Scientific Review
ISSN(e): 2412-2599, ISSN(p): 2413-8835
Vol. 9, Issue. 2, pp: 10-14, 2023
URL: https://arpgweb.com/journal/journal/10
DOI: https://doi.org/10.32861/sr.91.10.14

Academic Research Publishing Group

Original Research **Open Access**

Improving the Accuracy of Speech Recognition Models for Non-Native English Speakers using Bag-of-Words and Deep Neural Networks

Van-An Tran (Corresponding Author)
Institute of Information and Communication Technology, Le Quy Don Technical University, Ha Noi, Viet Nam
Email: tavistu@gmail.com

Dinh-Son Le
Institute of Information and Communication Technology, Le Quy Don Technical University, Ha Noi, Viet Nam

Ha Huy Hung
Faculty of Aerospace Engineering, Le Quy Don Technical University, Ha Noi, Viet Nam

Dinh- Quan Nguyen
Faculty of Aerospace Engineering, Le Quy Don Technical University, Ha Noi, Viet Nam

Article History
Received: 1 February, 2023
Revised: 2 April, 2023
Accepted: 3 May, 2023
Published: 13 May, 2023

Abstract

This letter presents a novel error correction module using a Bag-of-Words model and deep neural networks to improve the accuracy of cloud-based speech-to-text services on recognition tasks of non-native speakers with foreign accents. The Bag-of-Words model transforms text into input vectors for the deep neural network, which is trained using typical sentences in the curriculum for elementary schools in Vietnam and the Google Speech-to-Text data for those sentences. The trained network is then used for real-time error correction on a humanoid robot and yields 18% better accuracy than Google Speech-to-Text.

Keywords: Speech to text; Elementary school students in Vietnam; Deep neural network; Bag of words; English language; Error correction module.

1. Introduction

Speech Recognition (SR), also referred to as automatic speech recognition (ASR), computer speech recognition (CSR), and speech-to-text (STT) Waibel and Lee [1], is the conversion of spoken words into text. Speech is the predominant mode of communication among humans, and hence, machines that could understand and interact through speech have always intrigued humanity. ASR has been the subject of active research for the past several decades [2] and fits into the broader research goal of developing intelligent machines that can be instructed using natural languages.

Robots have proliferated into our homes and everyday surroundings. The era of personal computing is making way for an era of personal robots. Intelligent humanoid robots like the famous ASIMO [3] allow for a more natural way of interacting with humans that most often includes speech recognition and synthesis. Robots are increasingly becoming more social, i.e., they are helping with elderly care and are finding their way into education [4] as well. Social robots are required to interact with humans in different ways, including via speech.

Over the years, speech recognition technology has improved significantly [5] making speech a reliable mode of communication for human-machine interaction. However, there are still some areas that need further improvement. One such area is the degradation in the accuracy of speech recognition systems whose models are trained using speech data of native speakers but are required to recognize speech by speakers with a foreign accent [6, 7]. A similar problem is being considered in this study. An intelligent, humanoid robot [8] is being used for assistance in teaching English to elementary school students in Vietnam whose native language is Vietnamese and who speak English with a particularly distinct Vietnamese accent. This distinct Vietnamese accent causes deterioration in the accuracy of the standard speech-to-text models used for the recognition of spoken English. For instance, our survey of around 200 Vietnamese students, aged 7 to 10 years, each reading 500 basic sentences in English from their elementary school curriculum, has revealed that Google's Speech-to-Text was able to correctly recognize only 62% of the sentences. This is significantly low and can cause young kids to lose interest in learning English. Hence, we propose a modified data processing pipeline for a humanoid robot that features an error correction module to improve the speech recognition accuracy of cloud-based speech-to-text services such as Google Speech-to-Text. Our proposed methodology improves the speech recognition accuracy of Google Speech-to-Text and helps in making both interacting with the robot and learning English fun activities for the kids.

The contributions of this study are a new data processing pipeline is proposed for improving the speech recognition accuracy of a humanoid robot that features a novel error correction module, which uses a Bag-of-Words model and deep neural networks for the detection and correction of errors in speech transcripts of non-native speakers of English language, which are generated by commercial, cloud-based speech to text services.

2. The proposed methodology is implemented and tested in real-time on an actual humanoid robotic platform, which is used for supporting the teaching of the English language in elementary schools in Vietnam.

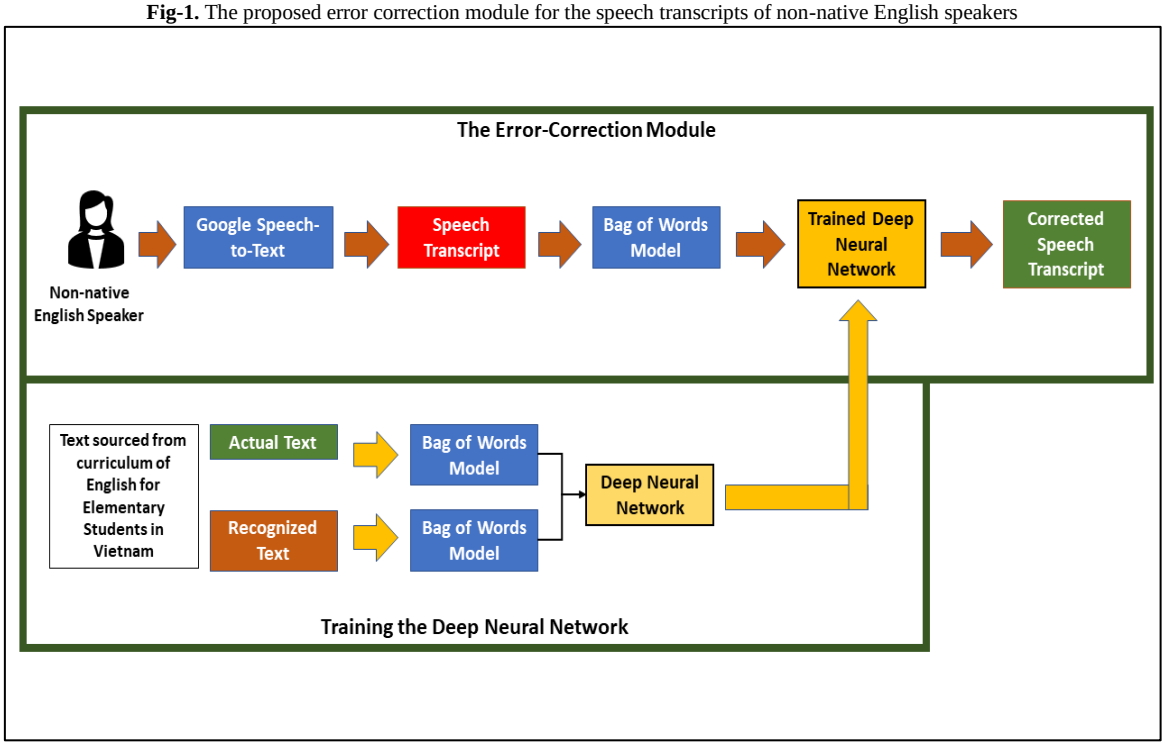
The rest of this letter is organized as follows: In Section 2 we present the proposed methodology for improving speech recognition accuracy of cloud-based Speech-to-Text services for non-native speakers of the English language with distinct foreign accents. In Section 3 we talk about the data and the experimental setup used for testing the proposed methodology, whereas in Section 4 we discuss our results. Section 5 concludes our study.

2. The Proposed Methodology

This section presents the proposed methodology for improving speech recognition accuracy of cloud-based speech-to-text services for non-native speakers of the English language. These cloud-based services use speech models developed predominantly using speech data for native speakers. In this study, the cloud-based speech-to-text service is used by a humanoid robot to transcribe the speech of elementary school students in Vietnam. The humanoid robot is meant to assist native Vietnamese students to learn English. The transcribed text is subsequently used by the robot to perform the desired action, which can be an action, movement, or gesture performed by the robot, some image or text displayed on the robot's screen, or some sound produced by the speakers mounted on the robot. However, the transcribed text may not be very accurate, especially, if the speaker is non-native and speaks English with a particular accent such as those from countries in the Asia-Pacific like Vietnam, who speak English with a very unique and distinct accent. For such situations, we propose a novel error detection and correction module to improve the accuracy of the cloud-based speech-to-text service.

The proposed error correction module for speech transcripts of non-native English speakers, i.e., elementary school students in Vietnam whose native language is Vietnamese, is illustrated in Figure 1. As shown in Figure 1, the speech is first transcribed using Google Speech-to-Text service, which, as suggested by our empirical findings, is going to contain errors, examples of which are given in Table 1. This is due to the peculiar accent of Vietnamese speakers, which is significantly different from that of native speakers of English on which the models of Google Speech-to-Text are usually trained. The speech transcript is vectorized using a Bag-of-Words model, and the resultant vectors are then used as input to a trained deep neural network (DNN), which generates a corrected version of the input text as output. Before integrating into the error correction module, the DNN is first trained using the actual text and the same text spoken by non-native English speakers and transcribed by Google Speech-to-Text, i.e., recognized text. Before being used as input to the DNN, both the actual and recognized texts are converted into vector form using a Bag-of-Words model as described in Section 2.1. Moreover, the text for training is selected from the curriculum for the English language for elementary schools in Vietnam.

Fig-1. The proposed error correction module for the speech transcripts of non-native English speakers



2.1. Vectorizing Text using the Bag-of-Words Model

We use the Bag-of-Words model to transform the text into vector form so that it can be used as input to train the deep neural network. The BoW is extensively used in natural language processing to convert a text into a bag of its words or corpus without considering the grammar or word order preserving only the multiplicity of words [9]. The BoW model represents a text as a vector with n components, where the jth component of this vector stores the jth word in the bag or corpus and its frequency in the text.

More formally, let us consider a set T comprising of m texts and a dictionary with n words, then T can be represented as a table D with m x n rows, where the i^{th} row of the table is the vector that represents the i^{th} text. Once the text has been converted into vectors then we can extract only the frequency information from these vectors to

further simplify their representation and make them more suitable for use as input to the deep neural networks. Example of such a vector is given in Equation 1, where the numbers show the frequencies of particular words in the text and the indices of these numbers correspond to the particular word in the bag or corpus.

$$B_1 = [1\ 2\ 2\ 2\ 2\ 1\ 1] \quad (1)$$

Vectors of the form given in Equation 1 are used as input to the deep neural network. However, the dimensionality of those vectors is significantly larger than the one given in Equation 1 because in that case the bag or corpus spans over the entire dictionary of words for a particular problem thereby resulting in very sparse vector representations for individual sentences.

2.2. Classification using Deep Neural Network

Once the text is converted into vector form then it can be used to train the deep neural network (DNN) to detect and correct errors in the output of Google Speech-to-Text. The DNN is a feedforward artificial neural network (ANN) but with many more hidden layers packed between the input and output layers than the typical ANN, which is often referred to as a shallow network. The higher number of hidden layers enables the DNN to capture the complex non-linear relationships between the inputs and the outputs. The many layers of the DNN create a hierarchical representation of inputs and map it to the outputs. This hierarchical representation allows the combination of features from the lower layers to generate higher-order features in the upper layers, which enables complex data modeling with relatively fewer units as compared to a similar shallow network [2].

In our study, we use the softmax function as the activation function for the output layer of the DNN. The softmax function assigns a probability ai to each input vector x as given in Equation 2, which represents the probability of the input x belonging to the i^{th} class. Since a_i is a probability, hence it is always positive and in the range $0 \rightarrow 1$. The value of a_i is calculated using Equation 2, which considers all the inputs, i.e., z_j, to the output layer.

$$a_i = \frac{\exp(z_i)}{\sum_{j=1}^{C} z_j} \quad i = 1,2,..,C \quad (2)$$

The training process for the DNN involves tuning the network parameters to minimize the loss function. We use the adaptive moment optimizer, i.e., Adam, for optimizing our loss function and thereby training our DNN [10].

3. Data and Experimental Setup

Figure 2 shows the intelligent humanoid robot, named BonBon [8], which was used for implementing and testing the proposed novel error detection and correction module for improving the accuracy of speech recognition for non-native speakers of English, i.e., the elementary school students in Vietnam. BonBon was developed to help elementary school students in Vietnam learn the English language. It comprises an upper body with a head, a ribcage, a pair of arms and hands, and a mobile platform that rests on three omnidirectional wheels. It weighs 35 kilograms and stands 1.2 meters tall. Inside its body, it houses two computers: one is an embedded computer with an Intel Core i7-8559U CPU that is used for providing the safety functions and motion control tasks of the robot, whereas the other is an NVIDIA Jetson AGX Xavier with a 512-core GPU, an 8-core 64-bit ARM CPU and 16Gb RAM to manage the speech and vision processing systems in addition to the task planning function of the robot. We use Google's TensorFlow to train and test our DNN, and the natural language toolkit (NLTK) for pre-processing the text using the Bag-of-Words model.

Fig-2. The intelligent, humanoid robot BonBon that was used for implementing the proposed error detection and correction module

The text corpus used for this study comprised around 500 sentences in the English language that were collected from the curriculum for primary schools in Vietnam. The speech data comprised 100, 000 samples recorded by 200 primary school students, with ages ranging from 7 to 10 years. All the 200 students involved in this study were native Vietnamese, who spoke English with a distinct Vietnamese accent.

4. Results and Discussion

In order to test our proposed error detection and correction module, speech data for our text corpus as mentioned in Section 3 was collected for 200 primary school students totaling 100.000 samples in all. These samples were then transcribed to text using Google Speech-to-Text. The output of Google Speech-to-Text for the speech data corresponding to each of the 500 sentences in our text corpus was organized using special data structures. This data structure comprised of two components, i.e., a tag that basically served as a class label for the input data or sentence, and patterns that stipulated the different ways in which that sentence is recognized by Google's Speech-to-Text. The data stored in these data structures were then converted into vectors using the Bag-of-Words model and then used to train our DNN. The accuracy of recognition of each sentence is measured using Equation 3.

$$\alpha = \left(\frac{S_1}{N_1} + \frac{S_2}{N_2}\right)/2 \quad (3)$$

where α represents the recognition accuracy, S_1 represents the number of correct words in the recognized sentence compared to the original sentence, N_1 is the number of words in the original sentence, S_2 is the number of correct words in the original sentence compared to the recognized sentence, and N_2 is the number of words in the recognized sentence. The recognition accuracy for each sentence as listed in Table 1 is calculated using the relation in Equation 3. Table 1 lists some sentences in both their original form, and the way they are recognized by Google Speech-to-Text.

Table-1. Analysis of accuracy for different sentences, both original and recognized

S.No.	Original Sentence	Recognized Sentence	Accuracy (%)
1	Do you have a smartphone?	Do you have a microphone?	80
2	Do you have a smartphone?	Do you have my phone?	60
3	What makes you relax?	What makes you real life?	67.5
4	What is your hobby?	What is your Hobby Lobby?	67.5
5	What is your favorite character?	What is your favorite artist?	80
6	What is your favorite character?	What is your favorite character?	100
7	How does cake taste?	How does cake pic	75
8	Do you have any questions for me?	Do you have a nickname for me?	85.7

Our empirical data suggests that Google Speech-to-Text is not very effective in correctly transcribing the speech of Vietnamese students due to the peculiar accent of those students. For example, when the students are asked to read the sentence "Do you have a smartphone?", it is recognized as either "Do you have a microphone?" or "Do you have my phone?" by Google Speech-to-Text. This recognition error can only be attributed to the difference in the accent of the students, and that of the speech used to train the Google Speech-to-Text models because the students know what they are reading and they read it correctly. Table 1 lists other instances of sentences, which are otherwise read correctly by the students but recognized incorrectly by the Google Speech-to-Text service due to differences in accent and pronunciation of the students. Situations like these, where students utter a sentence correctly, but the speech recognition system does not recognize it, can be frustrating for students and can adversely affect their learning. This is where the proposed error detection and correction module is used to compensate for the student's deficiencies in pronunciation and accent.

Since, the proposed methodology is implemented on a real, humanoid robot, both its speed and accuracy are of paramount importance. Hence, we measure its performance both in terms of speed and accuracy. Table 2 reports the time that the proposed method takes to correctly recognize different sentences. It can be observed in Table 2 that all the sentences take a few milliseconds, 4.0ms at most, which makes it suitable for real-time implementation on a humanoid robot.

Table-2. The processing speed of the proposed sentence recognition and correction algorithm

S.No	Identified Sentences	Time (milliseconds)
1	What's your phone number?	4.0
2	Are you tired?	2.99
3	Do you have a smartphone?	4.0
4	What is your favorite kind of book?	2.98
5	Hello	3.0

Table 3 compares the average accuracy of the proposed methodology and Google's Speech-to-Text on the text corpus and speech recognition data considered for the purpose of this study. It is clearly evident from Table 3 that the recognition accuracy of the proposed methodology is 80%, which is 18% higher than the 62% accuracy yielded by Google Speech-to-Text.

Table-3. Accuracy of the proposed error-detection and correction module, and Google Speech-to-Text

S.No	Method	Average Accuracy (%)
1	Google Speech-to-Text	62
2	Proposed Methodology	80

The proposed error detection and correction module compensates for the difference in the accent of the Vietnamese students and those on which the Speech-to-Text models are developed. The resulting improvement in accuracy makes the robot more effective in communicating with primary school students to help them learn English more effectively.

5. Conclusion

Speech recognition models may do poorly on recognition tasks involving non-native speakers with foreign accents. This study investigated a similar problem in the context of elementary school students who were native speakers of Vietnamese and a humanoid robot meant to help those students learn English. Our study indicated a significant decline in the accuracy of Google's Speech-to-Text, i.e., an average of 62% accuracy on speech data of around 200 students and a set of 500 basic English sentences. To improve the accuracy of Google's Speech-to-Text for Vietnamese speakers of English, we proposed a modified data processing pipeline for our humanoid robot, which incorporated a novel error correction module using the Bag-of-Words model and a deep neural network. The Bag-of-Words model was used to transform the text into vectors that could be used as input to the deep neural network. The deep neural network was first trained using typical sentences in the curriculum for elementary schools in Vietnam and the Google Speech-to-Text data for those sentences. The trained deep neural network was integrated into the error correction module and used for error correction in real time. The proposed methodology was implemented and tested on a humanoid robot. The proposed methodology with the novel error correction module yielded an average speech recognition accuracy of 80%, an 18% improvement over Google Speech-to-Text. Moreover, the proposed methodology executed in real-time demonstrates its potential for use in practical systems.

References

[1] Waibel, A. and Lee, K. F., 1990. *Readings in speech recognition*. San Mateo, CA: Morgan Kaufmann Publishers.

[2] Naumov, M., Mudigere, D., Shi, H. J. M., Huang, J., Sundaraman, N., Park, J., Wang, X., Gupta, U., Wu, C. J., *et al.*, 2019. "Deep learning recommendation model for personalization and recommendation systems, arXiv preprint arXiv." vol. 1906, p. 00091.

[3] Hirose, M. and Ogawa, K., 2007. "Honda humanoid robots development, philosophical transactions of the royal society a: Mathematical." *Physical and Engineering Sciences*, vol. 365, pp. 11–19.

[4] Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B., and Tanaka, F., 2018. "Social robots for education: A review." *Science Robotics*, vol. 3, p. eaat5954.

[5] Wang, D., Wang, X., and Lv, S., 2019. "An overview of end-to-end automatic speech recognition." *Symmetry*, vol. 11, p. 1018.

[6] Xie, X. and Jaeger, T. F., 2020. "Comparing non-native and native speech: Are l2 productions more variable?" *The Journal of the Acoustical Society of America*, vol. 147, pp. 3322–3347.

[7] Yoneyama, K. and Munson, B., 2017. "The influence of lexical characteristics and talker accent on the recognition of English words by speakers of Japanese." *The Journal of the Acoustical Society of America*, vol. 141, pp. 1308–1320.

[8] Le, D. S., Phan, H. H., Hung, H. H., Tran, V. A., Nguyen, T.-H., and Nguyen, D. Q., 2022. "Kfsenet: A key frame-based skeleton feature estimation and action recognition network for improved robot vision with face and emotion recognition." *Applied Sciences*, vol. 12, p. 5455.

[9] Sun, X., Xiao, Y., Wang, H., and Wang, W., 2015. "On conceptual labeling of a bag of words." In *The 24th International Joint Conference on Artificial Intelligence.*

[10] Kingma, D. P. and Ba, J., 2014. "Adam: A method for stochastic optimization, arXiv preprint arXiv." vol. 1412, p. 6980.