1. Introduction

Advances in artificial intelligence (AI) have allowed machines to take over certain tasks in medical diagnosis, aerospace, electronic trading platforms, robotics, transportation, manufacturing industries, education, etc.\textsuperscript{[1–7]} The increasing ubiquitous collaborative intelligence between humans and machines is forcing human–machine communication (HMC) to be more human and less machine like to accomplish the task at hand. Humans communicate with other humans mainly through speech signals because they can convey more information than other means of communication, such as gestures, and it may be the ubiquitous, natural mode of communication for humans and machines in the near future. Advances in speech recognition and speech synthesis through recent developments in deep learning methods have already created a voice-guided user interface (VUI) in many products, including Siri, Google Assistant, and Amazon Alexa. However, the development of single-modal communication using VUI deprives speech-impaired people from accessing the benefits of AI-integrated machines. Moreover, HMC via speech can also be easily affected by background noise, which can lead to ineffective HMC in noisy environments.\textsuperscript{[8,9]} Therefore, a multimodal HMC that is suitable for all people and works in all types of environments is needed. In recent years, to give voice to speechless individuals and reduce the impact of noisy environments, the advantages of computer vision and deep learning have been used to decode speech from recorded lip movements.\textsuperscript{[10]} Lip reading technology, which is also commonly known as visual speech recognition, has other advantages, such as ensuring the security of HMC with or without the audio signal, maintaining a noise-free environment, and biometric authentication.\textsuperscript{[11]} Therefore, lip-reading technology is considered as a possible candidate for multimodal HMCs.

However, lip-reading technology works well only in a well-lit environment. Poor illumination can affect feature extraction and the deep learning of lip movements, which consequently affects the accuracy of the model. Therefore, the application of lip-reading technology for HMC is completely ineffective in the dark where the possibility of illuminating the face is limited. Increasingly, humans accompanied by AI-integrated machines are exploring the darkness...
of deep oceans, deep space, tunnels, and caves; a novel HMC model integrated into a multimodal communication system could prove useful in facilitating human–machine interaction in such dark environments. \[12,13\] Furthermore, a possible energy crisis in the future would also require dark-mode HMCs. Therefore, to enable humans to communicate with intelligent machines in dark environments, this study presents a novel HMC model that can be used by anyone (regardless of ability) and is not affected by background noise.

HMC in complete darkness may be inspired by bioluminescent communication, where deep-sea creatures and cave creatures generate bioluminescence (a form of cold light) to transmit and receive large amounts of information within a species and sometimes between species.\[14\] The application of light sources to modulate robotic manipulation actions has been recently studied using flexible photodetector textiles, introducing a new interactive method for optical mechanical communication.\[15\] A form of cold light could be used to set up the dark-mode HMC via visualization of lip movements during speech, and deep learning algorithms could be used to decode speech in a manner similar to that used in lip reading technology. Thus, no external source of illumination such as a light-emitting diode (LED) is required to visualize the lips. A simple form of cold light placed around the lips could facilitate realization of dark-mode HMC. The possible source of cold luminescence to synchronize lip movement could be either persistent luminescence (PL) or mechanoluminescence (ML).\[16–18\] PL, also known as long-lasting phosphorescence, is light emission from certain materials that can persist for an extended time after the excitation sources, such as UV, have ceased. Following the pioneering work of Matsuzawa, who discovered an ultralong green PL over 24 h in SrAl2O4:Eu,Dy (SAO) phosphors, several PL materials with different peak emission wavelengths have been discovered and widely used for emergency signs, indicators, in vivo optical imaging, and toys in the past few decades.\[19\] In contrast, ML is the emission of light when a solid material is subjected to mechanical stimuli such as stress, pressure, friction, and vibration. Among the various ML materials, the self-reproducible ML material, in which ML is generated and maintained under repeated mechanical stimuli without the need for UV preirradiation, may be ideal for visualizing lip movements. Materials exhibiting obvious self-reproducible ML include ZnS:Cu/Mn\(^{2+}\), CaZnOS: Ln\(^{3+}\), SrZn2S2O:Ln\(^{3+}\), and SrZn2S2O:Ln\(^{3+}\). Various applications such as display systems, light sources, heartbeat detection, hybrid sensors, and motion-driven ML fibers have been proposed.\[20\]

To prove the concept of dark-mode HMC, PL is chosen in this study to decode the speech from the skin surrounding the lips using convolutional neural network (CNN) models. Ultrasoft PL–polymer composite patches were used to record the movement pattern of the skin during speech in a dark environment. Ten different classes, including seven words, two letters, and a smiling gesture, were considered to evaluate the performance of CNN models with different feature extraction layers. The present work will help all people, including disabled people with limited dexterity, voice tremors, or poor vision, to communicate with artificial intelligence machines in a dark environment.

2. Results

2.1. Visualization of Lip Movements in a Dark Environment

Initially, eight ultrasoft and flexible SAO–Ecoflex composite patches were applied directly to the lips to visualize lip movements in response to spoken words in a dark environment. However, we found that there was frequent contact between the tongue and the pasted PL patches on the lips. Therefore, in view of possible health problems, we selected the skin above the orbicularis oris muscle to apply the PL patches. The orbicularis oris muscle surrounds the lips and controls the shape of the lips during speech. Therefore, the skin above the orbicularis oris muscle could be an ideal location to decode speech based on the movement patterns of the PL patches. On this basis, eight PL patches were applied to the skin with biocompatible soft glue, as shown in **Figure 1a**. In a dark environment, the luminous PL patches can be seen surrounding the lips, as schematically presented in **Figure 1a**. It should be noted that the PL patches were naturally charged under fluorescent room light. Thus, no special UV source was used to charge the PL patches.

Because this study is mainly focused on improving the decoding of speech in dark environments, the number of distinguishable classes was limited to ten. Considering a small-scale database allows a comprehensive presentation of the model performance for each class, which should aid understanding among general readers. Furthermore, care has been taken in the selection of classes to include short words, long words, like-sounding words, visually similar words, visually different words, and gestures (**Figure 1a**).

The video for a particular class was recorded continuously, while the words were uttered repeatedly 20 times with a short interval in between. In this manner, 40 video clips, representing four participants and ten classes (\(4 \times 10\)), were obtained. A
portion of a video clip for each class is included as supplementary information, showing the movements of the PL fields in response to speech. To facilitate the extraction of the corresponding frames for each repetition, the short interval between pronunciations was synchronized by blocking the camera with a sheet of A4 paper. Also, ten representative frames from each repetition were selected for further analysis and feature extraction for deep learning. To extract ten frames for each repetition, each video clip was first converted into images of size 210 × 210. Then, the maximum intensity of each frame was determined and plotted against the number of frames, as shown in Figure 1b. The plot illustrates sudden increases and decreases in the intensity in cyclic mode, where the maximum value of intensity decreases continuously. The sudden rise and fall results from exposure to PL and blocking of the camera lens, whereas the falling intensity is a decaying afterglow of SAO as mentioned earlier.[19] Figure 1b was further processed by defining the threshold intensity and then converting it into a binary plot, as shown in the inset of Figure 1b. Based on the binary data, the first and last frames of each cycle were determined, representing the beginning and end of the exposure. After determining the frames for the beginning and end of the exposure, ten frames were extracted for each cycle considering equal sampling intervals, as shown in Figure 1c. Using this strategy, 800 frames (selected frame number X repetitions X participants) were extracted for each class, and 8000 frames were generated for ten classes.

We investigated whether the number of frames considered for each pronunciation (i.e., 10) retains sufficient movement patterns. Increasing the number of frames would certainly increase the information but requires more computational effort in

---

**Figure 1.** Illustration of data generation. a) Schematic illustration of video recording with a mobile phone in a dark environment while speaking or gesturing a specific set of words, with eight PL composite patches on the skin surrounding the lips (The cartoon woman head has been reproduced with permission from TurboSquid). b) A framework to extract ten frames for each repetitive action in a video. c) A sequence of ten frames obtained using the method shown in (b).
training the deep learning model. To visualize the motion patterns, ten frames were merged into a single image, as shown in Figure 2a. In some classes such as in “around,” one can clearly see multiple shifts in the position of the patches. In other classes, such as in “strain,” such shifts are very difficult to see owing to the small movement of the PL patches. For a closer look, two different regions of interest (ROIs) were selected in the merged image, as shown in Figure 2b as ROI-1 and ROI-2. It can be clearly seen that both ROIs show a clear representation of patch motion for most of the classes. Even similar words such as “around” and “ground” show different patterns in both ROIs. Even though there is similarity in one patch, other patches show a unique representation of lip movement. For instance, “smile” shows a similar pattern with strain and fracture at ROI-1 but has a distinct pattern at ROI-2. This suggests that there is local variation in the deformation of the skin surrounding the lips and that these deformations are highly correlated with the words spoken. Because the PL patches helpfully reveal the local variations in skin deformation, it is possible to see how effectively deep learning algorithms can learn these patterns and decode them into speech.

2.2. Decoding Motion Patterns of PL Patches into Spoken Words

In this study, a CNN was used to learn the connection between the motion patterns of PL patches and spoken words. A CNN is a type of deep learning model for processing data with grid-like topology (such as images) and is designed to automatically and adaptively learn spatial hierarchies of features from low-to high-level patterns. CNNs are typically composed of three distinct operational layers: a convolutional layer, a pooling layer, and a fully connected layer. The convolution layer applies a convolution operation to the input data busing a convolutional filter (kernel) to create a feature map. The result of the convolution operation is passed through a nonlinear activation function, such as a rectified linear unit (ReLU), before being passed to the pooling layer. The nonlinear activation function gives the CNN the

Figure 2. a) Visualization of movement patterns of PL patches for each class in respective merged images using ten representative frames. b) Illustration of the motion patterns of PL patches for each class at ROI-1 and ROI-2.
ability to learn particularly complicated objects. In contrast, the pooling layer is mainly responsible for reducing the in-plane dimensionality of the feature map. This helps to reduce computational requirements and combats the problem of overfitting when training the network. The output feature maps from the last layer of the convolutional layer or the pooling layer are converted into a 1D vector and connected to one or more fully connected layers (dense layers), in which each neuron in one layer is connected to each neuron in another layer by a learnable weight. The output of each neuron in the fully connected layers is passed through a nonlinear activation function, such as ReLU. The last fully connected layer usually has the same number of output nodes as the number of classes, and the activation function applied to the last fully connected layer is typically different from the others. An activation function applied to the multiclass classification task is a softmax function that normalizes the real output values of the last fully connected layer to the target class probabilities, where each value is between 0 and 1, and all values sum to 1.

The number of convolutional and pooling layers can be increased in the CNN architecture depending on the complexity of the images; a deep CNN can more effectively capture low-level details but at the cost of higher computational requirements. In recent years, owing to significant increases in computing power, many deep CNNs have been developed, such as AlexNet, GoogleNet, DenseNet, VGGNet, ResNet, and Mobilenet.[23]

In the present study, different CNN models inspired by VGGNet were tested with increasing depth in the feature-extracting layer. The fully connected layer was fixed in all CNN models. The depth of the feature-extracting layer was increased by simply repeating the smallest unit of the feature-extraction layer, which consisted of two convolutional layers and a max-pooling layer. The fully connected layer consists of three different layers. The first and second fully connected layers contained 256 and 128 neurons with a ReLU activation function and 50% dropout layer, respectively. The dropout layers were included to minimize overfitting. The final fully connected layer contained ten neurons for ten classes and was assigned a softmax activation function. Each convolutional layer consisted of a $3 \times 3$ kernel, $1 \times 1$ stride, and ReLu activation function, whereas the number of filters varied depending on the position of the feature-extraction layer, as shown in Figure 3a. The models were tested using the Adams optimizer to determine the global minimax of the categorical cross-entropy loss function with a given learning rate of 0.0001. The model consisting of one feature-extracting layer is referred to as the VGG-1 model. Similarly, models with two, three, four, and five feature-extraction layers are, respectively, called VGG-2, VGG-3, VGG-4, and VGG-5. It is important to note that the feature-extracting layer consisted of two convolutional layers and a max-pooling layer. The architecture of VGG-5 is illustrated in Figure 3a. Moreover, a deep CNN model has been implemented using the ResNet-34 architecture to validate the performance of the VGGNet-based models. The architecture of ResNet-34 is illustrated in Figure S3, Supporting Information.

To facilitate feature extraction from the representative 10 frames containing the spatial and temporal information of lip movements, the frames were superimposed after conversion to grayscale, as shown in Figure 3b, which provided a total of 800 data points (class $\times$ participants $\times$ repetitions) with channel dimensions of 10. The augmentation technique was applied in this work to artificially increase the number of data points by five times to reduce the overfitting problem, which resulted in 4000 data points from the augmentation. Augmentation was performed by random rotation ($\pm 20^\circ$), random scaling (0.9–1.1), flipping (horizontal), and random translation (horizontal $= \pm 10$ pixels, vertical $= \pm 10$ pixels) on each stacked image, as shown in Figure 3b. All data from the augmentation were used for training the model, whereas the 800 original data points were equally divided into validation and test datasets. The split was made based on the class, so that each class was represented equally in the validation and test datasets.

The results of the five different VGGNet-based models and ResNet-34 are presented in Table 1, which lists the train accuracy, validation accuracy, and test accuracy. VGG-1 showed a training accuracy of $\approx 90\%$ and validation and test accuracy of 72% and 74%, respectively. The lower percentage of accuracy in validation and test compared with the training accuracy percentage clearly shows the problem of overfitting that comes from the shallow feature-extraction layer. The improvement in overfitting was somewhat reduced in VGG-2, where the accuracy of the model in the validation and test datasets remarkably increased to 85.5% and 90%, respectively. Moreover, the additional feature-extracting layer in VGG-2 increased the training accuracy by 6% as compared with the VGG-1 model. The trend of increasing training accuracy and decreasing overfitting problems in response to the additional feature-extracting layer is observed up to VGG-5. In contrast, ResNet-34 showed a training accuracy of 100% with validation and test accuracies of 99% and 98.75%, respectively, which is quite similar to the performance of VGG-5 (training accuracy $= 99.02\%$, validation accuracy $= 99.02\%$, and test accuracy $= 98.75\%$). Thus, it can be clearly concluded that the deeper the feature extraction layers, the better the model performs to increase accuracy and combat overfitting problems.

It is interesting to observe the performance of some models on individual classes with test datasets using the confusion matrix plot, as shown in Figure 4. The confusion matrix of VGG-1 shows that the model is most frequently confused between “strain” and “stress” and “t” and “b.” Similarly, the model also shows some confusion in distinguishing between “around” and “ground.” VGG-1 predicted “smile” with a lower error percentage of 2.5%, whereas the model’s accuracy in predicting “fracture” was the worst, at only 27.5%. The poor prediction of “fracture” is due to incorrect predictions of other classes. VGG-2 significantly improved the prediction accuracy for each class compared with VGG-1. For example, the confusion between “strain” and “stress,” “b,” and “t,” and “around” and “ground” was significantly reduced. The accuracy of predicting “smiles” was 100% for VGG-2. The prediction percentage of “fractures” was also improved to 72.5%. The confusion matrix of VGG-5 is excellent, as shown in Figure 4. Similar to the percentage accuracy of the models from VGG-1 to VGG-5, the performance of the models in terms of the percentage accuracy for each class also increased, with VGG-5 showing the best performance. The model was able to predict “around,” “b,” “ground,” “mechanics,” “smile,” “stress,” and “t” with 100% accuracy, while “fracture” and “science” were predicted with 97.5% and 95% accuracy, respectively. VGG-5 occasionally predicted “stress” for a given true class.
strain,” which resulted in precision of “strain” to 92.5%. Interestingly, VGG-5 was able to predict “stress” with 100% precision. Furthermore, the ResNet-34 confusion matrix shows a similar tendency of confusion in certain classes. For example, the confusion of VGG models between “strain” and “stress” is also observed in ResNet-34. Moreover, both VGG-5 and ResNet-34 showed a prediction accuracy of 100% for the classes “around,” “b,” “t,” “ground,” “mechanics,” and “smile.” Thus, it can be concluded that there is minimal difference between the two models in terms of the tendency of confusion for a particular class despite the different principles in architecture.

All models classified “smiles” with the highest accuracy. With the VGG-based models, 100% accuracy was achieved immediately after VGG-2. This is likely due to the distinct motion patterns of the majority of patches, which result from stretching of the lips. The other classes, such as smiling, do not involve stretching of the lips, which results in distinct patterns that can be distinguished by shallow CNN models. The confusion matrices also show that the models are confused based on similarity of lip deformation rather than on pronunciation duration. For example, “mechanics” takes the longest time to pronounce compared with “b” and “t,” but the VGG models performed well despite their pronunciation durations. In contrast, two points considered more closely in “stress” and “strain” in Figure 2b clearly show some degree of similarity in the motion patterns of PL patches. It is likely that some of the remaining PL patches might also show some degree of similarity owing to the similarity of lip movements when these two words are pronounced. As a result of the similarities, even the deep CNN

| Method  | Train accuracy | Validation accuracy | Test accuracy |
|---------|----------------|---------------------|--------------|
| VGG-1   | 89.9%          | 72%                 | 74%          |
| VGG-2   | 95.7%          | 85.5%               | 90%          |
| VGG-3   | 96.68%         | 96.25%              | 94.25%       |
| VGG-4   | 98.5%          | 98.75%              | 98.5%        |
| VGG-5   | 99.02%         | 99%                 | 98.75%       |
| ResNet-34 | 100%          | 99%                 | 98.75%       |

Figure 3. Illustration of the a) architecture of the CNN based on VGGNet and b) data augmentation.
(DCNN) tended to be confused. One way to improve the discriminability of the features between the classes could be to increase the spatial resolution of the skin deformation by adding more PL patches around the lips. Increasing the spatial resolution of skin deformation could potentially increase the number of distinguishable patterns, allowing CNN to classify classes with higher precision. Another way to improve performance is to improve the CNN model itself. Thus, in addition to the VGGNet and ResNet architectures, other deep CNN models, such as AlexNet, GoogleNet, DenseNet, and MobileNet, can be used to determine the best complementary model for this study.

3. Discussion

To our best knowledge, no previous study has attempted to decipher speech from the movement of the skin surrounding the lips. However, our current work has successfully demonstrated that it is possible to use the skin near the lips to decode speech by simply classifying the motion patterns of the PL patches using a deep learning algorithm. Conventional lip reading technologies require well-lit images of lip deformation to generate features for deep learning to decode speech, unlike the present study, where deep learning does not consider geometric features of the lip. Instead, the deformation of the orbicularis oris muscle during speech is transferred to the PL patches to decode speech. More importantly, speech is decoded in a dark environment. Therefore, it can be useful for many people, including disabled people with limited dexterity, vocal tremors, and poor vision. This dark-mode communication can be used in various fields of science and technology, as shown in Figure 5. Recently, voice-activated vehicle control systems have been developed for automobiles.[24] However, these systems must contend with background engine noise, and passengers have demanded improved performance from lip-reading technology. However, the requirement for illumination in lip-reading technology has limited its application to daylight hours, as driving at night with the interior light illuminated is not safe. Therefore, we believe that our findings will help solve this problem. Dark-mode communication can also be integrated with Google Assistant, Siri, and Amazon Alexa to provide secure private interaction in the dark, and speech-impaired individuals can use it without restriction. Furthermore, dark-mode communication can be

![Figure 4. Comparison between the confusion matrices of four different models: VGG-1, VGG-2, VGG-5, and ResNet-34, where the column instances are the predicted classes and the row instances are the true classes.](image-url)
used in combination with physical biometrics such as fingerprints, palm prints, iris shape, etc. or alone in physical access control systems (e.g., to open doors to private or business properties) by setting up a lip motion password. Dark-mode communication between humans and robots in the darkness of deep oceans, deep space, and tunnels could be very promising for building a robust multimodal communication system. Cellphone applications developed for dark-mode communication can be used in nighttime emergencies such as when a person loses their voice during a fire incident or when a house is attacked by intruders.

In reality, people move their heads while talking. In the present work, however, head positions were fixed as much as possible when facing the camera. Because the deep learning models in the present work were trained considering random image rotation, image translation, and scaling, a certain degree of flexibility was provided for head movements when facing the camera. However, to accommodate head rotation away from the camera, the models must be trained with new experimental datasets that consider such head motions. Some studies have considered multiple cameras to record multiple views of the speaker while speaking to train the model. Applying a multiview database in various cameras to record multiple views of the speaker while considering such head motions. Some studies have considered the models must be trained with new experimental datasets that otherwise would result from PL patches directly adhered to the lips. Moreover, the PL patches used in this study are biocompatible and ultrasoft, which further increase the convenience of using them. The solution suggested in the present work is designed to be helpful in harsh environments such as deep oceans and deep space, and the visible range wavelength of a PL patch could be a better signal to measure than an IR signal as it can penetrate the protective glass shields of helmets.

In this study, PL has been considered to establish the concept of dark-mode HMC. However, PL materials require an external energy source such as UV or daylight to recharge on a regular basis. To overcome this drawback of PL materials, ML materials are an ideal substitute. Various studies have reported intense emissions from ultrasoft ML composites such as ZnS: Cu-polydimethylsiloxane (PDMS) and ZnS:Mn–PDMS under weak stimuli of moving facial skin. It would therefore be interesting to decode speech using skin-driven ML in the present study. Decoding speech using skin-driven ML will be the subject of our future work. Furthermore, our future work will consider a larger number of distinguishable words, as large-scale databases with multiple speakers play a major role in the development of lip-reading technology. In addition, other more tonal languages such as Mandarin will be considered in future work.

4. Conclusion

In this study, we present a possible method of communication between humans and artificial intelligence machines based on deep learning of PL motion patterns of the skin surrounding the lips. PL sources of ultrasoft and flexible SAO–Ecoflex composite patches were considered to record lip motion in a darkroom using a cell phone camera for ten different classes. Based on the motion patterns, it was found that there are local variations in the deformation of the skin surrounding the lips and that the local variation in deformation is strongly correlated with the respective class. Various VGG-based CNN models were tested by considering the feature extraction layer as a variable. It has been shown that the deeper the feature extraction layers, the better the model performs to increase accuracy and combat overfitting problems. A ResNet-34 model, which is significantly different from the VGG-based model, was also implemented to verify the performance of the VGG-based models. It was found that the VGG-5 and ResNet-34 models did not differ significantly in terms of the tendency of confusion for a given class and model.

Figure 5. Schematic illustration of applications of dark-mode communication in various fields of science and technology.
accuracy percentage. All models classified “smile” with the highest accuracy due to the unique motion patterns for the majority of PL patches. Dark-mode communication can be useful for all people, including the disabled with limited dexterity, vocal tremors, and poor vision. It can also be used to control automobiles, converse with robots, and for physical access control systems such as opening doors.

5. Experimental Section

Preparation of PL Composite Patches: PL–polymer composite patches were prepared from a homogeneous mixture of SAO microparticles and Ecoflex at a ratio of 20:80 by weight. SAO was purchased from Nemoto and Co., Japan, whereas Ecoflex was purchased from Smooth-On, Inc., USA. As Ecoflex had a mechanical compliance as high as that of human skin, it was selected for embedding the SAO microparticles. Mixing was performed using a planetary mixture for ≈2 min at 450 RPM. To improve the homogeneity and avoid agglomeration, ZrO2 balls with a diameter of 10 mm were added to the container before mixing. The well-mixed composite was then degassed for 5 min to remove the trapped air bubbles. Thereafter, the composite was poured onto the central part of one of the surfaces of the 3D-printed plate, which has several circular molds with a diameter of 2 mm and a thickness of 0.25 mm. A plate with circular molds was printed using CARIMA-IM-96. The spin coater was then used to spread the composite for 20 s at 1000 RPM. The surface was then wiped with a sharp knife, leaving the composite only in the molds. Subsequently, the plate was transferred to an oven and allowed to solidify for 1 h at 60 °C. Finally, the solidified composite patches were extracted. The entire process is shown in Figure S4, Supporting Information. The optical microscope (OM) image illustrating the distribution of the SAO in the patch is shown in Figure S1, Supporting Information.

Experimental Setup: Video recordings were made in a dark room, as shown in Figure 1a, using a cell phone camera (ISO 3200) at a frame rate of 30 fps. Four individuals whose native language was not English participated in the voice recording: three Koreans and one Nepalese. During the recording, the participant’s head was kept fixed as much as possible while facing the camera. The speech included ten different classes: two English letters (b and t) and eight English words (fracture, mechanics, science, strain, stress, ground, around, and smile). It should be noted that “smile” was chosen to show a gesture performed by stretching the lips rather than a pronunciation.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

S.T. and H.C.S. contributed equally to this work. This research was supported by Kyungpook National University Research Fund, 2021.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Keywords

dark-mode human–machine communication, deep learning, human–machine communication, luminescence, visual speech recognition

Received: February 7, 2022
Revised: March 22, 2022
Published online: April 29, 2022

[1] J. G. Richens, C. M. Lee, S. Johri, Nat. Commun. 2020, 11, 1.
[2] D. Izzo, M. Märtens, B. Pan, Astrodynamics 2019, 3, 287.
[3] S. Singh, P. K. Sharma, B. Yoon, M. Shojafar, G. H. Cho, I.-H. Ra, Sustain. Cities Soc. 2020, 63, 102364.
[4] S. Panesar, Y. Cagle, D. Chander, J. Morey, J. Fernandez-Miranda, M. Kliot, Ann. Surg. 2019, 270, 223.
[5] J. Li, H. Cheng, H. Guo, S. Qiu, Autom. Innov. 2018, 1, 2.
[6] B.-H. Li, B.-C. Hou, W.-T. Yu, X.-B. Lu, C.-W. Yang, Front. Inf. Technol. Electron. Eng. 2017, 18, 86.
[7] I. Roll, R. Wylie, Int. J. Artif. Intell. Educ. 2016, 26, 582.
[8] S. Yin, C. Liu, Z. Zhang, Y. Lin, D. Wang, J. Tejedor, T. F. Zheng, Y. Li, Eurasis J. Audio Speech Music Process. 2015, 1.
[9] T. Yoshioka, M. J. Gales, Comput. Speech Lang. 2015, 31, 65.
[10] S. Fenghour, D. Chen, K. Guo, B. Li, P. Xiao, IEEE Access 2021.
[11] C. Wright, D. W. Stewart, Eurasis J. Inf. Secur. 2020, 2020, 1.
[12] H. Liu, T. Fang, T. Zhou, Y. Wang, L. Wang, Procedia CIRP 2018, 72, 3.
[13] J. Flanagan, I. Marsic, A. Medl, G. Burdea, J. Wilder, in Multimedia Communications Springer, New York 1999, p. 502.
[14] S. Martini, L. Kuhnz, J. Mallefet, S. H. Haddock, Sci. Rep. 2019, 9, 1.
[15] L. Li, D. Wang, D. Zhang, W. Ran, Y. Yan, Z. Li, L. Wang, G. Shen, Adv. Funct. Mater. 2021, 31, 2104782.
[16] R. Gao, M. S. Kodaimiti, D. Yan, Chem. Soc. Rev. 2021, 50, 5564.
[17] H. Tan, T. Wang, Y. Shao, C. Yu, L. Hu, Front. Chem. 2019, 7, 387.
[18] J.-C. Zhang, X. Wang, G. Marriott, C.-N. Xu, Prog. Mater Sci. 2019, 103, 678.
[19] T. Matsuzawa, Y. Aoki, N. Takeuchi, Y. Murayama, J. Electrochem. Soc. 1996, 143, 2670.
[20] Y. Zhuang, R. J. Xie, Adv. Mater. 2021, 33, 0205925.
[21] Y. Hu, Y. Yang, X. Zhang, X. Wang, X. Li, Y. Li, T. Li, H. Zhang, J. Phys. Chem. C 2020, 124, 24940.
[22] J. Bottermann, P. F. Smet, Opt. Express 2015, 23, A868.
[23] A. Dhillon, G. K. Verma, Prog. Artif. Intell. 2020, 9, 85.
[24] Z. Huang, X. Huang, AIP Conf. Proc. 2018, 1955, 040074.
[25] R. K. Shubhangi, N. B. Balbhim, presented at Int. Conf. Comput. Methodol. Commun. ICCMC 2017, 2017 (accessed: July 2017).
[26] M. Hao, M. Mamut, N. Yadikar, A. Aysa, K. Ubil, IEEE Access 2020, 8, 204518.
[27] T. Saitoh, R. Konishi, presented at SICE-ICASE Int. Joint Conf. 2006 (accessed: October 2006).
[28] S. J. Anderson, A. C. M. Fong, J. Tang, IEEE Trans. Consum. Electron. 2013, 59, 352.
[29] X. Qian, Z. Cai, M. Su, F. Li, W. Fang, Y. Li, X. Zhou, Q. Li, X. Feng, W. Li, Adv. Mater. 2018, 30, 1800291.