MEAD: A Large-scale Audio-visual Dataset for Emotional Talking-face Generation

Supplementary Material

Kaisiyuan Wang[1*] [0000−0002−2120−8383] Qianyi Wu[1*] [0000−0001−8764−6178]
Linsen Song[1,3,4⋆] [0000−0003−0817−2600] Zhuoqian Yang[2] [0000−0002−5410−8282]
Wayne Wu[1⋆⋆] [0000−0002−1364−8151] Chen Qian[1] [0000−0002−8761−5563]
Ran He[3,4] [0000−0002−3807−991X] Yu Qiao[5] [0000−0002−1889−2567]
Chen Change Loy[6] [0000−0001−5345−1591]

[1] SenseTime Research
[2] Robotics Institute, Carnegie Mellon University
[3] Center for Research on Intelligent Perception and Computing, CASIA
[4] University of Chinese Academy of Sciences
[5] Shenzhen Institutes of Advanced Technology, Chinese Academy of Science
[6] Nanyang Technological University

Abstract. This document provides supplementary information which is not elaborated in our main paper due to the constraints of space: Section 1 illustrates the details about audio-visual data of our *MEAD*; Section 2 introduces the setting about actor and multi-view system. Section 3 introduces our user study result on our captured and generated audio video data. Section 4 introduces our ablation study result of different loss terms. Section 5 introduces our compound emotion experiment result. Section 6 provides our designed speech corpus for collection.

1 Datails of MEAD

Audio-visual Data. For the subset of each emotion, we survey the sentence and word duration time and draw their histogram as Fig. 1 and Fig. 2 respectively. The duration time of one sentence ranges from 1 to 7 seconds, and emotion variation doesn't obviously influence the duration distribution of sentences and words. We also demonstrate the sentence and word duration distribution of different emotional intensities as Fig. 3. For simplicity, we only demonstrate one emotion *angry* as an example in Fig. 4. Empirically, emotion intensities affect speech speed, which is verified in Fig. 4, too. From Fig. 4, we can see that the intensities of emotion visibly affect the sentence and word duration. In general, the higher the emotional intensity, the shorter the sentence and word duration.

2 Characters of Mead

Actor. The sixty speakers who made contributions to Mead are gender-balanced and from different continents covering more than 15 countries. Thus, the portraits of speakers have obvious regional characteristics, increasing the diversity of this dataset.
Multi-view System. There are seven cameras located at seven orientations to capture actor's motion. As shown in Fig. 6, five cameras were horizontally put in -60, -30, 0, 30, 60 degrees, and the other two cameras were put in 30 and -30 degree vertically. This setup provides large pose information which could benefit many other research like talking face generation in large pose.

3 Audio-video user study result

Except silent videos, we implement another user study experiment based on audio videos in emotion category classification as shown in Tab. 1. The audio-visual clips under test consists of two parts, the captured data and the generated videos by our method. As mentioned in the paper, the participants made fewer mistakes in some special pairs of emotion, such as disgust and contempt, fear and surprise, in the audio video experiments. Moreover, the accuracy in each category increases compared to the silent video experiment result.

Table 1: User study on emotion category accuracy of captured and generated audio video.

Emotion	angry	disgust	contempt	fear	happy	sad	surprise	neutral	mean
Captured	0.95	0.91	0.84	0.93	0.96	0.96	0.88	0.68	0.89
Generated	0.86	0.81	0.71	0.85	0.86	0.91	0.87	0.58	0.81

4 Ablation study result

In this section, we provide detailed ablation study results to demonstrate the effectiveness of different loss terms mentioned in the main paper. First, we remove L_{con1} and L_{rec} in turn to train the Neutral-to-Emotion Transformer, and the generated intermediate emotional images are illustrated in Fig.7. Both these two results suffer from severe information loss, and the perceptual loss L_{con1} contributes more to the texture style, while L_{rec} focuses more on the details reconstruction. The result generated with both loss terms is more satisfying in terms of both texture recovery and reconstructed details.

⋆ Equal contribution.
⋆⋆ Corresponding author (wuwenyan@sensetime.com).

Given the output of Neutral-to-Emotion Transformer, we continue to conduct experiments of different loss terms used in the Refinement Network, and the FID results are reported in Tab.2. We simply remove each loss term(*e.g.* **w/o** L_{adv}, **w/o** L_{mou}) during training the Refinement Network. As illustrated in Tab.2, the performance without L_{TV} suffer from the most loss which results from the generated checkerboard artifacts, and the FID results without L_{mou} and L_{em} obtained similar degree of reduction due to the artifacts in the mouth region and the unnatural emotion generation.

Table 2: Ablation study results of FID scores on different loss terms

Emotion	angry	disgust	contempt	fear	happy	sad	surprised	neutral	avg
w/o L_{adv}	33.72	37.70	42.66	37.65	32.11	33.19	23.81	33.76	34.31
w/o L_{mou}	36.23	42.69	44.14	39.50	33.07	34.96	25.77	37.99	36.79
w/o L_{TV}	53.49	63.50	51.15	50.98	48.78	49.10	35.19	45.33	49.69
w/o $L_{c_{em}}$	35.94	41.11	41.52	38.95	32.03	34.60	25.43	41.12	36.34
w/o $L_{c_{in}}$	35.32	40.53	43.10	34.63	32.23	32.67	24.05	37.01	34.94
full pipeline	36.14	36.99	43.02	35.06	32.81	32.64	25.97	28.06	33.84

5 Compound Emotional Talking-face.

Besides manipulating emotion category and intensity for the full face mentioned in the main paper, our method is also capable of generating talking-face video with compound emotion. We design such an experiment to examine the effectiveness on separated manipulation on the upper face and mouth region, which derives from compound emotion analysis research. In real life, people can recognize many distinct emotions, and we are able to create a new emotion by combining eight basic emotions. For instance, pleasantly surprised can be treated as a combination of happiness and surprise, and terrified can be regarded as a combination of fear and surprise. The compound emotion experiment aims to explore the scope of possible emotion which could be generated by the separated manipulation mechanism in our baseline method.

We conduct this experiment by first generating a intermediate image with one kind of emotion and then modifying the mouth area with another kind of emotion in accordance with the emotion of input audio. For example, when we set the input audio as a fearful one to drive a disgusted face generated by the neutral-to-emotion transformer, our method generates a compound emotional face that contains many frowns, squeeze and a wider mouth, as shown in the left column of Fig. 8. The right column of Fig. 8 demonstrates the face images of a sad mouth (mouth corners sag) combined with the upper face of other emotions. The generation of compound emotional proves that our method can generate face videos that never appear in the dataset, such as, a man is laughing and crying at the same time (result of driving happy face by sad audio).

6 Speech Corpus of Mead

We provide our designed speech corpus in following. As described in paper, our corpus includes several sentences expressed in each category, which can be divided into three parts, *e.g.* common, generic, and emotion-related. All the three common sentences and generic sentences are shared in eight emotions, and each category includes seven emotion-related sentences and another ten specified generic sentences. As there is no intensity difference is neutral category, thus we set ten more generic sentences for neutral compared to other emotions.

Common Sentences Read in Eight Emotions

1. She had your dark suit in greasy wash water all year
2. Don't ask me to carry an oily rag like that
3. Will you tell me why

Generic Sentences Read in Eight Emotions

1. Todd placed top priority on getting his bike fixed
2. One even gave my little dog a biscuit
3. I'll have a scoop of that exotic purple and turquoise sherbet
4. His superiors had also preached this saying it was the way for eternal honor
5. The plaintiff in school desegregation cases
6. Land based radar would help with this task
7. It was not whatever tale was told by tails
8. No the man was not drunk he wondered how he got tied up with this stranger
9. No price is too high when true love is at stake
10. The revolution now under way in materials handling makes this much easier

Angry

1. Who authorized the unlimited expense account
2. Destroy every file related to my audits
3. The cat's meow always hurts my ears
4. Why else would Danny allow others to go
5. Why do we need bigger and better bombs
6. Nuclear rockets can destroy airfields with ease
7. You're so preoccupied that you've let your faith grow dim
8. Cory and Trish played tag with beach balls for hours
9. He will allow a rare lie
10. Withdraw all phony accusations at once
11. Right now may not be the best time for business mergers
12. Kindergarten children decorate their classrooms for all holidays
13. A few years later the dome fell in
14. But in this one section we welcomed auditors
15. Lot of people will roam the streets in costumes and masks and having a ball
16. In many of his poems death comes by train a strongly evocative visual image
17. Then he would realize they were really things that only he himself could think

Disgust

1. Please take this dirty table cloth to the cleaners for me
2. The small boy put the worm on the hook
3. You're not living up to your own principles she told my discouraged people
4. Don't do Charlie's dirty dishes
5. Will Robin wear a yellow lily
6. Young children should avoid exposure to contagious diseases
7. Military personnel are expected to obey government orders
8. Basketball can be an entertaining sport
9. How good is your endurance
10. Barb burned paper and leaves in a big bonfire
11. December and January are nice months to spend in Miami
12. If people were more generous there would be no need for welfare
13. If the farm is rented the rent must be paid
14. Laboratory astrophysics
15. Pretty soon a woman came along carrying a folded umbrella as a walking stick
16. How much and how many profits could a majority take out of the losses of a few
17. Does society really exist as an entity over and above the agglomeration of men

Contempt

1. Are your grades higher or lower than Nancy's
2. This was easy for us
3. Only lawyers love millionaires
4. It's illegal to postdate a check
5. He stole a dime from a beggar
6. His failure to open the store by eight cost him his job
7. Let us differentiate a few of these ideas
8. The big dog loved to chew on the old rag doll
9. Family loyalties and cooperative work have been unbroken for generations
10. Withdraw only as much money as you need
11. The way is to rent a chauffeur driven car
12. No one material is best for all situations
13. Mosquitoes exist in warm humid climates
14. We of the liberal led world got all set for peace and rehabilitation
15. Can your insurance company aid you in reducing administrative costs
16. She sprang up and went swiftly to the bedroom
17. He ate four extra eggs for breakfast

Fear

1.Call an ambulance for medical assistance
2. Tornado's often destroy acres of farm land
3. Destroy every file related to my audits
4. Would you allow acts of violence
5. The high security prison was surrounded by barbed wire
6. His shoulder felt as if it were broken

7. The fish began to leap frantically on the surface of the small lake
8. Straw hats are out of fashion this year
9. That diagram makes sense only after much study 10. Special task forces rescue hostages from kidnappers
11. The tooth fairy forgot to come when Roger's tooth fell out
12. Will Robin wear a yellow lily
13. Their props were two stepladders a chair and a palm fan
14. This is a problem that goes considerably beyond questions of salary and tenure
15. The pulsing glow of a cigarette
16. One looked down on a sea of leaves a breaking wave of flower
17. We will achieve a more vivid sense of what it is by realizing what it is not

Happy

1. Those musicians harmonize marvelously
2. The eastern coast is a place for pure pleasure and excitement
3. Tim takes Sheila to see movies twice a week
4. They used an aggressive policeman to flag thoughtless motorists
5. When you're less fatigued things just naturally look brighter
6. By that time perhaps something better can be done
7. She found herself able to sing any role and any song which struck her fancy
8. That noise problem grows more annoying each day
9. Project development was proceeding too slowly
10. The oasis was a mirage
11. Are your grades higher or lower than Nancy's
12. Serve the coleslaw after I add the oil
13. By that one feels that magnetic forces are as general as electrical forces
14. His artistic accomplishments guaranteed him entry into any social gathering
15. He would not carry a brief case
16. Obviously the bridal pair has many adjustments to make to their new situation
17. Both the conditions and the complicity are documented in considerable detail

Sad

1. The prospect of cutting back spending is an unpleasant one for any governor
2. The diagnosis was discouraging however he was not overly worried
3. We can die too we can die like real people People never live forever
4. He didn't figure her at all and if he found out a woman it'd be bad
5. There would still be plenty of moments of regret and sadness and guilty relief
6. She drank greedily and murmured thank you as he lowered her head
7. There's no chance now of all of us getting away
8. Before Thursday's exam review every formula
9. They enjoy it when I audition

10. John cleans shellfish for a living
11. He stole a dime from a beggar
12. Jeff thought you argued in favor of a centrifuge purchase
13. However the litter remained augmented by several dozen lunchroom suppers
14. American newspaper reviewers like to call his plays nihilistic
15. But the ships are very slow now and we don't get so many sailors any more
16. It is one of the rare public ventures here on which nearly everyone is agreed
17. No manufacturer has taken the initiative in pointing out the costs involved

Surprise

1. The carpet cleaners shampooed our oriental rug
2. His shoulder felt as if it were broken
3. The patient and the surgeon are both recuperating from the lengthy operation
4. He ate four extra eggs for breakfast
5. While waiting for Chipper she crisscrossed the square many times
6. I just saw Jim near the new archeological museum
7. I took her word for it but is she really going with you
8. The viewpoint overlooked the ocean
9. I'd ride the subway but I haven't enough change
10. The clumsy customer spilled some expensive perfume
11. Please dig my potatoes up before frost
12. Grandmother outgrew her upbringing in petticoats
13. Salvation reconsidered
14. Properly used the present book is an excellent instrument of enlightenment
15. Lighted windows glowed jewel bright through the downpour
16. But this doesn't detract from its merit as an interesting if not great film
17. He further proposed grants of an unspecified sum for experimental Hospitals

Neutral

1. Bridges tunnels and ferries are the most common methods of river crossings
2. The moment of truth is the moment of crisis
3. The best way to learn is to solve extra problems
4. Thereupon followed a demonstration that tyranny knows no ideological confines
5. Calcium makes bones and teeth strong
6. Catastrophic economic cutbacks neglect the poor
7. Allow leeway here but rationalize all errors
8. Greg buys fresh milk each weekday morning
9. Agricultural products are unevenly distributed
10. The nearest synagogue may not be within walking distance
11. As such it was beyond politics and had no need of justification by a message

12. He always seemed to have money in his pocket
13. No return address whatsoever
14. Keep your seats boys I just want to put some finishing touches on this thing
15. He ripped down the cellophane carefully and laid three dogs on the tin foil
16. Who authorized the unlimited expense account
17. Destroy every file related to my audits
18. Please take this dirty table cloth to the cleaners for me
19. The small boy put the worm on the hook
20. Call an ambulance for medical assistance
21. Tornado's often destroy acres of farm land
22. The carpet cleaners shampooed our oriental rug
23. His shoulder felt as if it were broken
24. The prospect of cutting back spending is an unpleasant one for any governor
25. The diagnosis was discouraging however he was not overly worried
26. Those musicians harmonize marvelously
27. The eastern coast is a place for pure pleasure and excitement

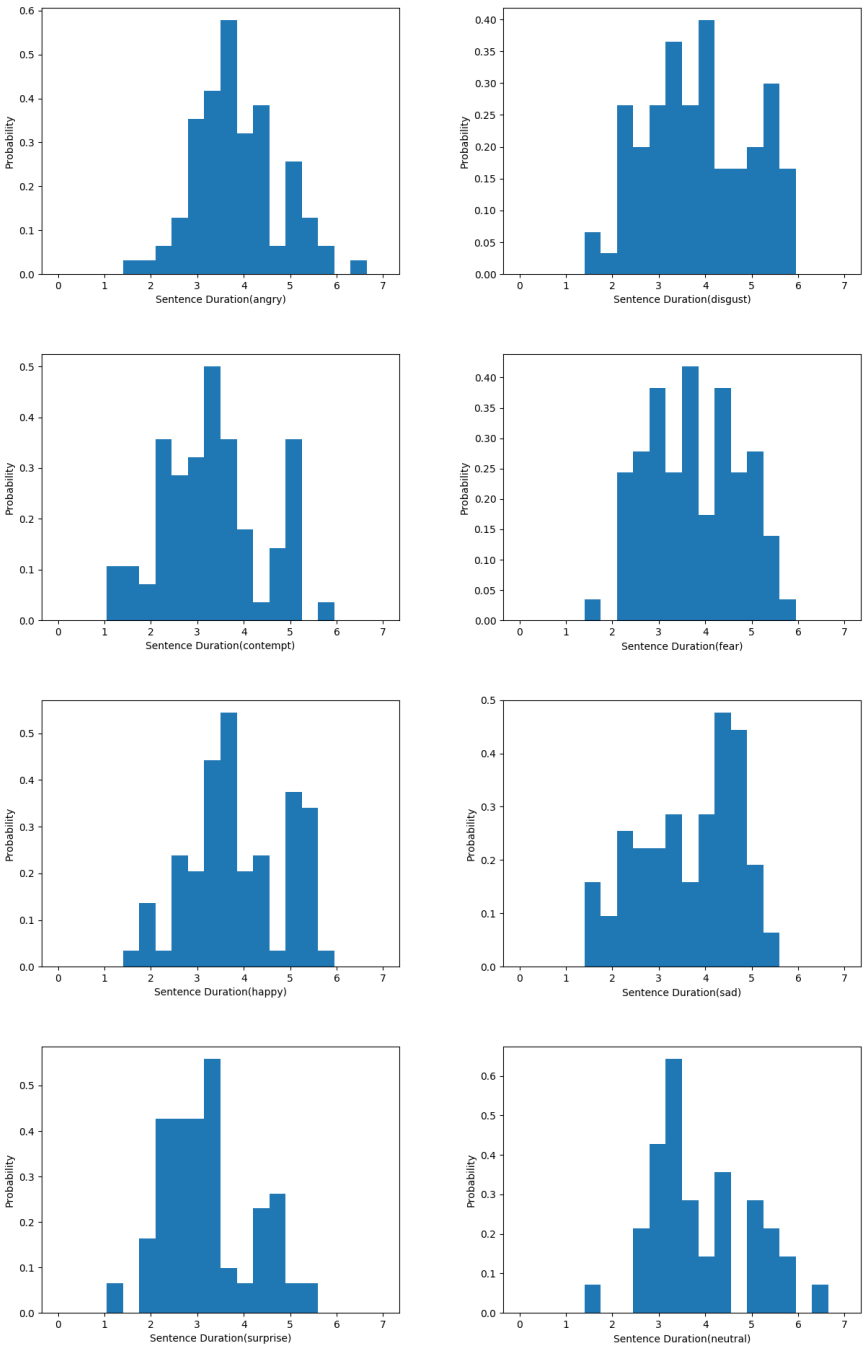

Fig. 1: The sentence duration time histogram of different emotions and all emotions. The histogram of each emotion is conducted on the data of all 3 emotional intensities.

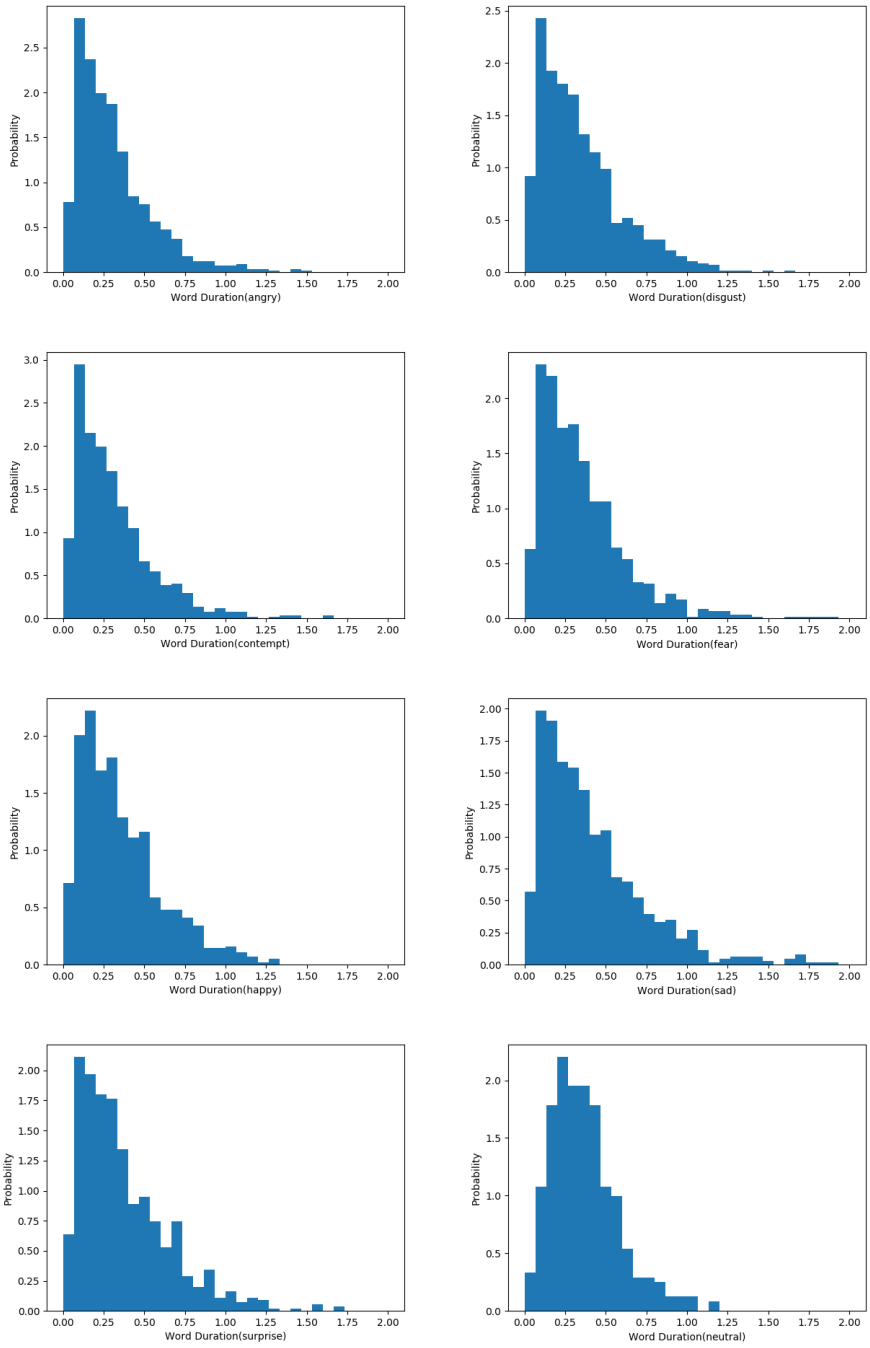

Fig. 2: The word duration time histogram of different emotions and all emotions. The histogram of each emotion is conducted on the data of all 3 emotional intensities.

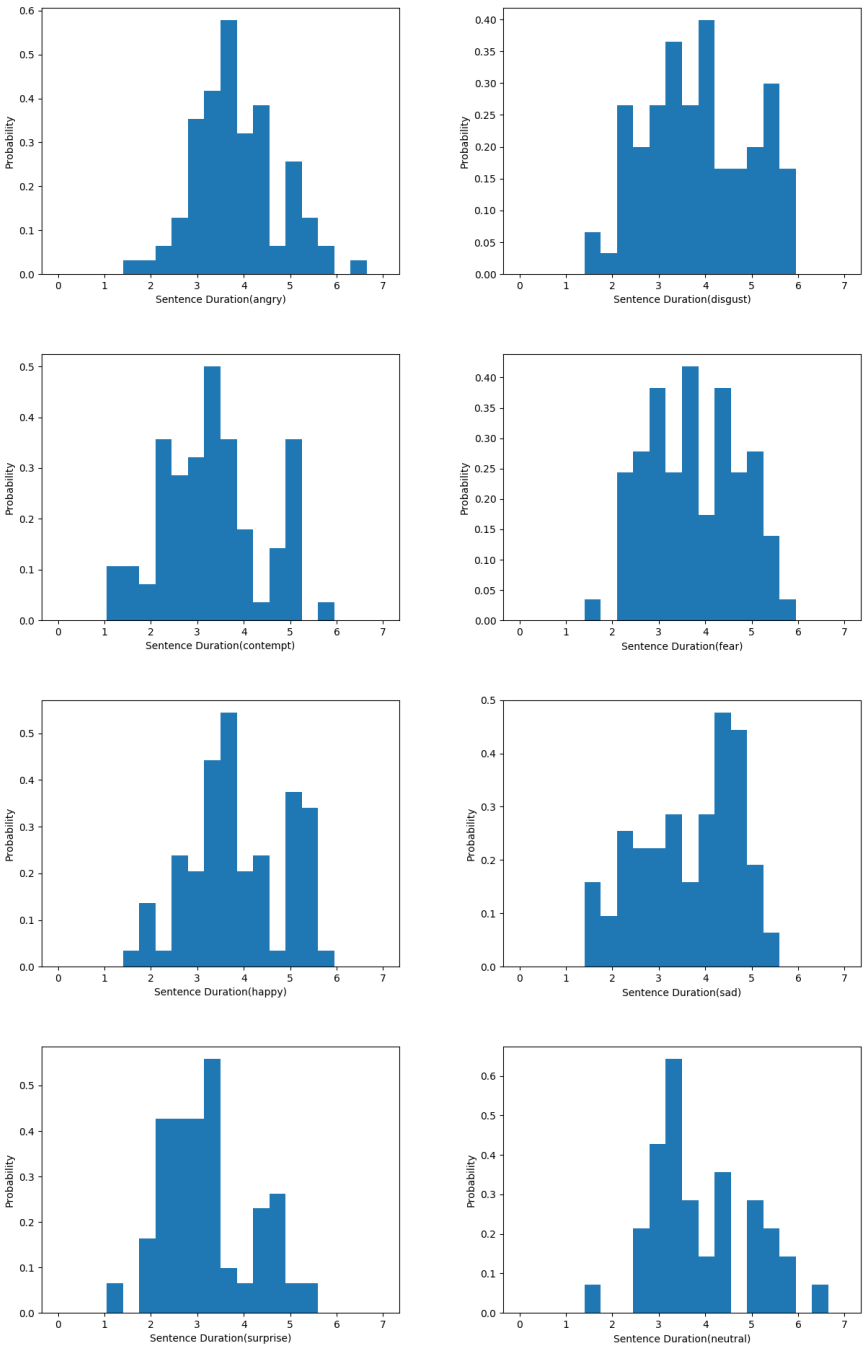

Fig. 3: The sentence duration time histogram of different emotions and all emotions. The histogram of each emotion is conducted on the data of all 3 emotional intensities.

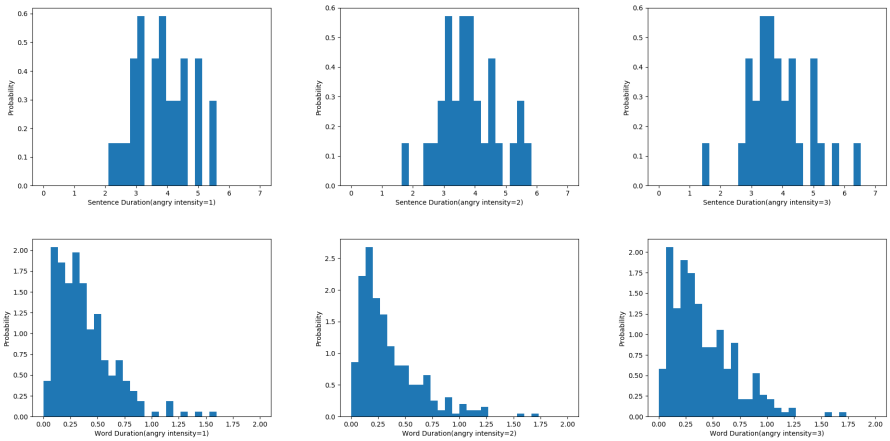

Fig. 4: The sentence and word duration time histogram of different emotional intensities, using *angry* as example.

Fig. 5: **Environment of data capture.** Our guidance team would help the actor get into better emotional state and express more accurate intensity.

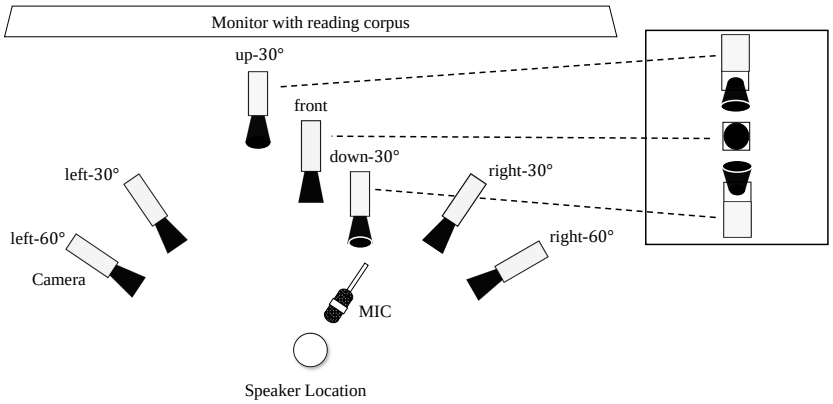

Fig. 6: **Multi-view setup of our capture system.** Five cameras were put in -60, -30, 0, 30, 60 degrees horizontally, and the other two cameras located in 30 and -30 degree vertically.

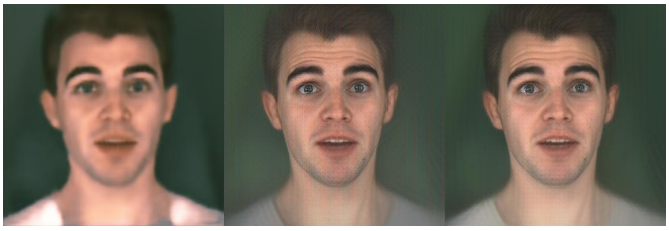

Fig. 7: Ablation study results of Neutral-to-Emotion Transformer.

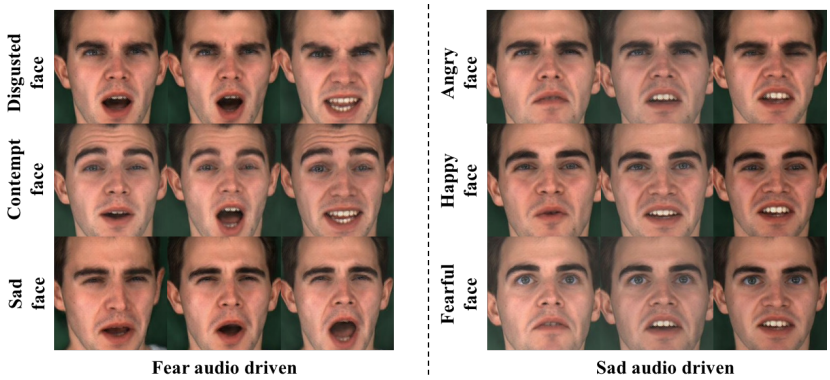

Fig. 8: **Compound emotion generation.** Compound talking-face with different emotions on the upper face and lower face respectively