Spoken English Repetitive Correction Retrieval Based on Syllable Unit WFST Web Search Filter

Shuang Yu*
Nanyang Institute of Technology, Nanyang, Henan, China, 473000

*Corresponding author e-mail: wanqiu0122@163.com

Abstract. Aiming at the difficulty in computer-assisted language teaching of repeated correction and error detection in spoken English, a fault-tolerant alignment and search filtering algorithm based on the syllable unit WFST (weighted finitestate transformer) network is proposed. This algorithm uses the script corresponding to the adjacent matching words in the recognition results to establish the above grammatical network for fault-tolerant alignment under secondary recognition. The candidate modified parts and replacement parts are used as the query and template for search filtering. In the end, the result of repeated correction error detection is determined by the confidence of the search filtering algorithm. To this end, the k-difference algorithm based on sequential assumptions and the n-gram algorithm based on random assumptions are proposed. Experiments show that the multi-n-gram hybrid search filtering with syllables as the modeling unit achieves relatively optimal results without using quadratic fault-tolerant alignment; when using quadratic fault-tolerant alignment, F-measure can obtain 3 to 4 further increase in percentage.

Keywords: Syllable Unit, Spoken Repetitive Correction Error Detection, Filtering Algorithm

1. Introduction

With the rapid development of speech recognition technology in recent years, computer-assisted language teaching has achieved unprecedented development in both technical and application levels. The work of automatically assessing the spoken language fluency of spoken learners has been extensively studied. The study mainly starts from the indicators of fluency in subjective perception, trains it with artificial scores, and constructs nonlinear regression such as multivariate linear or neural networks model.[1] Such as Cucchiarini'S early research in this field, and the TOFELiBT system currently used in the ETS exam.

Fluent error detection is a method that is closer to the application. It refers to the use of speech
recognition technology to automatically detect the location, duration, and type of errors of the learner's non-fluent parts, so as to provide fast and accurate feedback. Generally speaking, fluency can be divided into error pauses, repetitive corrections, and tone of insertion, etc. This article only focuses on repetitive corrections and error detection. The current research on repeated corrections is mostly based on the definition given by Sherberg.\cite{2} Meaning is divided into 3 parts: the modified part (reparandum), Insertion (reporting), replacement (repair), the first two parts are separated by IP (interrupting point). As shown in Figure 1. The purpose of this article is to use recognition technology to automatically detect modified parts and provide learners with fast and accurate feedback.

\[\text{IP} \quad \text{We need two tickets, I no, three tickets to Boston.}\]

\[\text{Modified parenthesis} \quad \text{Replacement part}\]

\textbf{Figure 1.} Repeatedly correct the composition of non-fluent sentences

2. Grammatical network based on syllable modeling

The search filtering algorithm of Section is based on a single recognition. Although the preprocessing part also does the conversion of words into syllables and phonemes, the algorithm always uses words as the unit during the decoding process, which will cause false negatives for the unfluent parts caused by incomplete words.\cite{3} Although the establishment of a ternary language model with syllables as a unit can theoretically simulate the phenomenon of incomplete words caused by repeated corrections by the speaker in pronunciation, it also has the following disadvantages:

1) The search space becomes larger and runs Slow down;

2) As the complexity of the language model increases, the generated word graphs become more dense. During dynamic matching, the pronunciation of some normal complete words will be replaced by different syllables, resulting in mismatches and false positives;

3) and false positives. Compared with the reported rate, the false negative rate is higher.

When the word map is too dense, some unsound pronunciations that originally belong to incomplete words will be aligned to a certain syllable in the correct complete word pronunciation during matching. \cite{4} To this end, this paper proposes a two-time fault-tolerant alignment algorithm based on the syllable unit WST network: based on one-pass recognition at the word level, by establishing a syllabic unit WFS network with jumps, the possible unfluent phenomenon is simulated to solve the problem of fault-tolerant alignment of fluent speech under the premise of limited script. Usually in the speech recognition process of children's speech or speech-impaired people, there will be a problem of recognition errors. In this regard, the algorithm does not focus on errors, but first focuses on matching the correct words. In general, when the readers pronounce correctly, due to strong language model constraints, the possibility of false rejection is small, and between these matching words, it is usually the part of the speech segment with a high probability of repeated modification.

Assume that when repeated corrections occur, "Bay" is not performed in order. This paper proposes an-gram search algorithm. The algorithm is also retrieved by music melody the algorithm evolved. In
this question, transform the melody factor into fluency factor, how many votes get according to the
tuple of different orders (gram) Row weighted accumulation.

Figure 2."I have choose a color" into sequential syllable WFST format

When measuring system performance, it needs to be compared with the repeated correction
phenomenon of manual detection. In the reading question type, because the reading content is limited,
the position of repeated corrections can be directly marked in the script, that is, if fluency occurs
between two words, use <m, the start time of the modified part, the mark of duration>Mark some of
the first words before changing them. Therefore, finding the first part of the replacement part is helpful
for comparison with the manually labeled results. For the above syllable-based detection, when it is
determined that a repeated correction phenomenon occurs in step 6, the syllables in the corresponding
result replacement part also need to be mapped back to their feet.

3. Data sets and experimental results

The experimental data used 13 to 16-year-old students' SETCS speaking test speech. Speakers'
mastery of English varies. All voices are recorded with standard ear pulses, and the voice
signal-to-noise ratio is above 12dB. The test is divided into 16 questions, each of which is a essay of
about 200 words. First, students are given 60 seconds to familiarize themselves with the essay, and
then read aloud for 90 seconds.

The data set is divided into training and development sets with 500 voices each, and the test set
contains 1,300 voices, one for each student. In terms of evaluation, its scores were marked as
described in the article; the error detection was more difficult to mark, mainly due to the unification
and disambiguation of the standard, the operability of the marking software, manual and Consistency
measurement of machine performance, etc. In fluent error detection, students are required to reach a
certain level in order to make the diagnosis itself practical. In the experiment, 1040 pieces of data with
a total score greater than 3.4 were selected for labeling.

| Testing standards | System Configuration                  | The detection rate/% | Accuracy/% | F value | Question s relativity |
|-------------------|--------------------------------------|----------------------|------------|---------|-----------------------|
| Comparison of machine and standard | Recognition + phoneme | 43.8 | 54.7 | 48.7 | 0.695 |
|                    | Recognition + music                  | 45.5 | 56.8 | 50.6 | 0.702 |
|                    | Quadratic fault-tolerant alignment+music | 44.2 | 67.6 | 53.3 | 0.725 |
| Comparison of people and standards   | 83.0 | 80.0 | 81.4 | 0.841 |

The final labeling result as table 1 reflected is that one voice is labeled by 2 experts, and the
arbitration modification result of the third expert is used as the final standard answer. In this data set, a
total of 3,507 unscrupulous points were marked, of which 1,364 were repeated correction errors, 1,560 errors were paused, and 483 were other inserted garbage. The repeated correction error is selected as the research object to test the detection performance of this system.

4. Conclusions
The grammar network and search filtering algorithm in the diagnosis of fluent areas based on children's English speech recognition are studied. This paper proposes an algorithm for establishing secondary fault-tolerant alignment based on syllable unit WWF network after one pass recognition; and a series of search filtering algorithms to improve accuracy. The experimental results on large-scale student test data show that the two-pass recognition using syllables as the modeling unit and the hybrid n-gram search strategy can achieve the best performance. The detection rate is 44.2% and the accuracy rate is 67.6%. How to improve the positioning accuracy, apply more rule-driven and fault-tolerant alignment techniques are the focus of future research.

References
[1] Cucchiarini C, Strik H, Boves L. Quantitative assessment of second language learners’ fluency by means of automatic speech recognition technology[J]. The Journal of the Acoustical Society of America, 2000, 107(2):989.
[2] Zechner K, Higgins D, Xi X, et al. Automatic scoring of non-native spontaneous speech in tests of spoken English[J]. Speech Communication, 2009, 51(10):883-895.
[3] Marx A W, Sisk T D. Democratization in South Africa: The Elusive Social Contract[J]. Political Science Quarterly, 1996, 110(3):493.
[4] Vergyri D, Stolcke A, Gadde V R R, et al. Prosodic knowledge sources for automatic speech recognition[C]// IEEE International Conference on Acoustics. IEEE, 2003.
[5] Johnson M, Charniak E. [Association for Computational Linguistics the 42nd Annual Meeting - Barcelona, Spain (2004.07.21-2004.07.26)] Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, - ACL '04 - A TAG-based noisy channel model of speech repairs[C]// Meeting of the Association for Computational Linguistics. DBLP, 2004:33-es.
[6] Mohri M, Pereira F, Riley M. Weighted finite-state transducers in speech recognition[J]. Computer Speech and Language, 2002, 16(1):69-88.