**Project Pipeline Write Up**

**Privacy Preserving Algorithms**

In order to build a Machine Learning (ML) model that can classify real voices from synthetically generated ones, real life voice data from speakers is required, along with certian information that may be helpful to the data analysts performing varying degrees of analysis. This data requires anonymization, and other procedures that protect data providing participants.

The main task is to develop a data pipeline that not only protects an employee/user's identity, but also anonymizes and preserves the privacy of the participants involved in training/ updating the ML and analysis. For this, a secure pipeline is required that does not risk compromising participant data.

When real voice data is collected from a sample of the population, it is labeled with certain criteria that could trace back and possibly compromise the identity of the data providers (speakers). The primary goals are to:

* Protect the identities of any speakers who are providing data.
* Preserve the privacy of individuals whom the classifier is intended for (e.g. enterprise employees).

The pipeline that addresses these goals will vary in its privacy preserving methods. Firstly, k-anonymization is applied to the speaker information in order to ensure that analytics can be provided without risking compromising any specific user information. Next, audio data is extracted and converted low resolution spectrograms. This step allows for saving the spectrogram image data without the fear of identity compromisation, since the images cannot be upscaled accurately, but will suffice to provide accurate analytics and features for a Convolutional Neural Network (CNN).

The final step is to fabricate misleading image data with popular oversampling techniques and combine them with the original real voices, while also generating an encrypted table that allows for seperating the synthetic oversampled data from the original. Figure (1.1) outlines the entire pipeline.

A picture containing diagram

Description automatically generated

Figure 1.1

**The Dataset**

For the demo provided, the data was extracted from two sources on kaggle. One source provided the real voice audio files from the "English Multi-speaker Corpus for Voice Cloning", (Fekadu 2019) and the synthetic voice audio files from the "Augmented Extended Train Robots", (Fabawi 2020).

The real voices consisted of about 44,000 audio files of varying lengths, while the synthetic voices consisted of approximately 187,000 audio files. The real voice data was recorded in English by 108 speakers from varying backgrounds, some native english speakers, and some non-native english speakers. The synthetic voice data was generated using Google cloud's "Text-to-Speech". To balance the data, 44,000 random samples of the synthetic voices were extracted.

The speaker information table for the real voice data was slightly anonymized, however, some speakers' anonymization was not handled correctly. Therefore, K-anonymization was applied.

**K-anonymity**

In order to rectify the lack of anonymization in the real voice data, k-anonymization was utilized using the Pandas library in Python. Originally, the data was represented by figure (1.2).

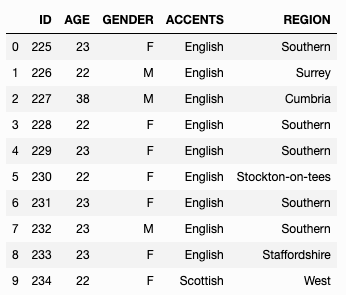


Figure 1.2

The following procedures were then applied to the table in order to further anonymize.

* The "Age" column was generalized into a single age group, since only 3 people were above the age of 29.
* The "Gender" column was kept as is, since the data contained 61 female and 47 male speakers.
* The "Accents" column was also generalized into two groups, native english speakers and non-native english speakers.
* The "Regions" column was generalized into US/Canada/EU/UK and Other.

The results of the k-anonymization, where k = 4, are represented in Figure (1.3).

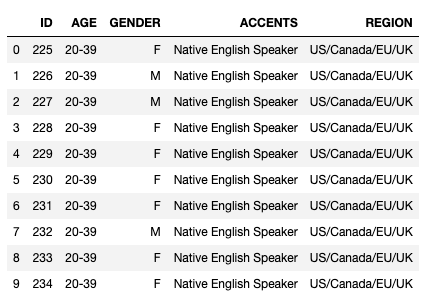


Figure 1.3

**Spectrogram Conversion and Low resolution blurring**

After anonymization, the next step is to convert the .wav files into image data. Spectrograms in this case are able to provide vital information required for data driven solutions without the need for high resolution images. Therefore, the images were saved after their low resolution conversion, which also preserves the voice identity of the speakers, since they cannot be upscaled. A sample of the different low resolution spectrograms can be seen in figure (1.4).

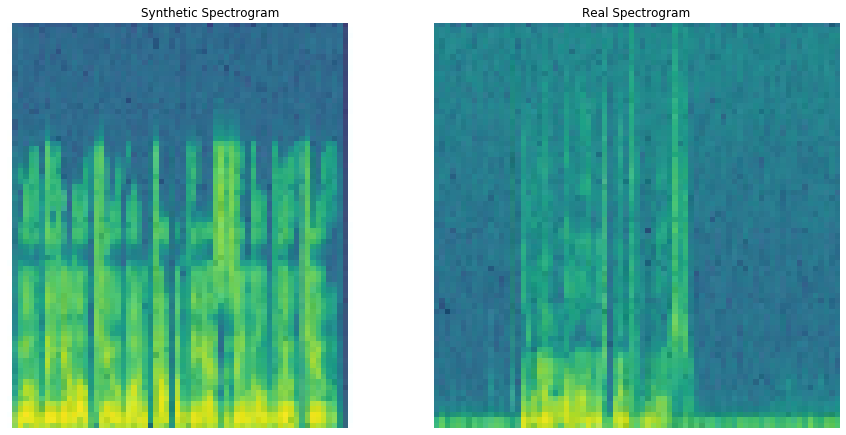


Figure 1.4

**Further Preservation of Privacy Using SMOTE & Encryption**

Talk about smote, encryption and other possible methods

**Data Science: Analysis and Machine Learning**

Brief introduction

**Descriptive Analysis**

**t-SNE Clustering Analysis**

**Machine Learning: Inception v3**

**Assumptions and Limits**

Phonetics, speaker variation,