1.Introduction

Testing oral proficiency is an important part for the ESL (English as Second Language) learners. Most of the spoken language analyzers are developed to measure this high-level skill. This report compared 3 analyzers developed by different research group with the same data source.

Our final analysis addressed such aspects as:

- Descriptive statistics of different analyzers' results
- Comparison of audio quality assessment
- Comparison of accurate record
- Comparison of total score

Data

Material presented in this report is based on information gathered from K-12 students' speaking practice under ALO7 learning environment. This dataset, in general, consists of 12557 students reading records and evaluation specification. The following items are included in the result:

- Score of native expert：7 native experts rate one audio on the established standards. Score range from 0 to 4. Mean score is employed in this analysis.
- Result specification of **Carnegie Speech Audio Analyzer**（is called **CS**）：recordings are evaluated base on ReadingFluency. Score range from 0 to 1.
- Result specification of **A Audio Analyzer** (is called **A**)：recordings are evaluated base on SentenceType. Scores range from 0 to 100.
- Score of **B Audio Analyzer** (is called **B**)：recording are evaluated base on SentenceType. Scores range from 0 to 5.

> A Audio Analyzer is developed by an artificial intelligence technology service provider specializing in K12 education.
>
> B Audio Analyzer is developed by national key software enterprise dedicated to the research of intelligent speech and language technologies.

2.Method & Result

2.1 Descriptive statistics of CS

Several different types of analysis can be performed with CS. In this context, ReadingFluency analyses was adopted to evaluate reading proficiency along several dimensions. Result specification consists total score, audio quality and other details.

2.2.1 Features

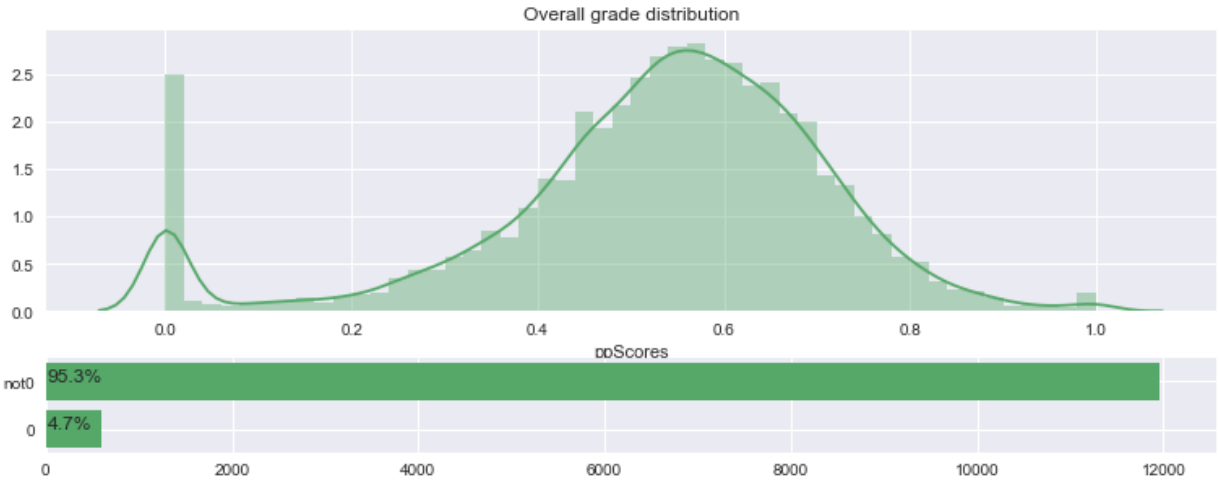
Result specification returned by CS audio analyzer include the number of words correctly spoken (Corr), the errors as per minimum edit-distance alignment (Err), as well as three kinds of errors——Substitution (referred to as "S"), Deletion (referred to as "D") , Insertion (referred to as "I"). Subsequently, we converted these features to numerical ratio. Following variables were identified:

- Total score: ranging from 0 to 1
- Correct ratio：Corr / (Corr + Err) * 100%，ranging from 0 to 1
- Error ratio：Err / (Corr + Err) * 100%，ranging from 0 to 1
    - Substitution ratio：S / (Corr + Err) * 100%，ranging from 0 to 1
    - Deletion ratio：D / (Corr + Err) * 100%，ranging from 0 to 1
    - Insertion ratio：I / (Corr + Err) * 100%，ranging from 0 to 1
- WordAvgScores：word average score.
- Volume：the volume (loudness) of the recording as a number between 0 and 1. Somewhere around 0.5 is ideal. however, [0.2—0.9] is tolerated.
- Noisy：ranging from 0 to 132
- AQstatus：a hexadecimal value encoding all the audio quality problems. In this dataset, 6 types lable information are presented:
    - 0: normal
    - 2: excessive background noise
    - 8: speech truncated at the end
    - 10: recording too loud
    - 12: recording too loud & excessive background noise
    - 18: recording too loud & speech truncated at the end
- Mode：there are two modes, one is for 'child', one is for 'Adult'. If this parameter is not in the original request, CS analyzer will recognize the sound source automatically.

2.2.2 Preliminary results

- **Total score**

  On the whole, most of the recording were graded 0, accounting for about 4.7% of the total population. The distribution of scores shows the characteristics of Normal / Gaussian distribution. Scores are mainly concentrated within [0.5—0.7].

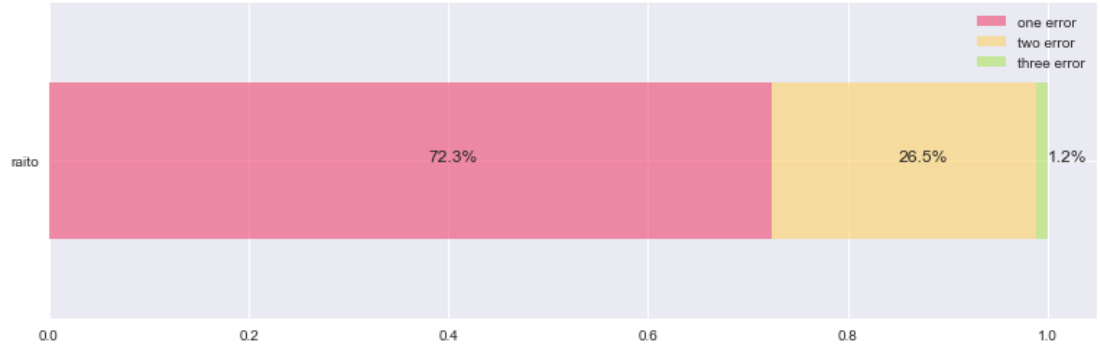


- **Details**

  1. Completely correct records accounted for 64.6% of the total records, 4.7% for completely incorrect records, and 30.6% for partially correct records
  2. The distribution of word average score is similar to the total score
  3. The probability of 3 kinds of errors：

     ```
     P(Deletion) > P(Substitution) > P(Insertion)
     ```

  4. The number of errors in one sentence：


     ```
     one error：72.3%
     two errors：26.5%
     three error：1.2%
     ```

- **Audio quality**

  Normal recording accounts for 60.1%, while the abnormal accounts for 39.9%. Excessive background noise, loud audio volume and truncated speech are the dominant factors influencing audio quality.

2.2.2 Q & A

- **Why there are so many 0 point records? What are the factors? What are the differences between 0 point record and non-zero record?**

  Comparing 0 point records with non-zero records:

```
1. Volume of most 0 point records are below 0.2, which is not acceptable
2. Speech truncated and recording too loud are frequently occuring status
among 0 point records
3. Nearly all 0 point records are labeled with 'Adult' mode.
```

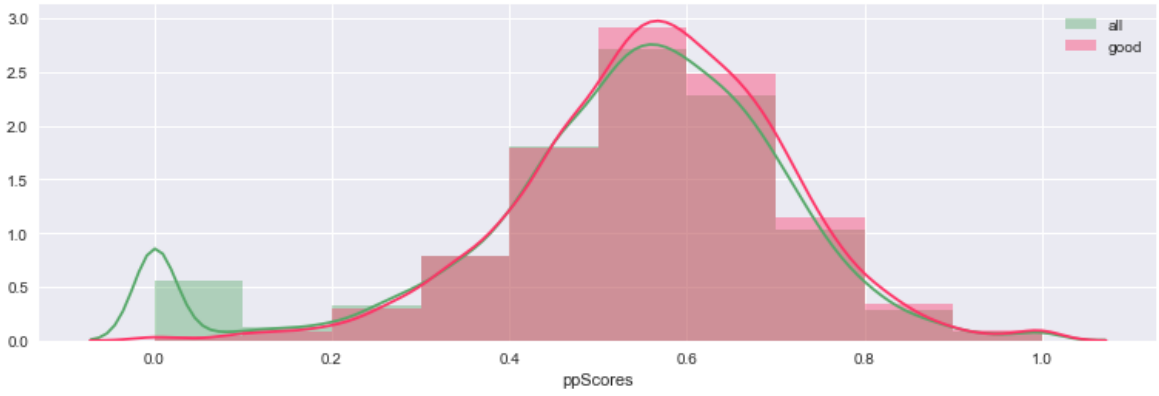
Consequently, most 0 point records exhibits poor audio quality. Meanwhile, there is the possibility that because most records are recognized as 'Adult' mode, which might be the reason for lower scores.

- **How to distinguish the accurate evaluation result from the whole records?**

  According to the audio quality status, all records are divided into 2 classes:

```
Accurate (good) : AQStatus == 0, account for 60.1%;
Inaccurate (bad) : AQStatus ! = 0, account for 39.9%;
```

  Excluding the poor quality audio, total score distribution represented via the red area:




```
It was found that the accurate part still has some 0 point records, but has
reduced considerablely after the filtering. Proportion of high scores has
increased.
```

2.2 Descriptive statistics of A

In the face of different scenarios, similar to the CS, AA put forward different kinds of service. Here SentenceType evaluation were deployed. Result specification consists of total score, details of pronunciation——Integrity, Accuracy, Fluency and Rhythm, as well as information related to the audio quality——signal-noise ratio, volume, status.

2.2.1 Features

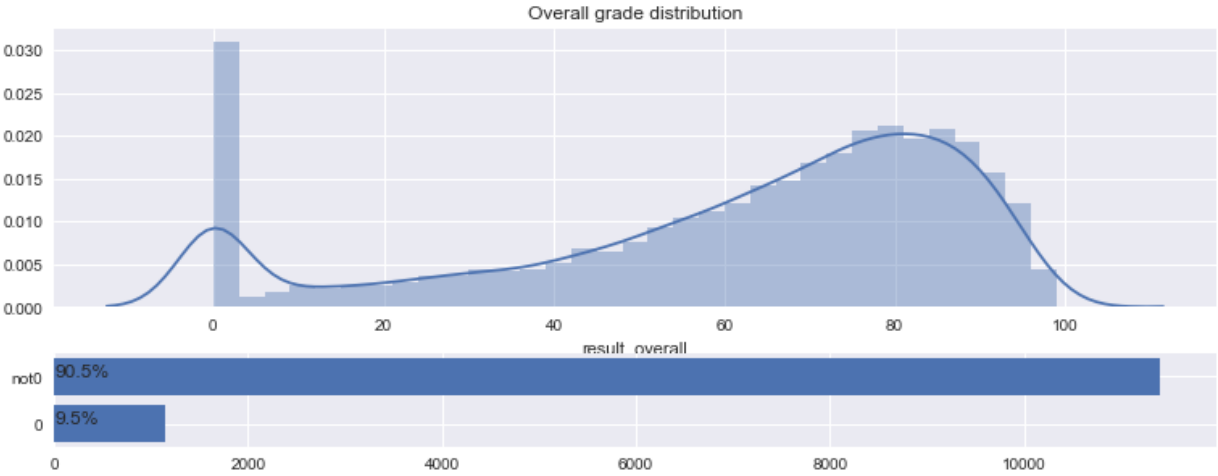
- Total score：ranging from 0 to 100
- Integrity：ranging from 0 to 100
- Accuracy：ranging from 0 to 100
- Fluency：ranging from 0 to 100
- Rhythm：ranging from 0 to 100

- SNR：signal-noise ratio, ranging from 0 to 40. The greater the value, the clearer the recording.
- Volume：ranging from 0 to 180
- TipID：5 types lable information are presented to show audio quality:
  - 0：normal
  - 10004：recording too soft
  - 10005：recording too loud
  - 10006：excessive background noise
  - 10008：truncated at the amplitude

2.2.2 Preliminary results

- **Total score**

  On the whole, most of the recordings are graded 0, accounting for about 9.5% of the total population, non-zero records account for 90.5%. Distribution is skewed to the right, scores are mainly concentrated within [0.5—0.7].



- **Details**

  1. The distribution of integrity, accuracy, and fluency is similar to the distribution of total scores, showing a right-skewed pattern. The final score can be approximated as (excluded the rhythm):

```
Total score = Integrity * Accuracy * 0.8 + Integrity * Fluency * 0.2
```

  2. Harshness / leniency of the score：

```
Integrity > Accuracy > Fluency > Rhythm
```

- **Audio quality**

  As the audio status (tipID) shows, normal audio accounts for 60%, the remaining 40% is affected by the sound quality. High noise and low volume are the main factors affecting the final score.

2.2.3 Q & A

- **Why there is so many 0 point records? What are the factors? What are the differences between 0 point record and non-zero record?**

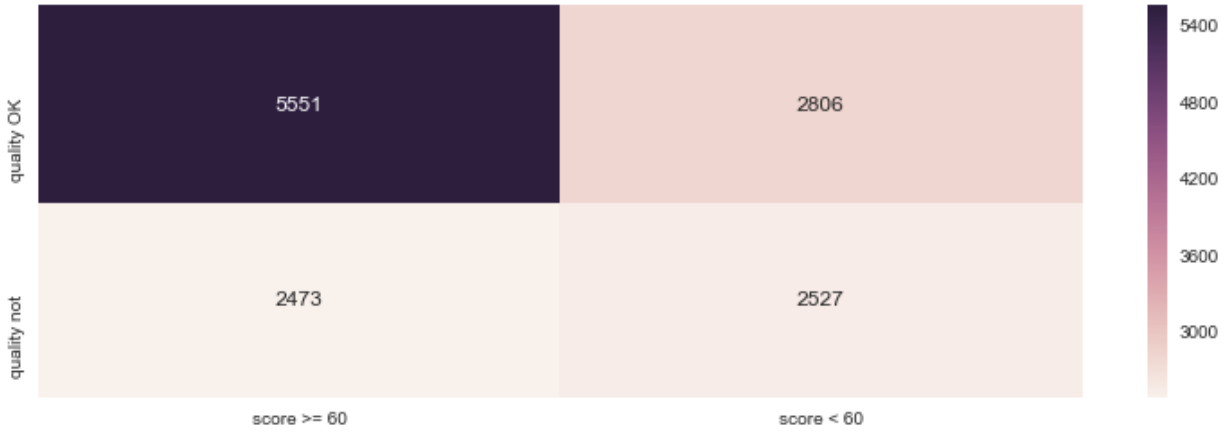
  Comparing 0 point records with non-zero records:

  ```
  1. Most of the 0 point records are also 0 on the evaluation of Integrity, Accuracy and fluency
  2. Nearly all 0 point records show low signal-noise ratio
  3. Most of the 0 point records show low volume
  ```

  Therefore, most of the 0 point records are resulted from poor quality. However, it is not ruled out that there are some records are "true zero", that is, user's incorrect reading lead to the low score, not audio quality.

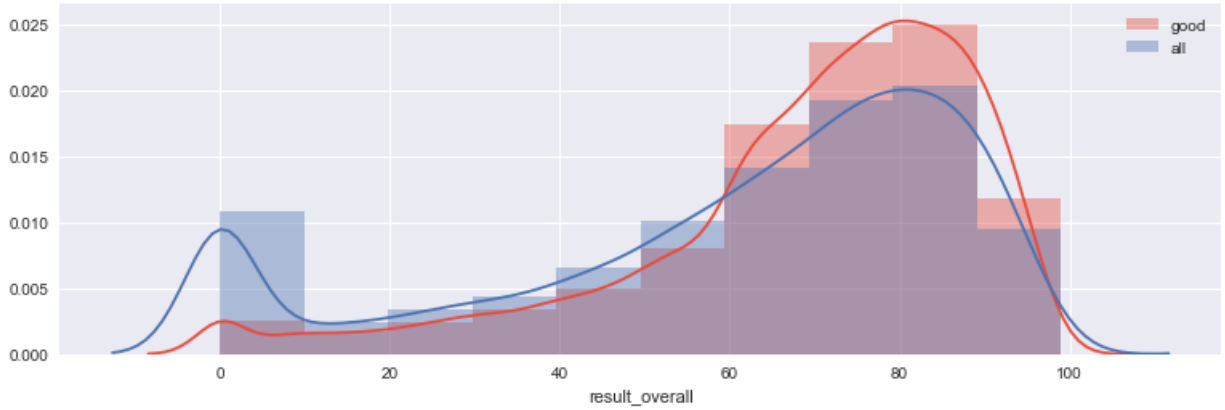
- **How to distinguish the accurate evaluation result?**

  Based on the suggestion by A analyzer, we reclassify all records based on threshold—60 points.



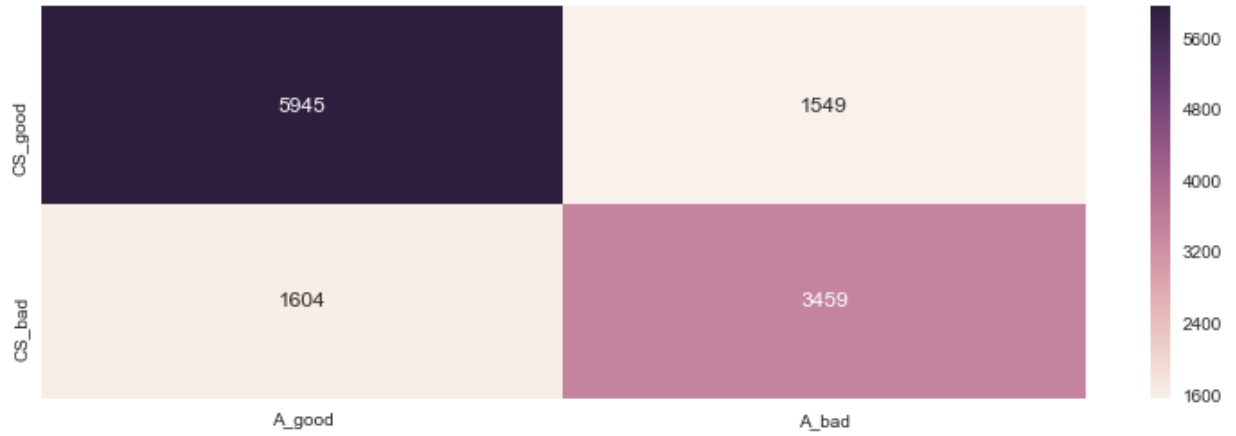

```
Accurate (good) : normal audio, and record that the total score >= 60 points (which may be not affected by sound quality) , account for 81.1%;
Inaccurate (bad) : abnormal audio, and record that the total score < 60 points, account for 18.9%
```

Excluding the inaccurate records, total score distribuion is as represented by red area:



```
It can be found that the accurate part still has some 0 point records, but has reduced a lot after filtering. Proportion of high scores has increased.
```

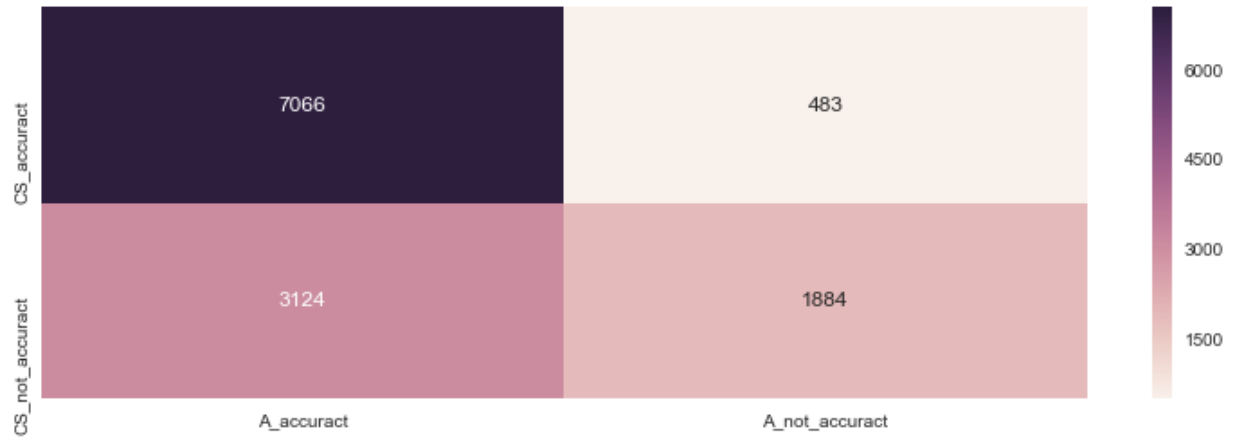
2.3 CS & A—Comparison of audio quality assessment



The consistency of sound quality evaluation is 74.89%.

In the evaluation of CS, poor audio quality record accounts for 40.32%； while in A, it accounts for 39.88%. CS is slightly strict in the evaluation of sound quality.

2.4 CS & A—Comparison of accurate record



The consistency of accurate division is 71.27%.

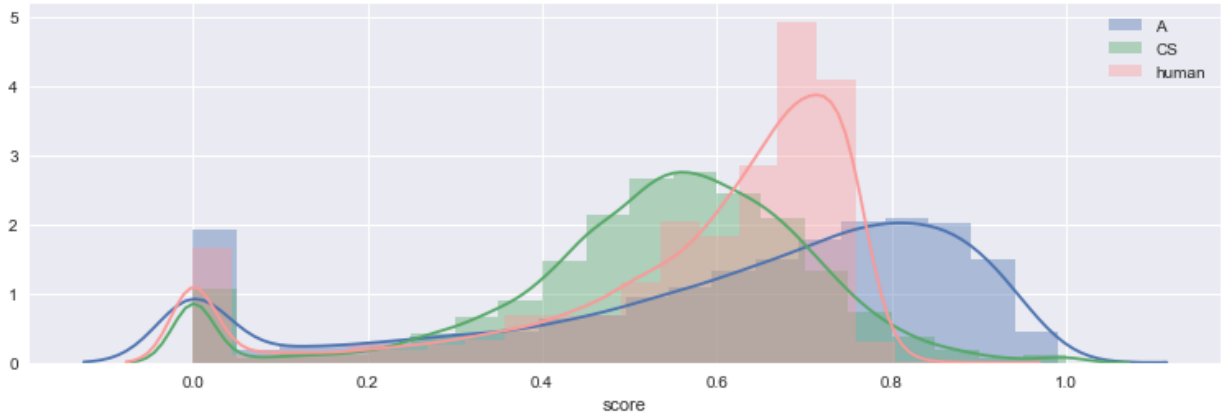
Accurate record in CS accounts for 60.12%, and 81.15% in CS. Besides, 24.88% is accurate in A but inaccurate in CS.

2.5 Comparison of total score

In this section, total score of B alalyzer is added. Discussion revolves around the following aspects：

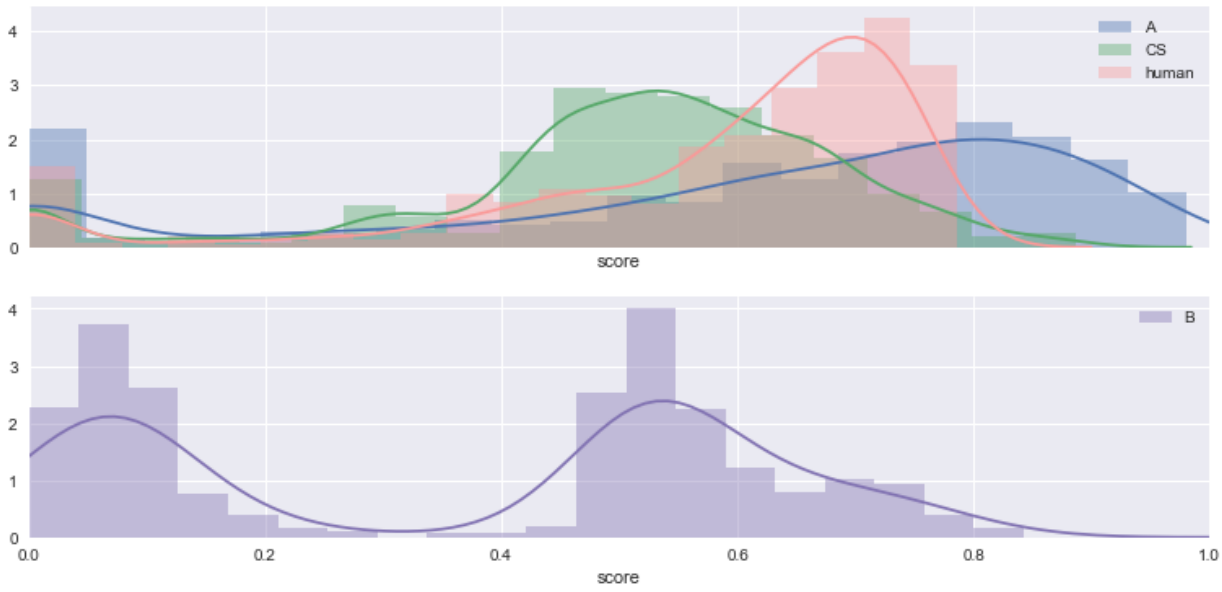
- **Total score distribution**
- **Which analyzer is most accurate**

As scores from different analyzers varies in range, we conformed all intervals to [0—1].



However, as these analyzers are based on different evaluation criterions, they should be compared under the same pattern. Consequently, we took expert score as the standard, and made some transformation on the score of CS analyzer and A analyzer.

The process can be simplified as finding a map function between analyzer and expert evaluation. Therefore, we employed symbolic regression to identify an underlying mathematical expression that best describes their relationship. Mean Squared Error (MSE) was adopted to measure its deviation.



The picture above shows the total score distribution of 1699 records, which is similar to the whole dataset. Meanwhile, it can be found that CS and B are similar in range [0.4—1].

MSE between expert score and analyzers (the smaller the value, the higher the similarity):

```
CS—human:  0.035
A—human: 0.037
B—human: 0.147
```

MSE of the analyzers:

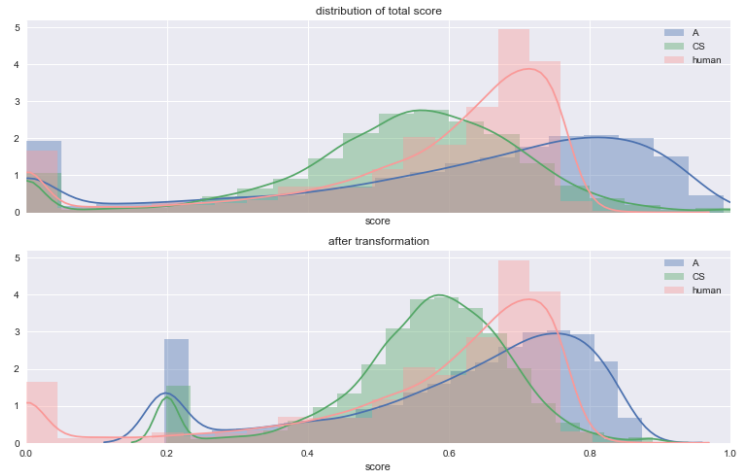
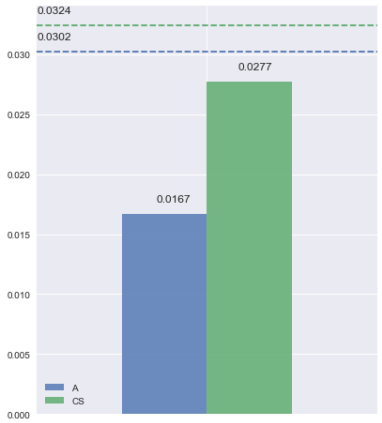
```
A—B: 0.216
CS—B: 0.125
```

2.5.1 Full data

- Transform function

```
F = 0.196 + A * 0.685
G = 0.199 + CS * 0.691
```

- Changes of distribution and MSE

Changes of distribution	Changes of MSE
	

After transformation, range of CS is [0.199—0.89], range of A is [0.196—0.874]

- In general, no matter whether transformed or not, A is more similar to the expert score (smaller MSE). New map function reduced MSE effectively.
- The range of score was narrowed after the transformation.
- The decline of MSE in CS and A are mainly caused by the accurate records.

2.5.2 Accurate record

- Transform function

```
F = 0.247 + A * 0.531
G = 0.378 + CS * 0.398
```

- Changes of distribution and MSE

Changes of distribution	Changes of MSE
	

After transformation, range of CS is [0.378—0.776], range of A is [0.247—0.773]

- In general, no matter whether it is transformed or not, A is more similar to the expert score (smaller MSE). New map function makes the MSE declined effectively.

3.Conclusion

3.1 Analysis type

In CS analyzer, ReadingFluency was chosen for measuring user reading proficiency, the object to be analyzed can be a sentence or a paragraph. Meanwhile, most of the ALO7's recording are short sentences, which may lead to low score.

In contrast, the evaluation type of A analyzer can be specified down to word, sentence, or paragraph, which is more credible.

3.2 Analysis mode

In CS analyzer, there are optional evaluation modes——'Child' and 'Adult'. In the analysis request, we did not assign any parameter. So it will recognize the sound source automatically. From the result, most records that were recognized as 'Adult' shows more 0 point on total score. This portion of records ('Adult' mode) account for 40%, which may be too high. In addition, some recordings are labeled as 'Adult' mode, but finally confirmed as child.

3.3 Analysis detail

In CS, oral proficiency can be reflected by reading fluency, which is a multidimensional concept, several types of error can lead to disfluency, such as mispronouncing, skipping, strong accent and etc. Referring to this, the details are produced by minimum edit-distance or Levenshtein distance algorithm when comparing the target text and the ASR output.

For A analyzer, the details refer to integrity, accuracy, fluency and rhythm. And the total score can be approximated as: integrity * accuracy * 0.8 + integrity * fluency * 0.2

3.4 Audio quality

In terms of audio quality, CS analyzer is slightly stricter than A, but the difference is not significant.

3.5 Total score

Due to the different evaluation standard, we make some transformation of CS and A anlayzer's total score. Via symbolic regression, a more suitable mapping relationship was screened out and result showed A analyzer is closer to expert rating than CS.