MAScreen: Augmenting Speech with Visual Cues of Lip Motions, Facial Expressions, and Text Using a Wearable Display

Hyein Lee
KAIST
Daejeon, Republic of Korea
hyein.l@kaist.ac.kr

Yoonji Kim
KAIST
Daejeon, Republic of Korea
yoonji@kaist.ac.kr

Andrea Bianchi
KAIST
Daejeon, Republic of Korea
andrea@kaist.ac.kr

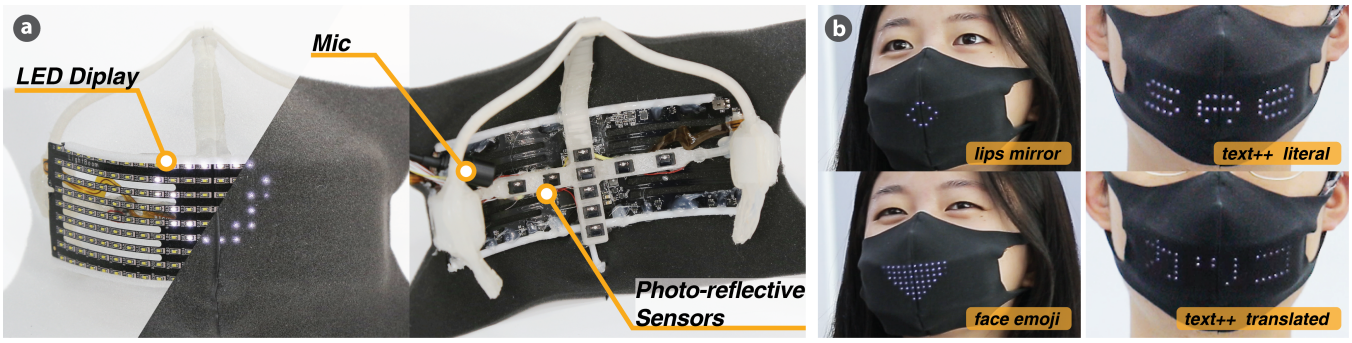

Figure 1: *MAScreen* (left) is an wearable augmentation device for improving speech intelligibility. MAScreen supports real-time lip motion detection (*lips mirror*), facial expressions (*face emoji*), and Speech-To-Text with translation (*text++*).

ABSTRACT

Personal protective equipment, particularly face masks, have become increasingly common with the rise of global health issues, such as fine-dust storms and pandemics. Face masks, however, also degrade speech intelligibility by effectively occluding visual cues, such as lip motions and facial expressions. In this paper, we propose *MAScreen*, a wearable LED display in the shape of a mask, which is capable of sensing lip motion and speech and provides real-time visual feedback of the mouth behind the mask.

CCS CONCEPTS

• **Human-centered computing** → Human computer interaction.

KEYWORDS

Face mask, speech augmentation, lip motion, wearable display

ACM Reference Format:
Hyein Lee, Yoonji Kim, and Andrea Bianchi. 2020. MAScreen: Augmenting Speech with Visual Cues of Lip Motions, Facial Expressions, and Text Using a Wearable Display. In *SIGGRAPH Asia 2020 Emerging Technologies (SA '20 Emerging Technologies), December 04-13, 2020*. ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3415255.3422886

SA '20 Emerging Technologies, December 04-13, 2020, Virtual Event, Republic of Korea

ACM ISBN 978-1-4503-8110-9/20/12.
https://doi.org/10.1145/3415255.3422886

1 INTRODUCTION

The usage of face masks has recently surged due to the rise of health and hygiene awareness concerning air pollution and the spread of a pandemic. Furthermore, in many countries, wearing a mask indoors is not limited to a recommendation or etiquette but a mandatory practice with legal consequences for transgressors[1]. However, by visually occluding the lips, masks strongly affect the quality and expressiveness of verbal communication, as well as decreassing speech intelligibility [Sumby and Pollack 1954]. In fact, numerous researches have demonstrated the importance of visual cues in communication, such as mouth and lip motion [Summerfield 1989] – a problem only exacerbated by any hearing loss or impairment. While previous work aimed to enhance speech intelligibility and to mitigate face occlusion by proposing, for example, a transparent mask [Atcherson et al. 2017], in this work, we see an opportunity to utilize the mask as a wearable augmentation device that enhances real-time speech in ways that are not possible with current masks.

MAScreen (figure 1) is a wearable augmentation device that is capable of mapping lip motions through a machine learning algorithm that classifies facial expression and mouth position (i.e., how sounds are articulated). These are rendered in real-time on the front surface of the mask using an LED display, therefore giving immediate feedback to bystanders of how the mouth is actually moved behind the mask. Furthermore, the voice input through a microphone is converted to text via a Speech-To-Text (STT) detection software, and the LED display shows the resulting text or its translation to another language in real-time. While we acknowledge that the usage of LED displays on the face is not original for enhancing communication or entertainment[2], the application of lip

[1] https://www.nytimes.com/2020/07/16/us/coronavirus-masks.html
[2] LED matrix face mask: https://www.lumencouture.com

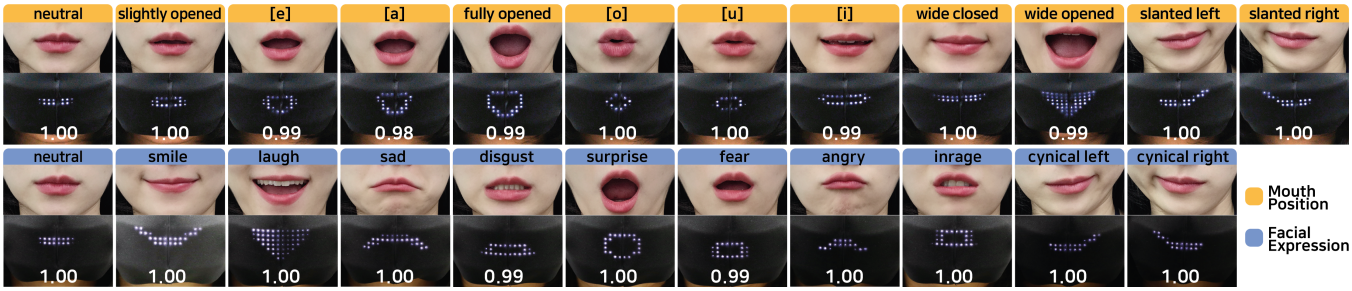

Figure 2: Classification result, f1-score indicated below, for 12 unique mouth positions and 11 facial expressions.

and speech detection using a combination of machine learning and STT with immediate feedback on a mask-shaped display is a novel contribution in the domain of augmented wearables.

2 MASCREEN SYSTEM

MAScreen consists of a custom lip sensing hardware with infrared reflective sensors facing the mouth and a flexible 9x24 LED display facing outward (Chemion LED smart glass by FunIOT), covered by a disposable hygiene mask.

Nine infrared reflective sensors (QRD1114) are attached to a custom 3D-printed cross frame (105mm x 90mm x40mm) at a distance up to 3cm from the lips — a layout that was empirically determined. When the mouth moves, the distance between the sensors and the skin changes, allowing sensors to detect lip motions, similarly to the system presented in [Sakashita et al. 2017]. The sensors are wired to an Arduino Mega (placed in a controlling box together with the display battery), which samples at an average of 640Hz the values for each sensor and transmits them to a PC for classification via serial (19200 bps). A small condenser mic (Boya BY-M3) is also attached to the 3D-printed case and directly connected to a PC. A Bluetooth module is mounted on the LED display for data to be transmitted back from the PC to the mask. The final prototype weights 32g.

Sensor data is received by a PC hosting a Python application running a machine learning classifier. After normalization of the raw sensing data collected from the photo-reflective sensor array and initial calibration, the sensor values are fed to a Support-Vector-Machine (SVM) classifier via Scikit-learn library.

For mouth positions we empirically selected 12 unique arrangements based on literature (such as vowels and various apertures), whereas for the facial expressions we selected 11 lip configurations, inspired by [Ekman and Friesen 2003]. To test the reliability of the classifier, for the mouth position, we trained and tested the system with the same user, using 90% of the collected data for training. We achieved a 99% accuracy based on 15,000 samples for each of the 12 lip configurations. For the facial expression, we achieved 100% accuracy based on 10,000 samples for 11 configurations. Details for each lip configuration with f1-scores are shown in Figure 2.

3 APPLICATIONS

Three applications were designed based on the results of a formative study with 12 participants (6 males, aged 22-28, M=25.5, SD=1.97). During the study, participants were asked to elaborate on their own experiences of wearing a face mask in public and how it affected their communication. Three patterns have emerged. Firstly, participants reported a degraded level of speech intelligibility as they *"could not see the lips"*. Secondly, they described that it was difficult to infer the real meaning of a sentence due to the lack of facial expression data. For example, one participant said that *"I heard 'nice' and I thought it was a compliment but it was a sarcasm (P6)"*. Lastly, one participant, referring to his experience of talking with a mask in a noisy environment suggested that *"the best way [to enhance speech] would be to show text, like subtitles"*.

Based on these findings we developed three applications written in Python: 1) *lips mirror*, 2) *face emoji*, and 3) *text++*. In the *lips mirror* application, the mask displays twelve distinct mouth positions, based on how the lips are moved in real-time. This application allows the users to see the speaker's lips and to determine when a person is speaking. It also provides an opportunity for lip reading. The *face emoji* application works similarly, but it aims to convey facial expressions by providing additional feedback other than speech cues (e.g., smile, grin...), and allowing to contextualize the real meaning of a sentence. Finally the *text++* application generates real-time subtitles and speech translation displayed on the mask. This application utilizes the Google Cloud API for both STT synthesis and translation to various languages.

ACKNOWLEDGMENTS

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the G-Information Technology Research Center support program (IITP-2020-2015-0-00742).

REFERENCES

Samuel R. Atcherson, Lisa Lucks Mendel, Wesley J. Baltimore, Chhayakanta Patro, Sungmin Lee, Monique Pousson, and M. Joshua Spann. 2017. 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* 28, 1.

Paul Ekman and Wallace V Friesen. 2003. *Unmasking the face: A guide to recognizing emotions from facial clues.* Ishk.

Mose Sakashita, Tatsuya Minagawa, Amy Koike, Ippei Suzuki, Keisuke Kawahara, and Yoichi Ochiai. 2017. You as a Puppet: Evaluation of Telepresence User Interface for Puppetry. In *Proceedings of UIST'17.* 217–228.

William H Sumby and Irwin Pollack. 1954. Visual contribution to speech intelligibility in noise. *The journal of the acoustical society of america* 26, 2 (1954), 212–215.

Quentin Summerfield. 1989. Lips, teeth, and the benefits of lipreading. *Handbook of research on face processing* (1989), 223–233.