IMAGE APPROACH TO ENGLISH DIGITS RECOGNITION USING DEEP LEARNING

*Fatin NA Zainol, Mohd Zamri Ibrahim**

Faculty of Electrical and Electronics Engineering Technology, Universiti Malaysia Pahang, Pekan Campus,
26600 Pekan, Pahang, Malaysia.
**E-mail: zamri@ump.edu.my*

Keywords: ARTIFICIAL NEURAL NETWORK, CONVOLUTIONAL NEURAL NETWORK, DEEP NEURAL NETWORK, MEL FREQUENCY CEPSTRAL COEFFICIENT, AUTOMATIC SPEECH RECOGNITION

Abstract

Despite good progress in speech recognition, various challenges still exist due to differences in how they speak, age, gender, emotions, and dialects when perceived by the ear. There is a proverb "I hear, and I forget; I see, and I remember". The image would be another solution to recognize what we hear. The main objective of this paper is to investigate the graphic method to learn digit English using the Deep Learning technique. In this work, Mel-frequency cepstral coefficients (MFCC) in the form of an image will be used as input to the system. Convolutional neural network (CNN) will be used to extract features from the image and an artificial neural network (ANN) will be used to classify those features into 10-digit English classes. By using the Speech Command dataset, the performance of the system will be compared with a conventional method that uses MFCC features in the form of a signal. The experiments showed that the image approach improves the recognition rate from 49% to 84%. It can be concluded that image approach can be used as an alternative method for digit recognition.

1 Introduction

Speech can be thought of as an asset or a need to communicate with machines. Even with the good progress, automatic speech recognition (ASR) still faces many challenges. These problems are due to a person's different characteristics (such as age, gender, and personality) that cause them to talk to different people differently. Most of the time, one can connect to the noise [1]. For five years, ASR has been a strong research center. It has been observed as an important bridge to promote good man-to-man and man-to-machine communication. In the past, language has not been an important part of human-machine communication. This was somewhat due to the technology at the time wasn't so good that user boxes would work for most real users under most real usage conditions, and partly because of the different ways. Many different speech modalities, such as keyboard and mouse, are better at speaking in terms of speech efficiency, flexibility, and accuracy [2].

ASR applications are used in business and government products such as Alexa, Google speak, and Siri are few of the design things related to ASR [3]. Speech is always treated like an important biometric mode and has an expressive study concept. It is found that although no modeling is used, maintaining a large and enough speakers can be a new challenge in the future to obtain the most accurate results. To communicate with machines, people can use language as a resource. Men need to complete natural counting, capture, and equality. Elham S. examined the influence of eyewitnesses on the implementation of the language recognition system of human confusion [4]. Voice recognition is allowed when text is converted from speech. Due to the importance of language knowledge, the backpropagation algorithm is widely used in artificial neural technology (ANN) today and has been developed for a direct application, for example, basic knowledge and information organization, through the process. Neural systems have seen a trend of interest in recent years and have been successfully applied over a wide range of problem areas, in various areas such as impact, design, topography and economic science. Nowadays, most of the language technology vendors work with users to understand the critical human issues related to product usage and application. The development of certain segments of the interface market will reflect the growth trends of the interface product. However, requests to get to know a person by their voice can be a daunting task. In the forensic field, the environment and the things involved in the process are very different, depending on the professional applications. Forensic requests of verbal information should be conducted under the guise of caution. Speech samples are affected under a variety of conditions. The recording environment, the speaker's age, medical condition, and emotional state can all alter the cognitive and cognitive process. Important features of the language are needed to improve the performance of the ASR system.

Conceptually, speech can be seen accurately from a numerical wave. However, due to the large variability of the speech signal over time, it is better to do some of the features to reduce that variability using speech data. in the picture. Importantly, it eliminates various sensory sources, such as sound or non -sound and, if sounded, eliminates the effect of periodicity or pitch, amplitude of the excitation signal, and base frequency [5]. In many cases, MFCC has been chosen as one of the best ways to translate features into an audio information system. However, it has been observed that the performance of MFCC-based systems degrades drastically with changing noise levels and noise types [6]. Typical speech