I. INTRODUCTION

An automatic speaker verification (ASV) is a more challenging problem than the ordinary speaker verification [1]. Nowadays, there are many machine learning models those can synthesize text and turn it into speech with the same characteristics as the train samples [2]. In other hands, there is also machine learning models who can turn some character voices into another someone voice character [3], [4], those techniques also called as Logical Access (LA) as a speaker verification attack. But, there is also another attack technique use Playback-Recorded that called as Physical Access.

Based on that problem aforementioned, inspired us to build an Automatic Speaker Verification model that robust to attacking using d-vectors with Deep Neural Networks (DNNs). And for the rest, we propose a spoof detection system for identifying Logical Access.

Text-to-speech system nowadays are really easy to find. Deepvoice 3 is one of them it trained on different kind of datasets, such as VCTK, LJSpeech, NIKL, and JUST [2].

In other hands, [3] perform voice cloning system with just a few sample. This single system could learn to reproduce thousands of speaker identities, with less than half an hour of training data for each speaker

The general procedure of speaker verification consists of three phases: Development, enrollment, and evaluation. For development, a background model must be created for capturing the speaker-related information. In enrollment, the speaker models are created using the background model. Finally, in the evaluation, the query utterances are identified by comparing to existing speaker models created in the enrollment phase [5], [6], [7].

For a long time, Gaussian Mixture Model (GMM) has been used for many approaches in Speaker Verification [8], [9]. Another traditional model also purposed to work with speaker verification case such as SVM, i-vector, and HMM [9].

Some other new research also proposed DNNs to work with this problem [9], [10], [11], the output of DNNs is called d-vectors and they are claimed that d-vectors has better performance rather than another state-of-the-art method such as GMM-UBM, SVM, etc.

In [12] DNNs has employed to working with Text-Dependent Speaker Verification. At the end of their observation, they said that d-vectors outperform the i-vectors 14% Equal Error Rate (EER) for clean condition and 25% EER in noisy condition.

Another approaches also written in [9], [6],[13], the CNN also employed as a feature extractor they are claimed CNN also has better performance compared to state-of-the-art methods such as GMM and SVM.

However, there is also another interesting problem to solve such as speaker diarization [14], automatic speaker recognition [15], speech enhancement using Generative Adversarial Networks (GAN) [11], [16].

In some speech recognition problem, one of the crucial problems is the lack of data training. It would be lead into overfitting and impacted hard to tackle unseen data. In order to tackle this problem,[17] come and introduce some simple data augmentation.

In this paper, we propose an augmentation technique applied to state-of-the-art method trained with data and also working with MFCC as a low-dimensional feature. Deep Neural Networks (DNNs) model and Gaussian Mixture Model-Universal Background Model (GMM-UBM) was employed to work on this problem.

The rest of the paper is organized as follows: Section 2 describes the details of the dataset. Section 3 discusses the details of baseline GMM-UBM and DNNs. The result and analysis of the model performance is given in Section 4 followed by the conclusion in Section 5.