Flipkart GRID 2.0

Solving for Voice Interactions in Indian Houses & Neighborhoods

Team Name : Wheels of Zeus

Institute Name: Indian Institute of Technology, Kanpur

Team members details

Team Name	Wheels of Zeus		
Institute Name	Indian Institute of Technology, Kanpur		
Team Members >	**1 (Leader)**	**2**	**3**
Name	Utkarsh	Tanmay Yadav	Gaurav Kumar
Batch	3rd Year	3rd Year	3rd Year

Deliverables/Expectations for Phase I (Idea Submission)

An **algorithm/approach with block diagrams and detailed explanation** to accomplish the below:

Given an audio which may or may not have background audible speech -

i) Identify if there are more than one speakers in the audio

ii) Identify and separate the primary speaker audio data based on a classifier that is either learned from data and/or assisted by some hand-engineered features.

Please provide definitions and working of the core components of the system. You can give references if any part of work is inspired by some previous work.

The solution should work for both a returning user as well as a new user. Considerations of different settings, edge cases will be given extra points.

Background speech may be from TV, music or humans. You can assume the impact of environmental noise like traffic, wind or other non-human household noise to be minimal. All other assumptions should be clearly stated.

The solution should run in real time. So **a brief discussion on computational complexity** would be expected.

Glossary

- Describe/ Expand abbreviations if you have used any in the slides below

  DSP - Digital Signal Processing

  CNN - Convolutional Neural Networks

  LSTM - Long Short Term Memory RNNs

  RNN - Recurrent Neural Networks

  ReLU - Rectified Linear Unit

Instructions (You Can Delete this Slide)

Dear Team,

Congratulations on reaching this stage - We look forward to some amazing & innovative solutions.

Please find some important instructions before you begin to prepare your submission decks.

Slide Limit	:	10 Slides of Content **post (after)** this Slide
Saving Format	:	Save the file as a PDF to ensure your formatting remains intact
Submission Guide	:	Only the **'Team Leader'** will be able to submit the Deck. Only the latest submission will be considered as final (You can keep updating your deck within the deadline)

Wishing you all the very best !

Team Flipkart GRiD

Use-cases

- As smart home devices are increasing as per U.S. Smart Speaker Consumer Adoption Report, so we propose **customer care services** to be on voice assistants so that people can ask questions instantly.
- Voice assistant for **order status enquiries**
- Apply voice assistance for shopping new products can save time in terms of **comparison** with other products, will allow people to place order on the move(ex: order while driving), etc.

Solution statement/ Proposed approach

1. General Overview

Since the problem statement described mainly address that human voice source separation should happen from various background noises, we will address primarily how to achieve it.

There are two ways to approach these problem statements:

1. **Digital Signal Processing**
2. **Deep Learning and Neural Networks**

Although conventional DSP could be used for cleaning and pre-processing the dataset we would be trying a Deep Learning approach to solve this problem with the advantages of **“Transfer Learning”**.

Transfer learning is a machine learning method where a model developed for a task is reused as the starting point for a model on a second task. It is a popular approach in deep learning where pre-trained models are used as the starting point on computer vision and natural language processing tasks given the vast compute and time resources required to develop neural network models on these problems and from the huge jumps in skill that they provide on related problems.

Continued..

2. Sub-Problems

2.1. Obtaining required data for model training

Since we primarily want to address scenario in common households, getting accurate and precise data which captures our use-case would be good task. Due to the limitations of the dataset, we will combine noise and human voice to generate our dataset.Extract background noise data and human voice data separately. We could obtain these data separately and combine these to generate our training data. Some of the common libraries which could help with it are:

1. **Background Noise**
   - 1.1. UrbanSound8K(https://urbansounddataset.weebly.com/urbansound8k.html)
   - 1.2. AudioSet Ontology by Freesound Datasets (https://annotator.freesound.org/fsd/)
2. **Human Voice**
   - 1.1. AudioSet(Google) (https://research.google.com/audioset/)
   - 1.2. Common Voice (https://voice.mozilla.org/en/datasets)
   - 1.3. LibriSpeech (http://www.openslr.org/12/)

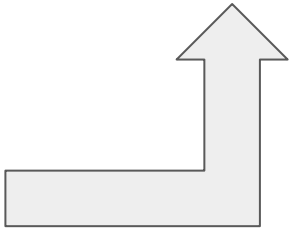
Generating Training/Testing Data

We can combine noise and human voice data to generate our training and testing data. Standard DSP Techniques could do it. A brief explanation of how we could achieve it is as follows.

We Combine the two speech sources, Ensuring that the sources have equal power in the mix. Normalize the mix so that its max amplitude is one.

A pseudo-code is displayed as follows:

```
noise = noise/norm(noise);
hSpeech = hSpeech/norm(hSpeech);
ampAdj  = max(abs([noise;hSpeech]));
noise = noise/ampAdj;
hSpeech = hSpeech/ampAdj;
mix     = noise + hSpeech;
mix     = mix ./ max(abs(mix));
```



Working with raw data

The raw file is in .wav format, typically representing Amplitude vs Time relationship. To be able to extract more features in speech synthesis, the **Spectrogram** analysis is widely used. Spectrograms are two-dimensional complex-valued graphs, with a third dimension represented by colors. Time is along the horizontal axis. The vertical axis represents frequency. The third dimension, color represent the amplitude of a particular frequency at a particular time. The Spectrogram could be simply obtained by using the **STFT technique** from standard DSP packages available. We are also able to convert it back to the time domain via **iSTFT**.

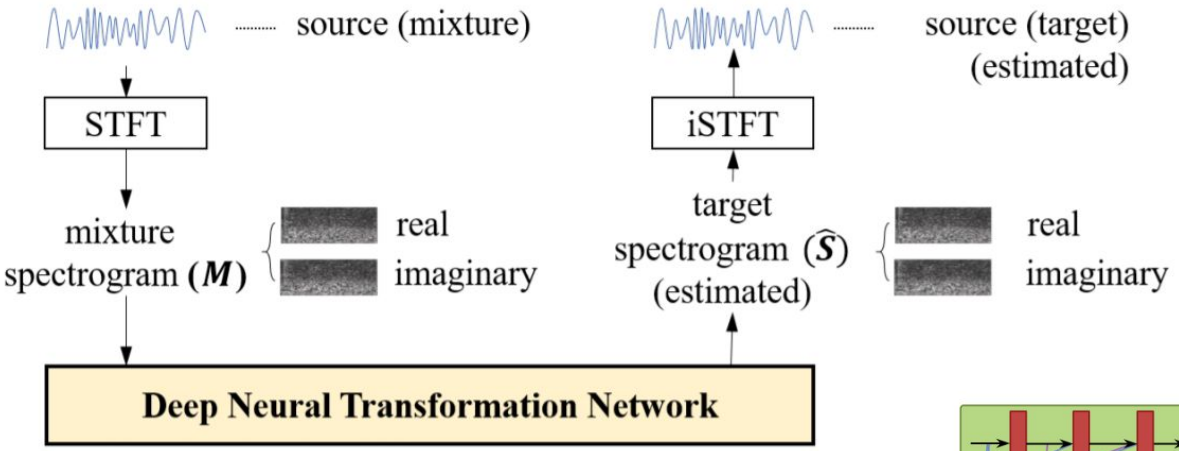
Sample Code using Librosa library

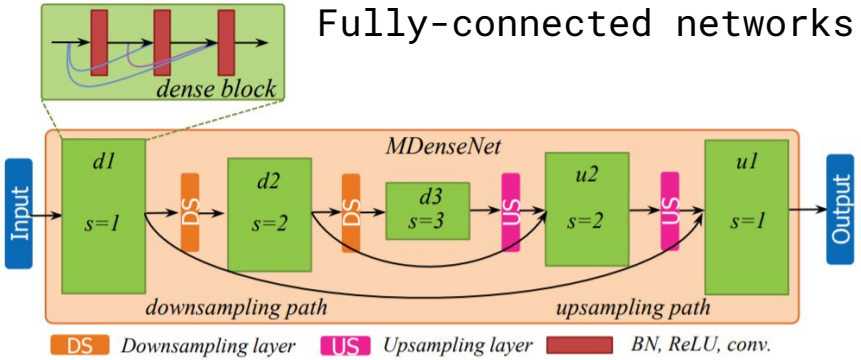
```
import librosa
import librosa.display

filename = 'test.wav'
y, sr = librosa.load(filename)
# trim silent edges
whale_song, _ = librosa.effects.trim(y)
librosa.display.waveplot(whale_song, sr=sr);
D = np.abs(librosa.stft(whale_song, n_fft=n_fft,
hop_length=hop_length))
librosa.display.specshow(D, sr=sr, x_axis='time',
y_axis='linear');
DB = librosa.amplitude_to_db(D, ref=np.max)
librosa.display.specshow(DB, sr=sr,
hop_length=hop_length, x_axis='time', y_axis='log');
plt.colorbar(format='%+2.0f dB');
```

DenseNet Architecture with Neural Transform & up/downsampling (probably LSTMs too?)

Let us discuss the problem of image segmentation first. In Image Segmentation, the machine has to partition the image into different segments, representing a separate entity. We need to exact specific features and classify them, and select which of them is useful and classify them accordingly. This is conventionally done with the help of CNNs. The image is converted into a vector that is used further for classification and reconstruct an image from this vector. This is done with the help of **multi-scale dense-net architecture** and particularly the **U-net architecture**.
Our problem could be vastly classified to Audio Source Separation(ASS) with some nuances pertaining to use-cases we are handling. We propose the use of **MDenseNet architecture.** We will use a Neural Transformation Layer along with downsampling to scale down our features. Then we upsample our vectors from the Neural Transformation Layer to predict our target voice.

Proposed Architecture Diagram





This is a representative model, The actual no. of dense blocks, no. of hidden layers, activation function and other hyperparameters will tuned accordingly with training and testing data during implementation.
The dense block comprises of Time-Frequency Convolutions with Time-Invariant Fully-connected networks.

Limitations

1. **Identifying more than one human voices:** Since our model's primary aim is to separate the primary source of a human voice, the model might fail when two people speak with the same intensity. If we need to handle that use-case, we will need separate approach called cocktail-party problem, which separates and identifies different vocals speaking with same intensity.
2. **Separating everyday household background noises:** Some of the typical background noises we encounter in our household are from TV, music or humans. Our Model should be able to classify the background noises and extract present human voice.
3. **Longer Audio Clips:** Typically audio files will have 5-10 secs of output, which scales with our model and use-case perfectly, but we may need to account longer audio clips for identification.
4. **Losses during conversion:** There might be typical losses when converting to-and-fro from the mel-spectrum.
5. **Sound from a distance/low noise:** Sound from a large distance may appear to be noise that sometimes our model might ignore.

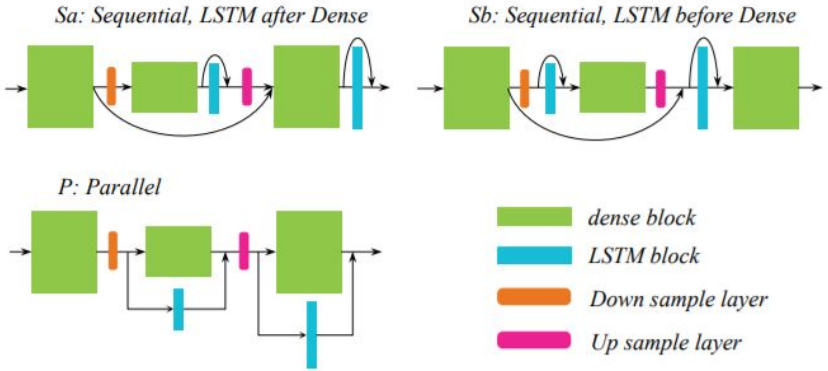
Future Scope

1. Use of LSTMs

LSTMs are a class of RNNs which typically address prediction of data and accounts for sudden modification in the data in real-time. Unlike typical RNNs, the model does not need to be re-run again if there's some information about the data but can be accounted in real-time. This might prove beneficial in speech analysis when identifying typical audio patterns of users and discarding/adding vital information in pattern recognition when using feature extraction in Neural Transformation Layer. Since LSTMs are computationally expensive, it might not be worth to add everywhere in NTL, but in strategic places by identifying suitable models.

2. Account for same intensity voices

This problem could be addressed by using reference Voice of the speaker which typically using the as-sitant or capturing particular phrases which we typically encounter in shopping experience of user.



References

- https://arxiv.org/abs/1805.02410
  (MMDenseLSTM: An efficient combination of convolutional and recurrent neural networks for audio source separation)
- https://towardsdatascience.com/understanding-semantic-segmentation-with-unet-6be4f42d4b47
  (Unet Architecture)
- https://towardsdatascience.com/getting-to-know-the-mel-spectrogram-31bca3e2d9d0
  (Mel spectrogram using Librosa)
- https://www.mathworks.com/help/deeplearning/ug/cocktail-party-source-separation-using-deep-learning-networks.html
  (Cocktail Party Source Separation using Deep Learning)
- https://arxiv.org/abs/1912.02591
  (Investigating Deep Neural Transformations for Spectrogram-based Musical Source Separation)


Flipkart
GRID 2.0