On Spoken English Phoneme Evaluation Method Based on Sphinx-4 Computer System

https://doi.org/10.3991/ijet.v12.i12.7957

Li Qin
Jilin Institute of Chemical Technology, Jilin, China
missqinli@qq.com

Abstract—In oral English learning, HDPs (phonemes that are hard to be distinguished) are areas where Chinese students frequently make mistakes in pronunciation. This paper studies a speech phoneme evaluation method for HDPs, hoping to improve the ability of individualized evaluation on HDPs and help provide a personalized learning platform for English learners. First of all, this paper briefly introduces relevant phonetic recognition technologies and pronunciation evaluation algorithms and also describes the phonetic retrieving, phonetic decoding and phonetic knowledge base in the Sphinx-4 computer system, which constitute the technological foundation for phoneme evaluation. Then it proposes an HDP evaluation model, which integrates the reliability of the speech processing system and the individualization of spoken English learners into the evaluation system. After collecting HDPs of spoken English learners and sorting them into different sets, it uses the evaluation system to recognize these HDP sets and at last analyzes the experimental results of HDP evaluation, which proves the effectiveness of the HDP evaluation model.

Keywords—oral English phoneme, HDP, Sphinx-4 computer system, evaluation model, evaluation algorithm

1 Introduction

With the implementation of national economic and political development strategies like “Going-out” and “Belt and Road”, Chinese are facing more opportunities of international communication and trade, and English, as an international language, is being more widely learnt in China [1]. Spoken English is the most direct form of English learning and communication; however, for most Chinese English learners, pronunciation has become their bottleneck in improving their spoken English, especially HDP, which are the areas where Chinese spoken English learners frequently make mistakes [2]. HDP is the symbolic feature of spoken English, reflecting the liaison of pronunciations and phonemic feature in spoken English. These two are the important guarantee for barrier-free communication with foreigners [3]. Performing HDP evaluation on English learners’ spoken English and giving continuous feedbacks will greatly improve their learning efficiency.

This paper aims to study the phoneme evaluation algorithm for spoken English to achieve phonetic feature evaluation on HDP. First of all, it introduces the typical computer-based automatic speech recognition technology and spoken language evaluation methods and describes the front-end, decoding and knowledge base functions of the key computer system Sphinx-4 applied in this paper [4], which has laid a theoretical and technical foundation for the spoken English phoneme evaluation algorithm. Then, this paper discusses the steps to establish the HDP evaluation model and the problems it is facing. By analyzing the model algorithm and system reliability, it establishes the HDP evaluation model. By integrating the computer system Sphinx-4 and the HDP evaluation model, this paper makes some modifications and supplements to the spoken English speech evaluation model, and gives some HDP evaluation example. In the experimental verification section, this paper chooses the pronunciations of English native speakers, English major students and computer major students as the HDP sample sets, and by establishing reasonable verification test, it proves the effective evaluation results of the evaluation model.

The establishment of an HDP model is also of great guiding significance to the evaluation on other features of spoken English. It is an extension of computer-assisted teaching. Studying and promoting this model will help spoken English learners in China improve their English skills in individualized ways through computer technology.

2 Language processing technology

The development of computer technology and voice processing technology provides new ideas for automatic speech recognition. At present, the basic process of computer automatic speech recognition is as follows: (1) establishing an acoustic library; (2) collecting speech signals, including signal noise reduction and phonetic feature extraction; (3) decoder outputting the recognition results [5] The establishment of a model library is the key to the speech recognition technology. In practice, different models may be established based on words, syllables and phonemes. For example, the pronunciation recognition of Chinese takes syllable as the basic unit while that of English takes phoneme as the basic unit. Model training methods commonly used include dynamic time warping technology (DTW), hidden Markov model (HMM) and artificial neural network (ANN). At the same time, with the maturing of computer speech recognition technologies, a lot of decoding software and platforms are also emerging, such as the CSLU automatic speech recognition toolkit, HTK (Hidden Markov model toolkit) and the Sphinx series.

2.1 Spoken English evaluation method

The essence of spoken speech evaluation is to compare the speech of the spoken English learners with the standard speech database and obtain the similarity, and based on that, give a corresponding rating.

Waveform comparison is the most common spoken speech recognition method. Its advantages are simple algorithm and fast operation; and its disadvantages are that it cannot be applied in the learning of massive sentences and lacks flexibility. Considering these disadvantages, people have paid great attention to studying how to apply the computer automatic speech recognition system in assisting speech evaluation. The evaluation process is shown in Table 1.

In the evaluation process, it is found different learners have different problems in their spoken English learning. However, numerical rating cannot help learners improve their pronunciation, nor can it help them find their inherent pronunciation flaws. This paper selects the recognition engine as the experimental platform, and integrates the rating process into fuzzy logic to give spoken English learners linguistic scores. In this way, different pronunciation feature ratings can be given for specific grammar phenomena and for different spoken language learners.

Table 1. The evaluation process of computer automatic speech recognition system

Step	Content
1	Speaking practitioner input voice
2	The system cuts the input voice signal
3	Comparison of Speech Fragment and Learning Model
4	Calculate the similarity with the standard acoustic model
5	Analysis of input speech speed, cycle, rhythm and other characteristics
6	Combined with different scores, get the final score
7	Show the results to the practitioner in different forms

2.2 Computer system Sphinx-4

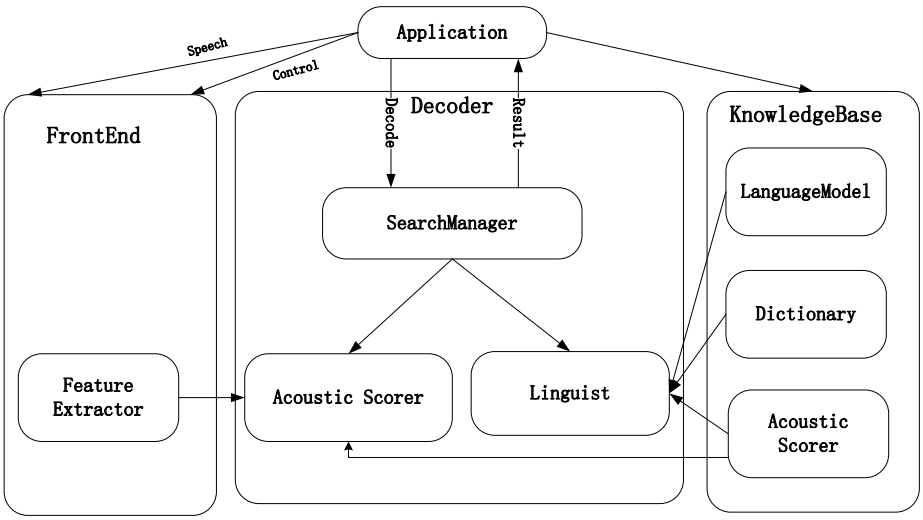

Fig. 1. The system architecture of Sphinx-4

The computer system Sphinx-4 is a continuous automatic speech recognition system written by Java language on Java platform. It was mainly designed and developed by Carnegie Mellon University. The highly modular and integrated model makes the system very efficient in speech recognition and configurable [6].

Figure 1 is a system structure diagram of the computer system Sphinx-4.

Sphinx-4 front end. he Sphinx-4 front end is mainly responsible for extracting feature frames of the speech signals. Each processing unit consists of an input, an output and the processing module. Each processing unit obtains data from the previous processing unit, which is processed by the processor and then buffered into the output buffer, waiting for the next processing unit to read this piece of speech information. At present, a front end processing unit can include sub-processing units such as preprocessing, windowing, spectrum analysis, discrete cosine transform, linear predictive coding and feature extraction according to the user's needs. The processing unit is configured to extract the required speech feature data [7].

Sphinx-4 decoder. The working process of the decoder is the construction process of a Sphinx-4 search graph. The steps are as follows:

1. The decoder acquires grammar from the knowledge base;
2. Expanding to a word-level network according to grammar rules;
3. Breaking down to a phoneme-level search network;
4. Phoneme sets form the search network with phonetic features.

After the establishment of the search network, the search algorithm is used to recognize the phonetic feature data to find the best matching results [8].

Decoder's work process is Sphinx-4 search map of the construction process, the steps are divided into:

1. decoder from the knowledge base to obtain grammar;
2. According to grammar rules to expand, constitute the word level network;
3. Step by step, constitute the phoneme level of the search network;
4. Through the phoneme collection constitute a phonetic feature search network.

After the search network is completed, the speech feature data is identified by the search algorithm, and the optimal matching result is found [8].

Sphinx-4 knowledge base. phinx-4 knowledge base consists of a language model, a dictionary and an acoustic model, supporting N-garm grammar and simple word list grammar and so on. Since this paper focuses on the phonemic features of spoken language, the main function of the Sphinx-4 dictionary is to map the grammar and words into basic phonetic symbol strings of phonemes.

3 Spoken English HDP evaluation model

In spoken English learning, there are always a considerable set of phonemes that are hard to distinguish for a particular group of English learners, which are called HDPs. For Chinese learners, phonemes like /t/ and /d/ and /w/ and /v/ are difficult to distinguish. If a model can master the pronunciation skills of different HDP sets and

give accurate HDP feedbacks and evaluation, it will be very helpful for spoken English Learners [9]. Different native language groups have different HDP sets. This paper mainly discusses the HDP evaluation model for English learners with Chinese as the mother tongue.

3.1 HDPs in Spoken English

HDP sets can be divided into two types. One is vowel confusion, such as the /e/ in bed and the /æ/ in bad and the other is the consonant confusion, like the /v/ in vet and the /w/ in wet. When a phoneme x is included in the samples for a spoken language learner, the system will provide all HDP sets containing this phoneme x and then obtain the most similar phoneme by searching for similarity [10].

For the example sentence "this is where I work", Figure 2 shows the candidate evaluation paths for all HDPs. Through the system evaluation, the accuracy of the learner's pronunciation can be given by the system after the recognition result comes out [11].

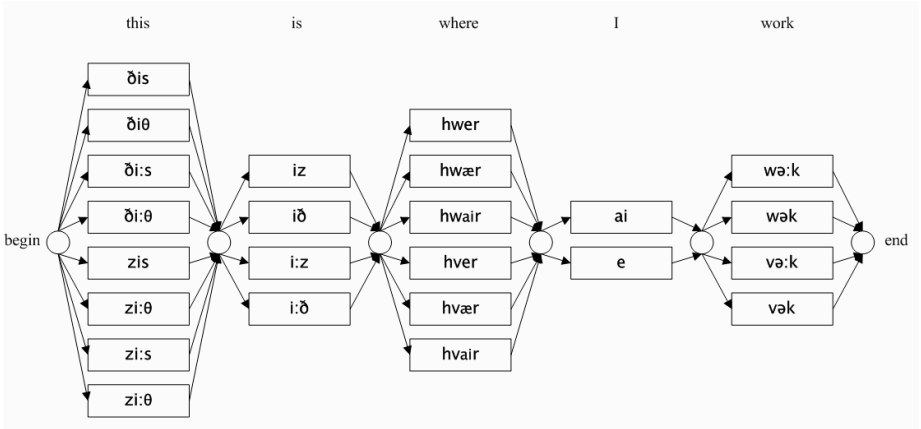

Fig. 2. The HDP graph for the sentence given in example

The evaluation model selects N=10 as the evaluation task constant and selects an HDP set containing 10 phonemes as the HDP cluster. A cluster is a unit of HDP evaluation [12].

3.2 Problems for HDP evaluation

The instability of the speech processing system is the main problem in HDP evaluation. The system takes the records of English native speakers as the standard corpus, compares them with the records of learners and obtains the system recognition results by taking statistics. In order to improve the reliability of HDP recognition, it is necessary to introduce the correct recognition rate r and the wrong recognition rate [13]. Their exact definitions are shown in Formula (1) (2):

$$r = N_{right}/(N_{right} + N_{error}) \tag{1}$$

$$e_i^j = N_c(i,j)/n_t(i) \tag{2}$$

indicates that the number of phonemes correctly recognized by the system, and indicates the number of phonemes incorrectly recognized by the system.

3.3 Reliability of the HDP evaluation model

In order to make the evaluation model more convincing, the concept of reliability is introduced to provide more robust evaluation results. The two metrics r and mentioned in the previous section are used to achieve reliability.

When the phoneme x is correctly identified by the system, the reliability of the phoneme can be defined per Formula (3):

$$f(x)=r(x) \tag{3}$$

When the phoneme x is recognized as the i-th phoneme in the same phoneme set, rather than x itself, its reliability can be defined per Formula (4):

$$f(x)= \tag{4}$$

3.4 HDP evaluation model

The HDP evaluation model is established according to the following steps:

1. Provide spoken English learners an HDP cluster statement script;
2. Record the speeches of the learners regarding these statements;
3. Generate all possible pronunciation extensions based on the HDP set and script;
4. Output HDP recognition results;
5. Calculate the reliability of HDP according to Formula (1) or (2);
6. Calculate the evaluation result of HDP cluster g. If g<0.5, and the switch variable R="Mediumness", input "NI" and exit; if g<0.5, and R="Goodness", exit; and if g<0.5, and R="Excellent", output "GOOD" and exit;
7. If g0.5, and the switch variable R="Mediumness", set R to "Goodness" and go to Step 5; if g0.5, and the switch variable R="Goodness", set R to "Excellent" and go to Step 5; and if g0.5, and the switch variable R="Excellent", then output "Excellent";
8. Exit.

In the algorithm, we use "0.5" as the key value for rating and determine the output value of phoneme evaluation through membership relation, providing stable model support for HDP phoneme recognition.

4 Implementation of the spoken English evaluation system

4.1 System implementation scheme

Based on the Sphinx-4 computer platform, this paper adds the HDP tagging module and HDP processing module. The speech evaluation system structure formed is shown in Figure 3.

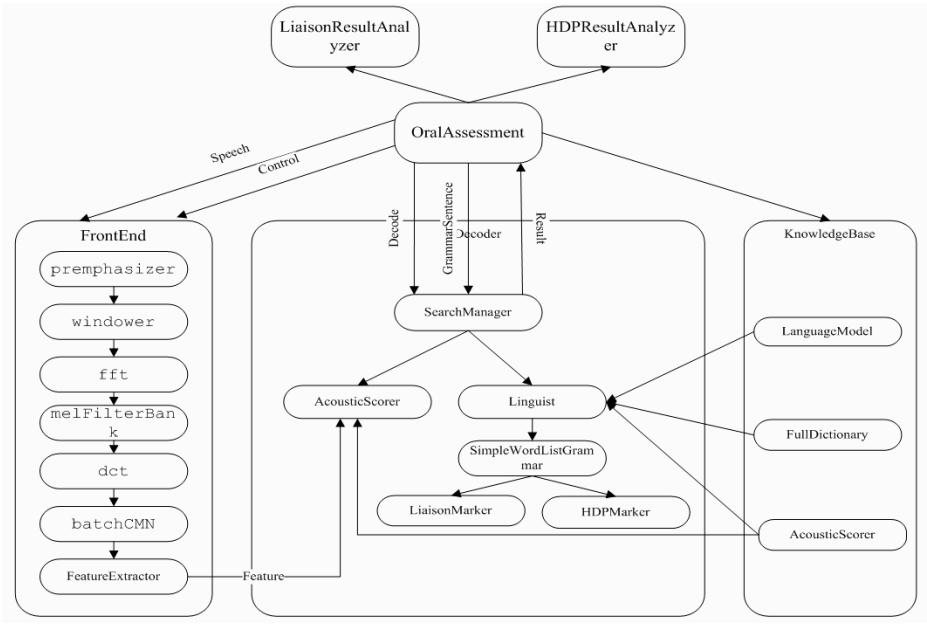

Fig. 3. The oral assessment system architecture

The system front-end configures recording pattern recognition to facilitate sound corpus test and identification. The system front-end configures the recording pattern recognition using such sub-processing units as pre-emphasis, windowing, ftt transform, discrete cosine transforms, batch processing and feature extraction [14].

4.2 HDP tagging and HDP recognition result processing

According to the needs of HDP evaluation, new words are added to the system for subsequent evaluation. The phonemes of each word are extended by replacing them with all HDPs one by one, and all the possibilities of pronunciations are obtained, and new words are formed [15].

For example, for the sentence “this is where I work”, as shown in Figure 2, the phonetic symbols of new words extended from each original word are linked with the acoustic model, and added to the system corpus. If the spoken English learner makes

a mistake, the system will recognize the incoming speech as a new word and gives an evaluation result.

The system in default only recognizes more than 60% of the matching words. With HDP cluster as the evaluation unit, the system performs maximum matching of each HDP in the cluster and then performs the evaluation on continuous clusters and gives the evaluation of language variables. Take the word "his" as an example. In the phoneme mode, this word is extended to {HH_IH_Z, HH_IY_Z, HH_IH_DH, HH_IY_DH}. The recognition system needs to find the maximum match such as "HH_IY_DH" and "HH_IH _Z", and then compares the standard statement and recognition statement, compares the actual pronunciation of the phonemes in the HDP set and calculates the reliability. After the system completes the evaluation of the entire statement, it will give the evaluation results regarding the individual word and phonemes.

4.3 Experiment description and result analysis

The training corpus T1 consists of 1064 pieces of recordings by native speakers, and T2 represents 122 HDP clusters recorded. For better comparison, students from the Department of English and Department of Computer Technology are selected as experimental samples, and their recordings of specific statements are denoted as T3 and T4. T1 is used for model training experiment, and T2, T3 and T4 for model verification experiment.

Since T2, T3 and T4 are from groups of people with different levels of spoken English, significantly different assessment results are expected. Table 2 and Figure 4 show the comparison of their speech evaluation results.

Table 2. The experiment results for the pronunciation evaluation taken on corpora T2-T4

	T2	**T3**	**T4**
Excellent	54%	19%	5%
Good	37%	53%	50%
Medium	7%	14%	25%
NI	2%	14%	20%

The system gives the T2 HDP set a 54% "excellent" rating, and gives 19% for T3 and 5% for T4. "Medium" rating is given to 24% of T4 and "NI" rating is given to 20% of it, rather different from the results of T2. The evaluation results of T3 are just between those of T2 and T4 and are generally consistent with the expectation. The pronunciation accuracy of English native speakers is higher than that of the Chinese students majoring in English and that of the latter is better than that of computer major students.

Figure 4 shows the comparison of the open test results of the corpus. It can be seen that the system shows an average phoneme recognition rate of more than 84%, proving the reliability and effectiveness of the spoken English phoneme evaluation model.

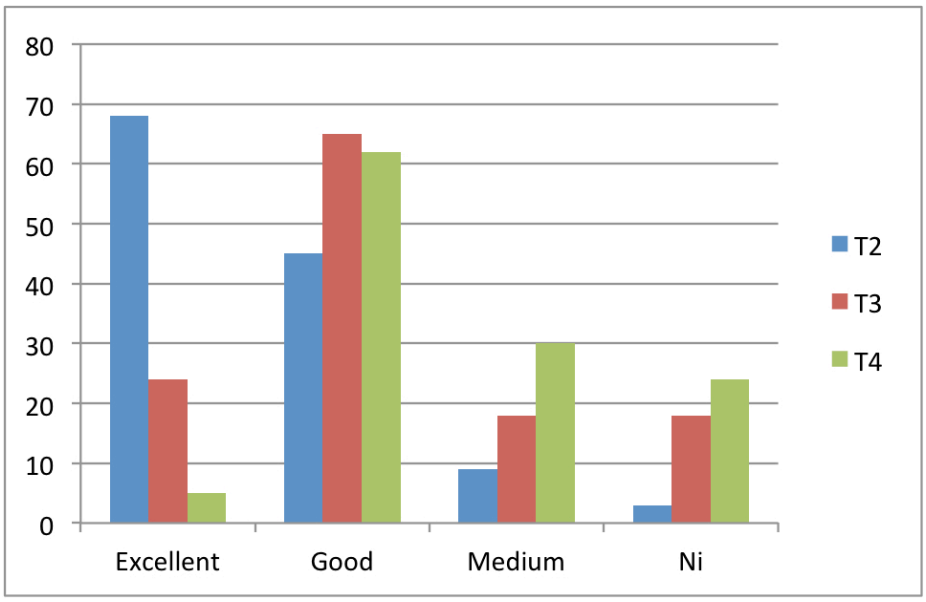

Fig. 4. The comparison of opening test results for test corpora T2-T4

5 Conclusions

Based on the application and development of the Sphinx-4 automatic speech recognition system, this paper introduces and analyzes in detail the phonemic features of HDPs in spoken English learning, aiming to improve the HDP pronunciation of spoken English learners in China. First of all, this paper describes the basis for speech processing and the computer Sphinx-4 platform and then establishes the HDP evaluation model. Through the integration of the Sphinx-4 platform and the evaluation model, this paper establishes the spoken English evaluation system and analyzes relevant evaluation results. This paper has the following conclusions and guiding significance:

1. The evaluation method can help spoken English learners improve their phoneme pronunciation in individualized ways and the evaluation results are reliable and stable.
2. The establishment of the evaluation method and the model is of great guiding significance to the promotion of the evaluation on other special features of spoken English.

6 References

[1] Scroyen, I., Bastelica, D., Poggi, M., Simonin, Y., Hibner, U., Rooijen, N., et al. (2015). Oral communication abstracts. Osteoporosis International, 29(2): 130-142.

[2] Ulijn, J.J., Strother, J.J. (1995). Technical and business negotiation: the listening and speaking processes in international communication. Journal of Fish Biology, 29(sA): 189-197.
[3] Lafontaine, H. (2012). Role and activation time course of the phonological and orthographic information during phoneme judgment. Neuropsychologia, 50(12), 2897-2906. https://doi.org/10.1016/j.neuropsychologia.2012.08.020
[4] Kong, Y.O., Cleuren, L., Latacz, L. (2009). Developing a reading tutor: Design and evaluation of dedicated speech recognition and synthesis modules. Speech Communication, 51(10): 985-994. https://doi.org/10.1016/j.specom.2009.04.010
[5] Lee, KaiFu. (1988). Large-vocabulary speaker-independent continuous speech recognition: the sphinx system. Speech Communication, 7(4): 375-379.
[6] Vojtko, J., Kacur, J., Rozinaj, G. (2007). The training of Slovak speech recognition system based on sphinx 4 for gsm networks. Australian Veterinary Journal, 85(3): 147-150. https://doi.org/10.1109/ELMAR.2007.4418818
[7] White, E.C., Dikangadissi, J.T., Dimoto, E., Karesh, W.B., Kock, M.D., Abiaga, N.O., et al. (2010). Home-range use by a large horde of wild mandrillus sphinx. International Journal of Primatology, 31(4): 627-645. https://doi.org/10.1007/s10764-010-9417-3
[8] Hicks, W.T., Yantorno, R.E. (2004). Determining the threshold for usable speech within co−channel speech with the sphinx automated speech recognition system. Journal of the Acoustical Society of America, 116(4): 2480-2480. https://doi.org/10.1121/1.4784905
[9] Zechner, K., Higgins, D., Xi, X., Williamson, D.M. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. Speech Communication, 51(10): 883-895. https://doi.org/10.1016/j.specom.2009.04.009
[10] Dahl, D.A., Norton, L.M., Scholz, K.W. (2000). Commercialization of natural language processing technology. Communications of the Acm, 43(11es): 7. https://doi.org/10.1145/352515.352525
[11] Kuo, C.C. (2011). Pronunciation assessment method and system based on distinctive feature analysis. Journal of the Acoustical Society of America, 130(6): 4183. https://doi.org/10.1121/1.3669375
[12] Rosenquist, C.J., Jr, H.J., Friedland, G.W., Gray, G.M., Zboralske, F.F. (1971). Assessment of a radiographic method for diagnosis of intestinal lactase deficiency: a prospective study. Investigative Radiology, 6(1): 40. https://doi.org/10.1097/00004424-197101000-00005
[13] Hicks, W.T., Yantorno, R.E. (2004). Determining the threshold for usable speech within co−channel speech with the sphinx automated speech recognition system. Journal of the Acoustical Society of America, 116(4): 2480-2480. https://doi.org/10.1121/1.4784905
[14] Scott, R.W., Korczak, B., Watkins, B.A., Sonis, S.T. (2012). Evaluation of a non-peptidic mimic of host defense proteins, pmx30063, in an animal model of oral mucositis. Journal of Clinical Oncology, 16(2): 80-90.
[15] Kranendonk, G., Hopster, H., Van, E.F., Van, R.K., Fillerup, M., De, G.J., et al. (2005). Evaluation of oral administration of cortisol as a model for prenatal stress in pregnant sows. American Journal of Veterinary Research, 66(5): 780-790. https://doi.org/10.2460/ajvr.2005.66.780

7 Author

Li Qin is with the School of Foreign Languages, Jilin Institute of Chemical Technology, Jilin 132022, China.

Article submitted 09 November 2017. Published as resubmitted by the authors 07 December 2017.