**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

**REMINDER** Talk about data lifecycle diagram in this context.

**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).

Table

Description automatically generated

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).

Table

Description automatically generated

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 (a 75x75 resolution), 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).

A picture containing background pattern

Description automatically generated

Figure 1.4

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

At this stage, the data has already been safeguarded. However, with the new emerging AI technologies, upsampling images of this nature accurately may be possible in the near future (REFERENCE). Therefore, as an additional layer of privacy protection, there are various methods that could be applied to the spectrogram image data.

The first of these techniques employs the SMOTE (Synthetic Minority Oversampling TEchnique) algorithm. SMOTE is often used in cases where an imbalance in data exists, whereby it is used to generate synthetic data that is similar to existing data. However, in this case, one could oversample the real voice image data, combine it with the stored original data, then develop an encrypted reference table that is furthermore stored securely. This allows for misleading data to exist among the original data, should it be leaked at any point.

Another technique would be to encrypt the images themselves using the Rubix's Cube principle. This clearly allows for more secure storage, and could possibly be combined with other encryption and secure storage techniques.

**Data Science: Analysis and Machine Learning**

Now that the data has been extracted, processed and protected, the first goal of protecting the data participants has been achieved. The idenity protection system in the form of Aritificial Intelligence and data analysis achieves the final goal of safeguarding employee/user privacy. The following were the three primary steps conducted in achieving the goal.

**Descriptive Analysis**

In order to discover pattern differences between real and synthetic voices, an analysis on the different color channels of the spectrograms was conducted, revealing some significant RGB differences between the real and synthetic spectrograms. Figure (1.5) provides this result.

Histogram

Description automatically generated

Figure 1.5

It is vital to note that this type of data analysis clearly does not infringe upon a participant's privacy, and that any other type of data analysis (e.g. age, gender and specific demographic) should deeply consider any privacy risks associated.

**t-SNE Clustering Analysis**

One form of Machine Learning conducted in this study is t-SNE (t-distributed stochastic neighbor embedding). This algorithm is based on Stochastic Neighbor Embedding and is used primarily to visualize high-dimensional data. This type of analysis can provide significant insight without generating or comprosmising participant, employee or user data.

For this task the MobileNetV2 feature extraction model was used to extract non-personally identifable information from the low resolution spectrograms before t-SNE was applied. The results are represented by figure 1.6.

Note that at this stage, only 6,000 samples were utilitized in training and visualization, due to hardware constraints. However, as will be noted at a later stage, this sample does represent the entire data set.

Chart, scatter chart

Description automatically generated

Figure 1.6: t-SNE for 6,000 spectrogram samples

**Machine Learning: Inception v3**

Finally, the identity protection tool was developed in the form of a classifier. Google's Inception v3 model was selected with some additional customized layers. The full architecure of the model can be found in the appendix(??).

After a train-test split of 75-25%, and varying validation split of 20%, the results of training on an even 2,250 samples of real and synthetic voices yielded some optimistic results. However, the degree of overfitting is uncertain at this stage and will require further tests.

The trained inception model was later validated on the rest of the large data set (over 80,000 samples) with extremely optimistic results as well. **(Add the results later when finalized)**

**Assumptions and Limits**

Phonetics, speaker variation,