Proceedings of the Federated Conference on Computer Science and Information Systems pp. 497–500
DOI: 10.15439/2018F13
ISSN 2300-5963 ACSIS, Vol. 15

Voice control in mixed reality

Dawid Połap
Institute of Mathematics
Silesian University of Technology
Kaszubska 23, 44-100 Gliwice, Poland
Email: Dawid.Polap@polsl.pl

***Abstract*—The gameplay in the augmented or virtual reality is based on the use of external equipment such as glasses/telephones and possibly the use of additional sensors or controllers. In both cases, the interaction involves pressing the keys on the phone or controller. An interesting aspect is the control of objects created in a virtual way using voice commands. In this paper, we propose a solution to manipulate objects in the augmented reality using player's voice. The user can move the object using pre-programmed commands. The solution is based on speech processing and artificial neural networks. The technique has been tested and the results presented and discussed.**

I. Introduction

AUGMENTED, merged or virtual realities are leading technological and research directions in the field of human machine interaction, image processing, games and many more. The rapid development of technology is visible through numerous achievements in these areas. The last examples are getting rid of mobile phones for extended goggles with motion controllers that allow not only to much more interaction, but a much longer immersion (the goggles do not overheat as fast as smartphones). Not only virtual reality, but augmented one is widely used. It is especially used in games, learning or even medicine what is visible in the effects of scientific work of scientists from around the world.

The biggest commercial achievement of these technologies was the mobile application for catching virtual monsters in 2016, i.e. *Pokemon GO*. The game based on searching and catching was based on a combination of reality (using a camera and GPS locator in smartphones) with 3D models that caused young people to leave the houses for fresh air and cross the streets. To this day, game is considered as a good change in the gaming industry when it comes to long time spent in front of the computer [1]. These types of applications not only have a good effect on health, but also make it possible to meet new people and cooperate [2]. Support for augmented reality is performed using a certain application with access to the camera. It is a kind of human–machine interaction [3], [4].

The transition from augmented to virtual reality is quite smooth. It is only necessary to turn off the camera and set up goggles to be able to move into a fully artificially created world. Studies on immersion are dispelled in increasing the feeling of detachment from the environment [5]. Enabling and capturing movements is quite a significant problem. However, it is gradually solved, as shown by the authors of [6], where the virtual keyboard was presented. Health problems such as stress or focusing too close to the display are constantly analyzed so that in the future everyone can safely and fully immerse into virtual world [7]. Nowadays, important issue is Internet of Things and its impact on our life [8], [9]. Similar, this technology and methods can be applied to different problems like some of geological [10], [11].

An important aspect is also research in the field of voice analysis and processing. In [12] deep learning was used to quickly detect the accent for English language. Again in [13], the authors focused on recognizing emotions on the basis of voice samples using classical processing and classification techniques. Moreover, the techniques of voice processing recorded in the sound file to the text version using artificial intelligence methods are presented and widely described in terms of numerous applications in practice [14].

In this paper, we focus on introducing a voice to this type of technology. Selected, recent developments in the signal processing field can be viewed in [15], [16]. My proposition is based on the analysis of the voice and its transformation into a text command that can be realized in an augmented/virtual reality.

II. Voice Controller

The voice controller operates on the basis of downloading a sound sample, its processing, classification and processing of the output decision. Described controller is based on converting a speech signal into its graphical form, cutting and classifying it, and then transferring information to an application supporting augmented or virtual reality.

A. Voice processing

The sound analysis is quite problematic for several reasons. The pronounced sound has the continuous form, by saving it in bit form, it simplifies the signal and is still incapable to analyze. Let $s(n) = (s_0, s_1, s_2, \ldots, s_{N-1})$ is a signal in discrete form. Unfortunately, this number sequence is still not possible to be analyzed. To remedy this, use a selected transformation such as Fourier defined as follows

$$S_k = \sum_{n=0}^{N-1} s_n \exp\left(-\frac{2\pi i n k}{N}\right) \quad 0 \leq k \leq N-1, \qquad (1)$$

where $S_k \in \mathbb{C}$ is a discrete value in $(S_0, S_1, S_2, \ldots, S_{N-1})$. For the purpose of machine calculations, in practice the above

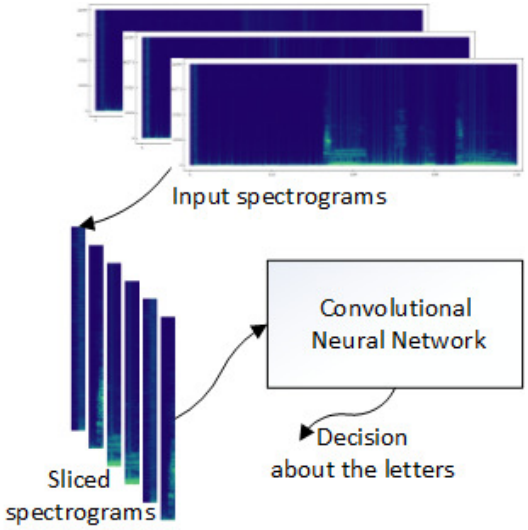


Fig. 1: Model of the proposed technique of converting the voice to the text.

equation is not used, and its recursive form called Fast Fourier Transform is used. It is defined as follows

$$\begin{aligned} S_k &= \sum_{n=0}^{N-1} s_n \exp\left(-\frac{2i\pi nk}{N}\right) \\ &= \sum_{m=0}^{N/2-1} s_{2m} \exp\left(-\frac{2i\pi km}{N/2}\right) \\ &+ \exp\left(-\frac{2i\pi k}{N}\right) \sum_{m=0}^{N/2-1} s_{2m+1} \exp\left(-\frac{2i\pi km}{N/2}\right). \end{aligned} \tag{2}$$

It is possible to present a sound sample with the help of an image so-called spectrogram. It is a flattened three-dimensional graph that is spanned on two axes – time and frequency. The flattened dimension is the intensity, which is depicted by the shadow of a given color. The formula for that is defined as

$$spectrogram\{s(t)\}(t,f) \equiv |S(t,f)|^2, \tag{3}$$

where $S(\cdot)$ is a short-time Fourier transform understood as

$$S(m,f) = \sum_{n=-\infty}^{\infty} s[n]w[n-m]\exp(-jfn), \tag{4}$$

where $s[n]$ is a signal in discrete form, $w(\cdot)$ is a window function like sine window described as

$$w(n) = \sin\left(\frac{\pi n}{N-1}\right) \tag{5}$$

B. Convolutional Neural Network

Convolutional neural network are an example of neural structures adapted to receive graphic images in the input [17]. The idea of operation this type of classifier is based on the cells in the primary cortex. The network structure is composed of three types of layers. First type is convolution layer where some feature are extracted from input image. In practice, some filter ω is applied to the image (for instance blur or sharpening defined as a matrix of 3×3 with step size S). Output from this layer is an image called feature map. The second type of layer is pooling, which reduces the size of the incoming image by calculation the mean, maximum or minimum value in a given neighborhood area. The third type is called fully connected and the architecture of these is similar to classical neural network structure. Each pixel returned from the last layer is considered as a single neuron which form the input layer. Then, hidden and one output layers are created.

These type of structure can be trained using backforward propagation algorithm [18], [19]. Let me introduce some designation – $f(\cdot)$ as an error function, output value from neuron at position (i,j) in l layer as $\frac{\partial f}{\partial y_{ij}^l}$. The error at the end of the network is known and marked as $\frac{\partial f}{\partial y_{ij}^l}$. Algorithm is based on chain rule what is understood as sharing weight with one another can be defined as

$$\frac{\partial f}{\partial \omega_{ab}} = \sum_{i=0}^{N-m}\sum_{j=0}^{N-m} \frac{\partial f}{\partial x_{ij}^l}\frac{\partial x_{ij}^l}{\partial \omega_{ab}} = \sum_{i=0}^{N-m}\sum_{j=0}^{N-m} \frac{\partial f}{\partial x_{ij}^l} y_{(i+1)(j+b)}^{l-1}. \tag{6}$$

Using above equation, error $\frac{\partial f}{\partial x_{ij}^l}$ can be calculated as

$$\frac{\partial f}{\partial x_{ij}^l} = \frac{\partial f}{\partial y_{ij}^l}\frac{\partial y_{ij}^l}{\partial x_{ij}^l} = \frac{\partial f}{\partial y_{ij}^l}\frac{\partial\left(\sigma(x_{ij}^l)\right)}{\partial x_{ij}^l} = \frac{\partial f}{\partial y_{ij}^l}\sigma'(x_{ij}^l), \tag{7}$$

where $\sigma(x)$ is activation function in classic reasoning. Having defined a formula for an error on the current layer, it is necessary to define formula for an error in earlier layers. Note that the gradient for the convolutional layer can be determine as

$$\begin{aligned} \frac{\partial f}{\partial y_{ij}^{l-1}} &= \sum_{a=0}^{m-1}\sum_{b=0}^{m-1} \frac{\partial f}{\partial x_{(i-a)(j-b)}^l}\frac{\partial x_{(i-a)(j-b)}^l}{\partial y_{ij}^{l-1}} = \\ &\sum_{a=0}^{m-1}\sum_{b=0}^{m-1} \frac{\partial f}{\partial x_{(i-a)(j-b)}^l}\omega_{ab}, \end{aligned} \tag{8}$$

and this equation can be used to define error as

$$\frac{\partial x_{(i-a)(j-b)}^l}{\partial y_{ij}^{l-1}} = \omega_{ab}. \tag{9}$$

It worth to note, that algorithm does not work for pooling layer which are skipped during training process.

C. Proposed technique for object manipulation in augmented reality

The built-in microphone on smartphone or tablet can record the sound in real time. Assume that the obtained audio will be saved every 2 seconds, it is possible to represent such a sample as spectrogram that will be cut every 0.5 seconds (these values are chosen in an empirical way). As a result, four graphics will be obtained, which will be subjected to training

and classification through a convolutional neural network (classical 5x5 filters were used – Gaussian blur and emboss). A schematic model is presented in Fig. 1.

This action allows to classify the analyzed spectrogram in order to change the sound into a text form. The text is simple in order to pass the parameter to the program, and more accurately enables the voice manipulation of the object placed in the augmented reality.

TABLE I: Values of statistical coefficients.

	0.1	0.01	0.001	0.0001
Γ	0.435	0.575	0.755	0.835
Λ	0.362	0.560	0.756	0.832
Ψ	0.221	0.388	0.608	0.713
Υ	0.416	0.581	0.752	0.845
Φ	0.447	0.570	0.758	0.825

III. Experiments

In order to test the proposed speech processing techniques and object manipulation, a simple model was created and placed on the screen of smartphone. The classifier was trained with 150 samples (75 per command) which were arranged in a random manner in the ratio of 70 : 30 (learn:test samples). The classifier was trained to obtain an error of 0.1, 0.01, 0.001 and 0.0001. The correctness of classification with respect to the error has been presented in Fig. 2-5. The best results were achieved for the smallest error, screenshots from application where these proposition was implemented are shown in Fig. 6. These results allow to calculate some statistical coefficient like accuracy Γ, Dice's coefficient Λ, overlap Ψ, sensitivity Υ, specificity Φ and calculated values of these parameters are presented in Tab. I. It shows that the accuracy increases very fast, which is a good indicator. Other values also increase, such as sensitivity, which means the probability of indicating that this is a mistaken command among all erroneous commands. Again, the coefficient of specificity points to value of incorrect samples that had a negative result.

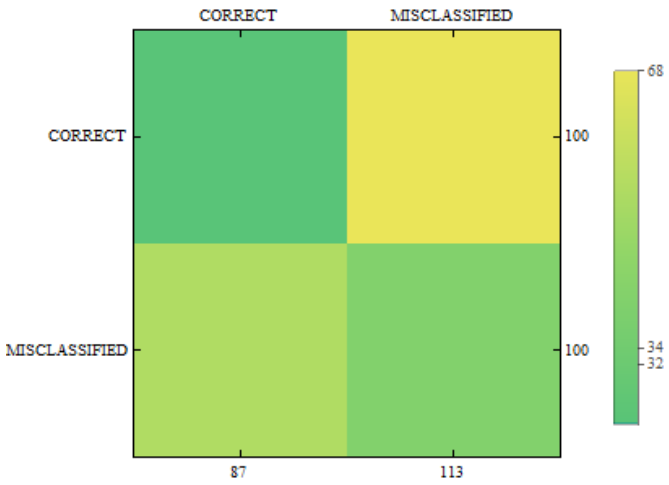


Fig. 2: Accuracy in relation to obtained error 0.1.

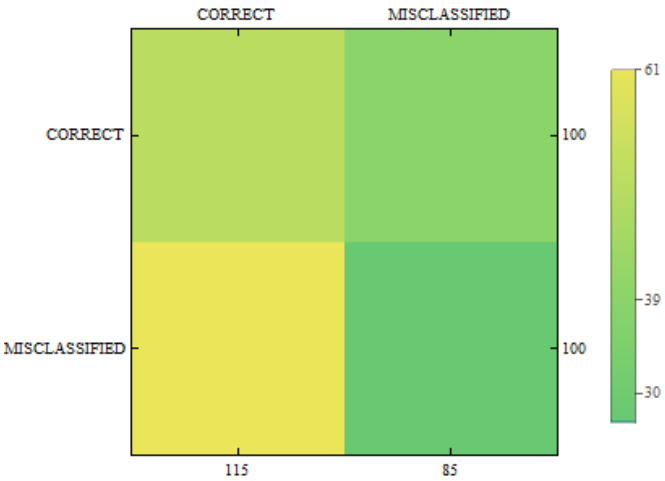


Fig. 3: Accuracy in relation to obtained error 0.01.

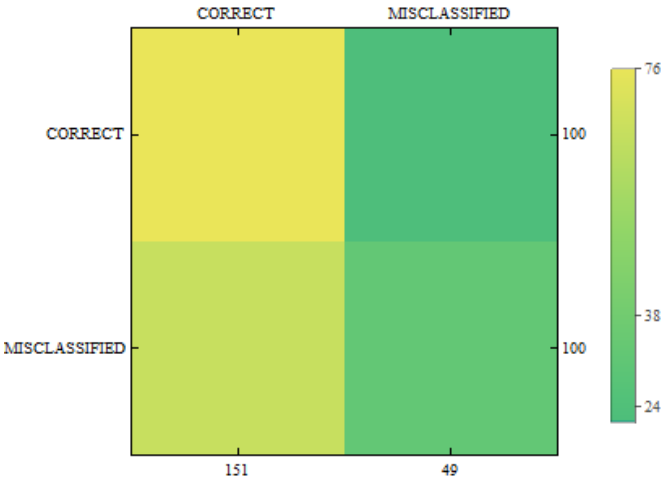


Fig. 4: Accuracy in relation to obtained error 0.001.

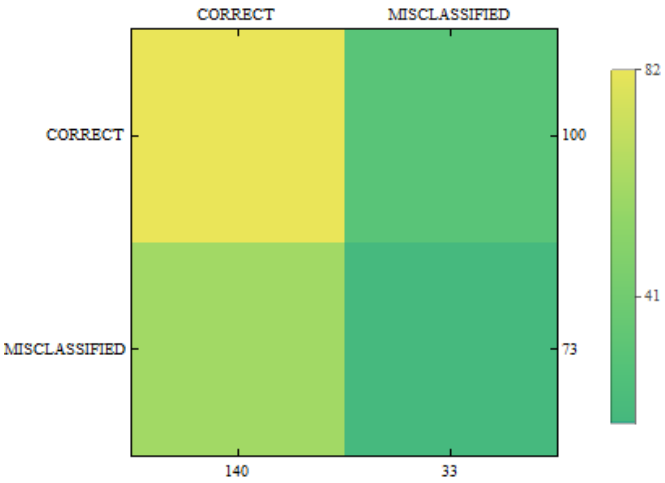


Fig. 5: Accuracy in relation to obtained error 0.0001.

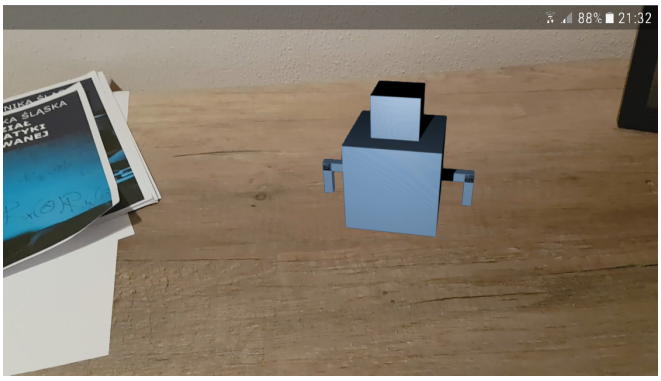
IV. Conclusion

The proposed solution can diversify, and above all, improve user interaction with artificially created objects. This type of solution is the first approach to this type of activity. The obtained results indicate a high potential, however, the proposed technique has several parameters that should have been taken into account. Particularly problematic is saving audio sample every few seconds, loading processing as well the length of sliced elements from spectrograms.

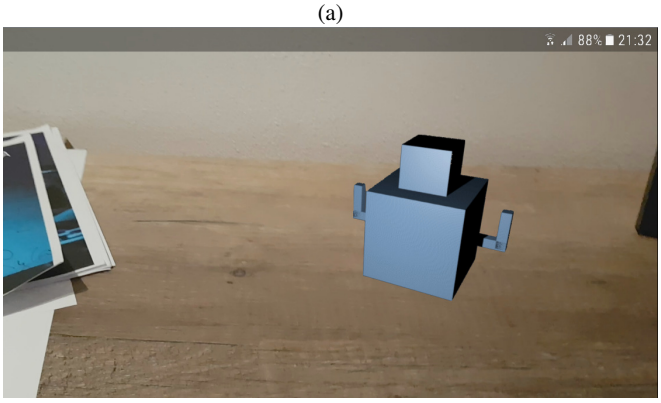
In this paper, two simple commands such us *UP* and *DOWN* were tested. The trained classifier allowed to correct manipulation of the object. The effectiveness of this model indicates the high flexibility of use in games like *Pokemon Go*, which can increase the playability and refresh the classic operation of augmented reality technology. In my future work, we plan to focus on improving these issues as well as improving the classifier's operation on longer text commands.

V. Acknowledgment

Author acknowledge contribution to this project of the "Diamond Grant 2016" No. 0080/DIA/2016/45 from the Polish Ministry of Science and Higher Education and the Rector pro-quality grant No. 09/010/RGJ18/0033 at the Silesian University of Technology, Poland.



(a)



(b)

Fig. 6: Screenshots of the application's operation – (a) object in the initial state, (b) object after executing the sound command "*UP*".

References

[1] A. G. LeBlanc and J.-P. Chaput, "Pokémon go: A game changer for the physical inactivity crisis?" *Preventive medicine*, vol. 101, pp. 235–237, 2017.

[2] B. Morschheuser, M. Riar, J. Hamari, and A. Maedche, "How games induce cooperation? a study on the relationship between game features and we-intentions in an augmented reality game," *Computers in human behavior*, vol. 77, pp. 169–183, 2017.

[3] S. Chodarev, "Development of human-friendly notation for xml-based languages," in *Computer Science and Information Systems (FedCSIS), 2016 Federated Conference on*. IEEE, 2016, pp. 1565–1571.

[4] Z. Sroczyński, "Actiontracking for multi-platform mobile applications," in *Computer Science On-line Conference*. Springer, 2017, pp. 339–348.

[5] L. Wang, F. Forni, R. Ortega, Z. Liu, and H. Su, "Immersion and invariance stabilization of nonlinear systems via virtual and horizontal contraction," *IEEE Transactions on Automatic Control*, vol. 62, no. 8, pp. 4017–4022, 2017.

[6] C.-M. Wu, C.-W. Hsu, T.-K. Lee, and S. Smith, "A virtual reality keyboard with realistic haptic feedback in a fully immersive virtual environment," *Virtual Reality*, vol. 21, no. 1, pp. 19–29, 2017.

[7] D. Cho, J. Ham, J. Oh, J. Park, S. Kim, N.-K. Lee, and B. Lee, "Detection of stress levels from biosignals measured in virtual reality environments using a kernel-based extreme learning machine," *Sensors*, vol. 17, no. 10, p. 2435, 2017.

[8] S. Deniziak, T. Michno, and P. Pieta, "Iot-based smart monitoring system using automatic shape identification," in *Federated Conference on Software Development and Object Technologies*. Springer, 2015, pp. 1–18.

[9] J. Protasiewicz, W. Pedrycz, M. Kozłowski, S. Dadas, T. Stanisławek, A. Kopacz, and M. Gałężewska, "A recommender system of reviewers and experts in reviewing problems," *Knowledge-Based Systems*, vol. 106, pp. 164–178, 2016.

[10] M. Sołtysiak, D. Dabrowska, K. Jałowiecki, and V. Nourani, "A multi-method approach to groundwater risk assessment: a case study of a landfill in southern poland," *Geological Quarterly*, vol. 62, no. 2, pp. 361–374, 2018.

[11] V. Nourani, G. Andalib, and D. Dabrowska, "Conjunction of wavelet transform and som-mutual information data pre-processing approach for ai-based multi-station nitrate modeling of watersheds," *Journal of hydrology*, vol. 548, pp. 170–183, 2017.

[12] K. Li, S. Mao, X. Li, Z. Wu, and H. Meng, "Automatic lexical stress and pitch accent detection for l2 english speech using multi-distribution deep neural networks," *Speech Communication*, vol. 96, pp. 28–36, 2018.

[13] R. Maskeliunas, V. Raudonis, and R. Damasevicius, "Recognition of emotional vocalizations of canine," *Acta Acustica united with Acustica*, vol. 104, no. 2, pp. 304–314, 2018.

[14] R. Shadiev, T.-T. Wu, and Y.-M. Huang, "Enhancing learning performance, attention, and meditation using a speech-to-text recognition application: Evidence from multiple data sources," *Interactive Learning Environments*, vol. 25, no. 2, pp. 249–261, 2017.

[15] A. Venckauskas, A. Karpavicius, R. Damaševičius, R. Marcinkevičius, J. Kapočiūte-Dzikiené, and C. Napoli, "Open class authorship attribution of lithuanian internet comments using one-class classifier," in *Computer Science and Information Systems (FedCSIS), 2017 Federated Conference on*. IEEE, 2017, pp. 373–382.

[16] M. S. Elmahdy and A. A. Morsy, "Subvocal speech recognition via close-talk microphone and surface electromyogram using deep learning," in *Computer Science and Information Systems (FedCSIS), 2017 Federated Conference on*. IEEE, 2017, pp. 165–168.

[17] M. Matsugu, K. Mori, Y. Mitari, and Y. Kaneda, "Subject independent facial expression recognition with robust face detection using a convolutional neural network," *Neural Networks*, vol. 16, no. 5-6, pp. 555–559, 2003.

[18] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 1717–1724.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.