Speech recognition for people with dysphasia using convolutional neural network

Bo-Yu Lin
Dept. of Computer Science and Information Engineering
National Taipei University
New Taipei, Taiwan, R.O.C
h5611239@gmail.com

Hung-Shing Huang
Dept. of Computer Science and Information Engineering
National Taipei University
New Taipei, Taiwan, R.O.C
msn9110@gmail.com

Ruey-Kai Sheu
Dept. of Computer Science
Tunghai University
Taichung City, Taiwan, R.O.C.
rickysheu@thu.edu.tw

Yue-Shan Chang
Dept. of Computer Science and Information Engineering
National Taipei University
New Taipei, Taiwan, R.O.C
ysc@mail.ntpu.edu.tw

***Abstract*—As the advance of technology, it is increasingly speech recognition tools on mobile devices, such as Google Voice and Apple Siri, those have been widely used and have high recognition rate of people's speech. However, these speech recognition tools cannot work well for people with the disease of "Dysphasia" and has very low recognition rate. It is important issue to develop a speech recognition toot for dysphasia, such as Cerebral Palsy(CP) and Amyotrophic Lateral Sclerosis(ALS), to assist those people communicating with others well. Recently, there are various open source programs has been announced, such as Google's Tensorflow, which is used to develop speech recognition based on Deep Neural Networks (DNNs). In this paper, we propose a Convolutional Neural Networks (CNNs) model to perform speech recognition for ALS. The model consists of two hidden CNN layers, each includes one CNN, Rectified Linear Unit (ReLU), Dropout Unit, and MaxPooling Layer. We implement the CNN for dysphasia speech recognition using Google's Tensorflow and collect 33 pronunciations from a dysphasia, each pronunciation at least has 350 training voice files. The highest accuracy is about 63% for one word. And it will be up to totally 94% for top five words. The result shows that it can effectively recognize speech for dysphasia. It can be expected to construct a speech recognition system for assisting dysphasia communicating with others.**

Keywords—Speech Recognition, Dysphasia, Convolutional Neural Networks, Deep Learning,

I. INTRODUCTION

People can be easy to communicate with others because there are thousands of pronunciations as well as tones in their language. However, for people with the disease of "Dysphasia", such as "Cerebral palsy(CP)[1]" or "Amyotrophic lateral sclerosis(ALS)[2]", they cannot normally talk to people because their voices are vague and pronunciations are not clear, that causes people difficult to understand what they are speaking. The disease of ALS is one of motor neuron disease, which refers to the general muscle atrophy caused by progressive degeneration of motor neurons. Therefore, some of the people with the disease of ALS, their typewriting and handwriting ability due to injured organs is also gradually decreased. They do not get pleasure to communicate with their friends and family.

Although there are increasingly speech recognition tools on mobile devices that have been widely used and have high recognition rate of people's speech, such as Google Voice[3] and Apple Siri [4]. These tools cannot work well for the people with the disease of "Dysphasia" with very low recognition rate. We let people with ALS symptom use these tools and try to say "The weather is good today." These tools' translation is completely wrong. Therefore, how to propose an effective method to help improve the speech recognition rate of people with ALS symptoms will be an important issue.

Deep Neural Networks (DNN) [5], especially Convolution Neural Networks (CNN) [6], has been widely used as a method of speech recognition and it is witnessed has good results. Google's TensorFlow [7], based on CNN, also provide an open source for training and testing voice files. It can work well for general voice recognition. But, it might not be applied well for the speech recognition for the people with dysphasia because their vocalized muscle damaged due to motor neuron disease. Applying CNN to recognizing speech[8,10,21] for the case may have following issues needs to be fixed. The first is the length of the pronunciation is different from normal people. Finding the proper length of speech for CNN to train a better model is an important issue. The second is the similarity of many phonetic pronunciations is higher than normal people; so that cannot obtain higher accuracy of recognition. For example, the pronunciation of "chih" is often recognized as "yi" or "jian". The third is that they cannot speak too long sentence because of muscle atrophy, they can only speak one or two words at a time. The finally is that how to build a tool for them using CNN with keyword spotting technique[9] to assist they accurately recognizing voice is important.

As mentioned issues above, in this paper, we will design and implement a tool to overcome the issues with following methods. We will find the proper length of speech through conducting some experiments. We use the pronunciations of the top five rates as our output that can be 50% more accurate than taking a single pronunciation. Using 33 pronunciations to train and its accuracy can be as high as 94%. In addition, we build a single pronunciation recognition model trained by CNN with keyword spotting technique, convert pronunciation to Chinese word, and use a series of words to make a sentence. Finally, we build speech database and design a complete speech recognition system for dysphasia.

The rest of the paper is organized as follows. Section II describes the background of this study, such as speech preprocessing, deep learning methods, convolutional neural networks, and related work. Section III presents the MFCC

2577-1655/18/$31.00 ©2018 IEEE
DOI 10.1109/SMC.2018.00372

architecture, the CNN architecture, and overall architecture. Section IV presents the implementation of speech recognition for Dysphasia. Section V is the experimental results of the proposed approach. Finally, we give a concluding remarks and future work in Section VI.

II. Background and Related Work

A. Mel-scale Frequency Cepstral Coefficient (MFCC)

First, it is very important to find acoustic features before building acoustic models. For speech recognition, the most popular method is using mel-scale frequency cepstral coefficient (MFCC). MFCC [11] takes human perception sensitivity with respect to frequencies into consideration, so it is best way for speech recognition. The greatest feature of the MFCC is that it can evenly distribute the frequency bands on the mel-scale, which means that the method is closer to the non-linear auditory system of human ear.

First, the MFCC will use pre-emphasizing technique to suppress the upper part of lips or vocal cords. Then it uses the frame blocking to divide the sound into several segments, and hanming windowing will make the spectrum more continuous. Using the Fourier transform to convert the signal in the time domain to the energy distribution in the frequency domain, and the triangular bandpass filter converts it to mel-spectrum. Then, it takes the logarithm of the mel-spectrum, and then use discrete cosine transform(DCT) to transpose the frequency domain of the mel-spectrum into the time domain, we can get continuous vector values based on acoustic features and time domain to form a two-dimensional spectrum feature map.

In this paper, MFCC is a significant pre-processing to make speech to acoustic spectrum features. These features can be used in our training.

B. Deep Learning

Deep learning [6] is part of machine learning. In 2012, Geoffrey Hinton and his students won the first place using deep learning in the ImageNet [12] Image Recognition Competition, and therefore deep learning is becoming more popular. Machine learning is to let the computer analyze the data by itself to find out the "eigenvalue" of the data. A deep neural network(DNN) is a neural network with multiple hidden layers. DNN can also provide modeling for complex nonlinear systems, this makes the model more complicated and diverse.

There are three most well-known basic deep learning methods: Deep Neural Networks (DNN), Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), most methods are based on these three methods to make improvements. Speech recognition system can also be developed using deep learning, we can use the acoustic model to find out corresponding words or sentences according to the speech.

C. Related Work

There are few studies for people with Dysphasia, but there is a similar study in [13]. It mentioned that children's speech has greater complexity and variability than adult speech, and their data is very scarce to train. Then, getting children's speech data is a difficult task. Therefore, this paper uses Vocal Tract Length Normalization (VLTN) to deal with this problem, which has proven to be effective.

In [14], authors apply ensemble learning to different deep learning systems. Ensemble learning can combine DNNs, RNNs or CNNs to improve recognition accuracy. In various deep learning architectures, each learning architecture has its own obvious advantages and disadvantages. Therefore, this paper would like to explore ensemble learning using linear or logarithmic methods to optimize mathematical formulas.

Convolutional neural networks (CNNs) have been shown to improve automatic speech recognition (ASR). In [15], it uses the weighted prediction error (WPE) to preprocess in noisy environments, and use CNN acoustic models with 10 layers. The result shows that it reduces the word error rate(WER) by 20%. Microsoft revised its system developed in 2016 with CNN-BLSTM technology in 2017[16]. This model includes character-based and conversational LSTM language models, this system utilizes a two-phase approach and has good results.

Many studies in speech recognition have shown that the performance of CNNs is better to that of fully connected DNNs. In [17], authors propose the use of CNN in far-field speech recognition for dealing with reverberation. Its experimental results show that a CNN coupled with a fully connected DNN can fast model the correlations in feature vectors with fewer parameters than a DNN, and this method can explore unseen testing environments.

However, the above studies are all referring to general people's speech recognition, there is very little research on speech recognition and CNN acoustic model for people of dysphasia. Because CNN [21] is suitable for speech recognition, we use it to build the acoustic model, which can construct the best speech recognition for people with dysphasia.

III. CNN Architecture

A. Typical Convolutional Neural Networks

CNN is comprised of at least one convolutional layers and following at least one fully connected layers in neural networks. Its architecture has advantage to train 2D structured things, such as image or speech signal. CNN is easier to train and have fewer parameters than DNNs with the same number of hidden units. In CNN, it has two core layers: convolutional layers and pooling layer.

- The purpose of convolutional layer is to extract different features of the input using kernel. Its equation for one point of output is following :

$$y[\,m,n\,] = \sum_j \sum_i x[\,i,j]k[m-i,n-j] \qquad (1)$$

In Fig. 1, we assume the kernel size is 3 × 3. Use (1) to calculate and get output as below:

$$\begin{aligned} y[1,1] = {} & x[0,0]k[1,1] + x[1,0]h[0,1] + x[2,0]h[-1,1\,] \\ & + x[0,1]k[1,0] + x[1,0]h[0,0] + x[2,1]h[-1,0\,] \\ & + x[0,2]k[1,-1] + x[1,2]h[0,-1] + x[2,2]h[-1,-1] \end{aligned}$$

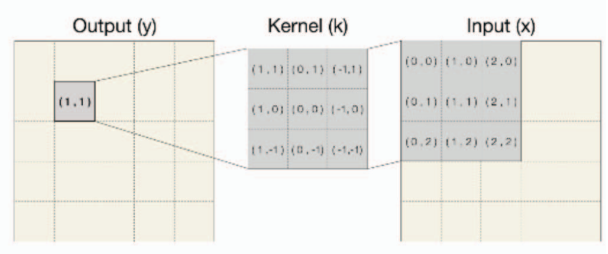

Fig. 1. Convolutional layers

- The pooling layer is a form of subsampling. It reduces the size of data, therefore the parameters decrease and the calculation can be more efficient. The most popular pooling function is "Max pooling". In Fig. 2, it show the max pooling operation:

$$y[1,1] = \text{Max}(x[2,2] , x[3,2] , x[2,3] , x[3,3])$$

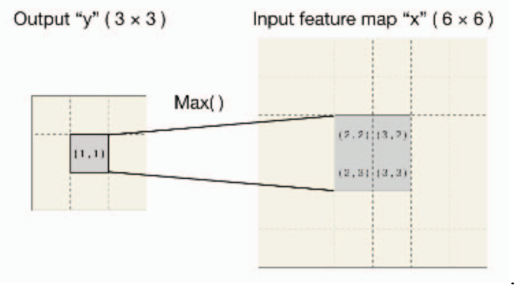

Fig. 2. Max pooling layer

B. CNN for Speech Recognition

Most studies[8,10,13,14,15,16,17,21] of speech recognition is used CNN to build model, it has fewer parameters and fast calculation. But these speech recognitions are not suitable for dysphasia. We use Tensorflow as development environment and refer this paper's [9] method to build acoustic model, it uses the keyword spotting (KWS)[18] task and CNN to achieve speech recognition. This method is more efficient to train 2D structure things and has low false reject rate .

CNN architecture, as shown in Fig. 3, consists of 2 convolutional layers with 64 kernels, 2 dropout layers with 50% probability, 2 (rectified linear unit) Relu layers, 2 max pooling layers and 1 softmax layer. The softmax output layer contains one output target for each of pronunciation in the keyword phrase to be detected, and if the pronunciation does not belong to any of pronunciations in the keywords, it would be assigned to "unknown". The weights of network uses gradient descent to optimize the cross-entropy criterion, another word is to improve overall accuracy.

We use CNN [21] instead of DNN [22] because CNN is obviously better than DNN for speech recognition. And RNN system is not good for people with dysphasia, they can't talk too long at once because of their insufficient vital capacity. Therefore, we use CNN and this paper [9] method to build the speech recognition for these people.

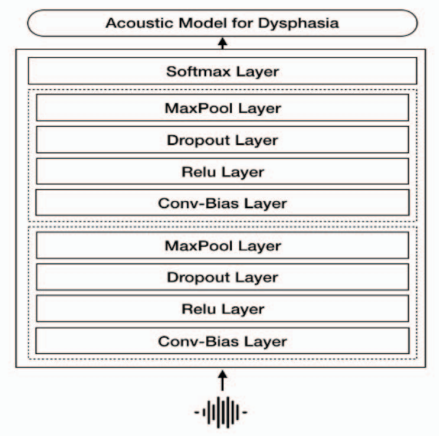

Fig. 3. CNN architecture for speech recognition

C. Parameters and data

At default, the length of the training speech data is 1.0 second for general. But the people with dysphasia have muscle atrophy or muscle disease, the length of speech would be shorter than general people. They have different speech waveform, therefore we try to use different lengths of speech, that we can compare the training result to find out best model. Another issue is training times, the number of training times mainly affects the accuracy of model and training efficiency. So, finding the right number of training times is a very critical task.

We need to do pre-processing for pronunciations. The Chinese words have five tone: high level, mid-rising, mid-falling-rising, high-falling and neutral tone, but speaking the correct tone is difficult for people with dysphasia. So, we integrate the words with same pronunciation but differen tone. For example, there are five pronunciations: Fa, Fá, Fǎ, Fà and Fạ, they are unified into high level pronunciation "Fa".

IV. Speech Recognition System for Dysphasia

A. System Architecture

The overall speech recognition system architecture uses a mobile phone as a client, which user can operate the user interface. The server is used to compute data, train model and save data, the framework is illustrated in Fig. 4.

The client (smart phone or tablet) has a recording, uploading and API communication module. At the server, there is a database for accessing speech data and Chinese words data. Use speech data to train the model through CNN, that server can produce the acoustic model for dysphasia. After acoustic model is generated, server can receive the client's request through the RESTful API [19]. Client upload the speech and command to server through RESTful API, the speech data access into database and go into acoustic model to be recognize. Acoustic model produces top five rate Chinese pronunciations to post-processing module called "pronunciations-to-words module", which can convert the pronunciation to word. Finally, server get the lists of words to response to client, and user interface show the results. After the user operates, the RESTful API can be used to update the frequency of words.

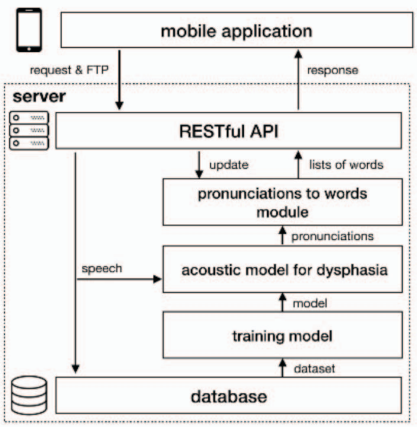

Fig. 4. Overall system architecture

B. System Flow Chart

As shown in Fig. 5, we illustrate in detail the steps of training and user operations.

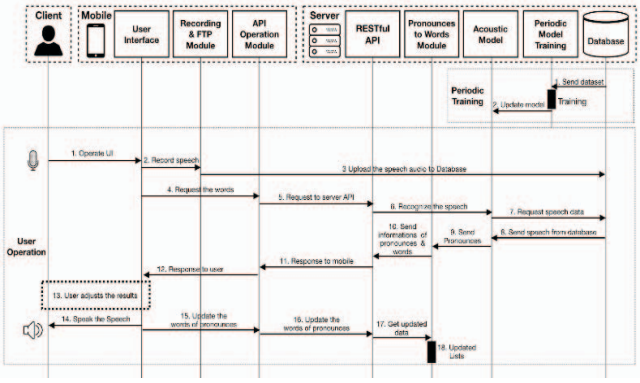

Fig. 5. System flow chart.

1) The training of acoustic models:

Step.1: The preprocessed speech data is sent from the database to the deep learning model.

Step.2: After receiving the data, set various parameters of the training module, including sampling rate, speech length and so on. After the training is beginning for a period of time, it generates the latest acoustic model.

2) User operations:

Step.1: The user operate speech recognition application through the user interface.

Step.2: Make speech recording.

Step.3: Use FTP to upload speech data to the server's database as recognized data and future training data.

Step.4: Send instructions to the API operation module and request to send speech recognition request and data.

Step.5: Instructions and data are transmitted to the server through the RESTful API.

Step.6: The server sends instructions to the acoustic model and requires preparation for speech recognition.

Step.7: The acoustic model asks database for the newly uploaded speech data.

Step.8: Send the speech data to the acoustic model for recognition.

Step.9: After speech recognition is completed, the result obtains top five pronunciations from speech, and these pronunciations send to the pronunciations-to-words module.

Step.10: Find out all words of pronunciations, and module makes the lists to transmit RESTful API.

Step.11: The RESTful API sends a response to the client so that it can receive the results.

Step.12: Receive the results, which delivers to user interface.

Step.13: The user interface shows different words according to the pronunciations.

Step.14: After the adjustment is complete, application can speak the general Chinese speech to user.

Step.15: At the same time, the selected frequency of the words of the pronunciation would be updated, therefore updated data passes to the API operation module.

Step.16: Transfer updated data to the server via RESTful API.

Step.17: Server receives it and sends the updated data to the pronunciations-to-words module immediately.

Step.18: The words can be updated the frequency in the pronunciation.

C. Pronunciations-to-words Module

This module has two main points, there are initialization task and frequency of words update task. First work is the initial word frequency, this job use web crawler [20] to get the thousand articles and count the frequency of Chinese words. Finally, we make words and words frequency of the pronunciation into json file to access.

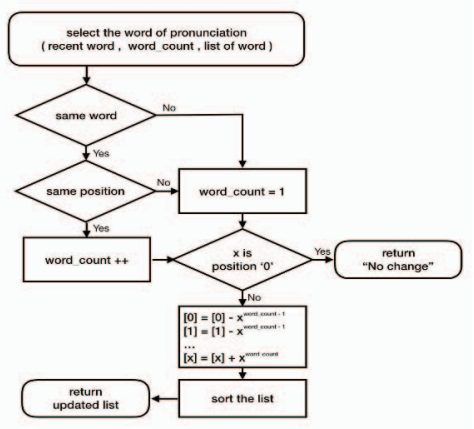

Fig. 6. Updated words frequency system

Another is frequency of words update task (Fig. 6). If the number of occurrences of the same word in the pronunciation is higher, its score is more higher in the pronunciation, and we use exponential growth because of increasing effectiveness. And the word score in front of selected word would be reduced, if it is already ranked first that its score no more increase.

Assume there is a list [‘A’: 600, ‘B’: 160, ‘C’: 140, ‘D’: 100], We select ‘C’ word 3 times and show the step:

Step.1: The word_count is 1 and positon ‘x’ is 2.

$$[0] = 599 - 2^{2-1} = 599$$
$$[1] = 159 - 2^{1-1} = 159$$
$$[2] = 140 + 2^{1} = 142$$

Step.2: The word_count is 2 and positon ‘x’ is 2.

$$[0] = 599 - 2^{2-1} = 597$$
$$[1] = 159 - 2^{2-1} = 157$$
$$[2] = 142 + 2^{2} = 146$$

Step.3: The word_count is 3 and positon ‘x’ is 2.

$$[0] = 597 - 2^{3-1} = 593$$
$$[1] = 157 - 2^{3-1} = 153$$
$$[2] = 146 + 2^{3} = 154$$

The positions would be sorted like this [‘A’: 593, ‘C’: 154, ‘B’: 153, ‘D’: 100].

V. Experimental Results and Implementation

A. Environment

This part introduces hardware and software development. The client’s application uses Zenpad z580kl which version is Andriod 7.0, and it can record speech and show the user interface. The part of server, the operating system is Ubuntu 16.04 , and this server is equipped with graphics processing unit called GeForce GTX 1080 Ti, which is mainly used for large data operations. Programming language is used Python 3 and Tensorflow 1.4.0 to develop system.

B. Data

The source of the speech signal is taken from people with dysphasia. The speech data trained in these experiments consisted of 33 pronunciations with no tone, this means we distinguish high-level, mid-rising, mid-falling-rising, high-falling and neutral tone into the same class. Because the people of dysphasia cannot speak speech clearly. And the speech data is 13,000, so there are 400 data for each pronunciation.

C. Evaluation

The number of training is 29,000 times, and 100 samples are taken at random each time for training. The first 20,000 times are 0.001 learning rate. Following 6,000 times are 0.0001 learning rate, the last 3,000 times are the 0.00001 learning rate. This setting mainly allows learning to be more convergent.

Table I shows the accuracy of training after taking different speech lengths of time. For the getting number of words, We take only a maximum 63% of the top 1 pronunciation, that accuracy is not enough to construct the system. So, we take top 3 and top 5 pronunciations to compare, that show the top 5 words having maximum 94% accuracy. Taking top 5 pronunciations can design a system that user selects pronunciation and words to speech wanted speech.

Another information in Table I is to find out the best time length of speech. Experimental results show that the accuracy of 0.8 and 0.9 seconds are better than other time lengths. Although there are two equally good results, we choose a length of 0.8 seconds as our time length of speech because of its top 1 rate is more higher than 0.9 seconds.

TABLE I. Related Speech Results

	Different Time Length of Speech				
	0.7 sec	***0.8 sec***	***0.9 sec***	***1.0 sec***	***1.25 sec***
Top 1 pronunciation	61.07%	**63.82%**	61.07%	59.85%	53.59%
Top 3 pronunciations	86.41%	**88.70%**	**88.70%**	87.18%	82.75%
Top 5 pronunciations	94.05%	94.05%	**94.20%**	93.44%	91.15%

Table II shows the relationship between the training times and the accuracy, the time length of the speech is 0.8 seconds and each time 100 samples are taken. The learning rate divided into two segments is 0.001 and 0.0001, and the learning rate divided into three segments is 0.001, 0.0001, 0.00001. We take top 5 pronunciations as our results. It can be seen that the training times does not need too many times, the 29,000 training times are the best.

TABLE II. Training Times and Accuracy

Times	***15000,3000 (18000)***	***20000,6000,3000 (29000)***
Accuracy	94.05%	**94.08%**
Times	***20000,10000,5000 (35000)***	***20000,15000,10000 (45000)***
Accuracy	93.13%	93.89%

VI. Conclusions and Future Work

This paper mainly deals with the communication issue for people with dysphasia. First, the tone problem is solved by integrating different tone in same pronunication. We refer the CNN-KWS method and adjust its model that find out 0.8 seconds which is best time length of speech, it improves the accuracy of recognition. And we use the top 5 pronunciations to increase the accuracy to 94% and design the complete system for people with dysphasia. These people can use system to recognize their speech and speak out using mobile phone or tablet.

In the future, we will continue to increase the amount of data and pronunciations, the model will be improved using different methods. In addition, we also want to combine Hidden Markov Model [24] (HMM) that build the Automatic Speech Recognition (ASR). This can be used on many ways, such as writing articles or message input. We believe this is a very valuable research.

References

[1] Colver. Allan, Fairhurst. Charles, Pharoah. Peter O D,“Cerebral palsy” The Lancet. London, vol. 383, iss. 9924, Apr 2014.

[2] Kiernan. Matthew C, Vucic. Steve, Cheah. Benjamin C, Turner. Martin R, Eisen. Andrew, “Amyotrophic lateral sclerosis” The Lancet. London, vol. 377, iss. 9769, Mar 2011.

[3] Google Inc., “Google Voice”, on web.

[4] Apple Inc., “iOS - Siri - Apple”, on web.

[5] Istituto Dalle Molle di Studi sull’Intelligenza Artificiale, “Deep learning in neural networks: An overview” Elsevier, vol. 61, pages 85-117, January 2015.

[6] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks”, Neural Information Processing Systems Foundation Inc., 2012.

[7] Google Inc., “Tensorflow”, on web.

[8] 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., vol. 22, iss 10, Oct. 2014.

[9] Tara N. Sainath, Carolina Parada, “Convolutional Neural Networks for Small-footprint Keyword Spotting”, Google Inc., INTERSPEECH, 2015.

[10] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Gerald Penn, “Applying Convolutional Neural Networks concepts to hybrid NN-HMM model for speech recognition”, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012.

[11] Parwinder Pal Singh, Pushpa Rani, “An Approach to Extract Feature using MFCC”, IOSR Journal of Engineering (IOSRJEN), vol. 04, iss 08, pages 21-25, August. 2014.

[12] ImageNet Org. “ImageNet”, on web.

[13] Mengjie Qian, Ian McLoughlin, Wu Guo, Lirong Dai, “Mismatched Training Data Enhancement for Automatic Recognition of Children’s Speech using DNN-HMM”, 10th International Symposium on Chinese Spoken Language Processing (ISCSLP), 2016.

[14] Li Deng, John C. Platt, “Ensemble Deep Learning for Speech Recognition”, Microsoft Research, One Microsoft Way, Redmond, WA, USA, 2014.

[15] Sunchan Park, Yongwon Jeong, Min Sik Kim, Hyung Soon Kim, “Linear prediction-based dereverberation with very deep convolutional neural networks for reverberant speech recognition”, International Conference on Electronics, Information, and Communication (ICEIC), 2018.

[16] W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, A. Stolcke, “The Microsoft 2017 Conversational Speech Recognition System”, Microsoft AI and Research, Technical Report MSR-TR-2017-39, August 2017.

[17] Takuya Yoshioka, Shigeki Karita, and Tomohiro Nakatani, “Far-field speech recognition using CNN-DNN-HMM with convolution in time”, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), August 2015.

[18] Guoguo Chen, Carolina Parada, Georg Heigold, “Small-footprint keyword spotting using deep neural networks”, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014.

[19] Armin Ronacher, "Flask - RESTful", in document.

[20] Christopher Olston and Marc Najork, “Web Crawling”, Foundations and Trends in Information Retrieval, vol. 4, No. 3, 2010.

[21] 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, vol. 22, no. 10, October 2014.

[22] George E. Dahl, Dong Yu, Li Deng, Alex Acero, “Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition”, IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, iss. 1, 2012.

[23] Joseph Kesheta, David Grangierb, Samy Bengioc, “Discriminative keyword spotting”, Speech Communication, vol. 51, iss. 4, pages 317-329, Apr 2009.

[24] L. R. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition”, Proceedings of the IEEE, vol. 77, Iss. 2, pages. 257 - 286, 1989.