Characterizing Speakers Using Spectrograms

Rahul Gupta[1]*, ***Saijal Agarwal***[2], ***Ravina Rewar***[3]
[1]Computer Science, Poornima Institute of Engineering and Technology/Rajasthan Technical University, India
[2]Computer Science, Poornima Institute of Engineering and Technology/Rajasthan Technical University, India
[3]Computer Science, Poornima Institute of Engineering and Technology/Rajasthan Technical University, India

Abstract

Deep learning models are good at seeing but not so good at hearing. So, when it comes to audio or speech data, we don't get expected results. Basically, our need was to train a deep learning model, which is able to classify the speeches on the basis of various characteristics of the speaker. In this paper, we tried to visualize differences among speakers on the basis of his Liveliness, Speech Rate and Vocal Depth in the form of images. We did this by taking the audio signals to spatial domain by using a method called Short Term Fourier Transformation. To better visualize the differences we used concept of Mel Spectrograms, which better correlates with human hearing. And we successfully visualized the differences with the help of mel spectrograms. And these differences were very significant and considerable which can be observed by anyone. These observations can be proven very useful in training deep learning models that can evaluate speakers.

Keywords: Spectrograms, Mel Spectrograms, CNNs, Characterizing Speakers

**Author for Correspondence* E-mail : 2015pietcsrahul083@poornima.org, Tel : 9461455913

INTRODUCTION

Human being can hear the differences between two speeches, but can he visualize those differences? Probably not. Because speech is a phenomenon, which can be sensed by ears, not eyes. But what if we could see those differences? It would not make any difference, because our task is to differentiate between two speeches, whether we do it by seeing or hearing, doesn't matter. But surely, it would make a lot of difference if we want to train a deep learning model that can determine the characteristics of a speaker. However, there are varieties of Artificial Neural Networks (ANNs) which can accept Visual (CNNs) or even Audio (RNNs) data. But RNNs have not produced much success. The reason being, that the RNNs work in temporal domain, and it is quite difficult task to handle data in temporal domain. But CNNs have produced considerable benchmarks in various image and audio classification tasks. So, if the audio data can be visualized, then CNNs can generate much better and accurate results in classifying speeches. Hence, in this work, our main focus is to visualize speech data as images and observe the differences in between the images of lively & monotonous speech, fast & slow speech and high pitch & low pitch speech, which we have discussed further.

TIME DOMAIN VS SPATIAL DOMAIN

Generally, the audio we hear is in time domain. It basically contains amplitude of an audio signal at each time instant. The x-axis is Time and the y-axis is Amplitude. But it doesn't contain much information about audio signal. And it is insufficient to train a deep learning model. So, spatial domain comes to our rescue. It basically contains the information about frequency contents of the audio. We apply a method called Fourier Transformation to the time domain signal, and it takes the signal to Frequency domain or spatial domain. The x-axis is Frequency and the y-axis is Intensity or Amplitude. And as the frequencies are the most important features in speeches, audio in spatial domain is more appropriate input for the deep learning models.

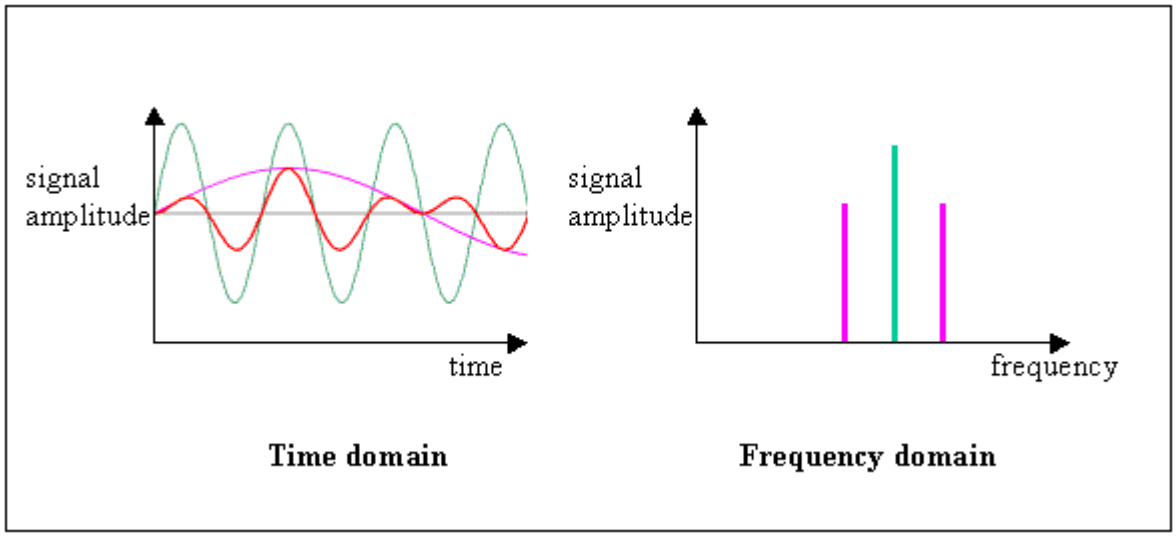

Figure1 : Plots of signal in Time-domain and Frequency-domain

SHORT TERM FOURIER TRANSFORMATION (STFT)

As described above, Fourier Transformation is a method that takes the time domain signal to frequency domain. But Fourier Transformation can only work for uniform signals (which changes uniformly with time). But real life audio signals are highly non-uniform in nature. Hence, to overcome this drawback, a new method was developed called Short Term Fourier Transformation (STFT). What it does is, it divides the audio signal into small pieces of specified window size. And then the Fourier Transformation is applied on each of those small pieces. It works on the principle that, signal of sufficiently small time duration can be considered as uniform signals. That's why it is known as Short Term Fourier Transform. The output of STFT is a spectrogram (Figure 2), which is basically like a 2D image. The x-axis indicates the time slices and the y-axis indicates the frequency bins. The spectrogram contains some bright and dark regions. High intensity regions indicate that content of corresponding frequency bin at corresponding time slice is very high and vice versa.

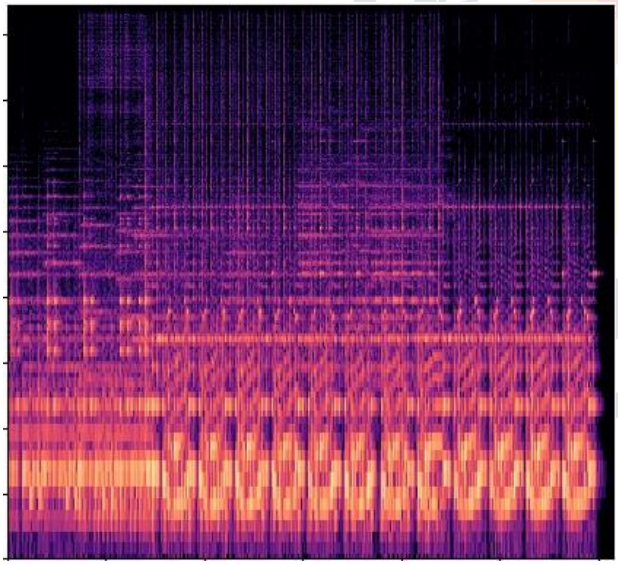

Figure 2 : Spectrogram(STFT) of a 5-sec audio clip

VISUALIZING SPEECHES AND MEL SPECTROGRAMS

We have discussed above that by taking the STFT we can visualize an audio signal by plotting its spectrogram. So, we can also visualize the audio signals of human speech also. But generally we have to use Mel spectrograms, which is a better option. Non-technically, the reason is, we expect our deep learning model to perceive the speech same as human ears perceive, and that's what mel spectrograms are used for. And technically, the reason is, Mel frequency spacing approximates the mapping of frequencies to patches of nerves in the cochlea, and thus the relative importance of different sounds to humans (and other animals). Thus, binning a spectrum into approximately mel frequency spacing widths lets you use spectral information in about the same way as human hearing. Hence, we are going to visualize human speech as perceived by human ears with the help of mel spectrograms.

AUDIO DATA

We collected some speeches from youtube, specially from the channels like Ted Talks, The Moth, Ink Talks, The Veritas Forum. And then we extracted 5-10 audio clips from each video of approx 5 seconds each. For this task, we used an audio processing software, Audacity. And then we plotted mel spectrograms of each clip one-by-one and made some observations, which will be discussed later in this paper. We used an audio processing library of python called, Librosa, to plot audio signals and mel spectrograms.

LIVELY VS MONOTONOUS SPEECH

Liveliness is one of the most important characteristics of a speaker that drives the attention of the audience. Liveliness basically means, using a voice that varies in pitch, speed and loudness. It denotes the energy and enthusiasm of a speaker. A lively speaker tend to engage the audience more than a monotonous speaker. We plotted the mel spectrograms of various pairs of lively and monotonous audio clips and we observed considerable differences in each pair. The mel spectrograms of lively speeches had low or maybe high intensity curvatures while the mel spectrograms of monotonous speeches had high intensity flat lines (Figure 3 & 4).

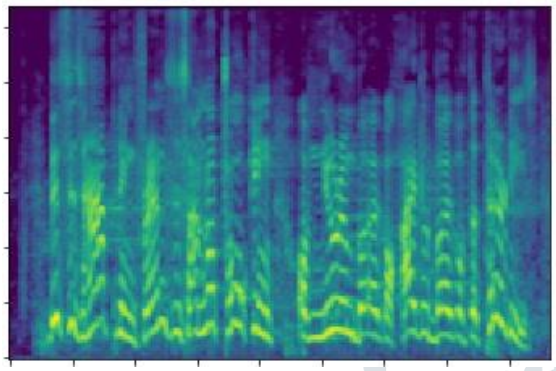

Figure 3: Mel-spectrogram of Lively Voice

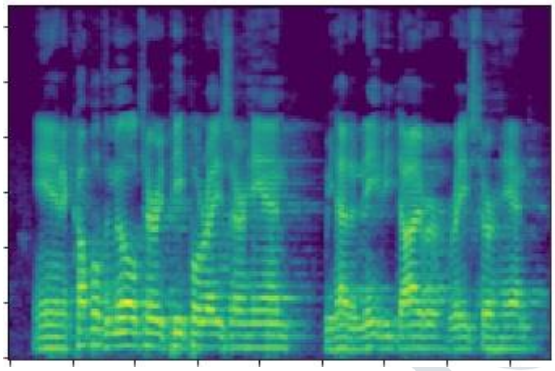

Figure 4: Mel-spctrogram of Monotonous Voice

HEAVY VS LIGHT SPECCH (LOW PITCH VS HIGH PITCH)

Heaviness or vocal depth in a speech also is an important characteristic, which can be used to determine the gender of the speaker, as males tend to have heavy or deep voice, that's why its also called as manly voice, and females tend to have light voice, that's why it is also called as feminine voice. We observed the significant differences between mel spectrograms of the heavy and light speech pairs. The mel spectrograms of heavy voice had the bright lines or curves very closely spaced whereas the mel spectrograms of light voice had the dull lines or curves very highly spaced (Figure 5 & 6).

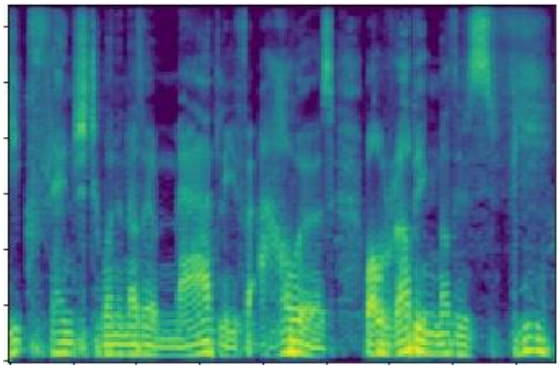

Figure 5: Mel-spectrogram of Heavy Voice

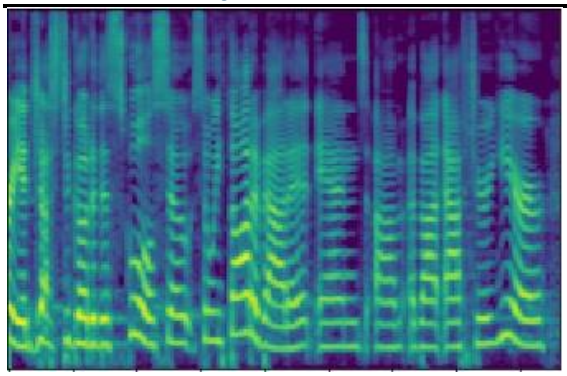

Figure 6: Mel-spectrogram of Light Voice

FAST VS SLOW SPEECH

The speech rate of the speaker plays a very important role in determining the quality of speech. However, speech rate is a very controversial dimension when it comes to judging a speaker. The judgement varies largely with speaker to speaker as well as with person to person in the audience. But overall, as far as we understood, when it comes to drawing attention of the audience, the fast speech plays the role, and when it comes to clearly understanding and getting the context of the speech, then slow speech is more desirable. We plotted the mel spectrograms of fast and slow speeches in pairs and we observed the differences between those plots. The mel spectrograms of fast speeches had curvatures of small sizes whereas the mel spectrograms of slow speeches had curvatures of large sizes (Figure 7 & 8).

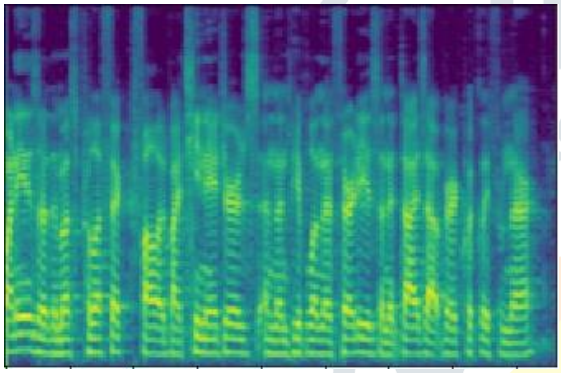

Figure 7: Mel-spectrogram of Fast Speech

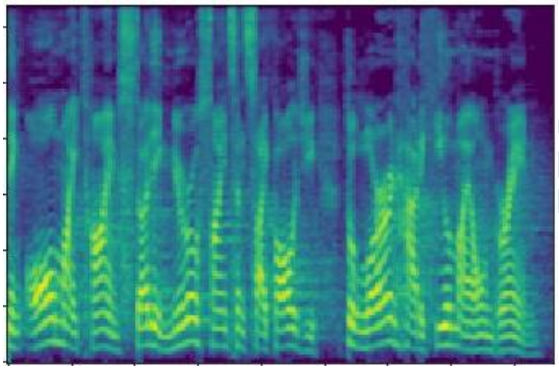

Figure 8: Mel-spectrogram of Slow Speech

CONCLUSIONS

Human speech and perception of that speech by human ear is a highly complex phenomenon. And if we want to mimic this perception into an artificial model, then it becomes even more tedious task. So, we found a new method to visualize the various characteristics of speech in the form of an image. And we were able to visualize the differences in speeches on the basis of these following characteristics; Liveliness, Speech Rate and Vocal Depth. Now, in future, we can easily built an artificial model to recognize or classify speech quality using a famous architecture of deep learning called Convolutional Neural Networks (CNNs).

LIMITATIONS

Above discussed methods have enabled us to visualize various characteristics of speeches in the form of images but not in every case we are able to see same differences in mel spectrograms as we saw above. It is because there are various other characteristics of speaker which are playing role while plotting the spectrograms and each characteristic affect the lines and curves in spectrograms differently. And however, the spectrograms are basically images, but CNNs do not work very well on spectrograms, because x- and y- axis of spectrograms

denote different parameters (namely time and frequency), so unlike normal images, spectrograms are not independent of rotation, stretching and shifting. Hence, we cannot expect much accuracy while using CNNs for this task.

REFERENCES

[1] Rosenberg and J. Hirschberg, "Acoustic/prosodic and lexical correlates of charismatic speech," in Interspeech 2005. Proceedings of Eurospeech'05, September 2005, pp. 513–516.
[2] E. Strangert, "What makes a good speaker? subjective ratings and acoustic measurements," in Proceedings from Fonetik 2007: speech, music and hearing, quarterly progress and status report, TMH-QPSR, Vol 50, 2007, 2007, pp. 29–32.
[3] E. Strangert and J. Gustafson, "What makes a good speaker? subject ratings, acoustic measurements and perceptual evaluations." in Interspeech 2008, vol. 8, 2008, pp. 1688–1691.
[4] Soroosh Mariooryad, Anitha Kannan, Dilek Hakkani-Tur and Elizabeth Shriberg, "Automatic characterization of speaking styles in educational videos" in 2014 IEEE International Conference On Accostic, Speech And Signal Processing (ICASSP).