J Ambient Intell Human Comput (2017) 8:913–924
DOI 10.1007/s12652-016-0406-z

ORIGINAL RESEARCH

CHEAVD: a Chinese natural emotional audio–visual database

Ya Li[1] · **Jianhua Tao**[1,2,3] · **Linlin Chao**[1] · **Wei Bao**[1,4] · **Yazhu Liu**[1,4]

Received: 30 March 2016 / Accepted: 22 August 2016 / Published online: 10 September 2016

Abstract This paper presents a recently collected natural, multimodal, rich-annotated emotion database, CASIA Chinese Natural Emotional Audio–Visual Database (CHEAVD), which aims to provide a basic resource for the research on multimodal multimedia interaction. This corpus contains 140 min emotional segments extracted from films, TV plays and talk shows. 238 speakers, aging from child to elderly, constitute broad coverage of speaker diversity, which makes this database a valuable addition to the existing emotional databases. In total, 26 non-prototypical emotional states, including the basic six, are labeled by four native speakers. In contrast to other existing emotional databases, we provide multi-emotion labels and fake/suppressed emotion labels. To our best knowledge, this database is the first large-scale Chinese natural emotion corpus dealing with multimodal and natural emotion, and free to research use. Automatic emotion recognition with Long Short-Term Memory Recurrent Neural Networks (LSTM-RNN) is performed on this corpus. Experiments show that an average accuracy of 56 % could be achieved on six major emotion states.

Keywords Audio–visual database · Natural emotion · Corpus annotation · LSTM · Multimodal emotion recognition

✉ Ya Li
yli@nlpr.ia.ac.cn

Jianhua Tao
jhtao@nlpr.ia.ac.cn

Linlin Chao
linlin.chao@nlpr.ia.ac.cn

Wei Bao
jsnubw@163.com

Yazhu Liu
yzliu90@163.com

[1] National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, Beijing, China

[2] CAS Center for Excellence in Brain Science and Intelligence Technology, Institute of Automation, Chinese Academy of Sciences, Beijing, China

[3] School of Computer and Control Engineering, Graduate University of Chinese Academy of Sciences, Beijing, China

[4] Institute of Linguistic Sciences, Jiangsu Normal University, Jiangsu, China

1 Introduction

Human emotion analysis and recognition have attracted a lot of interest in the past decades and extensive research has been done in neuroscience, psychology, cognitive sciences, and computer sciences. In which, emotional corpus is crucial in the building of emotion recognition model. This paper reports our work on constructing a multimodal emotional corpus.

The existing emotional corpus could be divided into three types: simulated/acted, elicited and natural corpus (Ververidis and Kotropoulos 2006; Wu et al. 2014) according to how the emotions are expressed. There are many acted emotional corpora (Ververidis and Kotropoulos 2006), which is usually acted by professional actors in a controlled conditions. Three widely used emotional corpora are introduced here: the Berlin Emotional Speech Database (Burkhardt et al. 2005), the CASIA Chinese Emotional Corpus and the Danish Emotional Speech Corpus (Engberg and Hansen 1996). The Berlin Emotional Speech Database contains segments recorded by ten subjects in seven

emotional classes: neutral, anger, fear, happiness, sadness, disgust and boredom with emotional scripts. The CASIA Chinese Emotional Corpus contains fifty linguistically neutral sentences portrayed by four subjects in six emotions, namely, *angry*, *fear*, *happy*, *neutral*, *sad*, and *surprise*. The Danish Emotional Speech Corpus contains semantically neutral utterances recorded by four subjects in four emotions, *anger*, *joy*, *surprise* and *sadness*. The two main limitations of these corpora are that the prototypical emotions were recorded by limited speakers, and the emotions were acted in isolated sentences.

Besides the acted corpus recorded in laboratories, there are several corpora attempt to record near natural emotional data. The emotion is elicited by partners, figures and/or videos in such kind of corpus. The Interactive emotional dyadic motion capture database (IEMOCAP) (Busso et al. 2008) and FAU Aibo Emotion Corpus are successful efforts to record spontaneous emotional states. IEMOCAP database is recorded in the condition of ten skilled actors performing selected emotional scripts. The FAU Aibo Emotion Corpus (Steidl 2009) consists of children's spontaneous emotions when they play with the Sony robot in the chosen scenarios. Concerning elicited emotion corpus in Chinese, Yuan et al. (2002) designed eight emotion-induction stories and use them to elicit emotions. Their emotion database of total 288 sentences is collected from nine speakers.

Recently, the demanding for real application forces the emotion research shift to natural and spontaneous corpus. In addition, this research shift also includes combining multiple modalities for analysis and recognition of human emotion. In all the modalities, speech and facial expression are the two important approaches in human emotional expression and interaction (Jaimes and Sebe 2007). The combination of these two modalities will no doubt improve the robustness of emotional recognition models. The past decades have seen a substantial body of literature on multimodal emotion recognition (Jaimes and Sebe 2007; Tao and Tan 2005). The natural and spontaneous emotional corpus is the recording in a real scenario, e.g., in human daily life. Morrison et al. (2007) collected a small speech emotional corpus from a call-center which includes 388 utterances. Most of the utterances are neutral, and some of them are angry. The distributions of happiness, sadness, fear, disgust, and surprise are very low. Because of the copyright and privacy issues, most of the other existing natural emotion corpus is collected from films and TV programs (Yu et al. 2001), etc. Although films and TV programs are often shot in controlled environments, they are significantly closer to real-world environments than the lab-recorded datasets. Some of the successful examples are the Belfast Natural Database (Douglas-Cowie et al. 2000), the Vera am Mittag German Audio–visual Emotional Speech Database (VAM) (Grimm et al. 2008), EmoTV Database (Devillers et al. 2006) and the SAFE (Situation Analysis in a Fictional and Emotional Corpus) Corpus (Clavel et al. 2006). The Belfast Natural Database contains audio–visual recordings extracted from English TV programs and interviews. There are 125 subjects and around 10–60 s of each clip. The VAM database contains audio–visual clips segmented from the German TV talk shows. The EmoTV Database (Devillers et al. 2006) consists of audio–visual interactions from TV interviews—both sedentary interactions and interviews 'on the street', and there are only 48 subjects in this corpus. The SAFE corpus aims to automatically detect extreme emotions occurring in abnormal situations, especially fear, that extracted from movies. HUMAINE Database (Douglas-Cowie et al. 2007) is a collection of several databases, and contains naturalistic and induced multimodal emotional data. Acted Facial Expressions In The Wild (AFEW) (Dhall et al. 2012) is also selected from films etc. Two annotators participated in the emotion annotation. However, this corpus focuses on the facial expression; some clips only contain background noise, which are difficult to process from the aspects of audio processing. Recently, another way to collect natural emotional data is under the human–computer or human–human interaction tasks (Wu et al. 2014). They talked and acted freely while the multimodal data were recorded. Some examples of these corpora are SEMAINE (McKeown et al. 2012), RECOLA (Ringeval et al. 2013), and AVDLC (Valstar et al. 2013). Regarding Chinese natural emotional corpus, Yu et al. (2001) selected 721 utterances from Chinese teleplays to build a corpus with anger, happiness sadness and neutral states. There are also some emotional corpora selected from movies (Clavel et al. 2004), which indicates that this strategy is practicable in natural emotion data collection.

Apart from the emotional data selection, emotion annotation is another important task in building emotion database. Emotion annotation strategy is related to how to describe emotion. But, emotion is a very complex phenomenon (Cowie and Cornelius 2003), as Fehr and Russell stated (1984), "Everyone knows what an emotion is, until asked to give a definition". Therefore, finding appropriate methods is not straightforward. According to the research in the area of psychology, two major groups can be distinguished (Wöllmer et al. 2008): (1) categorical approach, (2) dimensional approach. The categorical approach is based on a small number of basic and distinct emotions. But there has been considerable variability in published lists of the basic emotions. In this approach, the big six (Ekman 1999) is the principal approach, namely, *happiness*, *sadness*, *surprise*, *fear*, *anger* and *disgust*, which is suggested by Ekman by watching facial expressions. The dictionary of affect in language listed approximately 4500

English words (Whissel 1989). Averill (1975) listed as much as 558 emotional words to describe everyday emotions. Fehr and Russell reported 196 emotion words in English (Fehr and Russell 1984), while Plutchik gave 142 emotion words and created a wheel of emotions. He also suggested eight primary bipolar emotions: joy versus sadness; anger versus fear; trust versus disgust; and surprise versus anticipation (Plutchik 1980). Shaver et al. (1987) considered 135 words by manually sorting, and classified them into six types, including *love*, *joy*, *surprise*, *anger*, *sadness* and *fear,* which is a little bit different from the big six. Russell proposed 28 emotional words in a circle by a principal components analysis of 343 subjects self-report of their current affective states (Russell 1980). Differences within languages directly correlate to differences in emotion taxonomy. As for the emotions in Chinese, Xu and Tao (2003) also discussed the emotion words in Chinese. Gao and Zhu use 13 emotional words to describe emotions in broadcasting news (Gao and Zhu 2012). In the second dimensional approach, two or three numerical dimensions (Mehrabian 1996; Wöllmer et al. 2008) are used to describe emotions. Pleasure-Arousal-Dominance (PAD) Emotional State Model was proposed by Mehrabian and Russell in 1974 to describe human perceptions of physical environments (Mehrabian and Russell 1974). The pleasure dimension refers to how positive or negative the emotion is and the arousal dimension refers to how excited or apathetic the emotion is. Dominance is related to feelings of control and the extent to which an individual feels restricted in his behavior. Dominance is also referred as *Potency* in other researchers. Russell (1980) further simplified the model into two dimensions: valence and arousal. The comparison of the discrete and dimensional approaches is extensively studied (Barrett 1998), and the mapping between these two approaches is also investigated.

People may have the experience that, for some purpose, attempt to regulate their emotions, for instance by intensifying/weakening, masking, or completely hiding them (Kashdan and Breen 2008), which is called emotion regulation (Gross 2002). In this case, fake/suppressed emotion (Gross 2002) is defined as inhibiting ongoing emotion-expressive behavior, which is essential to socialization. The research and theory on emotion regulation, suppressed emotion has received considerable attentions in the recent decades (Butler et al. 2007), because emotion exchange is crucial in social interactions. Suppressed emotion should be treated difficulty in multimodal emotion annotation (Douglas-Cowie et al. 2007; Schröder et al. 2006).

As we introduced above and reported in some other review literatures (Douglas-Cowie et al. 2003; El Ayadi et al. 2011; Ververidis and Kotropoulos 2006; Wu et al. 2014), there was a great effort on building emotional databases in the past decades. However, three major problems (Cowie and Cornelius 2003; El Ayadi et al. 2011) still exist: Firstly, the large-scale emotional data in real application scenarios is relatively few compare with the large demanding of real applications. The emotion recognition performance on acted data is good, for instance, more than 80 % accuracy could be obtained and reported in several literatures (Eyben et al. 2009; Li et al. 2015). However, these models would fail in practical applications, because the test scenario is different from the trained scenario. Secondly, the numbers of subjects in these corpora still need to increase to make speaker independent emotion analysis and recognition infeasible, which is also inevitable when transforming this research into real applications. Thirdly, and more importantly, cultural differences have been observed in the way in which emotions are expressed (Rose 1991), valued, regulated (Butler et al. 2007) and perceived (Elfenbein and Ambady 2002). However, there are few natural emotional databases available in Chinese.

To overcome the above problems in Chinese multi modal emotional analysis and recognition, we have built a multi-modal emotion database named CASIA Natural Emotional Audio–Visual Database (CHEAVD). The main advantages of CHEAVD are as follows: Firstly, This is the first large scale corpus dealing with Chinese natural audio–visual emotions, which includes 140 min data selected from a variety of TV programs and films. Secondly, there are 238 speakers in this corpus, and the large number of speakers renders speaker-independent emotion analysis and recognition possible. Compared with the review of the existing emotional corpus in (El Ayadi et al. 2011; Ververidis and Kotropoulos 2006), CHEAVD includes the largest number of speakers. Thirdly, 26 non-prototypical emotional labels were provided instead of prototypical emotional labels. This annotation is valuable because the basic big six emotions are not enough to describe all the emotional expression in real life.

The other characteristics of this corpus are: (1) we provide multi-labels to reduce the ambiguity of the emotion perception in the real world (Douglas-Cowie et al. 2003); (2) we also provide fake/suppressed emotion (Gross 2002) labeling in this audio–visual corpus. In the real world environment, emotion is often complex and it is difficult to classify a segment into one specific emotion state. There are many superpositions of multiple emotions and masking of one emotion by another one. Therefore, we use multiple emotion labels and also provide fake/suppressed emotion labels in CHEAVD.

To get an overall impression of CHEAVD, we also carried out a baseline audio–visual emotion recognition experiments on this corpus. Long Short-Term Memory Recurrent Neural Networks (LSTM-RNN) (Hochreiter and Schmidhuber 1997), which is one of the state-of-art

machine learning techniques used to capture the temporal relation of input features, is utilized to model the audio and visual streams (Chao et al. 2015; He et al. 2015) in our work. The fact that multi modal emotion recognition surpasses single modality emotion recognition is verified by many previous work (Busso et al. 2004; He et al. 2015; Russell et al. 2003); however, this elicits difficulty in modality fusion. Currently, most of the studies combine the two modalities in decision level (Liu et al. 2014), in which, the single-modal recognition results obtained from individual features are combined in the final decision. Correspondingly, feature level fusion is extracting audio and visual features, etc., separately, pooling these features to single vectors for each feature sets and then concatenating these vectors into one single feature vector for classification (Busso et al. 2004). The apparent problem of this method is the temporal coupling among each modality is ignored. To overcome this problem, Tripled HMM (Song et al. 2004) and Multi-stream Fused HMM model (Zeng et al. 2005) are proposed to make use of the correlation between audio and visual streams in the unified HMM model framework. While the audio and visual signals always have different frame rates, temporal alignment before audio and visual features fed into these models is necessary, which is often manually operated. Recently, soft attention mechanism (Bahdanau et al. 2014; Mnih et al. 2014) is introduced to choose more important units within contexts, by which, models can learn alignment between different modalities. Therefore, we use soft attention mechanism to re-weight and combine the hidden representations of audio–visual feature sequences in LSTM-RNN for final classification. Experiments show that an average accuracy of 56 % could be achieved on the six major emotion states.

The paper is organized as follows. Section 2 gives the main principles of the raw data selection. Section 3 describes the details about emotion annotation and distribution. Section 4 presents the emotion recognition results with LSTM-RNN on this corpus. Finally, Sect. 5 gives the conclusion.

2 Chinese natural emotional audio–visual data selection

This section first gives the principles of raw emotion materials selection and the segmentation process, and then presents the distribution of age and gender group in database.

2.1 Principles of data selection

Several factors should be considered in the emotional data selection (El Ayadi et al. 2011). Many television shows display relatively strong emotion. We selected the raw data by watching a range of television programs over a period of one and a half months. The principles considered in the raw data selection are listed below.

a. The shows of the candidates should be close to real life scenarios, which means science fiction and the like are not accepted.
b. Films should contain actors' own voice rather than being dubbed into Chinese.
c. Mandarin is preferred, be cautious with the selection of materials with strong accent.
d. The shows should contain real interactions rather than acted materials.

The show types we select from are (a) films and television series tracing the real life scenarios, (b) chat shows, (c) impromptu speech shows and (d) talk shows.

Films and television series are the primary sources of CHEAVD. To avoid acted emotion, we select those films and television series reflecting real-life environment and skilled actors engaged in their roles. 34 films and 2 television series are selected for our corpus, for example, *We Get Married*, *Lost in Thailand*, *Silent Witness*, *Life is a Miracle*, *One Night in Supermarket*, *Under the Hawthorn Tree, Finding Mr. Right, Letter from an Unknown Woman* and *So Young*. All of these films are based on real stories and real life rather than science fiction and ancient life.

Chat shows provide many positive emotional materials. The emotional range tended to be limited to negative emotions that were determined by the politeness habits and cultural behaviors. Each show consists of several dialogues between two persons. The talk is moderated by one host. Discussion usually involves personal experiences and feelings. Two television shows are eventually selected, namely *Yang Lang One On One* and *Ke Fan Qing Ting*.

Impromptu speech shows and talk shows contain relative emotional materials. Various topics are covered in the shows, culture, sport, politics, society and art and so on. *Super Speaker* and *Tonight 80's Talk Show* are eventually selected as a counterbalance to the single emotion type expressed in the chat shows. These four types of shows act as complements for each other in positive and negative emotions.

Finally, CHEAVD has selected 34 films, 2 television series, 2 television shows, 1 impromptu speech and 1 talk show, in which, films and television series constitute the majority part.

2.2 Process of data segmentation

After the raw data selection, we need to segment all the long video into segments which show emotion. Firstly, we make a record of the start and end time of each segment,

accurate to 10 ms. This step is carried out manually by one of the authors. Then both original video files and the corresponding time stamp files were used as a reference to divide the original video into video and audio segments. Finally both video files and audio files were reserved.

Before the final emotion annotation, all the segments are manually checked again to guarantee the quality of facial expression and audio signal. The selection requirements are as follows.

a. Segments with high background music, noise and voice overlap are removed.
b. Each segment should contain only one speaker's speech and facial expression.
c. Video background should not be too dark.
d. Speech segment should contain a complete utterance. Segment without utterance is removed.

Many segments were finally discarded to ensure the quality of this corpus, and only 141 min were selected, which is 2 % of the raw data. The duration of these 2600 segments is from 1 to 19 s and the average duration is 3.3 s. The video files extracted from shows are captured as AVI files. All video segments are saved in 640 × 480 pixels with a frame rate of 25 fps. The wave files extracted from video files were stored at 44.1 kHz, stereo and 16 bit.

2.3 Age and gender distribution

To make speaker-independent research possible, CHEAVD provides about 238 different speakers' audio–visual data for this purpose. The partition of the recordings to gender is as follows: 52.5 % were male speakers, 47.5 % were female speakers.

The 238 speakers were between 11 and 62 years old. Considering that emotional expressions are not the same among each age group (Gross et al. 1997), 6 age categories were defined for balance as shown in Table 1. The percentage for each age group is shown as Fig. 1.

Table 1 Six age groups in database

Age	Description
a: <13	Child
b: 13–16	Mutation
c: 16–24	Youth
d: 25–44	Young
e: 45–59	Quinquagenarian
f: ≥60	Elder

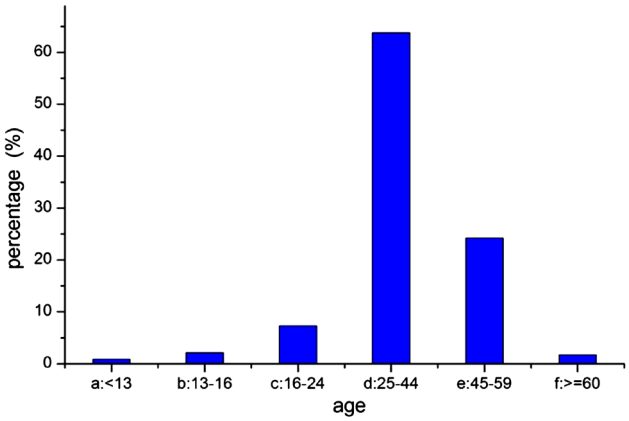

Fig. 1 The distribution of six age groups in CHEAVD

3 Emotion annotation

This section first discusses the appropriate emotion category for Chinese, and then elaborates the annotation strategy adopted in CHEAVD.

3.1 Emotion taxonomy

To construct a natural emotional corpus, reasonable emotional annotation method is a prerequisite. In this work, we adopted the category method for two reasons, although dimensional method is widely used to describe complex emotions in affective computing area. Firstly, our corpus is selected from TVs and movies, and the segments are very short. The short-segment makes it difficult to choose dimensional labeling method. In the experiment of assessing the reliability of the widespread dimensional labeling instrument, FEELTRACE, the typical durations of the clips are from 15 to 30 s (Cowie et al. 2000), which are much longer than our segments. Another reason, but less important, is category method is easy to understand, which have more demands from applications. Therefore, we adopted the category labeling method.

Regarding to the emotion category, as we introduced above, there is no unified basic emotion set. Most emotional databases address prototypical emotion categories, such as *happy*, *angry* and *sad*. However, people exhibit non-prototypical, subtle emotional states in everyday interactions. The common sense and psychological studies suggest that the full spectrum of human emotion cannot be expressed by a few discrete classes. So recognizing non-prototypical emotions; e.g., *shy*, *nervous*, *worried* and *anticipated*, is useful in human–computer interaction.

Take all of the above problems into consideration, we allowed non-limited emotional categories before annotation. During the annotation process, we added the emotional categories gradually based on all annotators' subjective feelings. However, this is not an unlimited renewal. The

Table 2 Non-prototypical emotions in CHEAVD

Category	Description
Shy	A character of implicative
Nervous	A feeling of slight fear or showing anxiety
Proud	A feeling of pleased with something has been done or possess
Frustrated	Disappointingly unsuccessful
Worried	Mentally upset over possible misfortune
Anticipated	Expected hopefully
Sarcastic	Laugh grimly
Helpless	Do not have the strength or power to do anything useful
Confused	Perplexed by many conflicting situations
Aggrieved	Cause to feel sorrow
Contemptuous	Show disrespect to others
Hesitant	A feeling of uncertainty

annotator were asked to discuss every new "emotion label" before it was added in. After several rounds, the emotion label list was fixed. Different from the other emotional databases, we provided several properties especially for Chinese. For example, the emotion "shy" in our corpus reflects the implicative character of Chinese. Table 2 shows the description of non-prototypical emotional categories in CHEAVD.

Both the segmentation and annotation were carried out by four native Chinese speakers. Video and audio files were provided for these annotators. Annotation is based on watching videos and listening to the corresponding segments with no access to the contextual information.

3.2 Multi-emotion annotation

As described by Devillers et al. (2005), emotional manifestations depend on both the content and the speaker. In the real life scenarios, there is a common stereotype that pure emotions make people either speechless or incoherent (Wöllmer et al. 2008). Clearly, the natural and spontaneous emotional state is complex and mixed at the same time, so a single emotion label is not sufficient. To overcome the ambiguity of emotion perception, we proposed to label primary and secondary labels to one emotional segment (Devillers et al. 2006; Tao et al. 2009). Table 3 shows some frequently occurred multi-emotion labels in CHEAVD. We find that when a person is angry, disgust, surprise, worried, sad or anxious may accompany by.

3.3 Fake/suppressed emotion annotation

As we discussed in the introduction, we provide "fake/suppressed emotion" in CHEAVD. Annotators were required to label the utterances that they felt were ambiguous from the facial expressions and acoustical signals. The common occurred fake/suppressed emotions in CHEAVD are listed in Table 4. We found that when there is a suppressed emotional state, in most cases, the internal state is negative, for instance, sad, angry, but the external expressions are happy or neutral. This phenomenon is widely occurred in our daily life.

Table 3 Multi-emotion labels in CHEAVD

Main emotion	Accompanying emotions
Angry	Disgust, surprise, worried, sad, anxious, blamed
Anticipated	Happy, serious
Embarrassing	Hesitant, confused, happy
Happy	Proud, surprise
Sad	Fear, angry
Surprise	Happy, angry
Worried	Sad
Nervous	Fearful
Disgust	Sad

Table 4 Commonly occurred fake/suppressed emotions in CHEAVD

Annotated label	Internal	External expression
Happy	Sad	Happy
Happy	Angry	Happy
Happy	Embarrassing	Happy
Neutral	Blamed	Neutral
Neutral	Sad	Neutral

Annotator 4					
Segment	Speaker	Age	Emotion 1	Emotion 2	Notes
001	F001	d	neutral	angry	suppressed
002	F001	d	sarcastic		
003	F001	d	neutral		

Fig. 2 A sample of the emotion annotation in CHEAVD

3.4 Corpus overview

Figure 2 shows the label format of CHEAVD of one annotator. The first column is the index of the segment, and "F" in the second column indicates this is a female speaker. Age "d" in the third column indicates the age of the speaker is about 25–44 years old. The next two columns are the emotion labels. "Emotion 1" is the primary label and "Emotion 2" is the secondary label. Secondary emotion label is not necessary for all the segments. The last column is the "Notes", and "suppressed" emotion is labeled here, which is rare in the whole corpus.

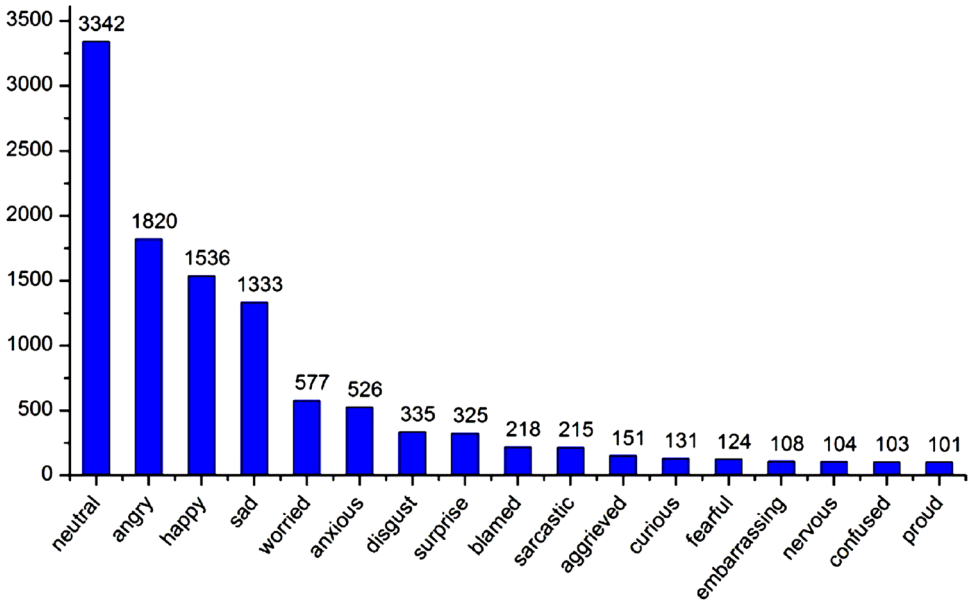

Fig. 3 The number of segments of the major emotions in CHEAVD

Table 5 The distribution of the other 9 emotions in CHEAVD

Emotion	Count	Emotion	Count
Helpless	98	Anticipated	41
Hesitant	74	Exclamation	36
Contemptuous	47	Shy	34
Frustrated	47	Guilty	19
Serious	43		

Table 6 The pairwise kappa coefficients of the four annotations

Annotators	A1	A2	A3	A4
A1		0.58	0.55	0.43
A2	0.58		0.52	0.41
A3	0.55	0.52		0.42
A4	0.43	0.41	0.42	

In total, 26 emotions were annotated in CHEAVD: neutral, angry, happy, sad, worried, anxious, disgust, surprise, blamed, sarcastic, aggrieved, curious, fearful, embarrassing, nervous, confused, proud, helpless, hesitant, contemptuous, frustrated, serious, anticipated, shy and guilty. Figure 3 and Table 5 show the final emotional labels in CHEAVD. The multiple emotion labels are counted separately.

As shown in Fig. 3 and Table 5, we can see that 17 emotions listed in Fig. 2 occur more frequently in our database and the other 9 emotions account for 2 % of the total. The largest number of utterances is neutral emotion which is consistent with our daily life. We believe this distribution of the emotions in our database reflects their frequency in the real world.

Since there are four annotators, the annotation consistency is evaluated by kappa statistic, which is shown in Table 6. All the pairwise kappa coefficients are around 0.5, showing that the four annotations are relatively consistent.

4 Automatic emotion recognition

In order to get an overall impression about the collected database, we also carried out multimodal emotion recognition experiments. LSTM-RNN is utilized in this work to model the audio and visual streams. Particularly, soft attention mechanism is employed for audio and visual streams alignment. This mechanism enables LSTM-RNN to learn to align audio and visual streams and predict the final emotion type jointly.

4.1 Feature extraction

The audio data is first resampled to 16 kHz. YAFFE (Mathieu et al. 2010) is utilized to extract audio features. All the 27 features of the default setting are extracted, including spectrum, envelope, energy, etc., and their transforms, e.g., their derivates. Finally, 939 dimensions features are extracted for each frame and the frame length is 1024. The audio features are then PCA whitened (Bengio 2012), and only 50 dimensions are kept finally.

For video features, we mainly focus on the face part. As the face shape provides important clues for facial expression, we use the landmarks' location of the face as a face shape feature. After feature normalization for each segment, these features can also reflect the head movement

and head pose. The 49 landmarks' locations are then PCA whitened, with the final 20 dimensions are kept.

For face appearance feature, we utilize the features extracted from a convolutional neural network (CNN) (LeCun et al. 1998) model. Features generated by a pre-trained CNN using face recognition data can be considered as face representation for expression recognition (Liu et al. 2014). The architecture of the CNN is the same as AlexNet and the training data is collected from CFW (Zhang et al. 2012) and Facescrub database (Ng and Winkler 2014). The deeper layers extract more abstract and robust features (Zeiler and Fergus 2014). Thus, we extract the feature from the 5th pooling layer as appearance feature. Since the dimension of features from the 5th pooling layer is 9216 in this work, which is too large for the training data of CHEAVD, Random Forest method is performed to select the most important features based on static images extracted from video data. Finally, 1024 features are kept for the appearance feature set.

4.2 Emotion recognition method

Figure 4 shows the architecture of the multimodal emotion recognition by LSTM-RNN. In the proposed architecture, the audio sequence is first learned and encoded into hidden representation from audio feature sequence by a LSTM layer. Then, the visual representation dynamics and audio visual coupling are encoded together by another LSTM layer. In this layer, soft attention mechanism is utilized to align the audio and visual streams. Window technique is applied in soft attention, in which, only a sub-sequence of audio visual features are considered in the soft attention mechanism to reduce complexity. The alignment scores are normalized by softmax. The output hidden units are then fed into the audio–visual LSTM, which learns the correlation of the audio and visual streams. At each time step, hidden representation of the audio–visual LSTM is calculated, which encodes the input features from the start to the current time step. Usually, the final emotion state could be obtained by the last hidden representation or the average of the whole sequence hidden representations. To improve the above two methods, we propose to use attention model to automatically selected the salient part of the sequence, and use these salient part to decide the final emotion state.

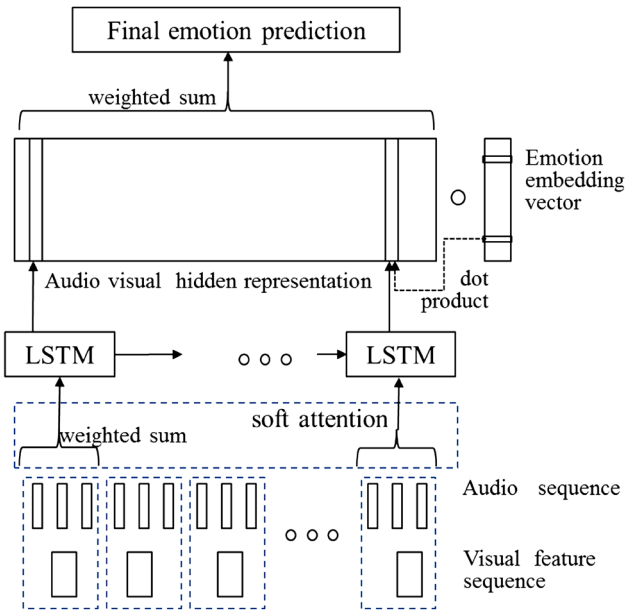

Fig. 4 Architecture of the multimodal emotion recognition by LSTM-RNN

To locate the emotional perception attentions, emotion embedding vector is introduced into the proposed model, which works as an anchor to select the salient emotional parts from the audio–visual stream. Multiplying the emotion embedded vector and the audio–visual hidden representation constitutes the attention score defined in this work. The whole audio–visual hidden representation matrix along the whole sequence can be calculated by summarizing the audio–visual hidden representation of every time step weighted by attention score. The final emotion state could be obtained by the weighted hidden representation. The detailed description of our method can be found in (Chao et al. 2016).

In the LSTM-RNN training, there are 64 memory cells utilized for both audio LSTM and audio–visual LSTM. The dimensions of all the hidden layers before LSTM layers are equal to the dimension of LSTM layers. The maximum training epoch is 50 with dropout regularization technique utilized in all layers except the LSTM layer. The drop rate is 0.5. Weight decay in all the layers with the parameter 0.0005 is applied to prevent over fitting. Early stopping technique is also employed. The best results for testing set are chosen by the best prediction accuracy in the validation set.

4.3 Emotion recognition results

The training set, validation set and testing set contain 1160, 462 and 700 segments, respectively. Since some of the emotion categories are very rare in CHEAVD, in this baseline experiment, we selected the major categories of the emotion labels, and also taking the commonly used emotion labels into consideration. Finally, *happy*, *angry*, *sad*, *surprise*, *disgust* and *neutral* are selected as the prediction targets in our experiments.

In the first experiment, we only use the primary emotion labels as the ground truth. The prediction results on training corpus, valid corpus and test corpus are listed in Table 7. Acoustic and visual features alone and multimodal emotion recognition results are all reported. Apart from LSTM-RNN, we also use Random Forest to carry out

Table 7 Emotion recognition results by LSTM-RNN

Features	Train acc	Valid acc	Test acc
Acoustic alone	0.54	0.50	0.53
Visual alone	0.57	0.56	0.57
Multimodal	0.65	0.58	0.55

Table 8 Emotion recognition results by Random Forest

Features	Train acc	Valid acc	Test acc
Acoustic alone	0.47	0.37	0.40
Visual alone	0.42	0.35	0.44
Multimodal	0.49	0.39	0.48

emotion recognition experiments on this dataset as a comparison, which is a conventional machine learning method. The number of trees is set as 100. In the multimodal emotion recognition, we use feature level fusion, which is combining the acoustic and visual features into a long feature vector, and then use Random Forest to classify the emotion labels. Table 8 shows the emotion recognition results by Random Forest as a supplementary comparison to LSTM-RNN. By Tables 7 and 8, we could see that LSTM-RNN outperforms conventional machine learning method to a large extent. In addition, multimodal usually could obtain a better performance than single modality emotion recognition. These results agree with the previous work presented in AVEC (Ringeval et al. 2015) and EmotiW challenges (Dhall et al. 2015).

Since the multi emotion label is allowed in the corpus construction, we also carried out recognition experiments to test the effect of the multi-label in emotion recognition. Label weight α is introduced to balance the effects of primary and secondary emotion labels. The final label of each segment is obtained by Eq. (1).

$$final\_label = label_1 + \alpha label_2 \quad (1)$$

Table 9 Multimodal emotion recognition results with different label weights by LSTM-RNN

Weight	Train acc	Valid acc	Test acc
0	0.65	0.58	0.55
0.7	0.68	0.58	0.58
0.6	0.74	0.61	0.56

Table 9 shows the average accuracy of multimodal emotion recognition with different label weights. We can see that when multi-label information is used in emotion recognition, the results can be improved in both training and testing.

Tables 10 and 11 show the confusion matrix of the training set and test result with $\alpha = 0.6$, because with this configuration, the performance is relatively fair to the emotions with a low share in the corpus.

From Tables 10 and 11, we can see that *happy* and *neutral* are two easy-recognized states, for their recall and precision are relatively higher than others in both training set and test set. While *surprise* and *disgust* are the worst-recognized two states in CHEAVD, especially, no segment of *disgust* is predicted correctly. One of the reasons for low accuracy is the muscular movements of these expressions are similar to each other, which is a problem in emotion production that finally become the intrinsic problem in automatic multimodal emotion recognition. The other reason might be the ambiguity of emotion perception, which leads to the inconsistence in ground truth data

Table 11 Confusion matrix on the testing set by LSTM-RNN with emotion weight = 0.6

Classified as →	Ang	Hap	Sad	Neu	Sur	Dis	% Recall
Ang	**58**	6	9	79	1	0	37.9
Hap	8	**78**	1	34	0	0	64.5
Sad	11	4	**21**	57	0	0	22.6
Neu	33	15	10	**235**	0	1	79.9
Sur	3	2	2	16	**1**	0	4.2
Dis	6	0	1	8	0	**0**	0.0
% Precision	48.7	74.3	47.7	54.8	50.0	0.0	

ang angry, *hap* happy, *sad* sad, *neu* neutral, *sur* surprise, *dis* disgust

Numbers in bold indicate the correct predicted emotions of each class

Table 10 Confusion matrix on the training set by LSTM-RNN with emotion weight = 0.6

Classified as →	Ang	Hap	Sad	Neu	Sur	Dis	% Recall
Ang	**162**	1	11	79	1	0	63.8
Hap	6	**164**	0	31	0	0	81.6
Sad	13	4	**76**	60	1	0	49.4
Neu	25	8	9	**447**	0	0	91.4
Sur	12	0	2	15	**9**	0	23.7
Dis	6	1	8	8	0	**1**	4.2
% Precision	72.3	92.1	71.7	69.8	81.8	100.0	

ang angry, *hap* happy, *sad* sad, *neu* neutral, *sur* surprise, *dis* disgust

Numbers in bold indicate the correct predicted emotions of each class

labeling. The samples with these emotional states are very few is also the reason leads to the low accuracy.

5 Conclusion

In this paper, we present the collection and annotation of the CASIA Natural Emotional Audio–Visual Database (CHEAVD) which extracted from 34 films, 2 TV series and 4 other television shows. In total, 140 min natural emotional segments extracted from 238 speakers were built. 26 non-prototypical emotional categories were finally annotated and multi-emotional labels were given for some segments. In addition, we provided several "fake/emotion" segments in our database. Such natural emotional data is of great interest to all research communities focusing on multi-modal emotion recognition, natural language understanding and speech synthesis and so on.

We also introduced our work on multimodal emotion recognition with corpus in this paper. LSTM-RNN, which is one of the state-of-art machine learning techniques is utilized to model the audio and visual streams and predict the final emotion state of each segment. We utilize the soft attention mechanism to temporally align the audio and visual streams and fuse these streams to improve the performance of multimodal emotion recognition. Multi-emotion label was used in the emotion recognition experiments, and the results show that an average accuracy of 56 % could be achieved on six major emotion states.

The corpus is chosen to utilize in the first multimodal emotion recognition challenge (Li et al. 2016), which is held by 2016 Chinese Conference on Pattern Recognition (CCPR). After the emotion recognition challenge, the corpus is free to research. Future work includes enlarging the scale of the database and investigating the suppressed emotion. Dimensional emotion annotation strategy is also an ongoing project.

Acknowledgments This work is supported by the National High-Tech Research and Development Program of China (863 Program) (No. 2015AA016305), the National Natural Science Foundation of China (NSFC) (Nos. 61305003, 61425017), the Strategic Priority Research Program of the CAS (Grant XDB02080006), and partly supported by the Major Program for the National Social Science Fund of China (13&ZD189). We thank the data providers for their kind permission to make their data for non-commercial, scientific use. Due to space limitations, providers' information is available in http://www.speakit.cn/. The corpus can be freely achieved at ChineseLDC, http://www.chineseldc.org.

References

Averill JR (1975) A semantic atlas of emotional concepts. Catalog of selected documents in psychology. American Psychological Association, Washington DC

Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate arXiv preprint arXiv:1409.0473

Barrett LF (1998) Discrete emotions or dimensions? The role of valence focus and arousal focus. Cognit Emot 12:579–599

Bengio Y (2012) Deep learning of representations for unsupervised and transfer learning. Unsuperv Transf Learn Chall Mach Learn 7:19

Burkhardt F, Paeschke A, Rolfes M, Sendlmeier WF, Weiss B (2005) A database of German emotional speech. In: INTERSPEECH, pp 1517–1520

Busso C et al (2004) Analysis of emotion recognition using facial expressions, speech and multimodal information. In: Proceedings of the 6th international conference on Multimodal interfaces. ACM, pp 205–211

Busso C et al (2008) IEMOCAP: interactive emotional dyadic motion capture database. Lang Res Eval 42:335–359

Butler EA, Lee TL, Gross JJ (2007) Emotion regulation and culture: are the social consequences of emotion suppression culture-specific? Emotion 7:30

Chao L, Tao J, Yang M, Li Y, Wen Z (2015) Long short term memory recurrent neural network based multimodal dimensional emotion recognition. In: Proceedings of the 5th international workshop on audio/visual emotion challenge. ACM, pp 65–72

Chao L, Tao J, Yang M, Li Y, Wen Z (2016) Audio visual emotion recognition with temporal alignment and perception attention. arXiv:1603.08321

Clavel C, Vasilescu I, Devillers L, Ehrette T (2004) Fiction database for emotion detection in abnormal situations. Paper presented at the international conference on spoken language processing, pp 2277–2280

Clavel C, Vasilescu I, Devillers L, Richard G, Ehrette T, Sedogbo C (2006) The SAFE corpus: illustrating extreme emotions in dynamic situations. Paper presented at the first international workshop on emotion: corpora for research on emotion and affect, pp 76–79

Cowie R, Cornelius RR (2003) Describing the emotional states that are expressed in speech. Speech Commun 40:5–32. doi:10.1016/s0167-6393(02)00071-7

Cowie R, Douglas-Cowie E, Savvidou S, Mcmahon E, Sawey M, Schröder M (2000) 'FEELTRACE': an instrument for recording perceived emotion in real time. Proc ISCA workshop on speech and emotion

Devillers L, Vidrascu L, Lamel L (2005) Challenges in real-life emotion annotation and machine learning based detection. Neural Netw 18:407–422

Devillers L, Cowie R, Martin JC, Douglas-Cowie E, Abrilian S, Mcrorie M (2006) Real life emotions in French and English TV video clips: an integrated annotation protocol combining continuous and discrete approaches. Paper presented at the international conference on language resources and evaluation, pp 1105–1110

Dhall A, Goecke R, Lucey S, Gedeon T (2012) Collecting large, richly annotated facial-expression databases from movies. IEEE MultiMedia 19:34–41

Dhall A, Ramana Murthy O, Goecke R, Joshi J, Gedeon T (2015) Video and image based emotion recognition challenges in the wild: Emotiw 2015. In: Proceedings of the 2015 ACM on international conference on multimodal interaction. ACM, pp 423–426

Douglas-Cowie E, Cowie R, Schröder M (2000) A new emotion database: considerations, sources and scope. In: ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion

Douglas-Cowie E, Campbell N, Cowie R, Roach P (2003) Emotional speech: towards a new generation of databases. Speech Commun 40:33–60. doi:10.1016/S0167-6393(02)00070-5

Douglas-Cowie E et al (2007) The HUMAINE database: addressing the collection and annotation of naturalistic and induced emotional data. In: Affective computing and intelligent interaction. Springer, pp 488–500

Ekman P (1999) Basic emotions. In: Handbook of cognition and emotion. John Wiley & Sons, Sussex, UK

El Ayadi M, Kamel MS, Karray F (2011) Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recogn 44:572–587

Elfenbein HA, Ambady N (2002) On the universality and cultural specificity of emotion recognition: a meta-analysis. Psychol Bull 128:203

Engberg IS, Hansen AV (1996) Documentation of the Danish emotional speech database (DES) vol Internal AAU report. Center for Person Kommunikation, Department of Communication Technology, Institute of Electronic Systems, Aalborg University, Denmark

Eyben F, Wollmer M, Schuller B (2009) OpenEAR—introducing the munich open-source emotion and affect recognition toolkit. Paper presented at the international conference on affective computing and Intelligent Interaction, pp 1–6

Fehr B, Russell JA (1984) Concept of Emotion Viewed From a Prototype Perspective. J Exp Psychol Gen 113:464–486

Gao Y, Zhu W (2012) How to describe speech emotion more completely-an investigation on Chinese broadcast news speech. In: 2012 8th international symposium on Chinese spoken language processing, pp 450–453

Grimm M, Kroschel K, Narayanan S (2008) The vera am mittag German audio–visual emotional speech database. Paper presented at the international conference on multimedia computing and systems/international conference on multimedia and expo, pp 865–868

Gross JJ (2002) Emotion regulation: affective, cognitive, and social consequences. Psychophysiology 39:281–291

Gross JJ, Carstensen LL, Pasupathi M, Tsai J, Skorpen CG, Hsu AY (1997) Emotion and aging: experience, expression, and control. Psychol Aging 12:590–599

He L, Jiang D, Yang L, Pei E, Wu P, Sahli H (2015) Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks. In: Proceedings of the 5th international workshop on audio/visual emotion challenge. ACM, pp 73–80

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9:1735–1780

Jaimes A, Sebe N (2007) Multimodal human–computer interaction: a survey. Comput Vis Image Underst 108:116–134

Kashdan TB, Breen WE (2008) Social anxiety and positive emotions: a prospective examination of a self-regulatory model with tendencies to suppress or express emotions as a moderating variable. Behav Ther 39:1–12

LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86:2278–2324

Li Y, Liu Y, Bao W, Chao L, Tao J (2015) From Simulated Speech to Natural Speech, What are the Robust Features for Emotion Recognition? In: 2015 International conference on affective computing and intelligent interaction (ACII), Xi'an, pp 368–373

Liu M, Wang R, Li S, Shan S, Huang Z, Chen X (2014) Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild. In: Proceedings of the 16th international conference on multimodal interaction. ACM, pp 494–501

Li Y, Tao J, Schuller B, Shan S, Jiang D, Jia J (2016) MEC 2016: The multimodal emotion recognition challenge of CCPR 2016. In: Chinese Conference on Pattern Recognition (CCPR), Chengdu, China

Mathieu B, Essid S, Fillon T, Prado J, Richard G (2010) YAAFE, an easy to use and efficient audio feature extraction software. In: ISMIR, pp 441–446

McKeown G, Valstar M, Cowie R, Pantic M, Schröder M (2012) The semaine database: annotated multimodal records of emotionally colored conversations between a person and a limited agent. IEEE Trans Affect Comput 3:5–17

Mehrabian A (1996) Pleasure-arousal-dominance: a general framework for describing and measuring individual differences in temperament. Curr Psychol 14:261–292

Mehrabian A, Russell JA (1974) An approach to environmental psychology. The MIT Press, Cambridge, Massachusetts

Mnih V, Heess N, Graves A (2014) Recurrent models of visual attention. In: Advances in neural information processing systems, pp 2204–2212

Morrison D, Wang R, De Silva LC (2007) Ensemble methods for spoken emotion recognition in call-centres. Speech Commun 49:98–112. doi:10.1016/j.specom.2006.11.004

Ng H-W, Winkler S (2014) A data-driven approach to cleaning large face datasets. In: IEEE international conference on image processing (ICIP). IEEE, pp 343–347

Plutchik R (1980) Emotion : a psychoevolutionary synthesis. Harper & Row, New York

Ringeval F et al (2015) Av + EC 2015: The first affect recognition challenge bridging across audio, video, and physiological data. In: Proceedings of the 5th international workshop on audio/visual emotion challenge. ACM, pp 3–8

Ringeval F, Sonderegger A, Sauer J, Lalanne D (2013) Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In: IEEE international conference and workshops on automatic face and gesture recognition (FG). IEEE, pp 1–8

Rose H (1991) Culture and the self: implications for cognition, emotion, and motivation. Psychol Rev 98:224–253

Russell JA (1980) A circumplex model of affect. J Personal Soc Psychol 39:1161–1178

Russell JA, Bachorowski J-A, Fernandez-Dols J-M (2003) Facial and vocal expressions of emotion. Annu Rev Psychol 54:329–349

Schröder M, Pirker H, Lamolle M (2006) First suggestions for an emotion annotation and representation language. In: International conference on language resources and evaluation. Citeseer, pp 88–92

Shaver P, Schwartz J, Kirson D, O'connor C (1987) Emotion knowledge: further exploration of a prototype approach. J Pers Soc Psychol 52:1061

Song M, Bu J, Chen C, Li N (2004) Audio–visual based emotion recognition-a new approach. In: Computer vision and pattern recognition. CVPR 2004. Proceedings of the 2004 IEEE computer society conference on 2004. IEEE, pp 1020–1025

Steidl S (2009) Automatic classification of emotion related user states in spontaneous children's speech. University of Erlangen-Nuremberg Erlangen, Germany

Tao J, Tan T (2005) Affective computing: A review. In: Tao J, Picard RW (eds) Affective computing and intelligent interaction. Springer, Berlin, pp 981–995

Tao J, Li Y, Pan S (2009) A multiple perception model on emotional speech. Paper presented at the international conference on affective computing and intelligent interaction and workshops, pp 1–6

Valstar M et al (2013) AVEC 2013: the continuous audio/visual emotion and depression recognition challenge. In: Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge. ACM, pp 3–10

Ververidis D, Kotropoulos C (2006) Emotional speech recognition: resources, features, and methods. Speech Commun 48:1162–1181

Whissel CM (1989) The dictionary of affect in language. In: Emotion: theory, research and experience, (Vol 4, The measurement of emotions). Academic Press, San Diego, CA

Wöllmer M, Eyben F, Reiter S, Schuller B, Cox C, Douglas-Cowie E, Cowie R (2008) Abandoning emotion classes—towards continuous emotion recognition with modelling of long-range dependencies. Paper presented at the INTERSPEECH, pp 597–600

Wu C-H, Lin J-C, Wei W-L (2014) Survey on audiovisual emotion recognition: databases, features, and data fusion strategies APSIPA transactions on signal and information processing 3:12

Xu X, Tao J (2003) Research on emotion classification in Chinese emotional system. The Chinese affective computing and intelligent interaction, pp 199-205

Yu F, Chang E, Xu Y-Q, Shum H-Y (2001) Emotion detection from speech to enrich multimedia content. In: Advances in multimedia information processing—PCM 2001. Springer, pp 550–557

Yuan J, Shen L, Chen F (2002) The acoustic realization of anger, fear, joy and sadness in Chinese. Paper presented at the INTERSPEECH, pp 2025–2028

Zeng Z et al (2005) Audio–visual affect recognition through multistream fused HMM for HCI. In: Computer vision and pattern recognition. CVPR 2005. IEEE Computer Society Conference on 2005. IEEE, pp 967–972

Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. In: Computer vision—ECCV 2014. Springer, pp 818–833

Zhang X, Zhang L, Wang X-J, Shum H-Y (2012) Finding celebrities in billions of web images. IEEE Trans Multimedia 14:995–1007