Pronunciation Tutor for Deaf Children based on ASR

Yunling Bai

Glasgow College, University of Electronic Science and Technology of China, Sichuan, China

Abstract. ASR, whose full name is Automated Speech Recognition, is a technology that converts human speech into text. Speech recognition, a multidisciplinary field, is closely related to acoustics, phonetics, linguistics, digital signal processing theory, information theory, computer science and other disciplines. ASR has been applied in educational technology such as deaf children's education in this day and age. This paper makes a preview of a project in which a computer-aided tutor for deaf children instruction based on the speech recognition technology. This tutor utilizes three effective models and is combined with data mining technology. Two evaluation approaches and overview of embedded experiment are also detailed in this paper.

Keywords: ASR; Deaf Children Education; Machine Learning.

1. Introduction

According to a sample survey conducted by relevant institutions, there are about 60 million disabled people in China, among which about 13 million are deaf and dumb, and about 1 million deaf people under 18 years old should be educated, which is a very large disadvantaged group (Google,2022). They have many unimaginable hardships and obstacles in learning, working and living, and have to work many times harder than normal people on their way to grow and develop.

Some economically developed countries and regions such as Europe, the United States, and Hong Kong have an early start and high level of deaf education[1]. At this stage, domestic and foreign deaf education and rehabilitation institutions are more focused on the research and equipment of special equipment. But in the application of modern teaching technology, especially in the network, computer, Automatic Speaking Recognition technology and other latest technological achievements in the application of China is basically in the initial stage. As a matter of fact, the application of new theories, new technologies and new equipment in modern information and other fields of science and technology, medicine, etc. to the rehabilitation and education of deaf people will certainly will certainly change the concept of deaf education. This will have a great impact on the educational reform of deaf schools.

This paper previews an application which firstly combines modern educational technology and the special needs of deaf education, by applying the results of information technology and software development to deaf education. This project aims to use modern educational technology and tools, based on Automatic Speaking Recognition (ASR) technology, to research, reform, and practice in the technical level of deaf education, focusing on cultivating deaf children's interest in learning, enabling them to master the ability to collect, analyze, and process information, to master modern means of self-learning and lifelong learning, and to improve their own quality and ability to adapt to social life, so that they can be on an equal footing to enable them to participate in social life on an equal footing and with equal opportunities, and to share the material and cultural achievements of society.

2. Related Work

The research on speech training for deaf people began in the mid- 1960s abroad. The auditory feedback training system, which uses the residual hearing of deaf people to correct their own pronunciation with the help of hearing aids, and the visual feedback training system, which corrects pronunciation by observing the characteristic parameters obtained through the processing of their own pronunciation on the CRT. The former is low cost, but the effect is poor or completely ineffective. In the early development of the visual feedback system components are high, with the development of computer and large-scale integrated circuit technology, especially the emergence of speech-

specific chips and micro-controllers, the cost has been greatly reduced. A variety of visually assisted speech training systems have been reported from abroad. These systems basically process the speech pronounced by the trainees and extract the characteristics of the speech (e.g., intensity, duration, spectrum, fundamental frequency, resonance peaks, etc.) with the characteristics of the standard pronunciation simultaneously displayed on the CRT, allowing the trainees to compare their pronunciation with the standard ones and gradually correct their pronunciation. Unfortunately, the information displayed in this system is too transferable for the average trainee to understand, especially for deaf children, and therefore affects the effectiveness of the training. This affects the effectiveness of the training.

As for the focus of this topic, Automatic Speech Recognition research, it started in the early fifties when electronic signal spectrum analysis instruments began to be used to recognize simple, small amounts of syllables and phonemes from speech signals. With the rapid development of computer technology, the research on speech recognition further heated up after entering the nineties, and many practical research directions emerged in addition to continuous speech dictation machines. via Voice, pioneered by IBM, marked that large vocabulary, non-person-specific, continuous speech recognition technology was maturing. There are also many more mature speech ASR products in the market, and most of them support secondary development, such as Microsoft's Speech Application SDK (SASDK) and JavaSpeechAPIBM's Dutty++ advocated by SUN. Most of them can recognize languages of different countries such as English, Japanese and Chinese, and Dutty++ can even recognize dialects of certain regions, such as Guangdong dialect a Cantonese.

3. Methodology

3.1 Overview

The tasks can be divided into three parts: the first part is to carry out speech recognition, capture the trainees' pronunciation, then put the pronunciation into the sentence for positioning, and finally give the correctness judgment and corresponding instruction. For the first speech recognition task, we propose to use CLDNN structure and discuss its advantages in following passage. In this project the SVD model was used to judge pronunciation correctness. Besides, we decide to use the confidence measure to reduce the error rate by words corresponding.

3.2 Models

3.2.1 CLDNN Model

To start with, a voice capture structure based on ASR needs to be established with a purpose of recognizing the pronunciation of the trainees. In this structure the input is the original waveform captured by the microphone, and the output is the text version of the input, which is the data we need.To accomplish speech recognition task, an improved Deep Neural Networks is proposed to be utilized, called CLDNN.

In the past few decades, compared with Gaussian Mixture Model/Hidden Markov Model (GMM/HMM) systems[2], Deep Neural Networks (DNNs) have shown significant advantage in large vocabulary continuous speech recognition (LVCSR) tasks[3]. In recent years, various neural networks such as Convolutional Neural Networks (CNNs) [4] and Long-Short Term Memory Recurrent Neural Networks (LSTMs) [5]] have been created based on advanced DNNs. The above-mentioned architectures all have distinctive merits: CNNs have an advantage in reducing frequency variations, LSTMs are applicable in temporal modeling and DNNs are good at mapping features to a more separable space. Then, an improved Deep Neural Networks structure, CLDNN, which combined these three architectures for their complementarity was constructed[2].This model can be used in ASR and according to the data provided in the paper, CLDNN provides a 4-6% relative improvement in word error rate (WER) over a LSTM, which is the strongest model in CNNs, DNNs

and LSTM (Google, 2015). This makes it one of the best neural networks so far and this is the reason why we chose it to do ASR for the application.The structure of CLDNN is shown in following figure.

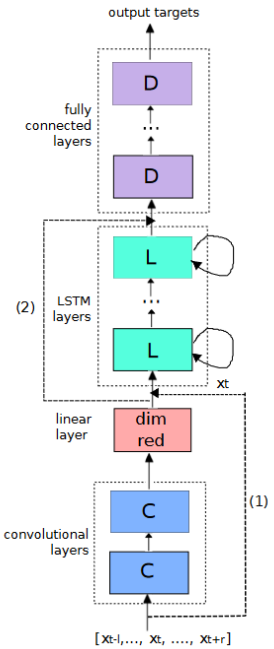

Fig 1. The structure of CLDNN

3.2.2 Support Vector Machine (SVM) Model

In the application, we provide multiple methods to help the user to adjust his/her pronunciation including play video clips of a child's mouth saying that word, showing the phonogram of the word, showing other words that rhyming with the original word, separated the word by few pauses or showing the spectrum-gram of their voice. In order to maximize the efficiency of the application, the application needs a model to decide which model is the most effective way to help a particular user. For instance, play video clips of a child's mouth saying that word may be the best way for some users to adjust their pronunciation while other users may prefer other method. SVM, as a machine learning algorithm can do it dynamically. The application can gather the data of the users while they are using it and make itself more suitable for this particular user.

SVM is a supervised classification algorithm, its general idea is that, suppose there are two classes of points on the sample space, we want to find a division hyperplane to separate these two classes of samples, and the division hyperplane should choose the one with the best generalization ability [6]. In practice, the application would first collect the data of the user as the variable in SVM. For instance, the application chooses "spectrum-gram" as the method to help the user when the user pronounces a word incorrectly. If the next time the user read the word correct, this data will be marked as '1', otherwise, marked as 0[6]. Suppose the application has n different methods, than the data could form a n-dimensional space and SVM can give us a n-1 dimensional space which classifies these data if the data is linearly dividable. This n-1 dimensional space contain the weight of the method as desired. However, if the data can't be divided linearly, we need to find an appropriate kernel function to map the data into a higher dimensional space in order to classify it.After that, the application can find the method that is most suitable for the current user and use it as the major method[6].

3.2.3 Confidence Measure

In practice, when students read sentences, they may make a small mistake of reading the wrong word order first. The main purpose of our application is to correct students' pronunciation, so we should reduce the sensitivity of the decision algorithm to such errors. In order to solve this problem, we intend to borrow the idea in [7] and use a constrained language model generated from the

sentence[8]. This model can reduce the false alarm rate by ignoring the word's sequence mistakes in the sentence. Without confidence measures[7], the determination will be a 0-1 hard decision.

As long as some hypothesized word h_t that match target word w_t is aligned against w_t the application classifies w as read correctly as shown in Fig 3. Instead, confidence measure is proposed to transform hard decision into soft decision between [0,1] by estimating Pr (Word was read correctly| Features) (or Pr(W|F). This improvement will provide more information to judge the accuracy of word pronunciation and can improve application flexibility by adjusting the threshold.

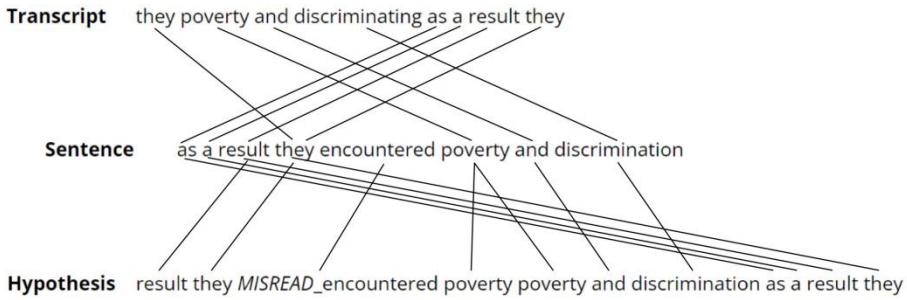

Fig 2. Alignments of a hypothesis (H) against a sentence (S), and a transcript (T) against S[7].

T	C	C	C	C	O	C	C	S
S	as	a	result	they	encountered	poverty	and	discrimination
H	C	C	C	C	S	C	C	

Fig 3. Classifications of words from sentence (S) based on alignments against hypothesis (H) and transcript (T) in Fig.2 (where c=correct, o=omission, s=substitution)[7]

4. Data Mining

Nowadays, data mining technology, which blooms rapidly, has been widely used in education field. Recording and analyzing data logged by tutors helps to dig out potentially useful information, which renders education dramatically more efficient [9]. In this deaf-oriented tutoring project, suitable data mining applications will bring considerable benefits.

4.1 Time Events

Target student of this tutor is relatively special. To make education more responsive to the needs of deaf students, personalizing the service, the number of times students ask for help when they make mistakes with different consonants and how long they spend to correct their errors can be recorded[9]. By analyzing the time-related data, individual pronunciation weaknesses can be identified. In this way, targeted and enhanced tutoring can be implemented.For individual student, the tutor can develop a schedule for frequent training and testing to eliminate personal errors. For all entire hearing-impaired customers, we can provide intensive tutoring on usual weaknesses, such as providing visual pronunciation animations from more angles and parsing of pairs of confusing consonants.

4.2 Import Student Data

According to [9], normal speech tutor will require students to input relevant information, such as gender, age, and IQ. The tutor oriented to special population should collect more detailed information, such as whether they have residual hearing or normal hearing history. With personal information, students can be classified into groups, enjoying more considerate services. Specifically, we will set up a more appreciate evaluating system for groups of students with different listening statement, in order to protect students' learning enthusiasm and self-confidence. For example, for students who are

born without their entire hearing, we use the most relaxed criteria to score their speech. Certainly, all criteria are based on data derived from the similarity between normal adults and students.

4.3 Probe Student Knowledge

Despite that normal explicit probes are equipped, additional probes helps to mine more information[9]. With a purpose of mastering students' learning progress, popover tests are generated the day after learning the consonant sounds of a word. Students' pronunciation is evaluated as True or False by ASR, and these data are collected together with the knowledge tracking diagram, which is shown as Fig 4 and utilized in [10], to judge the mastering progress of word pronunciation for students. the results can be used in evaluation of this tutoring app.

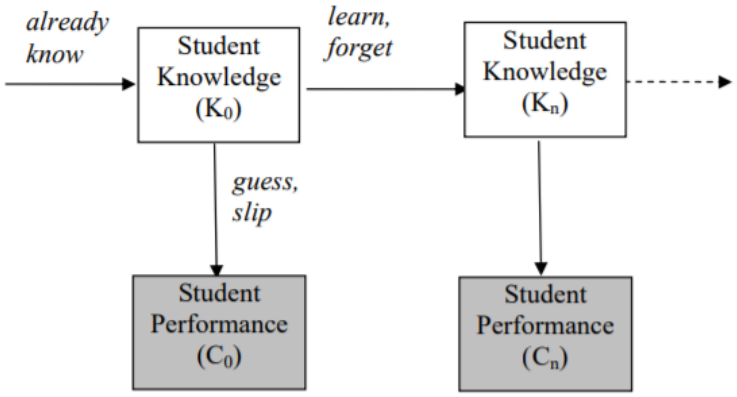

Fig 4. Diagram of knowledge tracing cited from [9]

5. Experiment

The effect of having previously accepted systematic pronunciation tutoring should be excluded [11]. Therefore, dozens of deaf people with a sound foundation who had never used a pronunciation correction app were invited to participate in the experiment.

First, all students will be given the same word reading test before the experiment, which will test their accuracy of pronunciation of these word consonants through the ASR. Then, the students were equally divided into three groups, one receiving artificial pronunciation correction, one using the consonant correction app, and another group to the control condition. All students received 30 minutes of consonant correction daily without additional tutoring and students using the app were not allowed to see the paper materials used by students who received manual tutoring. After a period of education, all students were tested again for pronunciation of vocabulary consonants.

By above-mentioned evaluation, for the students tutored by the app, the comparison of accuracy rates before and after education can indicate whether the app is useful for consonant correction. In addition, the teaching efficiency of the app is evaluated by comparing with the growth of correct rate of students receiving manual teaching. Through these two evaluations approach, one of which is intuitive and one is objective, we can have a appraise of tutor's teaching efficiency.

6. Conclusion

In order to help the deaf students to adjust their pronunciation, we design an app that can help them to do it. In this paper, we design and introduce the basic model of app, experiment setting and data mining. It can be seen that this project has feasibility and practicability, deserving to be put in to practice.

References

[1] C. Y. Hin, A. Yu On Lam and A. Wong Yiu Leung, "Translanguaging in Hong Kong Deaf Signers: Translating Meaning from Written Chinese," Sign Language Studies, vol. 22, (3), pp. 430-483, 2022.

[2] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, et al., "Deep Neural Networks for Acoustic Modeling in Speech Recognition", IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 82-97, 2012.

[3] T. N. Sainath, O. Vinyals, A. Senior and H. Sak, "Convolutional, Long Short-Term Memory, fully connected Deep Neural Networks," *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4580-4584, doi: 10.1109/ICASSP.2015.7178838.

[4] T. N. Sainath, A. Mohamed, B. Kingsbury and B. Ramabhad-ran, "Deep Convolutional Neural Networks for LVCSR", Proc. ICASSP, 2013.

[5] H. Sak, A. Senior and F. Beaufays, "Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling", Proc. Interspeech, 2014.

[6] Chen, P. H., Lin, C. J., & Schölkopf, B. (2005). A tutorial on ν‐support vector machines. Applied Stochastic Models in Business and Industry, 21(2), 111-136.

[7] Tam, Yik-Cheung & Mostow, Jack & Beck, Joseph & Banerjee, Satanjeev. (2003). Training a confidence measure for a reading tutor that listens. 10.21437/Eurospeech.2003-790.

[8] J. Mostow, S. Roth, A. G. Hauptmann, and M. Kane, “A Prototype Reading Coach that Listens [AAAI-94 Outstanding Paper Award],” in Proceedings of the Twelfth National Conference on Artificial Intelligence, pp. 785–792, 1994. Seattle, WA: American Association for Artificial Intellige.

[9] J. MOSTOW and J. BECK, "Some useful tactics to modify, map and mine data from intelligent tutors," Natural Language Engineering, vol. 12, (2), pp. 195-208, 2006.

[10] Beck, Joseph E., et al. "Does help help? Introducing the Bayesian Evaluation and Assessment methodology." International conference on intelligent tutoring systems. Springer, Berlin, Heidelberg, 2008

[11] Mostow, Jack, et al. "Evaluation of an automated Reading Tutor that listens: Comparison to human tutoring and classroom instruction." Journal of Educational Computing Research 29.1 (2003): 61-117.