ISSN No:-2456-2165

Targeted Voice Separation

Aakanksha Desai, Varsha Kini, Vrunda Mange, Prof. Suvarna Chaure
Department of Computer Engineering
SIES Graduate School of Technology Navi Mumbai, India

Abstract:- Speech is the preferred means of communication between people. It is starting to be the primary means of contact between machines and humans. Machines are increasingly able to imitate many of the conversational exchange capabilities for well-defined tasks. As a result, the ability of sophisticated machines can be used to meet social needs without burdening the consumer beyond the experience of natural spoken language. Speaker separation is a task to distinguish the target speaker's voice from interference. This interference can be the voices of other speakers in the background. In this paper, we present a method for obtaining a solution to the cocktail party problem by using neural networks. The input is an audio file containing voices of multiple speakers talking at the same time, and the clean speech of the target speaker. The output will be target speech separated from mixed audio in input.

Keywords:- *Cocktail Party Problem, Neural Networks, Voice Separation.*

I. INTRODUCTION

In a crowded room, where multiple people are speaking at the same time, our human ears are quite capable of discerning each individual speaker's voice. As machines are trying to be at par with humans, researchers have tried to implement the problem of speech separation on the virtual medium as well. However, in the field of computation, this problem has been persistent for several decades now, as it is a tedious task and cannot be solved easily. E. C. Cherry in his 1953 paper on "Some experiments on the recognition of Speech" coined the term Cocktail Party Problem for this particular issue [7]. Various methods have been employed to find a solution to this problem from spectral subtraction to principal component analysis (PCA).

Generally, speech separation is studied as a problem of signal processing. According to modern trends speech separation can be classified as a Machine Learning problem as it can be solved using supervised learning algorithms. In this approach the model is trained to classify speech and noise based on their discriminatory features. The task of speech separation using deep learning has gained wide popularity due to expedited development and considerable improvement in separation efficiency from training data. Over the last decade, a lot of progress has been made in solving the task of speech separation by utilising appropriate supervised learning algorithms. Our system proposes a method for speech separation by only separating the target speaker's voice from the mixed signal.

II. RELATED WORK

[1] divides the task of voice separation into two classes namely positive and negative classes. The positive class represents the required speaker's voice, while the negative class represents the background noise and interference. The result is achieved by training two separate networks: a speaker recognition network and a spectrogram masking network. This system is not open source and is trained on huge amounts of data. Also,there is currently no dataset available solely for the speaker separation task.

[2] proposed a source extraction process performed to extract all the sources using RNN. It can determine the required iterations. The system uses a block online approach in which only a single source is obtained at a given time. This process is repeated until all extraction of all the sources is complete.The system has no drawbacks, but can only be theoretically applied for any number of random speakers.

[4] proposed a system that uses PIT i.e, permutation invariant training which directly recognizes the multiple speech streams in a single source without first separating them. The system will provide the number of speakers in the audio input but will not separate the different speech streams which are necessary for targeted voice separation.

[5] proposed a system using four separate DNNs to estimate a combined time-frequency mask that gives better estimates for the sources as compared to using each DNN individually. Each mask has varying amounts of distortion and interference which is reflected in the output. The shortcomings of this approach are that prior information of the number of sources is required which is difficult in real-life scenarios, and it might eliminate audio that is required to be the output.

Speech separation being a challenging task, [6] proposes an end-to-end system for signal approximation, which improves the performance of the system greatly. The system is implemented using Bi-LSTM and hard clustering. The only shortcoming of using hard clustering is that some of the entities do not belong entirely to one class, as we may encounter several unknown classes in the real world.

[10] In this paper Deep U-Net architecture has been implemented for separating vocals from music. So we have implemented our system using this reference to perform speaker separation using our own dataset.

ISSN No:-2456-2165

III. PROPOSED WORK

A. Dataset

The Librispeech train-clean-360 subset[8] consists of 10-15 seconds long excerpts from audio books of 921 speakers. On merging the samples by applying permutation and combination to them, we created a dataset of 1000 audio files with 2 speakers. It took us around 10-15 hours to create the dataset.

B. Proposed System

Our system is divided into two parts:

a) Speaker Separation:

Initially the mixed input audio file will be processed to separate into individual speaker's voices.Due to the high computational demands involved in working with audio files, the training of the model becomes a tedious task. Hence we have downsampled the input to 8192 Hz. Further the Fourier Transform of this signal is calculated with a window size of 1024 and hop length as 768. The magnitude spectrograms obtained from the preprocessed data will be normalised to the range[0,1] and saved in a .npz format.

The model is trained based on U-Net architecture. It generates a soft mask which can then be multiplied with the mixture audio file to give separated source files.

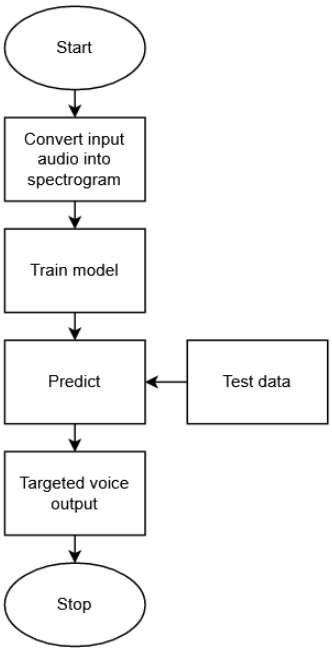

Fig. 1: System Flowchart

b) Targeted Voice Identification:

After completing separation of speaker voices from the mixed audio file, the next step is to identify the target speaker. The target voice from the separated voices is identified by comparing the reference file with obtained voices. We were able to carry out the comparison using the python library Resemblyzer. It compares different voices by computing the cosine similarity and gives a value on how similar they sound. The signal with higher value is then given as the output by the system.

IV. METHODOLOGY

The U-Net architecture is used to carry out the task of speaker separation in this work. Unet is an architecture evolved from a deep neural networks class called convolutional neural networks.Unlike CNN that generally focuses on image classification tasks(with one input image and one output label), U-Net can localise and differentiate borders by doing classification on each pixel, which increases precision.

The U-Net architecture is symmetric and is a type of fully convolutional network (FCN) which has only convolutional layers (i.e. layers are not fully connected). The architecture is divided into two parts:

- An encoder that learns a highly abstract representation of the input.
- A decoder that takes the encoder's input and maps it into segmented ground truth data.

The architecture was originally aimed at processing biomedical images. It improved the precision and localization of their microscopic neuronal structures[10]. The architecture is based on a completely convolutional network and resembles a deconvolutional network. Deconvolutional network or transposed convolutional network is basically a convolutional network that is reversed. A series of convolutional layers make up the de-convolutional network. These layers condense the image while increasing the number of channels, resulting in a small and deep representation. The up sampling layers are used to decode this compressed representation.

Jitter and other distortions can become evident when the time domain changes. As a result, maintaining a high degree of detail in the reproduction is critical. In order to allow flow of low-level data directly from high-resolution input to the corresponding high-resolution output, the U-Net architecture has skip connections across the layers at the same level in the hierarchy. These connections are established in both the encoder and the decoder.

V. ARCHITECTURE

The objective of the U-Net architecture is to indirectly identify the individual speakers from the given audio input. The soft mask obtained from the model is multiplied with the mixed spectrogram to give the final prediction of the target speaker. The network architecture is depicted in the diagram below.

ISSN No:-2456-2165

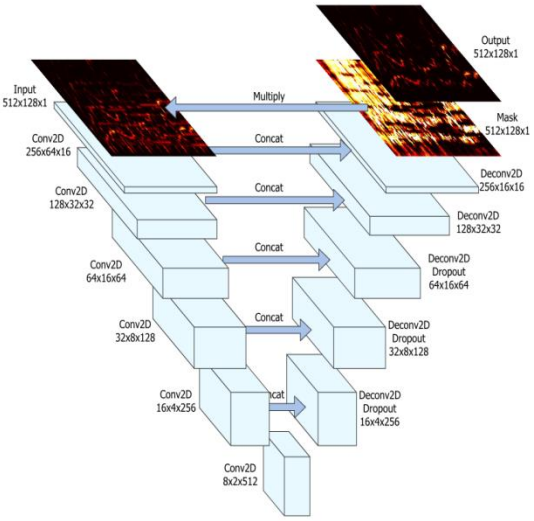

Fig. 2: U-Net Architecture[10]

Our U-Net implementation is similar to[10]. In the encoder, each layer consists of a strided 2D convolution with a stride length of 2 and a kernel size of 5x5, batch normalisation, and the activation function is a leaky rectified linear unit (ReLU). The strided deconvolution layer is similar to the convolution layer, except that the activation function used is a plain ReLU and has a 50 percent dropout. In the last layer we utilise a sigmoid activation function. The ADAM optimizer is used to train the model.

The neural network model primarily processes the magnitudes of the spectrograms obtained from the corresponding audio file. The spectrogram has 2 components associated with it, namely its magnitude and phase. The magnitude of the output is determined by multiplying the original magnitude with the mask that is obtained from the model, whereas the phase of the output is unchanged. This implementation has been effective in obtaining the target voice from the mixture, and the results are specified in the following section.

VI. RESULTS AND DISCUSSIONS

We have developed a system by training a model: that separates the two speakers successfully, identifies the target speaker.

The following is a spectrogram of an audio sample from the test set.

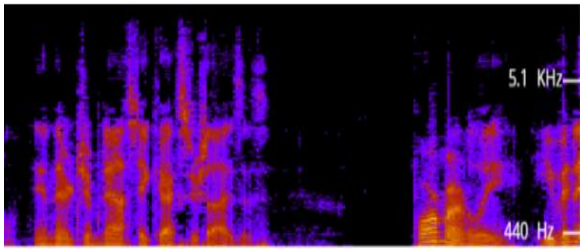

Fig 3. Spectrogram for mixture of 2 speakers

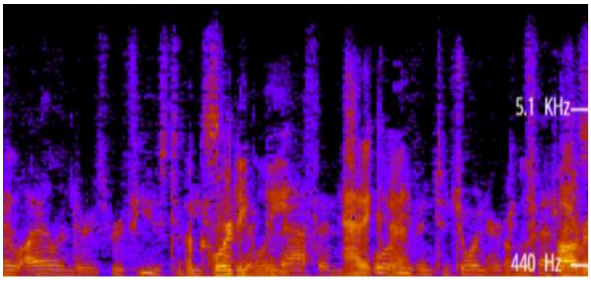

Fig. 4: Spectrogram for audio of target speaker

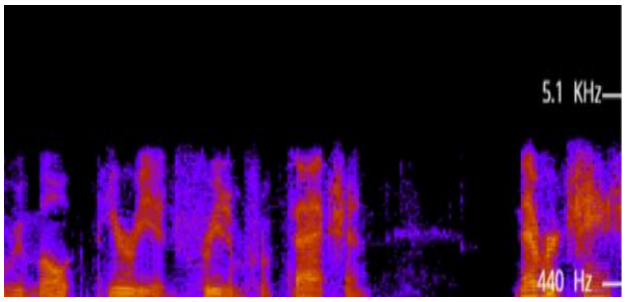

Fig. 5: Spectrogram for audio of first speaker (downsampled)

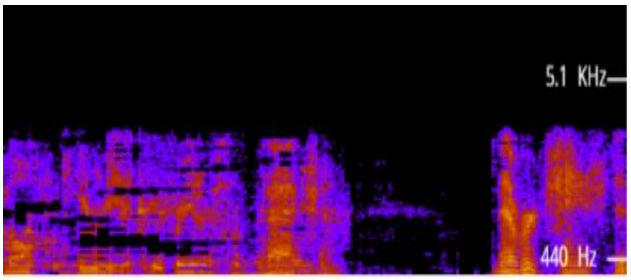

Fig. 6: Spectrogram for predicted audio of target speaker (downsampled)

A. Evaluation Metric

For evaluating our system, we have computed the Signal to Distortion ratio from the mir_eval python library. The bss_eval_sources method, which takes parameters reference audio and predicted audio, was used to compute SDR. The SDR obtained is 7.09 dB.

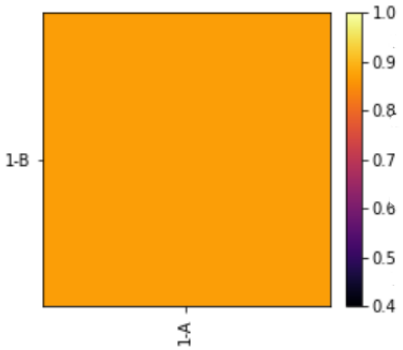

Fig. 7: Similarity matrix representation for similar audio files

ISSN No:-2456-2165

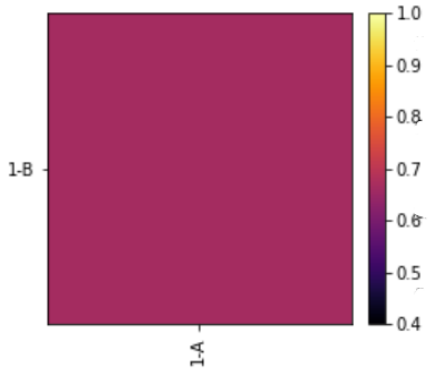

Fig. 8: Similarity matrix representation for dissimilar audio files

With the help of resemblyzer module we compare the output files with the reference file and a similarity metric is obtained. The audio file with greater similarity value will indicate the target voice.

Fig 7 shows comparison of one of the separated sources with reference audio. Its similarity percentage is 87.33782%.

Fig 8 shows comparison of the other separated source with reference audio. Its similarity percentage is 66.58919%.

We have also implemented our system which will present a simple interface for uploading the input audio files. The final output audio file will be downloaded onto the user's machine.

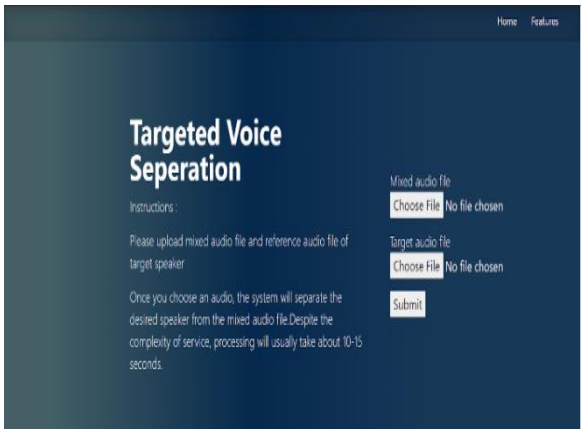

Fig. 9: Implementation

VII. CONCLUSION AND FUTURE WORK

We investigated the U-Net architecture for the task of separating voices of speakers. The U-Net method that we have employed for the problem of speech separation is unique in nature. Although different systems have been implemented such as Voicefilter, which provides state-of-art results, our system is one of its kind and has been successful in separating the mixed sources with minimal distortion. Since we have made use of a limited dataset, we believe that the system could be evaluated by adding more training data. We speculate that more training data will provide more accurate results, however the impact of training data size on quality of separated audio is yet to be investigated.

REFERENCES

[1.] Quan Wang, Hannah Muckenhirn, Kevin Wilson, Prashant Sridhar, Zelin Wu, John Hershey, Rif A. Saurous, Ron J. Weiss, Ye Jia, Ignacio Lopez Moreno "VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking" (19 June 2019)

[2.] Thilo von Neumann, Keisuke Kinoshita, Marc Delcroix, Shoko Araki, Tomohiro Nakatani, Reinhold Haeb-Umbach "All-Neural Online Source Separation, Counting and Diarization for Meeting Analysis" ICASSP2019.

[3.] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 4879–4883.

[4.] Dong Yu, Xuankai Chang, Yanmin Qian "Recognizing Multi-talker Speech with Permutation Invariant Training" INTERSPEECH 2017 August 20–24, 2017, Stockholm, Sweden

[5.] Emad M. Grais, Gerard Roma, Andrew J.R. Simpson, Mark D. Plumbley "Single Channel Audio Source Separation using Deep Neural Network Ensembles" AES 140th Convention, Paris, France, 2016

[6.] Yusuf Isik, Jonathan Le Roux, Zhuo Chen, Shinji Watanabe, John R. Hershey "Single-Channel Multi-Speaker Separation using Deep Clustering" Interspeech September 2016.

[7.] E.C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," J. Acoust. Soc. Am., vol. 25, pp. 975-979, 1953.

[8.] Vassil Panayotov, Guoguo Chen , Daniel Povey , Sanjeev Khudanpur "Librispeech: An ASR Corpus based on Public Domain Audio Books" Speech and Signal Processing (ICASSP), 2015 IEEE International Conference

[9.] Daniel Stoller, Sebastian Ewert, Simon Dixon "Wave-U-Net: A Multi-scale Neural Network for end-to-end Audio Source Separation" 19th International Society for Music Information Retrieval Conference, Paris, France, 2018.

[10.] Andreas Jansson, Eric Humphrey, Nicola Montecchio, Rachel Bittner, Aparna Kumar, Tillman Weyde "Singing Voice Separation with Deep U-Net Convolutional Networks" 18th ISMIR Conference.