International Journal of Engineering & Technology, 8 (1.11) (2019) 158-160

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper



Automatic Selection of Speech Data based on Confidence Measure

Mustafa Abdallah[1], Abdullah M. Moussa[1], Sherif M. Abdou[2], Mohsen Rashwan[1], Hassanin Al-Barhamtoshy[3*]

[1]Faculty of Engineering, Cairo University, Giza Egypt
[2]Faculty of Computers and Information, Cairo University, Giza Egypt
[3]Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia
**Corresponding author E-mail: hassanin@kau.edu.sa*

Abstract

The amount of training data used in automatic speech recognition and pronunciation aiding systems is one of the most important factors that can significantly affect the quality of the resulting systems. However, as the amount of training data increases, a huge effort in transcribing the data by professional linguists is needed. This task is usually expensive in terms of time and money. In this paper, we present an algorithm to automatically select more accurate subsets of speech data with high accuracy. The suggested algorithm utilizes confidence measures and posterior probabilities to extract parts of the data based on a confidence score. Experimental results and comparisons with a manually verified selection process and a random selection process show that the proposed algorithm is Robust and effective

Keywords: *data selection; confidence measure; speech processing; posterior probabilities; machine learning*

1. Introduction

Automatic Speech Recognition (ASR) and Computer Aided Pronunciation Learning (CAPL) are among the most important applications of machine learning. As of any machine learning application, the training data play a significant role in the resulting system quality. Regarding ASR and CAPL systems, the amount of training data is of special importance, however, the more training data used in the system, the more effort needed to transcript and validate the data. This process can be expensive in terms of time and money. While traditional speech data can be transcribed by non-specialized data entry operators, some CAPL systems –e.g. a Holy Quran recitation aiding system-need professionals who should master a special knowledge. This may make the process even more costly to accomplish. So, there is always a need to robust algorithms that can extract more accurate subsets from training data to meet both high accuracy and cost effective requirements.

Automatic selection of speech data has gained a wide attention in the last years. Several algorithms have been proposed to tackle the problem. For examples, in [1] the authors developed a data selection technique that is based on the maximum entropy principle to select speech utterances that contribute to a uniform distribution across different speech units. Also in [2], the authors presented a technique that uses principal component analysis to map the variance of the utterances in a speech database into a low-dimensional space, followed by clustering and a selection procedure to automatically choose subsets from the database. Other methods suggest a speech data subset selection approach based on submodular functions [3, 4]. Automatic selection can be also applied to speakers [5] or age groups [6]. This may be used when there are large variations in the speakers' spaces or to improve speech recognition accuracy using age group-specific acoustic models respectively. Speak correctly system [7] have been implemented to estimate the correct words to speak using HMM's states to find the acoustic features and back propagation algorithm to train that networks.

In this paper, we present a new algorithm to automatically select high quality subsets of speech data with high accuracy. The selection criterion is based on confidence measures and posterior probabilities to choose subsets of the data based on a prespecified confidence score. The experiments and comparisons with a manual selection criterion and a random selection criterion show that the proposed algorithm is promising. The rest of the paper is organized as follows: Section describes the proposed automatic selection algorithm. Section III illustrates the experimental and comparisons results. And finally the conclusions will be presented in section IV.

2. Proposed Algorithm

The inputs of the algorithm are a set of speech utterances from which we need to select a more accurate subset, a threshold of the confidence (i.e. the minimum confidence for each accepted recognized phoneme) and a Hidden Markov acoustic Model.

Each utterance in the aforementioned set is decoded using the acoustic model and recognition lattice is extracted for this utterance. After that, the N-best paths (i.e. best hypotheses generated from recognition) are generated from the lattice.

Then, backward probability is computed for each node in the N-best paths. Backward probability of a node is the summation of log probabilities of all paths from this node to the end of the N-best lattice. In addition, forward probability is also calculated for the node, which is the summation of log probabilities of all the paths from the beginning of lattice to that node.

After that, the total summation of log probabilities of all possible paths in the lattice is subtracted from the summation of forward and backward probabilities for each node. The result of such divi-

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

sion is called the confidence score of the node. Now, the best path in the lattice is selected as the path with the largest summation of confidence measures. Each confidence score for each node in the selected path is compared with the threshold. If confidence scores of all the nodes in the best path exceed the specified threshold, the utterance is accepted and included in the selected subset. Otherwise, the full utterance is rejected. Finally, the accepted ratio is calculated as the number of accepted utterances divided by the number of utterances in the input set. The proposed algorithm is summarized in figure 1.

Algorithm 1: Automatic selection of speech data

input : An HMM acoustic model *M*, a set *S* of *n* utterances, a threshold τ
output: A selected subset S_s, an acceptance ratio *R*.

1 Initialize $K = 0$
for each utterance *U* in *S* **do**

2 Using *M*, Extract the lattice of *U*
3 Generate the N-best alternatives in recognition **for each** node in the lattice **do**
4 Compute the log forward probability of the node
5 Compute the log backward probability of the node
6 Subtract total score of the whole lattice paths from the sum of forward and backward probabilities

end

7 Add the confidence scores of each node for each complete path and select the best path.

8 **for each** node in the best path **do**

if (the confidence score of the node > τ)

Include *U* in S_s

$K = K + 1$
else
Reject *U*

end
end
end

9 $R = K/n$

10 Return S_{hq} and *R*

Fig. 1 Data selection proposed algorithm

3. Experimental Results

In order to evaluate the proposed algorithm, several experiments have been conducted. First, we have built a triphone-based HMM acoustic model using 24 training hours dataset of a recitation of the Holy Quran. We have used another dataset for testing consists of a half an hour of a Holy Quran recitation. The total number of phonemes in the testing dataset is about 5,000 phones. This dataset was manually transcribed and revised by Holy Quran recitation experts. The features extracted from speech utterances were Mel frequency cepstral coefficients (MFCCs) and energy, along with their first and second temporal derivatives. Figure 2 shows the accuracy percentage ranges of the subgroups of the testing data.

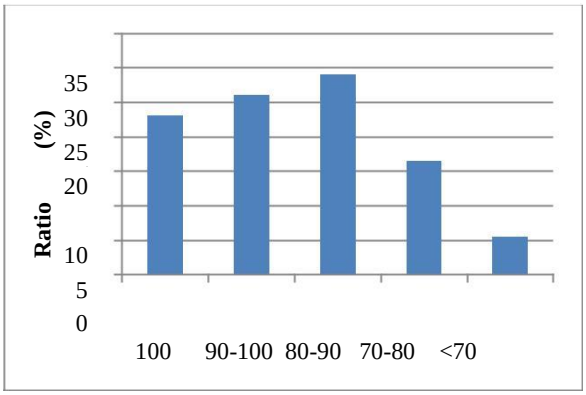


Fig. 2 Accuracy of sub-groups of testing dataset utterances

To check the sensitivity of the confidence score parameter, we have applied the proposed technique on the testing data and calculated the accuracy in terms of Phoneme Error Rate (PER) obtained after changing the value of the confidence score. The results are summarized in Table I.

Table I. Acceptance Ratio And Per Obtained Under Confidence Score Variations

Confidence Score	Acceptance Ratio (%)	PER (%)
0.7	40	93.4
0.8	33	94.34
0.85	32	94.69
0.9	30	95.16
0.95	27	95.93

The confidence score here is calculated on the phoneme level. e.g., if confidence score equals 0.7, this means that if any utterance contains any recognized phoneme with score less than 0.7, the full utterance will be rejected. This is a tough criterion since it can lead to many rejections that could be avoidable if the confidence score was applied on the utterance level.

It is clear from Table I that if the confidence score is small, the acceptance ratio of the selected data will be large. On the other hand, we will not be confident of our recognition results.

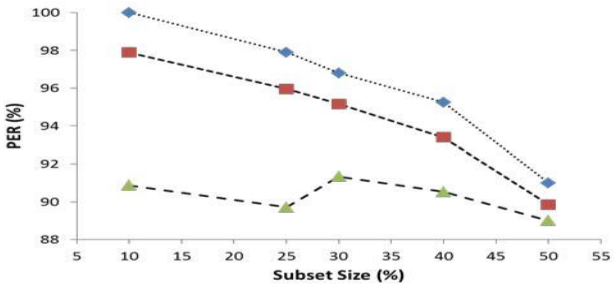
However, if the confidence score is large, the acceptance ratio of the selected data will be small, and our confidence in the selected data will be higher.

Another experiment has been done to evaluate the effectiveness of the presented technique. We have compared the proposed algorithm with a manual selection process that extracts the highest quality subsets from the data. We have also compared the proposed algorithm with a random selection process. Both of the three techniques (proposed, manual and random) have been used to extract 10%, 25%, 30%, 40% and 50% from the data then we have calculated the accuracy of the extracted subsets in terms of PER. The results are summarized in figure 3.

selections and random selections show that the suggested technique is effective and efficient. Future work may include investigating confidence score improvements using different probabilistic models. In addition, it is interesting to investigate the proposed algorithm to extract a small ratio from a speech dataset with very high confidence score and merge the selected subset with the data used to build the current acoustic model. Such acoustic model is used to build a new model and repeating such process recursively in the sake of a more accurate acoustic model that can select more data with higher confidence.

Acknowledgment

The authors thank King Abdulaziz University (KAU), Jeddah for providing necessary facilities to carry out the research.
The simulations in this work were performed at King Abdulaziz University's High Performance Computing Center (Aziz Supercomputer) (http://hpc.kau.edu.sa).



BeforeFig.3 you Accuracy begin comparisontoformatinyourtermspaper,ofPER firstbetweenwritethe andproposedsave algorithm (red points), a manual selection process (blue points) and a

As we cansee,randomwhileselectiontherisprocessasignificant(greenpoints)difference. in

As we can see, while there is a significant difference in accuracy between the manual selection process and the random selection one, the proposed technique has a comparable accuracy with the manual selection criterion. Moreover, our proposed technique is far from random selection method.

4. Conclusions

In this paper, we have presented an algorithm for automatically select subsets from speech dataset. We have conducted several experiments to validate the proposed algorithm. Experimental results and comparisons with manual

References

[1] Wu Y., Zhang R., and Rudnicky A.. Data selection for speech recognition. In Automatic Speech Recognition & Understanding, ASRU. IEEE Workshop, pp. 562-565, 2007.
[2] Nagórski A., Boves L., and Steeneken H. J.. Optimal selection of speech data for automatic speech recognition systems. In INTERSPEECH, 2002.
[3] Wei K., Liu Y., Kirchhoff K., and Bilmes J.. Unsupervised submodular subset selection for speech data. In Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference, pp. 4107-4111, 2014.
[4] Wei K., Liu Y., Kirchhoff K., Bartels C., and Bilmes J.. Submodular subset selection for large-scale speech training data. In Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference,
[5] 3311-3315, 2014.
[6] Christensen H., Casanueva I., Cunningham S., Green P., and Hain T. Automatic selection of speakers for improved acoustic modelling: recognition of disordered speech with sparse data. In Spoken Language Technology Workshop (SLT), 2014 IEEE, pp. 254-259, 2014.
[7] Hämäläinen A., et al.. Improving Speech Recognition through Automatic Selection of Age Group–Specific Acoustic Models. In International Conference on Computational Processing of the Portuguese Language,
[8] 12-23, Springer International Publishing, 2014.
[9] Al-Barhamtoshy H., Abdou S., and Jambi K. Pronunciation Evaluation Model for None Native English Speakers, http://www.lifesciencesite
[10] .com/lsj/life1109/030_24719life110914_216_226.pdf, pp. 216-226, 2014.