Deep learning approach for automatic speech recognition in the presence of noise

DOI: 10.46932/sfjdv5n7-009

Received on: Jun 10th, 2024
Accepted on: Jun 28th, 2024

Jane Oruh
PhD in Computer Sciences
Institution: Michael Okpara University of Agriculture
Address: Umudike, P.M.B., 7267, Umuahia, Abia State, Nigeria
E-mail: janeoruh@mouau.edu.ng

Serestina Viriri
PhD in Computer Sciences
Institution: University of KwaZulu-Natal
Address: King George V Avenue, Durban, South Africa
E-mail: viriris@ukzn.ac.za

Maliki Olaniyi
PhD in Mathematics
Institution: Michael Okpara University of Agriculture
Address: Umudike, P.M.B., 7267, Umuahia, Abia State, Nigeria
E-mail: maliki.niyi@mouau.edu.ng

ABSTRACT
Accurate recognition of speech in noisy environment is still an obstacle for wider application of speech recognition technology. The robustness of a speech recognition system is heavily influenced by the ability to handle the presence of background noise. In this research work, we propose a model based on Deep Fourier Neural Network (DFNN) for Automatic Speech Recognition (ASR) using LibriSpeech dataset. Most of the existing speech recognition techniques lack the robustness of handling background noise, as a result these techniques are not applicable in real-time. In order to mitigate the challenges of background noise, this research work proposes an efficient recognition technique which analyses in detail the raw audio waveforms using the Deep Fourier Neural Network (DFNN). This novel deep learning approach has a concise architecture and is an efficient approach for automatic speech recognition. The proposed deep learning approach embeds the Fourier transform, which is one of the most popular feature representations transform for audio signal processing. The Fourier transform extracts the core information from waveforms in the form of short term spectra of the speech signal as a function of time. The extracted short term spectra are analyzed deeply in the proposed DFNN model for accurate speech recognition in the presence of noise.

Keywords: LibriSpeech Dataset, Deep Learning, Deep Fourier Neural Network (DFNN), Automatic Speech Recognition (ASR), Noise Reduction.
1 INTRODUCTION

Speech is used to communicate information from a speaker to one or more listeners. The speaker produces a speech signal in the form of pressure waves that travels from speaker’s mouth to listener’s ears. The signal consists of variations in pressure as a function of time and is measured directly in front of the mouth, the primary sound source. The conversion of speech signal into a useful message (its corresponding text) is called Automatic Speech Recognition (ASR) (Nasreen, 2016).

Enhancement and recognition of speech signals in the presence of reverberation and noise, remains a challenging problem in many applications. Many past methods are prone to generating artifacts in the enhanced speech, and must trade off noise reduction against speech distortion. Recent approaches have started to address this issue, demonstrating improvements in both objective speech quality and automatic speech recognition (Delcroix, 2014).

Basically, the recorded speech signal in a real-world application may be corrupted by various noise types, interferences, echoes and reverberation resulting from the acoustic environment and enclosure. This degradation can significantly reduce the intelligibility of the speech signal by human listeners and also deteriorate the performance of speech coding and recognition systems (Cohen, 2009).

Improving intelligibility and/or overall perceptual quality of degraded speech signals using signal processing techniques, is the objective of speech enhancement (Parchami M. /.-P., 2016). Hence, a spectral filtering method based on the Short Time Fourier Transform (STFT) filtering are necessary for high performance speech enhancement system, as demonstrated in this work (Oruh, 2021).

1.1 PROBLEM STATEMENT

From our investigation, existing research on speech and audio processing has demonstrated the ability to obtain excellent performance when learning directly from raw audio waveforms using Fourier Neural Networks (FNNs). However, the exact inner operations of a FNN remain obscure to a large extent, which in a way hinders further developments and improvements towards this direction. In this work, we theoretically analyze and explain how deep FNNs learn from raw audio waveforms and identify potential limitations of existing network structures. In order to analyze raw audio wave forms in detail, we will employ Fourier analysis. To this end we propose novel network architecture called Deep Fourier Neural Network (DFNN), which will offer a simple but concise architecture with high model interpretability.
1.2 OBJECTIVES OF THE FOURIER NEURAL NETWORK

In applying the Fourier Neural Network to signal processing of the audio data, the specific objectives were considered:

1. to understand how Neural network operates;
2. to understand Fourier series and its transformation;
3. to study the Fourier neural network model.

1.3 GENERAL OBJECTIVES OF THE STUDY

In this work, we wish to extract features from a data-driven perspective by directly inputting raw sound waveforms into an end-to-end deep FNN architecture incorporating Fourier analysis for speech recognition purposes. We develop an algorithm for this type of FNN which we called Deep Fourier Neural Network (DFNN). The training of the DFNN will enable us to extract the weight from the first hidden layer and apply Fourier transform on the weight. Our objective is to be able to reconstruct the signal by reversing the first hidden layer and study its properties. In particular, we focus on applying deep models to problems that are realistic in the domain of speech recognition.

2 RELATED WORK

Many recent studies on speech enhancement have focused on spectral analysis of the speech signal. The analysis of the speech signal typically starts with a time dependent Fourier transform, also called Short-Time Fourier Transform (STFT), which is a method to analyze signals whose Fourier transform (i.e. spectrum) changes over time, as it is the case of the speech signal (Toledano, 2018).

Speech analysis for ASR systems typically starts with a STFT, which implies selecting a fixed point in the time-frequency resolution trade-off. In state-of-the-art ASR systems, the most widely used acoustic features; Mel Frequency Cepstral Coefficients (MFCC), Perceptual Linear Prediction Coefficient (PLP)) are based on the Short-Time Fourier Transform (STFT) (Tüske, 2011).

The short-time spectrum of the signal is the magnitude of a Fourier Transform of the waveform after it has been multiplied by a time window function of appropriate duration. To obtain a more useful representation of the speech signal in terms of parameters that contains relevant information in an efficient format, Short time Fourier transform amongst the other features, is best used to obtain the energy spectrum of the speech signal (Nasreen, 2016).
For STFT application, Parachami (Parchami M., 2016) stated that in the frequency domain speech enhancement, the spectrum of a clean speech signal is estimated through the modification of its noisy speech spectrum and then it is used to obtain the enhanced speech signal in the time domain, this can be compared to the approach deployed for the audio speech enhancement in this paper.

The STFT is an advanced method for ASR enhancement. It extracts the core information from waveforms in the form of short-time spectra of the speech signal as a function of time known as Short Time Fourier Transform (STFT). Simply, in the continuous-time case, the function to be transformed is multiplied by a window function which is non-zero for only a short period of time.

The Basic concept for Short-Time Fourier Transform is to break up the signal in time domain to a number of signals of shorter duration, then transform each signal to frequency domain. A window function of finite length is chosen, and the signal using the window is truncated. The Fourier Transform of the truncated window is computed and the result saved. Formulating the process of a continuous STFT is given by:

\[ STFT \ x(t)(\tau, \omega) \equiv X(\tau, \omega) = \int_{-\infty}^{\infty} x(t)w(t - \tau)e^{-j\omega t}dt \quad (1) \]

where:

- \( x(t) \) is the time-domain signal to be transformed,
- \( \tau \) is the time (slow time; lower resolution than \( t \)), and
- \( \omega \) is the frequency,
- \( W(t) \) is the window function, commonly a Hann window or Gaussian window bell centred around zero, and
- \( X(\tau, \omega) \) a complex function representing the phase and magnitude of the signal over time and frequency (this is essentially the Fourier Transform of \( x(t)\omega(t - \tau) \) a complex function representing the phase and magnitude of the signal over time and frequency (Ahmadizadeh, 2014).

One of the classic methods for joint time-frequency analysis is the short time Fourier transform (STFT). STFT is a mathematical transformation associated with \( FT \) to determine the frequency and phase of a sine wave in a local region of the time-varying signal. The concept of STFT is to first choose a window function with time-frequency localization. Then assume that the analysis window function \( \omega(t) \) was stationary over a short time, which ensures \( f(t)\omega(t) \) is a stationary signal within different time widths. STFT uses fixed window functions, the most commonly used include the Hanning window, the Hamming window, and the Blackman-Harris window (Jurafsky, 2020). The Hamming window, a generalized cosine window, is used in this article. It is usually represented as:
\[
\omega(t) = a_0 + (1 - a_0) \cos\left(\frac{2\pi t}{T}\right), 0 \leq t \leq 0
\] (2)

where:

\[a_0 = 0.53836.\] This function is a member of both the cosine-sum and power-of-sine families.

The Hamming window can efficiently reflect the attenuation relationship between energy and time at a certain moment (Solovyev, 2020). Although speech is a non-stationary signal, it is generally assumed to be quasi-stationary i.e., one with approximately constant statistics over short periods of time and, therefore, can be processed through a short-time Fourier analysis. Note that the modifier 'short-time', implies a finite-time window over which the properties of speech may be assumed to be stationary; it does not refer to the actual duration of the window. In speech processing, the Hamming window function is typically used and its width \(T_w\) is normally 20-40ms (Paliwal, 2005).

3 METHODS AND TECHNIQUES

3.1 DEFINING THE DATASET

Train-clean-360.tar.gz audio file from LibriSpeech ASR corpus was specifically chosen as the benchmark dataset for this work (Dutta, 2016). LibriSpeech is an ASR corpus of approximately 1000 hours of 16 kHZ read English speech based on public domain audio books. All the speech samples contained in the dataset are in .wav format, and have been carefully segmented and aligned. The corpus is freely available under the very permissive CC BY 4.0 license (Commons, {Accessed: 2017-11-07}), and there are example scripts in the open source Kaldi ASR toolkit that demonstrate how high-quality acoustic models can be trained on this data (Panayotov, 2015).

3.2 DATASET SAMPLING AND PRE-PROCESSING

Speech is an analog signal which should be converted to digital form. Sampling and quantization are performed to convert the continuous signal into a series of discrete values. The train-clean-360 is read as an audio file and sampled as a speech waveform, as shown in Figure 1.
3.3 PROCESSING OF THE SPEECH SIGNAL

Once the speech signal has been sampled, the next phase is signal processing to separate the speech signals from background noise (Sarma, 2019). Here, the audio speech signal is filtered using STFT technique, and converted to a mono signal waveform as shown in Figure 2. The standard STFT was used to perform a filtering operation on the speech waveform as demonstrated in Figure 2 to enhance the speech signal. In this case the analysis window $w[n]$ plays the role of the filter impulse response.

To illustrate the view, we fix the value of $\omega$ at $\omega_0$ and rewrite it as:

$$X(n, \omega_0) = X \infty m - \infty (x[m]e^{-j\omega_0 m})w[n - m]$$  

(3)

Which can be interpreted as the convolution of the signal $(x[n]e^{-j\omega_0 n})$ with the sequence
\[ w[n]: X(n, \omega_0) = (x[n] e^{-j\omega_0 n}) * w[n] \]  

(4)

and the product \( x[n] e^{-j\omega_0 n} \) can be interpreted as the modulation of \( x[n] \) up to frequency \( \omega_0 \) (i.e., per the frequency shift property of the FT) (Gutierrez-Osuna, 2016).

4 METHODOLOGY

The seven principles below define to a large extent our deep learning methodology. Ideally, deep learning/machine learning models must:

- have multiple compositional levels of non-linear feature extractors;
- learn distributed representations (Rumelhart, 1986) of their input;
- be sufficiently expressive and have large enough capacity to solve interesting practical problems;
- have learning procedures that scale close to linearly with the number of training cases;
- scale to thousands of dimensions (although not necessarily to thousands of intrinsic dimensions);
- be able to make use of unlabeled data and weakly labeled data;
- allow efficient approximate inference procedures.

The above restrictions can make the task of developing acceptable algorithms very difficult. However, our approach will be systematic, direct and simple enough.

We will employ Fourier transforms in this research work, and it is worthy to note that Fourier transform is one of the most popular feature representation transforms for audio signal processing (Hinton, 2012). In stationary signals where all frequency components exist throughout the entire duration, Fourier transform could extract all the required information from waves. However, for non-stationary signals, some frequency components do not appear during the entire span of the wave (Weng, 2015). In order to deal with this problem, short-time Fourier transform (STFT) was developed to extract feature components assuming that the signal is relatively stable within a single short time window. Still, there is a trade-off between the resolution and temporal dynamics depending on the window size.

Many methods have been proposed to deal with the task of the spectral analysis of the speech signal. Some of them have a psychophysical foundation, that is, they are based on physiological research on human hearing. Others have arisen in other fields of engineering but have proved to be adequate for this task. Certainly one of the simplest but also more powerful approaches is computing a short term Fourier spectrum of the speech signal. An aspect of Fourier analysis applicable to speech signals was discussed briefly in section 4.1.
4.1 FOURIER ANALYSIS

Given a data set \( x = (x(0), x(2), ..., x(n-1)) \), it is the task of Fourier analysis to reveal its periodic structure. We can think of the data set as function \( X \) evaluated at the points 0, 2, ..., \( n-1 \). The function \( X \) can be written as a linear combination of the basis functions;

\[
f_0(t) = \frac{1}{\sqrt{n}} \cos \left( \frac{2\pi t}{n} \right) - i \sin \left( \frac{2\pi t}{n} \right) = \frac{1}{\sqrt{n}} (\omega_n^*)^0 t
\]

\[
f_1(t) = \frac{1}{\sqrt{n}} \cos \left( \frac{2\pi t}{n} \right) - i \sin \left( \frac{2\pi t}{n} \right) = \frac{1}{\sqrt{n}} (\omega_n^*)^1 t
\]

\[
\vdots
\]

\[
f_{n-1}(t) = \frac{1}{\sqrt{n}} \cos \left( \frac{2\pi t}{n} \right) - i \sin \left( \frac{2\pi t}{n} \right) = \frac{1}{\sqrt{n}} (\omega_n^*)^{(n-1)-t}
\]

(5)

where:

\( \omega_n \) denotes the \( n \)th complex root of unity \( \omega_n = \exp \left( -\frac{2\pi}{n} \right) \).

Writing the data set as a linear combination of these functions amounts to finding which of the given frequencies is present in the data. Denote by \( F_n^* \) the \( n \times n \) matrix whose columns are the basis functions evaluated at \( t = 0,1,\ldots,n-1 \) that is the element at row \( i \) and column \( j \) of \( F_n^* \) is \( (\omega_n^*)^j \) for \( i, j = 0,1,\ldots,n-1 \).

We are looking for a vector \( a \) of amplitudes such that

\[
F_n^* a = x
\]

(6)

The \( n \) dimensional vector \( a \) is the spectrum of the speech signal. The matrix \( F_n^* \) defined as:
\[ F_n = \frac{1}{\sqrt{n}} \begin{pmatrix} \omega_n^0 & \omega_n^0 & \cdots & \omega_n^0 \\ \omega_n^0 & \omega_n^1 & \cdots & \omega_n^{n-1} \\ \vdots & \vdots & \ddots & \vdots \\ \omega_n^0 & \omega_n^{n-1} & \cdots & \omega_n^{(n-1)(n-1)} \end{pmatrix} \]  

(7)

is the transpose conjugate of the matrix \( F_n^* \). Since the basis functions \( f_1, f_2, f_{n-1} \) are mutually orthogonal this means that \( F_n^* \) is unitary and in this case

\[ F_n F_n^* a = x \quad \Rightarrow \quad a = F_n x \quad (8) \]

The expression \( F_n x \) is the discrete Fourier transform of the vector \( x \). The inverse Fourier transform is given of course by

\[ F_n^* F_n = F_n^* a \quad \Rightarrow \quad x = F_n a \quad (9) \]

The speech signal is analyzed as follows: a window of length \( n \) is used to select the data. Such a window can cover for example 10 milliseconds of speech. The Fourier transform is computed and the magnitudes of the spectral amplitudes (the absolute values of the elements of the vector, \( a \) ) are stored. The window is displaced to cover the next set of \( n \) data points and the new Fourier transform is computed. In this way we get a short term spectrum of the speech signal as a function of time. To that extent our speech recognition algorithms should recover from this kind of information the correct sequence of phonemes.

4.2 FAST FOURIER TRANSFORMATIONS

Since we are interested in analyzing the speech signal in real time it is important to reduce the required number of numerical operations. A Fourier transform computed as a matrix vector multiplication requires around \( O(n^3) \) multiplications. A better alternative is the Fast Fourier Transform (FFT) which is just a rearrangement of the matrix vector multiplication. This rearrangement of the Fourier matrix is the basis of the Fast FFT which we hope to employ in the course of this research work. It is interesting to know that many speech recognition systems use some kind of variation of the Fourier coefficients.
4.3 DEEP NEURAL NETWORKS

It is worth mentioning the meaning of the term Deep Learning. The purpose of the earliest artificial intelligence algorithms was to model how learning happened in the brain, thus these algorithms were called artificial neural networks (ANNs). Deep Learning (DL) is the term applied to the algorithms not necessarily inspired by the biological neural networks (Goodfellow, 2016) and contains many more processing layers than a traditional neural network.

Most definitions of DL highlight the use of models with multiple layers of nonlinear processing units that process and transform the input data. These sequential transformations create a feature representation hierarchy, which can be supervised or unsupervised (Deng, 2014). These models can be seen as an infinitely flexible function, which can, for example, translate a language to another or recognize faces in pictures. To allow this function to undertake its task, fitting its parameters is required, which is achieved by the technique called backpropagation. Finally, and the reason why DL is so popular today is the recent availability of, first, devices that allow to fit these parameters quickly, Graphical Processing Units (GPUs), and, second, large databases that allow to scale the algorithms.

4.4 THE MODIFIED DISCRETE FOURIER NEURAL NETWORK

Consider the discrete Fourier transform model given by the following equation:

\[ f(x_1, ..., x_2) = \sum_{y_1, ..., y_n} c(y_1, ..., y_n) e^{i(y_1 x_1 + \cdots + y_n x_n)} \]

(10)

We modify the above formula in two ways in order to make it more suitable for computer implementation.

First we employ the cosine representation for the exponential \( e^{i(y_1 x_1 + \cdots + y_n x_n)} \), and we will also filter the output through the sigmoid function \( \Phi \) in order to obtain outputs in the interval \([0; 1]\). So the final form of the output function for our neuron is:

\[ f(x_1, ..., x_2) = \Phi\left(\sum_i c_i \prod_{j=1}^n \cos(w_{ij} x_j + \phi_i)\right) \]

(11)

The diagram for the deep FNN is depicted in Figure 3.
SUMMARY

The approach in this work aims to improve the speech quality alongside the intelligibility of the train-clean-360 audio file by time-frequency analysis, filtering, and noise reduction of the speech signal using the STFT techniques analysis. Since we are interested in analyzing the speech signal in real time it was necessary to reduce the required number of numerical operations. This was achieved using a Fast Fourier Transformation. The enhanced speech signal obtained in Figure 2 was subjected to further analysis using the proposed DFNN technique. The training with DFNN enabled us to extract the weight from the first hidden layer and apply Fourier transform on the weight.

CONCLUSION

In this work, we had theoretically analyzed and explained how DFNNs learn from raw audio waveforms of train-clean-360 of the LibriSpeech ASR dataset and identified potential limitations of existing network structures. Furthermore, a novel network architecture called Deep Fourier Neural Network (DFNN) was proposed that will offer a simple but concise architecture with high model interpretability. The proposed DFNN will be used to analyze in detail the raw audio waveforms thereby reduce the challenges of background noise. This research hopes to make significant research contributions to speech recognition in the domain of deep learning and machine learning applications. The primary focus of the study is to apply deep learning models to problems that are realistic in the domain of speech recognition.
REFERENCES


Commons, C. (Accessed: 2017-11-07). Creative Commons Attribution 4.0 International(CC BY 4.0) License.


Gutierrez-Osuna, R. (2016). Introduction to speech processing. *CSE@ TAMU*.


