Chapter 4
Application of Deep Learning in Speech Recognition

Rekh Ram Janghel
NIT Raipur, India

Satya Prakash Sahu
NIT Raipur, India

Yogesh Kumar Rathore
NIT Raipur, India

Shraddha Singh
NIT Raipur, India

Urja Pawar
NIT Raipur, India

ABSTRACT

Speech is the vocalized form of communication used by humans and some animals. It is based upon the syntactic combination of items drawn from the lexicon. Each spoken word is created out of the phonetic combination of a limited set of vowel and consonant speech sound units (phonemes). Here, the authors propose a deep learning model used on tensor flow speech recognition dataset, which consist of 30 words. Here, 2D convolutional neural network (CNN) model is used for understanding simple spoken commands using the speech commands dataset by tensor flow. Dataset is divided into 70% training and 30% testing data. While running the algorithm for three epochs average accuracy of 92.7% is achieved.

DOI: 10.4018/978-1-5225-7862-8.ch004

INTRODUCTION

Speech is "the vocalized form of communication used by humans and some animals, which is based upon the syntactic combination of items drawn from the lexicon. Each spoken word is created out" of the phonetic combination of a limited set of vowel and consonant speech sound units (phonemes). The 30 words included in the database differ from person-to-person such that their accent, their speaking frequency differentiates one person from the other. Speech recognition is the inter-disciplinary sub-field of computational linguistics. It develops methodologies and technologies that enable the recognition and translation of spoken language into text by computers shown in Figure 1.

It is also known as "automatic speech recognition" (ASR), "computer speech recognition", or just "speech to text" (STT). It incorporates knowledge and research in the linguistics, computer science, and electrical engineering fields."

The "spectrogram is a basic tool in audio spectral analysis and other fields. It has been applied extensively in speech analysis (Deller, Proakis & Hansen, 1993; Schafer & Markel, 1979). The spectrogram can be defined as an intensity plot (usually on a log scale, such as dB) of the Short-Time Fourier Transform (STFT) magnitude. The STFT is simply a sequence of FFTs of windowed data segments, where the windows are usually allowed to overlap in time, typically by 25-50% (Allen & Rabiner, 1977). It is an important representation of audio data because human hearing is based on a kind of real-time spectrogram encoded by the cochlea of the inner ear (O'Shaughnessy, 1987). The spectrogram has been used extensively in the field of computer music as a guide during the development of sound synthesis algorithms. When working with an appropriate synthesis model, matching the spectrogram often corresponds to" matching the sound extremely well. In fact, spectral modeling synthesis (SMS) is based on synthesizing the short-time spectrum directly by some means (Zölzer, 2002).

Fast "Fourier Transform (FFT)-based computations are more accurate than the other slow transforms as the functions applied are different in FFT. Discrete Fourier transforms computed through the FFT are more accurate than slow transforms and the convolutions computed with the help of FFT are more accurate than the directly acquired results." Nonetheless, these results are critically dependent on the employed FFT software's accuracy, which should generally be considered suspect. Due to inherent instability, some popular recursions for fast computation of trigonometric table (or twiddle factors) are inaccurate. FFT is highly stable even in the higher dimensions (Schatzman, 1996).

Mel frequency cepstral coefficient (MFCC) has become a standard speech recognition system and is most popular due to the high efficiency of computation schemes available for it and due to its robustness in the presence of different types of noises. In the computation process of MFCC, we pass the voice signal through various triangular filters. These triangular filters are placed in a perceptual Mel scale linearly

Figure 1. Voice recognition