eYSIP2017

Speech Spoofing and Verification

Arvind Kumar
Yash Agrawal
Aditya Panwar
Vamshi Krishna
Fayyaz

Duration of Internship: 22/05/2017 – 07/07/2017

2017, e-Yantra Publication

Speech Spoofing and Verification

Abstract

The project contains 3 separate modules that were built using Machine Learning Techniques.

- **Speaker Recognition:** Identifying a speaker's voice from a database of 10 speakers with considerable accuracy.
- **Speech Spoofing:** Voice of a source speaker is converted to a target speaker using Recurrent Neural Networks.
- **Speech To Text:** Given a speech input, the corresponding text is printed.

Completion status

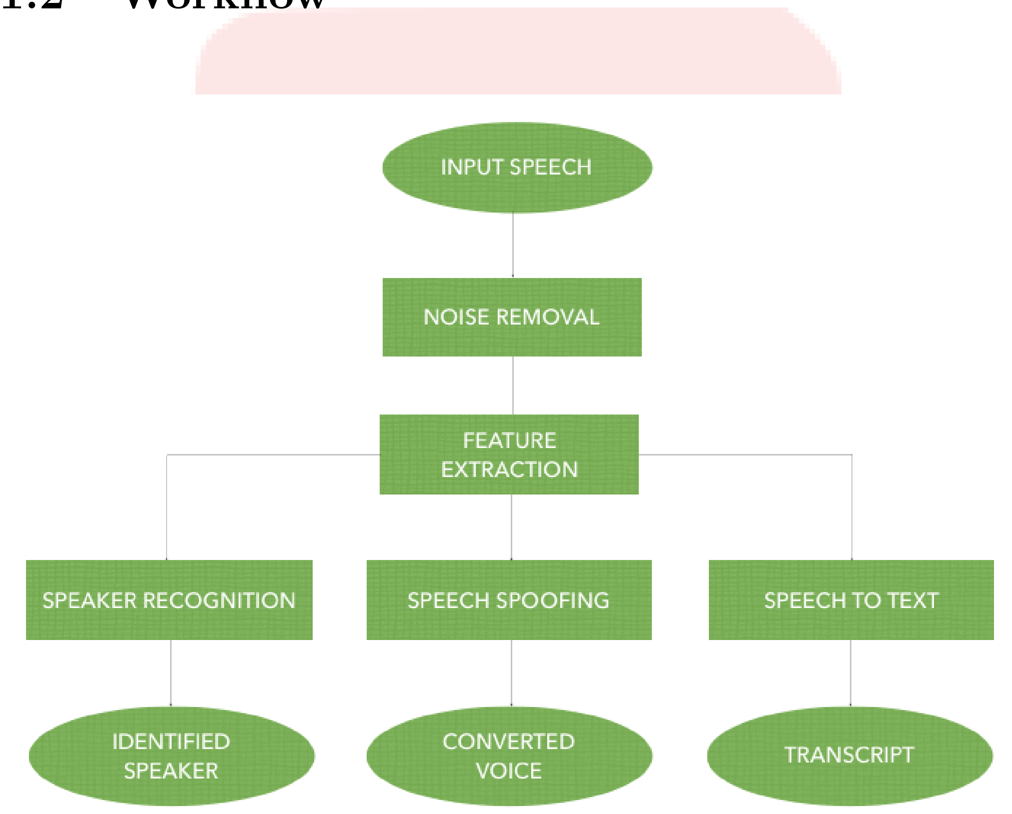
- Speaker Recognition is completed successfully with 100% accuracy for 10 different speakers.
- Speech to text model gives desirable phoneme level output.
- In spoofing, A female voice is changed to a male voice with some loss of pitch.
- Tutorials and documentation have been done.

1.1 Software used

- Python3
- Tensorflow 1.2.0 download link
- Keras download link

- hmmlearn download link
- librosa download link
- pysptk download link
- numpy download link
- pyrenn download link

1.2 Workflow



1.3 Use and Demo

GitHub Link of Tutorial

- Speaker Recognition

```
no_of_speakers = 13  # TODO: Enter Number of Speakers in Training Set (Number of gmodel files)
test_speech1 = 'MREM_Sr8.wav'  # TODO: Enter Test File Name Here.
```

Figure 1.1: Input File to Classifier

```
Log Probabilities...

FAML : -65138.3296259
MASM : -63999.0901936
FEAB : -64042.6893204
MFKC : -64656.27665
MCBR : -69850.0881777
FDHH : -66759.1974197
FHRO : -68033.5597198
FJAZ : -66269.1535362
MKBP : -65422.5558796
MLKH : -66903.4476359
SRK : -94843.5884862
FMEL : -66644.6317108
MREM : -59260.5123821

Closest Match : MREM

Process finished with exit code 0
```

Figure 1.2: Log Probabilities and Output

- Speech Spoofing
  Source Training Audio
  Target Training Audio
  Source Test Audio
  Converted Audio (Note: Reduce Volume!)

- Speech To Text

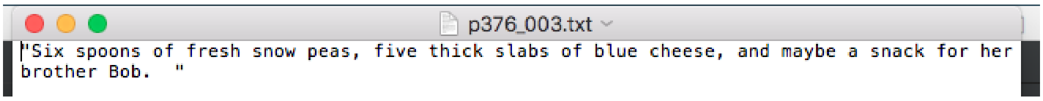


Figure 1.3: Transcript of Input Speech

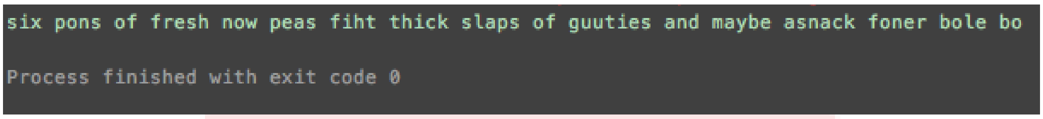


Figure 1.4: Decoded Speech at Phoneme Level

1.4 Software and Code

Github link for the repository of code

- **Speaker Recognition** We use Hidden Markov Models with Gaussian Mixture Emissions to model a speaker. Standard HMM Algorithms are used to perform Training and Classification.
  - Obtaining Features

    ```
    file_name = "speech.wav"
    rate, x = librosa.core.load(file_name)
    feature_vectors = mfcc(x, samplerate=rate)
    ```

  - Building GMMHMM Model

    ```
    model = GMMHMM(n_components=3, n_mix=128, covariance_type='diag')
    ```

  - Training Saving

    ```
    model.fit(feature_vectors)
    pickle.dump(model, f)
    ```

  - Classification

    ```
    prob = model.score(feature_vectors)
    ```

- **Speech Spoofing** Voice features of a source speaker is mapped to that of a target speaker using Recurrent Neural Networks. Speech signal is then reconstructed using these features.
  - Feature Extraction

    ```
    mcep_vectors = pysptk.mcep(frames, order=25, alpha=0.42)
    ```

  - RNN Building Training

    ```
    net = pyrenn.CreateNN([order+1, order+5, order+5, order+1])
    net = pyrenn.train_LM(source, target, net, k_max=100, verbose=True)
    ```

  - Mapping

    ```
    new_mcep = pyrenn.NNOut(mcep_vectors.transpose(), net).transpose()
    ```

  - Reconstruction

    ```
    logspec = pysptk.mgc2sp(new_mcep, alpha=0.42, frameLenth=1024)
    spec = np.exp(logspec).T
    out = librosa.core.istft(spec, 256, 1024, pysptk.blackman(frameLength))
    ```

- **Speech To Text** Dilated convolution neural network is implemented with CTC loss function using TensorFlow.
  - Building Model

    ```
    graph = tf.Graph()
    with graph.as_default():
    ```

  - Training Storing Model

    ```
    with tf.Session(graph=graph, config=config) as session:
    saver.save(session, 'saved_models/final/s2t2')
    ```

  - Restoring

    ```
    path_f = "/Users/..."  # Enter Folder Path
    path_m = "/Users/..."  # Enter Model Path
    new_saver = tf.train.import_meta_graph(path_m)
    new_saver.restore(session, tf.train.latest_checkpoint(path_f))
    ```

1.5 Future Work

- **Speaker Recognition**
  - Identifying a new Speaker.
  - Adapting to Noisy Environment so that it can be used in classrooms for attendance purposes.
  - Creating a UI to ease the training process.
- **Speech Spoofing**
  - Using better tools (STRAIGHT / PSOLA) for reconstructing speech signal.
  - Tackling the need for Time Aligned data.
- **Speech To Text**
  - A language model can be made to improve the accuracy.
  - More data can be used to Train the model better.

1.6 Bug report and Challenges

- Understanding Tensorflow, Keras and Math behind HMM.
- How to use LSTM RNN and dilated convolution neural network and to implement it using TensorFlow.
- The amount of data required for training Speech to Text Model is high. So, we obtained many different data sets and trained it.
- Input Dimensions: Some functions need input arrays that are 2D while others need data that is 3D. So, a modular approach is done where each segment is tested separately and all variables shapes are verified every step.
- Training Data for Speaker Recognition: It is a trade off between number of speakers and training data for each speaker. For our training set of 10 speakers and our goal of 100% accuracy, 90 to 100 seconds of training data was ideal.
- Neural Network Architecture: We scored various architectures before settling for 2 Hidden Layers for Speech Spoofing.

Bibliography

[1] Alex Graves and Navdeep Jaitly, *Towards End-to-End Speech Recognition with Recurrent Neural Networks*

[2] Fisher Yu and Vladlen Koltun, *MULTI-SCALE CONTEXT AGGREGATION BY DILATED CONVOLUTIONS*

[3] *L. R. Rabiner, A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, Proceedings of IEEE, vol. 77, no. 2, pp. 257-286, Feb. 1989.*

[4] *Rodrguez E., Ruz B., Garca-Crespo ., Garca F. (1997) Speech/speaker recognition using a HMM/GMM hybrid model. In: Bign J., Chollet G., Borgefors G. (eds) Audio- and Video-based Biometric Person Authentication. AVBPA 1997. Lecture Notes in Computer Science, vol 1206. Springer, Berlin, Heidelberg*

[5] *Voice Conversion with Deep Learning by Miguel Varela Ramos, 2016.*