“Select language, modality or put on a mask!” Experiments with multimodal emotion recognition

Paweł Bujnowski[1], *Bartłomiej Kuźma*[1], *Bartłomiej Paziewski*[1], *Jacek Rutkowski*[1], *Joanna Marhula*[1], *Zuzanna Bordzicka*[1], *Piotr Andruszkiewicz*[1,2]

[1]Samsung Research Poland, Warsaw
[2]Warsaw University of Technology, Warsaw

{p.bujnowski,b.kuzma,b.paziewski,j.rutkowski2,j.marhula,
z.bordzicka,p.andruszki2}@samsung.com

Abstract

We propose a system designed for multimodal emotion recognition. Our research focuses on showing the impact of various signals in the emotion recognition process. Apart from reporting the average results of our models, we would like to encourage individual engagement of conference participants and explore how a unique emotional scene recorded on the spot can be interpreted by the models - for individual modalities as well as their combinations. Our models work for English, German and Korean. We show the comparison of emotion recognition accuracy for these 3 languages, including the influence of each modality.

Our second experiment explores emotion recognition for people wearing face masks. We show that the use of face masks affects not only the video signal but also audio and text. To our knowledge, no other study shows the effects of wearing a mask for three modalities. Unlike other studies where masks are added artificially, we use real recordings with actors in masks.

Index Terms: multimodal emotion recognition, multilingual models, masked faces

1. Introduction

Multimodal Emotion Recognition (MER) has recently been explored more extensively as many annotated video datasets have been made publicly available [1, 2]. The majority of research focuses on models that can effectively process signals of various modalities [3]. Some of them reduce computations to deal with rich input effectively, e.g. by the use sparse modules [4, 5].

Although many methods have been presented so far, there is still a lack of extensive research on the influence of video, audio and text signals on the recognition of emotions analyzed both jointly and separately. Our motivation is to understand MER by studying the real impact of single modalities and their combinations for various emotions. We would like to present our average results as well as compare them with on-the-spot recordings of emotional scenes.

As a part of our study, we also investigate the effects of adding noise to the model input. We have decided to use a natural noise in the form of partly covered face, for example, by mask-wearing or covering the head with a scarf. Mask-wearing is more common in some collectivistic countries (e.g. South Korea, Thailand, the United Arab Emirates, Mexico), but also has become more popular worldwide due to the spread of Covid-19 [6]. Even though the pandemic has receded, masks have not disappeared completely from the public space. In our experiment, we show that mask-wearing has an impact on three modalities in a standard MER system: apart from the image, the audio and text signals are also affected, which leads to lower MER accuracy. In comparison with other research on emotion recognition for masked faces [7, 8], our approach includes all 3 signals (video, audio and text) and uses real recordings of actors in masks.

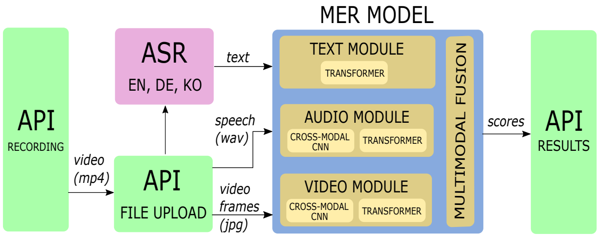


Figure 1: *System architecture.*

Discussion of our MER models can engage many conference participants who can also record emotional scenes themselves (with or without masks) or select one of sample videos for analysis. Importantly, our MER system supports not only English but also German and Korean, which increases the cultural scope of the research and the significance of our experiments. We also plan to increase the number of languages in the future.

2. System architecture

Our demonstration MER system is built of a few components: API with signal recording, automatic speech recognition (ASR), multimodal deep-learning model for emotion recognition, API showing detection results. Figure 1 shows the architecture. Our API is a tool adapted to recording people (both image and speech) in MP4 video format. It can also upload any movies in supported formats. The app selects 10 to 15 frames per second from the video and pushes them to the MER model. In parallel, the system separates the audio file using the FFMPEG library, `https://ffmpeg.org`, transforming sampling rate to 16kHz and leaving 1 channel. The audio file is sent as an input to the ASR model. Depending on the language, we use various ASR Pytorch libraries: for English - ”Shinji Watanabe/gigaspeech_asr_..._bpe5000” model from the ESPnet library [9], for German - ”speechbrain/asr-crdnn-commonvoice-de” model, `https://speechbrain.github.io`, and for Korean - the commercial ASR from `https://cloud.google.com`. As a MER model we have modified the multimodal end-to-end sparse architecture from [4] by improving threshold mechanism, changing training hyperparameters and fixing some minor errors in the code. We have also adapted a text pretrained transformer to various languages: we use BERT-multimodal uncased model for German and Korean, and Albert XLarge for English from `https://huggingface.co` library.

With a series of images, a speech WAV file and an ASR text output, the model is able to run with various multimodal signal setups - from unique modalities to all their combinations. Subsequently, in a short time ($\sim$ 1 s), the system presents radar charts for 7 emotion categories, expressing the confidence of the model for selected modality groups (see Figure 2).

3. Empirical results

3.1. Multilingual and multimodal models

The first experiment explores the impact of signal combinations in the model input on MER results. Depending on the language, we use various datasets. For English we combined CMU-MOSEI [2] and IEMOCAP [1] data to build a more complex dataset than any of them alone. For a better quality of the MOSEI corpus, we selected videos that were annotated with the same emotion by at least 2 judges. From IEMOCAP we picked emotion videos with the same labels as in MOSEI. The whole collection includes about 10k video scenes ($\sim$17h) with 7 emotions: neutral, happiness, anger, sadness, fear, disgust and surprise. For German and Korean we used an internal collection of multimodal data from open license movies that were cut into short videos and annotated by 3 judges. We received approx. 3k videos for German ($\sim$4.3h with 160 actors) and 1.2k for Korean ($\sim$2.5h with 140 actors).

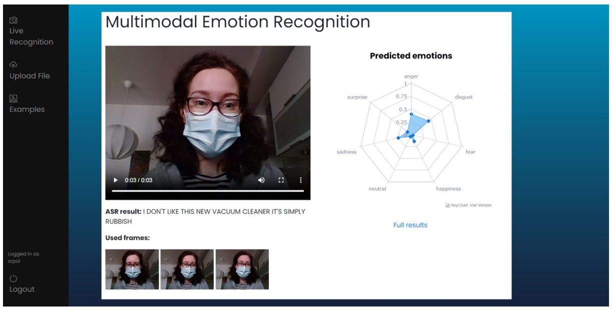
We divided data into 3 parts: train, valid and test, respecting the ratio: 70%, 15%, 15%. For each language we trained separate models with various combinations of modalities. We run the fully end-to-end model up to 30 epochs to optimize the scores. Our results are presented in the left part of Table 1. For all languages the video signal appeared the most important. For English combining video with audio gives the highest F1, whereas for German and Korean single modality video model achieves the best results. Possibly the dominance of text modality in the other studies, like [4], was caused by a low number of video frames used or too short model training.

Table 1: *Ablation study - F1 measure (%) of the MER models splitted for modality combinations. SIGN stands for signals type, EN-English, DE-German, KO-Korean; EN-NM/M - model test on new English-speaking actors without and with a mask.*

SIGN	EN	DE	KO	EN-NM	EN-M
V	66.8	**56.8**	**63.7**	27.0	16.3
A	54.6	49.6	55.8	27.4	24.0
T	62.5	46.7	56.5	43.8	40.8
VA	**68.9**	55.5	62.7	28.5	23.8
VT	65.8	49.1	60.0	**48.7**	41.2
AT	63.8	48.8	55.7	42.5	41.0
VAT	67.4	49.7	56.6	43.7	**41.7**

3.2. Experiments with people wearing masks

The second experiment focuses on the effect of mask-wearing on MER. We collected about 400 English recordings in masks and 200 without a mask from 14 actors for the 7 emotions. We evaluated models for English from Section 3.1 trained on data with uncovered faces. In this study, we used ASR text output (instead of human transcription) as a more realistic input for the AI system. We conducted 2 examinations with subsets of actors without and with a mask (see the right part of Table 1). Depending on the modality, F1 results for unmasked actors turned out to be 1.5% to 10% better than for actors wearing masks. Both the video and audio signals were affected by mask-wearing. For new data, contrary to the previous experiment, text gives the highest F1 among single modalities.



Multimodal Emotion Recognition

ASR result: I DON'T LIKE THIS NEW VACUUM CLEANER IT'S SIMPLY RUBBISH

Used frames:

Figure 2: *Application for multimodal emotion recognition*

4. Demonstration

Our tool is an interactive system that runs in a web browser (see Figure 2). The demo session is designed to explore 2 experiments. Users can record their own emotional scenes in English, German or Korean, or they can select one of the provided video samples. Additionally, the user can record a video wearing a mask and compare the results for the video with uncovered face (recorded data will not be stored). The system will show the MER results on radar charts for different signal combinations and illustrate the most important discrepancies for videos with and without masks. Users can compare their own results versus the mean and analyze typical errors for various emotion types (not printed in this article due to lack of space). We believe the demo will provide the audience with interesting feedback on MER challenges for selected languages, signal types and mask wearing.

5. References

[1] C. Busso *et al.*, “IEMOCAP: Interactive emotional dyadic motion capture database,” *Language Resources and Evaluation*, vol. 42, pp. 335–359, 12 2008.

[2] A. Zadeh *et al.*, “Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph,” in *ACL*, jul 2018, pp. 2236–2246.

[3] A. B. Zadeh *et al.*, “Multi-attention recurrent network for human communication comprehension,” in *AAAI*, 2018.

[4] W. Dai *et al.*, “Multimodal end-to-end sparse model for emotion recognition,” in *NAACL:HLT*, 2021, pp. 5305–5316.

[5] J. Cheng *et al.*, “Multimodal phased transformer for sentiment analysis,” in *EMNLP*, 2021, pp. 2447–2458.

[6] J. Lu, P. Jin, and A. English, “Collectivism predicts mask use during covid-19,” *National Academy of Sciences*, vol. 118, 05 2021.

[7] P. Ross and E. George, “Are face masks a problem for emotion recognition? Not when the whole body is visible.” *Frontiers in Neuroscience 16:915927*, 2022.

[8] Z. Yang *et al.*, “Multimodal emotion recognition with surgical and fabric masks,” in *ICASSP*, 2022, pp. 4678–4682.

[9] S. Watanabe *et al.*, “ESPnet: End-to-end speech processing toolkit,” in *INTERSPEECH*, 2018, pp. 2207–2211.