Utilizing Voice Assistance and Wearable Interactive Mixed Reality Solutions for Industrial Training Workflows

2022 IEEE 4th Eurasia Conference on IOT, Communication and Engineering (ECICE) | 978-1-6654-8208-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/ECICE55674.2022.10042870

R. S. D. Putera*
Institute of Industrial Engineering and Management,
National Formosa University
Yunlin, Taiwan 632
rafly.s.putera@gmail.com

T. M. Cheng
Institute of Industrial Engineering and Management,
National Formosa University
Yunlin, Taiwan 632
rtmc@nfu.edu.tw

***Abstract*—Field workers under training often need assistance locating, recognizing, and familiarizing themselves with different machines and equipment. Whenever further guidance is needed, an expert is expected to assist using visual and auditory means. Thus, we employed an intelligent voice assistant system by integrating Augmented Reality (AR) and Natural Language Processing (NLP) methods to alleviate workflows and construct an interactive approach for field workers in training or practice. Microsoft HoloLens 2 was used as the Head Mounted Display (HMDs) to overlay necessary AR visuals and speech synthesis for user assistance. for its implementation, the multinomial Naive Bayes classification algorithm was performed at different sequential run-times throughout a standard operating procedure. This resulted in preserving training speeds and maintaining high accuracy when each and any unique attribute of a mechanical part is uttered. As the objective is to design an interactive training solution, further development in this research can be made by enhancing the user interface, interaction, and overall experience.**

Keywords—Natural Language Processing, Augmented Reality, Human-Computer Interaction, Smart Voice Assistant, Microsoft HoloLens 2

I. Introduction

With the increase of smart factories operating and the shortage of skilled labor growing, difficulties are bound to happen during any company's hiring process. Moreover, grasping and familiarizing workloads on varieties of tasks leave new workers perplexed as on-the-job skills and knowledge are still yet to be mastered. Although necessary training is conducted for workers beforehand, further assistance is still needed for guidance when complex problems occur. As demands for assistance technologies grow under the influence of Industry 4.0 [1], Supporting tools such as voice assistance (VA) and augmented reality (AR) is seen implemented on tasks requiring procedural support [2] which can be beneficial for training purposes.

A. Voice Assistance

VAs are programs that can be integrated into devices such as computers or smartphones to understand human speech and respond to users using synthetic voices [3]. In contrast to the traditional approach of clicking or dragging a button, the ability to interact with speech revolutionizes communication between users and their computing systems [4]. VAs has been used in many different areas of application ranging from carrying out simple tasks like fetching the weather report [5] to even controlling autonomous vehicles [6]. As VAs enables an alternative approach to the interaction between users and their computers, implementations provide further convenience for several applications in an industrial environment.

B. Augmented and Mixed Reality

AR is a growing technology that allows virtual information generated by computers to be superimposed onto an actual environment or scene during user operation [7]. AR applications have been emerging with implementations being continuously adopted across various industries such as education [8], manual assembly and manufacturing [9], and maintenance [10]. With its usage expanding, the benefits from integrating have been observed to assist users in understanding complex and trivial problems. AR systems are expected to communicate effectively through procedural projections of AR instructions when dealing with manual assemblies or maintenance tasks. Moreover, AR systems also play an active role in providing visual pointers to meet the user's cognitive needs [11]. Virtual objects shown in a given AR system can be operated on devices such as computers, mobile phones, and wearable head-mounted displays (HMDs). Using Microsoft's 2nd generational mixed reality (MR) headsets, Microsoft Hololens 2, AR objects are superimposed onto the display of the headsets engulfing the user's peripheral vision. By doing so, computer-generated information can then be combined with live feeds of real-world environments.

C. Speech Interactions with AR

With instruction-based AR systems, cognitive and communication efficiency improvements can be attained [12]. Furthermore, by incorporating an additional module where speech recognition is equipped with artificial intelligence (AI) algorithms, a multi-modal human-computer interaction (HCI) can be achieved. As Microsoft Hololens 2 accommodates the input of speech along with its main purpose to overlay virtual objects, integrating an intelligent voice assistance module delivers a more interactive approach when conducting AR-assisted training activities. As the system is assigned to recognize what the user says, it also needs to understand the context and meaning by classifying a user's utterance through natural language processing (NLP).

II. Related Work

Although a lot of research was done on AR head-wear for assistance through procedural instructions, few have yet to see the benefits of incorporating a voice assistant system into it. Efforts such as exploring AR in assistance and training [13], as well as implementations for its usage on an aseptic bottling line [14], provide several benefits when working with workers. As for using voice interactions with AR, a study was conducted on its usage for car components

978-1-6654-8208-0/22/$31.00 ©2022 IEEE

assembly [15] using windows speech recognition application programming interface (API) where voice interaction was proven to be an effective method for tasks in AR applications as it allows the manipulation of 3D objects while being hands-free. In addition, another study where natural language understanding (NLU) was integrated for the use of AR navigation interfaces [16] showed that not only it proves to be more flexible but also accurate when compared to all speech-to-text approaches.

III. Recognizing User Speech

Before processing speech as text inputs onto a system, sentences or commands uttered by users must first be recognized. With windows speech recognition API, developers can take advantage of applying a straightforward approach to recognizing human speech. As experiments were done on Unity's game engine, three classes of recognizers were provided to be of use depending on their purposes. out of the three, the dictation-recognizer was used as it matches the objective of this research in recognizing any utterances made by the user. Following with implementation of the dictation-recognizer class, Microsoft published a journal for modern speech recognition where its described techniques are still applied [17]. Microsoft expressed that a speech recognition system is composed of two primary domains: the communication channel and the speech recognizer. The communication channel revolves around the generation of speech and how its waveform is depicted and processed as a signal through a vocal apparatus. Within the speech recognizer, the processed signal is passed on to a speech decoder where it is decoded into a word sequence that can be understood. The system architecture of a speech recognition system can be seen in Fig. 1 showing how processes, applications, and models perform during the recognition of speech.

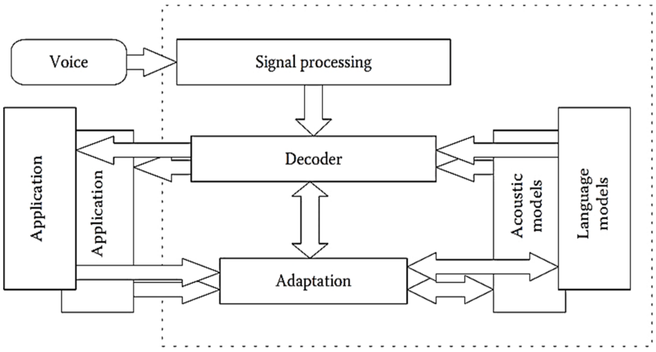


Fig. 1. Basic system architecture of a speech recognition system.

A signal-processing module extracts important feature vectors from a speech signal for it to be processed into a decoder module. Once the necessary input feature vectors are present, the decoder uses two types of models to generate a word sequence with the highest maximum posterior probability; language and acoustic. The adaptation component allows modifications to be made to any of the two models to improve further performance.

A. Acoustic and Language Models

Acoustic models are the essential part of achieving a reliable speech recognition system, its purpose is to determine what words have been said by a user through comparisons made in the model. As its purpose is to deliver recognized words with high accuracy, acoustic models are exposed to many factors such as different speaker variations, context variations, and environmental noise. The acoustic model is composed of phonetic features to match the input feature vectors. On the other hand, language models are composed of data that acts as a set of rules to be used when an input feature vector is awaiting to be decoded. The structure of a chosen language is laid out and built as a model where grammar and parsing techniques are adopted. Furthermore, analyzing methods are used through parsing techniques to verify that a given sentence complies with the grammar of the chosen model. Both models add up the probabilistic result where the decoder module eventually accepts the highest posterior probability of determining the given structural sentence

IV. Natural Language Processing

Naturally, speech recognition is a method where speech is converted to text. Several other processes like implementing deep learning or machine learning techniques onto a speech recognition system are usually done whenever one decides to build a responsive VA system. In this research, a machine learning approach was conducted using Natural Language Processing (NLP). Natural Language Processing is the analysis of linguistic data where it aims to create a representation of texts that adds structure to an unstructured natural language [18]. An NLP structure can be applied in two ways [18].

1) Syntactic: Captures grammatical relationships of different components in a text.

2) Semantic: Captures the conveyed meaning from a text.

A. Tokenizer, Sentence Splitters, and Part of Speech Tagger

A tokenizer also known as a process called tokenization means that the raw data of a sentence is converted into meaningful data strings. Sentences are segmented in this process where each word is then considered as a single entity. By doing so, a system understands words in a larger context when a large string of data is present such as sentences. In contrast to tokenization, sentence splitter is the proceeding process where extensive data of raw strings in a paragraph is split and categorized into a sentence.

Part of speech tagger also known as POS tags takes in recognized words and determines what part of speech they are. Tags used to identify given words are based on the Penn Treebank Project [19] which consists of a model containing approximately 7 million words of tags to classify adjectives, adverbs, nouns, and many more.

B. Topical Classifier

Topical Classifier is an NLP technique that extracts and comprehends a collection of text data by assigning a topic or category. With the breaking down of human language through the process of tokenization, sentence splitting as well as speech taggers, classification algorithms take place by finding patterns and using a semantic structure for a system to understand the given topic a user is talking about. As topic classification is a supervised machine learning technique, topics used as tags are predefined before any classifications are made. Due to this, using topical classification allows a system to understand user speech under a well-thought-out context. On the other hand, an unsupervised learning approach to this would be topic modeling which is less precise as it operates through clustering techniques.

V. Methodology and Experimentation

With NLP techniques used on speech recognition systems, its application can further on be implemented on Hololens 2 to improve AR instruction-based assistance through a multimodal interactive system with voice assistance capabilities. To do this, we proposed the Augmented Reality Voice Assistance system (ARVA) to provide interactive assistance for inexperienced workers in an industrial training environment. Procedures in ARVA's application begin when a user requests help to complete certain tasks or to remember the names of complicated industrial equipment and its whereabouts.

ARVA's workflow can be seen in Fig. 2 where the given diagram depicts the starting and ending process of how a user communicates with the system using speech as input. When human speech is detected as input by the system, the process of speech recognition using Microsoft's dictation recognizer begins. This process operates by comparing the captured essential feature vectors of the raw waveform with the acoustic and language models. Once the comparisons end with the highest probable result, the system then stores this as a string for it to be classified. after the speech is recognized, a topical classifier using the Multinomial Bayes classification algorithm is used. Predefined processes of topics consisting of keywords are written beforehand to determine what the worker needs assistance. Computations after running the algorithm result in a confidence level. If it is low, the system communicates through speech synthesis requesting another input. If the resulting confidence is high, the procedural instruction initiates along with the necessary speech synthesis needed when doing a task. This procedure repeats until a specific task is completed.

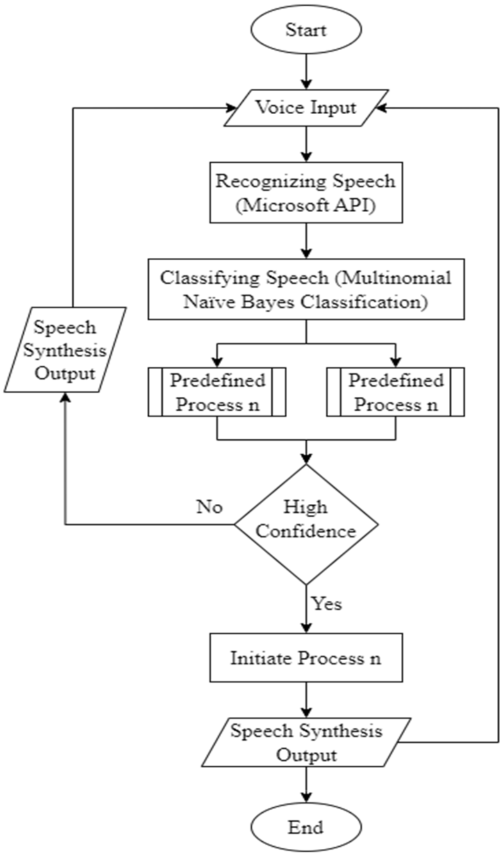


Fig. 2. ARVA's workflow diagram.

A. Multinomial Naïve Bayes Classification

Multiple algorithms accommodate the classification process needed. As we focus on alleviating training workflows with VA, the Multinomial Naive Bayes algorithm is adopted as it has several advantages under an industrial field of application. The algorithm is great for small sample sizes such as short sentences which makes it suitable for ongoing communication between the system and the user. Moreover, with small sample sizes, the computation cost is lower allowing for easier implementation and configuration as well as resulting in a quick and accurate approach to prediction.

Equation (1) shows the probabilistic model of naive Bayes.

$$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\, p(\mathbf{x} \mid C_k)}{p(\mathbf{x})} \tag{1}$$

where *P* represents the posterior probability, *C* represents outcomes or classes, and *x* represents a vector accompanied by *n* amount of features.

In the multinomial naive Bayes however, this equation is expressed in simple terms.

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}} \tag{2}$$

$$p(\mathbf{x} \mid C_k) = \frac{(\sum_{i=1}^{n} x_i)!}{\prod_{i=1}^{n} x_i!} \prod_{i=1}^{n} {p_{ki}}^{x_i} \tag{3}$$

With the multinomial model, classification can be achieved as it takes in the frequency of the features present in a text, where *x*, is a feature vector as a histogram. With the algorithm being lightweight, training a model based on the selected keywords is not time-consuming. It is remarkably fast enough to where the training and classification process was executed during the system's run-time

A predefined process consists of a training model which holds a group of keywords that makes up a topic or class. Once the speech is recognized, a process initiates training its model and eventually applies it to the recognized speech as input to classify and predict its class. This process is repeated throughout the entire procedure for workers requiring assistance during training. As shown in Fig. 3, *C* is represented as classes where it is defined by a set of *n*, *T* represents the set of training models, and *n* represents an iterative number of the following process.

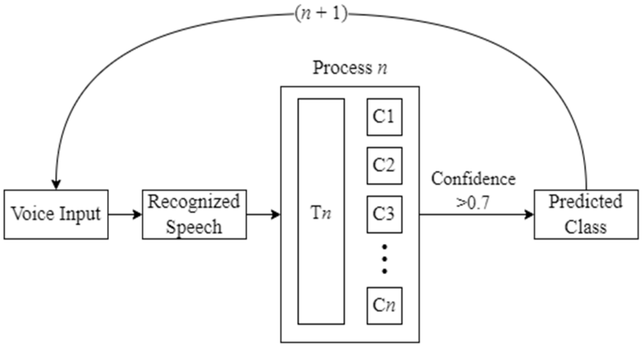


Fig. 3. Procedural process classification.

B. ARVA System Architecture

ARVA was built using two main components, Microsoft Hololens 2 and Unity engine. ARVA's components are separated into two segments, one accommodating the system hardware and the other being the application software. Within the Hololens domain, multiple peripherals are present such as a microphone, camera, and all the necessary sensors needed to track and register the user's movement. On the other hand, the unity engine holds all the necessary functions and processes to make use of the hardware. This is represented as a high-level system architecture of ARVA (Fig. 4).

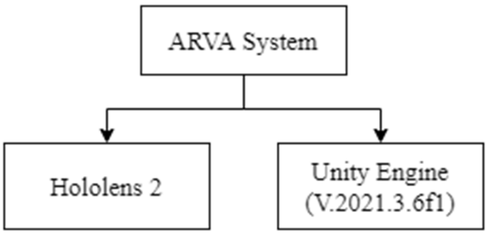


Fig. 4. High-level system architecture of ARVA.

These two domains are further broken down into a lower-level architecture which helps to understand what components and functions are being used. As shown in Fig. 5, all the necessary components and functions are needed to build the system. Each of the modules between both domains communicates with one another such as recognizing speech from a microphone, classifying the given speech, and eventually using the built-in speakers to output any speech synthesis made. With ARVA, displaying computer-generated objects aims to increase user experience as well as satisfaction. This can be achieved with a well-thought-out user interface as well as a seamlessly smooth interaction between users and the system since Hololens is equipped with well-developed tracking sensors and displays.

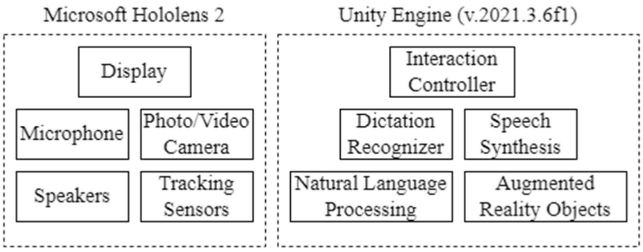


Fig. 5. Medium-level system architecture of ARVA.

C. Experimentation

In the experiment, we created a schematic that mimicked a training procedure where tasks like maintenance and assembly were present. As illustrated in Fig. 6, two processes were implemented. The 1st process attempted to recognize what the worker requires assistance with, this fell under the category of either maintenance or assembly. Once the 1st process was completed, the system waited for the next phrase of words the worker says. With assembly and maintenance having their schema, the second process consisted of classes where a worker's tasks were conducted. For example, a worker needed assistance in performing maintenance for machines A, B, or C. Likewise, this applied to the assembly schema as well. The topical classification was not used in the first process to predict a class as it was direct. On the other hand, the next process consisted of a training model holding different characteristics that described each of both maintenance and assembly classes.

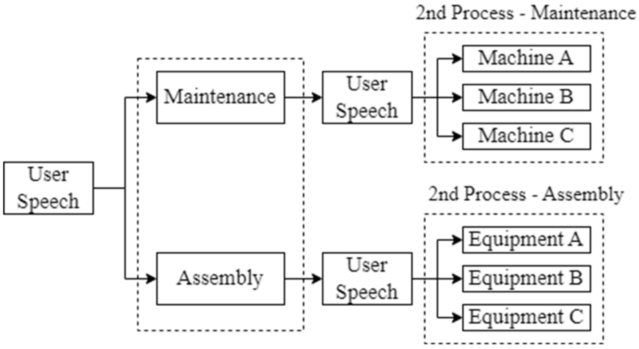


Fig. 6. Experiment workflow scenario of ARVA.

The training data used in this method were stored as a JSON file and seen in the following tables. Attributes shown in the tables described the machines' physical properties. This was because, during training procedures, inexperienced workers were prone to forget the complex names of several industrial objects. By doing this, workers could say the physical aspects of an object whereas the system determined and predicted that particular object (Tables I and II).

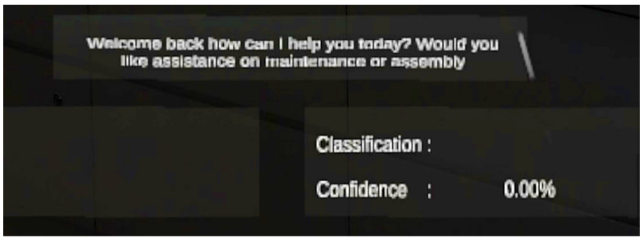
TABLE I. ATTRIBUTE DESCRIBING EACH MACHINE OBJECT

MACHINE		
A	*B*	*C*
Section A Circuits Bulky Green	Section B Pistons Metal Compact	Section C Hydraulics Rubber Bulky

TABLE II. ATTRIBUTE DESCRIBING EACH EQUIPMENT OBJECT

EQUIPMENT		
A	*B*	*C*
Compact Round Metal Thick	Compact Rectangular Glass Fragile	A and B

With the program executing its training and prediction during runtimes, three initial screenshots were captured with the Hololens 2. A simple user interface was made to accommodate essential features such as what the user has said, the system's speech synthesis, and finally the classification window showing its predicted class and confidence (Fig. 7).



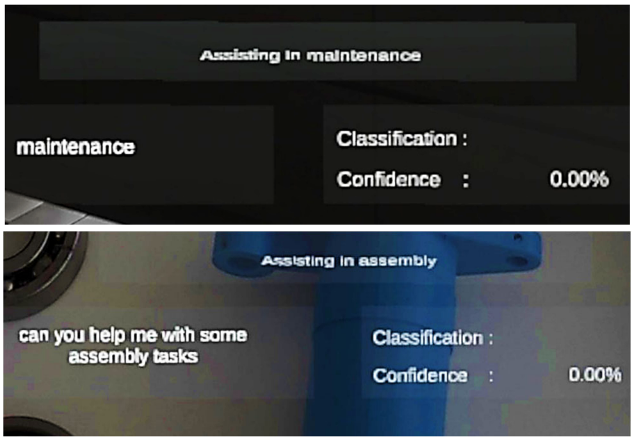
Welcome back how can I help you today? Would you like assistance on maintenance or assembly

Classification :

Confidence : 0.00%

Fig. 7. ARVA on standby for user procedure selection.

Using a straightforward approach containing keywords of "maintenance" and "assembly", users simply uttered the keyword or form a complete sentence to proceed (Fig. 8). Speech made by the user was displayed on the bottom left whereas speech given by the system is placed above. Predicted labels as well as their computed confidence was seen in the right-hand corner.



Assisting in maintenance

maintenance

Classification :

Confidence : 0.00%

Assisting in assembly

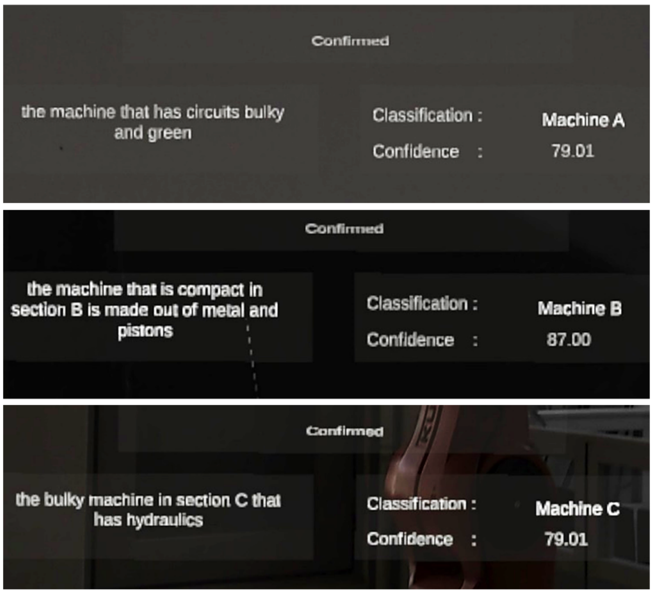
can you help me with some assembly tasks

Classification :

Confidence : 0.00%

Fig. 8. Selection of the maintenance and assembly procedure.

After the 1st process of prompting the user to choose between executing a maintenance or assembly task, the system awaited the next set of sentences to be classified. Figure 9 shows three classified labels of machines A, B, and C along with the sentences that were uttered. To achieve the correct classifications of both machines A and B, three of their physical attributes were used. This resulted in a confidence percentage of 79.1%. For Machine B, a confidence of 87.0% was obtained due to mentioning all physical properties of the machine.



Confirmed

the machine that has circuits bulky and green

Classification : Machine A

Confidence : 79.01

Confirmed

the machine that is compact in section B is made out of metal and pistons

Classification : Machine B

Confidence : 87.00

Confirmed

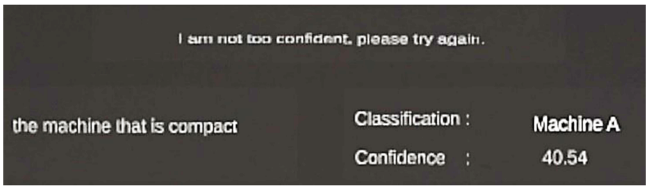
the bulky machine in section C that has hydraulics

Classification : Machine C

Confidence : 79.01

Fig. 9. Classification of machines in the maintenance procedure.

In a case where the exact name of either machines A, B, or C was uttered, the system executed the same approach of proceeding to any further process similar to the first process of selection. On the other hand, if the name of a machine was not mentioned or too few of its physical properties were uttered, the system failed to achieve a high confidence classification. As seen in Fig. 10, the system failed to predict Machine B where one of its physical properties was "compact" and resulted in a confidence level of 43.54%. Whenever too few essential keywords were mentioned in the system, classifications with high confidence were harder to achieve. The worker needed to be able to specify two or more of a given object's physical properties when using this training model. Moreover, for a predicted class to be accepted and continue on any further process, a confidence level above 70.0% must be obtained. If not, the system then prompted the user to repeat their sentence.



I am not too confident, please try again.

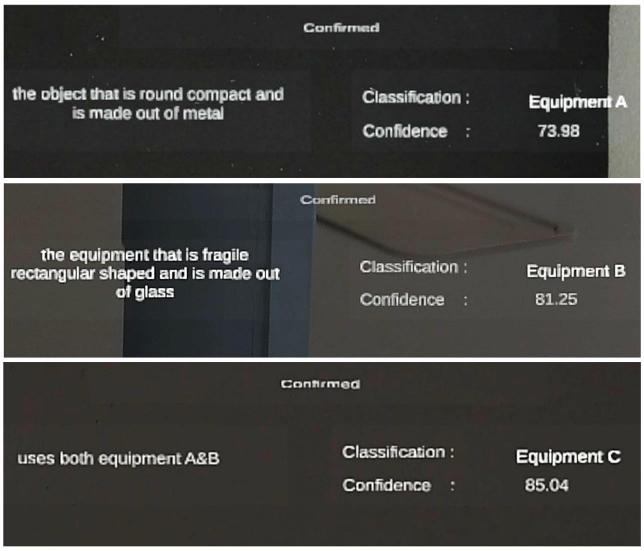
the machine that is compact

Classification : Machine A

Confidence : 40.54

Fig. 10. Output of low confidence predicted class.

Similarly, classifying the equipment in the assembly procedure followed the same pattern. High confidence in classifications was obtained for all equipment by mentioning more than two physical properties. However, unlike the maintenance procedure, one of the classes in the assembly schema was dependent on two other classes: equipment A and B. By doing this, workers identified equipment with their subpart which allowed a more complex schema to be made. Figure 11 shows each classification, confidence level, and what the user uttered when performing an assembly procedure.



Confirmed

the object that is round compact and is made out of metal

Classification : Equipment A

Confidence : 73.98

Confirmed

the equipment that is fragile rectangular shaped and is made out of glass

Classification : Equipment B

Confidence : 81.25

Confirmed

uses both equipment A&B

Classification : Equipment C

Confidence : 85.04

Fig. 11. Classification of equipment in the assembly procedure.

VI. Discussion

Differences in scaling were present when the program was being conducted on Unity compared to when the application was running. Views from the person using the Hololens 2 were closer and brighter compared to all the screenshots used and taken above. Furthermore, the implementation of NLP was feasible on an HMD by using the multinomial Naive Bayes classification algorithm. The benefits of adopting this method are its quick processing time when training a model and its great usage when predicting small sample sizes of sentences. This allows workers to have ongoing communication with their system. Although it has its fair share of advantages, this model achieves a high confidence level when too many keywords are present thus lowering its accuracy of prediction. Therefore, when applied, keywords used for training should only be those that are essential to maintain high resulting accuracy.

VII. Conclusion

With the rapid growth of the 4th industrial revolution, much training is needed for inexperienced workers. ARVA, a system where NLP techniques are integrated into an AR headset allows an interactive and reliable approach to instructional guidance for training tasks. Using this proposed method, workers can familiarize with remembering and understanding complex machines and equipment. To enhance the user experience for workers under training, improvements can always be made to this research by building a better user interface, interaction, and user experience.

REFERENCES

[1] E. Ras, F. Wild, C. Stahl, and A. Baudet, "Bridging the skills gap of workers in Industry 4.0 by Human Performance Augmentation Tools," Proceedings of the 10th International Conference on Pervasive Technologies Related to Assistive Environments, 2017.

[2] The next generation of search: Voice OK google, how can my site be the single answer? - seoclarity. [Online] Available: https://go.seoclarity.net/hubfs/docs/research/seoclarity_whitepaper_next-generation-search-voice.pdf

[3] M. B. Hoy, "Alexa, Siri, Cortana, and more: An introduction to voice assistants," Medical Reference Services Quarterly, Vol. 37, No. 1, Pp. 81–88, 2018.

[4] K. Rybinski and E. Kopciuszewska, "Will artificial intelligence revolutionize the student evaluation of teaching? A big data study of 1.6 Million student reviews," Assessment & Evaluation in Higher Education, Vol. 46, No. 7, Pp. 1127–1139, 2020.

[5] H. Segi, R. Takou, N. Seiyama, T. Takagi, Y. Uematsu, H. Saito, and S. Ozawa, "An automatic broadcast system for a weather report radio program," IEEE Transactions on Broadcasting, Vol. 59, No. 3, Pp. 548–555, 2013.

[6] G. Lugano, "Virtual assistants and self-driving cars," 2017 15th International Conference on ITS Telecommunications (ITST), 2017.

[7] E. Dubois and L. Nigay, "Augmented reality," Proceedings of DARE 2000 on Designing augmented reality environments - DARE '00, 2000.

[8] D. Chytas, E. O. Johnson, M. Piagkou, A. Mazarakis, G. C. Babis, E. Chronopoulos, V. S. Nikolaou, N. Lazaridis, and K. Natsis, "The role of Augmented Reality in Anatomical Education: An overview," Annals of Anatomy - Anatomischer Anzeiger, Vol. 229, P. 151463, 2020.

[9] A. Y. C. Nee and S. K. Ong, "Virtual and augmented reality applications in manufacturing," IFAC Proceedings Volumes, Vol. 46, No. 9, Pp. 15–26, 2013.

[10] Q. Loizeau, F. Danglade, F. Ababsa, and F. Merienne, "Evaluating added value of augmented reality to assist aeronautical maintenance workers—experimentation on on-field use case," Virtual Reality and Augmented Reality, Pp. 151–169, 2019.

[11] Z. Wang, X. Bai, S. Zhang, W. He, X. Zhang, L. Zhang, P. Wang, D. Han, and Y. Yan, "Information-level AR instruction: A novel assembly guidance information representation assisting user cognition," The International Journal of Advanced Manufacturing Technology, Vol. 106, No. 1-2, Pp. 603–626, 2019.

[12] Z. Wang, X. Bai, S. Zhang, M. Billinghurst, W. He, P. Wang, W. Lan, H. Min, and Y. Chen, "A comprehensive review of Augmented Reality-based instruction in manual assembly, training and Repair," Robotics and Computer-Integrated Manufacturing, Vol. 78, P. 102407, 2022.

[13] M. Moghaddam, N. C. Wilson, A. S. Modestino, K. Jona, and S. C. Marsella, "Exploring augmented reality for worker assistance versus training," Advanced Engineering Informatics, Vol. 50, P. 101410, 2021.

[14] E. Bottani, F. Longo, L. Nicoletti, A. Padovano, G. P. Tancredi, L. Tebaldi, M. Vetrano, and G. Vignali, "Wearable and interactive mixed reality solutions for fault diagnosis and assistance in manufacturing systems: Implementation and testing in an aseptic bottling line," Computers in Industry, Vol. 128, P. 103429, 2021.

[15] D. Aouam, S. Benbelkacem, N. Zenati, S. Zakaria, and Z. Meftah, "Voice-based augmented reality interactive system for car's Components Assembly," 2018 3rd International Conference on Pattern Analysis and Intelligent Systems (PAIS), 2018.

[16] J. Zhao, C. J. Parry, R. dos Anjos, C. Anslow, and T. Rhee, "Voice interaction for augmented reality navigation interfaces with natural language understanding," 2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ), 2020.

[17] X. Huang and L. Deng, "An overview of modern speech recognition," Microsoft Research, 17-Oct-2018. [Online]. Available: https://www.microsoft.com/en-us/research/publication/an-overview-of- modern-speech-recognition/.

[18] K. Verspoor and K. Cohen, ”Natural Language Processing”, Encyclopedia of Systems Biology, Pp. 1495-1498, 2013.

[19] A. Taylor, M. Marcus, and B. Santorini, "The penn treebank: An overview," Treebanks, Pp. 5–22, 2003.