Acoust. Sci. & Tech. **46**, 1 (2025)

INVITED PAPER

Contributions of audio and visual modalities to perception of Mandarin Chinese emotions in valence-arousal space

Yongwei Li[1], Aijun Li[2,*], Jianhua Tao[3,4,†], Feng Li[5], Donna Erickson[6] and Masato Akagi[7]

[1]*CAS Key Laboratory of Behavioral Science, Institute of Psychology, Chinese Academy of Sciences, Beijing, 100045, China*
[2]*Corpus and Computational Linguistics Center, Institute of Linguistics, Chinese Academy of Social Sciences, Beijing, 100732, China*
[3]*Department of Automation, Tsinghua University, Beijing, 100190, China*
[4]*Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing, 100190, China*
[5]*School of Management Science and Engineering, Anhui University of Finance and Economics, Anhui, 233030, China*
[6]*Haskins Laboratories, New Haven 06511-6695, U.S.A.*
[7]*Graduate School of Advanced Science and Technology, Japan Advanced Institute of Science and Technology, Ishikawa, 923–1292 Japan*

(*Received 9 April 2024, Accepted for publication 17 August 2024*,
J-STAGE Advance published date: 24 August 2024)

Abstract: Emotions are usually perceived by multimodal cues for human communications; in recent years, emotions have been studied from the perspective of dimensional approaches. Investigation of audio and video cues to emotion perception in terms of categories of emotion has been relatively extensively conducted, but the contribution of audio and video cues to emotion perception in dimensional space is relatively under-investigated, especially in Mandarin Chinese. In this present study, three psychoacoustic experiments were conducted to investigate the contributions of audio, visual, and audio-visual modalities to emotional perception in the valence and arousal space. Audio-only, video-only, and audio-video modalities were presented to native Chinese subjects with normal hearing and vision for perceptual ratings of emotion in the valence and arousal dimensions. Results suggested that (1) different modalities contribute differently to perceiving valence and arousal dimensions; (2) compared to video-only modality, audio-only modality generally decreases arousal and valence at lower levels, and increases arousal and valence at higher levels; (3) the video-only modality plays an important role in separating anger and happiness emotions in the valence space.

Keywords: Dimensional emotion, Valence-arousal, Emotion perception, Audio-visual

1. INTRODUCTION

Emotions play an important role in human-to-human communication; human emotions are expressed through various modalities (speech, facial expression, gesture, etc.). In daily face-to-face conversations, emotions are frequently perceived from the speaker's voice and facial expressions. Thus, investigating emotional perception from both the audio and visual perspective is important for understanding human emotional communication.

The perception of emotions can be examined from the perspective of a categorical or dimensional approach. In the categorical approach, emotions are represented by discrete states, such as six basic emotions (e.g., happiness, anger, sadness) [1]. The discrete categorical approach may not reflect the complexity of emotions in our daily life, because different degrees/intensities of an emotional state may change over time depending on the communication situation [2]. In the dimensional approach, emotions are represented by a value on continuous dimensional space,

*e-mail: liaj@cass.org.cn
†e-mail: jhtao@tsinghua.edu.cn
[doi:10.1250/ast.e24.41]

such as, for instance, valence and arousal emotional space [3]. Valence refers to how positive or negative the listener perceives an event, and arousal refers to how calm or excited the listener perceives an event. Dimensional descriptions of human emotion are reported by a number of studies [4–6].

Neuroimaging studies have suggested that human emotion perception from one sensory channel is affected by information processing in another [7]. McGurk and MacDonald demonstrated that clips of faces in movement affect speech perception, in their well-known Mcgurk effect study [8]. Moreover, several studies have suggested that there is also an emotional McGurk effect between audio and visual information [9–12]. For example, De Gelder suggested the existence of mandatory bidirectional links between emotion detection structures in vision and audition, with cross-modal stimuli having a decisive influence on emotion perception of happy and sad [13]. Takagi *et al.* investigated the audio-visual interactions among six emotion categories in the Japanese language [14]. Similarly, Arimoto and Okanoya have investigated audio-visual interactions with emotion perception along valence-arousal-dominance dimensions for the Japanese language [15]. Using six-emotion description, Dang *et al.* reported that about a 50% variance between Japanese and Chinese cultures [16]. A preliminary study by Li [17] looking at emotional categories in Mandarin Chinese and Japanese language, reported facial expression may play a more important role in sadness and anger emotions.

To summarize, the investigation into the influence of auditory and visual cues on the perception of emotional categories and dimensions for Japanese emotions has been comprehensively explored. However, for tonal languages, particularly Mandarin Chinese, Li [17] has examined the effects of audio-only, video-only, and combined audio-video stimuli on the perception of discrete emotions (e.g., happiness, sadness) but did not extend this investigation to dimensional emotion perception (e.g., valence, arousal), which is the theoretical basis for the multimodal dimension emotion recognition task.

Building upon [17], our study further explores the effects of single audio-only, single video-only, and combined audio-video stimuli on human dimensional emotion perception, especially in valence and arousal dimensional spaces. In this way, we address the existing research gap and contribute to a more comprehensive understanding of multimodal human dimensional emotion perception in tonal languages.

In this paper, three psychoacoustic experiments were conducted in this present study. In experiment I, the audio-only modality stimuli were presented to subjects for perceptually rating emotions in terms of valence and arousal. In experiment II, the video-only modality stimuli were presented to subjects for perceptually rating emotions in terms of valence and arousal. In experiment III, audio-video modality stimuli were presented to subjects for perceptually rating emotions in terms of valence and arousal. The experimental results suggested that (1) different modalities contribute differently to perceiving valence and arousal dimensions; (2) Compared to video-only modality, audio-only modality generally decreases arousal and valence at lower levels, and increases arousal and valence at higher levels; (3) the video-only modality plays an important role in separating anger and happiness emotions in the valence space.

2. EXPERIMENT I: EMOTION PERCEPTION OF AUDIO-ONLY IN VALENCE-AROUSAL SPACE

The task of experiment I is to investigate human emotion perception in V-A space from the audio-only modality.

2.1. Method

2.1.1. Stimuli

Ten sentences, each with seven different emotions (i.e., happiness, sadness, anger, disgust, fear, surprise, and neutral) were uttered by a professional female actress from Bejing Film Academy, who speaks standard Chinese. Canon Power Shot TX1 recorded the actor's performance in the sound-proof room at the Institute of Linguistics, Chinese Academy of Social Sciences. As a result, 70 utterances (10 sentences $\times$ 7 emotions $=$ 70) were recorded as stimuli. Note that the actor's intended emotions were correctly perceived by listeners in a previous study [17].

2.1.2. Subjects

Twenty normal-hearing listeners (ten females and ten males) participated in this experiment. All subjects were native Chinese-speaking listeners. The subjects were undergraduate students at the Anhui University of Finance and Economics, ranging in age from 18 to 22 years old.

2.1.3. Procedure

Seventy emotional speech utterances (audio-only) were presented to each subject through headphones in a quiet room, and a Matlab graphical user interface (GUI) was used for the experiment. Each subject listened to a total of 70 emotional utterances, the presentation orders of the stimuli were randomized across each subject. Before the test, the basic information of emotion dimensions and the meanings of valence and arousal were introduced to the subjects. In the test, each subject was asked to give a score for each stimulus based on her/his perceptual impression of valence and arousal. The scores for the perceptual evaluation of emotions in the valence-arousal (V-A) space ranged from 1 to 5 with a step of 0.2. For the evaluation of valence, the score 1 indicates very negative, 5 as very

positive; for the evaluation of arousal, the score of 1 indicates very calm, 5 as very excited.

2.2. Results

The mean perceptual scores of the 70 emotional utterances in the valence-arousal space across 20 subjects are plotted in Fig. 1. As shown in Fig. 1, the scores of neutral, fear, and disgust emotional speech were near the center $(3, 3)$ in the V-A space. It was found that the perceptual scores for the anger emotional speech were the highest in both valence and arousal, whereas the sadness emotional speech was perceptually scored the lowest in both valence and arousal. Surprise and happiness emotional speech were perceptually scored almost the same for valence, while the scores of surprise emotional speech were a little higher than those of happiness in the arousal space.

In order to check whether the arousal and valence perceptual scores of the audio-only modality were significantly different, we did a two-way repeated-measures analysis of variance (ANOVA) for the 20 subjects; the perceptual scores were the dependent variable and the seven emotion categories and ten sentences, the two within-subject factors. The results indicated that for the arousal ratings there was a significant effect of the emotion category $[F(6, 114) = 142.214,\ p < 0.001]$. For the valence ratings, there was a significant effect of the emotion category $[F(6, 114) = 89.508,\ p < 0.001]$. Thus, we see that both for the arousal and valence ratings, the emotional categories were perceived as significantly different.

2.3. Discussion

The distribution of the perceptual scores was different in the valence-arousal space for different emotion categories. As shown in Fig. 1, the average perceptual scores of the neutral stimuli were a little lower than 3 (middle level) in the arousal space, with the perceptual scores of fear, neutral, and sad grouped together closely in the arousal space. The perceptual scores for anger, disgust, and surprise were high in the arousal space, which suggests that arousal may mainly be affected by the audio modality, similar to that found in [18].

Fundamental frequency (F_0) is well-known to be important in emotional speech perception [19–22]. A possible reason for anger, disgust and surprise to be perceived as highly aroused maybe related to F_0. The F_0 of emotional speech is calculated by STRAIGHT (speech transformation and representation by adaptive interpolation of weighted spectrogram) [23]. In order to examine the relationships between F_0 and perceptual scores in the valence-arousal spaces, F_0 *vs.* scores in the arousal space and F_0 *vs.* scores in the valence space are presented in Figs. 2 and 3. Figure 2 shows that F_0 affects the perceptual scores in the arousal space, an increase in F_0 corresponds to a higher level of scores in the arousal space. These

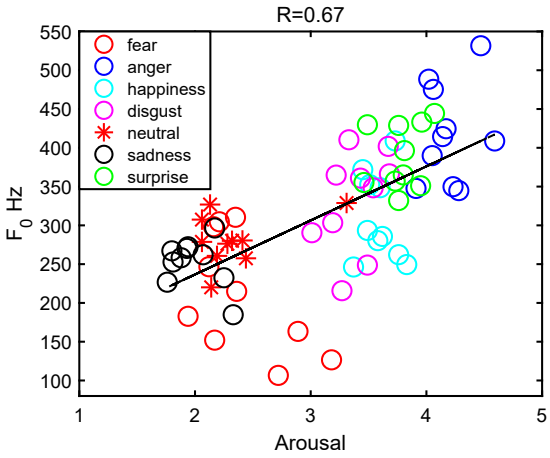

Fig. 2 F_0 (y-axis) *vs.* scores in the arousal space (x-axis), regression analysis indicates $R = 0.67$.

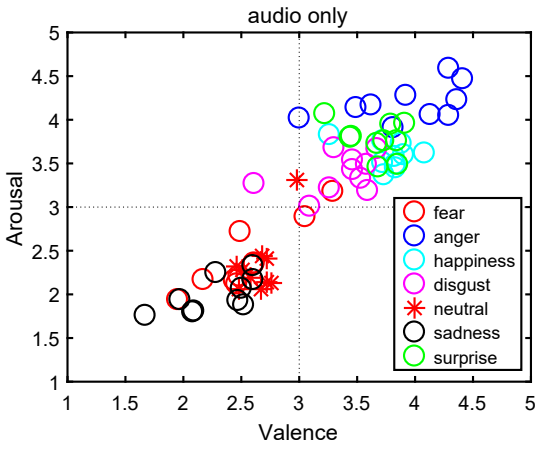

Fig. 1 The perceptual scores of the audio-only modality in the valence-arousal space.

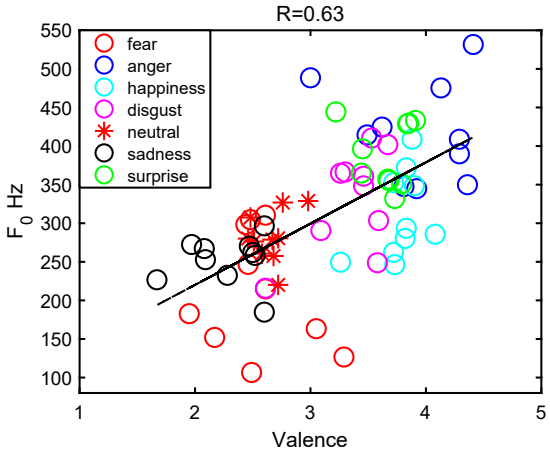

Fig. 3 F_0 (y-axis) *vs.* scores in the valence space (x-axis), regression analysis indicates $R = 0.63$.

findings align with the results reported by [24,25]. Figure 3 demonstrated that the perceptual scores in the valence space exhibits variation based on F_0 values, as increasing F_0 values coincide with higher scores in the valence space, consistent with previous findings by [26].

3. EXPERIMENT II: EMOTION PERCEPTION OF VIDEO-ONLY IN VALENCE-AROUSAL SPACE

The task of experiment II is to investigate human emotion perception in V-A space from the video-only modality.

3.1. Method

The same 70 stimuli in the experiment I were used again. To avoid participant memory effects, we conducted experiment II a week after experiment I. In this experiment, 70 stimuli were video-only without audio, which helped to explore the contribution of video-only modality to emotion perception in V-A space.

The 20 subjects who participated in experiment I also participated in this experiment. In this experiment, subjects were required to watch the female speaker's face, and each subject was asked to give a score for each stimulus based on her/his perceptual impression of valence and arousal. The experiment procedure was the same as experiment I.

3.2. Results

The average perceptual scores of the 70 emotional videos in the valence-arousal space across 20 subjects are plotted in Fig. 4. As shown in Fig. 4, the perceptual scores of neutral, fear, anger, disgust, sadness, and surprise emotional videos were located near the center $(3, 3)$ in the valence-arousal, while happiness of emotional videos was perceptually scored in the highest in both of valence and arousal.

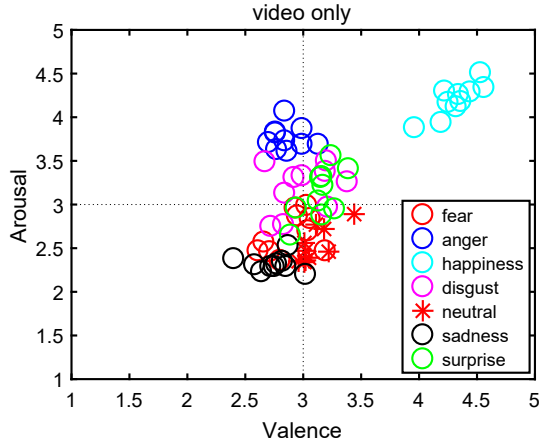

Fig. 4 The perceptual scores of the video-only modality in valence-arousal space.

In order to check whether the arousal and valence perceptual scores of the video-only modality were significantly different, we did a two-way repeated-measures analysis of variance (ANOVA) for the 20 subjects; the perceptual scores were the dependent variable and the seven emotion categories and ten sentences, the two within-subject factors. The results indicated that for the arousal ratings, there was a significant effect of the emotion category [$F(6, 114) = 53.847$, $p < 0.001$]. For the valence ratings, there was a significant effect of the emotion category [$F(6, 114) = 28.060$, $p < 0.001$]. Thus, we see that both for the arousal and valence ratings, the emotional categories were perceived as significantly different.

3.3. Discussion

The average perceptual scores of different emotions were close to each other in the valence-arousal space (except for happiness), especially perceptual scores in valence. The happiness emotional videos were perceptually scored the highest in both valence and arousal space, indicating the dominant contributions of video-only modality for happiness in the valence-arousal space. The perceptual scores of happiness in the valence-arousal space were consistent with results from a previous study [27]. The perceptual scores of neutral, fear, anger, disgust, sadness, and surprise emotional videos are similar in terms of the valence space; in the arousal space, the perceptual scores range from low to high: sadness, neutral, fear, disgust, surprise, anger, and happiness. A similar finding has been reported in precious studies [4,5].

In the field of visual emotion analysis, facial action units (AUs) are widely employed for describing facial behaviors and serve as frequently utilized features in facial expression recognition [28–30]. To explore the relationships between AUs and perceptual scores in the valence-arousal spaces, the parameter values of AUs (AU1, AU2, AU4, AU5, AU6, AU7, AU9, AU10, AU11, AU12, AU13, AU14, AU15, AU16, AU17, AU18, AU20, AU22, AU23, AU24, AU25) are calculated using the facial action coding system (FACS) [31], which is the most widely used method for analyzing facial movements [32]. The linear correlation coefficients between each AU and perceptual scores in valence-arousal space are calculated and presented in Table 1, values that are bold and exceed 0.6 signify a higher coefficient of determination. As listed values in Table 1, a total of five correlation coefficients exceeding 0.6 were observed between Action Units (AUs) and perceptual scores within the arousal and valence dimensions, respectively, we calculated the average of the five coefficients above 0.6. The average correlation coefficient between AUs and perceptual scores in arousal space was 0.62, and the average correlation coefficient between AUs and perceptual scores in valence space was 0.698. The

Table 1 Correlation coefficients between facial action units and perceptual scores in valence-arousal space. The values in bold, which are larger than 0.6, denote a higher coefficient of determination.

AUs	Description	Arousal	Valence
AU1	Inner brow raiser	0.42	0.51
AU2	Outer brow raiser	0.42	0.48
AU4	Brow lowerer	0.11	−0.07
AU5	Upper lid raiser	0.16	0.27
AU6	Cheek raiser	−0.31	−0.19
AU7	Lid tightener	−0.60	**−0.62**
AU9	Nose wrinkler	0.45	0.55
AU10	Upper lip raiser	−0.14	−0.50
AU11	Nasolabial deepener	0.020	−0.08
AU12	Lip corner puller	0.47	**0.63**
AU13	Cheek puffer	**0.63**	0.33
AU14	Dimpler	0.16	−0.10
AU15	Lip corner depressor	**0.61**	0.29
AU16	Lower lip depressor	0.07	−0.04
AU17	Chin raiser	−0.23	−0.24
AU18	Lip puckerer	0.53	**0.68**
AU20	Lip stretcher	**0.62**	0.46
AU22	Lip funneler	0.42	0.59
AU23	Lip tightener	0.38	0.13
AU24	Lip pressor	**0.63**	**0.79**
AU25	Lips part	**0.61**	**0.77**

video-only modality played a crucial role in distinguishing between these emotions in the valence space. This aligns with the work of De Gelder and Vroomen [13], which highlighted the strong impact of visual cues on emotional recognition, particularly for distinguishing positive and negative emotions.

4. EXPERIMENT III: EMOTION PERCEPTION OF AUDIO-VIDEO IN VALENCE-AROUSAL SPACE

The task of experiment III is to investigate human emotion perception in V-A space from the audio-video modality.

4.1. Method

The same 70 stimuli in experiments I and II were used again. In this experiment, 70 stimuli were audio-video modality, which helped to explore the contribution of audio-video modality to the emotion perception in V-A space.

The 20 subjects who participated in experiments I and II also participated in this experiment. To avoid participant memory effects, we conducted experiment III a week after experiment II. In this experiment, subjects were required to watch the female speaker's face while listening to her speech, and each subject was asked to give a score for each stimulus based on her/his perceptual impression of valence and arousal. The experiment procedure was the same as experiments I and II.

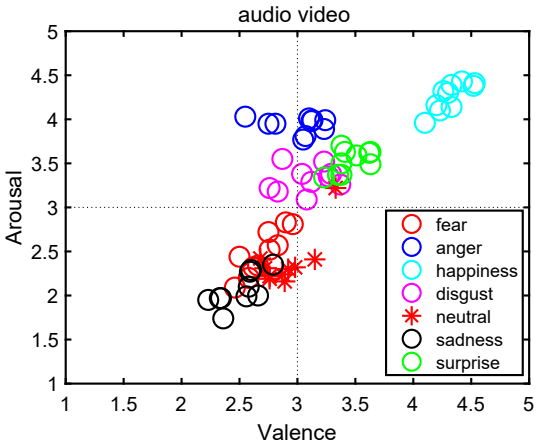

Fig. 5 The perceptual scores of the audio-video modality in the valence-arousal space.

4.2. Results

The average perceptual scores of the 70 emotional audio and video in the valence-arousal space across 20 subjects are plotted in Fig. 5. As shown in Fig. 5, the perceptual scores of each emotion category form a cluster in the valence-arousal space. More specifically, the perceptual scores of happiness emotion were the highest in both arousal and valence, while sadness emotion was perceptually scored the lowest in both valence and arousal. The perceptual scores of anger, surprise, and disgust emotion were high in arousal, but fear and neutral emotion were perceptually scored as low in arousal. It was also noted that anger, disgust, and neutral emotions were perceptually scored near the center (3) in the valence space.

In order to check whether the arousal and valence perceptual scores of the audio-video modality were significantly different, we did a two-way repeated-measures analysis of variance (ANOVA) for the 20 subjects; the perceptual scores were the dependent variable and the seven emotion categories and ten sentences, the two within-subject factors. The results indicated that for the arousal ratings, there was a significant effect of the emotion category [$F(6, 114) = 77.612$, $p < 0.001$]. For the valence ratings, there was a significant effect of the emotion category [$F(6, 114) = 34.443$, $p < 0.001$]. Thus, we see that both for the arousal and valence ratings, the emotional categories were perceived as significantly different in valence and arousal space.

4.3. Discussion

The average perceptual scores of each emotion category form a cluster in different locations in the valence-arousal spaces. The average perceptual scores of the neutral stimuli were a little lower than the middle level in arousal, which may be caused by the F_0 in the audio modality as discussed in Experiment I. The relative

Table 2 Average and standard deviation of perceptual scores of 7 emotions in valence-arousal space for three modalities.

Dimension	Arousal			Valence		
Modality	Audio-only	Video-only	Audio-video	Audio-only	Video-only	Audio-video
Fear	2.41 (0.39)	2.67 (0.23)	2.48 (0.25)	2.55 (0.38)	2.89 (0.19)	2.70 (0.16)
Anger	4.19 (0.21)	3.77 (0.13)	3.94 (0.09)	3.90 (0.45)	2.86 (0.13)	3.01 (0.22)
Happiness	3.59 (0.15)	4.20 (0.18)	4.26 (0.15)	3.78 (0.21)	4.32 (0.17)	4.32 (0.13)
Disgust	3.38 (0.21)	3.19 (0.27)	3.32 (0.14)	3.36 (0.31)	3.00 (0.23)	3.09 (0.21)
Neutral	2.33 (0.36)	2.53 (0.19)	2.38 (0.31)	2.60 (0.15)	3.10 (0.14)	2.90 (0.20)
Sadness	1.99 (0.20)	2.31 (0.09)	2.06 (0.18)	2.27 (0.31)	2.74 (0.17)	2.50 (0.18)
Surprise	3.78 (0.19)	3.13 (0.27)	3.52 (0.12)	3.65 (0.22)	3.15 (0.14)	3.45 (0.13)

positions in the valence-arousal space of seven emotions obtained in this experiment are similar to the results shown in [3], who originally proposed dimensional emotion as a method to represent a valence-arousal space. It was also noted that the perceptual scores of different sentences in each emotion category were different; this difference might be due to the different emotional degrees for the different sentences due to the linguistic content of the sentence.

5. COMPARISON OF THE PERCEIVED SCORES OF THREE MODES IN VALENCE-AROUSAL SPACE

To compare the perceived scores among different modalities in the valence-arousal space, the average and standard deviation of perceptual scores of 10 sentences in each emotion in valence-arousal space were calculated for three modalities, and are shown in Table 2.

To further examine the effect of audio-only, video-only, and audio-video modalities, the perceptual scores in the arousal and valence space for each modality in each emotion category were subjected to statistical analysis using the scores as the dependent variable, and modalities and sentences as two within-subject factors. The modalities (audio-only, video-only, and audio-video) and sentences as the two within-subject factors. For the perceptual scores of 20 subjects in arousal, two-way repeated-measures analysis of variance (ANOVA) indicated the significant effect of fear [$F(2, 38) = 3.746$, $p = 0.0328$], anger [$F(2, 38) = 3.894$, $p < 0.005$], happiness [$F(2, 38) = 40.087$, $p < 0.001$], disgust [$F(2, 38) = 3.089$, $p > 0.05$], neutral [$F(2, 38) = 1.135$, $p > 0.05$], sadness [$F(2, 38) = 9.484$, $p < 0.005$], and surprise [$F(2, 38) = 32.551$, $p < 0.001$]. For the perceptual scores of 20 subjects in valence, two-way repeated-measures ANOVA indicated the significant effect of fear [$F(2, 38) = 12.063$, $p < 0.001$], anger [$F(2, 38) = 15.413$, $p < 0.001$], happiness [$F(2, 38) = 27.368$, $p < 0.001$], disgust [$F(2, 38) = 8.328$, $p < 0.005$], neutral [$F(2, 38) = 20.118$, $p < 0.001$], sadness [$F(2, 38) = 16.724$, $p < 0.001$], and surprise [$F(2, 38) = 12.164$, $p < 0.001$].

Table 3 Statistical analysis of perceptual scores of 20 subjects for three modalities, A-V (audio *vs.* video), A-AV (audio *vs.* audio-video), V-AV (video *vs.* audio-video), significance level = 0.05.

Dimension	Arousal			Valence		
Modality	A-V	A-AV	V-AV	A-V	A-AV	V-AV
Fear	s.	n.s.	n.s.	s.	n.s.	s.
Anger	s.	n.s.	n.s.	s.	s.	n.s.
Happiness	s.	s.	n.s.	s.	s.	n.s
Disgust	n.s.	n.s.	n.s.	s.	s.	n.s.
Neutral	n.s.	n.s.	n.s.	s.	s.	s.
Sadness	s.	n.s.	s.	s.	s.	s.
Surprise	s.	s.	s.	s.	n.s.	s.

As listed scores in Table 2, the average perceptual scores were different in different modalities in both the arousal and valence space.

In order to explore whether these scores are significantly different, statistical test at 0.05 significance level was performed for audio *vs.* video, audio *vs.* audio-video, and video *vs.* audio-video. The results as shown in Table 3.

In arousal space, significant differences were identified between audio and video modalities for emotions such as fear, anger, happiness, sadness, and surprise, indicating that these modalities contribute differently to the perception of arousal. When comparing audio with audio-visual modalities, significant differences were observed only for happiness and surprise. Similarly, comparisons between video and audio-visual modalities revealed significant differences solely for sadness and surprise. These findings suggest that the audio modality tends to decrease arousal of happiness while increasing surprise, whereas the video modality increases the arousal of sadness but decreases surprise.

Regarding valence space, significant differences were found between audio and video modalities for all emotions, further suggesting distinct contributions of each modality to the perception of valence. Significant differences between audio and audio-visual modalities were noted for anger, happiness, disgust, neutral, and sadness. Additionally, comparisons between video and audio-visual modal-

ities revealed significant differences for fear, neutral, sadness, and surprise. The results suggested that the audio modality decreases the valence of happiness, neutral, and sadness emotions, while increasing the valence of anger and disgust. Conversely, the video modality increases the valence of fear, neutral, and sadness emotions while decreasing surprise.

It is noteworthy that the average perceptual scores for surprise in the audio-video modality occupy an intermediate position between the scores of the audio-only and video-only modalities within the arousal space. Similarly, in the valence space, the average perceptual scores for neutral and sadness in the audio-video modality are situated between those of the audio-only and video-only modalities. This suggests a complex interaction between auditory and visual cues in the perception of emotional valence and arousal. There are two possible reasons for this result, one of which is similar to Arimoto and Okanoya's [15] findings that incongruent emotional expressions in audio and visual modalities can be integrated as the brain synthesizes a middle ground between the conflicting cues from each modality, often resulting in the perception of a third emotion. Such an integration could explain why the audio-video modality results fell between the unimodal scores. A second reason may be cultural differences. Kawahara *et al.* [33] highlight how cultural shaping from audio and videos affects emotion perception, which also may factor in our findings. The findings from this study contribute to drawing a more comprehensive picture of multimodal emotion perception in Mandarin Chinese.

Figures 1, 4, and 5 show perceptual scores of different modalities in the valence-arousal space. The perceptual scores of the video-only modality were centralized (except for happiness), whereas the perceptual scores of the audio-only modality were more dispersedly distributed, which indicates the important contributions of the audio-only modality on the perception of emotions. In comparison with the perceptual scores of the video-only modality, the perceptual scores of the audio-video modalities were more dispersedly distributed, which further indicates the important contributions of the audio modality in the valence-arousal space. This result is in line with previous audio-visual modality findings that the audio modality plays a dominant role in emotion perception of audio-visual modalities [34,35]. This is also consistent with findings by studies [13,36] that highlight the prevalence of auditory cues for certain dimensions of emotion perception, particularly for arousal perception. More importantly, Mower *et al.* [36] found that the distribution of the audio-video modalities are the largest in valence-arousal spaces. It is further noted that the perceptual differences (the distances in the valence space) for anger and happiness in the audio-only modality were much smaller than those in the video-only modality. More specifically, in the video-only modality, the perceptual scores of anger showed movements toward the low valence space, whereas the perceptual scores of happiness showed movements toward the high arousal and high valence space. This further indicates the important contributions of the video-only modality in separating anger and happiness emotions in the valence space.

In summary, audio-only, video-only, and audio-video modalities each uniquely affect the perception of emotional arousal and valence. Compared to video-only modality, audio-only modality generally decrease arousal and valence at lower levels, and increase arousal and valence at higher levels. This may be specific to Mandarin Chinese.

We only used data from one female speaker in this paper, the results may also contain effects related to the speaker's individuality. Therefore, future work is to capture the common characteristics of multiple speakers.

6. CONCLUSION

This study investigated the contributions of audio, visual, and audio-visual modalities to the perception of Mandarin Chinese emotions in the valence-arousal space. Three psychoacoustic experiments were conducted with native Chinese subjects using audio-only, video-only, and audio-video stimuli. The following findings were obtained:

- Modality-Specific Contributions: Different modalities contributed differently to the perception of valence and arousal. The audio-only modality had a more dispersed distribution of perceptual scores, indicating its significant role in emotion perception. The video-only modality, however, played a crucial role in distinguishing emotions, particularly in separating anger and happiness in the valence space.
- Contributions of audio-only and video-only modalities: Compared to video-only modality, audio-only modality generally decreases arousal and valence at lower levels, and increases arousal and valence at higher levels, the findings may be specific to Mandarin Chinese.
- Influence of F_0 and AUs: The study found that F_0 significantly affects perceptual scores in the arousal space, with higher F_0 values corresponding to higher arousal scores. Additionally, AUs showed significant correlations with perceptual scores in the valence-arousal space, especially for valence space, highlighting the importance of visual cues in emotional perception.

The findings of the present study are valuable for gaining insight into multimodal emotion perception in the valence-arousal space for the Mandarin Chinese. Furthermore, they contribute to the comprehension of multimodal dimensional emotion recognition, with potential implica-

tions for the integration of multimodal fusion strategies into deep learning methods for the Mandarin Chinese dimensional emotion recognition systems.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grants 62201571 and U21B2010, and in part by the Key Laboratory of Linguistics, Chinese Academy of Social Sciences (Project #2024SYZH001).

REFERENCES

[1] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *J. Pers. Soc. Psychol.*, **17**, 124–129 (1971).

[2] R. Elbarougy and M. Akagi, "Improving speech emotion dimensions estimation using a three-layer model of human perception," *Acoust. Sci. & Tech.*, **35**, 86–98 (2014).

[3] J. A. Russell, "A circumplex model of affect," *J. Pers. Soc. Psychol.*, **39**, 1161–1178 (1980).

[4] A. Mollahosseini, B. Hasani and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affect. Comput.*, **10**, 18–31 (2017).

[5] L. Zhang, D. Tjondronegoro and V. Chandran, "Representation of facial expression categories in continuous arousal–valence space: Feature and correlation," *Image Vis. Comput.*, **32**, 1067–1079 (2014).

[6] Z. Peng, J. Dang, M. Unoki and M. Akagi, "Multi-resolution modulation-filtered cochleagram feature for lstm-based dimensional emotion recognition from speech," *Neural Networks*, **140**, 261–273 (2021).

[7] B. Kreifelts, T. Ethofer, W. Grodd, M. Erb and D. Wildgruber, "Audiovisual integration of emotional signals in voice and face: An event-related fMRI study," *NeuroImage*, **37**, 1445–1456 (2007).

[8] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, **264**, 746–748 (1976).

[9] S. Fagel, "Emotional McGurk effect," *Proc. Int. Conf. Speech Prosody*, Vol. 1, Citeseer (2006).

[10] S. Karuppali, J. S. Bhat, P. Krupa and Shreya, "An auditory-visual conflict of emotions-evidence from McGurk effect," *Adv. Life Sci. Technol.*, **4**, 51–57 (2012).

[11] A. Li, Q. Fang, Y. Jia and J. Dang, "Emotional McGurk effect? a cross-cultural investigation on emotion expression under vocal and facial conflict," *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, NLP-NABD CCL 2013, pp. 214–226 (2013).

[12] E. M. Provost, Y. Shangguan and C. Busso, "Umeme: University of michigan emotional McGurk effect data set," *IEEE Trans. Affect. Comput.*, **6**, 395–409 (2015).

[13] B. De Gelder and J. Vroomen, "The perception of emotions by ear and by eye," *Cogn. Emot.*, **14**, 289–311 (2000).

[14] S. Takagi, S. Hiramatsu, K.-i. Tabei and A. Tanaka, "Multisensory perception of the six basic emotions is modulated by attentional instruction and unattended modality," *Front. Integr. Neurosci.*, **9**, Article 1 (2015).

[15] Y. Arimoto and K. Okanoya, "Dimensional mapping of multimodal integration on audiovisual emotion perception," *Proc. Auditory-Visual Speech Processing* (*AVSP*) *2011*, pp. 93–98 (2011).

[16] J. Dang, A. Li, D. Erickson, A. Suemitsu, M. Akagi, K. Sakuraba, N. Minematsu and K. Hirose, "Comparison of emotion perception among different cultures," *Acoust. Sci. & Tech.*, **31**, 394–402 (2010).

[17] A. Li, *Encoding and Decoding of Emotional Speech: A Cross-cultural and Multimodal Study between Chinese and Japanese* (Springer, Berlin, Heidelberg, 2015).

[18] R. Banse and R. Klaus, "Acoustic profiles in vocal emotion expression," *J. Pers. Soc. Psychol.*, **70**, 614–636 (1996).

[19] M. Goudbeek and K. Scherer, "Beyond arousal: Valence and potency/control cues in the vocal expression of emotion," *J. Acoust. Soc. Am.*, **128**, 1322–1336 (2010).

[20] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Commun.*, **40**, 227–256 (2003).

[21] T. Banziger and K. R. Scherer, "The role of intonation in emotional expressions," *Speech Commun.*, **46**, 252–267 (2005).

[22] M. Bulut and S. Narayanan, "On the robustness of overall f0-only modifications to the perception of emotions in speech," *J. Acoust. Soc. Am.*, **123**, 4547–4558 (2008).

[23] H. Kawahara, I. Masuda-Katsuse and A. De Cheveigne, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech Commun.*, **27**, 187–207 (1999).

[24] T. Johnstone and K. R. Scherer, "Vocal communication of emotion," in *Handbook of Emotions*, 2 (2000), pp. 220–235.

[25] Y. Li, J. Li and M. Akagi, "Contributions of the glottal source and vocal tract cues to emotional vowel perception in the valence-arousal space," *J. Acoust. Soc. Am.*, **144**, 908–916 (2018).

[26] T. Waaramaa, A.-M. Laukkanen, M. Airas and P. Alku, "Perception of emotional valences and activity levels from vowel segments of continuous speech," *J. Voice*, **24**, 30–38 (2010).

[27] K. Sun, J. Yu, Y. Huang and X. Hu, "An improved valence-arousal emotion space for video affective content representation and recognition," *Proc. IEEE Int. Conf. Multimedia and Expo. 2009*, pp. 566–569 (2009).

[28] S. Velusamy, H. Kannan, B. Anand, A. Sharma and B. Navathe, "A method to infer emotions from facial action units," *IEEE Int. Conf. Acoust. Speech Signal Process.* (*ICASSP*) *'11*, pp. 2028–2031 (2011).

[29] P. Khorrami, T. Paine and T. Huang, "Do deep neural networks learn facial action units when doing expression recognition?" *Proc. IEEE Int. Conf. Computer Vision Workshops*, pp. 19–27 (2015).

[30] E. L. Rosenberg and P. Ekman, *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System* (*FACS*) (Oxford University Press, Oxford, 2020).

[31] P. Ekman and W. V. Friesen, "Facial action coding system," *Environ. Psychol. Nonverbal Behav.* (1978).

[32] W. Yan and Y. Chen, "Measuring dynamic micro-expressions via feature extraction methods," *J. Comput. Sci.*, **25**, 318–326 (2018).

[33] M. Kawahara, D. A. Sauter and A. Tanaka, "Culture shapes emotion perception from faces and voices: Changes over development," *Cognition Emotion*, **35**, 1175–1186 (2021).

[34] L. Piwek, F. Pollick and K. Petrini, "Audiovisual integration of emotional signals from others' social interactions," *Front. Psychol.*, **6**, Article 611 (2015).

[35] K. Petrini, P. McAleer and F. Pollick, "Audiovisual integration of emotional signals from music improvisation does not depend on temporal correspondence," *Brain Res.*, **1323**, 139–148 (2010).

[36] E. Mower, M. J. Mataric and S. Narayanan, "Human perception of audio-visual synthetic character emotion expression in the presence of ambiguous and conflicting information," *IEEE Trans. Multimedia*, **11**, 843–855 (2009).

Yongwei Li received the M.S. and Ph.D. degrees in information science from the Japan Advanced Institute of Science and Technology, Nomi, Japan, in 2014 and 2018, respectively. He is currently an Assistant Professor with the CAS Key Laboratory of Behavioral Science, Institute of Psychology, Chinese Academy of Sciences, Beijing, China. His research interests include speech emotion analysis, and modeling of speech production.

Aijun Li is a research fellow, the deputy director of the Institute of Linguistics, Chinese Academy of Social Sciences (CASS), and the head of the CASS Key Laboratory of Linguistics. She received her master's and bachelor's degrees from the Department of Computer Science at Tianjin University, as well as her Ph.D. from the School of Human Information Science at JAIST, and. She currently serves as the vice president of the Chinese Linguistic Society, a member of the Oriental COCOSDA steering committee, a member of the Academic Committee of the Ministry of Public Security's Key Laboratory of Intelligent Speech Technology. She is also the chief editor of the Chinese Journal of Phonetics and on the editorial boards for Studies of Chinese Language, Minority Languages of China and China Scientific Data. Her recent research has focused on intonation typology, first and second language acquisition, and the production and perception of speech in interaction that is closely related to AI.

Jianhua Tao received the M.S. degree from Nanjing University, Nanjing, China, in 1996 and the Ph.D. degree from Tsinghua University, Beijing, China, in 2001. He is currently a Professor with the Tsinghua University, Beijing, China. He has authored or co-authored more than 200 papers on main journals and proceedings, including the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING. His current research interests include speech recognition, speech synthesis, human computer interaction, affective computing, and pattern recognition. He is the Board Member of ISCA, the Chair or Program Committee Member for several main conferences, including Interspeech, ICPR, ACII, ICMI, and ISCSLP. He was the Steering Committee Member of the IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, and is an Associate Editor for the *Journal on Multimodal User Interfaces and International Journal of Synthetic Emotions*. He was the recipient of several awards from the important conferences, such as Interspeech and NCMMSC.

Feng Li received the Ph.D. degree in information science from Graduate School of Advanced Science and Technology, Japan Advanced Institute of Science and Technology (JAIST), Ishikawa, Japan in 2019. He is now an Associate Professor with the Department of Computer Science and Technology, Anhui University of Finance and Economics, Bengbu, China. He is also a Postdoctoral Researcher with the School of Information Science and Technology, University of Science and Technology of China (USTC), Hefei, China. His research interests include audio source separation, speech emotion recognition, and the signal processing of speech.

Donna Erickson received the B.A. degree from The Ohio State University, Columbus, OH, USA, in 1966, the M.A. degree from the University of Michigan, Ann Arbor, MI, USA, in 1968, and the Ph.D. degree from the University of Connecticut, Storrs, CT, USA, in 1976. Her Ph.D. thesis was on the laryngeal electromyographic activity underlying the tones of Thai. She taught linguistics and English as a Second Language with Earlham College and The Ohio State University from 1982 to 1996. During this time, she was also a Research Scientist with the Center for Cognitive Linguistics, The Ohio State University, working with Professor Osamu Fujimura on voice quality and analyzing X-ray Microbeam data. In 1998, she was a Visiting Professor with Kanazawa University, Japan, from 2000 to 2006, a Professor with Gifu City Women's College, Japan, and from 2006 to 2012, a Professor with the Show University of Music, Kawasaki, Japan. She retired from full-time teaching in 2012, and is currently an Affiliated Research Scientist with Haskins Laboratories, New Haven, CT, USA. She continues to be active in research, focusing on the acoustic and articulatory characteristics of voice production, as it pertains to emotional and social affective expressions, and also voice production changes during singing.

Masato Akagi (Life Member, IEEE) received the B.E. degree from the Nagoya Institute of Technology, Nagoya, Japan, in 1979, and the M.E. and Ph.D. (Eng.) degrees from the Tokyo Institute of Technology, Tokyo, Japan, in 1981 and 1984, respectively. In 1984, he joined the Electrical Communication Laboratories, Nippon Telegraph and Telephone Corporation. From 1986 to 1990, he was with ATR Auditory and Visual Perception Research Laboratories. Since 1992, he has been a Faculty Member of the School of Information Science, JAIST, and is currently a Professor Emeritus. His research interests include speech perception, modeling of speech perception mechanisms in humans, and signal processing of speech. Dr. Akagi was the recipient of the IEICE Excellent Paper Award from the IEICE in 1987, Best Paper Award from the Research Institute of Signal Processing in 2009, and Sato Prize for Outstanding Papers from the Acoustical Society of Japan in 1998, 2005, 2010, and 2011. And he was a president of the Acoustical Society of Japan (ASJ) in 2011–2013.