A subjective study on the English speech intelligibility in Chinese classroom with machine learning algorithm

Xuhao Du

ABSTARCT

Speech intelligibility in classrooms affects the learning efficiency of students directly, especially for second language students. In previous research, the speech intelligibility value is determined by several major parameters like speech level, signal to noise ratio, and reverberation time in the rooms. This paper investigates the importance of each factors in determining the speech intelligibility with subjective tests and advance machine learning algorithms. Four major algorithms were used to analysis the data including the traditional average method, artificial neural network, support vector machines, random forest and decision tree algorithms. By doing the subjective experiment of 60 volunteers and processing the data with machine learning algorithms, it is found that the most important factor is signal to noise ratio and then the sound pressure level. Different factors take effect under different acoustics condition. The gender is also a factor that influence the speech intelligibility. By comparing the importance of each factors, the combination of sound pressure level and signal to noise ratio has better prediction performance than the combination of background noise level and signal to noise ratio. Among all the algorithm, the Random Forest do the best job in predicting the speech intelligibility.

Key words: speech intelligibility; machine learning; speech level; signal to noise ratio; reverberation time

1. Introduction

Speech intelligibility (SI) is defined as the measure of the comprehensible quality of speech, which can be evaluated by the percentage of correctly understood words or sentences in a specific list under controlled conditions. SI can be used to quantify the speech perception in classrooms, which directly affects the learning efficiency of students, especially younger students and the students who are working on a second language. Poor perception in classrooms has been reported to cause perceptual and cognitive problems among students. With a number of linguistic and phonetic reasons, the perception results of native and non-native listeners might differ with each other. It has been indicated that SI is affected by a number of acoustic parameters such as signal to noise ratio (SNR), speech level (L_s), and reverberation time (T_{60}).

L_s affects the perception of listeners directly, especially in noisy environment. As early as in 1922, it was found that the syllable articulation decreases with L_s when L_s is relatively high or low. In 1956, it was also found that SI deteriorates with extremely weak or strong vocal forces, and SI decreases when L_s increases from 80 dB regardless to the variance of SNR. In recent studies, SI and listening difficulty ratings were used simultaneously to found the optimum L_s or acceptable range of L_s.

SNR is defined as the difference between L_s and L_{BN}, where the background noise in classrooms is mainly contributed by ventilation systems, students' activities, and noise sources outside classrooms. It has been demonstrated that SI increases with SNR, and the detrimental effect of interfering noise can be eliminated if SNR is larger than 15 dB in classrooms.

T_{60} depends on the room geometry and the amount of absorption in the room. SI is usually negatively related to T_{60}, and the optimal T_{60} for SI should be closed to 0 s. The maximum allowed T_{60} for different kinds of classrooms has been proposed in some standards and acoustic design guides.

In previous work, researchers used traditional method to process the subjective experiment sample like averaging the data, using the least means square to calculate

the empirical regression equation, figure all the data to observe the variation trends to find out the phenomenon of the SI with other classroom acoustics factors and so on. In this work, the advanced data processing algorithms, including the artificial neural network (ANN), support vector machines (SVM), random forest (RF) and decision tree (DT) algorithms will be used in processing the data samples and compared the result with the traditional method. By using the advance algorithm, a more accurate prediction system is developed, the importance of each factors are presented.

2. Methodology

The SI test was carried out with professional headphones (Type: AKG 518). The test material was generated by convoluting the sections of Coordinate Response Measure (CRM) corpus with the measured room impulse responses and then mixing them with the background noise. In the test, the listeners were asked to listen to the test material and answer the questions on what they had heard. After the test was completed, the SI results were obtained by collecting and grading the answer sheets.

2.1. Processing of the test material

The widely used CRM corpus is adopted to generate the test material, which consists of 256 short phrases in English, and each phrase is made up of 3 parts: a signal, a color and a single number. There are 8 English signals (arrow, baron, charlie, eagle, hopper, laker, ringo and tiger), 4 colors (blue, green, red and white) and 8 single numbers (1 to 8) in total. With different combinations, 256 different short phrases were made following the structure: “ready (a signal) go to (a color) (a number)”. These phrases were spoken by 8 native English speakers (4 males and 4 females) and recorded in an anechoic chamber.

The test material was made up of 64 different sections, and each section was constituted by 20 phrases which were randomly chosen from the CRM corpus. There is a 2.5-second time interval between adjacent phrases and a 5-second pause between

adjacent sections. The L_s, the SNR, and the T_{60} of each section were processed to be different from each other, and there were 4 different L_s (50, 60, 70 and 80 dBA), 4 different SNRs (0, 5, 15 and 20 dB) and 4 different reverberation conditions (T_{60} = 0.77, 1.07, 1.29 and 1.52 s) in total. In order to prevent listeners from adapting, the order of the 64 sections was randomly arranged. Since the whole test material lasts about 96 minutes, it was divided into 4 parts with equal lengths. During the listening tests, each volunteer only takes one part. Because the order of the 64 sections was randomly arranged, dividing them into 4 parts did not introduce new errors.

The reverberation effect of each processed section was implemented by convoluting the standard CRM section with the room impulse responses measured in a real room with dimensions 6.0 m × 6.6 m × 4.5 m. As shown in Fig. 1, a loudspeaker was placed at the corner of the room to excite as many modes as possible, and the measurement microphone (Type: BAST 100291) was placed 5.7 m away from the loudspeaker (about 1.5 m, 1.4 m and 3.2 m away from the walls). In the measurements, a B&K Pulse 3560D system generated pink noise to drive the loudspeaker and recorded the data with a sampling frequency of 65536 Hz. Each impulse response was measured three times and averaged, and then the corresponding T_{60} was calculated from them. The T_{60} values used in the tests were the average values of 500 Hz, 1000 Hz and 2000 Hz octaves.

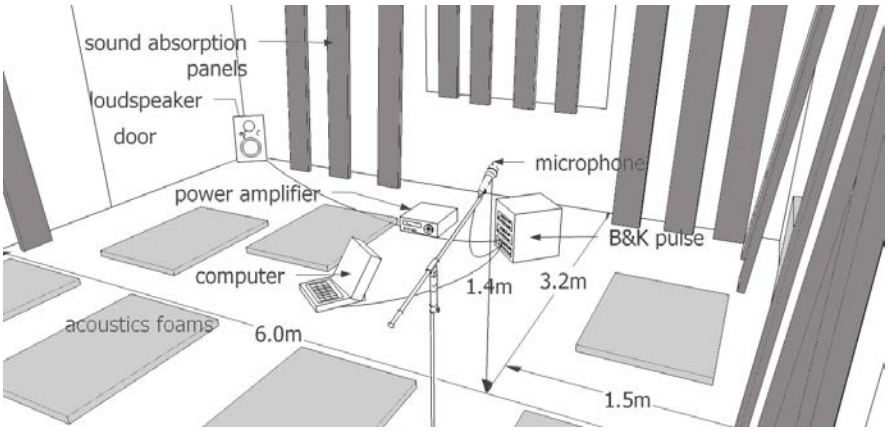


Fig. 1 Measurement setup in a room with sound absorption panels against the walls and acoustics foams on the floor

By changing the amount of sound absorption panels and foams decorated in the room, the impulse response and the corresponding T_{60} of the room were adjusted. According to previous survey, the T_{60} was controlled to vary from 0.77 to 1.52 s in the measurements.

The CRM sections with reverberation effects were then mixed by using the *Adobe Audition* software with the typical noise measured in a classroom at Xianlin Campus of Nanjing University during a class break. The L_s and SNR of the playback sound were adjusted by changing the amplitudes of the input signals of the CRM sections and the typical noise and monitored by using a B&K Head and Torso Simulator (HATS, Type: 4128).

2.2. Listening test

Sixty normal-hearing volunteers with English as the second language (24 females, 36 males, aged between 20 and 24) participated the listening test. The 60 volunteers were randomly divided into 4 groups to listen to the 4 parts of the test material, respectively, so there were 15 complete samples after all volunteers completed the test and 991 individual sample sets.

To help the volunteers to get familiar with the test procedure, they were asked to listen to a standard section of CRM corpus before the formal test. This standard section doesn't contain any background noise or reverberation information and the sound pressure level is adjusted to 70 dBA during the playback. During the test, the volunteers were asked to choose the color and the number that they had heard. The L_{BN} in the test room was lower than 35 dBA, which was at least 15 dB lower than the playback sound pressure level. So the background noise effect of the playback environment can be neglected. After the volunteers completed the test, the answer sheets were collected and graded.

3. Data analysis and discussion

3.1. Modal Comparison

3.1.1 Traditional Average algorithm

In the traditional average algorithm, the percentages of correct answers (SI) of the 15 samples are averaged and presented in Table 1 with the corresponding standard deviations (SD) in Table 2. Compared with the results done on the first language user, the perception accuracy shown in Table 1 is a little lower, while the corresponding SDs in Table 2 are larger. The reason is that the volunteers who participated the test are non-native English listeners, whose perception accuracy might be poorer than native listeners under the same acoustic conditions. Previous studies have also found that the phonological factors such as some features of the first language and the phonetic characteristics of some sound of the second language may significantly reduce the speech intelligibility and increase listening difficulty of non-native listeners.

Table 1 Average percentage of correct answers (%) of the test at different SNR, T_{60}, and L_s

T_{60} (s)	0.77				1.07				1.29				1.52			
SNR (dB) / L_s (dBA)	0	5	10	15	0	5	10	15	0	5	10	15	0	5	10	15
50	73.7	79.0	86.3	86.0	71.0	84.0	86.0	92.3	72.0	72.3	80.0	86.7	76.3	75.3	76.7	73.3
60	80.3	83.7	92.7	90.7	78.0	85.7	92.7	90.7	82.3	84.0	92.3	90.7	78.0	90.7	88.3	87.7
70	79.0	87.7	94.0	96.7	81.3	87.3	92.7	90.3	90.3	84.0	92.7	92.7	72.6	88.0	88.3	93.0
80	75.7	88.3	95.7	89.3	76.0	87.3	93.0	97.0	85.7	81.7	91.3	92.3	79.7	86.0	87.0	92.0

Table 2. Standard deviations (SD) of the average percentage of correct answers (%) at different SNR, T_{60}, and L_s

T_{60} (s)	0.77				1.07				1.29				1.52			
SNR (dB) / L_s (dBA)	0	5	10	15	0	5	10	15	0	5	10	15	0	5	10	15
50	9.9	10.2	9.7	12.4	13.0	12.4	9.8	10.0	14.1	7.5	10.4	10.1	12.2	18.5	16.2	24.8
60	9.9	8.1	5.3	7.5	10.1	10.0	7.3	12.1	10.3	7.8	8.2	10.8	11.6	7.3	9.0	13.5
70	6.6	9.8	6.3	3.6	6.9	7.5	9.0	7.7	7.4	8.5	6.2	6.5	8.6	14.9	13.2	8.6
80	9.06	7.0	5.6	10.7	10.2	7.5	8.0	5.6	8.8	12.6	10.9	9.8	7.7	9.7	11.0	8.8

Based on the average values in Table 1, the best-fit regression equation involving L_s, SNR and T_{60} is obtained by

$$\mathrm{SI} = -17.6 + 1.58\ \mathrm{SNR} - 0.0508\ \mathrm{SNR}^2 + 2.90\ L_s - 0.0204\ L_s^2 - 3.66\ T_{60}\ \ (\%), \quad (1)$$

From Equation (1), it is clear that with constant SNR and T_{60}, the highest SI achieved when L_s reach 71 dBA. The SI decreases with the T_{60} increases and in the SNR range we studied, the best SI achieved when SNR reach 15 dB. The average algorithm gives us a brief SI variation trends with L_s, SNR and T_{60} and other machine learning algorithms show similar pattern with a little different.

3.1.2 Artificial Neural Network

The ANN algorithm is a kind of machine learning algorithm that inspired by the biologic neural action achieved by several hidden and layers computation. By the forward propagation computation, the defined cost function will be minimum by adjusting the parameter between each layer that estimate the real function of the system based on the input and output variables. In this experiment, there are totally 991 samples with 5 inputs and 1 output as the training data set. The input variables including the L_s, SNR, T_{60}, L_{BN}, gender and the output variable is SI. The following algorithms used the same sample data as the training data set. In ANN, 80% of the samples are selected randomly as the training data and 20% of the samples are the test data. With comparing the training data cost and the test data cost, 1 hidden and 24 layers were used in this ANN system. After the computation, Fig. 2 shows the parameters' diagram of the system. The accuracy and the cost of the neural network system is shown in the following part together with other algorithms.

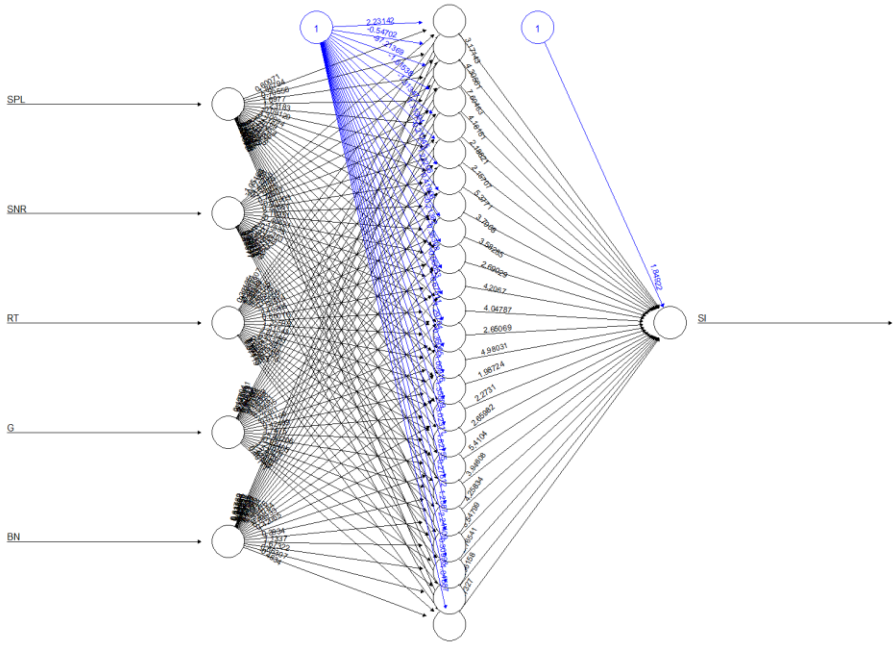


Fig. 2 The neural network system

3.1.3 Support Vector Machine

SVM is a supervised machine learning algorithm used for classification and regression analysis. Given a set of training examples, each marked for belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into certain category. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. In choosing the two major parameters in the SVM, the degrees of freedom and the cost, 80% of the samples are selected as the training data set and 20% are the test data set. The training cost and the test cost under different combinations of the parameters are present in Fig. 3.

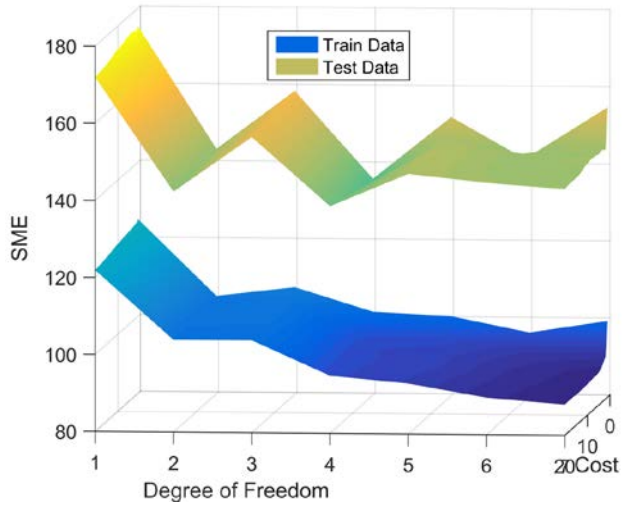


Fig. 3 The cost of the SVM algorithm under different parameters combination

Comparing the cost of the training data and the test data, the SVM parameters can be chose to avoid overfitting or under fitting because of higher of lower degrees of freedom. From Fig. 3, the error of training data decrease with the increase of the degree of freedom while focus on the error of the test data, the error reach its minimum value when degrees of freedom are 2 and 4. So in this simulation, the support vector machine will set as 2 degrees of freedom and 21 cost as cost doesn't make much different in our training data. From the parameter selection, the degrees of freedom is 2, the same as the result using traditional average method, which means the SVM algorithm do fit the former verified result.

3.1.4 Random Forest

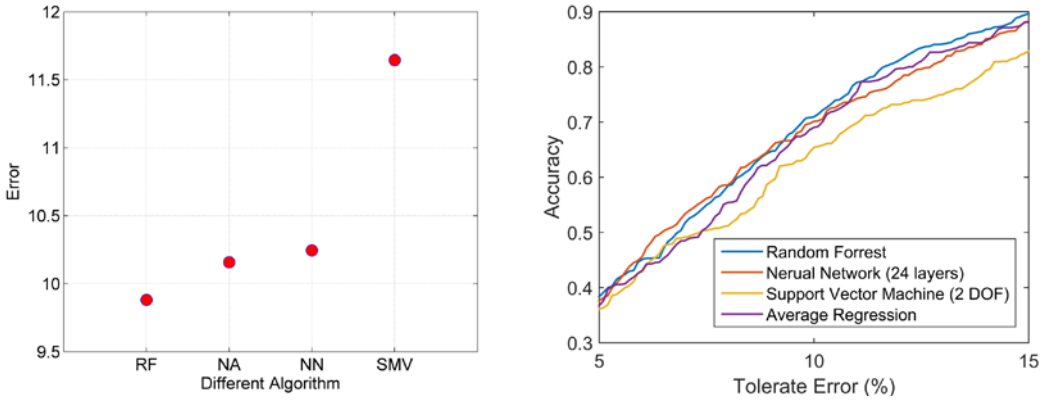
Breiman proposed random forests, which add an additional layer of randomness to bagging. In addition to constructing each tree using a different bootstrap sample of the data, random forests change how the classification or regression trees are constructed. In standard trees, each node is split using the best split among all variables. In a random forest, each node is split using the best among a subset of predictors randomly chosen at that node. This somewhat judger intuitive strategy turns out to perform very well compared to many other classifiers, including

discriminant analysis, support vector machines and neural networks, and is robust against overfitting. In addition, it is very user-friendly in the sense that it has only two parameters (the number of variables in the random subset at each node and the number of trees in the forest), and is usually not very sensitive to their values.

In this study, similar to previous algorithms, 80% of the data is selected as the training data and 20% as the test data. Using the RF algorithm, the importance of each factor to the SI will be shown according to the range of SI change with the factor. The result will be shown in the following section.

3.1.5 Comparison of 4 different algorithms

The accuracy and the errors of 4 different algorithms are presented in Fig. 4. The error means the mean prediction errors of the estimated function. Since the experiment in this work is a subjective test, the result may be variate among person to person, so the accuracy in Fig. 4 (b) is adjusted to be when the error is less than the tolerate error, the prediction is judged to be correct. For example, if the real SI is 60 while the prediction SI is 55, the prediction is judged to be correct when the tolerate error is higher than 5%, otherwise it is judged to be wrong when the tolerate error is lower than 5%. Fig. 4 present the accuracy result of 4 different algorithms.



(a) (b)

Fig. 4 (a) The mean cost error of the training data set (b) the accuracy versus the tolerate error of 4 different algorithms

From Fig. 4 (a), the RF algorithm presents the lowest cost error and the error of the ANN is slightly higher than the traditional average algorithm. The SMV has the highest error among all the algorithms. The accuracy under certain tolerate error presents the similar result with the error comparison that the RF algorithm does the best job among all other algorithms and the SVM perform worst. The traditional average and neural network algorithm show the similar performance. In a word, in the cost error and accuracy comparison, the RF perform best, and then the traditional average and neural network, SVM does the worst job. In the following factors analysis, the result from RF will be studied more carefully.

3.2. Relationship among SI and other factors

3.2.1. The importance of each factors to speech intelligibility

Produced by the RF algorithm, the importance of 5 different variables, L_s, SNR, T_{60}, L_{BN}, gender to SI is studied. The L_{BN} is calculated by subtracting the SNR from L_s. As in two different papers, different combinations are used as the combination of SNR and L_s, L_{BN} and L_s, so by using RF, the importance of L_{BN} and L_s can be compared. The result is shown in Fig. 5. The lower two figures resent the importance of other factors when no L_s or L_{BN} included, respectively.

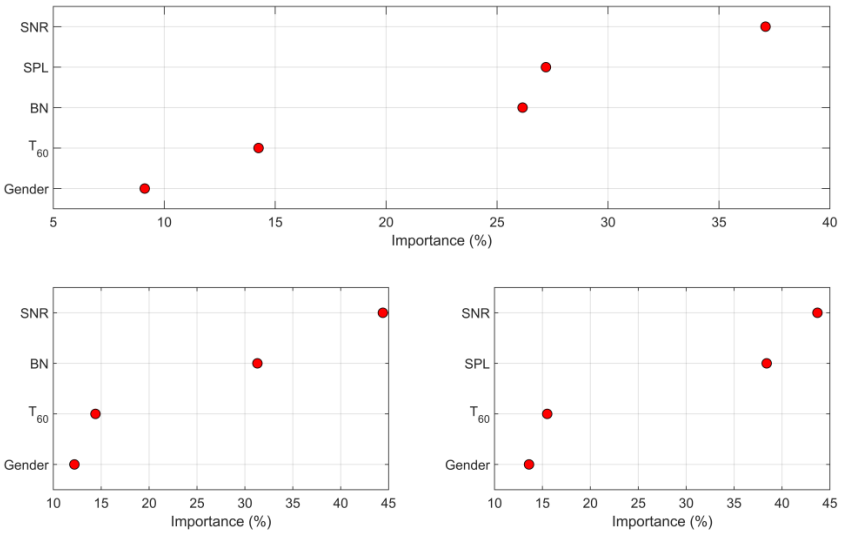


Fig. 5 The importance of each factors

It is clear that no matter in what circumstances, the SNR is the most importance factor among all the variables, and it takes 37.0%, 44.8% and 44.5% in three different variables combination. The second importance factor is the L_s and $L_{BN,}$ 27.5% and 26% when all the factors included, 37% and 31% when without one the other, respectively. No matter in what condition, the importance of L_s is higher than the one of L_{BN}. So the combination of SNR and L_s presents the SI better than the combination of SNR and L_{BN}. And the Third importance factor is T_{60}, around 15% in all situations. The gender effect the SI worst, with only 9%, 12% and 14% respectively. To learn more about the influence of the factor under different situations, the decision tree algorithm is used to analysis the data. The decision tree is presented in Fig. 6.

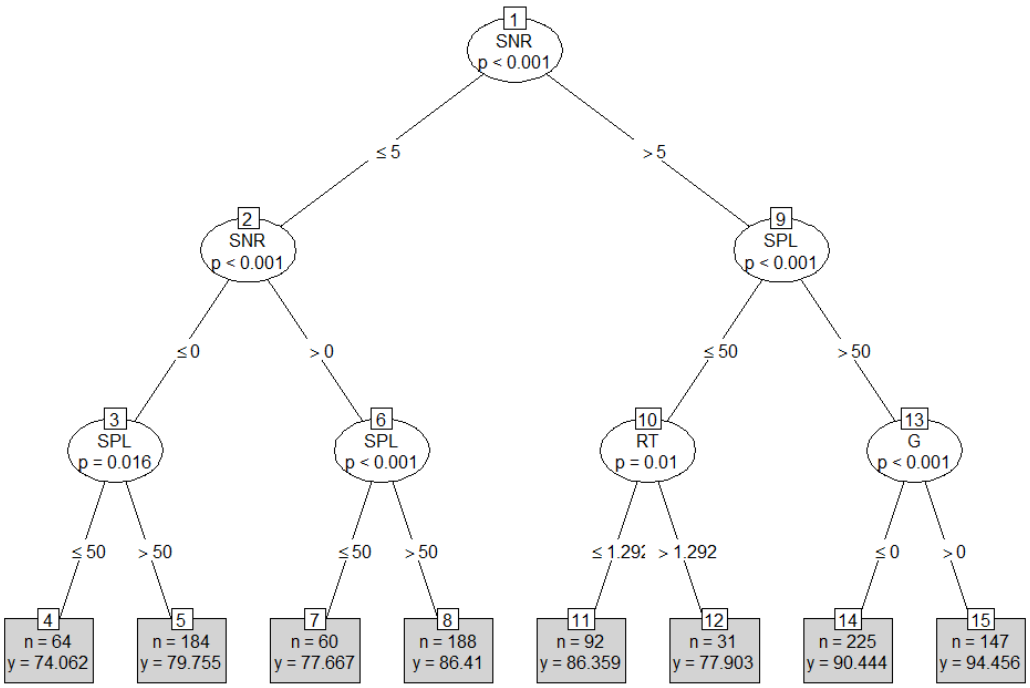


Fig. 6 The decision tree among all the factors

The decision tree draw the similar conclusion with the RF algorithm as the SNR take the most important role in deciding the SI, and the L_s. When the SNR lower than 5 dB, the T_{60} and gender do not take much effect on the SI. When SNR higher than 5 dB, the T_{60} takes effect on SI when L_s lower than 50 dBA and the gender takes effect

on SI when L_s higher than 50 dBA.

3.2.2. Relationship among SI, SNR and T_{60}

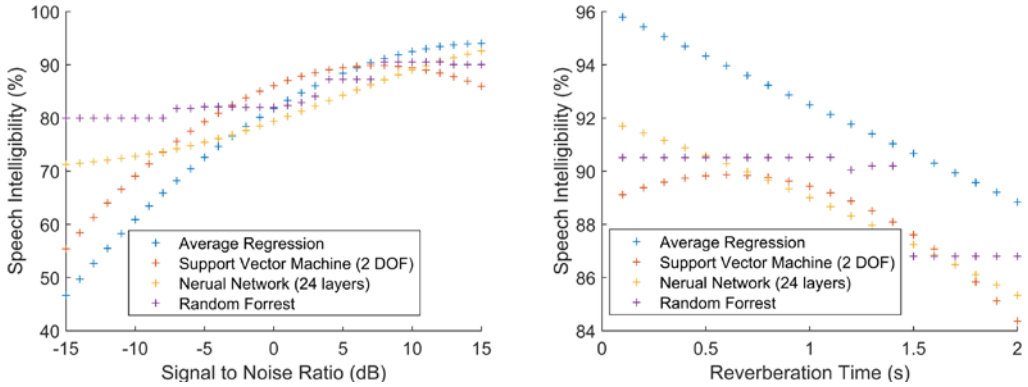


Fig. 7 The variation trends of SI versus (a) SNR and (b) T_{60} under other fixed factor

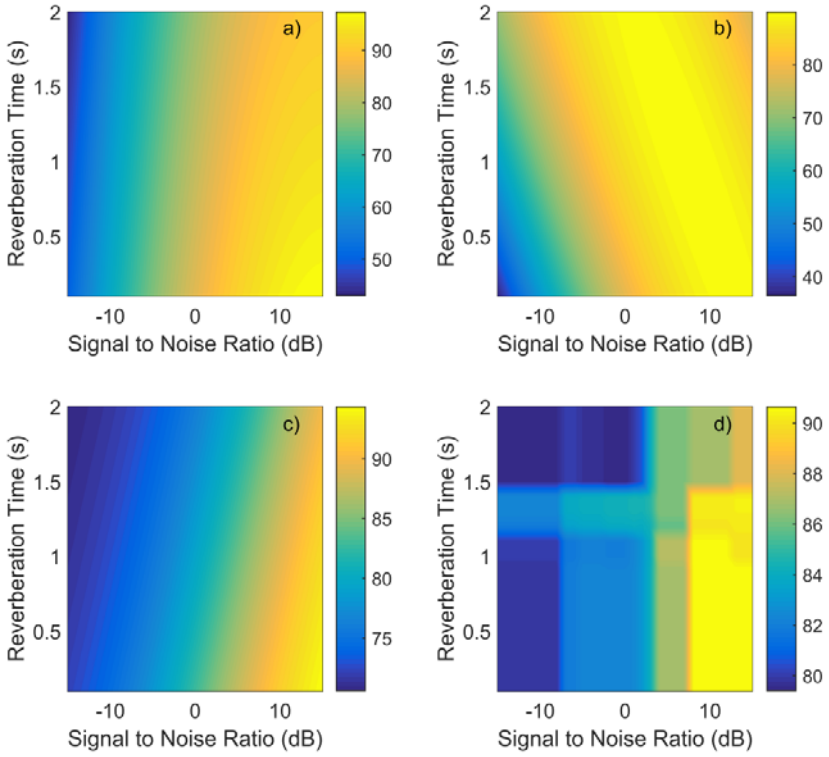


Fig. 8 The SI versus the SNR and T_{60} calculated by (a) traditional average method; (b) support vector machines algorithm; (c) artificial neural network algorithm and (d) the random Forest algorithm.

Based on the different machine learning algorithms, the relationships between the

SI and SNR, T_{60} when Chinese listeners are listening English was illustrated in Fig. 7 (a) and Fig. 8 (a) to (d), where it is clear that SI is positively related to SNR regardless to the influence of T_{60}. In the traditional method, SVM and Random Forest algorithms, the SI rises with a decreasing rate with the increase of SNR, which is in good agreement with previous research of native listeners. The optimal SNR value lies on 15 dB, 10 dB and 5 dB under the traditional method, SVM and Random Forest algorithms, respectively. From Fig. 7 (b), the general trends of the curves under all algorithms indicate that SI declines roughly when T_{60} grows, which agrees with the previous results that shorter T_{60} leads to higher SI. From Fig. 8, the phenomenon can be observed that longer T_{60} required lower SNR to achieve the best listening performance.

3.2.3. Relationship among SI, L_s and SNR

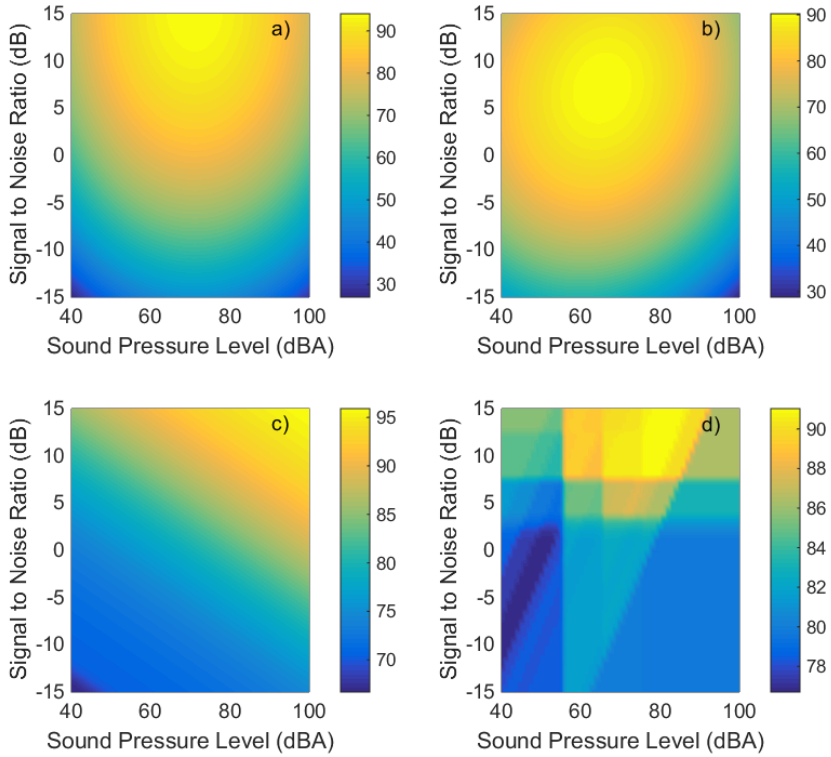


Fig. 9 The SI versus the SNR and L_s calculated by (a) traditional average method; (b) support vector machines algorithm; (c) artificial neural network algorithm and (d) the random Forest algorithm.

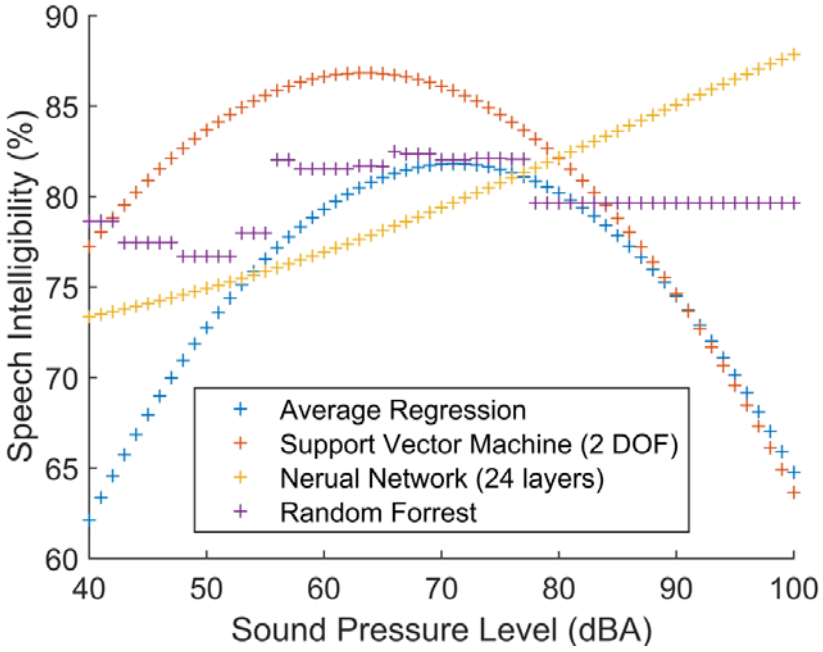


Fig. 10 The variation trends of SI versus L_s

Based on different algorithms, the relationship between the SI and L_s at certain T_{60} gender and SNR when Chinese listeners are listening English is illustrated in Figs. 9 and 10, respectively. The general trends in Figs. 9 (a) (c) (d) and 10 indicate that SI increases with L_s at first, and then decreases after it reaches its highest value under all the algorithm except the artificial neural network. The trends are consistent with the previous result that the syllable articulation increases drastically with L_s when L_s is lower than 50 dB, while it drops with a slow rate when L_s exceeds 80 dB. According to Peng's research, the optimal L_s for Chinese Mandarin speech intelligibility locates within 70 – 85 dBA, which match the result from random Forest algorithm. From the research reported in this paper, the optimal L_s for the speech intelligibility when Chinese listeners are listening English locates around 70 dBA, depending on the values of T_{60} and SNR. While from the result of the random Forest algorithm, present in Fig. 9 (d), this phenomenon is more and more obvious when SNR is higher. In the opposite side, when SNR is low, the SI will increase with the L_s increase. The optimal

L_s will appear when the SNR is around 7 dB.

3.2.4. Relationship between SI, T_{60} and L_s

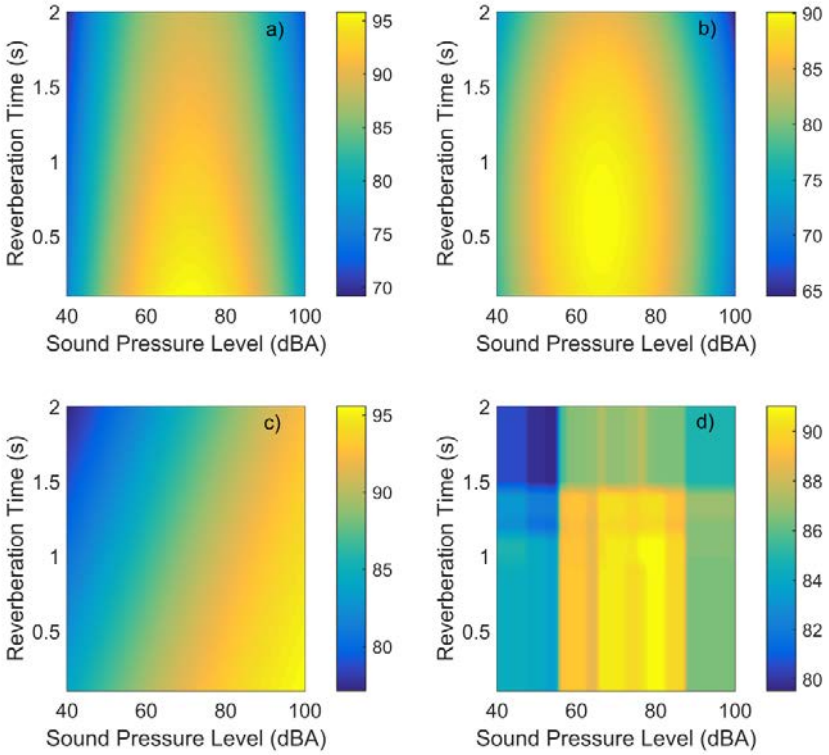


Fig. 11 The SI versus the T_{60} and L_s calculated by (a) traditional average method; (b) support vector machines algorithm; (c) artificial neural network algorithm and (d) the random Forest algorithm.

Fig. 11 presents the combination relationship among the SI, T_{60} and L_s. From the random Forest algorithm, the variation of the SI with L_s is much larger when T_{60} is longer.

3.2.5. Relationship between SI, SNR, T_{60} L_s and gender

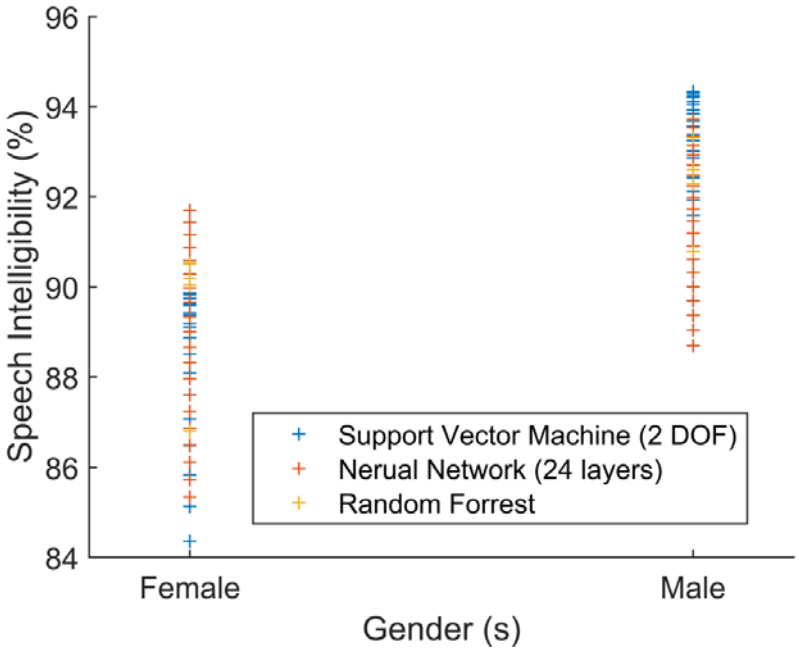


Fig. 11. The relationship between SI and gender under three machine learning algorithm

From Fig. 11, it is obvious that the female score a little lower than male do under all the algorithms.

4. Conclusions

A speech intelligibility test was conducted by playing back the prepared English test material with headphones in a quiet room to Chinese students who use English as the second language and the collected data was processed with machine learning algorithms and are compared with the traditional method. Three major machine learning algorithm were used in prediction include the artificial intelligent, support vector machine and random forest algorithm. In addition, to study the importance of each factor, the decision tree was used as well. The algorithms reveal that the signal to noise ration plays the most important role in determining the speech intelligibility, and the speech pressure level, reverberation time and gender. When signal to noise ratio is low, the gender and reverberation time do not take much effect and when signal to noise ratio is high enough, the gender takes effect when speech level is high as well ,

otherwise the reverberation time takes effect. Via the comparison of all the algorithms, the random forest does the best job in speech intelligibility prediction both tin the accuracy and cost error. The test results show that the optimum speech level is about in the range of 70 to 80 dBA when SNR and T_{60} keep constant. Both too high and too low speech levels result in poor speech intelligibility when SNR is high enough.