. Therefore, a research question naturally comes up: *Can we characterize head motion with stable features which are invariant in an utterance to reflect the emotion state?*

In our study, we use the cues of the static and dynamic features to represent head motion, as these cues have the nature of stability in a lasting period. They could be a sound candidate to reflect the emotion state in an utterance. To explore the link between these cues and the emotion state, temporal head rotation angles are transformed into a set of static and dynamic features using Fourier Transform.

## 3 Related Work

This section briefly describes Discrete Fourier Transform (DFT) and then reviews previous works on human motion data processing using DFT. Next, we will describe previous works that explicitly consider emotional information to investigate head motion.

### 3.1 Discrete Fourier Transform

DFT has been widely used in signal processing community. It is able to decompose a one-dimensional temporal signal into a Direct Current (DC) component and a series of harmonically related sinusoidal components, which allows us to view a temporally-based sequence in a frequency-based domain. Assuming that a temporal sequence, $L$, consists of $T$ values at $T$ time frames with its sampling rate of $f_s$, $L$ is represented as $L = [l_1, l_2, ..., l_t, ..., l_T]$. Using DFT, $L$ is described as the sum of the DC component and a series of harmonically related sinusoidal components, which is represented as $l_t = A_0 + \sum_{n=1}^{N} A_n \sin(2\pi \cdot n \cdot f_0 \cdot t + \phi_n)$. $A_0$ is the DC component referred to as the averaged value of the temporal sequence. $n$ is the index of $s$ from 1 to $N$; and $N$ is the largest integer no more than $\frac{T}{2}$, denoted by $[\frac{T}{2}]$. $A_n \sin(2\pi \cdot n \cdot f_0 \cdot t + \phi_n)$ refers to as the $n\text{-}th$ sinusoidal component. $sin$ represents the sinusoidal function. $A_n$ and $\phi_n$ stand for the amplitude and the phase of the $n\text{-}th$ sinusoidal component, respectively. The $n\text{-}th$ sinusoidal component is a curve of sinusoidal signal with a frequency of $n \cdot f_0$. $f_0$ is defined by $\frac{f_s}{T}$ and referred to as the fundamental frequency. The $N\text{-}th$ sinusoidal component has the highest frequency of $\frac{f_s}{2}$ ($= N \cdot f_0 = \frac{T}{2} \cdot \frac{f_s}{T}$). For instance, if $f_s$ is $30Hz$, $N$ is 15 ($= \frac{f_s}{2}$) and $f_N$ is $15Hz$ ($= \frac{30}{2}$).

### 3.2 Discrete Fourier Transform for Human Motion

Several works have applied DFT to explore human motion data including head movement, walking motion, and shoulder vibration, which are briefly reviewed as follows. The work of [13] has employed DFT to analyze head rotation during walking and running. Their work shows that predominant frequencies of horizontal and vertical rotations are all within a 0.6-8.2$Hz$ range and that the predominant frequency of the pitch rotations is at least twice that of the yaw rotations during walking or running. Azuma and Bishop [14] have attempted to eliminate or reduce the effects of system delay in Head-Mounted Display systems. They made an effort to characterize head-motion using DFT. They take the advantage of frequency signals to quantify the ”jitter” often occurring in predicted signals. Ma et al. [15] shows that the head motions with high frequencies more than $12Hz$ are more likely perceived as unnatural. Their work suggests that the natural head motion is in the range between $0Hz$-$12Hz$. Tilmanne and Dutoit [16] reported that the dynamic features with low frequencies are predominantly observed in most styles of walking and that the other dynamic features with higher frequencies play a more important role in stylistic walking than in normal walking. Their work shows that various styles differ from each other

in the profile of the dynamic features. Niewiadomski et al. [17] employ DFT to separate torso leaning and shoulder vibration from human data captured with two markers on the top of two shoulders, which is driven by the shoulder and torso motions. In their work, torso leaning and shoulder vibration cannot be directly obtained from motion capture data, while they are distinguishable in the frequency-domain.

### 3.3 Head Motion in Emotion Perception

Several works [10] [12] explicitly take into account the emotional information when studying head motion with speech. The two works attempt to produce human-like head movements from provided speech input. The work of [10] explicitly utilizes the type of emotion as the input complementary to the given speech. The other work [12] validates the synthesized head motion data by explicitly taking into account its expressiveness of emotion. However, their works are not focused on understanding how head motions are linked to the expressiveness of emotion. Our work is to explore the relationship between head motion and, the expressiveness and the perception of emotion.

Other works have been dedicated to the study of the uni-modality of head motion in the perception of complex emotion, which typically does not occur in an utterance and is not related to expressive speech. For instance, the work of [11] demonstrated that the emotional information in head motion is complementary rather than redundant to the emotional content in facial expressions and that emotional expressibility of head motion is not limited to nodding and shakes but also other gestures such as head tilts, raises, etc. Moreover, the work of [18] shows that a raised head is correlated to happiness and 'superiority emotions' (e.g. pride, contempt), a bowed head is associated with 'inferiority emotions' (e.g. guilt, humiliation). The work of [19] shows that the head raised 15-20 degrees encourages the recognition of pride. The work of [20] reported that head position strongly affects the reaction to both anger and fear.

## 4 Data Acquisition and Processing

(a) pitch (b) roll (c) yaw

Fig. 1: Illustration of three-dimensional head rotation angles. The rotation angles are displayed with a human-like virtual character [21].

Our study relies on a human audiovisual dataset which collects the audiovisual expression data from a professional actress. The collection was performed in a laboratory setting. To record the utterance audio signals, the actress wore a close talking SHURE microphone with a 48k(*Hz*) sampling rate. She was equipped with a mocap system that captures 3-dimensional head rotation angles (see Figure 1) and facial expressions at 120*Hz*. In our work, DFT is applied to decompose each dimension of the captured temporal head rotation angles into the DC component called the static feature in our work, denoted by $s$, and a series of sinusoidal components. The amplitude of a sinusoidal component is viewed as a dynamic feature, denoted by $d$. According to the description of DFT in Section 3.1, $N$ stands for the total number of sinusoidal components and it is valued at $[\frac{T}{2}]$ for each utterance, where $T$ is the length of the utterance. The $N\text{-}th$ sinusoidal component has the highest frequency at $\frac{f_s}{2}$, where $f_s$ is the sampling frequency of the motion capture data.

In our dataset, the original value of $f_s$ is $120Hz$. It means that 60*Hz* ($\frac{f_s}{2}$) would be the highest frequency in the dynamic features extracted with DFT. On the other hand, the work of [15] has demonstrated that natural head motion is likely to be lower than $12Hz$ and that the head motions with frequencies more than $14Hz$ may lead to the perception of unnaturalness. This suggests that the human head moves probably with the frequencies lower than $12Hz$ or $14Hz$ and that head motions with higher frequencies may occur rarely. Moreover, video at 30*Hz* is able to adequately meet the requirement of human visual continuity [22]. Inspired by the works of [15], [22], our captured head motion data is down-sampled to 30*Hz* accordingly. Then, each dimension of the head motion is further decomposed into $s$ and a series of sinusoidal components with the highest frequency of $15Hz$, which is slightly higher than $14Hz$.

Considering that the motion capture data is collected from an actress who uttered towards the cameras in front of her, she could be assumed to be speaking in a two-party conversation. In a two-party conversation, the bias to right in yaw/roll rotation may lead to the same influence as that to left. As the first step, the current work focuses on the static feature of pitch rotation only (see Figure 1); the yaw and roll rotations are not investigated in our work. Particularly, $s^p$ stands for the static feature of pitch rotation. All the acronyms/abbreviations used in this paper are summarized in Table 1.

## 5 Objective Analysis

To learn more about the role of head motion in the expressiveness of emotion, objective analysis is carried out first. Similar to the work of [10], head motions are used to recognize the emotion of neutral, sadness, happiness, and anger. Different from [10], we utilize the static and dynamic features to characterize head motions. If the static and dynamic features result in a validated recognition accuracy, they would be useful features to characterize head motions with emotional information.

As introduced in Section 3, an angle stream with the length of $T$ is decomposed into $s$ and $[\frac{T}{2}]$ sinusoidal components. Then, 15 representative sinusoidal components are picked out and their amplitudes are viewed as the dynamic features, denoted by $d_{1\sim15} = [d_1, d_2, ..., d_{15}]$, for each utterance. $d_{1\sim15}$ are respectively extracted from the 15 non-overlapping frequency intervals with the width of $1Hz$ in the range from $0Hz$ to $15Hz$. $d_i$, $i = 1, ..., 15$, has the peak with the maximum amplitude in the interval from $(i-1)$ $Hz$ to $i$ $Hz$, which is done similarly in the works of [16], [17]. In particular, we use $d_i^p/d_i^y/d_i^r$ to denote $d_i$ in terms of pitch/yaw/roll rotation. Moreover, $d_{i1\sim i2}^p$ is defined as a

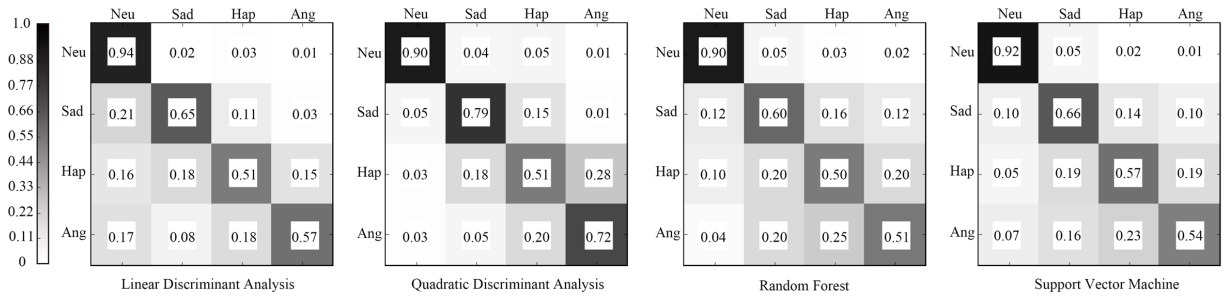



Fig. 2: The confusion matrices of the averaged accuracies of emotion recognition based on the static and dynamic features. The four confusion matrices report the accuracies of the four classifiers, respectively. Each confusion matrix reports the accuracies averaged over 50 experiments, which is characterized by the gray level and also the percentage.

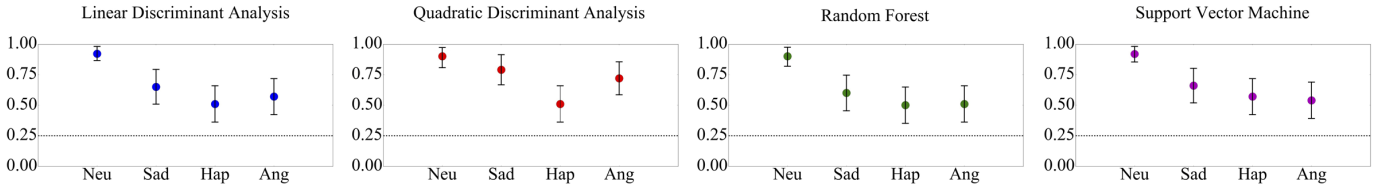



Fig. 3: The 95% confidence intervals of the recognition accuracies of the four classifiers. The dash lines highlight the chance level of recognition accuracy.

TABLE 1: Notation and Acronyms

| Acronym | Definition |
|---|---|
| DC | Direct Current (component) |
| $f_s$ | signal sampling rate |
| $s^p/s$ | the static feature of head pitch rotation |
| $d^i$ | the $i$-$th$ dynamic feature in the $i$-$th$ interval from ($i$-1)$Hz$ to $i$ $Hz$ |
| $d_i^p/d_i^y/d_i^r$ | $d_i$ in terms of pitch/yaw/roll rotation. |
| $d^{lo}/d^{in}/d^{hi}$ | the square root of 3D rotation angles. |
| sd | a concatenated vector of the static feature and the dynamic features |
| LF/IF/HF | low/intermediate/high frequency |
| $g_{LF}^d$ | the dynamic features with low frequency, referring to smooth-and-slow movement |
| $g_{IF}^d$ | the dynamic features with intermediate frequency, referring to moderate-fluidity-and-speed movement |
| $g_{HF}^d$ | the dynamic features with high frequency, referring to jerky-and-fast movement |
| $g^s$ | the still-frame movement, referring to $s^p$ |
| $g^{or}$ | the original motion, referring to the sum of the above four groups of $g_{LF}^d$, $g_{IF}^d$, $g_{HF}^d$ and $g^s$ |
| US-M | User Study on Perception of Motion Modality |
| US-Sta | User Study on Perception of Static Motion |
| US-Dyn | User Study on Perception of Dynamic Motions |
| $M^s/M^h$ | the uni-modality of speech/head motion |
| $M^{hs}$ | the bi-modality of head motion and speech |
| $C_s^{or}/C_h^{or}$ | Cond. of $M^s/M^h$ with original head motion data |
| $C_{hs}^{or}$ | Cond. of $M^{hs}$ with original head motion data |
| $C^{\overline{\mathrm{LF}}\cap\overline{\mathrm{S}}}$ | the rest by removing $g_{LF}^d$ and $g^s$ from $g^{or}$ |
| $C^{\overline{\mathrm{IF}}\cap\overline{\mathrm{S}}}$ | the rest by removing $g_{IF}^d$ and $g^s$ from $g^{or}$ |
| $C^{\overline{\mathrm{HF}}\cap\overline{\mathrm{S}}}$ | the rest by removing $g_{HF}^d$ and $g^s$ from $g^{or}$ |
| $C^{\overline{\mathrm{S}}}$ | the rest by removing $g^s$ from $g^{or}$ |

set of $d_i^p$ with $i$ ranging from $i_1$ to $i_2$; $d_{i1\sim i2}^y$ and $d_{i1\sim i2}^r$ are similarly defined. Finally, a concatenated vector, denoted by $sd$, is built for each utterance, including the 46 elements of $d_{1\sim15}^p$, $d_{1\sim15}^y$, $d_{1\sim15}^r$, and $s^p$. $\{sd\}$ denotes the set of $sd$, which is collected from all the utterances in the dataset.

Our objective analysis relies on the collected set of $\{sd\}$. To explore whether the static and dynamic features are associated with the emotional information, we build four classifiers separately to perform emotion recognition on $\{sd\}$, including Linear Discriminant analysis(LDA), Quadratic Discriminant Analysis(QDA), Random Forests (RF), and Support Vector Machine (SVM). In our experiments, the solver of LDA is singular value decomposition; RF consists of ten trees and takes Gini coefficient as the criterion to measure the quality of a split; SVM takes Radial Basis Function kernel as its kernel function.

**Experiments**: In each experiment, $\{sd\}$ is randomly split into 80% for training and 20% for testing, and each classifier performs a four-class classification. As mentioned above, four classifiers are respectively employed in our experiments. In total, $\{sd\}$ is randomly and independently split 50 times, each of which is used to conduct one experiment. Hence, 50 experiments are carried out with each classifier. Figure 2 reports the resulting confusion matrix that provides the averaged recognition accuracies over the 50 experiments for each classifier. To test whether the recognition accuracies are statistically higher than the chance level (25%, one out of four), the 95% confidence interval [23] of recognition accuracy is calculated for each emotion and for each classifier, reported in Figure 3.

**Results**: Figure 2 shows that the recognition accuracies of the four emotions are not identical. For example, while the recognition accuracy of neutral is in a range from 90% to 92% with the four classifiers, the recognition accuracy is from 60% to 79% for sadness, from 50% to 57% for happiness, and from 51% to 72% for anger. The highest accuracy for neutral, sadness, happiness, and anger is 92%, 79%, 57% and 72% among the four classifiers. Viewing the performances of the four classifiers, the averaged recognition accuracy is always higher than the chance level (25%), which

is validated by the 95% confidence intervals of recognition accuracy reported in Figure 3. In particular, it is observed that the recognition accuracy of neutral is higher than that of the other three emotions. On the other hand, Table 2 reports the 95% confidence intervals of error rate which are calculated based on the performances of the four classifiers. The results show that happiness is recognized as anger with the confidence interval from 14.7% to 26.3% and that anger is recognized as happy with the confidence interval from 15.6% to 27.4%. These two interval confidences contain the chance level of 25%; and the other ones are below the chance level. In other words, the confidence intervals of error rate are always no more than the chance level. The above results verify that the static and dynamic features can be used to characterize head motion and to recognize the type of emotion.

**Exploratory Study** (OB-Ex): The above objective experiments show the static and dynamic features provide important clues to identify emotion carried in head motion. We are aware that the promising recognition accuracy may be attributed to all the static and dynamic features or to a portion of them (e.g. only the static feature or a portion of the dynamic features). To figure it out, we validate the differences among the four emotions in the static and dynamic features. For example, if a significant difference in the static feature is validated, the static feature can be viewed as an important clue for emotion recognition; otherwise it is considered insignificant for emotion recognition.

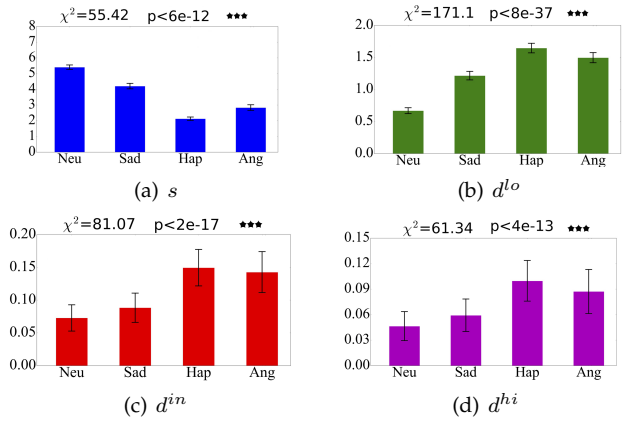



Fig. 4: Kruskal Wallis test results on the static feature ($s$), the dynamic features with low, intermediate, and high frequencies ($d^{lo}$, $d^{in}$, and $d^{hi}$). The error bars indicate the standard deviations. The Kruskal Wallis test results show that there is a significant difference among the four emotions for each feature.

During the validation, the Kruskal Wallis test for the difference among the four emotions is not performed for the 46 static and dynamic features. For the sake of simplicity, we calculate the square root of the 3-dimensional rotation angles at each time frame, and then use DFT to decompose the sequence of square root for each utterance into a static feature and a series of dynamic features with the lowest frequency of $\frac{f_s}{T}$ and the highest one of $15Hz$ (see Section 3). Finally, 4 representative static and dynamic features are selected including the static feature denoted by $s$, a low-dynamic feature denoted by $d^{lo}$, an intermediate-dynamic feature denoted by $d^{in}$, and a high-dynamic feature denoted by $d^{hi}$. $d^{lo}$, $d^{in}$, and $d^{hi}$ are taken as the ones with the highest frequency in the low frequency interval from $\frac{f_s}{T}$ to $5Hz$, in the intermediate frequency interval from $5Hz$ to $10Hz$, and in the high frequency interval from $10Hz$ to $15Hz$. Finally, $\{s\}$, $\{d^{lo}\}$, $\{d^{in}\}$, and $\{d^{hi}\}$ are gathered as the set of $s$, $d^{lo}$, $d^{in}$, and $d^{hi}$ from all the utterances in the dataset.

TABLE 2: The 95% confidence interval of error rate (%). The confidence intervals containing the chance level (25%) are highlighted with gray and the other ones are below the chance level.

| | Neu | Sad | Hap | Ang |
|---|---|---|---|---|
| **Neu** | - | [1 7] | [0.5 6] | [-0.5 3] |
| **Sad** | [7.2 16.8] | - | [8.9 19.1] | [2.8 10.2] |
| **Hap** | [4.4 12.6] | [13.1 24.4] | - | [14.7 26.3] |
| **Ang** | [3.8 11.7] | [7.5 17] | [15.6 27.4] | - |

The validation is performed four times separately on the four feature sets of $\{s\}$, $\{d^{lo}\}$, $\{d^{in}\}$, and $\{d^{hi}\}$. For example, $\{s\}$ are divided into four sub-groups according to the emotion of utterance. The validation is carried out to test the differences among the four sub-groups. Since we do not have prior knowledge on whether the distribution of $s$ in one emotion follows Gaussian or not. Therefore, the Kruskal Wallis test [24] is used to validate where the samples are independent of each other. Similarly, the validation is performed on the four feature sets. Figure 4 illustrates the results on the four feature sets. As can be observed in Figure 4, there are significant differences in $s$, $d^{lo}$, $d^{in}$, and $d^{hi}$ among the four emotions.

**Discussion**. Viewing the above results, head motion plays an important role in the expressiveness of emotion. The static and dynamic features provide meaningful information of the emotion simultaneously conveyed by speech. The promising recognition accuracies reveal that the static and dynamic features are informative clues for emotion recognition. This suggests that head motions perform with different patterns of the static and dynamic features for the four emotions, which is confirmed by our exploratory study. The recognition accuracy of neutral is the highest. It shows that head motions in the neutral state are more distinguishable from the other emotional states. This could suggest that humans are skilled in encoding emotional information into head motion, which is in line with the previous work of [10]. Our exploratory study demonstrates that the four categories of the static feature and the dynamic features with low, intermediate, and high frequencies are reasonable to explain the usefulness of head motion for emotion recognition. It suggests that humans intentionally move the head to express their specific emotions.

To further understand how the static and dynamic features are correlated to the emotional information in human perception, we conduct three intra-related subjective experiments in order to collect and analyze human perception of emotion on original or manipulated head motion data.

## 6 Subjective Experiments

Our objective experiments reveal that the emotional information may be decoded from head motion alone. In order to further understand whether humans are able to recognize emotion from head motion alone in an utterance, the first goal of our subjective experiments is to investigate whether human subjects are able to judge the type of emotion from head motion alone. The second goal is to look into how the static and dynamic features affect human perception of emotion.

As introduced in Section 3, a pitch, roll, or yaw head rotation sequence with the length of $T$ is the sum of a static feature and $T$ dynamic features. The $1^{st}/T\text{-}th$ dynamic feature has the lowest/highest frequency of $\frac{15}{T}/15Hz$ in our dataset. While the static feature indicates the still motion, the frequency of dynamic feature reflects the perception of fluidity (jerky or smooth) and speed (slow or fast) of movement. For instance, the work of [14] employs frequency to quantify the "jitter" occurring in movements. The dynamic features with low frequencies refer to jerky and slow movement; those with high frequencies reflect smooth and fast movement. Hence, according to the perception of movement fluidity and speed, the dynamic features in an utterance are categorized into three groups with low frequencies (LF), intermediate frequencies (IF), and high frequencies (HF), respectively, which are denoted by $g^d_{LF}$, $g^d_{IF}$, and $g^d_{HF}$:

- $g^d_{LF}$, referring to the components of smooth-and-slow movement, contains the dynamic features with low frequencies ranging from $\frac{15}{N}$(the lowest frequency) to $5Hz$;
- $g^d_{IF}$, referring to the components of moderate-fluidity-and-speed (neither smooth/slow nor jerky/fast) movement, contains the dynamic features with intermediate frequencies ranging from $5Hz$ to $10Hz$;
- $g^d_{HF}$, referring to the components of jerky-and-fast movement, contains the dynamic features with high frequencies ranging from $10^{th}$ to $15^{th}$(the highest frequency);

These categories enable us to investigate the link between the human perception of emotion and the human perception of movement fluidity and speed. In addition to the above groups of dynamic features, the other two groups are defined as follows:

- $g^s$, referring to the component of still movement, contains $s^p$, which is viewed as a group and refer to as a still-frame movement.
- $g^{or}$ refers to the original motion. It is viewed as the sum of the above four groups of $g^d_{LF}$, $g^d_{IF}$, $g^d_{HF}$ and $g^s$.

The second goal, mentioned before, of our subjective experiments is to investigate how $g^s$, $g^d_{LF}$, $g^d_{IF}$ and $g^d_{HF}$ affect human perception of emotion, which could reveal the perception of emotion from still movement, fluidity and speed of movement.

To achieve our goals, we conduct three intra-related perceptual experiments. The first user study is called *User Study on Perception of Motion Modality* (US-M). It is mainly to study whether human subjects are able to judge the emotion from the uni-modality of head motion data, denoted by $M^h$. US-M is concerned with the first goal. The second user study is called *User Study on Perception of Static Motion* (US-Sta). It looks into human perception's link with $g^s$. US-Sta is concerned with the second goal. The third user study is called *User Study on Perception of Dynamic Motions* (US-Dyn). It looks into human perception's link with the dynamic features. US-Dyn is also concerned with the second goal.

In the perceptual experiments, 20 participants are recruited, consisting of 8 males and 12 females with age ranging from 18 to 40 (M=27 years, SD=3.9 years), to rate their perceptions on a human-like virtual character displaying animations through designed online webpages. The participants are invited to complete the user studies in a quiet office. The participants can freely watch video clips of virtual character on one webpage as many times as they want. At the beginning of each user study, each participant fills out a demographic questionnaire concerning his/her age, gender, education level, occupation and country in which the participant spent the majority of his/her life. Our studies are focused on human perception of emotion linked with head motions but not on the appearance of virtual characters. The identical virtual character is employed in our perceptual experiments. In each animation clip, the virtual character displays head animation and/or simultaneously utters the corresponding acoustic speech. To suppress the potential effect of emotional information from facial expression (even neutral expression [25]), the facial expression region (including the eyes region) and the mouth region are intentionally masked with strong mosaic. Figure 8 shows the masked virtual character. We describe the three perceptual experiments as follows.

### 6.1 User Study on Perception of Motion Modality

US-M is conducted to learn whether human subjects are able to recognize the type of emotion according to $M^h$.

**Hypothesis**: in US-M, we formulate one hypothesis as follows:

- **M-H**: human subjects are able to reliably judge emotion from head motion alone in an utterance.

This hypothesis is formulated according to that the type of emotion can be reliably referred from head motion data (see the objective experiments). M-H would be confirmed if human subjects are able to judge emotion from $M^h$ with similar recognition accuracies as from the bi-modality of head motion and speech, noted by $M^{hs}$, or the uni-modality of speech, noted by $M^s$, as it is well known that humans are able to well infer emotion from $M^{hs}$ or $M^s$.

**Protocol**: The participant is instructed to watch clips of virtual character and then asked to judge the emotion of the virtual character (Figure 5). We provide the elements of the protocol we follow in this user study.

1) Stimuli: 12 utterances (3 sentences $\times$ 4 emotions) are randomly selected from the dataset. For each utterance, we take into account three conditions of $C^{or}_h$, $C^{or}_s$, and $C^{or}_{hs}$ where $C$ is the abbreviation of Condition; $h$, $s$, and $hs$ respectively refer to $M^h$, $M^s$, and $M^{hs}$; and $or$ refers to original data of motion and/or speech recorded in our dataset. $C^{or}_h$,

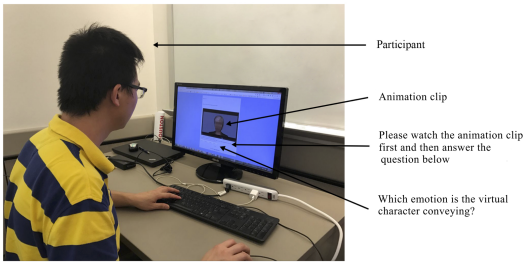



Fig. 5: A snapshot of a participant taking part in US-M.

$C_s^{or}$, and $C_{hs}^{or}$ stand for the original head data, the original speech data, and the original head-and-speech data. In total, 36 clips are created from 12 utterances for 3 different conditions, including 24 animation clips and 12 audio clips.

2) Procedure: The participant is invited to perceive the 36 clips, which are displayed randomly for each participant. After watching each clip, the participant is instructed to select the perceived emotion out of neutral, sadness, happiness, and anger. Figure 5 demonstrates a snapshot of a participant who is taking part in this user study.

**Results**: Each clip is judged 20 times by 20 participants. We average the recognition accuracies of the 3 clips which have the identical condition and emotion. In total, 9 averaged recognition accuracies are calculated and reported in Figure 6. The results show that $C_{hs}^{or}$ outperforms $C_h^{or}$ and $C_s$ and also $C_s$ performs better than $C_h$. In particular, although $C_s$ is scored slightly less than $C_{hs}$, the overlapping of the $95\%$ confident intervals of $C_{hs}^{or}$ and $C_s$ is observed for the four emotions. This means that there is no significant difference between $C_{hs}^{or}$ and $C_s$. Moreover, while the overlapping of the $95\%$ confident intervals of $C_s^{or}$ and $C_h$ is observed for neutral, sadness, and anger, it is unobserved for happiness. Particularly, although $C_h$ in neutral and anger has the accuracy more than the chance level ($25\%$), it is scored lower than the chance level ($25\%$) in sadness and happiness. The above observations indicate that *M-H is untrue.*

**Discussion-M**: The results above show that $M^{hs}$ and $M^s$ provide more emotional information than $M^h$. Humans are able to decode the emotional information from the coordinated head motion and speech audio. This shows that the speech audio predominantly affects the perception of emotion. Additionally, the results of $C_h$ reveal that humans are unable to reliably perceive the emotional information from $M^h$. This is inconsistent with the results of our objective experiments.

Humans are able to perceive anger and neutral better than happiness and sadness from $M^h$. It appears that humans encode more information into head motion when conveying neutral and anger than expressing sadness and happiness, or the human sensitivity to head motion depends on the perceived emotion. On the other hand, the objective results in Figure 4 show that the static and dynamic features in neutral and anger are not higher than those in sadness and happiness. Considering together the subjective and objective results mentioned above suggests that human perception is not straightforwardly correlated to the amplitude of the static and dynamic features. This also explains why humans are sensitive to subtle expressions with slight movement.

Since $M^h$ is inadequate to characterize the type of emotion, $M^{hs}$ is employed in US-Sta and US-Dyn. Moreover, considering that the speech audio is adequate to shape the type of emotion, in the other perceptual experiments we investigate the roles of $g^s$, $g_{LF}^d$, $g_{IF}^d$, and $g_{HF}^d$ in the perception of the level of emotion instead of emotion recognition.

### 6.2 User Study on Perception of Static Motion

US-Sta investigates the impact of $g^s$ on human perception of the level of emotion. $g^s$ refers to the static feature of pitch head rotation.

**Hypotheses**: US-Sta is carried out to investigate the hypotheses formulated as follows:

- **S-H$^{\text{Neu}}$/H$^{\text{Sad}}$/H$^{\text{Hap}}$/H$^{\text{Ang}}$** (intra-emotion): human subjects are sensitive to $g^s$ when perceiving the level of neutral/sadness/happiness/anger. For the sake of simplicity, **S-H** stands for a general symbol of the four hypotheses.
- **S-H$^{\text{InterEmo}}$** (inter-emotion): when perceiving the level of emotion, human sensitivity to $g^s$ is different depending on the perceived type of emotion.

S-H brings insight into $g^s$ for each individual emotion. It enables us to further understand the impact of $g^s$ on the perception of one emotion. Additionally, OB-Ex has shown that $g^s$ which human speakers encode has significant difference depending on the expressed emotion (see Figure 4(a)). S-H$^{\text{InterEmo}}$ takes insight into $g^s$ from the inter-emotions. It allows us to verify whether the human listener decodes $g^s$ in different manners depending on the perceived type of emotion.

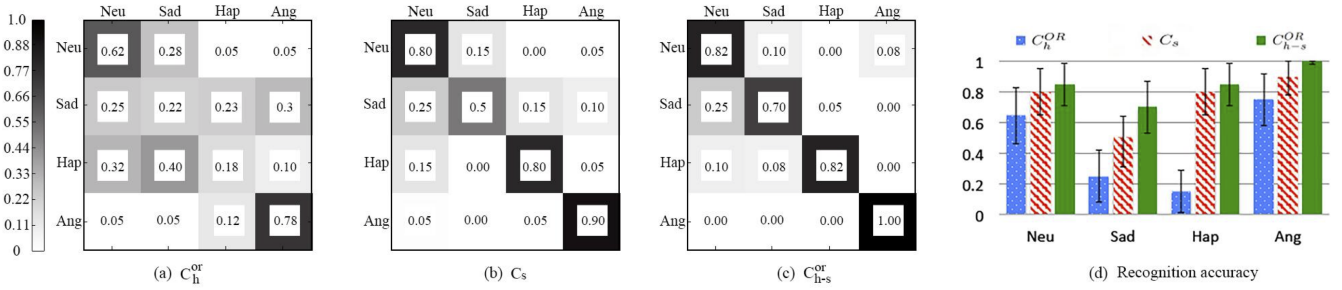
To confirm S-H, we compare the level of each of the four emotions with $g^s$ at different values (conditions). S-H would be confirmed if the human perception of the level of emotion is differential among the conditions of $g^s$. To confirm S-H$^{\text{InterEmo}}$, we compare the levels of the four emotions perceived across the conditions of $g^s$. S-H$^{\text{InterEmo}}$ would be confirmed if the human perception of the level of emotion is distinguishable among the four emotions.

If both S-H$^{\text{InterEmo}}$ and DC-H are verified, this would suggest that each emotion may have its own optimal value(s) of $g^s$ to enhance its expressiveness. Hence, an exploratory study is done in this user study to answer the question below:

- **DC-Ex**: What is the optimal value of $g^s$ (the static feature) that can most enhance the expressiveness of neutral/sadness/happiness/anger?

**Protocol**: The participants are asked to perceive the virtual character displaying human head data with manipulated $g^s$ and then to rate the level of perceived emotion. We provide the elements of the protocol we follow in this user study.

1) Stimuli: We investigate $g^s$ from $-15^\circ$ to $15^\circ$ with a step size of $5^\circ$, where $0^\circ$ denotes face straight forward. In total, $g^s$ has seven candidates (conditions)




(a) $C_h^{or}$ (b) $C_s$ (c) $C_{h\text{-}s}^{or}$ (d) Recognition accuracy

Fig. 6: Emotion Recognition in US-M. (a)-(c). The Confusion matrices of emotion recognition under $C_h^{or}$, $C_s$, and $C_{hs}^{or}$. (d). Recognition accuracies of $C_h^{or}$, $C_s$, and $C_{hs}^{or}$ (the values in the diagonal lines of the three confusion matrices.). The error bars indicate the $95\%$ confident interval.

Fig. 7: Reference image (front view and side view) of head pitch rotation ranging from $-15°$, $-10°$, $-5°$, $0°$, $5°$, $10°$, and $15°$. In our subjective experiments, the whole face region is intentionally masked with strong mosaic (see Figure 8).

visualized in Figure 7. To quantify the influence of $g^s$ on perception, twelve audiovisual utterances are selected from the dataset. They consist of three randomly selected sentences uttered four times with the four emotions. For each utterance, $g^s$ of the original head motion is removed and substituted by its seven conditions, respectively. Note that no other operation is performed on head pitch rotation, and no operation is performed on roll and yaw rotation values of the head as well as the speech audio signals. 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 $\times$ 4 emotions $\times$ 7 conditions) were created.

2) Procedure: US-Sta is designed to investigate the differences among the seven conditions of $g^s$. In this user study, the virtual character is used to display the 3-dimensional head rotation data under the seven conditions. Each webpage encloses seven animation clips under the seven conditions but with the same speech audio, in a random order. According to the aforementioned OB-Ex, the speech audio is adequate to characterize the conveyed emotion, it could be appropriate to assume that the seven animation clips on one webpage interface express the same emotion. This user study is to investigate the level of the expressiveness of emotion instead of recognition of the emotion type. To facilitate the participation, the participant is explicitly informed of the expressed emotion through texts.

In total, 12 webpages are created from the three sentences and their four emotions. The participant is instructed to view 12 webpages (see Figure 8). He/She is asked to rate the level of the expressed emotion for each animation clip using a 5-points Likert scale.

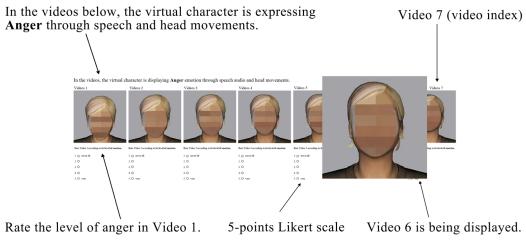



Fig. 8: A snapshot of the webpage used in US-Sta. One webapge encloses seven animation clips under the seven conditions but with the same speech audio. The clips are displayed separately with an individual button of Play. Only one clip can be displayed at one time. When a video clip is chosen to play, its visual dimension is automatically zoomed in. In the figure, the $6^{th}$ video clip is being played.

**Results on S-H$^{\text{InterEmo}}$**: To investigate the formulated hypothesis of S-H$^{\text{InterEmo}}$, we study whether the rated levels of emotion are differential among the four emotions. The rated level is computed under the seven conditions of $g^s$, respectively. For example, under the condition of $g^s$ at $10°$, the scores of the animation clips with $g^s$ at $10°$ are collected from the 12 webpages and then they are classified into four groups according to the type of emotion. Each group contains 60 scores from 3 utterances with the same emotion, which are rated by 20 participants. The validation of DC-H$^{\text{InterEmo}}$ is to test whether there is significant difference among the four groups of 60 scores for each condition of $g^s$. Each score is rated on one webpage and by one participant. These collected scores can be viewed as being independent of each other. Moreover, we do not know whether the distribution of the 60 scores in one group is Gaussian. Therefore, we use the Kruskal-Wallis Test, which is a rank-based nonparametric test to assess whether there are significant differences between two or more groups of an independent

TABLE 3: Statistical test results of the Kruskal-Wallis Test. The Kruskal-Wallis Test is used seven times for the seven conditions of $g^s$. Each time is to verify whether there is a significant difference in the score of the level of perceived emotion, under one condition of $g^s$, among the four emotions of neutral, sadness, happiness and anger. ** marks p<.01; *** marks p<.001.

| | $-15°$ | $-10°$ | $-5°$ | $0°$ | $5°$ | $10°$ | $15°$ |
|---|---|---|---|---|---|---|---|
| $\chi^2(3)$ | 88.85 *** | 49.44 *** | 34.34 *** | 30.21 *** | 28.89 *** | 21.09 *** | 14.09 ** |

variable [24]. In this user study, the independent variable is the type of emotion, which has four independent groups: 60 scores with neutral, 60 ones with sadness, 60 ones with happiness, and 60 ones with anger.

The Kruskal-Wallis Test is carried out seven times for the seven conditions of $g^s$. At each time, Kruskal-Wallis Test performs over the four emotions under one condition of $g^s$. The results are reported in Table 3. As can be seen, the Kruskal-Wallis tests show that there is a significant difference in the rated scores among the four emotions for the seven conditions of $g^s$. It shows that human perception of the level of emotion is distinguishable among the four emotions. *This verifies DC-H$^{InterEmo}$ that the human sensitivity of $g^s$ is different depending on the perceived emotion.*

**Results on S-H**: To validate the formulated hypothesis of S-H, we look into whether the rated levels of emotion are differential among the seven conditions of $g^s$. It is carried out four times for the four emotions, respectively. For example, for the emotion of anger, the scores of the animation clips with anger are collected from the 3 webpages with anger and then they are classified into seven groups, according to the condition of $g^s$. Each group contains 60 scores from 3 utterances with the same emotion, which are rated by 20 participants. The validation of DC-H is to test whether there is significant difference among the seven groups of 60 scores for each emotion. Considering that the participants can freely view the seven animation clips on one webpage and then rate them at one time, the seven scores are dependent of each other. Furthermore, we do not know whether the distribution of the 60 scores in one group is Gaussian. Therefore, we use the Friedman Test, which is a nonparametric test to assess whether there are statistically significant differences between groups of a dependent variable. In this user study, the dependent variable is the condition of $g^s$, which has seven dependent groups: 60 scores under $-15°$, 60 ones under $-10°$, 60 ones under $-5°$, 60 ones under $0°$, 60 ones under $5°$, 60 ones under $10°$, and 60 ones under $15°$.

The Friedman test is carried out four times for the four emotions. Each time, Friedman test is used over the seven conditions of $g^s$ with one emotion. The results are reported in Table 4(a). As can be seen, the Friedman tests show that there are significant differences in the rated scores among the seven conditions of $g^s$ for the four emotions. It reveals that human perception of the level of emotion is differential among the conditions of $g^s$ for each of the four emotions. *This validates DC-H$^{Neu}$/H$^{Sad}$/H$^{Hap}$/H$^{Ang}$ that human subjects are sensitive to the static feature in the perception of neutral/sadness/happiness/anger.*

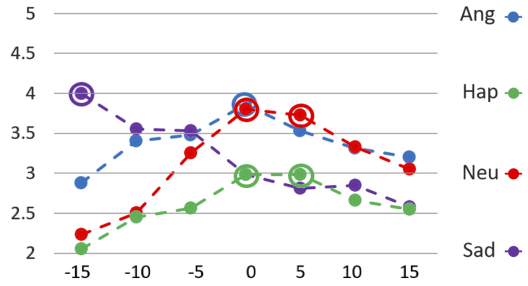



Fig. 9: Perception of emotion rated by participants. For each emotion, the point(s) with the highest value(s) is marked with circles. Particularly, no significant difference is statistically found between the two circled points for neutral and happiness. More details about statistically significant differences can be found in Table 4(b).

TABLE 4: Significant differences in the perception of neutral (Neu), sadness (Sad), happiness (Hap) and anger (Ang) between the conditions of $g^s$. The Friedman test is to test the differences among the seven conditions of $g^s$, which is reported in Table 4(a). Post hoc analysis with Wilcoxon signed-rank tests was conducted to verify the significant differences between the highest value(s) of $g^s$ (see Figure 9) and its other values. - marks no significant difference; * marks p<.05; ** marks p<.01; *** marks p<.001.

(a) Friedman test ($\chi^2$-values and significant differences)

| | Neu | Sad | Hap | Ang |
|---|---|---|---|---|
| $\chi^2(6)$ | 93.43 *** | 83.2 *** | 34.1 *** | 35.39 *** |

(b) Wilcoxon signed rank test (z-values and significant differences)

| **Neu** | $-15°$ | $-10°$ | $-5°$ | $0°$ | $5°$ | $10°$ | $15°$ |
|---|---|---|---|---|---|---|---|
| $0°$ | 5.22 *** | 5.06 *** | 3.55 *** | | 0.54 - | 2.85 ** | 3.87 *** |
| $5°$ | 5.35 *** | 5.22 *** | 3.16 ** | -0.54 - | | 2.16 * | 3.06 ** |

| **Sad** | $-15°$ | $-10°$ | $-5°$ | $0°$ | $5°$ | $10°$ | $15°$ |
|---|---|---|---|---|---|---|---|
| $-15°$ | | 2.96 ** | 2.49 * | 4.36 *** | 4.99 *** | 4.89 *** | 5.82 *** |

| **Hap** | $-15°$ | $-10°$ | $-5°$ | $0°$ | $5°$ | $10°$ | $15°$ |
|---|---|---|---|---|---|---|---|
| $0°$ | 4.11 *** | 2.86 ** | 3.04 ** | | 0.05 - | 2.59 ** | 2.25 * |
| $5°$ | 4.61 *** | 2.95 ** | 2.54 * | -0.05 - | | 2.69 ** | 2.32 * |

| **Ang** | $-15°$ | $-10°$ | $-5°$ | $0°$ | $5°$ | $10°$ | $15°$ |
|---|---|---|---|---|---|---|---|
| $0°$ | 4.57 *** | 2.89 ** | 3.18 ** | | 2.33 * | 3.23 ** | 3.74 *** |

**Results on S-Ex**: Since S-H is validated, we want to identify the optimal value(s) of $g^s$ to enhance the perception of emotion. Due to the validation of S-H$^{InterEmo}$, the optimal value(s) of $g^s$ should be reasonably different among the four emotions. So, we explore the optimal value(s) of $g^s$ for each emotion. The exploration is based on the averaged scores of

each condition of $g^s$.

Figure 9 shows the averaged scores rated by the participants for the four emotions. We aim to identify $g^s$ with the highest score(s) only which corresponds the optimal value(s) of $g^s$. To learn about it, the Wilcoxon signed-rank test is employed to carry out the pairwise comparisons between the conditions of $g^s$. The Wilcoxon signed-rank test allows us to pick out $g^s$ with the highest score(s) by verifying the statistical difference between the averaged scores. The results of statistical difference are shown in Table 4(b).

For **neutral** and **happiness**, the two highest scores are observed: $0°$ 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 $g_s$ valued at $0°$ or $5°$. For **sadness/anger**, $-15°/0°$ is observed as the value of $g_s$ 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 $g_s$ valued at $-15°/0°$. The highest scores above are marked by circles in Figure 9.

**Discussion-S**: The verification of S-H and S-H$^{\textbf{InterEmo}}$ shows that the static motion has non-trivial impact on the perception of emotion. The perceptual sensitivity is no more than $5°$. The optimal value(s) of $g^s$ suggests that the perception of neutral and happiness can be enhanced by adjusting the averaged pitch head rotation straight (e.g. $0°$) or slightly up (e.g. $5°$); the perception of sadness can be enhanced by adjusting the averaged value of head pitch rotation towards down (e.g. $-15°$); and the perception of anger can be enhanced by adjusting the averaged value of head pitch rotation towards straight (e.g. $0°$).

When observing the static features of the original human motion data reported in Figure 4(a), we find that the original static features tend to face up slightly and they are not strictly consistent with the above reported optimal values. For example, while the optimal value is $0°$ or $5°$ for neutral, the original static feature is about $5°$; while the optimal value is $-15°$ for sadness, the original static feature is about $3°$ or $4°$; the original static feature is about $2°$ or $3°$ for anger, which is approaching the optimal value of $0°$; the original static feature is about $2°$ for happiness between the optimal values of $0°$ and $5°$. These comparisons suggest that the expressiveness of emotion could be enhanced by consistent manipulations of adjusting the averaged value of the pitch rotation of the captured head motion data in an utterance (The interested reader can find more details about a further in-depth study of consistent manipulations to enhance the expressiveness of emotion in [26]).

### 6.3 User Study on Perception of Dynamic Motions

US-Dyn investigates the influence of $g^d_{LF}$, $g^d_{MF}$, and $g^d_{HF}$ on human perception of the level of emotion. To do it, we make several variations by removing $g^d_{LF}$ and $g^s$ from $g^{or}$, denoted by $C^{\overline{\text{LF}}\cap\overline{\text{S}}}$; by removing $g^d_{IF}$ and $g^s$ from $g^{or}$, denoted by $C^{\overline{\text{IF}}\cap\overline{\text{S}}}$; and by removing $g^d_{HF}$ and $g^s$ from $g^{or}$, denoted by $C^{\overline{\text{HF}}\cap\overline{\text{S}}}$. Additionally, another variation, noted by $C^{\overline{\text{S}}}$, is made by removing $g^s$ from $g^{or}$. $C^{\overline{\text{LF}}\cap\overline{\text{S}}}$, $C^{\overline{\text{IF}}\cap\overline{\text{S}}}$, $C^{\overline{\text{HF}}\cap\overline{\text{S}}}$, and $C^{\overline{\text{S}}}$ are the four conditions taken in US-Dyn. $C$ is the abbreviation of *Condition*.

The influence of $g^d_{LF}/g^d_{IF}/g^d_{HF}$ is estimated by comparing $C^{\overline{\text{LF}}\cap\overline{\text{S}}}/C^{\overline{\text{IF}}\cap\overline{\text{S}}}/C^{\overline{\text{HF}}\cap\overline{\text{S}}}$ with $C^{\overline{\text{S}}}$. In these comparisons, the impact of $g^s$ is excluded by removing $g^s$, as the role of $g^s$ has been studied in Section 6.2. We look into $C^{\overline{\text{LF}}\cap\overline{\text{S}}}$, $C^{\overline{\text{MF}}\cap\overline{\text{S}}}$, $C^{\overline{\text{HF}}\cap\overline{\text{S}}}$, and $C^{\overline{\text{S}}}$ in the perception of the level of emotion.

**Hypotheses**: US-Dyn is carried out to investigate the hypotheses formulated as follows:

- **D-H$^{\text{Neu}}$/H$^{\text{Sad}}$/H$^{\text{Hap}}$/H$^{\text{Ang}}$** (intra-emotion): the human sensitivity to low-dynamic, intermediate-dynamic, and high-dynamic motions ($g^d_{LF}$, $g^d_{IF}$, and $g^d_{HF}$) is differential in the perception of neutral/sadness/happiness/anger. For the sake of simplicity, **D-H** stands for a general symbol of the four hypotheses.
- **D-H$^{\text{InterEmo}}$** (inter-emotion): the sensitivity to low-dynamic, intermediate-dynamic, and high-dynamic motions ($g^d_{LF}$, $g^d_{IF}$, and $g^d_{HF}$) is different in the perception of different emotions.

D-H brings insight into $g^d_{LF}$, $g^d_{IF}$, and $g^d_{HF}$ from the intra-emotion. It enables us to further learn the influence of $g^d_{LF}$, $g^d_{IF}$, and $g^d_{HF}$ on the perception of one emotion. D-H$^{\text{InterEmo}}$ takes insight into $g^d_{LF}$, $g^d_{IF}$, and $g^d_{HF}$ from the inter-emotions. It allows us to verify whether the human listener is able to differently decode $g^d_{LF}$, $g^d_{IF}$, and $g^d_{HF}$ when perceiving the different types of emotion.

To confirm D-H and D-H$^{\text{InterEmo}}$, we compare the roles of $g^d_{LF}$, $g^d_{IF}$, and $g^d_{HF}$ in the perception of the level of emotion. SC-H would be confirmed if the roles of $g^d_{LF}$, $g^d_{IF}$, and $g^d_{HF}$ are differential in the perception of one emotion; moreover, SC-H$^{\text{InterEmo}}$ would be confirmed if the roles of $g^d_{LF}$, $g^d_{IF}$, and $g^d_{HF}$ are distinguishable between the four emotions.

While investigating the emotional expressiveness of head motion in speech, researchers looked into the naturalness of head motion in previous works [10] [12]. This user study is to investigate the roles of low-dynamic, intermediate-dynamic, and high-dynamic motions in the naturalness of head motion. Furthermore, although previous works often study the emotional expressiveness of head motion and the naturalness of head motion, this study presents unexplored aspect of the relationship between emotional expressiveness and naturalness. Hence, an exploratory user study is done to investigate the hypothesis as below:

- **SC-H-Ex**: the perception of emotion on head motion is consistent with that of the naturalness of head motion.

**Protocol**: Participants are instructed to watch 4 animation clips with the same speech on one webpage and then asked to rank these animation clips two times: according to the level of the expressiveness of emotion and the naturalness of head animation, respectively. We provide the elements of the protocol we follow for this user study.

1) Stimuli: 12 utterances (3 sentences $\times$ 4 emotions) are selected from the dataset. For each utterance, the four conditions of $C^{\overline{\text{LF}}\cap\overline{\text{S}}}$, $C^{\overline{\text{IF}}\cap\overline{\text{S}}}$, $C^{\overline{\text{HF}}\cap\overline{\text{S}}}$, and $C^{\overline{\text{S}}}$

are considered and 16 animation clips (4 emotions $\times$ 4 conditions) are created from its utterances. In total, 48 clips are created for this study.

2) Procedure: US-Dyn is designed to investigate the differences among the four conditions of $C^{\overline{\mathrm{LF}}\cap\overline{\mathrm{S}}}$, $C^{\overline{\mathrm{MF}}\cap\overline{\mathrm{S}}}$, $C^{\overline{\mathrm{HF}}\cap\overline{\mathrm{S}}}$, and $C^{\overline{\mathrm{S}}}$ for each emotion. The virtual character is used to display the 3-dimensional head rotation data under the four conditions. The animation clips are displayed through an online webpage interface. Each webpage encloses four animation clips under the four conditions but with the same speech audio, in a random order. According to OB-Ex, the speech audio is adequate to characterize the conveyed emotion, so we assume that the four animation clips on one webpage express the same emotion. This study is to investigate the perception of the level of emotion instead of emotion recognition. To facilitate the participation, the participants are explicitly informed of the expressed emotion through texts. In total, 12 webpages are created from the three sentences and the four emotions.

The participants are instructed to view 12 webpages. They are asked to rank the four animation clips twice: according to the level of perceived emotion and the naturalness of head motion, respectively. Figure 10 shows a snapshot of the webpage used in this user study.

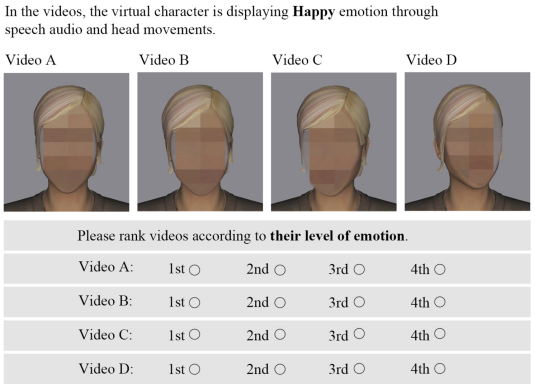



Fig. 10: A snapshot of the webpage used in US-Dyn. One webpage encloses four animation clips under the four conditions of $C^{\overline{\mathrm{LF}}\cap\overline{\mathrm{S}}}$, $C^{\overline{\mathrm{IF}}\cap\overline{\mathrm{S}}}$, $C^{\overline{\mathrm{HF}}\cap\overline{\mathrm{s}}}$, and $C^{\overline{\mathrm{S}}}$ of an utterance. The clips are displayed separately with the individual play button.

**Results on D-H$^{\mathbf{InterEmo}}$ and D-H**: To verify the formulated hypotheses of D-H$^{\mathbf{InterEmo}}$ and D-H, the Friedman test is first employed to determine the difference in one emotion between the ranks of $C^{\overline{\mathrm{LF}}\cap\overline{\mathrm{S}}}$, $C^{\overline{\mathrm{MF}}\cap\overline{\mathrm{S}}}$, $C^{\overline{\mathrm{HF}}\cap\overline{\mathrm{S}}}$, and $C^{\overline{\mathrm{S}}}$. It is used to test the differences between groups when the dependent variable being measured is ordinal [27]. It performs separately four times for the four emotions. The results of the Friedman test are reported in Table 5(a). The results show that there is no significant difference between $C^{\overline{\mathrm{LF}}\cap\overline{\mathrm{S}}}$, $C^{\overline{\mathrm{MF}}\cap\overline{\mathrm{S}}}$, $C^{\overline{\mathrm{HF}}\cap\overline{\mathrm{S}}}$, and $C^{\overline{\mathrm{S}}}$ for neutral while there are significant differences for sadness, happiness, and anger.

TABLE 5: Significant differences in the perception of neutral (Neu), sadness (Sad), happiness (Hap) and anger (Ang) between the conditions of $C^{\overline{\mathrm{LF}}\cap\overline{\mathrm{S}}}$, $C^{\overline{\mathrm{MF}}\cap\overline{\mathrm{S}}}$, $C^{\overline{\mathrm{HF}}\cap\overline{\mathrm{S}}}$, and $C^{\overline{\mathrm{S}}}$. Table 5(a) reports the results of the Friedman test among the five conditions for each emotion. Post hoc analysis with Wilcoxon signed-rank tests was conducted to verify significant differences of the pairwise comparisons. Table 5(b) reports the results of the pairwise comparisons between $C^{\overline{\mathrm{HF}}\cap\overline{\mathrm{S}}}$ and the other four conditions; Table 5(c) reports the results of the pairwise comparisons of $C^{\overline{\mathrm{LF}}\cap\overline{\mathrm{S}}}$, $C^{\overline{\mathrm{MF}}\cap\overline{\mathrm{S}}}$, and $C^{\overline{\mathrm{HF}}\cap\overline{\mathrm{S}}}$, which have significant differences from $C^{\overline{\mathrm{s}}}$ (see Table 5(b)). - marks no significant difference; * marks p<.05; ** marks p<.01; *** marks p<.001.

(a) Friedman test ($\chi^2$-values and significant differences)

| | Neu | Sad | Hap | Ang |
|---|---|---|---|---|
| $\chi^2(3)$ | 5.84 - | 71.34 *** | 98.78 *** | 79.04 *** |

(b) Wilcoxon signed rank test (z-values and significant differences) between $g^{\overline{\mathrm{HF}}\cap\overline{\mathrm{DC}}}$ and the other three conditions

| **Neu** | $C^{\overline{\mathrm{LF}}\cap\overline{\mathrm{S}}}$ | $C^{\overline{\mathrm{MF}}\cap\overline{\mathrm{S}}}$ | $C^{\overline{\mathrm{HF}}\cap\overline{\mathrm{S}}}$ |
|---|---|---|---|
| $C^{\overline{\mathrm{S}}}$ | 1.80 - | 0 - | 1.25 - |
| **Sad** | $C^{\overline{\mathrm{LF}}\cap\overline{\mathrm{S}}}$ | $C^{\overline{\mathrm{MF}}\cap\overline{\mathrm{S}}}$ | $C^{\overline{\mathrm{HF}}\cap\overline{\mathrm{S}}}$ |
| $C^{\overline{\mathrm{S}}}$ | 5.70 *** | 2.59 - | 2.02 - |
| **Hap** | $C^{\overline{\mathrm{LF}}\cap\overline{\mathrm{S}}}$ | $C^{\overline{\mathrm{MF}}\cap\overline{\mathrm{S}}}$ | $C^{\overline{\mathrm{HF}}\cap\overline{\mathrm{S}}}$ |
| $C^{\overline{\mathrm{S}}}$ | 5.59 *** | 6.12 *** | 2.01 - |
| **Ang** | $C^{\overline{\mathrm{LF}}\cap\overline{\mathrm{S}}}$ | $C^{\overline{\mathrm{MF}}\cap\overline{\mathrm{S}}}$ | $C^{\overline{\mathrm{HF}}\cap\overline{\mathrm{S}}}$ |
| $C^{\overline{\mathrm{S}}}$ | 6.27 *** | 5.29 *** | 3.64 * |

(c) Wilcoxon signed ranks test (z-values and significant differences) between $C^{\overline{\mathrm{LF}}\cap\overline{\mathrm{S}}}$, $C^{\overline{\mathrm{IF}}\cap\overline{\mathrm{S}}}$, $C^{\overline{\mathrm{HF}}\cap\overline{\mathrm{S}}}$ which significantly differ from $C^{\overline{\mathrm{S}}}$ (see Table 5(b)).

| | | |
|---|---|---|
| **Hap** | $C^{\overline{\mathrm{LF}}\cap\overline{\mathrm{S}}}$(3.65) < $C^{\overline{\mathrm{IF}}\cap\overline{\mathrm{S}}}$(2.90), z=4.25, p<3e-05 | *** |
| **Ang** | $C^{\overline{\mathrm{LF}}\cap\overline{\mathrm{S}}}$(3.57) < $C^{\overline{\mathrm{IF}}\cap\overline{\mathrm{S}}}$(2.63), z=3.95, p<8e-05 | *** |
| | $C^{\overline{\mathrm{LF}}\cap\overline{\mathrm{S}}}$(3.57) < $C^{\overline{\mathrm{HF}}\cap\overline{\mathrm{S}}}$(2.30), z=5.44, p<6e-08 | *** |
| | $C^{\overline{\mathrm{IF}}\cap\overline{\mathrm{S}}}$(2.63) < $C^{\overline{\mathrm{HF}}\cap\overline{\mathrm{S}}}$(2.30), z=1.86, p=0.06 | - |

To examine where the differences actually occur for sadness, happiness, and anger, a post hoc analysis with Wilcoxon signed-rank tests was conducted with a Bonferroni correction applied to the pairwise comparison. The partial results of Wilcoxon signed-rank tests are reported in Tables 5(b) and 5(c). Table 5(b) reports the results of pairwise comparing $C^{\overline{\mathrm{S}}}$ with $C^{\overline{\mathrm{LF}}\cap\overline{\mathrm{S}}}$, $C^{\overline{\mathrm{IF}}\cap\overline{\mathrm{S}}}$, or $C^{\overline{\mathrm{HF}}\cap\overline{\mathrm{S}}}$. As can be seen, for **neutral**, no significant differences are observed between three comparisons. This suggests that dynamic motions have no influence on the perception of neutral. For **sadness**, while no significant difference is observed in the comparison between $C^{\overline{\mathrm{S}}}$ and $C^{\overline{\mathrm{IF}}\cap\overline{\mathrm{S}}}$ and another comparison between $C^{\overline{\mathrm{S}}}$ and $C^{\overline{\mathrm{HF}}\cap\overline{\mathrm{S}}}$, there is a significant difference in the comparison between $C^{\overline{\mathrm{S}}}$ and $g^{\overline{\mathrm{LF}}\cap\overline{\mathrm{S}}}$. This suggests that $C^{LF}$ has significant impact on the perception of sadness and $C^{IF}$ and $C^{HF}$ have no effect on it. For **happiness**, there are significant differences between the comparison between $C^{\overline{\mathrm{LF}}\cap\overline{\mathrm{S}}}$ and $C^{\overline{\mathrm{S}}}$ and also another one between $C^{\overline{\mathrm{IF}}\cap\overline{\mathrm{S}}}$ and

TABLE 6: Wilcoxon signed ranks test on perception of naturalness in neutral (Neu), sadness (sad), happiness (hap) and anger (ang). This table reports the z-value of Wilcoxon signed ranks test. - marks no significant difference. * marks p<.05, ** marks p<.01,*** marks p<.001.

| **Neu** | $C_{hs}^{\overline{\mathrm{LF}}\cap\overline{\mathrm{DC}}}$ | $C_{hs}^{\overline{\mathrm{MF}}\cap\overline{\mathrm{DC}}}$ | $C_{hs}^{\overline{\mathrm{HF}}\cap\overline{\mathrm{DC}}}$ |
|---|---|---|---|
| $C_{hs}^{\overline{\mathrm{DC}}}$ | 5.99 *** | -0.34 - | 1.22 - |
| **Sad** | $C_{hs}^{\overline{\mathrm{LF}}\cap\overline{\mathrm{DC}}}$ | $C_{hs}^{\overline{\mathrm{MF}}\cap\overline{\mathrm{DC}}}$ | $C_{hs}^{\overline{\mathrm{HF}}\cap\overline{\mathrm{DC}}}$ |
| $C_{hs}^{\overline{\mathrm{DC}}}$ | 6.94 *** | 4.87 *** | 4.47 *** |
| **Hap** | $C_{hs}^{\overline{\mathrm{LF}}\cap\overline{\mathrm{DC}}}$ | $C_{hs}^{\overline{\mathrm{MF}}\cap\overline{\mathrm{DC}}}$ | $C_{hs}^{\overline{\mathrm{HF}}\cap\overline{\mathrm{DC}}}$ |
| $C_{hs}^{\overline{\mathrm{DC}}}$ | 5.91 *** | 1.52 - | 1.59 - |
| **Ang** | $C_{hs}^{\overline{\mathrm{LF}}\cap\overline{\mathrm{DC}}}$ | $C_{hs}^{\overline{\mathrm{MF}}\cap\overline{\mathrm{DC}}}$ | $C_{hs}^{\overline{\mathrm{HF}}\cap\overline{\mathrm{DC}}}$ |
| $C_{hs}^{\overline{\mathrm{DC}}}$ | 4.82 *** | 0.004 - | -1.22 - |

$C^{\overline{\mathrm{S}}}$. This suggests that $C^{LF}$ and $C^{IF}$ affect significantly the perception of happiness. Table 5(c) further reports the pairwise comparisons between $C^{\overline{\mathrm{LF}}\cap\overline{\mathrm{S}}}$ and $C^{\overline{\mathrm{IF}}\cap\overline{\mathrm{S}}}$. The result shows that $C^{\overline{\mathrm{LF}}\cap\overline{\mathrm{S}}}$ is ranked significantly more advanced than $C^{\overline{\mathrm{IF}}\cap\overline{\mathrm{S}}}$. This shows that $C^{LF}$ influences the perception more than $C^{IF}$. For **anger**, there are significant differences in all the three comparisons with $C^{\overline{\mathrm{LF}}\cap\overline{\mathrm{S}}}$, $C^{\overline{\mathrm{IF}}\cap\overline{\mathrm{S}}}$, and $C^{\overline{\mathrm{HF}}\cap\overline{\mathrm{S}}}$. This reveals that $C^{LF}$, $C^{IF}$, and $C^{HF}$ have certain affect on the perception of anger. Moreover, Table 5(c) reports the results of the pairwise comparisons between $C^{\overline{\mathrm{LF}}\cap\overline{\mathrm{S}}}$, $C^{\overline{\mathrm{IF}}\cap\overline{\mathrm{S}}}$, and $C^{\overline{\mathrm{HF}}\cap\overline{\mathrm{S}}}$. The results show that $C^{\overline{\mathrm{LF}}\cap\overline{\mathrm{S}}}$ is ranked significantly more advanced than $C^{\overline{\mathrm{IF}}\cap\overline{\mathrm{S}}}$ and $C^{\overline{\mathrm{HF}}\cap\overline{\mathrm{S}}}$ and that $C^{\overline{\mathrm{IF}}\cap\overline{\mathrm{S}}}$ is ranked more advanced than $C^{\overline{\mathrm{HF}}\cap\overline{\mathrm{S}}}$. These results reveal that $C^{LF}$ impact the perception of emotion more than $C^{IF}$ and $C^{HF}$ and that $C^{IF}$ have more influence than $C^{HF}$.

The results above show that $g_{LF}^d$, $g_{IF}^d$, and $g_{HF}^d$ differently impact the perception of sadness, happiness, and anger while they have no influence on the perception of neutral. Therefore, *$H^{Sad}$, $H^{Hap}$, and $H^{Ang}$ are validated while SC-$H^{Neu}$ is not confirmed*.

Observing Table 5(b) and taking insight into the differences of $g_{LF}^d$, $g_{IF}^d$, and $g_{HF}^d$ between the four emotions, we find that $g^{LF}$ has no influence on the perception of neutral while it make an impact on the perception of sadness, happiness, and anger. Moreover, $g_{IF}^d$ has no influence on the perception of neutral and sadness while it impacts the perception of happiness and anger. And, $g_{HF}^d$ has no influence on the perception of neutral, sadness, and happiness while it impacts the perception of anger.

These observations above reveal that the impact of $g_{LF}^d$, $g_{IF}^d$, or $g_{HF}^d$ is differential among the four emotions. *This verifies SC-$H^{InterEmo}$*.

**Results on D-H-Ex**: To verify the formulated hypothesis of D-H-Ex, we employ the Wilcon signed ranks test to verify significant differences in the perception of naturalness of head motion between $C^{\overline{\mathrm{LF}}\cap\overline{\mathrm{S}}}$, $C^{\overline{\mathrm{IF}}\cap\overline{\mathrm{S}}}$ or $C^{\overline{\mathrm{HF}}\cap\overline{\mathrm{S}}}$ and $C^{\overline{\mathrm{S}}}$. The results are reported in Table 6.

The significant difference reflects significant impact on perception of naturalness. For **neutral**, there is a significant difference between $C^{\overline{\mathrm{S}}}$ and $C^{\overline{\mathrm{LF}}\cap\overline{\mathrm{S}}}$, while there is no significant difference between $C^{\overline{\mathrm{S}}}$ and $C^{\overline{\mathrm{IF}}\cap\overline{\mathrm{S}}}$, and between $C^{\overline{\mathrm{S}}}$ and $C^{\overline{\mathrm{HF}}\cap\overline{\mathrm{S}}}$. This means that $g_{LF}^d$ has significant influence on perception of naturalness while $g_{IF}^d$ and $g_{HF}^d$ have no significant influence. For **sadness**, there are significant differences between $C^{\overline{\mathrm{S}}}$ and $C^{\overline{\mathrm{LF}}\cap\overline{\mathrm{S}}}$, between $C^{\overline{\mathrm{S}}}$ and $C^{\overline{\mathrm{IF}}\cap\overline{\mathrm{S}}}$, and between $C^{\overline{\mathrm{S}}}$ and $C^{\overline{\mathrm{HF}}\cap\overline{\mathrm{S}}}$. This means that $g_{LF}^d$, $g_{IF}^d$ and $g_{HF}^d$ have significant influence. For **happiness** and **anger**, there is a significant difference between $C^{\overline{\mathrm{S}}}$ and $C^{\overline{\mathrm{LF}}\cap\overline{\mathrm{S}}}$ while there is no significant difference between $C^{\overline{\mathrm{S}}}$ and $C^{\overline{\mathrm{IF}}\cap\overline{\mathrm{S}}}$, and between $C^{\overline{\mathrm{S}}}$ and $C^{\overline{\mathrm{HF}}\cap\overline{\mathrm{S}}}$. This shows that only $g^{LF}$ has influence on the naturalness of head motion.

These above observations indicate that $g_{LF}^d$ has impact on the perception of naturalness for all the four emotions while $g_{IF}^d$ and $g_{HF}^d$ have effect on naturalness for sadness and no influence for neutral, happiness, and anger. This means that $g_{LF}^d$ is crucial to the perception of naturalness for the four emotions and that the naturalness for sadness relies on not only $g_{LF}^d$ but also $g_{IF}^d$ and $g_{HF}^d$.

Tables 6 and 5(b) illustrate the roles of $g_{LF}^d$, $g_{IF}^d$ and $g_{HF}^d$ in the perception of the naturalness of head motion and in the perception of emotion, respectively. In Tables 6 and 5(b), the significant impact (difference) on the perception of emotion or naturalness is highlighted with gray.

As can be seen in Tables 5(b) and 6, in neutral, $g_{LF}^d$ has impact on the perception of naturalness rather than emotion; in sadness, $g_{IF}^d$ and $g_{HF}^d$ has impact on the perception of naturalness rather than emotion; in happiness, $g_{IF}^d$ and $g_{HF}^d$ has impact on the perception of emotion rather than naturalness; in anger, $g_{IF}^d$ and $g_{HF}^d$ has impact on the perception of emotion rather than naturalness. *These observations show that SC-H-Ex is not verified*.

**Discussion-Dyn**: The verification of $H^{Sad}$, $H^{Hap}$, and $H^{Ang}$ reveals that humans are sensitive to $g_{LF}^d$, $g_{IF}^d$, and $g_{HF}^d$ and it suggests that humans are very skilled in decoding $g_{LF}^d$, $g_{IF}^d$, and $g_{HF}^d$ depending on the perceived emotion. Moreover, it implies that $g_{LF}^d$, $g_{IF}^d$, and $g_{HF}^d$ provide useful cues to carry emotion information. The non-verification of D-$H^{Neu}$ shows that $g_{LF}^d$, $g_{IF}^d$, and $g_{HF}^d$ have no influence on human perception of emotion. Comparing the verified hypotheses with emotional states (sad, happy, and angry) and the non-verified hypothesis with non-emotional state (neutral), it reveals that $g_{LF}^d$, $g_{IF}^d$, and $g_{HF}^d$ play important roles in encoding the emotional information into head motion for the emotional states. The verification of SC-$H^{InterEmo}$ shows that humans decode the emotional information in a non-trivial manner, which depends on the perceived emotion. Indeed, when observing Figure 4(d) and Table 5(b), we find that, while $g_{HF}^d$ has no impact on the perception of happiness but affects the perception of anger, it has a higher amplitude in happiness than anger. It shows that a higher amplitude may have lower effect on the perception, which supports the verification of SC-$H^{InterEmo}$.

The non-verification of D-H-Ex reveals that humans perceive emotion through head motion and the naturalness of head motion in two different channels. When observing the non-naturalness, humans are capable of perceiving the emotional information. For example, while perceiving the degraded naturalness due to the lack of $g_{IF}^d$, and $g_{HF}^d$ in sadness, humans can perceive sadness accurately.

From the perception of emotion, we observe that $g_{LF}^d$ affects sadness, happiness, and anger; $g_{IF}^d$ has effect on

TABLE 7: Summary of the effects of the static and dynamic features on emotion expression (Table 7(a)) and naturalness perception (Table 7(b)). "Average" refers to the static feature on head pitch rotation. Smooth-and-slow (Sm-Sl), moderate-fluidity-and-speed (Mo-fl-sp), and jerky-and-fast (Je-fa) refer to the dynamic features on head pitch rotation.

(a) The static feature rotation angles that most enhance the emotions, and certain dynamic features, with Y (Yes), that affect emotion perception.

| | Neutral | Sadness | Happiness | Anger |
|---|---|---|---|---|
| Average | $0°/5°$ | $-15°$ | $0°/5°$ | $0°$ |
| Sm-Sl | | Y | Y | Y |
| Mo-fl-sp | | | Y | Y |
| Je-fa | | | | Y |

(b) Certain dynamic features, with Y (Yes), affect naturalness perception.

| | Neutral | Sadness | Happiness | Anger |
|---|---|---|---|---|
| Sm-Sl | Y | Y | Y | Y |
| Mo-fl-sp | | Y | | |
| Je-fa | | Y | | |

happiness and anger; and $g^d_{HF}$ has impact on anger only. For the perception of naturalness, we observe $g^d_{LF}$ influences all the four emotions; $g^d_{IF}$ and $g^d_{HF}$ have effect on sadness only. This suggests that dynamic motions with lower frequencies have more effect on the perception of emotion and the perception of naturalness of head motion, which is consistent with the results in Table 5(c). This can be explained by the amplitudes of dynamic motions reported in Figures 4(b)-4(d), which demonstrates that the dynamic motion with lower frequencies has higher amplitudes. On the other hand, although $g^d_{HF}$ has a low amplitude, it impacts the perception of anger and the perception of head naturalness for sadness.

## 7 Conclusions

In this paper, we propose the static and dynamic features, extracted with Fourier transform, to characterize head motion in expressive speech, and aim to analyze and understand the human perception of emotion in head motion and the human perception of the naturalness of head motion. The static feature reflects the rotation bias and the dynamic features describe the motions from smooth-and-slow, Table 7 summarizes the effects of the static and dynamic features on emotion expression and naturalness perception.

Our work is conducted through intra-related objective and subjective experiments. It reveals that human speakers encode the static and dynamic features for head motion differently, depending on the emotions they are expressing. Our study reveals that humans are skillful both at encoding and decoding the static and dynamic features and that these features provide sound explanations to the link between head motion and the perception of emotion in an utterance.

In the objective experiments, static and dynamic features are employed to recognize the emotion in an utterance. The promising recognition accuracies show that the static and dynamic features are effective in characterizing and recognizing emotions. We also learn that the uni-modality of head motion data provides adequate clue to reflect the emotional information. Based on the captured human motion data, statistical tests show that the static and dynamic features are coordinated in a non-trivial manner depending on the conveyed emotion. Our objective experiments suggest that the static and dynamic features would be useful features to characterize head motion to distinguish the type of emotion in an utterance. It demonstrates that humans are very skilled at coordinating the static and dynamic features to express specific emotions.

In accordance with the objective analysis, we conduct three subjective experiments (US-M, US-Sta, and US-Dyn). US-M shows that humans are unable to reliably perceive emotion from head motion alone, which is not in line with the finding of our objective experiments. It suggests that humans encode more information into head motion when conveying neutral and anger than expressing sadness and happiness. US-Sta reveals that the static feature is closely related to the perception of emotion. Particularly, the perceptual sensitivity to up-down still rotation is no more than $5°$. US-D shows that dynamic motions with lower frequencies have more impact on the perception of emotion and the naturalness of head motion but a higher amplitude may not lead to more effect on the perception. US-Dyn shows that humans are able to decode dynamic motions in different manners depending on the perceived emotion. It also reveals that humans perceive emotion carried in head motion and the naturalness of head motion in two different channels. On one hand, humans are able to perceive emotion carried in unnatural head movement; on the other hand, even if the emotion is degraded due to the certain manipulation of head motion, humans may still perceive that head movement is natural. US-Sta and US-Dyn together reveal that humans are sensitive to the static and dynamic motions in a non-trivial manner depending on the perceived emotion when perceiving emotion and the naturalness of head motion. US-Dyn and US-M indicate that human perception is unrelated to the amplitude of the dynamic features. A higher amplitude may affect the perception less. These results provide evidence that humans are adept at decoding the static and dynamic features to perceive emotion.

The reported objective and subjective results establish that the static and dynamic features are informative to characterize head motion in expressive speech. The static and dynamic features provide reasonable explanations to head motion linked to the expressiveness and perception of emotion in an utterance, based on a low-level description of motion instead of action or motion style. To the best of our knowledge, this work is the first to explore the low-level characterization of head motion in expressive speech.

Some limitations exist in our current work: 1) this work relies on the head motion data of a professional actress uttering a planned script. We will further validate whether our findings can be generalized to other persons and also whether the findings can be generalized to head motion in daily conversations (and even multi-party conversations); 2) Our investigations are based on short utterances in the dataset. It is worthy to further study utterances lasting for longer time; 3) our subjective experiments rely on human perception on the performance of a virtual character. It is

still unclear whether our findings can be validated on the performances of real people, due to the potential perceptual gap on the performances of real people and those of virtual characters.

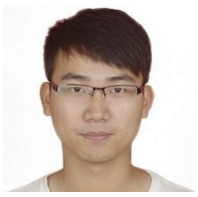
## 8 Acknowledgments

This research is supported in part by NSF IIS-1524782.

## References

[1] P. Ekman and W. V. Friesen, "Facial action coding system," 1977.
[2] M. Lhommet and S. C. Marsella, "Expressing emotion through posture," *The Oxford Handbook of Affective Computing*, pp. 273–285, 2014.
[3] A. C. C. Roether, L. Omlor and M. Giese, "Critical features for the perception of emotion from gait," *J. Vision*, vol. 8, no. 6-15, pp. 1–32, 2009.
[4] G. Castellano, M. Mortillaro, A. Camurri, G. Volpe, and K. Scherer, "Automated analysis of body movement in emotionally expressive piano performances," *Music Perception: An Interdisciplinary Journal*, vol. 26, no. 2, pp. 103–119, 2008.
[5] H. M. Paterson, F. E. Pollick, and A. J. Sanford, "The role of velocity in affect discrimination," *Proceedings of the Cognitive Science Society*, pp. 756–761, 2001.
[6] M. M. Gross, E. A. Crane, and B. L. Fredrickson, "Methodology for assessing bodily expression of emotion," *Journal of Nonverbal Behavior*, vol. 34, no. 4, pp. 223–248, 2010.
[7] N. Fourati and C. Pelachaud, "Perception of emotions and body movement in the emilya database," *IEEE Transactions on Affective Computing*, vol. xx, no. xx, p. (preprint online), (accepted in 2016).
[8] F. Durupinar, M. Kapadia, S. Deutsch, M. Neff, and N. I. Badler, "Perform: Perceptual approach for adding ocean personality to human motion using laban movement analysis," *ACM Trans. Graph.*, vol. 36, no. 1, pp. 6:1–6:16, Oct. 2016.
[9] A. Kleinsmith and N. Bianchi-Berthouze, "Affective body expression perception and recognition: A survey," *IEEE Transactions on Affective Computing*, vol. 4, no. 1, pp. 15–33, 2013.
[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] A. Adams, M. Mahmoud, T. Baltrušaitis, and P. Robinson, "Decoupling facial expressions and head motions in complex emotions," in *ACII*, 2015, pp. 274–280.
[12] Y. Ding, C. Pelachaud, and T. Artieres, "Modeling multimodal behaviors from speech prosody," in *International Conference on Intelligent Virtual Agents*, 2013, pp. 217–228.
[13] G. Grossman, R. J. Leigh, L. Abel, D. Lanska, and S. Thurston, "Frequency and velocity of rotational head perturbations during locomotion," *Experimental brain research*, vol. 70, no. 3, pp. 470–476, 1988.
[14] R. Azuma and G. Bishop, "A frequency-domain analysis of head-motion prediction," in *SIGGRAPH*, 1995, pp. 401–408.
[15] X. Ma, B. H. Le, and Z. Deng, "Perceptual analysis of talking avatar head movements: A quantitative perspective," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2011, pp. 2699–2702.
[16] J. Tilmanne and T. Dutoit, "Stylistic walk synthesis based on fourier decomposition," in *Intell. Tech. for Inter. Enter.*, 2013, pp. 71–79.
[17] R. Niewiadomski, M. Mancini, Y. Ding, C. Pelachaud, and G. Volpe, "Rhythmic body movements of laughter," in *International Conference on Multimodal Interaction*, 2014, pp. 299–306.
[18] A. Mignault and A. Chaudhuri, "The many faces of a neutral face: Head tilt and perception of dominance and emotion," *Journal of nonverbal behavior*, vol. 27, no. 2, pp. 111–132, 2003.
[19] J. L. Tracy and R. W. Robins, "The prototypical pride expression: development of a nonverbal behavior coding system." *Emotion*, vol. 7, no. 4, p. 789, 2007.
[20] U. Hess, R. B. Adams, and R. E. Kleck, "Looking at you or looking elsewhere: The influence of head orientation on the signal value of emotional facial expressions," *Motivation and Emotion*, vol. 31, no. 2, pp. 137–144, 2007.
[21] M. Thiebaux, S. Marsella, A. N. Marshall, and M. Kallmann, "Smartbody: Behavior realization for embodied conversational agents," in *AAMAS*, 2008, pp. 151–158.
[22] P. Read and M.-P. Meyer, *Restoration of motion picture film*. Butterworth-Heinemann, 2000.
[23] "http://onlinestatbook.com/2/estimation/proportion_ci.html."
[24] W. Kruskal and W. Wallis, "Use of ranks in one-criterion variance analysis," *Journal of the American Statistical Association*, pp. 583–621, 1952.
[25] D. N. A. M. Martinez, "Emotion perception in emotionless face images suggests a norm-based representation," *Journal of Vision*, 2009.
[26] Y. Ding, L. Shi, and Z. Deng, "Perceptual enhancement of emotional mocap head motion: an experimental study," in *International Conference on Affective Computing and Intelligent Interaction*, 2017, pp. 242–247.
[27] "https://statistics.laerd.com/spss-tutorials/friedman-test-using-spss-statistics.php."

**Yu Ding** is a postdoctoral researcher at the University of Houston (UH). He earned Computer Science Ph.D. (2014) at Telecom Paristech in Paris (France). He received his M.S. degree in Computer Science from Pierre et Marie Curie university (France) and B.S. degree in automation from Xiamen University (China). His research interests include nonverbal communication, and human computer interaction.



**Lei Shi** is a PhD student in the Computer Science department at the University of Houston(UH). He completed the B.S. and M.S. degrees in Computer Science from North China Electric Power University (China), in 2013 and 2016 respectively.

**Zhigang Deng** is currently a Full Professor of Computer Science at the University of Houston (UH). His research interests include computer graphics, computer animation, human computer interaction, and affective computing. He earned his Ph.D. in Computer Science at the Department of Computer Science at the University of Southern California in 2006. Prior that, he also completed B.S. degree in Mathematics from Xiamen University (China), and M.S. in Computer Science from Peking University (China).