IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Audible sound analysis based on state clustering through multiple deep Neural Network Modeling

Ms.Sneha S. Thune
PG Student:
Department of ETC engineering
Priyadarshini Bhagwati College of Engineering,Nagpur

Dr.N.K.Choudhari[1] Ms.A.P.Thakare[2]
Professor [1]Assistant Professor[2]
Department of ETC Engineering
Priyadarshini BagwatiCollege of engineering, Nagpur

Abstract ***Speech sign processing is an emerging area to observe the processing of speech signals with digital sign processing generation and phonological understanding. It is one of the core technologies inside the subject of records technological know-how studies. Passing records via voice is the maximum essential, simplest,most typically used and handy trade form of information. The Matlab language is a totally powerful computer software software program for facts evaluation and processing. It can rework sound files into discrete facts documents after which use their powerful matrix computing strength to process records along with digital filtering, Fourier remodel, domain and frequency domain evaluation, sound playback and a variety of pictures. Its sign processing and analysis toolbox for the voice signal evaluation gives a completely rich function, using these features can speedy and without difficulty entire the voice signal processing and analysis, in addition to sign visualization,making human-pc interplay greater convenient. Signal processing is one of the crucial applications of Matlab. This design is geared toward most of the voice processing software content, operation inconvenience and different issues. The use of MATLAB 7.Zero incorporated GUI interface design, a diffusion of characteristic calls to achieve the voice signal frequency, amplitude, fourier remodel and filtering, concise interface, clean to perform. All these have certain practical significance. Finally,this paper places forward his very own perspectives on the further improvement of speech sign processing.***

\

I. INTRODUCTION

The goal of text-to-speech (TTS) synthesis is to render a Naturally sounding speech waveform given a textual content to be synthesized. Figure 1 outlines a human speech manufacturing process. A textual content (or idea) is first translated into actions of articulators and organs. Using air-waft from a lung, vocal supply excitation signals containing periodic (via vocal wire vibration) and aperiodic (by turbulent noise) additives are generated.

By filtering the supply signals by way of timevarying vocal tract switch capabilities controlled via the articulators, their frequency traits are modulated. Finally, the filtered source indicators are emitted. The intention of TTS is to imitate this process through computers in a few way. Text-to-speech can be regarded as a chain-to-series mapping trouble; from a series of discrete symbols (text) to a realvalued time collection (waveform).

Typical TTS structures encompass text analysis and speech synthesis parts. The text evaluation component consists of some of herbal language processing (NLP) steps, along with word segmentation, text normalization, element-of-speech (POS) tagging, and grapheme-to-phoneme (G2P) conversion. This element plays a mapping from a chain of discrete symbols to any other collection of discrete symbols (e.G., sequence of characters to collection of words). The speech synthesis component plays mapping from a chain of discrete symbols to actual-valued time collection.

It consists of prosody prediction and speech waveform era. The former and latter parts are often known as “the front-give up” and “again-give up” in TTS, respectively. Although each of them are crucial to achieve fantastic TTS structures, this paper specializes in the latter one. Statistical parametric speech synthesis (SPSS) is one of the most important techniques in the returned-stop element. This approach makes use of an acoustic version to represent the relationship among linguistic and acoustic capabilities and a vocoder to render a speech waveform given acoustic capabilities. This technique offers various benefits over concatenative speech synthesis , which is some other major technique within the text (concept) frequency switch traits importance start--give up fundamental frequency air drift Sound supply voiced: pulse voiceless: noise speech Outline of speech production manner. Back-end part of TTS structures, consisting of small footprint and flexibility to change its voice characteristics However, the naturalness of the synthesized speech from SPSS isn't as accurate as that of the great samples from concatenative speech synthesizers. Zen et al. Pronounced three predominant factors that may degrade the naturalness exceptional of vocoder, accuracy of acoustic version, and effect of over smoothing. This paper addresses the accuracy of acoustic version. Although there had been many attempts to develop a extra accurate acoustic model for SPSS],

the hidden Markov model (HMM) is the most famous one. Statistical parametric speech synthesis with HMMs is referred to as HMM-based totally speech synthesis Inspired from the fulfillment in system getting to know and automated speech reputation, 5 distinct forms of synthetic neural community primarily based acoustic models were proposed in 2013.

Highly correct audio classifiers, if exist, have many realistic applications in all walks of our life, from medicinal drug to enterprise. Developing such correct classifiers, however, is onerous. Unlike in computer vision, improvements in laptop listening are in early degrees. Audio classifiers, in contrast to photo classifiers, usually have decrease accuracies. However, with the ready availability of curated, public audio datasets and ML type algorithms, it's far easier than ever to construct accurate classifiers.Several studies papers recently posted classifiers at the UrbanSound8k dataset. However, these classifiers most effective have 50-seventy nine% accuracy variety. In this project, with the aid of employing various ML techniques, I intention to noticeably enhance this accuracy from its modern-day high of seventy nine%. To acquire generalizable and dependable consequences, for ML version education, I will use industry gold-standard, okay-fold-move-validation at the schooling set, that's 80% of the supply records. The skilled model might be examined on the take a look at set, that's the final 20% unseen supply information. Experiments might be repeated and even run on different working structures to degree for variability.A standard convolutional neural community includes some of one of a kind layers stacked together in a deep architecture: an enter layer, a collection of convolutional and pooling layers (which may be mixed in numerous methods), a restrained number of absolutely linked hidden layers, and an output (loss) layer. The actual difference, while compared to the multilayer perceptron, lies within the advent of a combination of convolution and pooling operations A convolutional layer introduces a special manner of organizing hidden devices which ambitions to take gain of the local shape gift inside the two-dimensional enter information (by and large, but no longer limited to, photos). Each hidden unit, instead of being connected to all the inputs coming from the preceding layer, is confined to processing handiest a tiny part of the entire input area (e.G. Small 3 × three blocks of pixels), referred to as its receptive field. The weights of this kind of hidden unit create a convolutional kernel (filter) which is applied to (tiled over) the complete enter space, resulting in a characteristic map. This way, one set of weights may be reused for the entire enter space. This is based on the idea that regionally useful capabilities can be additionally useful in different locations of the enter area - a mechanism which now not most effective massively reduces the range of parameters to estimate, however improves robustness to translational shifts of the facts.

A typical convolutional layer will consist of numerous filters (feature maps). Further dimensionality reduction can be achieved through pooling layers, which merge adjacent cells of a feature map. The most common pooling operations performed are taking the max (winner takes all) or mean of the input cells. This downsampling further improves invariance to translation

II BLOCK DIAGRAM

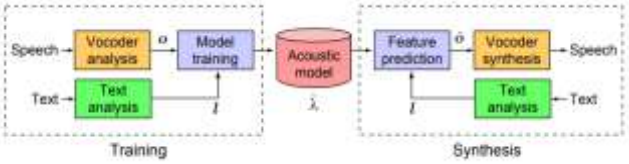

Fig 1 Block Diagram of Overall System

Background and significance of the subject

Speech signal processing is a more practical professional course of electronic engineering. Voice is an important source of human access to information and has an important means of using information. Passing messages to each other through language is one of the most important basic functions of mankind. Language is a human-specific function, it is to create and record thousands of years of human civilization. It is the fundamental means of history, there is no today's human civilization without language. Speech is the acoustical representation of language and most important means of transmitting information to each other. It is the most important, most effective, most commonly used and most convenient exchange of information form. Speech signal processing is a discipline that studies the processing of speech signals with digital signal processing technology. It is a new discipline, comprehensive multidisciplinary field and a wide range of interdisciplinary disciplines.

Basic overview of the system

Graphical user interface (GUI) also known as graphical user interface refers to the use of graphical display of computer operating user interface. Compared to the command line interface used by earlier computers, the graphical interface is visually more acceptable to the user. Wide usage of GUI in today's computer development is one of its major achievements. It greatly facilitates the use of non-professional users and people no longer need to rote a lot of orders since it has been replaced by the window, menu, buttons, etc. to facilitate to operate.

Neural Networks

Statistical parametric speech synthesis combines vocoder and acoustic models to render a speech waveform given a text. Although SPSS offers various advantages over concatenative speech synthesis, such as flexibility to change its voice characteristics and small footprint, the naturalness of synthesized speech from SPSS is still not as good as the best samples from concatenative one.

The accuracy of acoustic models is one of the factors that degrade the naturalness. This project reviewed the progress of acoustic models in SPSS from the acoustic trajectory and context modeling point of views. Although a number of different types of acoustic models have been applied to SPSS, the HMM has been the most popular one for the last two decades. However, recently proposed artificial neural networkbased acoustic models look promising and have started replacing HMMs in SPSS.

One major reason why the HMM has been a dominant acoustic model in SPSS is the existence of open-source software to build endto-end systems. As there are a number of open-source

software for deep learning, we expects that artificial neural networks will be the next dominant acoustic model in the very near future

III The Concept

Step 1. Extracting Features

Although deep learning eliminates the need for hand-engineered features, we have to choose a representation model for our data. Instead of directly using the sound file as an amplitude vs time signal we use a log-scaled mel-spectrogram with 128 components (bands) covering the audible frequency range (0-22050 Hz), using a window size of 23 ms (1024 samples at 44.1 kHz) and a hop size of the same duration.

This conversion takes into account the fact that human ear hears sound on log-scale, and closely scaled frequency are not well distinguished by the human Cochlea. The effect becomes stronger as frequency increases. Hence we only take into account power in different frequency bands. This sample code gives an insight into converting audio files into spectrogram images. We use glob and librosa library - this code is a standard one for conversion into spectrogram and you're free to make modifications to suit the needs.

Step 2. Choosing an Architecture

We use a convolutional Neural Network, to classify the spectrogram images.This is because CNNs work better in detecting local feature patterns (edges etc) in different parts of the image and are also good at capturing hierarchical features which become subsequently complex with every layer as illustrated in the image

Step 3. Transfer Learning

As the CNNs learn features hierarchically, we can observe that the initial few layers learn basic features like various edges which are common to many different types of images. Transfer learning is the concept of training the model on a dataset with large amounts of similar data and then modifying the network to perform well on the target task where we do not have a lot of data. This is also called ***fine-tuning*** - explains transfer learning very well.

Step 4. Data Augmentation

While dealing with small datasets, learning complex representations of the data is very prone to overfitting as the model just memorises the dataset and fails to generalize. One way to beat this is to augment the audio files into producing many files each with a slight variation.

We proposed a five-layer stacked CNN network for sound event recognition based on a special convolutional filter configuration with decreasing filter sizes and static and delta log-mel input features. The test results from three datasets, ESC-10, ESC-50, and Urbansound8k , indicated that the recognition performance of our model is higher than those of previous logmel-CNN models including Picazk Salamon and Bello and EnvNet We designed an end-to-end stacked CNN model for sound event recognition from raw waveforms without feature engineering. It has a special two-layer feature extraction convolution layer and convolutional filter configuration to directly learn features from raw waveforms.

Our models achieve a 2% and 16% improvement in recognition accuracy on the datasets of ESC-50 and Urbansound8k respectively, compared to the existing top end-to-end model EnvNet and the 18-layer convolutional neural network

We developed a novel ensemble environmental event sound recognition model, DS-CNN, by fusing logmel-CNN and end-to-end raw-CNN models using DS evidence theory to exploit raw waveform features as well as the log-mel features.

The goal of this project was to evaluate whether convolutional neural networks can be successfully applied to environmental sound classification tasks, especially considering the limited nature of datasets available in this field. It seems that they are indeed a viable solution to this problem. Conducted experiments show that a convolutional model outperforms common approaches based on manuallyengineered features and achieves a similar level as other feature learning methods.

Although, taking into consideration much longer training times, the result is far from groundbreaking, it shows that convolutional neural networks can be effectively applied in environmental sound classification tasks even with limited datasets and simple data augmentation. What is more, it is quite likely that a considerable increase in the size of the available dataset would vastly improve the performance of trained models, as the gap to human accuracy is still profound.

System framework and implementation

1) Voice sign acquisition Use the computer's sound card tool to acquire a voice signal and save it within the computer.

2) Processing of voice alerts

The processing of voice signals specially consists of signal extraction, signal adjustment, sign transformation and filtering. I.Time Domain Analysis of Speech Signals The voice sign is a non-stationary time-various sign, which contains a selection of data. In speech processing, speech synthesis, speech popularity, voice enhancement and other voice processing want to extract the voice which includes a variety of records with none exception. The reason of voice signal analysis is to extract the data carried by using the voice sign with the benefit and effective extraction. The speech signal evaluation can be divided into time domain and transform area. The time area evaluation is the handiest method. The time domain waveform of the speech sign is analyzed at once. The function parameters of the speech are mainly brief-time period energy, average 0-crossing price, shorttime autocorrelation characteristic and so forth. Extraction: thru the graphical consumer interface on the menu function keys to gather a computer equipment at the audio sign to finish the audio signal frequency, amplitude , different information extraction and get the voice signal waveform. Adjustment: the person interface inside the layout of the enter audio signal under a ramification of adjustments such as changes in amplitude and the frequency of operation to be able to reap the adjustment of the voice sign.

II. Frequency domain analysis of speech alerts

The Fourier representation of the signal plays an crucial role in the analysis and processing of the signal. Because of the linear system, it is simple to determine its reaction to sine or complex exponential sum. Therefore, the Fourier evaluation method can remedy a number of the sign evaluation and processing troubles. In addition, Fourier indicates that some of the characteristics of the signal come to be extra apparent so it can be greater in-depth description of the signal of the red physical phenomenon. Since the speech sign modifications over time, it's far commonly believed that speech is the output of a linear gadget this is excited via a quasi-periodic pulse or random noise source. The output spectrum is the made from the channel system frequency

reaction and the excitation source spectrum. The frequency reaction and the excitation source of the channel system all alternate with time. Therefore, the general popular Fourier illustration is not appropriate for speech sign although it is appropriate for the illustration of periodic and stationary random indicators. Since the speech sign can be taken into consideration in a short time, the approximation does not exchange so we are able to use short-time period analysis. Transformation: within the person graphical interface below the acquisition of the voice sign Fourier and different modifications then draw the before and after the transformation of the spectrum and the conversion of the cepstrum. Filter: filter out the noise part of the voice signal, can be used low-bypass filter, high-bypass clear out, band-pass clear out and band resistance filter then examine the outcomes of numerous filters. 3) The effect of voice signal show Play the voice of the processed signal and pay attention to the processing impact via the person's graphical interface output feature

IV Design and Implementation of Neural Networks

A neural network is a network or circuit of neurons, or in a modern sense, an artificial neural network, composed of artificial neurons or nodes. Thus a neural network is either a biological neural network, made up of real biological neurons, or an artificial neural network, for solving artificial intelligence (AI) problems. The connections of the biological neuron are modeled as weights. A positive weight reflects an excitatory connection, while negative values mean inhibitory connections. All inputs are modified by a weight and summed. This activity is referred as a linear combination. Finally, an activation function controls the amplitude of the output. For example, an acceptable range of output is usually between 0 and 1, or it could be −1 and 1.

Unlike von Neumann model computations, artificial neural networks do not separate memory and processing and operate via the flow of signals through the net connections, somewhat akin to biological networks.

These artificial networks may be used for predictive modeling, adaptive control and applications where they can be trained via a dataset. Self-learning resulting from experience can occurwithin networks, which can derive conclusions from a complex and seemingly unrelated set of information

One of the main things with training deep neural architectures in a supervised manner is the amount of computational effort and labeled data required for efficient learning. While the former is in some part addressed on a universal basis by hardware advances and general-purpose GPU computing, the latter is very domain-dependent.

Unfortunately, publicly available datasets of environmental recordings are still very limited - both in number and in size1. This is quite understandable, considering the high cost of manual annotation. Although the situation gradually improves with the introduction of new collections of recordings , it is still one of the major hindrances to the development of new data-intensive approaches in this field.

This is especially important, since the performance of supervised deep models is strongly influenced by the size of the dataset available for learning

VI CONCLUSION

In this paper, the layout of the speech signal processing gadget is introduced in detail. A series of photograph evaluation and processing techniques are used to realise the primary processing characteristic of the speech signal. After the test run, the design of the speech signal is completed. It is higher to complete the spectrum analysis of the speech signal thru the fft rework to achieve the spectrum of the voice signal. In the piece of the clear out, the principle task is to start from the digital filter out to layout the clear out. This essentially achieves the filter out and finished a ramification of filter out effect comparison and really constant with the requirements of the venture.

In this paper, the design of the voice sign processing gadget is added in element. A collection of photograph evaluation and processing techniques are used to realise the simple processing characteristic of the speech signal. After the take a look at run, the basic motive is done.

The layout has the following advantages:

1) The interface is concise. In the operation of the voice sign processing interface, the menu button clear and every feature corresponds to most effective one button which gets rid of the want for largescale operation of complicated software steps.

2) Processing quicker. Since there are more than one steps inside the whole operation and each step is not very near so the velocity of each step is very rapid.

Three) Occupies much less reminiscence space. The entire process simplest takes up to dozens of KB of bodily area which eliminates the need for software set up.

The layout also has many locations to improve, particularly within the following regions:

1) This method can best be achieved under the general application of the voice sign processing, the feature is incredibly easy, can't carry out complex voice sign processing.

2) As the gadget is positioned in the widespread situations of the voice signal processing, the calculation accuracy is enormously low, cannot perform extra correct voice signal processing.

3) The software is crafted from simple preparation, the operation interface is small, can not technique the bigger physical memory voice sign file.

REFERENCES

[1]. S.-R. Kuang and J.-P. Wang, “Design of power-efficient configurable booth multiplier,”IEEE Trans. Circuits Syst. I, Reg. Papers, vol. 57, no. 3, pp. 568–580, Mar. 2010.

[2]. W. Ling and Y. Savaria, “Variable-precision multiplier for equalizer with adaptive modulation,” in Proc. 47th Midwest Symp. Circuits Syst., vol. 1. Jul. 2004, pp. I 553–I556.

[3]. Xilinx Inc.: XAPP 290: Two flows for Partial Reconfiguration: Module Based or Difference Based.www.xilinx.com, Sept. (2004).

[4]. O. A. Pfander, R. Hacker, and H.-J. Pfleiderer, “A multiplexer-based concept for reconfigurable multiplier arrays,”in Proc. Int. Conf. Field Program. Logic Appl., vol. 3203. Sep. 2004, pp. 938–942.

[5]. H. Lee, “A power-aware scalable pipelined booth multiplier,” in Proc. IEEE Int. SOC Conf., Sep. 2004, pp. 123– 126.

[6]. Mesquita, D., Moraes, F., Palma, J., Moller, L., Calazanas, N.: Remote and Partial Reconfiguration of FPGAs: tools and trends. International Parallel and Distributed Processing Symposium,(2003).

[7]. Meyer-Baese, U.: Digital Signal Processing with Field Programmable Gate Arrays. Springer, (2001).

[8]. Xilinx Inc.: Development System Reference Guide. www.xilinx.com.

[9]. A. Bermak, D. Martinez, and J.-L. Noullet, “High density 16/8/4-bit configurable multiplier,” Proc. Inst. Electr. Eng. Circuits Devices Syst.,vol. 144, no. 5, pp. 272–276, Oct. 1997.

[10]. M. Hatamian and G. L. Cash, "A 70 MHz 8 bit x 8 bit parallel pipelined multiplier in 2.5 μm CMOS,"IEEE Journal of Solid-State Circuits, vol. 21, no. 4, pp. 505–513, 1986.

[11]. Yeong-Jae Oh, Hanho Lee, Chong-Ho Lee, "A Reconfigurable FIR Filter Design Using Dynamic Partial Reconfguration", IEEE,vol-06, pp. 4851–4854, ISCAS 2006.

[12]. S.Karthick, Dr. s. Valarmathy and E.Prabhu ," RECONFIGURABLE FIR FILTER WITH RADIX-4 ARRAY MULTIPLIER" , jatit, Vol. 57 No.3, pp.326-336 , Nov.2013.

[13]. K.Anandan and N.S.Yogaananth, "VLSI Implementation of Reconfigurable Low Power Fir Filter Architecture" , IJIRCCE, Vol.2, Special Issue 1, pp no 3514-. 3520 ,March 2014.

[14]. Martin Kumm, Konrad M¨oller and Peter Zipf "Dynamically Reconfigurable FIR Filter Architectures with Fast Reconfiguration", ieee journal of solid-state circuits, vol. 41, no. 4, april 2006.

[15]. Pramod Kumar Meher, Shrutisagar Chandrasekaran,and Abbes Amira , "FPGA Realization of FIR Filters by Efficient and Flexible Systolization Using Distributed Arithmetic", ieee transactions on signal processing ,pp no-1-9.

[16]. Xiaoxiao Zhang, Farid Boussaid and Amine Bermak, "32 Bit×32 Bit Multiprecision Razor-Based Dynamic Voltage Scaling Multiplier With Operands Scheduler", ieee transactions on very large scale integration (vlsi) systems, vol. 22, no. 4, april 2014.

[17]. K.Gunasekaran and M.Manikandan," High Speed Reconfigurable FIR Filter using Russian Peasant Multiplier with Sklansky Adder", Research Journal of Applied Sciences, Engineering and Technology 8(24): 2451-2456, 2014.

[18]. Shidhartha Das, David Roberts, Seokwoo Lee, Sanjay Pant, David Blaau, Todd Austin, Krisztián Flautner and Trevor Mudge , "Self-Tuning DVS Processor Using Delay-Error Detection and Correction", ieee journal of solid-state circuits, vol. 41, no. 4, april 2006.

[19]. J Britto Pari ,et al., "Reconfigurable Architecture Of RNS Based High Speed FIR Filter" , Indian Journal Of Engineering And Material Sciences,pp . 230-240, vol.21, april 2014.