Düzce University
Journal of Science & Technology

Review Article

A Convolutional Neural Network Model Implementation for Speech Recognition

Şafak KAYIKÇI [a,*]

[a,*] *Department of Computer Engineering, Faculty of Engineering, Abant Izzet Baysal University, Bolu, TURKEY*
** Sorumlu yazarın e-posta adresi: safak.kayikci@ibu.edu.tr*

DOI : 10.29130/dubited.567828

ABSTRACT

Speech recognition is the capability of an appliance to analyze vocable and diction in a phonetic language and turn them into a machine comprehensible arrangement. It is an interdisciplinary subfield of linguistics, computer science and electrical engineering that establishes processes and techniques that understands and converts speech to text. This paper presents a convolutional neural network model for recognition of speech data.

Keywords: *Speech Recognition, Deep Learning, Confusion Matrix*

Konuşma Tanıma için Bir Evrimsel Sinir Ağı Modeli Uygulaması

ÖZET

Konuşma tanıma, bir cihazın fonetik bir dilde kelime bilgisi ile diksiyonu analiz etme ve bunları makinenin anlaşılır bir düzenine dönüştürebilme kabiliyetidir. Konuşmayı anlayan ve metne dönüştüren süreç ve teknikleri oluşturan disiplinlerarası bir dilbilim olup bilgisayar bilimi ve elektrik mühendisliği alt alanıdır. Bu çalışmada konuşma verilerinin tanınması için evri bir sinir ağı modeli sunulmaktadır.

Anahtar Kelimeler: *Konuşma Tanıma, Derin Öğrenme, Karışıklık Matrisi*

I. INTRODUCTION

SPEECH recognition is the capability of an appliance to analyze vocable and diction in a phonetic language and turn them into a machine comprehensible arrangement. It is an interdisciplinary subfield of linguistics, computer science and electrical engineering that establishes processes and techniques that understands and converts speech to text. Speech recognition uses both acoustic and language modeling algorithms. Acoustic model demonstrates the association between linguistic elements of communication and audio signals whereas language model pairs voice with character arrays in order to differentiate between alike pronounced words. The development of big data analysis and deep learning has accelerated the work in this field. In deep learning, a convolutional neural

Received: 20/05/2019, Revised: 24/06/2019, Accepted: 06/07/2019

network (CNN or ConvNet) is deviation of multilayer perceptron which are influenced by connection stencils between biological neurons. CNNs built on preprocessing algorithms that learns from preceding knowledge and have other various application areas like video recognition, recommender systems, image classification and natural language processing.

II. History And Related Work

In 1950s, a digit recognition system named “Audrey” [1] was developed by S.Balashek, R.Bidduplh and H.Davis at Bell Research Labs. After that, Soviet researchers developed a time warping (DTW) algorithm which is able to process on 200 word vocabulary. In 1971, BBN, IBM, Carnegie Mellon and Stanford Research Institute aided in a project supported by DARPA. Its objective was to diagnose conversation at least in a thousand words. In the middle of 1980s, Fred Jelinek from IBM developed a sound actuated typist Tangora [2]. It was able to manage 20,000 vocals.

Two conversation programs (EARS and GALE) was developed at 2000s by DARPA. Google's first experience started with transferring Nuance scientists. Its voice search is currently available for thirty languages. At the beginning, speech recognition was controlled by conventional algorithms like Hidden Markov Models joined with neural networks. After all, most of speech recognition researches are substituted by long short-term memory (LSTM) deep learning method [3]. But this methods still have weaknesses among Gaussian mixture model/Hidden Markov model (GMM-HMM) which is consist of productive methods.

III. Model And Implementation

TensorFlow [4] latterly presented Speech Commands Datasets which has 65,000 with a second duration announcements of thirty little words [5]. In this study, only three outcomes of dataset used as classifiers which are “bed”, “cat” and “happy”.

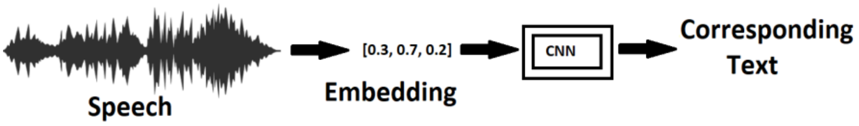


Figure 1. *Workflow diagram of model*

Embedding is mapping distinct entities like words to vectors of real numbers. Embeddings are essential inputs for deep learning and outcomes are depends on real numbers. All objects are assigned by compact vectors to contribute training. Embedding functions converts input items into continuous vectors effectively. Systems also use resemblance in vector area (like Euclidean distance or the intersection angle) for a concentrated and adjustable metric of object alikeness. Most common technique is finding nearest neighbors.

Nyquist rate [6] is the double value of bandlimited function or channel. It is defined either by as a lower bound for the sample rate for alias-free signal sampling or as an upper bound for the symbol rate across a bandwidth-limited baseband channel. There are consistently different countless number of continuous functions that suits the same example space with a continuous function, *x(t)* at a constant rate, fs samples/second. Nevertheless, only one of them is bandlimited to ½ fs cycles/second (hertz) which is its Fourier transform. The sound wave is an acoustic pressure wave. The speech signal is generated when the air blown through the lungs passes through the human voice production mechanism. In the human voice production mechanism are mainly vocal cords, palate, teeth and lip. The voice path starts from the larynx exit, ends on the lips, and the speech signal is formed along this path.

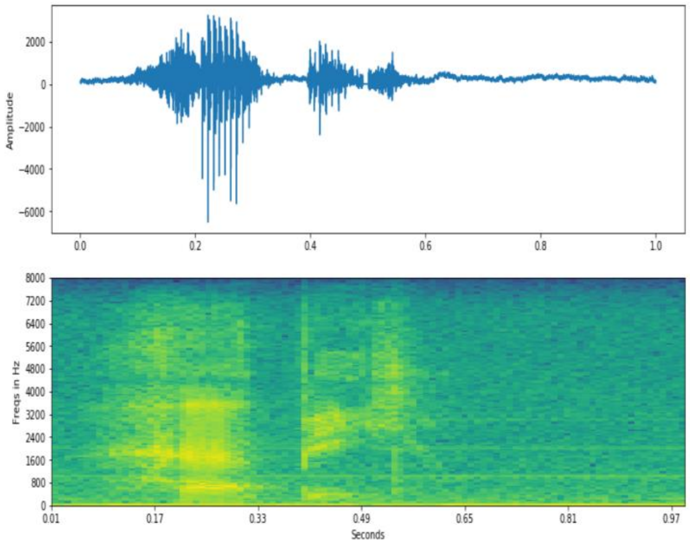


***Figure 2. (a)** Raw Wave and **(b)** Spectrum of Speech File*

In other words, the speech is formed according to the position and postures of the anatomical structures completely located on the sound path. This diversity is the basic element that allows different sounds to emerge. As the air from the lungs passes through the larynx, the vibrations it creates on the vocal cords create very weak sounds. Then, with the other factors on the audio path, the voice reaches the basic form and the production of the speech signal is completed. On the basis of the separation of the speeches, the events that take place in the voice production mechanism are involved. The mathematically good modeling of the human voice production mechanism is important for speech processing.

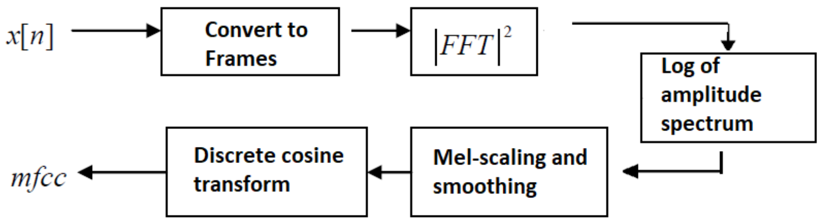

$x[n]$
Convert to Frames
$|FFT|^2$
Log of amplitude spectrum
Mel-scaling and smoothing
Discrete cosine transform
$mfcc$

***Figure 3.** Block diagram of revealing MFCC attributes*

MFCC, which is used effectively for the extraction of the speech signals, models the human ear rather than the human voice path. Modeling of the human ear was performed experimentally and Mel scale was formed. For the creation of this scale, 1000 Mels was taken as the reference for the 1000 Hz value. Herz-Mels values were determined experimentally according to the scale that the human ear perceived. As a result of this experiment, Herz-Mels values were linearly matched at low frequency (up to 1 kHz) and high frequency values were observed to be logarithmic. In the MFCC extraction process, first the portions of the default length, in which the speech signal is stationary, are selected to overlap in the successive windows. In this way, the transition between frames and change becomes more uniform. The speech signal transfer function is then passed through a filter with a finite impulse response (Finite Impulse Response - FIR). This process, called pre-emphasization, is used to make the high frequencies of the speech signal more pronounced.

A. CONVERT TO FRAMES

The aim of framing and windowing is to reduce the spectral effects. Discontinuity in frames is prevented by windowing. In this way, the central regions of the sound is strengthened while the edge regions are weakened. The mathematical expressions of the widely used are Hamming, Hanning, Blackman, Gauss, rectangular and triangular windowing functions.

B. FAST FOURIER TRANSFORM (FFT)

In obtaining the MFCC, the amplitude spectrum of the signal passed through the window is calculated by the FFT. Each frame in the time domain consisting of the FFT and N samples is converted into the frequency field. FFT, discrete fourier conversion. Discrete fourier transform of a frame equation is defined in equation (1).

$$X[k] = \sum_{n=0}^{N-1} x_n . e^{-2\pi jkn/N}, k = 0,1,2,...,N-1 \tag{1}$$

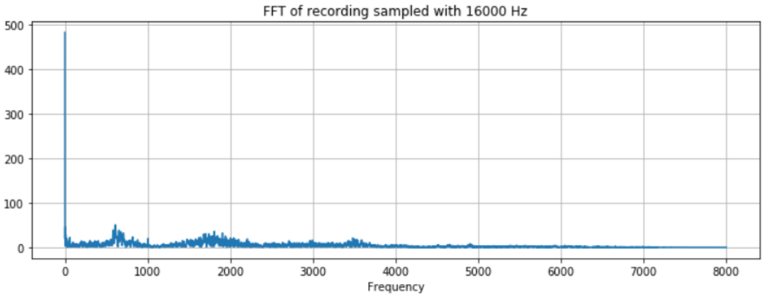


***Figure 4**. Fast Fourier Transform with 16000 Hz*

Generally *X[k]* refers to complex numbers. The resulting sequence *{X [k]}:* the zero frequency corresponds to $k = 0$ and the positive frequencies correspond to $(0 < f < fs / 2)$ and $1 \leq k \leq (N/2) - 1$ values, while the negative frequencies correspond to $(-fs / 2 < f < 0)$, $(N / 2) + 1 \leq k \leq N-1$ where fs is the sampling frequency. When calculating the FFT of a sign, the length of the sign must be 2^M $M \in N$, in other words it must be the power of two.

C. LOG OF THE AMPLITUDE SPECTRUM

Speech signal can be expressed *as* $|S(e^{jw})| = |X(e^{jw})| \ |F(e^{jw})|$. Here, S, X and F correspond to the speech signal, source and filter. The source represents the unmarked audio signal produced by the vocal cords. Strainer refers to the path followed by the sound path. Logarithm is used to separate the effect of the sound path from the source. By taking logarithms, the product is converted to the sum of the components $log|S(e^{jw})| = log|X(e^{jw})| + log|F(e^{jw})|$. The logarithmic spectrum can be considered as a combination of components with different frequencies. Then, by applying the reverse FFT to these two components, it is possible to have knowledge about the fast and slow changing components.

D. MEL SCALING AND SMOOTHING

The Mel scale kepstrum coefficients were first defined by Davis and Mermelstein [7]. They took the spectra of the sign's amplitude and passed through a triangular strainer array. The number of filters FS is defined as the selected signal bandwidth *[0, fs/2]* Hz and fs sampling frequency. Davis and Mermelstein defined by the center of the first 10 filters in the linear frequency, the next 10 filters were placed logarithmically. All filters have equal amplitude. Slaney's method of obtaining MFCC is widely used in speaker recognition applications in recent years [8]. Slaney placed 40 filters in the frequency range of 133-6854 Hz. The center frequency of the first thirteen strainers was in the range of 200-1000 Hz, spaced at 66.67 Hz. The center frequencies of the remaining twenty-seven filters were placed in the logarithmic step 1.0711703 in the range 1071-6400 Hz.

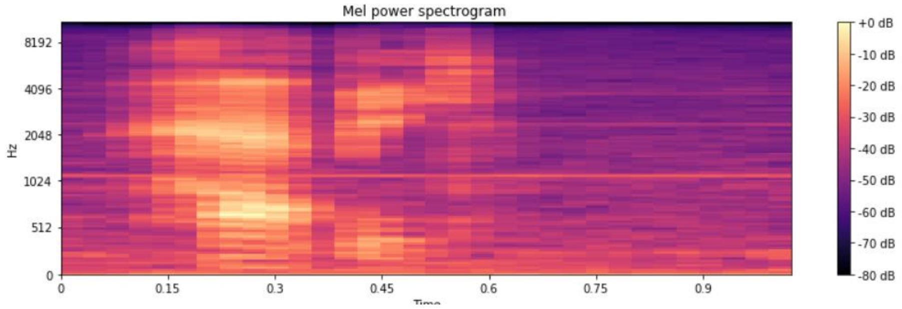


Figure 5*: Mel power spectrogram*

The amplitude of the filter sequences proposed by Slaney varies inversely with the bandwidth of the filter. That is, if the bandwidth of the filter is small (linear Mel scale region below 1000 Hz), the amplitude of the filter is large, if the bandwidth of the filter is large (logarithmic Mel scale region above 1000 Hz), the amplitude of the filter is small.

E. DISCRETE COSINE TRANSFORM

The cepstrum coefficients are calculated in the finalization of MFCC. With cepstral representation, spectral shape changes due to recording and transmission environment are removed. In addition, cepstral coefficients show a high degree of statistical independence and give a higher recognition rate than amplitude spectrum representation. The real cepstrum is defined as the inverse fourier transformation of the logarithmic amplitude spectrum and is calculated using the cosine transformation for real signs.

E. CONVOLUTIONAL NEURAL NETWORK IMPLEMENTATION

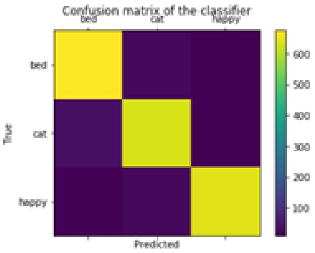
Convolutional neural network is comprised different layers like input, output and hidden layers. Each layer has different functionality [9]. Hidden layers generally related with activation functions, normalization, pooling and fully connectivity. Each convolutional layer executes a convolution function to its inputs and transfers the conclusion to next layer. Every neuron is responsible for its approachable data area. CNN architectures may composed of local or global bands. These bands cut down the magnitude of data by joining the output of neuron clusters to single band. Local pooling joins tiny arrays whereas global pooling deals with all neurons in model. Additionally, pooling may measure the maximum or average value from every previous neuron cluster. Fully connectivity is similar to conventional multilayer perceptron neural network (MLP). It bridges each neuron of a layer to each neuron of next layer. The smoothed data vector is transmitted all over fully connected layers.

III. Results And Discussion

Confusion matrix is a kind of grid outline and contingency table that visually indicates the performance of a model [10]. A confusion matrix is a table commonly used to describe the performance of a classification pattern on a set of test data, in which actual values are known.

Table 1. *Confusion Matrix of Table*

	bed	**cat**	**happy**
bed	675	25	10
cat	38	635	11
happy	15	23	644



Accuracy rate is a measure of how often the classifier guesses correctly and error Rate is how often the classifier has incorrectly estimated it. True positive rate shows how much the true positive value the classifier estimates. Sensitivity is also known as hit rate must be as high as possible. Here we calculated recall value as 0.95 and precision is 0.93. Specificity or selectivity is a measure of how much the true negative value of the classifier is correct. Positive predictive value is a measure of how accurately all classes are estimated. It should be as high as possible either. Prevalence is the measure of how often 1 is found at the end of the estimation. F Score is the harmonic mean of the ratio of the true positive values and the precision. It is a measure of how well the classifier performs and is often used to compare classifiers. We calculate an F-measure 0.94 which is a successful value for model.

Speech recognition discipline within the field of voice recognition is a system that tries to gain an important place in the process of developing technology and it is the process of recognizing and recognizing human voice by a computer through a microphone. This process is an important need in human-computer communication. Because now people want to print or do something on the computer without using the keyboard. In this paper, a convolutional neural network model for recognition of speech data is presented across Tensowflow Speech Commands Data Set as an review work.

V. References

[1] K. Davis , R. Biddulph, and S. Balashek "Automatic Recognition of Spoken Digits", *The Journal of the Acoustical Society of America*, vol. 24, no. 6 , pp. 637-642, 1952.

[2] S. Das, M. A. Picheny, *In Automatic Speech and Speaker Recognition*, Boston, USA: Springer, 1996, pp. 457-479

[3] S. Hochreiter, J. Schmidhuber, "Long short-term memory", *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 1997

[4] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean and M. Kudlur “Tensorflow: A System for large-scale machine learning”, *12th Symposium on Operating Systems Design and Implementation (OSDI),* Savannah, GA, USA, 2016, pp. 265-283

[5] Tensowflow Speech Commands Data Set v0.01 (2019, 01 April). [Online]. Erişim: https://www.kaggle.com/c/tensorflow-speech-recognition-challenge/data

[6] H. Nyquist, “Certain topics in telegraph transmission theory”, *Transactions of the American Institute of Electrical Engineers*, vol. 47, no. 2, pp. 617-644, 1928

[7] Davis, Steven, and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences”, *IEEE transactions on acoustics, speech, and signal processing*, vol. 28, no. 4, pp. 357-366, 1980

[8] Slaney, Malcolm, Michele Covell, and B. Lassiter, “Automatic audio morphing”, *International Conference on Acoustics, Speech, and Signal Processing Conference (IEEE)*, 1996, pp. 1001-1004

[9] S. Postalcioglu, “Performance Analysis of Different Optimizers for Deep Learning-Based Image Recognition”, *International Journal of Pattern Recognition and Artificial Intelligence*, 2019

[10] Townsend, T. James “Theoretical analysis of an alphabetic confusion matrix”, *Perception & Psychophysics*, vol. 9, no. 1, pp. 40-50, 1971