A new method to display the accuracy of results for each keyword of the KWS model

Nguyen Tuan Anh, Nguyen Thi Hang

***Abstract*— This study solves the accuracy problem of each keyword when training Keyword spotting (KWS) in non-aligned string results. This approach is called Keyword Detection Accuracy (KWA), which has been improved from the Levenshtein Distance algorithm, it is used to evaluate the accuracy of keywords in KWS by measuring the minimum distance between two strings. The main improvement algorithm is to display the status of each keyword in the training phase for predictive and true labels. In this study, the model used for training is LIS-Net, which is used in Speech Command Recognition. The results of the model are significantly improved compared to baseline models, and the results are displayed on graphs that can see the accuracy of each keyword.**

Keywords: Speech Keyword Spotting; KWS; Keyword Accuracy; Keyword Spotting Accuracy; KWA; Speech Recognition.

I. INTRODUCTION

The method of evaluating keyword accuracy is the goal of this study. The objective of the KWS problem is to detect key phrases in an input utterance. KWS has been developing for many years, getting more attention lately with significant algorithm advancement and quality. This research is inherited and developed from previous research of the author [1]. Currently there are many research methods, using only Audio, without labels [2]. The supervised learning method uses both audio and labels, from the use of traditional methods [3], to the basic forms of Deep Learning [4], and Deep Neural Network Based types are of great interest [4]–[8], with different methods of evaluating results, but all of them have not solved the KWS results as a string. KWS can be classified into two categories: classification and regression. KWS is classified into binary classification and multi-layer classification.

The first type, multiple-class classification, the goal of this type is to classify utterances into groups. Such as in game applications, keywords are forward, backward, left, right, up, down, etc. each keyword is an utterance in the data set with the same length. In 2017, Google has created a dataset with a list of these keywords, called Google Speech Command. This dataset contains 35 keywords, each of them has one-second long, classified into 36 separate groups [9]. With regression type, a data set consists of utterances, with different lengths, in each utterance that can be contained or not one or more keywords in a given keyword list. True labels are strings, they are not classified, and the position of each word in speech data also unknown. KWS's task is to check if the keywords are in utterances, if they are, then which keywords. In essence, this problem is similar to the Speech Recognition problem, but with a much smaller set of word as keywords, the remaining words are garbage [10].

The second type, binary classification, is usually a type of wake-up word, applied in electronic products such as smartphones and smart devices. Some companies are using this type such as Apple with "Hey Siri", Google with "Hello Google", Xiaomi with "Xiao Ai Tong Xue". In this type, it usually only has one keyword, the length of the keyword has little variation in speech data. The KWS's mission is to find out in an utterance that contains or not a keyword, so it is classified into binary classification problem. For example, with Google, a user said "OK Google, open Gmap", after the phrase "OK Google" is detected, a connection will be opened so that the device can communicate directly to a server, and then the server will do the task in the end of the command that converts "open Gmap" into text, understand the semantics and transfer the command to the device to serve the user.

To measure results, in the classification type, there are some methods to do, like confusion matrix, including true positive (TP), true negative (TN), false positive (FP), false negative (FN) and measures based on those values [11], in article [12], they used this method to present the results. Based on these methods, a model based on parameters is evaluated such as true positive rate (TPR), true negative rate (TNR), false positive rate (FPR), false negative rate (FNR), accuracy (ACC), F1 score. With these methods, it is easy to calculate the confusion matrix, but this method cannot apply to string results, because when only one-character changes, the comparison result is no longer accurate. In the regression type, there are some system assessment measures such as: Word Error Rate (WER), Token Error Rate (TER), Character Error Rate (CER), Word Accuracy (WACC). Speech Recognition (SR) accuracy measurement is

mainly based on Word Error Rate (WER) [13], it is calculated based on the Minimum Edit Distance algorithm, and calculations based on unit of word. WER is an effective tool to compare and evaluate the accuracy of different systems as well as the improvement of a system. In KWS, the concept of TER is also used, instead of using WER, it uses each keyword (possibly containing multiple words) as a unit of calculation. CER is used similarly to WER, but the unit of measurement is based on characters. These methods can evaluate the system accuracy, but if a systems with zeros-resource is developed, we will need more information, such as the number of utterances of each keyword, the accuracy of each keyword, the ratio between accuracy and the number of utterances (because of some languages, like Chinese, there are variation, changing the pronunciation according to the words standing next to each other), if using WER only, it is impossible to know exactly.

There are several methods for evaluating the KWS system based on accurate and inaccurate prediction calculations of predictive labels with real labels such as Term Weighted Values (TWV), Maximum Weighted Values (MTWV)[14]. In paper [15], they used Actual TWV (ATWV), they only consider whether or not the keyword is in the predictive label. In the article [16], they used $P@n$ method to present results of top n keywords. In the article [17], they introduced the DR/FA evaluation method for telephone speech, these methods can evaluate the models, but still evaluate the accuracy of entire keyword set, so the problem of estimating the accuracy of each keyword is still unresolved. it is hard to know how many keywords have correctly predicted, not predicted or missed, when the output of KWS model is a string and when training, only accuracy of entire data set is calculated, by calculating the minimum string distance of predicted labels by true labels. When studying the evaluation method of KWS problem, we found that it is difficult to measure the accuracy of each keyword on predicted results. Because KWS model returns the results as strings, so it is difficult to determine the accuracy in percent of each word. But this analysis is necessary, allowing us to know the distribution of each keyword in the dataset, especially with words that have multiple pronouncement ways, mutations and modifications as in Chinese or dialect in other languages, for example, see Table 1.

Table 1. Chinese characters, when reading and writing differently

Character (write)	Pingyin	Read/say
不行	Bùxíng	→ Bùxíng
不变	Bù biàn	→ bú biàn
不爱	Bù ài	→ Bú ài

The more variation, the more data is needed for a keyword during training. Evaluating a KWS model is to evaluate the accuracy of predicted outputs compared to the true labels in the form of string. This study focuses on solving this problem. Different from the existing assessment methods, the objective of this study is to provide a method for calculating the accuracy of each keyword in the output sequence of the Regression problem. Proposing a method to display the results on a new chart type so that we can observe the number of keywords in the data set, the number of correct predictive keywords, false predictions and unpredictable, that's also the reason because the name Keyword Accuracy is selected.

II. THEORY

In this section, several methods will be studied so that they can be compared. As mentioned above, existing method of expressing results can be classified into two categories, classification and regression. Classification type is easily calculating results into confusion matrix parameters such as true positive, false positives, false negatives, true negatives. The second type, regression, is a comparison between the predicted string labels and the true labels that currently applied by WER and the result is accuracy over the entire data set. In this study, the regression model is focused for strings predicted results.

Table 2. Typically Used Error Rates and Their Synonyms

Name	Acronym	Formula	Synonyms
False Positive Rate	FPR	$\frac{FP}{FP+TN}$	False Accept Rate (FAR), Fall-out
False Negative Rate	FNR	$\frac{FN}{FN+TP}$	False Reject Rate (FRR), False Alarm Rate
True Positive Rate	TPR	$\frac{TP}{TP+FN}$	True Accept Rate, Sensitivity, recall, Hit Rate True
True Negative Rate	TNR	$\frac{TN}{TN+FP}$	True Reject Rate, Detection, Rate, Specificity, Selectivity
Positive Predictive Value	PPV	$\frac{TP}{TP+FP}$	Precision
Accuracy	ACC	$\frac{TP+TN}{TP+TN+FP+FN}$	
F1 score	F1	$\frac{2TP}{2TP+FP+FN}$	

The first method, the confusion matrix and related formulas, aims to evaluate accuracy in binary and multiple-class classification. To classify results, with binary classifiers, predictive results is classified into one of the two classes that are real positive cases and real negative cases; With multi-keywords, the results are classified into n*n matrices with n being the number of keywords. In a dataset, the number of real positive cases is called condition positive (P), the number of real negative cases is called condition negative (N). Since then, the predicted results are classified into one of four categories, accurate predictions include true positive (TP) and true negative (TN), incorrect predictions include false positives (FP) and false negatives (FN). From the predicted results, the relevant results is calculated as in Table 2, equations obtained from [18]–[20] [21]. Finally, we have methods to evaluate results based on those formulas via receiver operating characteristic (ROC) curves, e.g. TPR/FPR [22], Precision/Recall [23], [24], False reject Rate/ False Alarm Rate ([25], [26]), False Negative Rate/Hourly False Positives [27].

The second method, $P@k$. In the article [23], the accuracy algorithm was used the formula (1) for evaluating method. The returned result is the accuracy of top k keywords in the system.

$$P@k = \frac{|\{W_r\} \cap \{kW_p\}|}{|\{kW_p\}|} \quad (1)$$

where Wr is relevant words, kWp is retrieved words, $P@k$ is a precision measurement. The result returns a number, representing the system's accuracy, for example, $P@6 = 0.617$

The third method, TWV. Term Weighted Value (TWV) is a measurement method of KWS system evaluation, introduced in [14], illustrated by the formula (2) - (5)

$$P_{miss}(\theta) = 1 - \frac{N_{corect}(\theta)}{N_{true}} \quad (2)$$

$$P_{fa}(\theta) = 1 - \frac{N_{incorect}(\theta)}{N_{Ninc}} \quad (3)$$

$$TWV(\theta) = 1 - (P_{miss}(\theta) + \beta P_{fa}(\theta)) \quad (4)$$

With:

$$\beta = \frac{E}{V}(P\,r^{-1} - 1) \quad (5)$$

where θ refers to detection threshold, $N_{correct}$, $N_{incorrect}$ refer to the number of keywords correct and incorrect detections, respectively. N_{true} refers to the number of occurrences of keywords in that utterance, N_{Ninc} refers to the number of incorrectly detected keywords in that utterance, $P_{miss}(\theta)$ and $P_{fa}(\theta)$ denote the probability of miss and false alarm, respectively. The cost/value ratio, C/V, is 0.1, thus the value lost by a false alarm is a tenth of the value lost for a miss. The prior probability of a term, Pr, is 10^{-4} [14]. Detection score is greater than or equal to θ. The result of this method returns a number to evaluate the system, such as TWV = 0.1962. Recently some articles, such as [28], also use this measure method to represent their results, and the value also returns a number to evaluate the accuracy of their model. In order to evaluate the number of keywords and their correlations, it is necessary to do more in another way. This method can evaluate the accuracy of the model, but in speech, it does not only simply consider that True Label and predicted label contain which keywords but also consider the order in which these words appear. So, the WER method is based on the Minimum Edit Distance, which is still used in many speech recognition systems. There are two other methods to calculate accuracy based on TWV method of Actual TWV (ATWV) and Maximum TWV (MTWV). ATWV uses actual decisions to represent the system's ability to predict the optimal operating point given by the TWV scoring metric. MTWV is a TWV value of θ yields the maximum TWV [14]. This method is used by some studies such as ([15], [29]). The fourth method, Minimum Edit Distance (MED). The Levenshtein algorithm ([30][31]) used to calculate the MED between two strings. Suppose the two strings given for comparison are s and t, the length of the strings is |s| and |t|, MED is calculated according to the formula (6) ([31], [30]):

$$MED_{s,t}(i,j) = \begin{cases} \max(i,j) & if\ \min(i,j) = 0 \\ \min\begin{cases} MED_{s,t}(i-1,j) + 1 \\ MED_{s,t}(i,j-1) + 1 \\ MED_{s,t}(i-1,j-1) + 1_{s \neq t} \end{cases} & otherwise \end{cases} \quad (6)$$

If $s_i \neq t_j$ then $1_{s_i \neq t_j} = 1$ and 0 otherwise, $MED_{s,t}(i,j)$ is the smallest distance of the first i characters of s compared to the first j characters of t To measure the accuracy of a model, Word Error Rate (WER) is used, calculated according to the formula (7) [13].

$$WER_{s,t} = \frac{S + I + D}{N} = \frac{MED_{s,t}}{N} \quad (7)$$

Where S, I and D represent the number of substitutions, insertions and deletions, N is the number of words in the reference.

In order to evaluate a KWS problem, we have four main methods as mentioned above, but in all of them, there is no one strong enough to calculate the accuracy of each keyword that one or more keywords are inside a string; Displays the balance distribution of each keyword in the data set. That is the motivation for us to carry out this research. Moreover, this study has provided a new way of displaying graphics, thereby fully demonstrating simultaneous information. That is the motivation for this research to be done.

III. Propose Method

In order to obtain comparable results, in this study, LIS-Net was used. The architecture of LIS-Net network is illustrated in Figure 1. The input layer for 16 kHz raw wave data using to create spectrogram image [32], the next numbers of blocks, called the Light Interior Search block (LIS-Block), and a classification block for creating the number of output classes (N_c) are stacked together. Each LIS-Block is stacked by number of LIS-Cores (core block of LIS network) and enclosed by two convolutions followed by Batch Normalization and activation layers. It aims to increase the ability to learn parameters through intermediate layers. Each output of LIS-Block is transited by a max polling block. Unlike ResNet, LIS-Net's architecture has the reduced width, height and the increased depth of feature tensor after each LIS-Block. In a block, the dimension of the LIS-Core's feature remains unchanged, but it is easy to change the number of cores. It leads to change of network depth easily and can use for different problems. The purpose of this design is aimed at optimizing the network for each further specific problem. Adjusting the width of the network between adjacent LIS-Blocks is done by two layers of convolution and max polling [33].

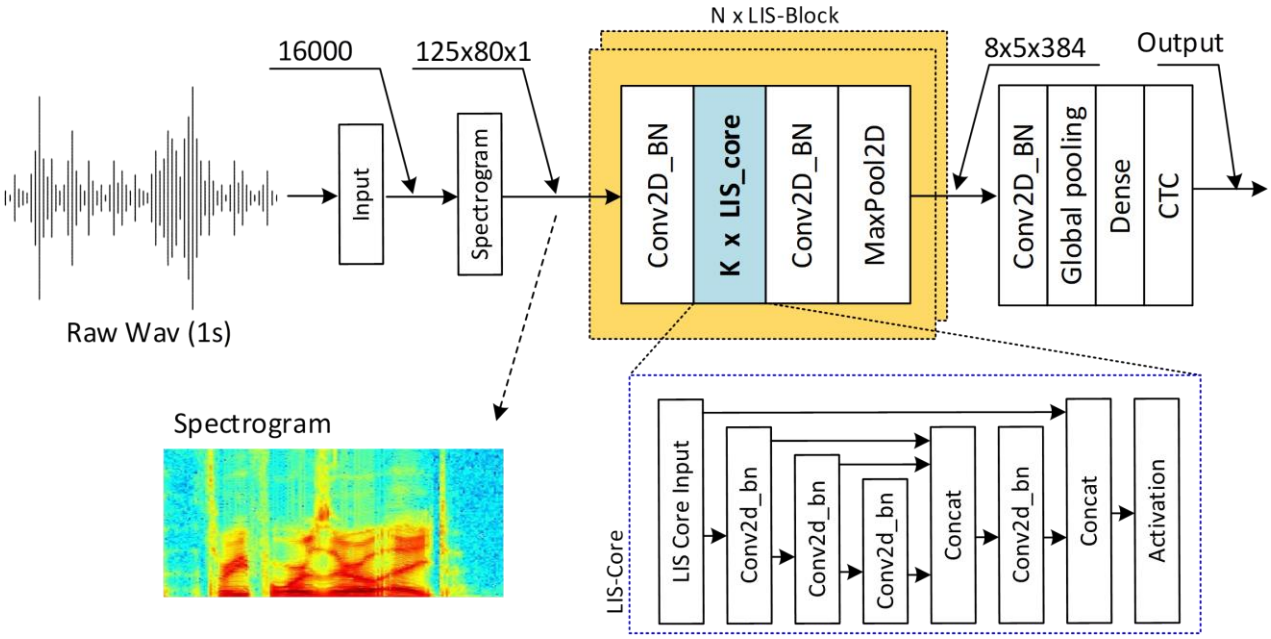

Figure 1. LIS-Net structure

In this study, we propose an algorithm that calculates the accuracy of the model according to the keyword, with the model output being a string of characters that can have keywords or not and proposes a new method of representing the results. This one is improved from the MED algorithm of Levenshtein for the KWS problem. The output of regression model is a string, to match the multi-lingual problem (like Chinese and Vietnamese, completely different from the structure of words). We introduce an algorithm in equation (8) so called Speech Keyword Accuracy (KWA), to determine the exactly editing position of each keyword, based on the MED. To be compatible in multiple languages, each label will be separated into a list of words, in Chinese, separated by each character, in Vietnamese separated by space between words.

$$MED_{s,t}(i,j) = \begin{cases} \begin{cases} i \\ TOC_{1..i,j} = M_{ins} \end{cases} & if\ j = 0 \\ \begin{cases} j \\ TOC_{i,1..j} = M_{ins} \end{cases} & if\ i = 0 \\ min \begin{cases} \begin{cases} MED_{s,t}(i-1,j) + 1 \\ TOC_{i,i} = M_{del} \end{cases} \\ \begin{cases} MED_{s,t}(i,j-1) + 1 \\ TOC_{i,i} = M_{inc} \end{cases} \\ \begin{cases} MED_{s,t}(i-1,j-1) + 1 \\ TOC_{i,j} = M_{sub} \end{cases} & if\ s_i \neq t_j \\ \begin{cases} MED_{s,t}(i-1,j-1) + 1 \\ TOC_{i,i} = M_{eq} \end{cases} & if\ s_i = t_j \end{cases} & otherwise \end{cases} \quad (8)$$

In the KWA algorithm in equation (8), the input is provided by two lists s, t and a list output TOC (abbreviation of type of changes), in which each element is equal, substitution, insertion or deletion, denoted by M_{eq}, M_{sub}, M_{inc} and M_{del}, respectively, each of them is a constant number. The result is updated to a global variable, from there, accuracy of each keyword is obtained as in equation (12), the accuracy of the whole model across the dataset as definition in equation (13). WER based on TOC also observed as in equation (7), where, in each utterance, parameters is calculated as in equation (9)-(11)

$$S_i = \sum_j (TOC_{i,j} == M_{sub}) \quad (9)$$

$$I_i = \sum_j (TOC_{i,j} == M_{inc}) \quad (10)$$

$$D_i = \sum_j (TOC_{i,j} == M_{del}) \quad (11)$$

Or $WER = MED_{s,t}/N$

This study also proposes a method to presenting results in a graph to easily observe the accuracy of each keyword in the keywords set. In Figure 2, The total number of each keyword occurrences denote as N_{kw}: $N_{kw} = N_{ny} + N_{cp}$. This representation method tells us the overall WER of that system, the number of keywords, the status of each keyword, how many percent each keyword predicted correctly, correlation in terms of number of keywords included in dataset and the number of incorrectly predicted words and not yet predicted. That information can be read along the vertical axis on the left. According to the vertical axis on the right, the results in accuracy as a percentage and WER can be observed, either of which may be missing.

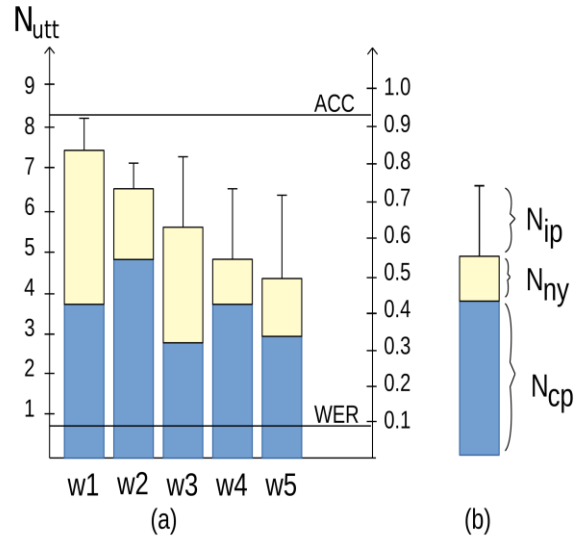

Figure 2. Example of presentation of Speech Keyword Accuracy algorithm

N_{utt}: Number of utterances, w_i $(i = 1,2\ldots)$: predefined keywords,
ACC: Model's accuracy,
WER: keyword error rate of model,
N_{ip}: Number of keywords incorrectly predicted (not in true label),
N_{ny} : The number of keywords not yet predicted,
N_{cp}: Number of keywords correctly predicted.

During training, incorrectly predicted words can have many reasons, which may be due to lack of data, imbalance in the data set (in classification of images dataset or isolated speech dataset maybe easier to identify than speech recognition dataset). From here, in training process, we will be known that which keywords is needed to prepare more training data so each keyword can be balanced on WER with others. The formula for calculating ACC [34] for each keyword (acc_i) is given in equation (12), and global ACC can be calculate as in (13).

$$acc_i = \frac{N_{cp} - N_{ip}}{N_{cp} + N_{ny}} \quad (12)$$

$$ACC = \frac{1}{N}\sum_{i=0}^{N-1} acc_i \quad (13)$$

where N_{cp}, N_{ip}, N_{ny} refer to number of correctly predicted, incorrectly predicted and not predictable, respectively. N denotes as the number of utterances in the dataset. Here, parameters is calculated as equation (14)-(16)

$$N_{cp_i} = j\ (TOC_i, j == M_{eq}) \quad (14)$$

$$N_{ip_i} = j\ (TOC_i, j == \{M_{del} | M_{sub}\}) \quad (15)$$

$$N_{ny_i} = j\ (TOC_i, j == M_{inc}) \quad (16)$$

IV. Experiments and Results

To do the experiment, we selected two small database sets, representing the low-resources languages, ViVos and THCH30.

A. Dataset

THCHS-30 corpus. THCHS-30 corpus is an open speech Chinese database [35], publicized in Openslr, for a total of up to 30 hours for free of reading audios with labels, recorded in a quiet room. To get results for the KWS problem, 10 keywords are selected and implemented by taking 10 words with the highest occurrence frequency in the entire data set to perform the test. After selecting, we have the following keyword list:

KW = [的, 一, 有, 人, 了, 不, 为, 在, 用, 是]

(De, yī, yǒu, rén, le, bù, wèi, zài, yòng, shì)

ViVos corpus. ViVos corpus is an open speech Vietnamese data set [36]. It includes 15 hours of voice recording for ASR purposes. published by AILAB, VNU's computer science laboratory - Hanoi University of Technology. The method of selecting keywords is the same as on THCH-30 dataset, and the keyword list has been selected including 6 keywords as:

KW= [Bật đèn, Tắt đèn, Kéo rèm, Đóng rèm, Mở cửa, Khóa cửa]

These two sets of data will be used to train with LSTM-CTC model based on [37], outputs of the model and true labels are saved to calculate KWA and display results.

B. Presentation Method

Both ViVos and THCH-30 data sets are trained by LSTM-CTC model, during training, the model is evaluated by CTC loss, based on [37]. CTC loss does not show us how much the accuracy of the model is, but it is possible to evaluate the same model, the same data set, which training session has lower loss, the weight is better. From there the training system can be optimized, to give out the predicted results of the model and combine it with true labels, calculate accuracy according to each keyword and overall accuracy. The formula (12) and (13) are used. The result of this step is shown on the graphic.

In Figure 3, we can observe, firstly, the number of each keyword is small, and therefore, the difference between the keywords is small, but the percentage is large. Secondly, although the model of accuracy results is quite high, but the percentage of incorrect prediction is also high, and finally, observing WER and accuracy of the system visually, giving us an overview of the model.

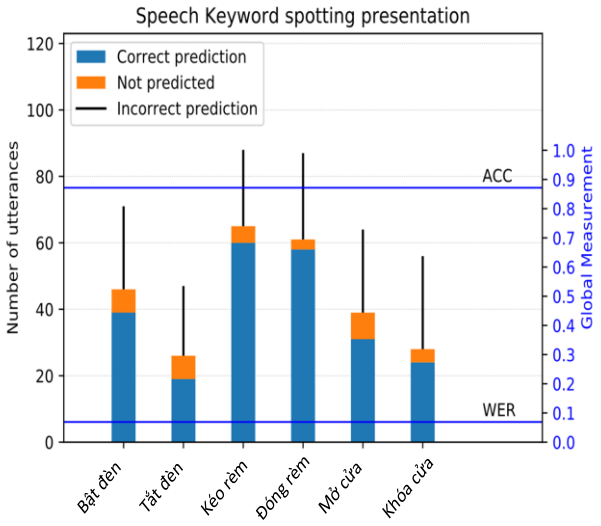

Figure 3. The graph shows the correlation of results between keywords of ViVos dataset

In the Figure 4, it can easily be observed that a huge difference in the number of keywords, the first keyword has approximately twice to sixth times the number of remaining keywords, this leads to difficult for training model to get higher accuracy for the entire set of keywords in the dataset.

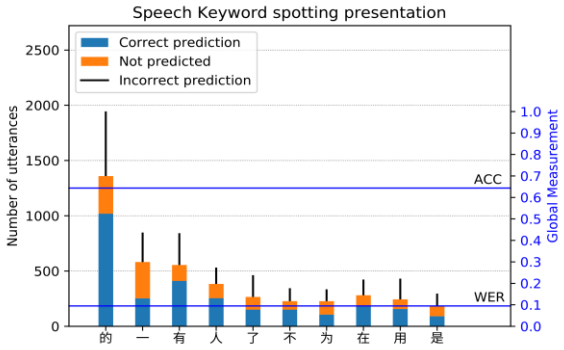

Figure 4. The graph shows the correlation of results between keywords of THCH-30 dataset

On the other hand, it is observed that in the second keyword bar, ACC of this keyword has not reached about 50%, while other keywords having higher ACC, thereby giving us a clue to understanding the cause of global ACC is not high.

V. CONCLUSIONS

This study has just presented a method of measuring the accuracy of keywords in the KWS problem and presented a method to display the accuracy of the whole system on the chart, provide useful information for deep learning models. To measure the accuracy of each keyword, improved MED method is used, in this method, each state such as substitution, insertion, deletion is recorded to evaluate results when comparing predicted strings with true strings. On the graph, the status of each keyword is displayed along with the number of keywords in the training, WER and Accuracy are also shown on the same image, this method makes it easy to observe the status of all model information. This method helps us understand the balance of keywords in the data set instead of WER or accuracy only. Despite many advantages, KWA still cannot avoid such complex drawbacks. Only string data should be used. In many cases it is not necessary to use an accuracy rating to each keyword. This method can be applied to Speech Recognition problem for almost zero-resource languages and semi-supervised ASR, which will be our future research work.

VI. ACKNOWLEDGMENTS

In this article, we would like to especially thank to Thai Nguyen University of Technology, Thai Nguyen, Viet nam for supporting us in the experimental process.

REFERENCES

[1] N. T. Anh and H. T. K. Dung, “Keyword Accuracy : A new method of calculation and representation keyword accuracy for speech keyword spotting in string results,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 2, pp. 658–663, 2020, doi: https://dx.doi.org/10.14569/IJACSA.2020.0110283.

[2] M. Awaid, A. H., and S. A., “Audio Search Based on Keyword Spotting in Arabic Language,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 5, no. 2, pp. 128–133, 2014, doi: 10.14569/ijacsa.2014.050219.

[3] H. F. C. Chuctaya, R. N. M. Mercado, and J. J. G. Gaona, "Isolated Automatic Speech recognition of Quechua numbers using MFCC, DTW and KNN," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 10, pp. 24–29, 2018, doi: 10.14569/IJACSA.2018.091003.

[4] M. K, I. A., and G. Onwodi, "Neural Network Based Hausa Language Speech Recognition," *Int. J. Adv. Res. Artif. Intell.*, vol. 1, no. 2, pp. 39–44, 2012, doi: 10.14569/ijarai.2012.010207.

[5] P. D. Hung, T. M. Giang, L. H. Nam, and P. M. Duong, "Vietnamese speech command recognition using Recurrent Neural Networks," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 7, pp. 194–201, 2019, doi: 10.14569/ijacsa.2019.0100728.

[6] J. Ren and M. Liu, "An Automatic Dysarthric Speech Recognition Approach using Deep Neural Networks," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 12, pp. 48–52, 2017, doi: 10.14569/ijacsa.2017.081207.

[7] M. Walid, B. Souha, and C. Adnen, "Speech recognition system based on discrete wave atoms transform partial noisy environment," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 5, pp. 466–472, 2019, doi: 10.14569/ijacsa.2019.0100560.

[8] M. A. A. Al- Rababah, A. Al-Marghilani, and A. A. Hamarshi, "Automatic detection technique for speech recognition based on neural networks inter-disciplinary," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 3, pp. 179–184, 2018, doi: 10.14569/IJACSA.2018.090326.

[9] P. Warden, "Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition," 2018.

[10] E. Chandra and K. A. Senthildevi, "Keyword Spotting: An Audio Mining Technique in Speech Processing – A Survey," *IOSR J. VLSI Signal Process. Ver. II*, 2015, doi: 10.9790/4200-05422227.

[11] Wikipedia contributors, "Confusion matrix - Wikipedia, the free encyclopedia." [Online]. Available: https://en.wikipedia.org/w/index.php?title=Confusion_matrix&oldid=881721342. [Accessed: 31-Mar-2019].

[12] Z. Wang, X. Li, and J. Zhou, "Small-footprint Keyword Spotting Using Deep Neural Network and Connectionist Temporal Classifier," 2017.

[13] Wikipedia contributors, "Word error rate." Wikipedia, The Free Encyclopedia., 2019.

[14] G. Fiscus, Jonathan G and Ajot, Jerome and Garofolo, John S and Doddington, "Results of the 2006 spoken term detection evaluation," *Proc. sigir*, vol. 7, pp. 51--57, 2007.

[15] Y. Bai *et al.*, "End-to-end keywords spotting based on connectionist temporal classification for Mandarin," in *Proceedings of 2016 10th International Symposium on Chinese Spoken Language Processing, ISCSLP 2016*, 2017, doi: 10.1109/ISCSLP.2016.7918460.

[16] H. Kamper, G. Shakhnarovich, and K. Livescu, "Semantic keyword spotting by learning from images and speech," *arXiv Prepr. arXiv1710.01949*, 2017.

[17] J. Nouza and J. Silovsky, "Fast keyword spotting in telephone speech," *Radioengineering*, vol. 18, no. 4, pp. 665–670, 2009.

[18] S. Marcel, M. S. Nixon, and S. Z. Li, Eds., *Handbook of Biometric Anti-Spoofing*. London: Springer London, 2014.

[19] D. M. W. Powers, "Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation," 2007.

[20] P. Wu, H. Liu, X. Li, T. Fan, and X. Zhang, "A Novel Lip Descriptor for Audio-Visual Keyword Spotting Based on Adaptive Decision Fusion," *IEEE Trans. Multimed.*, vol. 18, no. 3, pp. 326–338, 2016, doi: 10.1109/TMM.2016.2520091.

[21] T.~Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861--874, 2006.

[22] R. Menon, H. Kamper, J. Quinn, and T. Niesler, "Fast ASR-free and almost zero-resource keyword spotting using DTW and CNNs for humanitarian monitoring," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2018-Septe, pp. 2608–2612, 2018, doi: 10.21437/Interspeech.2018-1580.

[23] I. Pratikakis, K. Zagoris, B. Gatos, G. Louloudis, and N. Stamatopoulos, "ICFHR 2014 Competition on Handwritten Keyword Spotting (H-KWS 2014)," *Proc. Int. Conf. Front. Handwrit. Recognition, ICFHR*, vol. 2014-Decem, pp. 814–819, 2014, doi: 10.1109/ICFHR.2014.142.

[24] Y. Huang and W. Y. Wang, "Deep Residual Learning for Weakly-Supervised Relation Extraction," *Proc. 2017 Conf. Empir. Methods Nat. Lang. Process.*, pp. 1803–1807, 2017, doi: 10.18653/v1/D17-1191.

[25] G. Chen, C. Parada, and G. Heigold, "Small-footprint keyword spotting using deep neural networks," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 4087–4091, doi: 10.1109/ICASSP.2014.6854370.

[26] T. N. Sainath and C. Parada, "Convolutional Neural Networks for Small-footprint Keyword Spotting," *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH 2015*, 2015.

[27] A. AbdulKader, K. Nassar, M. Mahmoud, D. Galvez, and C. Patil, "Multiple-Instance, Cascaded Classification for Keyword Spotting in Narrow-Band Audio," no. Nips, 2017.

[28] M. J. F. Gales, K. M. . Knill, A. Ragni, and S. P. . Rath, "Speech recognition and keyword spotting for low resource languages: Babel project research at CUED," *Spok. Lang. Technol. Under-Resourced Lang.*, no. May, pp. 14–16, 2014.

[29] M. J. F. Gales, K. M. . Knill, A. Ragni, and S. P. . Rath, "Speech recognition and keyword spotting for low resource languages: Babel project research at CUED," in *Spoken Language Technologies for Under-Resourced Languages (SLTU)*, 2014, no. May, pp. 14–16.

[30] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals.," *Dokl. Akad. Nauk SSSR*, vol. 163, no. 4, pp. 845–848, 1965.

[31] "Levenshtein distance," *https://en.wikipedia.org/wiki/Levenshtein_distance*, 2018.

[32] T. Kim, J. Lee, and J. Nam, "Sample-level CNN architectures for music auto-tagging using raw waveforms," *2018 IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 366–370, 2018.

[33] N. T. Anh *et al.*, "LIS-Net : An End-to-End Light Interior Search Network for Speech Command Recognition," *Unpublished*.

[34] A. Ogawa, T. Hori, and A. Nakamura, "Estimating speech recognition accuracy based on error type classification," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 12, pp. 2400–2413, 2016.

[35] Z. Zhang, D. Wang, and X. Zhang, "THCHS- 30 : A Free Chinese Speech Corpus," 2015.

[36] H.-T. Luong, H. Chi Minh City, and H.-Q. Vu, "A non-expert Kaldi recipe for Vietnamese Speech Recognition System," 2016.

[37] Alex Graves and N. Jaitly, "Towards End-To-End Speech Recognition with Recurrent Neural Networks," *Proc. 31st Int. Conf. Mach. Learn.*, vol. 32, no. 1, pp. 1764–1772, 2014, doi: 10.1145/1143844.1143891.

Nguyen Tuan Anh: School of Electronic and Information Engineering, South China University of Technology, Guangzhou, 510641, P.R. China 0084988086099.

Nguyen Thi Hang, Faculty of Construction and Environment, Thai Nguyen University of Technology, Thai Nguyen, Vietnam, 0084968046789.