Speech-to-text conversion and Speaker Count extrapolation using vector quantization of speech characteristics

Suraj Pathak
College of Information and Computer Sciences
University of Massacheusetts
Amherst, United States Of America
suraj.s.pathak@gmail.com

Aditya Mujumdar
School of Computing, Informatics, and Decision Systems Engineering
Arizona State University
Tempe, United States Of America
aditya.mujumdar09@gmail.com

Saurabh Shaligram
Maker's Lab
Tech Mahindra
Pune, India
saurabh.shaligram@gmail.com

***Abstract*— Speech recognition in humans is far more advanced compared to speech recognition systems on machines available today. Our aim is to simulate the biological concepts responsible for the working of a human ear to create a speech recognition system that can perform speech to text and count the number of people speaking at every instant. The idea which we pursue is to use the Random Forest Classifier Algorithm to convert speech into text and calculate the number of people speaking at any given instant using a rule-based approach.**

Keywords— Automatic Speech Recognition, Random Forest Classifier, feature extraction, speaker count, speech-to-text

I. INTRODUCTION

Speech Recognition is on its way to becoming an integral part of our daily lives. From voice assistants to video captioning it has pierced through all the barriers and is now being integrated in most of the devices being developed today. It has helped to break the language barrier, by automating the process of language translation and video captioning. Thus, expanding the reach of science, education and technology.

Research suggests that most of the preprocessing of the audio signals happens in the cochlea which is located in the inner ear. Electrical impulses from the cochlea travel through the auditory nerve towards the brain to be interpreted as sound [1]. This paper suggests a method loosely based on the functioning of the cochlea to extract and interpret information from sound signals.

We replicated the preprocessing stage by creating a digital cochlea using and array of bandpass filters. The processed information outputted through this digital cochlea was then used to:

- Devise a rule-based model to count the number of speakers.
- Train a Random Forest Classifier algorithm to generate a speech recognition model.

The speaker count module finds the number of speakers at every sample. This system can be used in order to predict the number of speakers speaking in a specific interval. Some examples of the applications of this module are:

- News telecast rooms for predicting the number of speakers in an open debate speaking at the same time.
- Military applications to stealthily gather information and strategize plan of action.
- Increase accuracy of speaker diarization models.

The speech-to-text model in this paper is based on the concept of bio-mimicry [6]. It is modelled after the human ear, more specifically the cochlea. The system creates a vector representations of spoken words and then makes a prediction based on the training data.

II. EXPERIMENTS

A. Audio file format experiments

Audio is stored in multiple ways depending upon the type of an audio codec used [5]. The compilation of a table with information about audio codecs with varying sampling rates, bit rates and bit depths can be seen in Fig.1.

Major audio encoding formats include:

- Lossy audio compression format

With a considerable audio compression using a lossy compression codecs, the quality of the audio is compromised in order to reduce the overall size of the audio.

- Lossless audio compression format

  Without losing a noticeable amount of information, lossless audio compression helps in retaining the data with reduction in the size of the file.
- Uncompressed format

  Since there is no compression on the audio file, an uncompressed format generates a representation of the audio file without any loss of information.

An uncompressed audio format helps in retaining all the information at the time of a recording as there is no pre-processing involved. Using a PCM encoding, a .wav file was created for a given audio to store in an uncompressed format.

TABLE I. AUDIO FILE FORMAT COMPARISON

File format	Properties				
	Codecs	***Compression***	***Sampling Rate (KHz)***	***Bit Rate (Kbps)***	***Size per sec (KBps)[a]***
.wav	PCM	Uncompressed	44.1	705.6	88.2
			88.2	1411.2	176.4
.ogg	VORBIS	Lossy Quality - 10	44.1	224	29.3
		Lossy Quality - 10	88.2	256	30.8
		Lossy Quality - 5	88.2	160	10.6
.flac	FLAC	Lossless	44.1	379	46
			88.2	498	60.9
.dsf	DSD	Lossless	5600	5600	698.3

- approximate values

Fig. 1. Comparing audio codecs with varying properties

B. *Spectrogram experiments*

To view the frequency information present in an audio file, we viewed the spectrogram of the audio file[2][3] and conducted further experiments with varying sound samples. These samples consisted of:

- Same speakers speaking different words
- Different speakers speaking the same words.

The results of these experiments showed that different words are comprised of multiple frequencies depending upon the word and its pronunciation.

Fig.2 shows how different words spoken by the same person have visually dissimilar spectrograms.

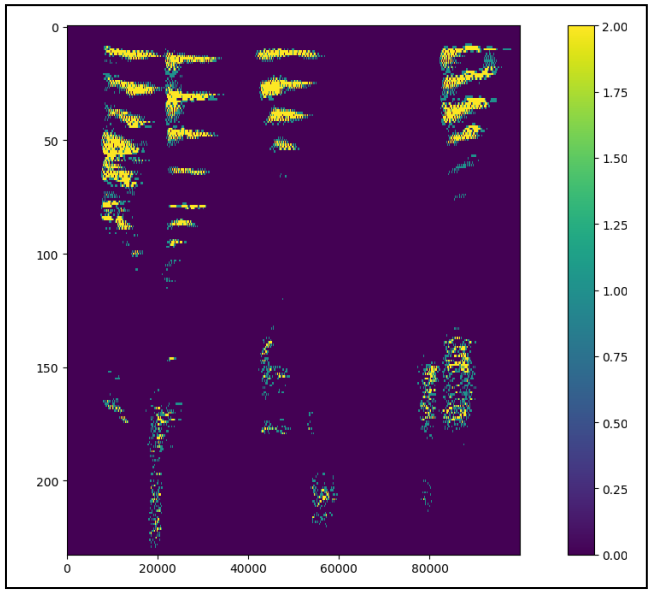

Fig. 2. Spectrogram of words ‘also’, ‘gives’, ‘shade’ spoken by a single person. Y axis denotes frequency bands. X axis denotes the sample number.

Fig.3. shows two low frequency band spectrograms which depict the comparison between two people speaking the same word.

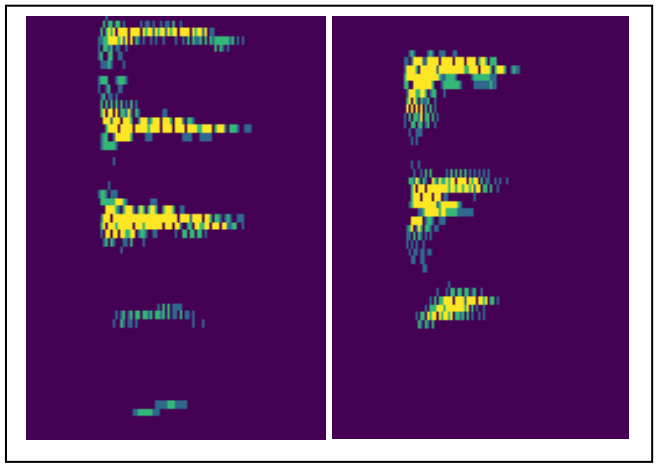

Fig. 3. Low frequency band spectrogram comparison of two people saying the letter ‘q’.

Conclusively, we came up with the hypothesis that by looking at particular bands in a spectrogram, we can determine the number of speakers as well as predict the words spoken. This relation is defined as:

$$speaker\ count \propto sound\ activity\ in\ low\ frequency\ bands \quad (1)$$

III. Implementation

A. *Digital cochlea*

The cochlea in our ear is a spiral shaped organ that is responsible for converting sound waves into electrical impulses. Tiny hair cells present in the cochlea help in the process as they have specific resonance frequencies and resonate when certain frequency waves are played. Hair cells of varying length are present in the cochlea and have resonance frequencies ranging from 20 Hz to 20 kHz [1] . To mimic these properties on a computer, Bandpass filters of varying center frequency and band width can be used.

The digital cochlea is made using an array of bandpass filters with center frequencies ranging from 20 Hz to 20 kHz [4], however, the entire range need not be considered. The output of this digital cochlea is a 2D array that looks similar to a spectrogram (obtained by performing a fast Fourier transform) of the input audio.

B. *Speaker count module*

The Speaker Count module takes in the output of the digital cochlea to estimate the number of speakers speaking at a particular sample in an audio file.

Speaker count is found by:

- Converting the raw spectrogram data using a sigmoid function into an array of 1s and 0s.
- Scaling the data received from the sigmoid function into values which represent the number of speakers.

Converting Spectrogram data using a Sigmoid function:

Taking the output of the digital cochlea which represents the raw data of the spectrogram, we further work towards processing this data into a more readable format consisting of only 1s and 0s. This process removes any low amplitude disturbances present in the audio file. The processing of the raw spectrogram data is done using a sigmoid function which has been found experimentally and is given by

$$S(x) = \frac{e^{0.2(x-137)}}{e^{0.2(x-137)}+1} \quad (2)$$

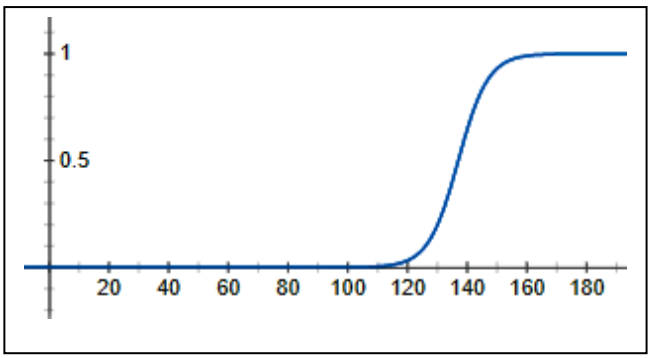

Fig. 4. Graph of the sigmoid function (2)

Scaling the Sigmoid function output:

Based on the hypothesis (1) for speaker count, we look at the lower frequency bands that we experimentally found to be between 20 Hz – 170 Hz. By scaling the output of the digital

cochlea at these bands, we find the final count of speakers in an audio file at a given sample. This scaling is done by using an experimentally derived equation.

$$C(x) = 0.0289x^2 + 0.5029x + 0.0874 \quad (3)$$

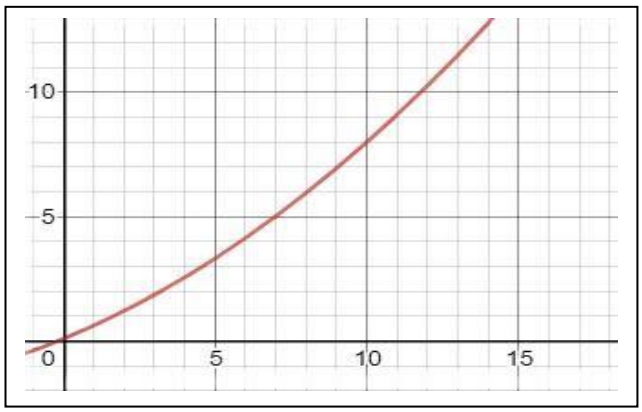

Fig. 5. Graph of the scaling function

C. Speech to text module

The speech-–to-text module takes the output of the digital cochlea as input and performs speech-to-text at intervals where sound activity is observed.

The speech-to-text module works by:

- Creating a unique vector for the interval where voice activity was detected
- Feeding the unique vector to a Random Forest Classifier algorithm, trained on a speech-to-text database, and predicting the class of the input vector – the word for the input audio.

Unique vector creation:

In order to use the digital cochlea's 2D output in a Random Forest Classifier algorithm, the output needs to be reduced to a 1D array. This dimension reduction is done by calculating the average amplitude of each band of the output from the digital cochlea (bandpass filter array) and returning the results as an array. Further processing is done on this array to summarize it into a vector with lesser elements

Random Forest Classifier:

The model to predict the spoken word was trained using a random forest classifier algorithm on a self-created labeled dataset. The dataset consisted of seven different speakers uttering 22 unique words compiled into nine audio files.

Each audio file in the dataset was split into words. Each of these words were converted to unique vectors and fed into the classifier algorithm for training. The training-testing data split was 80%-20% respectively.

IV. Conclusion

The digital cochlea effectively translates audio format representation. It mimics the function of a human cochlea by using a well-defined set of bandpass filters at regular intervals. The output produced from this digital cochlea is transformed into a 2D vector array which represents every sample in any recorded audio file. This vector array information is then used to solve two main problems of speaker count prediction and speech to text conversion. Speaker Count prediction uses a combination of sigmoid and scaling functions to give us an accurate dynamic estimate of the number of speakers at any instant. The Speech-to-Text conversion module uses this speech data acquired from the digital cochlea to encode the 2D vector into a format recognizable to an AI-based pattern recognition tool which eventually feed us with speech information converted to text.

Acknowledgment

This work was supported by Tech Mahindra. The authors would like to express their sincere thanks to Mr. Nikhil Malhotra for providing the infrastructure needed to conduct the research. Special thanks are also given to Ms. Jugnu Manhas, the team at Maker's Lab, Hinjewadi, and the internees for their utmost interest and support towards the completion of this project.

References

[1] Camhi, J. Neuroethology: nerve cells and the natural behavior of animals. Sinauer Associates, 1984.

[2] JL Flanagan, Speech Analysis, Synthesis and Perception, Springer-Verlag, New York, 1972

[3] Julius O. Smith III, "Mathematics of the discrete fourier transform (DFT), with audio applications --- second edition" 2007

[4] Rosen, Stuart, and Peter Howell. *Signals and Systems for Speech and Hearing*. Emerald, 2011.

[5] *R.C. Jaiswal, Audio-Video Engineering, Nirali Prakashan, Pune, Maharashtra, 2009.*

[6] Vincent, J., Bogatyreva, O., Bogatyrev, N., Bowyer, A., & Pahl, A. (2006). Biomimetics: Its practice and theory. *Journal of the Royal Society, Interface*, *3*(9), 471-482.