ICIC Express Letters
Volume **17**, Number **11**, November **2023**

ICIC International ©2023 ISSN 1881-803X
pp. 1287–1294

SMART GLASSES LIVE TRANSCRIPTION AND AUDIO CLASSIFICATION MODEL TO ASSIST HEARING IMPAIRMENT

Moch Kelvin RA, Levandio Yuvens Jonathan*
Daniel Christopher Kesuma, Ivan Sebastian Edbert
Silviya Hasana and Derwin Suhartono

Computer Science Department
School of Computer Science
Bina Nusantara University
JL. K. H. Syahdan No. 9, Kemanggisan, Palmerah, Jakarta 11480, Indonesia
{ moch.kelvin; daniel.kesuma; ivan.edbert }@binus.ac.id; { silviya.hasana; dsuhartono }@binus.edu
*Corresponding author: levandio.jonathan@binus.ac.id

Received November 2022; accepted January 2023

Abstract. *Hearing impairment is one of the most common disabilities worldwide. To assist hearing-impaired individuals in communication and maximize their Quality-of-Life (QoL), we proposed a model that utilizes audio recognition trained with machine learning to classify ambient sound and generate live transcripts. We built the model from audio streamed from SoundDevice and then trained it using Convolutional Recurrent Neural Network (CRNN) model against the UrbanSound8k dataset to achieve audio classification followed by a pre-trained Wav2vec 2.0 model to convert speech in the audio into text. Our classification model shows 86% accuracy with 76% validation accuracy.*

Keywords: Hearing impairment, Speech caption, Audio classification, Convolutional recurrent neural network, Real-time audio streaming

1. Introduction.

Hearing impairment that requires rehabilitation affects over than 5% of the world's population, or estimated 466 million people (432 million adults and 34 million children). Hearing impairment is projected to affect approximately 2.5 billion individuals or 1 in every 4 people by 2050 [1]. Hearing impairment is defined as the inability to hear, a condition where sound transmission from the outer ear to the brain is disrupted [2] which can affect either one or both ears. Hearing impairment ranges from mild (21-35 dB), moderate (50-65 dB), severe (65-80 dB), profound (80-95 dB) to total hearing loss (> 95 dB) [1]. While mild-to-severe hearing-impaired individuals can benefit from hearing aids, some profound hearing-impaired individuals and those with total hearing loss are unfit candidates for hearing aids and require sign language.

Hearing-impaired individuals are at huge disadvantage because this disability extremely restricts their daily activities. Such restrictions for instance: activity limitations, including understanding spoken news on various outlets or public announcements [3] and difficulty participation, especially in conversations [4] that could potentially lead to social isolation. Machine learning is one of the leading approaches to assist with various disabilities, including hearing impairment. Implementation of sound recognition using machine learning in wearable devices, smartphones and smart devices suggests that it is possible to assist hearing-impaired individuals and contribute to alleviating their disability burden. In this paper, we proposed a sound recognition model that can produce ambient sound classifications and live transcription.

Various researches have proposed implementation of sound recognition model for wearable devices or head-mounted-displays. A communication support system with Gaussian

DOI: 10.24507/icicel.17.11.1287

Mixture Model – Hidden Markov Model (GMM-HMM) acoustic model was proposed by Watanabe et al. [5] to translate human voice into text scripts that are displayed real-time on smart glasses. This system implemented a sound localization technique to help locate the speaker's direction. Assistance systems [6, 7] were proposed by implementing Google's Speech to Text service Application Programming Interface (API) on Microsoft Hololens and Google Glass. Another eyewear-based approach was proposed by Olwal et al. [8] by augmenting speech and audio with visual representations. Automatic Speech Recognition (ASR) using Google Chrome's Web Speech API is implemented by Yamamoto et al. [9] for real-time captioning in face-to-face communication mediated with a transparent display. This transparent display is designated to not occlude the perception of body language.

Most recent approaches are implemented with machine learning. Jain et al. [10] proposed Soundwatch, an implementation of Convolutional Neural Network (CNN) in smartwatch to assist hearing-impaired individuals by providing real-time, always-available visual and haptic feedback. An Internet of Things (IoT)-based approach has been proposed as well [11] by performing sound processing using CNN model with 16 layers depth known as VGG16. A mobile-based sensing system for recognition and notification of acoustic events using CNN model is proposed by Liu et al. [12] that enables location-independent event recognition and notification system. "Protosound" is proposed by Jain et al. [13] that uses a CNN model for mobile devices, MobileNetV2 architecture to train sound data. Protosound obtained average accuracy of 90.4%. Finally, Ava is recommended for hearing impaired individuals including for classroom use [14]. Ava combined artificial intelligence with a professional captioner (Scribes) for instant transcription.

The primary purpose of this study is to develop a smart glass live transcription and audio classification model to assist hearing-impaired individuals. This model featured notifications, warnings, and text. For classification, the suggested model used CNN and Recurrent Neural Network (RNN) [15, 16, 17]. This study is divided into four sections. The first section described the introduction to this research, including the primary problems that the researchers would like to address and related research in the topic. The methodology of this study is discussed in Section 2. The results are presented in Section 3 and Section 4 will summarize the study's limitations and suggestions for future research.

2. Methodology.

We developed a model for smart-glasses to classify and caption ambient sound and human speech. Smart-glasses are chosen over other devices due to its comfort and preferability among users [18]. Unlike mobile devices and desktop based applications that induce visual dispersion in information, resulting increased in effort, concentration, as well as information loss [19], smart-glasses will accommodate hearing-impaired individuals to observe their surroundings and communicate face-to-face naturally, that will increase their environmental awareness and perceived emotional connections [20].

In our experiment, we use CNN and RNN. CNN is a very popular deep learning technique to detect features automatically without any human supervision [21, 22]. It is built up of neurons that can be learned and enhanced in terms of weight and bias. CNN is made up of an input layer, numerous hidden layers, and an output layer. A multi-layer neural network is typically made up of one convolutional layer and one fully connected layer. The convolutional layer is an essential component of CNN that is made up of linear and non-linear mathematical processes that extract features using filters called kernels. Following some procedures, the pooling layer decreases the dimension of input to speed up calculation and avoid overfitting. Before being passed to fully connected layers, the output of the feature learning process is flattened. Fully connected layers, known as dense layers, are those in which every input is connected to every output by a learnable weight for classification. The final dense layer is connected to a normalization layer or activation function chosen to normalize the output to the target class probability [22, 23, 24].

CNN is widely used in image classification, while RNN is widely utilized in sound recognition and natural language processing. Because RNN includes internal memory that is utilized to recall the input received, this method is recommended to predict sequential data where the immediate past data is needed to predict the future data, such as audio or text. Unlike a feed-forward neural network, which only considers the current input, an RNN contains two inputs: the current and recent past inputs, each of which has a weight that can be updated using gradient descent and Backpropagation Through Time (BPTT). RNN can be extended as well to create a Long Short-Term Memory Network (LSTM). Figure 1 represents the illustration of RNN.

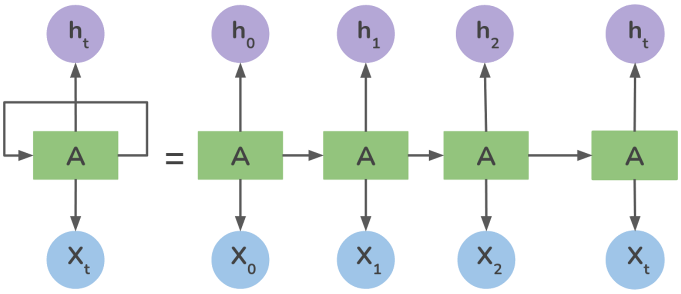


Figure 1. Rolled and unrolled RNN

We built our model using CNN architecture because, despite its popularity for image classification, CNN works for audio classification as well [25]. CNN may also be utilized for audio because audio data can be represented as raw audio waveforms [26] or spectrogram [15, 27] which can be processed similar to 2D image to find key features such as capturing time and frequency features from audio [16]. Despite the fact that audio can be represented in an image-like format, audio is a sequential data with a temporal component and CNN architecture is not deep enough to understand the complicated audio structure [28]. To address these concerns, we utilize RNN for classification by replacing the fully connected layer before the output layer with the RNN layer, which is then referred to as the Convolutional Recurrent Neural Network (CRNN).

We developed a model that combines a customized model to classify ambient sound with a pre-trained model from Wav2vec 2.0 to caption human speech [29]. Figure 2 illustrates our model flowchart. The custom model is built using PyTorch to analyze sound and build with CRNN architecture to classify ambient sound. This application will take audio streaming as input and transmit it to the transcript and ambient sound classification models. Output from both models will be displayed on the smart glasses.

The custom model was built using CRNN architecture made from CNN and RNN. CNN is made up of three layers: a convolutional 2D layer, a batch normalization 2D layer followed by an Exponential Linear Unit (ELU) activation function, and a max pooling 2D layer followed by a dropout layer with a dropout rate of 10%. The ELU mathematical formula is shown in (1). The RNN consists of LSTM with two layers and the hidden size is 64. The output layer is followed by a dropout layer with a 30% dropout rate and a batch normalization 1D layer.

$$\begin{array}{rl} fx = x & \text{if } x > 0 \\ \alpha \exp(x) - 1 & \text{if } x \leq 0 \end{array} \tag{1}$$

The dataset for the custom model was obtained from the UrbanSound8k which is a pre-sorted dataset from 10 classes organized into ten folders, including 8732 short-labeled sounds with each less than or equal to 4 seconds [30]. Table 1 displays the sound classes.

For this custom model, we tried to create and compare three different training data configurations and two different models to find which configuration is the best. The first

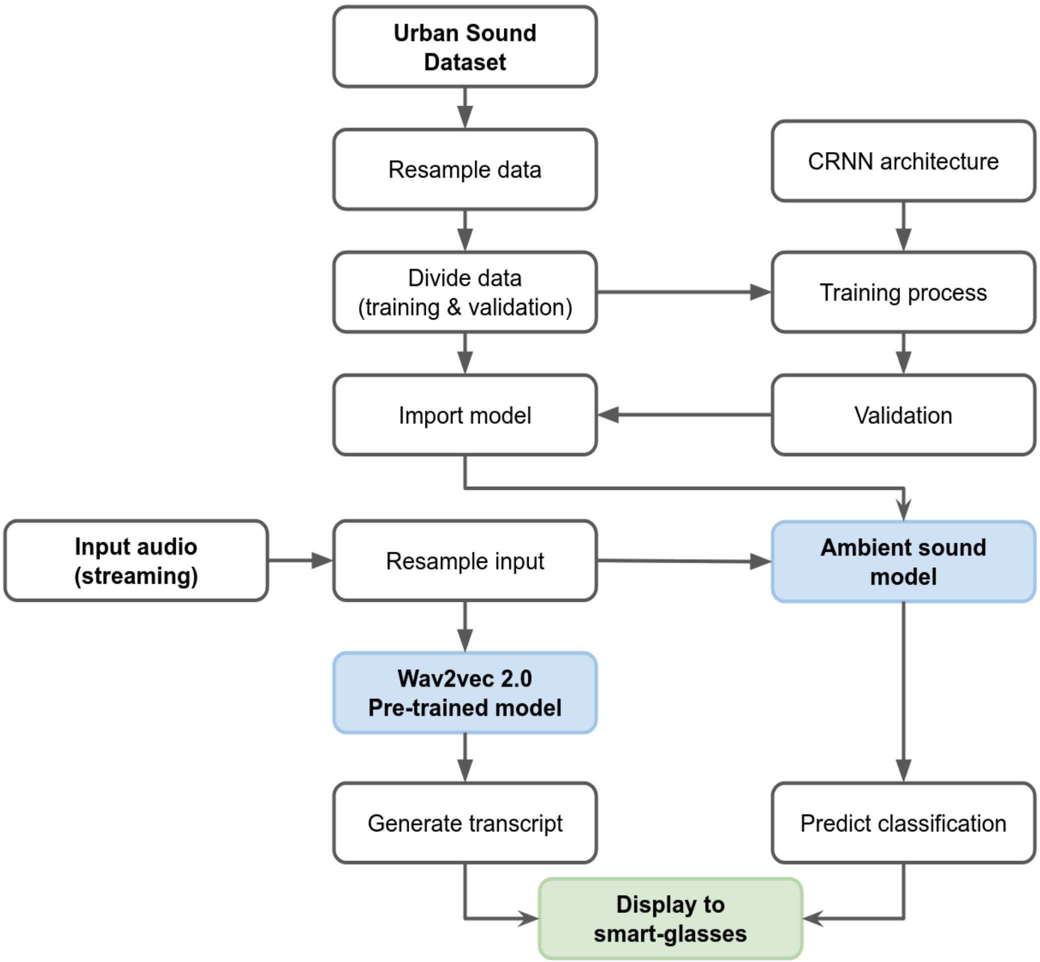


Figure 2. Flowchart model

Table 1. UrbanSound8k classes

No.	Class name	Number of samples										Total samples	Duration (seconds)
		Fold1	Fold2	Fold3	Fold4	Fold5	Fold6	Fold7	Fold8	Fold9	Fold10		
1	Air conditioner	100	100	100	100	100	100	100	100	100	100	1000	3994.928713
2	Car horn	36	42	43	59	98	28	28	30	32	33	429	1053.953286
3	Children playing	100	100	100	100	100	100	100	100	100	100	1000	3961.874525
4	Dog bark	100	100	100	100	100	100	100	100	100	100	1000	3148.749594
5	Drilling	100	100	100	100	100	100	100	100	100	100	1000	3548.244036
6	Engine idling	96	100	107	107	107	107	106	88	89	93	1000	3935.992542
7	Gun shot	35	35	36	38	40	46	51	30	31	32	374	616.796493
8	Jack hammer	120	120	120	120	120	68	76	78	82	96	1000	3610.974722
9	Siren	86	91	119	166	71	74	77	80	82	83	929	3632.701586
10	Street music	100	100	100	100	100	100	100	100	100	100	1000	4000
Total		**873**	**888**	**925**	**990**	**936**	**823**	**838**	**806**	**816**	**837**	**8732**	**31504.2155**

configuration included models 1, 2, and 3. Fold2 and 10 were used for test data on model 1. Fold3 and 10 were used for test data on model 2, and the third used only fold10 as test data. The rest of the data were used for training. To speed up data loading, the batch size for all models was set equivalent to 24 with 4 workers. If the dataset size was not divisible by the batch size, the last unfinished batch was dropped. Additionally, we reshuffled the data at every epoch so that the model did not learn unnecessary patterns. Before being fed to the model, the audio channel was mixed down to mono if necessary, added with noise to improve the model noise robustness and then trimmed with random length. We trained the model in Google Collab with 90 epochs and stop when the validation performance did not improve for 16 epochs (for the third model, we stop after 8 epochs) with Negative Log-Likelihood (NLL Loss) as the loss function and Adam as the optimizer with a starting learning rate of 0.002. The learning rate is then decayed by 0.5 per 10 epochs. The NLL Loss formula and the Adam optimizer function are defined in (2) and (3), respectively.

$$l(x,y) = \sum_{n=1}^{N} \frac{1}{\sum_{n=1}^{N} Wy_n} ln \sum_{n=1}^{N} ln \tag{2}$$

$$m_t = \beta_1 m_{t-1} + (1-\beta_1)\left[\frac{\delta L}{\delta w_t}\right] vt = \beta_2 v_{t-1} + (1-\beta_2)\left[\frac{\delta L}{\delta w_t}\right]^2 \tag{3}$$

The second configuration called model 4 utilized Stochastic Gradient Descent (SGD) as the optimizer with 0.002 learning rate and 0.01 weight decay, LinearLR as the learning rate scheduler with start factor 0.333 and end factor 1.0, also Cross Entropy Loss as the loss function. The Stochastic Gradient Descent and Cross Entropy Loss formula is shown in (4) and (5), respectively. The batch size equals 48 with 2 workers. This model trained with 100 epochs and stopped when the validation performance did not improve for 8 epochs. For comparison purpose, we use the same dataset used to train model 3.

$$\theta = \theta - \eta \cdot \nabla\theta_J\left(\theta; x^{(i)}; y^{(i)}\right) \tag{4}$$

$$L_{CE} = -\sum_{i=1}^{n} t_i \log(p_i), \text{ for } n \text{ classes} \tag{5}$$

We used a pre-trained model from Wav2vec 2.0 to develop the Live Transcript feature. This pre-trained model encodes speech sounds using a multi-layer CNN, then masks the resulting latent speech representations and feeds them to a Transformer network to produce contextualized representations. This model is pre-trained using 960 hours of unlabeled audio and fine-tuned with 100 hours of labeled data, both from Librispeech [29].

The Sounddevice library is used to record audio. When the application is launched on smart glasses, the audio recording begins automatically and converts the audio to a wav file every 3 seconds. The generated wav files will be used as input for the classification and live transcript models. To save memory, the audio file will be deleted before the next audio file is generated. If the confidence value is higher than 0.5, the prediction results will be displayed on the smart glass screen.

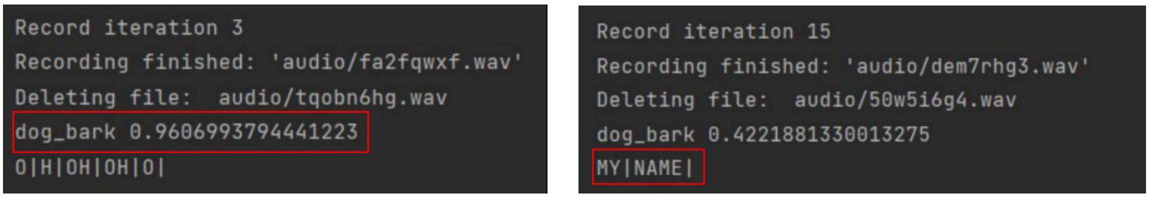
3. **Results and Discussions.**

After training all 139786 parameters in each model, we obtained model 3 as the best model. Table 2 shows that this model outperforms both of the other models in every term. With the same test data, the first model configuration produced better models, as shown in Table 2 that model 3 still outperforms model 4 in every aspect. As a result, we employed model 3 in our model, which has been trained on datasets ranging from fold1 to fold9. We conducted a test stream by playing the YouTube video “DOGS BARKING” followed by talking near the microphone saying “My name”. The models’ results are depicted in Figure 3.

We proposed a simple application that combined two models and displayed information such as the number iterations completed, newly generated audio file, audio file from the previous iteration, classification model results and speech transcript model results. As indicated in the red box in Figure 3, the classification model labeled the YouTube video as dog bark with 0.96 confidence. The speech transcript model in created “MY|NAME|” exactly as the words spoken, as seen in the red box. Figure 4 demonstrated results of our

TABLE 2. Training results

Class name	First configuration			Second configuration
	Model 1 (fold2 & 10)	Model 2 (fold3 & 10)	Model 3 (fold10)	Model 4 (fold10)
Best epochs	14	29	35	44
Total epochs	31	46	44	53
Loss	0.7054	0.5201	0.4699	0.9209
Accuracy	77.87%	84.82%	86.22%	70%
avg_precission	68.79%	74.98%	76.58%	67.03%
avg_recall	69.41%	75.69%	77.12%	67.55%
val_loss	1.1108	1.2257	0.8369	1.0034
val_accuracy	64.43%	63.01%	75%	68.99%
val_avg_precission	59.19%	57.84%	65.02%	66.25%
val_avg_recall	57.98%	57.17%	64.69%	65.68%



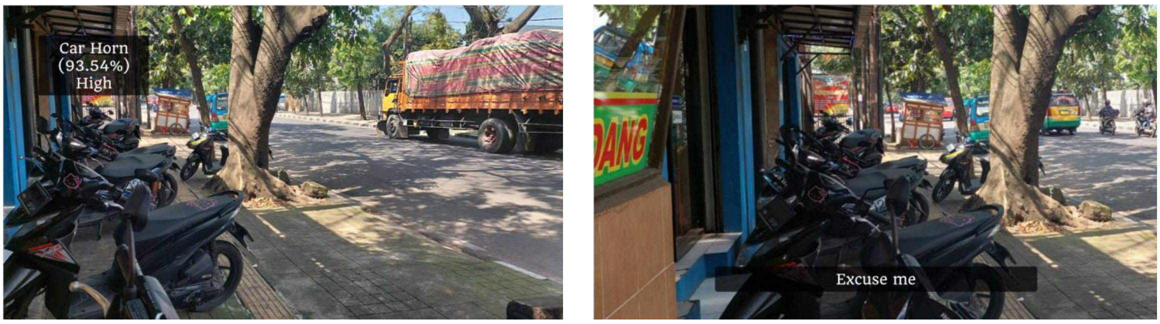
```
Record iteration 3
Recording finished: 'audio/fa2fqwxf.wav'
Deleting file:  audio/tqobn6hg.wav
dog_bark 0.9606993794441223
O|H|OH|OH|O|
```

(a) Classification result

```
Record iteration 15
Recording finished: 'audio/dem7rhg3.wav'
Deleting file:  audio/50w5i6g4.wav
dog_bark 0.4221881330013275
MY|NAME|
```

(b) Wav2vec 2.0 transcript result

FIGURE 3. Output models



(a) Output of sound classification

(b) Output of sound captioning

FIGURE 4. Output results of audio classification and audio captioning

sound classification of when a certain ambient sound is present and live transcription when a speech is heard but no unique ambient sound is heard. Despite good result in several experiments, the ambient sound classification model often gave false predictions as well that could be a foundation for further improvements. The results shown have confidence more than 50% with an indicator label to show the confidence category range from “Low” (50%-70%), “Mid” (70%-85%) and “High” (> 85%).

4. **Conclusions and Future Works.** We proposed a smart glasses application for ambient sound classification and live transcription to assist hearing-impaired individuals. We trained a CRNN model and obtained 86% accuracy and 75% validation accuracy. Our results often generated incorrect predictions that we plan to address in the future.

Another limitation is the audio length. Since the interval is quite short (3 seconds), some words might get cut and unrecognized. Non-voice audio could also trigger the model to produce an incomprehensible and meaningless output. On the other side, human speech could trigger the ambient sound classification to give high-confidence predictions. Currently, the UrbanSound8k dataset only has 10 classes, adding new data and classes will potentially increase the prediction accuracy.

REFERENCES

[1] World Health Organization and Others, *World Report on Hearing*, 2021.

[2] S. Anastasiadou and Y. Al Khalili, *Hearing Loss*, StatPearls Publishing, Treasure Island (FL), 2019.

[3] E. Amann and I. Anderson, Development and validation of a questionnaire for hearing implant users to self-assess their auditory abilities in everyday communication situations: The hearing implant sound quality index (HISQUI19), *Acta Oto-Laryngologica*, vol.134, no.9, pp.915-923, 2014.

[4] B. H. Timmer, L. Hickson and S. Launer, Adults with mild hearing impairment: Are we meeting the challenge?, *International Journal of Audiology*, vol.54, no.11, pp.786-795, 2015.

[5] D. Watanabe, Y. Takeuchi, T. Matsumoto, H. Kudo and N. Ohnishi, Communication support system of smart glasses for the hearing impaired, *International Conference on Computers Helping People with Special Needs*, pp.225-232, 2018.

[6] Y.-H. Peng, M.-W. Hsi, P. Taele, T.-Y. Lin, P.-E. Lai, L. Hsu, T.-C. Chen, T.-Y. Wu, Y.-A. Chen, H.-H. Tang et al., Speechbubbles: Enhancing captioning experiences for deaf and hard-of-hearing people in group conversations, *Proc. of the 2018 CHI Conference on Human Factors in Computing Systems*, pp.1-10, 2018.

[7] S. Alkhalifa and M. Al-Razgan, Enssat: Wearable technology application for the deaf and hard of hearing, *Multimedia Tools and Applications*, vol.77, no.17, pp.22007-22031, 2018.

[8] A. Olwal, K. Balke, D. Votintcev, T. Starner, P. Conn, B. Chinh and B. Corda, Wearable subtitles: Augmenting spoken communication with lightweight eyewear for all-day captioning, *Proc. of the 33rd Annual ACM Symposium on User Interface Software and Technology*, pp.1108-1120, 2020.

[9] K. Yamamoto, I. Suzuki, A. Shitara and Y. Ochiai, See-through captions: Real-time captioning on transparent display for deaf and hard-of-hearing people, *The 23rd International ACM SIGACCESS Conference on Computers and Accessibility*, pp.1-4, 2021.

[10] D. Jain, H. Ngo, P. Patel, S. Goodman, L. Findlater and J. Froehlich, SoundWatch: Exploring smartwatch-based deep learning approaches to support sound awareness for deaf and hard of hearing users, *Proc. of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, pp.1-13, 2020.

[11] D. Jain, K. Mack, A. Amrous, M. Wright, S. Goodman, L. Findlater and J. E. Froehlich, Homesound: An iterative field deployment of an in-home sound awareness system for deaf or hard of hearing users, *Proc. of the 2020 CHI Conference on Human Factors in Computing Systems*, pp.1-12, 2020.

[12] S. Liu, Z. Zhou, J. Du, L. Shangguan, J. Han and X. Wang, UbiEar: Bringing location-independent sound awareness to the hard-of-hearing people with smartphones, *Proc. of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol.1, no.2, pp.1-21, 2017.

[13] D. Jain, K. H. A. Nguyen, S. M. Goodman, R. Grossman-Kahn, H. Ngo, A. Kusupati, R. Du, A. Olwal, L. Findlater and J. E. Froehlich, ProtoSound: A personalized and scalable sound recognition system for deaf and hard-of-hearing users, *Proc. of the 2022 CHI Conference on Human Factors in Computing Systems*, pp.1-16, 2022.

[14] P. Millett, Accuracy of speech-to-text captioning for students who are deaf or hard of hearing, *Journal of Educational, Pediatric & (Re) Habilitative Audiology*, vol.25, 2021.

[15] M. Ahmed, T. I. Robin, A. A. Shafin et al., Automatic environmental sound recognition (AESR) using convolutional neural network, *International Journal of Modern Education & Computer Science*, vol.12, no.5, 2020.

[16] S. Li, Y. Yao, J. Hu, G. Liu, X. Yao and J. Hu, An ensemble stacked convolutional neural network model for environmental event sound recognition, *Applied Sciences*, vol.8, no.7, 1152, 2018.

[17] P. Gimeno, I. Viñals, A. Ortega, A. Miguel and E. Lleida, Multiclass audio segmentation based on recurrent neural networks for broadcast domain data, *EURASIP Journal on Audio, Speech, and Music Processing*, vol.2020, no.1, pp.1-19, 2020.

[18] L. Findlater, B. Chinh, D. Jain, J. Froehlich, R. Kushalnagar and A. C. Lin, Deaf and hard-of-hearing individuals' preferences for wearable and mobile sound awareness technologies, *Proc. of the 2019 CHI Conference on Human Factors in Computing Systems*, pp.1-13, 2019.

[19] A. Miller, J. Malasig, B. Castro, V. L. Hanson, H. Nicolau and A. Brandão, The use of smart glasses for lecture comprehension by deaf and hard of hearing students, *Proc. of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pp.1909-1915, 2017.

[20] D. Jain, R. Franz, L. Findlater, J. Cannon, R. Kushalnagar and J. Froehlich, Towards accessible conversations in a mobile context for people who are deaf and hard of hearing, *Proc. of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*, pp.81-92, 2018.

[21] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie and L. Farhan, Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions, *Journal of Big Data*, vol.8, no.1, pp.1-74, 2021.

[22] A. Dhillon and G. K. Verma, Convolutional neural network: A review of models, methodologies and applications to object detection, *Progress in Artificial Intelligence*, vol.9, no.2, pp.85-112, 2020.

[23] R. Yamashita, M. Nishio, R. K. G. Do and K. Togashi, Convolutional neural networks: An averview and application in radiology, *Insights into Imaging*, vol.9, no.4, pp.611-629, 2018.

[24] S. Albawi, O. Bayat, S. Al-Azawi and O. N. Ucan, Social touch gesture recognition using convolutional neural network, *Computational Intelligence and Neuroscience*, vol.2018, 2018.

[25] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold et al., CNN architectures for large-scale audio classification, *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp.131-135, 2017.

[26] T. Kim, J. Lee and J. Nam, Sample-level CNN architectures for music auto-tagging using raw waveforms, *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp.366-370, 2018.

[27] A. Polo-Rodriguez, J. M. Vilchez Chiachio, C. Paggetti and J. Medina-Quero, Ambient sound recognition of daily events by means of convolutional neural networks and fuzzy temporal restrictions, *Applied Sciences*, vol.11, no.15, 6978, 2021.

[28] J. Sang, S. Park and J. Lee, Convolutional recurrent neural networks for urban sound classification using raw waveforms, *2018 26th European Signal Processing Conference (EUSIPCO)*, pp.2444-2448, 2018.

[29] A. Baevski, Y. Zhou, A. Mohamed and M. Auli, Wav2vec 2.0: A framework for self-supervised learning of speech representations, *Advances in Neural Information Processing Systems*, vol.33, pp.12449-12460, 2020.

[30] J. Salamon, C. Jacoby and J. P. Bello, A dataset and taxonomy for urban sound research, *Proc. of the 22nd ACM International Conference on Multimedia*, pp.1041-1044, 2014.