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
Emails: zyang23@nyit.edu, kyn3603@rit.edu, zf2078@nyu.edu, hcao02@nyit.edu

BACKGROUND

- The COVID-19 pandemic has affected more than 200 countries on all continents. Wearing face masks becomes a daily behavior for most people.
- While face masks effectively reduce the risk of infection, it has created a new normal, changing how people communicate in fundamental ways.
  - Muffle the high-frequency sounds
  - Block facial expressions
  - Prevent people from seeing and reading lips

MOTIVATION

- To investigate the effect of face masks on the different modalities' automatic emotion classification
- To study how the muffled speech and the limited visibility of facial expression degrade the emotion classification performance
- To study how often and for which emotion the muffled audio and the occluded visual modalities exhibit complementarity, dominance and redundancy.

CONTRIBUTIONS

- Investigate how different types of masks affect automatic emotion classification in different modalities
- Re-generate data with fabric and surgical masks for each modality
- Train emotion recognition models on both original data and re-generated mask data
- Conduct the contribution analysis to study how muffled speech and occluded face interplay with each other
- Investigate how different modalities contribute to the prediction of emotion with and without mask
- Investigate cross-corpus emotion recognition across clear and mask data

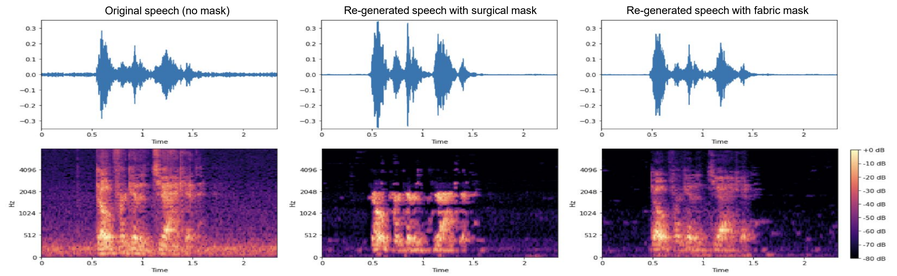
ACKNOWLEDGEMENT

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

DATASETS

Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D)

- Audiovisual corpus collected to explore human emotion expression and perception behaviors in different modalities
- Facial and vocal emotional expressions in sentences spoken in a range of basic emotional states (Anger, Disgust, Fear, Happiness, Neutral, and Sadness)
- Consists of 7, 442 clips (over 10 hours) and 91 actors with diverse ethnic background
- Re-generate the speech signal from the CREMA-D dataset with Surgical mask and Fabric mask
- Use the original CREMA-D videos but only focus the upper face



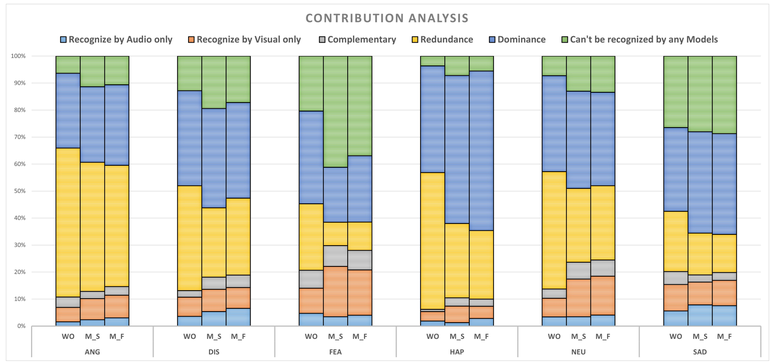
EMOTION CLASSIFICATION RESULTS

- Classification accuracy significantly degrades on the re-generated masked speech
- *Fear* and *happy* are affected the most
- *Anger* and *sad* being the least affected emotions
- BoAU NoMask outperforms BoAU Mask
- Two types of multimodal mask models achieve similar performance remarkably higher than unimodal

%	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

MULTIMODAL CONTRIBUTION ANALYSIS

- Based on figure, individual modality shows more contributions for emotion classification with mask, and much less redundant information.
- Contributions from different modality changed substantially with the muffled sound and blocked face.



FEATURES

Video Facial Features

- Select 17 AUs that are commonly involved in the coding of the six basic emotions and divide them into two groups.
- We estimate 21 High Level Statistical Functionals at the utterance level on LLDs for video features

AUs from upper face without the mask blocking	
AU 1: Inner brow raiser	AU 6: Cheek raiser
AU 2: 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	

Acoustic Features

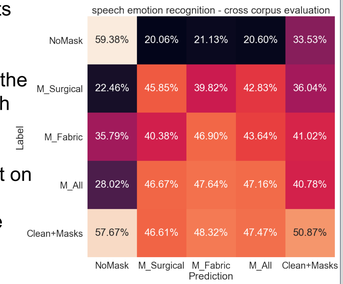
- ComParE 2016 acoustic features
- Extracted via openSMILE toolkit.
- Contains 6,373 static features resulting from the computation of functionals (statistics) over low-level descriptor (LLD) contours

Multimodal Features

- Combine the ComParE acoustic features and the Bag-of-AUs video facial action unit features together
- Features for Multimodal Analysis without masks: 6730 dimensions
- Features for Multimodal Analysis with masks: 6541 dimensions

CROSS-CORPUS EVALUATION

- Different mask datasets shows comparable results with the within-corpus evaluation
- confirms the similarity of the two types of mask speech
- Models trained with the Clean+Mask speech together perform the best on all datasets and achieve comparable performance with the within-corpus evaluations.



CONCLUSIONS

- Different types of masks yield similar accuracy, and they show substantial degradations compared with the emotion recognition without mask.
- More emotion-related information is portrayed in the mask occluded facial expressions than in the mask muffled speech.
- Combined audio-visual presentation further improves the emotion recognition performance
- Based on contribution analysis, individual modality is more important for emotion classification with mask, and much less redundant information
- Based on cross-corpus evaluation, model trained with clean and mask speech together is the most robust model against all types of speech