IJMLNCE JOURNAL

International Journal of Machine Learning and Networked Collaborative Engineering

Journal Homepage: http://www.mlnce.net/home/index.html

DOI : https://doi.org/10.30991/IJMLNCE.2020v04i04.004

Vietnamese Voice Classification based on Deep Learning Approach

[a]Bui Thanh Hung

[a]Faculty of Information Technology,Ton Duc Thang University, 19 Nguyen Huu Tho Street, Tan Phong Ward, District 7, Ho Chi Minh City, Vietnam, buithanhhung@tdtu.edu.vn

Abstract

In the digital era, it is undeniable that voice classification plays a meaningful task in various aspects of life. In this research, we propose a method of predicting the gender and region of the Vietnamese voice which is based on the spectrum of sound using the deep learning approach. From the raw dataset, we conducted the preprocessing stage to take the audio dataset to the same frequency and time standard. After that, we extracted Mel Spectrogram feature and then put into a deep learning model - Convolutional Neural Network to train and optimize. Our experiments on 37 samples taken from VIVOS corpus audio dataset achieve the accuracy of 86.48% for predicting gender and 51.45% for predicting the region of the voice.

Keywords

Voice classification,

Mel Spectrogram feature,

Deep Learning, Convolutional Neural Network

1. Introduction

Spoken communication is the most typical mean of communication in human lives. Voice carries a lot of information regarding the person who is speaking. To receive the voice from a person, there are certain features that exist in the signal of the voice. Because of its valuable information, there are many applications using voice recognition such as: chat, detecting person, business, security, etc.

To understand the characteristics of speech in media, voice recognition is an especially significant task. However, recognizing voice is a very challenging problem because of various ways expression of people and difficult to distinguish its unclear features. Many researchers focus on solving this problem, and the most effective approach is using deep learning. For the Vietnamese voice recognition, there are a few research, this task is still challenge because of its characteristics.

In this paper, we propose a deep learning method to classify Vietnamese voice. We extracted Mel Spectrogram feature and used Convolutional Neural Network to recognize the voice. We did experiment on VIVOS dataset and evaluated by accuracy. The organization of this paper is as follows: Section 2 introduces related work. Then Section 3 describes in detail our proposed method. Section 4 shows the experiments. Finally, summarizing our work and future directions are discussed in Section 5.

2. Related Work

Training low-level extracted features is conventional technique for solving voice recognition problem. However, how to extract good features is difficult, and optimization is even harder. As a result, focusing on the use of powerful strategies for semantic analysis and relying on model selection to optimize the results are the traditional trend in audio retrieval.

There are many different techniques in sound feature extraction presented in the various documents used in sound recognition and detection. Each technique has its own advantages and disadvantages depending on the acoustic environment [1].

A couple of the featured extraction techniques include:

- Mel-frequency Cepstral Coefficients (MFCC)
- Linear Prediction Coefficient (LPC)
- Perceptual Linear Predictive (PLP)
- Relative Spectral Processing (RASTA)
- Linear Prediction Cepstral Coefficient (LPCC)

The traditional classifications of Gaussian Mixture Models (GMM), Support Vector Machine (SVM) and Hidden Markov Models (HMM) were used a lot in the past in sound recognition. But these approaches often do not handle well when the audio data is recorded through many different devices and environments, or not under the same recording conditions. In contrast, Deep Learning approaches often give higher results and accuracy. Restricted Boltzmann Machine (RBM), Convolutional Neural Network (CNN) and Long-Short-Term Memory (LSTM) are among the deep learning approaches for voice recognition [2-5].

Tapas et al used CNN to identify Closed-Set Device-Independent Speaker [6]. Nidhi Srivastava et al proposed MFCC and Neural Networks for speech recognition [7]. Using phase encoded Mel filterbank energies and CNN method is used by Rishabh N. Tak et al. [8]. Ossama Abdel-Hamid [9] proposed Hybrid Deep Neural Network-Hidden Markov model (HMM) for speech recognition.

Based on the previous researches, we propose Convolutional Neural Network for Vietnamese voice classification.

3. Methodology

3.1 The proposed model

Figure l describes our proposed model. There are four main parts in this model: collecting data, preprocessing, extracting feature and training. From the self-collected raw audio dataset, the subject performed pre-processing methods such as converting audio formats into WAV, reducing sample size from 44100Hz to 16000Hz. Each sound file has a size of 3s each and is divided into 6 sets of samples corresponding to 2 genders: Male and Female and 3 regions of Vietnam: North, Central and South. Then we pass to the deep learning model- CNN to extract the features and identify. We will describe more detail how to preprocess and extract feature in next part.

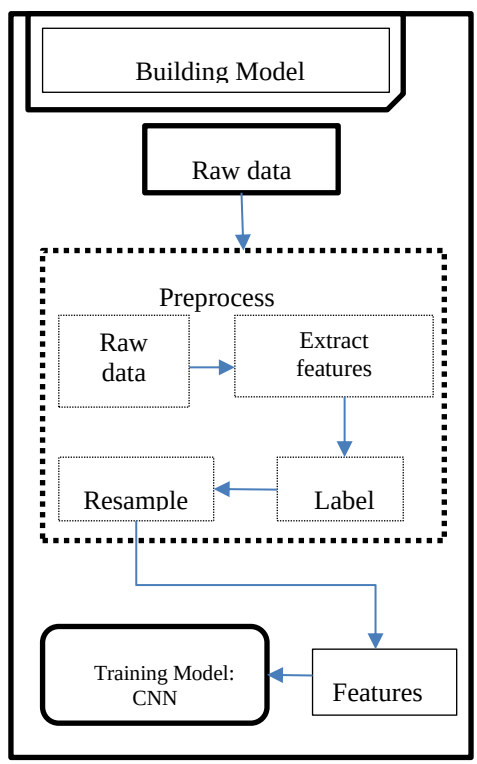

Figure 1:The proposed model

3.2 Mel Feature Extraction

In order to characterize the sound, the researchers typically will use MFCC (Mel Frequency Cepstral Coefficients). However, for frequency purpose, if using the MFCC feature extraction, it often results in serious problems. The reason is that transforming the discrete cosine which converts the spectral energy into another form might not be able to maintain the locality or inherent characteristics) of the frequency [9, 10]. In our proposed model, we use Log-Mel Spectrogram to extract sound characteristics for the training process. The process is described in Figure 2

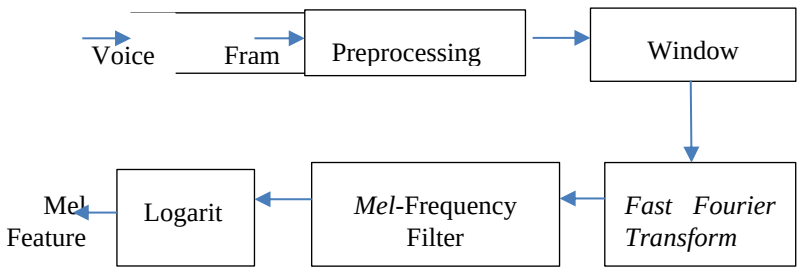

Figure 2: Procedure of Mel feature extraction

Observing the above process, we can see that the sound is divided into frames of fixed length. The purpose is to sample small (in theory stable) signal segments. The window function removes the extra effects and vectors implemented on each window frame. Fast Fourier Transforms of each frame are calculated and logarithm of spectral amplitude. Phase information is ignored since the spectral amplitude is more important than phase. It requires to carry out the logarithm of the spectral amplitude since the volume of the signal is approximately logarithmic. Next, we made the change in popularity according to Mel scale. Each section is described in details in the following sections.

In data sampling, we consider the digitized audio signal by disassembling the value over evenly spaced intervals; therefore, it is important to ensure that the sampling rate is large enough to describe the signal- waveform signal. The sampling frequency should be at least twice as much as the waveform frequency as Nyquist's theorem. Common sampling rates are 8000, 11025, 16000, 22050, 44000. Frequently, it is used the frequencies above 10Khz. Figure 3 presents example of sound by time and amplitude.

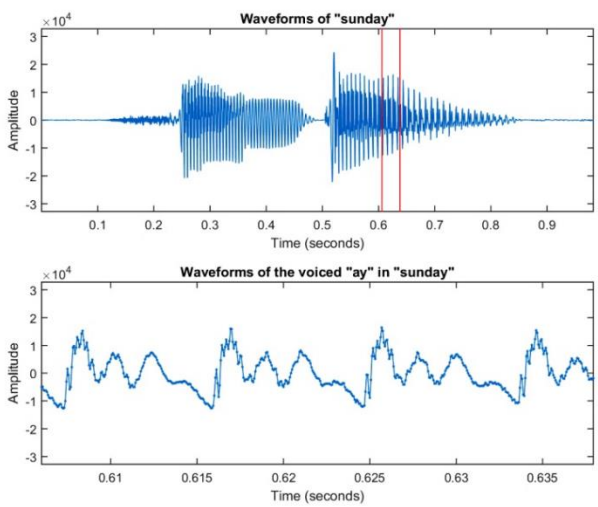

Figure 3: Example of sound sample

Frame signaling

Framing is the process of dividing a signal sample into a number of overlapping or non-overlapping frames. The purpose of framing is to sample small (theoretically stable) signal segments. The problem is that

the sound nature is unstable. So the Fourier transform will represent the frequency that occurs across the time domain instead of a specific time.

Since when the signal is unstable, it should be divided into discrete windows so that the signal in each window becomes static and the Fourier transformation is performed on each frame.

Getting the signal window

The next step is to take the window for each frame separately to reduce interruptions of the voice signal at the beginning and the end of each frame. Usually Hamming window is used, this window looks like the formula as follows:

$$w(n) = 0{,}54 - 0{,}46\cos\left(2\Pi\frac{n}{N}\right), 0 \le n < N \qquad (1)$$

In which the window length L = N + 1

Fast Fourier Transformation

Discrete Fourier Transform (DFT) or Fast Fourier transformation (FFT) is performed to convert each frame with N samples from time domain to frequency domain. The original signal should be performed Fourier transforms through a band pass filter to handle the Mel frequency deviation.

Converting to Mel scale

To accurately describe the frequency reception of the auditory system, another scale is built - the Mel scale. Converting the frequency to the Mel frequency domain helps to smoothen the spectrum and give rise to meaningful sensed frequencies. Fourier transforms the signal through a bandpass filter to simplify spectrum without losing data.

This is done by aggregating spectral components into a frequency band. Spectrum is simplified by using an array of filters to separate the spectrum into channels. The filters are evenly spaced on the Mel scale and taken logarithmic on the frequency scale; the low frequency channels are linear while the high frequency channels are logarithmic.

- The transition to the Mel frequency scale is performed in three steps:
- Fix the value area under each filter and sometimes set the scale to 1.
- Set M = number of required filter bands.
- Evenly distributed on the Mel frequency scale
- Convert from Hz to Wi on a linear scale.

After preprocessing and extracting feature, we use deep learning approach to train the model, the next part will present it.

3.3 Deep Learning approach

We use CNN- a Deep Learning method for voice classification [3, 6, 9, 11-14]. For model recognition, when using it, the input data needs to be arranged as feature maps to include CNN training. To visualize, we arrange the input as a 2-dimensional array where the horizontal and vertical values are the pixel values at the x and y coordinates. RGB color values (Red, Green, Blue) can be considered as 3 different 2 dimensional feature maps. At both training and test time, CNNs slide a small window on the input image, so the weight of the network can learn many features of the input data through this window.

It is essential to align the voice feature into feature maps to be suitable for CNN network processing and training. The input image involved in the processing of the CNN network can be considered a spectrum with the static, delta, and delta-delta characteristics playing the role of red, green, and blue. Following this way, we need to ensure that the input images maintain their inherent characteristics or "locality" on the two

frequency and time axes. Like previous speech recognition studies, each single window as input to CNN will include a certain amount of context (context 9-15 frames).

Typically to characterize the sound, researchers will use MFCC (Mel Frequency Cepstral Coefficients). However, for frequency purpose, if using the MFCC feature extraction, it often causes serious problems. The reason is transforming the discrete cosine that converts the spectral energy into another form may not be able to maintain the locality or inherent characteristics of the frequency [9].

The subject used the Log-Mel Spectrogram to describe the distribution of acoustic energy in each different frequency and to represent each voice frame.

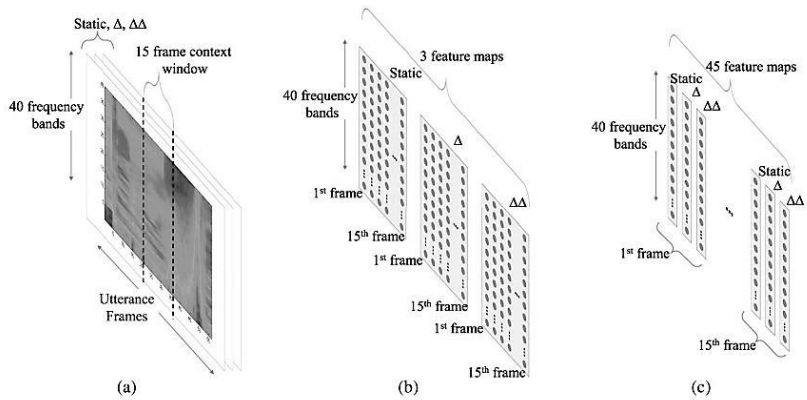

Figure 4: Two ways of the voice feature are input of a CNN network

There are a number of different ways to organize Log-Mel Spectrogram features into the features map. In Figure 4 (b) the voice features can be arranged into 3 2-dimensional feature maps, in each feature map representing Log-Mel Spectrogram (static, delta and delta-delta) features are classified together with both frequency (using frequency band index) and time (using the number of frames in each context window). So the 2-dimensional convolution is performed to normalize the changes in time and frequency simultaneously.

In addition, the frequency variation can also be normalized. So the Log-Mel Spectrogram features are arranged as one-dimensional feature maps as shown in Figure 4 (c). For example, we will create 45 1-dimensional feature maps, with each map having a size of 40 if the context window contains 40 filter bands and 15 frames. Therefore, one-dimensional convolution will be applied along the frequency axis.

When the input feature maps are formed, the convolutional layer and the pooling layer use independent operations to generate activating units in sequence on those layers. Similar to the input classes, units of pooling and convolution can also be arranged in the map. Figure 5 shows CNN model for voice classification.

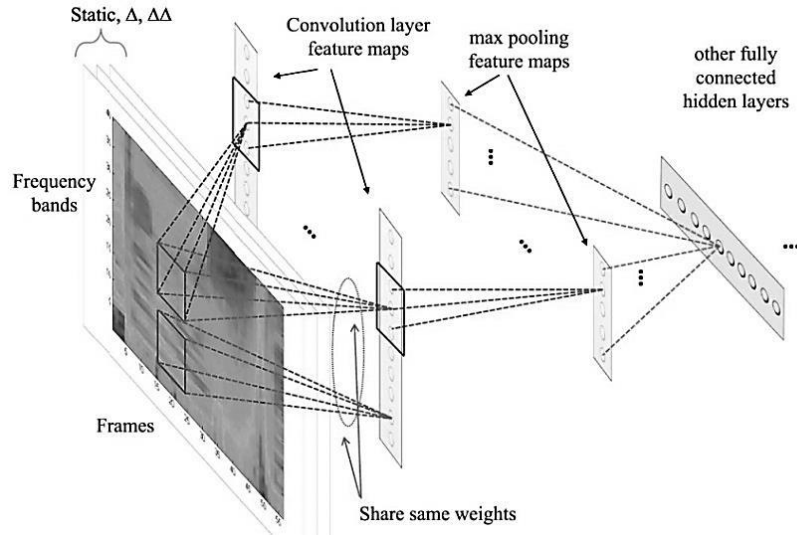

Figure 5: CNN model for voice classification

4. Experiments

4.1 Dataset

We did experiments with the collected data form Zalo AI contest which was divided into 6 sets of samples corresponding to 2 genders and 3 regions followed by converting the audio to the extension wav format. The total training data for the system is 270 samples. The dataset is described in details in the Table 1.

Table 1: The dataset

Area of Voice in Vietnam	Number of samples
North Male	42
North Female	45
Central Male	75
Central Female	44
South Male	27
South Female	37
Total	270

The test data contains 37 samples collected from Vivos Corpus data set of the Computer Science Lab [15], University of Sciences formatted with the wav extension. On each audio data file, we cut a file with the duration of 3s, then reduced a sound frequency to 16Khz in order to bring the training data sample to a common standard. e used the Pydub sound processing library [16] for the audio processing.

The purpose of this process is to standardize the input so that the CNN network can better learn voice features. In addition, it also ensures uniformity in training data. Next, put them in the project's data directory for the training process.

4.2. Experiment Result

We trained samples from the collected audio data. We used the Librosa sound processing library [17] which is installed in Python's library. Audio files are divided into frames as 250ms in length without duplication. Each frame is labeled corresponding to the sound file (including gender and region).

From each frame, we extracted characteristics of Log-Mel Spectrogram with 60 bands. The results are entered in the log amplitude function. The results are then characterized with the corresponding delta. After extracting the Log-Mel Spectrogram feature, we used the Numpy library to synthesize the features and labels of the sound. Keras [18] and Tensorflow library [19] supporting the Convolutional Neural Network are used with following parameters are shown in Table 2.

The model trained by the optimal function is Adam with learning rate: 0.0001, beta_1: 0.9, beta_2: 0.999. The training and testing ratio is 8:2 and 9:1, respectively.

After the training is completed, the model is saved for building the application. The predictive audio sample is divided into 250 frames each, using the featured extraction method described above, and then fed into the CNN network. The label of the file is selected by the voting majority strategy.

Table 2. Detail of model parameters

Layer	Parameter
Conv2D	64 kernels (7x7)
MaxPooling2D	Size 3x3, strides 2x2
Conv2D	128 kernels (5x5)
MaxPooling2D	Size 2x2
Conv2D	256 kernels (2x2)
MaxPooling2	Size 2x2
Flatten	
Dense	200
Dropout	0.2
Dense	Softmax

We use accuracy to evaluate the prediction result. The formula is calculated as follows:

$$accuracy\left(y,\hat{y}\right)=\frac{1}{n_{samples}}\sum_{i=0}^{n_{samples}-1}1\left(\hat{y}_i=y_i\right) \quad (2)$$

where, $\hat{y}$ is prediction result of i sample and y_i is the true label. The assessment is based on two factors: gender identity accuracy, and regional accuracy.

The results of gender and region predictions are shown in Table 3 and Figure 6. By the results, the gender detection accuracy is 32 samples/37 accurate samples and 19 samples/37 accurate samples is the accuracy of region identification.

Table 3. The result of our proposed model

Method	Accuracy	
	Gender	Region
CNN + Log-Mel Spectrogram	**86.48%**	**51.45%**

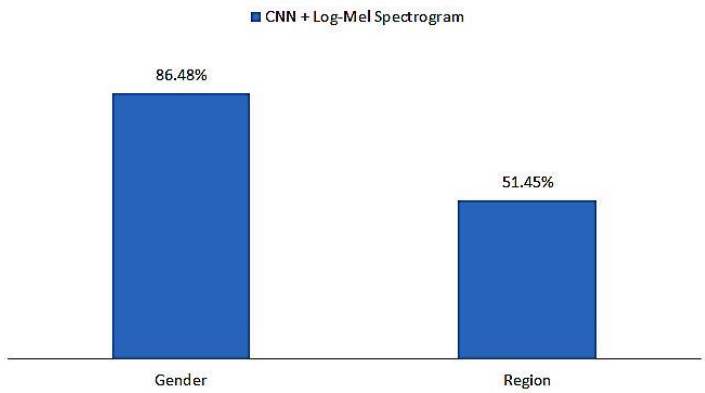

Figure 6: Our proposed model results

Following the results, we can see that the recognition system is quite good about gender, and low recognition rate of the region of the voice. Vietnamese is a tonal language with different dialects. Such variety of dialects brings up the challenges for Vietnamese automatic recognition systems.

In terms of pronunciation alone, the same word but in different regions can be pronounced differently. With two different dialects, they sound the same, but the content is understood differently by each dialect. This factor alone can cause confusion, significantly affecting the speech recognition system.

Although there has not been a widely recognized approach on how to divide Vietnamese dialect, the majority of researchers divide it into three main regions: Northern, Central and Southern. The division of dialectal regions is also relative, not completely separate. There are transitions between regions, sometimes within a locality, with a narrow geographical range such as between villages and communes. From that, we can see that the data used to train the learning convolutional neural network is not enough to cover the number of voices for each region, and in each region there will be many different small dialects for a region, this is an among the largest barriers for the system to be able to identify the most accurate domain. Another reason is the data still has quite a bit of noise in each audio file.

Thus, it can be seen that the gender identification system is quite good and in order to improve the ability to identify the region, the input data must be large and sufficient to cover a certain region, then accuracy of region identification increases.

5. Conclusion

We presented an approach to classify gender and region of Vietnamese voice by using deep learning - Convolutional Neural Network model in this research. We extracted Mel Spectrogram feature and use CNN to recognize the gender and region voice. We conduct experiment on Zalo AI contest and VIVOS dataset and evaluate by accuracy. By the experiment results, our proposed method got the best scores since it could recognize gender and region voice. We are looking to work more on the learning in depth to improve the accuracy in the voice classification.

References

[1]. Isra Khan, Rafi Ullah, Shah Muhammad Emaduddin. Robust Feature Extraction Techniques in Speech Recognition: A Comparative Analysis. Conference: International Conference on Computing & Information Sciences. 2019

[2]. M. Alam, M. D. Samad, L. Vidyaratne, A. Glandon, and K. M. Iftekharuddin. Survey on Deep Neural Networks in Speech And Vision Systems. 2019, arXiv:1908.07656.

[3]. D.Nagajyothi, P. Siddaiah. Speech Recognition Using Convolutional Neural Networks. International Journal of Engineering & Technology 7(4):133-137, 2018.

[4]. Khalid Hussain, Mazhar Hussain and Muhammad Gufran Khan. Improved Acoustic Scene Classification with DNN and CNN. 2017

[5]. Michele Valenti, Dario Tonelli, Fabio Vesperini, Emanuele Principi, Stefano Squartini. A Neural Network Approach for Sound Event Detection in Real Life Audio. 25th European Signal Processing Conference (EUSIPCO) 2017.

[6]. Tapas Chakraborty, Bidhan Barai, Bikshan Chatterjee, Nibaran Das, Subhadip Basu and Mita Nasipuri. Closed-Set Device-Independent Speaker Identification Using CNN. International Conference on Intelligent Computing and Communication, 2019.

[7]. Nidhi Srivastava. Speech Recognition using MFCC and Neural Networks. International Journal of Engineering Development and Research. Vol. 2, pp. 2122-2129, 2013

[8]. Rishabh N. Tak, Dharmesh M. Agrawal, and Hemant A. Patil. Novel Phase Encoded Mel Filterbank Energies for Environmental Sound Classification. International Conference on Pattern Recognition and Machine Intelligence 2017

[9]. Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu. Convolutional Neural Networks for Speech Recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing Volume: 22, Issue: 10, Oct. 2014

[10]. Jha, S., Kumar, R., Chiclana, F., Puri, V., & Priyadarshini, I.: Neutrosophic Approach for Enhancing Quality of Signals. Multimedia Tools and Applications, 1-32, 2019.

[11]. Bui Thanh Hung, Vijay Bhaskar Semwal, Neha Gaud, Vishwanth Bijalwa. Violent Video Detection by Pre-trained Model and CNN-LSTM Approach". Proceedings of Integrated Intelligence Enable Networks and Computing. Springer Series in Algorithms for Intelligent Systems, 2021.

[12]. Bui Thanh Hung, Le Minh Tien. Facial Expression Recognition with CNN-LSTM. Research in Intelligent and Computing in Engineering. Springer Series in Advances in Intelligent Systems and Computing. 2020.

[13]. Bui Thanh Hung. Face Recognition Using Hybrid HOG-CNN Approach". Research in Intelligent and Computing in Engineering. Springer Series in Advances in Intelligent Systems and Computing, 2020.

[14]. Bui Thanh Hung: Domain-Specific Versus General-Purpose Word Representations in Sentiment Analysis for Deep Learning Models. Frontiers in Intelligent Computing: Theory and Applications pp 252-264, Springer, 2019.

[15]. Hieu-Thi Luong and Hai-Quan Vu: A Non-Expert Kaldi Recipe for Vietnamese Speech Recognition System. In Proc. WLSI-3 & OIAF4HLT-2, 2016.

[16]. Pydub: https://github.com/jiaaro/pydub

[17]. McFee, Brian, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. Librosa: Audio and Music Signal Analysis in Python. In Proceedings of the 14th Python in Science Conference, pp. 18-25. 2015.

[18]. François Chollet et al.: Keras, 2015. https://keras.io/

[19]. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., and Zheng, X.: Tensorflow: A system for large-scale machine learning. Tech. rep. Google Brain arXiv preprint, 2016.

Author's Biography

Bui Thanh Hung received his M.S. degree and Ph.D. degree from Japan Advanced Institute of Science and Technology (JAIST) in 2010 and in 2013.

He is currently a lecturer of Faculty of Information Technology, Ton Duc Thang University.

He has completed 2 projects, published 9 journals, 11 book chapters, 32 International conference papers and 15 domestic conference papers.

He had three best paper awards of FICTA 2018, RICE 2020 and ICAIAA 2021 and is selected as Excellent Scientific Researcher of Thu Dau Mot University in 2019, 2020, 2021. He is a reviewer of many reputed journals and conferences.

His main research interests are Natural Language Processing, Machine Learning, Machine Translation, Text Processing, Data Analytics, Computer Vision and Artificial Intelligence.