MULTIMODAL EMOTION RECOGNITION WITH SURGICAL AND FABRIC MASKS

Ziqing Yang[1]*, Katherine Nayan*[2]*, Zehao Fan*[3]*, and Houwei Cao*[1]

[1]Department of Computer Science, New York Institute of Technology
[2]Department of Computing Security, Rochester Institute of Technology
[3]Department of Electrical and Computer Engineering, New York University

ABSTRACT

In this study, we investigate how different types of masks affect automatic emotion classification in different channels of audio, visual, and multimodal. We train emotion classification models for each modality with the original data without mask and the re-generated data with mask respectively, and investigate how muffled speech and occluded facial expressions change the prediction of emotions. Moreover, we conduct the contribution analysis to study how muffled speech and occluded face interplay with each other and further investigate the individual contribution of audio, visual, and audio-visual modalities to the prediction of emotion with and without mask. Finally, we investigate the cross-corpus emotion recognition across clear speech and re-generated speech with different types of masks, and discuss the robustness of speech emotion recognition.

Index Terms— multimodal emotion classification, mask, muffled speech, occluded facial expressions, cross-corpus evaluation

1. INTRODUCTION

Recently, the COVID-19 pandemic has affected more than 200 countries on all continents. Wearing face masks is one of the essential ways to control the spread of the virus and more than 50 countries require their citizens to wear face masks in public. While face masks effectively reduce the risk of infection, it has created a new normal, changing how people communicate in fundamental ways. On one hand, masks can muffle sounds, particularly the high-frequency sounds, which makes it harder to understand certain voices and speech, as well as the paralinguistic information. On the other hand, masks also block facial expressions and prevent people from seeing and reading lips, which can help people to better understand what they are hearing and the emotional state of the speaker. Several studies have been conducted recently to study the effect of face masks on hearing and speech recognition [1, 2, 3, 4]. For example, a study in [5] investigated how different types of masks affect speech recognition with different levels of background noise, and they found that different types of masks generally yield similar accuracy with low levels of background noise, but differences between masks become larger with high levels of noise. On the other hand, many recent studies investigated how masks affect human emotion perception and social judgments [6]. For example, a study in [7] investigated how emotion recognition, trust attribution and re-identification of faces differ when faces are seen without mask, with a surgical mask, and with a transparent face mask restoring visual access to the mouth region.

However, there is no existing study systematically investigating how face masks affect the automatic emotion classification in different modalities of audio, video and audio-visual. More specifically, we still do not have a full understanding of how the muffled speech and the limited visibility of facial expression degrade the emotion classification performance, as well as the interplay between the muffled audio and the occluded visual modalities. Moreover, it is of interest to know how often and for which emotion the muffled audio and the occluded visual modalities exhibit complementarity (i.e. when only the combination of the two modalities can perform the correct prediction), dominance (when the two modalities predict different emotions, and one dominating modality gives the correct prediction and matches the output from multi-modal prediction), and redundancy (when both modalities can predict the correct emotion).

In this study, we systematically investigate how different types of mask affect automatic emotion classification in different channels of audio, visual, and multimodal. We train emotion classification models for each modality with the original data and the re-generated mask data respectively, and investigate how face masks change the prediction of emotions in different modalities. Moreover, we perform the contribution analysis to study how muffled speech and occluded face interplay with each other and contribute to the emotion classification tasks. The rest of the paper presents the full details of our study. Section 2 describes the CREMA-D dataset used in our study and how we re-generate the mask data from that. Section 3 introduces different types of acoustic, facial, and multimodal features used in our study. Section 4 presents the multimodal emotion classification experiments and results, and discusses the unique contributions from different modalities. In Section 5, we study speech emotion recognition across the original clear speech and the re-generated speech with dif-

ferent masks, and discuss the robustness of speech emotion recognition, followed by the conclusion in Section 6.

2. DATASET

The dataset we use is *Crowd-sourced Emotional Multimodal Actors Dataset* (CREMA-D) [8], which is an audiovisual corpus collected to explore human emotion expression and perception behaviors in different modalities. It consists of facial and vocal emotional expressions in sentences spoken in a range of basic emotional states (Anger, Disgust, Fear, Happiness, Neutral, and Sadness). This corpus consists of 7, 442 clips (over 10 hours) of emotional sentences collected from 91 actors with diverse ethnic backgrounds. The task for the actors was to convey that they are experiencing a target emotion while uttering a given sentence.

In order to measure speech changes caused by sound absorption by the mask material, we use an artificial voice generator, a micro bluetooth speaker (Bose Sound Link micro), to re-generate the speech signal from the CREMA-D dataset with two types of mask: disposable surgical mask and fabric mask, and the re-generated mask speech was recorded using a digital sound recorder. As a result, three speech datasets are used for the speech emotion recognition experiments. Figure 1 shows the example speech waveforms and mel-spectrograms for the original speech without mask (left), the re-generated speech with surgical mask (middle), and the re-generated speech with fabric mask (right). As we can clearly see from the figure that both the surgical and fabric masks muffle speech, especially on the higher frequency bands. We use the following labels for the speech data:NoMask represents the original clear speech from the CREMA-D dataset, M_Surgical represents the re-generate speech from CREMA-D dataset with surgical mask, and M_Fabric represents re-generate speech from CREMA-D dataset with fabric mask.

To emulate the occlusion of facial expressions resulted from a mask, we only feed visual features extracted from the upper face not blocked by the mask to the prediction models, as will be detailed in the next section.

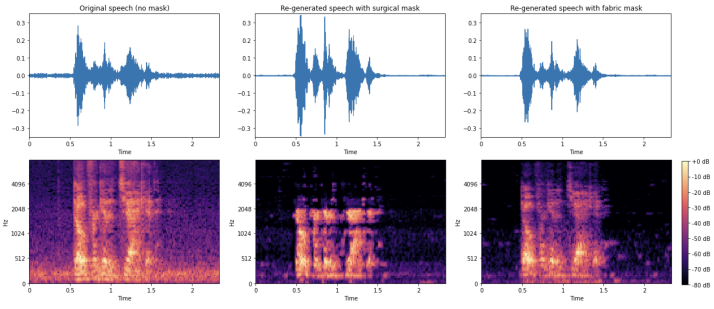

Fig. 1. Example speech waveforms and mel-spectrograms for the original speech without mask (left), the re-generated speech with surgical mask (middle), and the re-generated speech with fabric mask (right).

3. FEATURES

We investigate the state-of-the-art ComParE acoustic feature set for speech emotion recognition with and without masks. For the visual features, we investigate two sets of Bag-of-AUs features, one based on 17 AUs over the entire face area, and the other based on the AUs extracted only from the upper face not blocked by the mask. Finally, we combine the acoustic features and visual features together as the multi-modality features for multimodal emotion recognition.

3.1. Acoustic Features

We first use the openSMILE toolkit [9] to extract the ComParE acoustic features [10], which is the state-of-the-art feature set for many paralinguistic tasks including speech emotion recognition, speaker trait analysis, etc. This comprehensive set of acoustic features contains 6,373 static features resulting from the computation of functionals (statistics) over low-level descriptor (LLD) contours. It includes 130 Low Level Descriptors (LLDs), such as prosodic, spectral and voice quality features, from which we estimate the High Level Statistical Functionals (e.g., min, max, range, argmin, argmax, mean, standard deviation, skewness, kurtosis) at the utterance level.

3.2. Video Facial Features

Facial Action Units (AUs) are characterized by contractions of specific facial muscles that correspond to a displayed emotion. They have been widely used as features in facial expression analysis and emotion recognition [11, 12, 13]. In this study, we select 17 AUs that are commonly involved in the coding of the six basic emotions, and divide them into two groups. The first one only contains 8 AUs extracted from upper face without the mask blocking, and the second group includes the complete set of the 17 AUs over the entire face area. The descriptions of these 17 selected action units are listed in Table 1. Similar to the ComParE features used in acoustic analysis, we also extract the utterance-level Bag-of-AUs features. Specifically, for each video frame, we first estimate the intensity of the selected AUs by using the OpenFace facial behavior analysis toolkit [14, 15]. Those are the LLDs for video features. After that, we estimate 21 High Level Statistical Functionals (min, max, range, argmin, argmax, mean, standard deviation, three quartile values, three inter-quartile range values, skewness, kurtosis, and intercept, slope, quadratic error in linear regression, and R-square, p-value, standard error of estimated slope in linear regression) at the utterance level. Thus, for each video clip, we extract two sets of utterance-level facial action unit features: the BoAU_Mask with 168 static features and the BoAU_noMask contains 357 features.

3.3. Multimodal Features

Finally, we combine the ComParE acoustic features and the Bag-of-AUs video facial action unit features together as our

Table 1. The descriptions of the selected action units.

AUs from upper face without the mask blocking	
AU1: Inner brow raiser	AU 6: Cheek raiser
AU2: Outer brow raiser	AU 7: Lid tightener
AU 4: Brow lowerer	AU 9: Nose wrinkler
AU 5: Upper lid raiser	AU 45: Blink
AUs from lower face blocked by mask	
AU 10: Upper lip raiser	AU 20: Lip stretcher
AU 12: Lip corner puller	AU 23: Lip tightener
AU 14: Dimpler	AU 25: Lips part
AU 15: Lip corner depressor	AU 26: Jaw drop
AU 17: Chin raiser	

multimodal features. For the original data without mask, we concatenate the ComParE acoustic features extracted from the original clear speech with the BoAU_noMask features extracted from the complete set of AUs over the entire face, and the resulting Multi_NoMask features contain 6730 dimensions. For the re-generated data with mask, we concatenate the ComParE features extracted from the masked speech with the BoAU_Mask features extracted from subset of AUs on the upper face, and the corresponding Multi_Mask features contain 6541 dimensions.

4. EXPERIMENTS & RESULTS

In order to investigate how masks affect automatic emotion classification in different channels of audio, visual, and multimodal, we train emotion classification models for each modality with the original data and the re-generated mask data respectively. Following the standard setting, we use Support Vector Machine (SVM) with linear kernel for model training and classification. All the results reported in this paper are based on subject-independent 5-fold cross-validation.

4.1. Emotion Classification Results

Table 2 summarizes the classification accuracy of different types of acoustic, video facial, and multimodal features for the original (no mask) and the re-generated surgical (S) and fabric (F) mask data. We first discuss how speech emotion classification is affected by sound absorption of different mask materials. As expected, compared with the 59% unweighted averaged recall (UAR) obtained on original clear speech, the classification accuracy significantly degrades on the re-generated masked speech (47.0% on the surgical mask speech and 46% on the fabric mask speech). We also noticed that accuracy degradation varies for different emotions. For example, *fear* and *happy* are affected the most, followed by *disgust* and *neutral*, with *anger* and *sad* being the least affected emotions, where the reduced loudness on mask speech even increases the chance of predicting *sad* and slightly improves its classification accuracy. It's well known that *anger* and *sad* can be perceived most easily on the audio channel, while *fear*, *happy*, and *disgust* are more difficult to identify with acoustic in general. Our results indicate that masks further increase the difficulty to predict the emotions which are usually hard to predict by audio in normal conditions. We also noticed that the two types of masks (surgical & fabric) achieve comparable performance, which suggests that they show similar impacts on speech emotion classification and we can combine data with different masks together to build more robust models.

Table 2. The classification accuracy of different types of acoustic, facial, and multimodal features for original (no mask) and re-generated surgical (S) and fabric (F) mask data.

%	UAR	ANG	DIS	FEA	HAP	NEU	SAD
Acoustic Features							
NoMask	0.59	0.77	0.52	0.49	0.57	0.68	0.54
M_Surgical	0.47	0.71	0.43	0.29	0.30	0.52	0.57
M_Fabric	0.46	0.73	0.38	0.22	0.30	0.52	0.59
Video Facial Features							
BoAU_NoMask	0.63	0.68	0.72	0.48	0.90	0.65	0.37
BoAU_Mask	0.55	0.61	0.64	0.38	0.87	0.56	0.25
Multimodal Features							
Multi_NoMask	0.76	0.86	0.76	0.65	0.90	0.82	0.58
Multi_Mask (S)	0.66	0.77	0.66	0.36	0.85	0.69	0.54
Multi_Mask (F)	0.66	0.77	0.68	0.41	0.87	0.68	0.54

Next, we turn to discuss how emotion classification is affected by less visible cues of facial expressions due to wearing masks. Similar as what we observed in audio channel, the BoAU_NoMask features extracted from the entire face achieve 63% UAR, which substantially outperforms the 55% obtained by the BoAU_Mask features from the upper face only. However, compared with the acoustic analysis, less performance degradations are observed in general and on each individual emotions. The facial expression analysis shows the best prediction power on predicting *happy* and the worst results on *sad*, regardless of wearing mask or not.

Then we turn to discuss the multimodal emotion classification. The two types of multimodal mask models achieve similar performance, 66% UAR, which are remarkably higher than the performance of uni-modal analysis with re-generated mask data. It demonstrates that muffled speech and occluded facial expression can still provide complementary information to each other for emotion classification with mask. Moreover, compared with the no mask situation, the overall performance gap is 10%. *Fear* is the emotion being affected the most by masks, with 29% accuracy drop, and *happy* and *sad* are the two emotions that are most robust to different masks, with less than 5% performance degradation.

4.2. Multimodal Contribution Analysis

Next, in order to study the individual contribution of audio and video modalities to the prediction of emotion with and without mask, we examine the agreement among predictions based on individual modality (audio or visual) and the audio-visual prediction, and divide all the data in 6 groups: recognized by audio only; recognized by visual only; complemen-

tary (recognized by audio-visual only); dominance (when audio and visual predict different emotions, one of which gives the correct prediction and matches the prediction from multimodal); redundancy (can be recognized by both audio and visual modals); and can't be recognized by any models. We count the proportion of clips falling into each group and show the results in Figure 2.

Based on figure, we first notice that compared with emotion classification without mask, individual modality (audio or visual) shows more contributions for emotion classification with mask, and they also provide much less redundant information. We also notice that the contributions from different modality changed substantially with the muffled sound and blocked face.

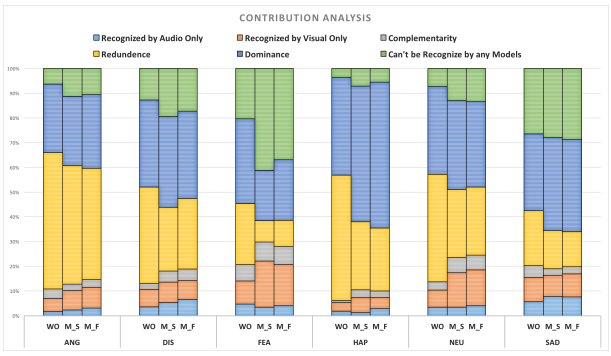

Fig. 2. Individual contribution of audio and video modalities to the prediction of corrected emotion without (WO) and with surgical (S) and fabric (F) mask.

5. CROSS-CORPUS EVALUATION

In this section we study speech emotion recognition across the original speech and the re-generated speech with different masks, and discuss the robustness of speech emotion recognition. Five SVM classifiers were trained with ComParE features on each of the following datasets respectively: NoMask, M_Surgical, M_Fabric, M_All combing two types of mask speech, and Clean+Mask combing the original clean speech and the re-generated mask speech together. The cross-corpus evaluation results are shown in Figure 3, where the results along the diagonal are from within corpus cross validation.

It's clear that the mask muffled speech is very different from the clean speech, and the performance from the cross evaluations between clean and mask datasets are very poor. The cross-evaluations across different mask datasets shows comparable results with the within-corpus evaluation, which further confirms the similarity of the two types of mask speech. We also notice that the models trained with the Clean+Mask speech together perform the best on all datasets, and achieve comparable performance with the within-corpus evaluations on both the clean and mask speech. This suggests that we may boost performance when the training data is noisy by making use of the additional reliable emotional speech from the other datasets for more precise prediction.

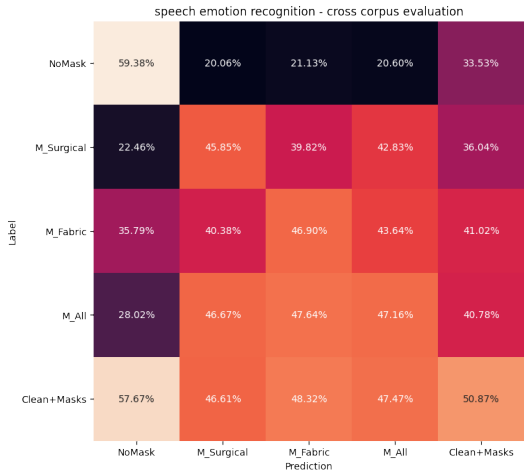

Fig. 3. Cross-corpus evaluation across NoMask, M_Surgical, M_Fabric, M_All, and Clean+Mask speech.

6. CONCLUSION

In this study, we trained emotion classification models for audio, visual, and audio-visual with the original CREMA-D dataset without mask and the re-generated mask data respectively, and investigated how muffled speech and occluded facial expressions change the prediction of emotions in different modalities. Our results suggest that different types of masks generally yield similar accuracy, and they show substantial degradations compared with the general unimodal and multimodal emotion recognition without mask. Similar to what we usually found in the no mask cases, more emotion-related information is portrayed in the mask occluded facial expressions than in the mask muffled speech, and the combined audio-visual presentation further improves the emotion recognition performance. Moreover, we perform the contribution analysis to study how muffled speech and occluded face interplay with each other and further analyze the individual contributions of audio, visual, and audio-visual modalities to the prediction of emotion with and without mask. It's interesting to observe that compared with the general cases without mask, the individual modality (audio or visual) seems more important for emotion classification with mask, and the muffled speech and occluded face also show much less redundant information with each other. Finally, we investigated the cross-corpus emotion recognition across the clear speech and the re-generated speech with different masks, and discuss the robustness of speech emotion recognition. Our results indicated that the model trained with clean and mask speech together is the most robust model against all types of speech.

7. ACKNOWLEDGMENT

This work is partially supported by the US National Science Foundation (NSF) EAGER Grant IIS-2034791 and REU Grant CNS-1852316.

8. REFERENCES

[1] Samuel R Atcherson, Lisa Lucks Mendel, Wesley J Baltimore, Chhayakanta Patro, Sungmin Lee, Monique Pousson, and M Joshua Spann, "The effect of conventional and transparent surgical masks on speech understanding in individuals with and without hearing loss," *Journal of the American Academy of Audiology*, vol. 28, no. 1, pp. 58–67, 2017.

[2] Carmen Llamas, Philip Harrison, Damien Donnelly, and Dominic Watt, "Effects of different types of face coverings on speech acoustics and intelligibility," 2009.

[3] Ryan M Corey, Uriah Jones, and Andrew C Singer, "Acoustic effects of medical, cloth, and transparent face masks on speech signals," *The Journal of the Acoustical Society of America*, vol. 148, no. 4, pp. 2371–2375, 2020.

[4] Christoph Pörschmann, Tim Lübeck, and Johannes M Arend, "Impact of face masks on voice radiation," *The Journal of the Acoustical Society of America*, vol. 148, no. 6, pp. 3663–3670, 2020.

[5] Joseph C Toscano and Cheyenne M Toscano, "Effects of face masks on speech recognition in multi-talker babble noise," *PloS one*, vol. 16, no. 2, pp. e0246842, 2021.

[6] Felix Grundmann, Kai Epstude, and Susanne Scheibe, "Face masks reduce emotion-recognition accuracy and perceived closeness," *PloS one*, vol. 16, no. 4, pp. e0249792, 2021.

[7] Marco Marini, Alessandro Ansani, Fabio Paglieri, Fausto Caruana, and Marco Viola, "The impact of facemasks on emotion recognition, trust attribution and reidentification," *Scientific Reports*, vol. 11, no. 1, pp. 1–14, 2021.

[8] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.

[9] Florian Eyben, Martin Wöllmer, and Björn Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[10] Felix Weninger, Florian Eyben, Björn W Schuller, Marcello Mortillaro, and Klaus R Scherer, "On the acoustics of emotion in audio: what speech, music, and sound have in common," *Frontiers in psychology*, vol. 4, pp. 292, 2013.

[11] Gianluca Donato, Marian Stewart Bartlett, Joseph C Hager, Paul Ekman, and Terrence J Sejnowski, "Classifying facial actions," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 21, no. 10, pp. 974, 1999.

[12] Maja Pantic and Marian Stewart Bartlett, "Machine analysis of facial expressions," in *Face recognition*. In-Tech, 2007.

[13] Arman Savran, Bulent Sankur, and M Taha Bilge, "Regression-based intensity estimation of facial action units," *Image and Vision Computing*, vol. 30, no. 10, pp. 774–784, 2012.

[14] T. Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency, "Openface 2.0: Facial behavior analysis toolkit," *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 59–66, 2018.

[15] T. Baltrusaitis, M. Mahmoud, and P. Robinson, "Cross-dataset learning and person-specific normalisation for automatic action unit detection," *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 06, pp. 1–6, 2015.