Contents

List of Figures

List of Tables

Chapter 1

Surrey Audio-Visual Expressed Emotion (SAVEE) Database

1.1 Introduction

The design of an automatic emotion recogniser is based on many factors, and one of important factors that can effect its performance is the emotional database used to build its models representing human emotions. Emotional behaviour databases of acted and spontaneous emotions in different modalities including audio, visual and audio-visual have been recorded for the analysis of emotions. The attributes of emotional database that effect the performance of emotion recogniser include emotion categories, number of speakers, modalities (e.g., audio, visual, and audio-visual), and quality of data.

For analysis of vocal expressions of emotions audio databases have been recorded including AIBO, Berlin Database of Emotional Speech, and Danish Emotional Speech Database [2, 3, 11]. AIBO database [2] is consisted of natural data recordings from children while interacting with robot in 11 emotion categories. Berlin Database of Emotional Speech [3] is an acted database recorded from 10 actors. The database consists of 10 German sentences recorded in anger, boredom, disgust, fear, happiness, sadness and neutral emotions. Another example is Danish Emotional Speech Database [11] which consists of recording from 4 actors. The recorded data consist of 2 words, 9 sentences and 2 passages in anger, happiness, sadness, surprise and neutral emotions. These databases have been found very useful for research on audio emotion recognition, but the main limitation is that they contain only audio.

Facial expressions databases have been recorded for analysis of facial emotional behaviour. Cohn-Kanade facial expression database [13] is popular acted database which consists of recordings from 210 adults in 6 basic emotions and AUs. MMI database is another comprehensive data set of facial behaviour [15] with acted and spontaneous expressions. It consists of 1250 videos and 600 static images in 6 basic emotions, single AU and multiple AUs. Bosphorus 3D database [17] is an acted database which consists of 3396 face scans from 81 adults. The database covers 6 basic emotions and neutral, 28 AUs, 14 head poses, and 4 occlusion. These are good quality facial expression databases but they do not contain any multimodal information.

Audio-visual emotional databases have been recorded to investigate different ways of fusing multimodal information to improve emotion recognition performance. Adult Attachment Interview database [16] is natural audio-visual database which consists of subject's interviews about their childhood experiences. It consists of recordings from 60 adults in 6 basic emotions along with embarrassment, contempt, shame, and general kinds of positive and negative emotions. Belfast Naturalistic database [7] consists of clips taken from television and realistic interviews conducted by a research team. It consists of 209 sequences from TV and 30 from interviews from total of 125 subjects. An example of acted database is Facial Motion Capture database [5] which consists of recordings from an actress in anger, happiness, sadness and neutral emotions. The actress's facial expressions were captured by attaching 102 markers to her face, and total of 612 sentences were recorded. IEMOCAP [4] has been recorded from 10 subjects (5 male, 5 female). Actors recorded three selected scripts and dialogues in hypothetical scenarios designed to elicit specific emotions. The data were recorded by attaching 53 markers on the face, 2 markers on wristbands, two markers on headbands, and one marker on each hand. It consists of 12 hours of data in anger, excited, frustration, happiness, sadness and neutral emotions.

We captured an audio-visual British English database suitable for multimodal emotion analysis. We adopted more controlled approach than Adult Attachment Interview database [16], Belfast Naturalistic database [7] and HUMAINE database [8]. Natural databases have certain limitations due to uncontrolled environment including adjustment of data capture equipment, lexical and emotional content, and acoustic and visual backgrounds. The wide distribution of natural databases is normally prevented due to copyright and privacy issues [6, 8]. We used phonetically-balanced sentences and 60 facial markers to obtain phone-level annotations and coordinates of points on actor's face. In comparison to Facial Motion Capture database [5], we aimed to increase the number of actors and affect classes to cover 6 basic emotions with even distribution. The multiple speakers data with wider range of emotions provide us the opportunity to perform speaker-independent emotion analysis. The quality assessment of data in terms of actor's expressed emotions was performed by subjective evaluation under audio, visual and audio-visual scenarios. IEMOCAP database [4] is good choice for multimodal emotion analysis although some basic emotions are missing (e.g. disgust, fear and surprise), but it was not available at the outset of present work. The following sections present corpus design, data capture, data processing and annotation, subjective quality evaluation, database dissemination and summary.

1.2 Corpus design

1.2.1 Subject selection

We recorded an audio-visual emotional database from four native English male speakers, one of them was postgraduate student and rest were researchers at the University of Surrey. Native English speakers were selected in order to avoid variation in accent due to different culture. The details of recorded subjects are given in Table 1.1. The accent of speakers were different from each other, two subjects (JE and JK) had Southern English accent, one subject (KL) had Scottish and one subject (DC) had Welsh. Age of speakers varied from 27 to 31 years with an average age of 30 years. Speakers participated in data capture on volunteer basis. SAVEE database is suitable for both speaker-dependent and speaker-independent emotion classification. But it contains only

Table 1.1: *Details of subjects recorded for SAVEE database.*

Speaker ID	Age (years)	Sex	Accent
KL	27	Male	Scottish
JE	29	Male	English
JK	31	Male	English
DC	31	Male	Welsh

male speaker's data, and some differences may exist from females as they experience emotions more intensively than men [18]. The current size of the database is limited but it can be extended in future. Data capture, processing and annotation are time consuming tasks and due to limited time we were unable to extend it further.

1.2.2 Emotion categories

Many psychologists have described emotion in terms of discrete theories [14] based on assumption of existence of some universal basic emotions. However, there is variation in number and types of emotions described by many researchers. The most popular example of discrete emotion theory is the classification of basic emotions into anger, disgust, fear, happiness, sadness and surprise. This idea was mainly supported by cross-cultural studies conducted by Ekman [10], which discovered that humans perception of some basic facial expressions was same in different cultures. Most of research in the field of emotion recognition is influenced by discrete theory of emotion, and has focused on recognising these basic emotions [20]. We selected the widely used Ekman's 6 basic emotions including anger, disgust, fear, happiness, sadness and surprise, plus neutral for our database.

1.2.3 Text material

The text material for our database was selected from standard TIMIT database. TIMIT corpus of read speech has been designed to provide speech data for acoustic-phonetic analysis, and development and evaluation of automatic speech recognition systems.

We selected text material from TIMIT database because it comprised of phonetically-diverse sentences of various types. The selected text material was consisted of 15 sentences for each of the 7 emotion categories. Sentences were selected in such a way to cover all phonemes for each emotion class. The 15 sentences were distributed into 3 common, 2 emotion-specific and 10 generic sentences that were different for each emotion. The 3 common and $2\times6=12$ emotion-specific sentences were recorded as neutral to result 30 neutral sentences. The distribution of sentences in this way resulted 120 utterances per actor.

The 3 common (-) and 14 emotion-specific sentences are listed below.

- She had your dark suit in greasy wash water all year.
- Don’t ask me to carry an oily rag like that.
- Will you tell me why?

A Who authorised the unlimited expense account?
A Destroy every file related to my audits.

D Please take this dirty table cloth to the cleaners for me.
D The small boy put the worm on the hook.

F Call an ambulance for medical assistance.
F Tornado’s often destroy acres of farm land.

H Those musicians harmonise marvellously.
H The eastern coast is a place for pure pleasure and excitement.

Sa The prospect of cutting back spending is an unpleasant one for any governor.
Sa The diagnosis was discouraging; however, he was not overly worried.

Su The carpet cleaners shampooed our oriental rug.
Su His shoulder felt as if it were broken.

N The best way to learn is to solve extra problems.
N Calcium makes bones and teeth strong.

The full list of sentences can be found in appendix A.

1.3 Data capture

1.3.1 Design of prompts

Emotion and text prompts were designed for data capture. The purpose of emotion prompts was to give idea of emotions to actors, and text prompts were consisted of sentences to be recorded. For each emotion category a slide was created, which was consisted of three facial expression pictures and a short movie clip for that specific emotion. The number of sentences for each of 6 basic emotions were 15 and for neutral emotion were 30. The text for each emotion category was split into 3 sets, which resulted 5 sentences per set for each of 6 basic emotions and 10 sentences per set for neutral emotion.

The emotion and text prompts were divided into 3 groups, where each group consisted of emotion prompts followed by sentences for each of 7 emotion categories. The three groups were combined in a single file. The set of emotion and text prompts for the first group is shown in Figure 1.1. Here we have shown only 5 sentences for the neutral emotion. Similar sets of emotion and text prompts were created for the other two groups as well, which followed the first group. Emotion prompts were same for each group but text prompts were different. The aim of splitting data into three groups was to minimise the bias due to fatigue.

1.3.2 Audio-visual capture system

SAVEE database was captured in 3D vision laboratory at Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, UK. The data capture sessions spanned over several months from February 2008 to November 2008. The data were captured from four British male subjects including KL, JE, JK and DC, as shown in Figure 1.2.

To extract facial features, actor's frontal faces were painted with 60 markers. Markers were painted on forehead, eyebrows, cheeks, lips and jaw. Busso and Narayanan [5] divided the face into upper, middle and lower face regions to extract features from facial markers, as shown in Figure 1.3 (right). The upper region included markers above eyes in forehead and eyebrow area, and lower region contained markers below upper lip, including mouth and jaw. The middle region covered cheek area between upper and lower regions. The placement of facial marker in our work was inspired from Facial Motion Capture database [5]. In Facial Motion Capture database 102 facial markers were attached to an actress face, and we used 60 marker but all parts of the face were covered. We used less number of markers due to two reasons, first some markers may be redundant when they lie very close to each other, and second marker tracking is more reliable. The use of facial markers is very helpful to observe facial movement for different emotions. These marker can be easily tracked through sequence of images and this simplifies the extraction of facial features.

The 3dMD's 4D capture system [1] was used to capture the 2D frontal colour video and Beyerdynamic microphone signals. The data capture setup is shown in Figure 1.4. The 3dMD's system covers 180 ° face capture at a speed of 60 fps for up to 10 minutes capture cycle. For our data capture, we asked actors to look in front of one colour camera to capture 2D colour image of frontal face. The distance between subject and camera was 105 cm and between microphone and face was 20 cm. The speaker was sitting in front of camera with black background. For lighting, five ambient light lamps were used: one in front and two on left and two on right at different heights and angles. The sampling rate was 44.1 kHz for audio, and 60 fps for video.

At the start of recording session, actors practised to express different emotions and feed back was provided by two members of recording session. Emotion and text prompts were displayed on monitor in front of actors during recordings. The recording session for each actor consisted of three sub sessions. Emotion and text prompts were split into three parts and in each sub session part of the data was recorded. In one capture cycle 5 sentences were recorded for an emotion with small pause between sentences. Actors then stopped for a while to get ready for recordings of next emotion. Some parts of data were re-recorded to achieve satisfactory level of expressions. During the recording session feed back was given to actors by two members involved in recordings regarding the quality of their expressed emotions. The number of sentences recorded per actor were 120, which resulted an audio-visual database of 480 utterances.

1.4 Data processing and annotation

1.4.1 Speech labelling

We recorded 5 sentences per emotion in one capture cycle with small pause between sentences. The sequence of sentences were manually splitted into individual sentences, and were labelled at phone level for extraction of duration features, as described below.

The speech data labelling was performed in semi-automated way in two steps. First, the automatic labelling of data was performed using the Hidden Markov Model Toolkit (HTK) [19]. HTK is sophisticated toolkit used for HMM training, testing and results analysis, and has been used mainly for research in the field of speech recognition. Second, Speech Filing System software [12] was used to manually correct any errors in the automatic phone boundaries resulted from HTK.

For automatic data labelling through HTK, data from one speaker of SAVEE database were used for training which were manually labelled by using Speech Filing System software [12]. The raw speech waveforms were parametrised by 13 MFCCs and delta features. The MFCCs features were extracted by using Hamming window of 25 ms with a step size of 10 ms. First order preemphasis was applied to signal using coefficient of 0.97, and filter banks of 26 channels had been used. To train HMMs a prototype model was defined. For phone-based system, we used topology of 3-state left-right with no skips. The states were initialised with zero mean and standard deviation of 1. Monophone HMMs were first initialised and then re-estimated in the next step using training data from one speaker of SAVEE database. The phone based system was then used to find phoneme boundaries for data from three speakers of SAVEE database based on forced alignment.

As a second step, the automatic phone boundaries obtained from HTK were rechecked manually by using Speech Filing System software [12]. Speech Filing System software was used to correct any errors in phone boundaries based on listening assisted by waveform and spectrogram, as shown in Figure 1.3 (left). The final output was phone level labels and boundaries for the audio data.

1.4.2 Marker tracking

To facilitate extraction of facial features, actor's face was painted with 60 markers. After data capture, markers were manually labelled for first frame of sequence and then tracked for remaining frames using marker tracker. The marker tracker threshold pixels in the blue channel (i.e. colour of markers) and then detected blobs greater than a certain size. This provided us set of candidate markers, which were matched to markers from previous frame by looking for minimal overall displacement, i.e. we found the closest markers between frames $n-1$ and n, took that as match and then proceeded to next closest until all markers had been tried. If there was no candidate marker within certain distance then that marker was assumed to be occluded and was frozen in place until it became visible again. Marker tracker was unable to track some markers at lower lip for small part of data after they disappeared due to lip movement. Those marker were labelled manually. This stage resulted set of 2D marker coordinates for each frame of visual data.

1.5 Subjective quality evaluation

The quality of recorded data was checked in terms of expressed emotions by performing subjective evaluation of database. These tests assess the database in terms of speaker's intended emotions, and can be used to evaluate the performance of emotion recognition systems on this database. We performed three kinds of subjective evaluation for our database: audio, visual and audio-visual.

1.5.1 Assessment protocol

We selected 10 subjects to evaluate SAVEE database in terms of expressed emotions. Out of 10 subjects, 5 were native English speakers and rest of them had been living in UK for more than a year. All evaluators were students at University of Surrey. It has been reported in some studies that women experience, express and perceive emotions more intensively than men [18], therefore to avoid gender biasing half of the selected evaluators were female.

The details of subjects are given in Table 1.2. Although each utterance of SAVEE database was evaluated by 10 subjects but some subjects evaluated a part of database which increased the number of evaluators to 20 (10 male, 10 female). Out of 10 male subjects, 5 were native English speakers and rest of them were non-native. The female subjects were consisted of 4 native English speakers and 6 non-native speakers. Ages of subjects were in the range of 21 to 31 years, with an average of 26 years for male subjects, 23 years for female subjects, and 25 years for all subjects.

The subjective evaluation was performed at utterance level in three ways: audio, visual and audio-visual. Slides were created with audio, visual and audio-visual clips of each utterance, as shown in Figure 1.5. There were 120 clips from each actor, which were divided into 10 sets with 12 clips per set. Sets were randomised to remove systematic bias from the responses of evaluators. For each evaluator, a different data set was created using Balanced Latin Square [9]. This process resulted 10 sets with different sequence of audio, visual and audio-visual clips for each actor's data.

The subjects were trained by using slides containing three facial expression pictures, a short movie clip and two audio files for each emotion, as shown in Figure 1.6. Any additional speaker-dependent training were not provided, although some actors were known to some of them. Subjects were asked to play audio, visual and audio-visual clips and select from one of 7 emotions on a response sheet, as shown in Figure 1.7. The responses were averaged over 10 subjects for each actor's data.

1.5.2 Analysis of results

The classification accuracies for audio, visual and audio-visual modalities for 7 emotion classes averaged over 4 actor's data are given in Table 1.3. Each actor's data were evaluated by 10 subjects (5 male, 5 female). Average classification accuracy for visual data was higher compared to audio, yet the overall performance improved by combining the two modalities. The results indicate that both audio and facial expressions play an important role to convey emotions.

For audio data, confusion existed between different emotions. Anger and disgust were confused with each other, and disgust was also confused with all other emotions spe-

Table 1.2: *Details of subjects who evaluated SAVEE database.*

Serial no.	**Subject ID**	**Age (years)**	**Sex**	**English speaking ability**
1	CB	23	Male	native (English)
2	DO	25	Male	native (English)
3	TS	29	Male	native (English)
4	JE	29	Male	native (English)
5	JK	31	Male	native (English)
6	MA	27	Male	lived for 1 year in UK
7	AS	21	Male	lived for 2.5 years in UK
8	AK	21	Male	lived for 2.5 years in UK
9	ZK	29	Male	lived for 1.5 years in UK
10	AA	29	Male	lived for 1 year in UK
11	SA	21	Female	native (English)
12	AH	22	Female	native (English)
13	NA	25	Female	native (English)
14	MH	23	Female	native (English)
15	NN	27	Female	lived for 3 years in UK
16	AB	24	Female	lived for 2 years in UK
17	SS	25	Female	lived for 7 years in UK
18	HN	22	Female	lived for 5 years in UK
19	BY	22	Female	lived for 7 years in UK
20	IS	22	Female	lived for 3 years in UK

cially surprise and neutral. Fear was confused with both sadness and surprise, and sadness was confused with neutral. Happiness and surprise were confused with each other, and in addition happiness was confused with anger, and surprise with fear. Results indicated higher classification accuracies for anger, sadness and neutral, and lowest for disgust emotion.

Overall classification accuracy achieved with visual data was higher compared to audio, but still confusion existed between some emotions. Disgust was confused with sadness, and fear was confused with both surprise and sadness. Surprise emotion was confused with fear. For visual data higher classification accuracy was observed for anger, happiness and neutral, and lowest for fear emotion.

The classification accuracy for each emotion was improved by using audio-visual data, which resulted overall higher classification accuracy. The two modalities facilitated each other to improve the overall classification performance. Confusion was observed between some emotions. Fear emotion was confused with both sadness and surprise. Average classification accuracy was higher for anger, happiness and neutral, and was lowest for fear emotion.

Average classification accuracy achieved with audio, visual and audio-visual data for 7 emotion classes and 4 emotion classes are summarised in Tables 1.4 and 1.5 respectively. Classification accuracy for 4 emotion classes were achieved by merging some emotions based on their confusion. The four emotion categories were displeased (anger, disgust), gloomy (fear, sadness), excited (happiness, surprise), and neutral(neutral). Results indicate higher classification accuracy for facial expressions compared to audio, and overall classification performance improved for audio-visual combined. The decrease in number of classes made the classification task easier, and improvement in classification accuracy was observed. Average classification accuracy for 7 emotion classes averaged over 4 actor's data and 10 evaluators was 66.5 % for audio, 88.0 % for visual, and 91.8 % for audio-visual data. For 4 emotion classes, average classification score was 76.3 % for audio, 91.3 % for visual, and 95.2 % for audio-visual.

Overall, expressed emotions for 441 out of 480 sentences were correctly classified by at least 8 out of 10 subjects under audio-visual conditions, indicating good agreement with actor's intended emotion over the database.

1.6 Database dissemination

To contribute in the field of emotion recognition, we decided to share our multimodal database with other researchers working in the same area of research. A web-based repository is designed which contains all details about SAVEE database.

1.6.1 Design of web-based repository

A website is designed to present SAVEE database and to provide all information related to database and the data itself: home, introduction, database, evaluation, references and download.

The home section of SAVEE website is designed to provide introduction of the database. The subsections are consisted of abstract, conclusion and acknowledgements.

The next section provides an overview of the different audio, visual and audio-visual databases that have been recorded for analysis of emotions. Some popular databases have been discussed in this section.

The database section of website explains the procedure of corpus design, data capture, data processing and annotation. The corpus design subsection contains information about the selection of subjects, choosing the number and types of emotion categories, and selection of text material. The data capture subsection explains the design of emotion and text prompts, 3dMD's capture system [1] and data recording procedure. The data processing and annotation subsection explains the procedure of speech data labelling and marker tracking.

The evaluation section describes the evaluation procedure that has been adopted for quality assessment of SAVEE database. The quality evaluation of audio, visual and audio-visual data was performed by 10 subjects including native and non-native speakers. This section also includes baseline results for speaker-dependent and speaker-independent scenarios for SAVEE database. These results have been listed for researchers to perform comparative evaluation of their methods against humans and baseline results on this database.

The references section lists our publications during the course of this study and other references that has been cited in this website.

The data is available for download from downloads section. This section is further subdivided into data, annotation and meta data. The data subsections contain the audio data, marker's data and audio-visual videos. Annotation subsection contains phone-level annotation of audio data. Meta data contain details of actors, complete list of sentences, emotion and text prompts slides, an example video of markers tracking, details of evaluators, a set of audio, visual and audio-visual slides used for data evaluation, slides used for training of subjects before data evaluation, response form for evaluators, humans classification rate for each audio, visual and audio-visual data clip, full list of audio and visual extracted features, and lists of selected feature sets.

1.6.2 Data preparation and release

The data is prepared at utterance level for download and it consist of following main parts: audio files, 2D marker coordinates, phone-level annotation of audio data, audio-visual videos, and face image sequences data. Each type of data is subdivided based on actor's identity. The audio files are in standard WAV file format. The tracked 2D marker coordinates are stored in marker locations section, and the phone-level labels of audio data are stored in phonetic labelling section. The audio-visual videos are prepared in AVI format for all data at utterance level. Facial expressions data in JPEG format are also available for download.

The database is available for download from SAVEE database website, and is free of charge for research purpose.

1.7 Summary

This chapter described the design, capture, annotation, quality evaluation and dissemination of SAVEE database of expressed emotions.

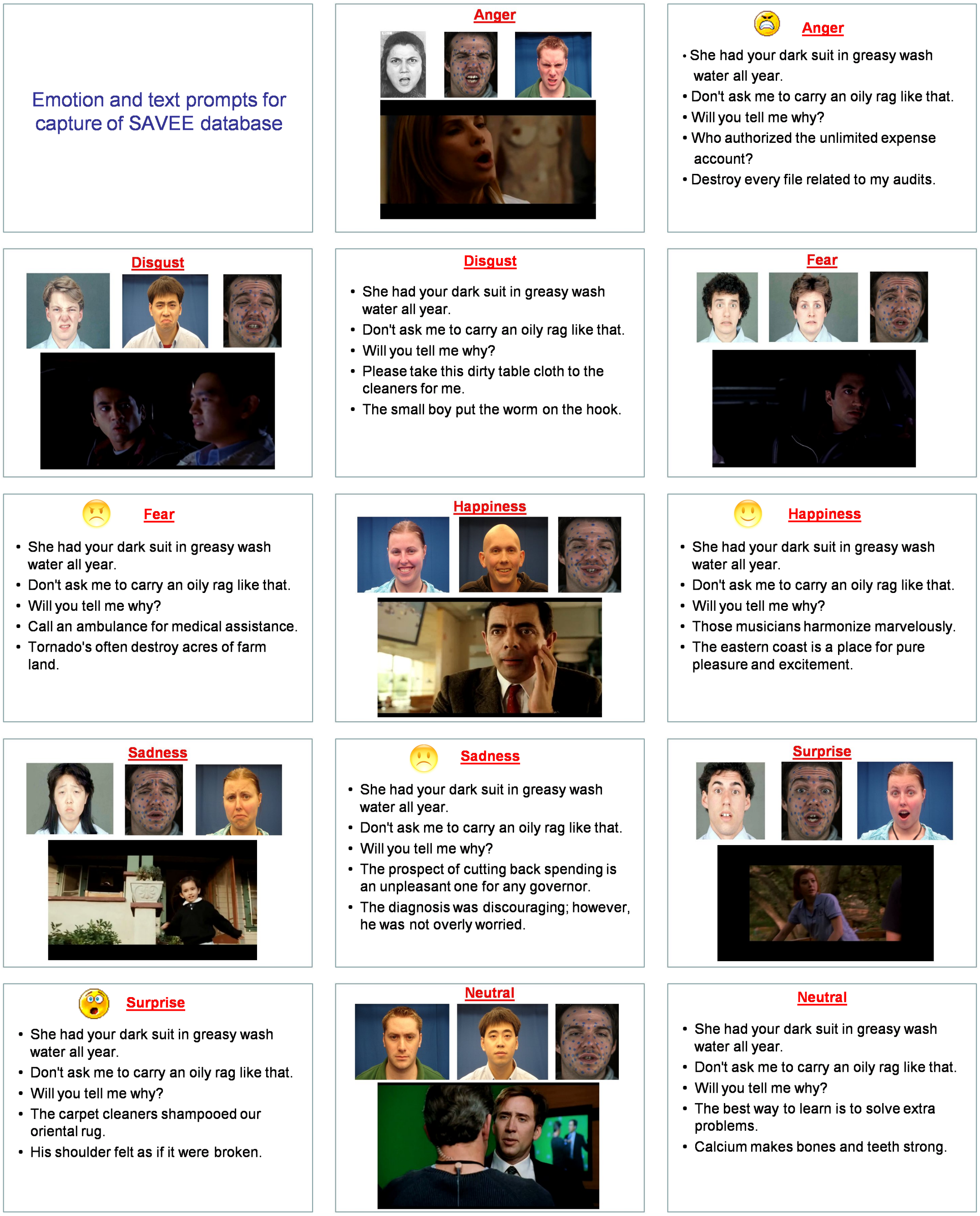
SAVEE database has been recorded from 4 native English speakers with an average age of 30 years. The emotion categories consisted of Ekman's 6 basic emotions including anger, disgust, fear, happiness, sadness and surprise, and neutral. The text material was selected from standard TIMIT database. The text material was consisted of 15 phonetically-balanced sentences per emotion, which comprised of 3 common, 2 emotion-specific and 10 generic sentences that were different for each emotion. The 3 common and 2×6=12 emotion-specific sentences were recorded as neutral which resulted 30 neutral sentences. The distribution of sentences resulted 120 utterances per actor and 480 in total.

For data capture, emotion and text prompts were created. Emotion prompt for each emotion was consisted of three facial expression pictures and a short movie clip for that specific emotion. Emotion and text prompts were divided in 3 groups, where each group consisted of all 7 emotions. The purpose was to minimise the bias due to fatigue. The data was captured at CVSSP, University of Surrey, UK over several months of year 2008. The 3dMD's capture system [1] was used to capture 2D frontal colour video and Beyerdynamic microphone signals. The sampling rate was 44.1 kHz for audio and 60 fps for video.

The speech data was labelled at phone-level in semi-automated way in two steps. First, the automatic data labelling was performed using HTK [19]. Second, Speech Filing System software [12] was used to correct any errors in the automatic phone labels based on listening assisted by waveform and spectrogram. The final output was phone level labels for the audio data. To facilitate extraction of facial features, actor's face was painted with 60 markers. After data capture, markers were manually labelled for the first frame of sequence and tracked for remaining frames using marker tracker. This stage resulted set of 2D marker coordinates for each frame of visual data.

The subjective quality evaluation of audio, visual and audio-visual data of each actor was performed by 10 subjects. Out of 10 subjects, 5 were native English speakers and rest of them had been living in UK for more than a year. To avoid gender biasing half of the evaluators were female [18]. Average classification accuracy for 7 emotion classes averaged over 4 actors was 66.5 % for audio, 88.0 % for visual, and 91.8 % for audio-visual data. Overall, expressed emotions for 441 out of 480 sentences were correctly classified by at least 8 out of 10 subjects under audio-visual conditions, indicating good agreement with actor's intended emotion over the database.

To share the database with other researchers, a web-based repository is developed. The website provides introduction to the database, background of emotional databases, details of corpus design and capture, data processing and annotation, subjective quality evaluation and data for download.



Emotion and text prompts for capture of SAVEE database

Anger

Anger

- She had your dark suit in greasy wash water all year.
- Don't ask me to carry an oily rag like that.
- Will you tell me why?
- Who authorized the unlimited expense account?
- Destroy every file related to my audits.

Disgust

Disgust

- She had your dark suit in greasy wash water all year.
- Don't ask me to carry an oily rag like that.
- Will you tell me why?
- Please take this dirty table cloth to the cleaners for me.
- The small boy put the worm on the hook.

Fear

Fear

- She had your dark suit in greasy wash water all year.
- Don't ask me to carry an oily rag like that.
- Will you tell me why?
- Call an ambulance for medical assistance.
- Tornado's often destroy acres of farm land.

Happiness

Happiness

- She had your dark suit in greasy wash water all year.
- Don't ask me to carry an oily rag like that.
- Will you tell me why?
- Those musicians harmonize marvelously.
- The eastern coast is a place for pure pleasure and excitement.

Sadness

Sadness

- She had your dark suit in greasy wash water all year.
- Don't ask me to carry an oily rag like that.
- Will you tell me why?
- The prospect of cutting back spending is an unpleasant one for any governor.
- The diagnosis was discouraging; however, he was not overly worried.

Surprise

Surprise

- She had your dark suit in greasy wash water all year.
- Don't ask me to carry an oily rag like that.
- Will you tell me why?
- The carpet cleaners shampooed our oriental rug.
- His shoulder felt as if it were broken.

Neutral

Neutral

- She had your dark suit in greasy wash water all year.
- Don't ask me to carry an oily rag like that.
- Will you tell me why?
- The best way to learn is to solve extra problems.
- Calcium makes bones and teeth strong.

Figure 1.1: *Emotion and text prompts for data capture of SAVEE database (slides start from left and move to right, and top to bottom).*

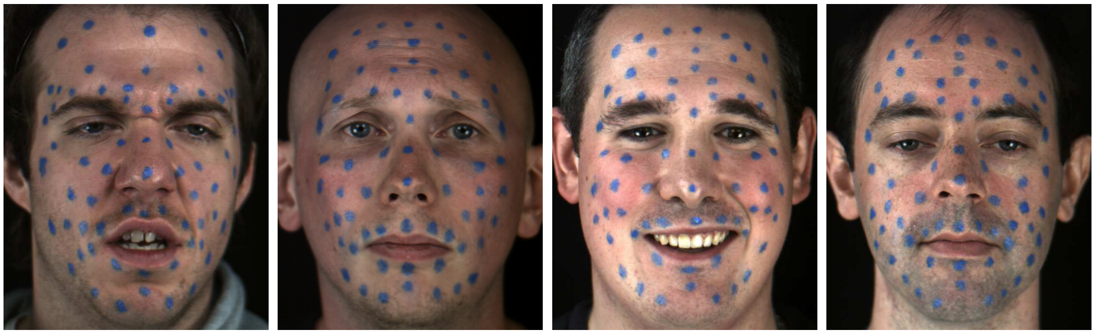


Figure 1.2: *Facial markers placed on four subjects with expressions (from left): Displeased (anger, disgust), Gloomy (fear, sadness), Excited (happiness, surprise) and Neutral (neutral).*

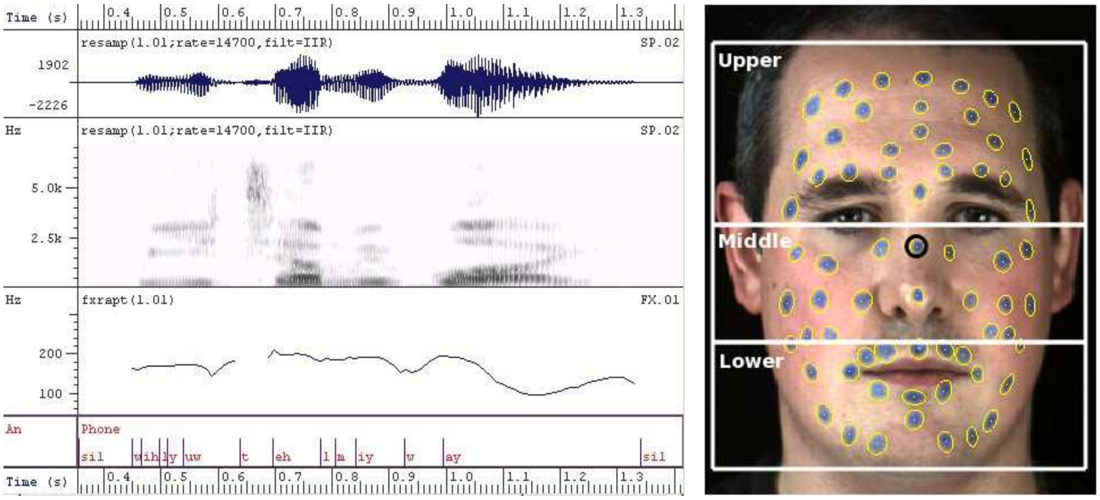


Figure 1.3: *Audio feature extraction with Speech Filing System software (left), and visual data (right) with tracked marker locations. Marker on the bridge of nose (encircled in black) was taken as a reference.*

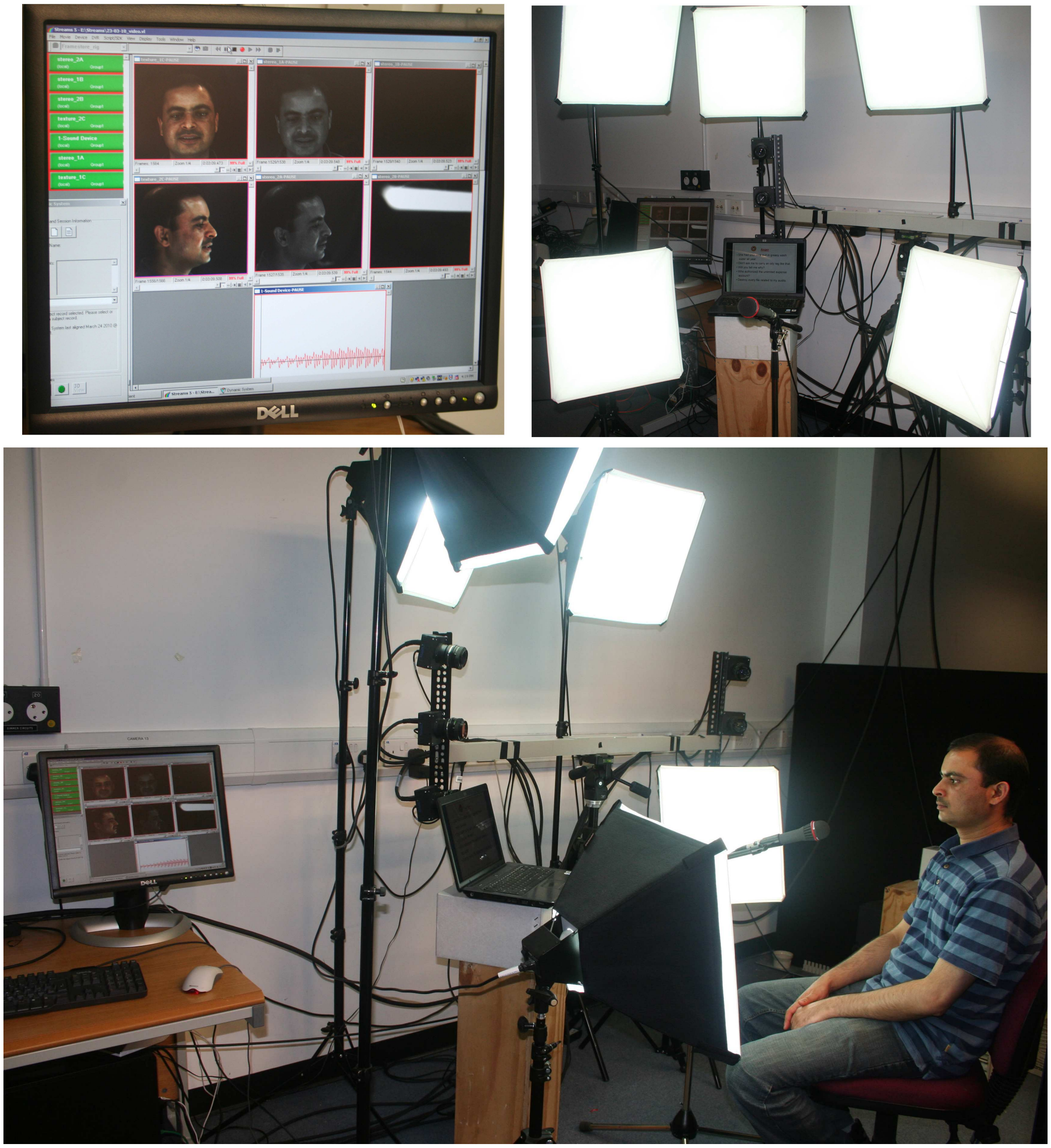


Figure 1.4: *Data capture setup for SAVEE database.*

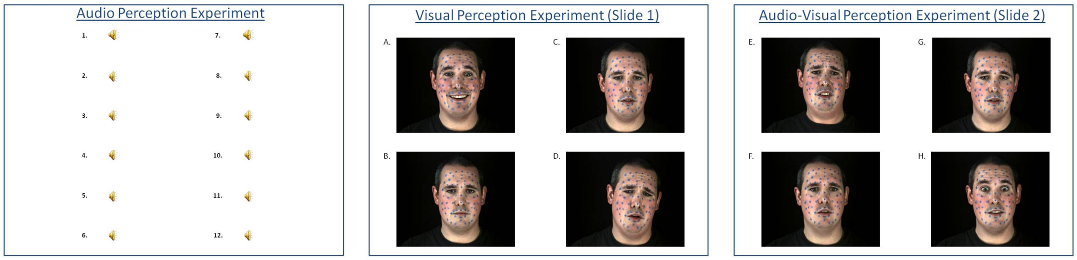

Audio Perception Experiment
1.
2.
3.
4.
5.
6.
7.
8.
9.
10.
11.
12.
Visual Perception Experiment (Slide 1)
A.
B.
C.
D.
Audio-Visual Perception Experiment (Slide 2)
E.
F.
G.
H.

Figure 1.5: *Subjective evaluation of audio, visual and audio-visual data of SAVEE database at utterance level.*

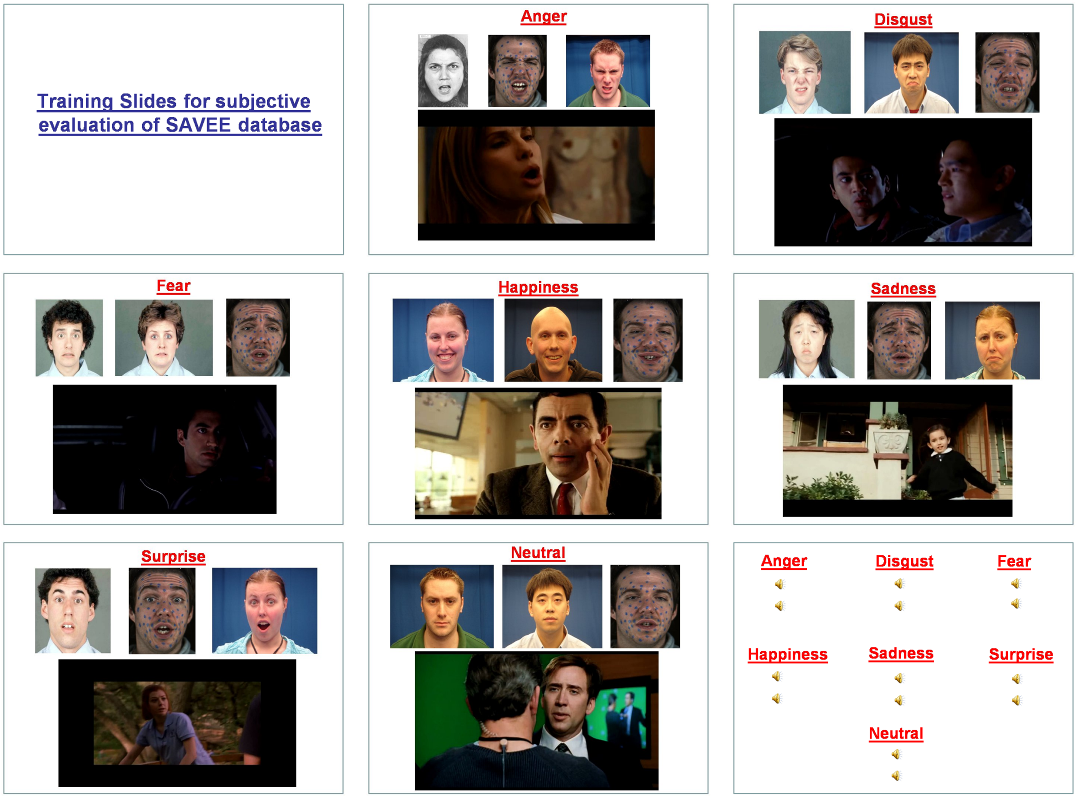
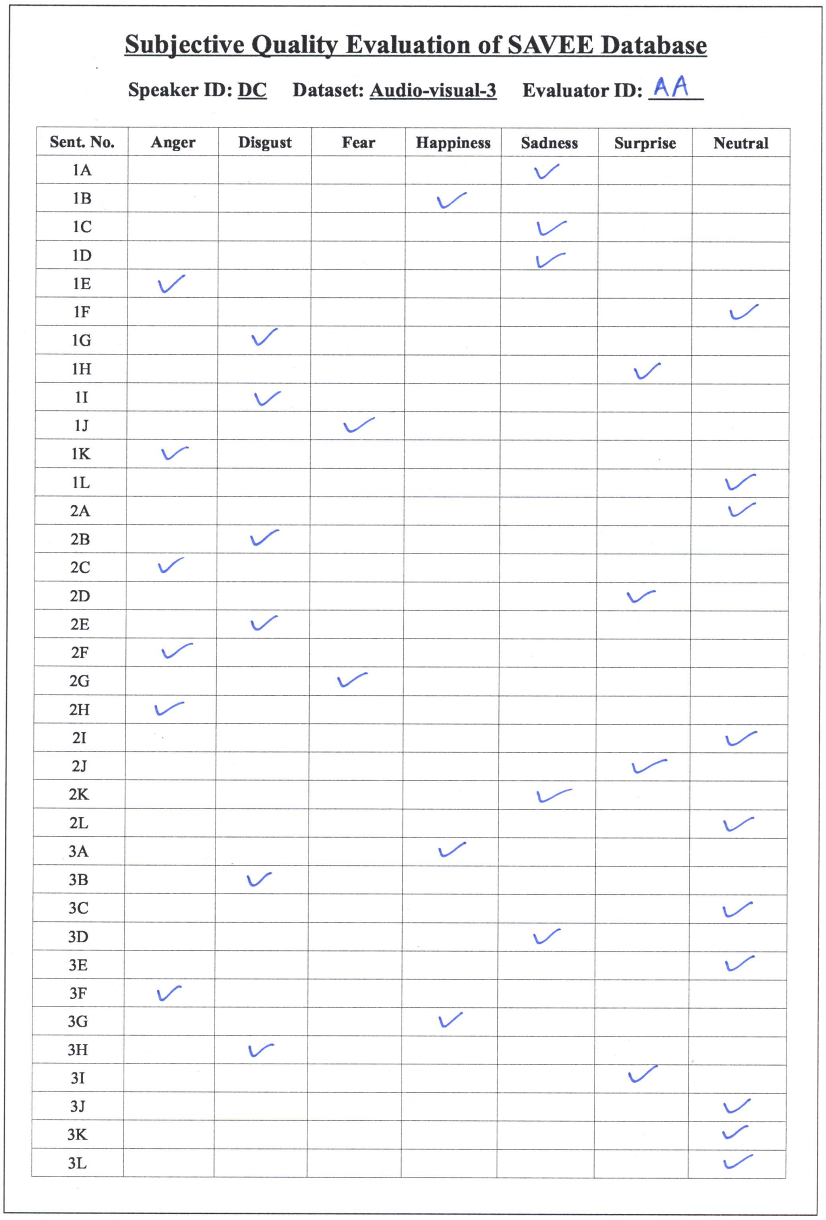


Figure 1.6: *Audio and visual emotion prompts used for training of subjects before evaluation of SAVEE database (slides start from left and move to right, and top to bottom).*



Subjective Quality Evaluation of SAVEE Database

Speaker ID: DC **Dataset:** Audio-visual-3 **Evaluator ID:** AA

Sent. No.	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Neutral
1A					✓		
1B				✓			
1C					✓		
1D					✓		
1E	✓						
1F							✓
1G		✓					
1H						✓	
1I		✓					
1J			✓				
1K	✓						
1L							✓
2A							✓
2B		✓					
2C	✓						
2D						✓	
2E		✓					
2F	✓						
2G			✓				
2H	✓						
2I							✓
2J						✓	
2K					✓		
2L							✓
3A				✓			
3B		✓					
3C							✓
3D					✓		
3E							✓
3F	✓						
3G				✓			
3H		✓					
3I						✓	
3J							✓
3K							✓
3L							✓

Figure 1.7: *Response sheet used for subjective evaluation of SAVEE database.*

Table 1.3: *Average human classification accuracy (%) for 7 emotion classes (***A***nger, ***D***isgust, ***F***ear, ***H***appiness, ***Sa***dness, ***Su***rprise, ***N***eutral): mean over 4 actors and 10 subjects with 95 % confidence interval (n=40). Confusion (%) between emotions in range of* $\geq 5 \& < 10$ *is in blue colour,* $\geq 10 \& < 15$ *is in magenta colour, and* ≥ 15 *is in red colour.*

Actual	**Recognised emotion**						
emotion	**A**	**D**	**F**	**H**	**Sa**	**Su**	**N**
Audio, average recognition rate = **66.5 ± 2.5**							
A	**78.0**	**10.2**	1.7	2.7	0.2	4.8	2.5
D	**8.3**	**38.1**	**5.7**	**6.7**	**8.8**	**11.5**	**20.8**
F	4.2	3.0	**53.8**	**6.5**	**16.3**	**14.2**	2.0
H	**10.0**	**6.3**	1.7	**57.7**	2.8	**17.0**	4.5
Sa	0.3	3.5	**11.8**	1.5	**71.2**	1.2	**10.5**
Su	**7.3**	**5.8**	**12.3**	**15.5**	3.3	**54.3**	1.2
N	1.2	1.6	0.8	0.5	**6.7**	0.2	**89.0**
Visual, average recognition rate = **88.0 ± 0.6**							
A	**94.3**	4.5	0.2	0.5	0.3	0.0	0.2
D	3.3	**80.7**	4.2	0.0	**11.7**	0.0	0.2
F	0.5	3.3	**62.0**	2.8	**8.7**	**16.2**	**6.5**
H	0.3	0.0	0.2	**97.5**	0.2	0.9	1.0
Sa	0.0	0.3	**8.0**	0.0	**90.0**	0.3	1.3
Su	0.2	0.5	**10.0**	1.3	0.0	**87.3**	0.7
N	0.7	0.0	0.0	0.7	2.3	0.1	**96.2**
Audio-visual, average recognition rate = **91.8 ± 0.1**							
A	**96.0**	3.6	0.3	0.0	0.0	0.0	0.0
D	2.6	**89.0**	2.5	0.0	**5.2**	0.3	0.3
F	0.9	0.5	**73.2**	1.4	**12.3**	**9.7**	2.2
H	0.0	0.2	0.0	**97.5**	0.0	1.6	0.7
Sa	0.2	0.3	**5.3**	0.0	**92.8**	0.0	1.3
Su	0.0	0.7	**8.0**	1.1	0.2	**89.7**	0.3
N	0.1	0.2	0.0	0.7	0.8	0.0	**98.2**

Table 1.4: *Average human classification accuracy (%) for 7 emotion classes, over 10 participants. Mean is averaged over 4 actor's data with 95 % confidence interval (CI) based on standard error (n=40).*

Human	**KL**	**JE**	**JK**	**DC**	**Mean (±CI)**
Audio	53.2	67.7	71.2	73.7	66.5 ± 2.5
Visual	89.0	89.8	88.6	84.7	88.0 ± 0.6
Audio-visual	92.1	92.1	91.3	91.7	91.8 ± 0.1

Table 1.5: *Average human classification accuracy (%) for 4 emotion classes, over 10 participants. Mean is averaged over 4 actor's data with 95 % confidence interval based on standard error (n=40).*

Human	**KL**	**JE**	**JK**	**DC**	**Mean (±CI)**
Audio	63.2	80.9	79.2	82.0	76.3 ± 2.4
Visual	90.6	97.2	90.0	87.4	91.3 ± 1.1
Audio-visual	94.4	98.3	93.5	94.5	95.2 ± 0.6

Appendix A

List of sentences for SAVEE database

Table A.1: *List of SAVEE database sentences for* **A***nger,* **D***isgust,* **F***ear,* **H***appiness,* **Sa***dness,* **Su***rprise and* **N***eutral emotion.*

Sentence number	**Emotion**	**Sentence**
1	A	She had your dark suit in greasy wash water all year.
2	A	Don't ask me to carry an oily rag like that.
3	A	Will you tell me why?
4	A	Who authorized the unlimited expense account?
5	A	Destroy every file related to my audits.
6	A	Cory and Trish played tag with beach balls for hours.
7	A	He will allow a rare lie.
8	A	Withdraw all phony accusations at once.
9	A	Right now may not be the best time for business mergers.
10	A	Kindergarten children decorate their classrooms for all holidays.
11	A	A few years later the dome fell in.
12	A	But in this one section we welcomed auditors.
13	A	Lot of people will roam the streets in costumes and masks, and having a ball.
14	A	In many of his poems, death comes by train: a strongly evocative visual image.
15	A	Then he would realize they were really things that only he himself could think.
16	D	She had your dark suit in greasy wash water all year.
17	D	Don't ask me to carry an oily rag like that.
18	D	Will you tell me why?
19	D	Please take this dirty table cloth to the cleaners for me.
20	D	The small boy put the worm on the hook.
21	D	Basketball can be an entertaining sport.
22	D	How good is your endurance?
23	D	Barb burned paper and leaves in a big bonfire.
24	D	December and January are nice months to spend in Miami.
25	D	If people were more generous, there would be no need for welfare.
26	D	If the farm is rented, the rent must be paid.
27	D	Laboratory astrophysics.
28	D	Pretty soon a woman came along carrying a folded umbrella as a walking stick.
29	D	How much and how many profits could a majority take out of the losses of a few?
30	D	Does society really exist as an entity over and above the agglomeration of men?

Table A.2: *List of SAVEE database sentences for* **A***nger,* **D***isgust,* **F***ear,* **H***appiness,* **Sa***dness,* **Su***rprise and* **N***eutral emotion.*

Sentence number	**Emotion**	**Sentence**
31	F	She had your dark suit in greasy wash water all year.
32	F	Don't ask me to carry an oily rag like that.
33	F	Will you tell me why?
34	F	Call an ambulance for medical assistance.
35	F	Tornado's often destroy acres of farm land.
36	F	Straw hats are out of fashion this year.
37	F	That diagram makes sense only after much study.
38	F	Special task forces rescue hostages from kidnappers.
39	F	The tooth fairy forgot to come when Roger's tooth fell out.
40	F	Will Robin wear a yellow lily?
41	F	Their props were two stepladders, a chair and a palm fan.
42	F	This is a problem that goes considerably beyond questions of salary and tenure.
43	F	The pulsing glow of a cigarette.
44	F	One looked down on a sea of leaves, a breaking wave of flower.
45	F	We will achieve a more vivid sense of what it is by realizing what it is not.
46	H	She had your dark suit in greasy wash water all year.
47	H	Don't ask me to carry an oily rag like that.
48	H	Will you tell me why?
49	H	Those musicians harmonize marvelously.
50	H	The eastern coast is a place for pure pleasure and excitement.
51	H	That noise problem grows more annoying each day.
52	H	Project development was proceeding too slowly.
53	H	The oasis was a mirage.
54	H	Are your grades higher or lower than Nancy's?
55	H	Serve the coleslaw after I add the oil.
56	H	By that, one feels that magnetic forces are as general as electrical forces.
57	H	His artistic accomplishments guaranteed him entry into any social gathering.
58	H	He would not carry a brief case.
59	H	Obviously, the bridal pair has many adjustments to make to their new situation.
60	H	Both the conditions and the complicity are documented in considerable detail.

Table A.3: *List of SAVEE database sentences for* **A***nger,* **D***isgust,* **F***ear,* **H***appiness,* **Sa***dness,* **Su***rprise and* **N***eutral emotion.*

Sentence number	**Emotion**	**Sentence**
61	Sa	She had your dark suit in greasy wash water all year.
62	Sa	Don't ask me to carry an oily rag like that.
63	Sa	Will you tell me why?
64	Sa	The prospect of cutting back spending is an unpleasant one for any governor.
65	Sa	The diagnosis was discouraging; however, he was not overly worried.
66	Sa	Before Thursday's exam, review every formula.
67	Sa	They enjoy it when I audition.
68	Sa	John cleans shellfish for a living.
69	Sa	He stole a dime from a beggar.
70	Sa	Jeff thought you argued in favor of a centrifuge purchase.
71	Sa	However, the litter remained, augmented by several dozen lunchroom suppers.
72	Sa	American newspaper reviewers like to call his plays nihilistic.
73	Sa	But the ships are very slow now, and we don't get so many sailors any more.
74	Sa	It is one of the rare public ventures here on which nearly everyone is agreed.
75	Sa	No manufacturer has taken the initiative in pointing out the costs involved.
76	Su	She had your dark suit in greasy wash water all year.
77	Su	Don't ask me to carry an oily rag like that.
78	Su	Will you tell me why?
79	Su	The carpet cleaners shampooed our oriental rug.
80	Su	His shoulder felt as if it were broken.
81	Su	The viewpoint overlooked the ocean.
82	Su	I'd ride the subway, but I haven't enough change.
83	Su	The clumsy customer spilled some expensive perfume.
84	Su	Please dig my potatoes up before frost.
85	Su	Grandmother outgrew her upbringing in petticoats.
86	Su	Salvation reconsidered.
87	Su	Properly used, the present book is an excellent instrument of enlightenment.
88	Su	Lighted windows glowed jewel-bright through the downpour.
89	Su	But this doesn't detract from its merit as an interesting, if not great, film.
90	Su	He further proposed grants of an unspecified sum for experimental hospitals.

Table A.4: *List of SAVEE database sentences for* **A***nger,* **D***isgust,* **F***ear,* **H***appiness,* **Sa***dness,* **Su***rprise and* **N***eutral emotion.*

Sentence number	**Emotion**	**Sentence**
91	N	She had your dark suit in greasy wash water all year.
92	N	Don't ask me to carry an oily rag like that.
93	N	Will you tell me why?
94	N	The best way to learn is to solve extra problems.
95	N	Calcium makes bones and teeth strong.
96	N	Catastrophic economic cutbacks neglect the poor.
97	N	Allow leeway here, but rationalize all errors.
98	N	Greg buys fresh milk each weekday morning.
99	N	Agricultural products are unevenly distributed.
100	N	The nearest synagogue may not be within walking distance.
101	N	As such, it was beyond politics and had no need of justification by a message.
102	N	He always seemed to have money in his pocket.
103	N	No return address whatsoever.
104	N	Keep your seats, boys, I just want to put some finishing touches on this thing.
105	N	He ripped down the cellophane carefully, and laid three dogs on the tin foil.
106	N	She had your dark suit in greasy wash water all year.
107	N	Don't ask me to carry an oily rag like that.
108	N	Will you tell me why?
109	N	Who authorized the unlimited expense account?
110	N	Destroy every file related to my audits.
111	N	Please take this dirty table cloth to the cleaners for me.
112	N	The small boy put the worm on the hook.
113	N	Call an ambulance for medical assistance.
114	N	Tornado's often destroy acres of farm land.
115	N	The carpet cleaners shampooed our oriental rug.
116	N	His shoulder felt as if it were broken.
117	N	The prospect of cutting back spending is an unpleasant one for any governor.
118	N	The diagnosis was discouraging; however, he was not overly worried.
119	N	Those musicians harmonize marvelously.
120	N	The eastern coast is a place for pure pleasure and excitement.

Bibliography

[1] 3dMD. 3dMD 4D Capture System. Online: http://www.3dmd.com, accessed on 10 September, 2010.

[2] A. Batliner, C. Hacker, S. Steidl, E. Nöth, S. D'Arcy, M. Russel, and M. Wong. "You Stupid Tin Box" - Children Interacting with the AIBO Robot: a Cross-Linguistic Emotional Speech Corpus. In *Proc. Int'l Conf. on Language Resources and Evaluation*, pages 171–174, 2004.

[3] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss. A Database of German Emotional Speech. In *Proc. Interspeech*, pages 1517–1520, 2005.

[4] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, and S.S. Narayanan. IEMOCAP: Interactive emotional dyadic motion capture database. *Journal of Language Resources and Evaluation*, 42(4):335–359, 2008.

[5] C. Busso and S.S. Narayanan. Interrelation between Speech and Facial Gestures in Emotional Utterances: A Single Subject Study. *IEEE Transactions on Audio, Speech and Language Processing*, 15(8):2331–2347, 2007.

[6] R. Cowie, E. Douglas-Cowie, and C. Cox. Beyond Emotion Archetypes: Databases for Emotion Modeling Using Neural Networks. *Neural Networks*, 18:371–388, 2005.

[7] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach. Emotional Speech: Towards a New Generation of Databases. *Speech Communication*, 40(1-2):33–60, 2003.

[8] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J. Martin, L. Devillers, S. Abrilian, A. Batliner, N. Amir, and K. Karpouzis. The HUMAINE Database: Addressing the Collection and Annotation of Naturalistic and Induced Emotional Data. In *Proc. Int'l Conf. on Affective Computing and Intelligent Interaction*, pages 488–500, 2007.

[9] A.L. Edwards. Experimental Design in Psychological Research. New York: Holt, Rinehart and Winston, 1962.

[10] P. Ekman, W.V. Friesen, M. O'Sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W.A. LeCompte, T. Pitcairn, P.E. Ricci-Bitti, K. Scherer, and M. Tomita. Universals and Cultural Differences in the Judgments of Facial Expressions of Emotion. *Journal of Personality and Social Psychology*, 53(4):712–717, 1987.

[11] I.S. Engberg and A.V. Hansen. Documentation of Danish Emotional Speech Database (DES). Center for Person Kommunikation, Dept. of Comm. Tech., Inst. of Elect. Sys., Aalborg Univ., Denmark, 1996.

[12] M. Huckvale. Speech Filing System. UCL Dept. of Phonetics & Linguistics, UK. Online: http://www.phon.ucl.ac.uk/resource/sfs/, accessed on 3 September, 2010.

[13] T. Kanade, J. Cohn, and Y. Tian. Comprehensive Database for Facial Expression Analysis. In *Proc. IEEE Int'l Conf. Face and Gesture Recognition*, pages 46–53, 2000.

[14] A. Ortony and T.J. Turner. What's Basic About Basic Emotions? *Psychological Review*, 97(3):315–331, 1990.

[15] M. Pantic, M.F. Valstar, R. Rademaker, and L. Maat. Web-Based Database for Facial Expression Analysis. In *Proc. ACM Int'l Conf. Multimedia*, pages 317–321, 2005.

[16] G.I. Roisman, J.L. Tsai, and K.S. Chiang. The Emotional Integration of Childhood Experience: Physiological, Facial Expressive, and Self-Reported Emotional

Response during the Adult Attachment Interview. *Developmental Psychology*, 40(5):776–789, 2004.

[17] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun. Bosphorus Database for 3D Face Analysis. In *Proc. First COST 2101 Workshop on Biometrics and Identity Management (BIOID 2008)*, 2008.

[18] M. Swerts and E. Krahmer. Gender-related differences in the production and perception of emotion. In *Proc. Interspeech*, pages 334–337, 2008.

[19] S. Young, G. Evermann, D. Kershaw, D. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. Hidden Markov Model Toolkit. Cambridge University Engineering Department, UK. Online: http://htk.eng.cam.ac.uk/, accessed on 3 September, 2010.

[20] Z. Zeng, M. Pantic, G.I. Roisman, and T.S. Huang. A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58, 2009.