Optimization of Computer-aided English Pronunciation Training Data Analysis System

Chaohui Liang [1] and Jiling Shang [2]

[1]School of International Education & Europe-Asia Jiaotong, Zhengzhou Railway Vocational and Technical College, Zhengzhou 450000, China, yjfgef@126.com
[2]School of International Education & Europe-Asia Jiaotong, Zhengzhou Railway Vocational and Technical College, Zhengzhou 450000, China, redballons99@163.com

Corresponding author: Chaohui Liang, yjfgef@126.com

Abstract. In this paper, we propose an audiovisual fusion method based on the optimization of a computer-aided English pronunciation training data analysis system, which is based on the Convolutional Neural Network (CNN). An independent CNN structure is utilized to achieve independent modeling and asynchronous information transfer of audiovisual perception and to obtain descriptions of audiovisual parallel data in high-dimensional feature space, and then the long-time dependencies of the audiovisual parallel data in higher dimensions are modeled through a shared full-connection structure immediately following the CNN. Constructing auditory features to visual features characteristics of the generative model; the generative model is then used to automatically generate many visual features, which are combined with the CNN-based audiovisual fusion method to perform bimodal modeling. The experiments show that when the generated model is trained and tested in the same acoustic environment, only a small amount of audio-visual parallel data is required, and in combination with the proposed bimodal method based on visual feature generation. The bimodal method based on visual feature generation can effectively solve the problem of missing visual information in the actual usage environment. The audiovisual fusion method proposed in this paper can model the independence, asynchrony, and long-term interdependence between audiovisual parallel data, which is of great significance for the further study of the audiovisual fusion method based on deep learning.

Keywords: computer assistance; English pronunciation; training data; analysis system; optimization.
DOI: https://doi.org/10.14733/cadaps.2021.S4.37-48

1 INTRODUCTION

With the rapid development of science and technology, electronic devices such as computers, mobile phones, and tablet computers are gradually entering people's daily lives. In recent years, people have put forward higher requirements on various electronic devices-intelligent development. The traditional way of human-computer interaction through the mouse and keyboard touch screen has been unable to meet people's actual needs [1]. It has long been a dream of human beings to have a voice communication with machines directly and let them understand what humans say, Speech recognition technology enables the machine to "understand" the human language and directly execute the instructions conveyed by the language, which not only simplifies the input operation but also makes the "communication" between the human and the machine more convenient and natural. Undoubtedly, speech recognition technology has incomparable advantages in the future development of science and technology [2]. Especially in recent years, with the wide application of deep learning and the rapid development of artificial intelligence, voice recognition technology has received increased attention [3]. As one of the key technologies, voice recognition technology must occupy an extremely important position in the future intelligent development and is a field that all the big business giants must compete. Due to the lack of balanced regional economic development, the lack of teacher resources, and the limitations of the English learning environment, traditional language teaching uses classroom learning, in which even the smallest class has dozens of students. Therefore, the lack of instructional time for oral learning is often a weakness of classroom learning. The inability to get corrective feedback on pronunciation exercises in class is a difficulty for learners to learn English pronunciation [4].

O’Brien studied the distinguishing features of the mispronunciation of Dutch by non-Dutch speakers [5]. The features selected for the detection of vowel and consonant errors included duration, resonance peak, and fundamental frequency. The Rate of Rising (ROR) feature and the energy spectra of different positions are used to obtain the discriminating features for different consonants or vowels [6]. Two methods, Linear Discriminant Analysis (LDA) and decision tree were used to discriminate the phoneme errors. Experiments showed that using energy spectra has a better discriminative effect. Ho et al. studied the discriminative features of the flat and warped tongues in Chinese [7]. The results of the study showed that there is a great difference between them in the peak energy segment of the spectrum. Huang constructed the correct pronunciation network (The Correct Pronunciation Nets (CPNs)) and validated CPNs using a data-driven approach to find error patterns from everyday Mandarin-speaking students in the process of learning Taiwanese [8]. Finally, they also applied this technique to a CAPT system. Pouyanfar et al. studied the use of multi distributed neural networks for pronunciation error detection and diagnostic feedback to overcome the difficulties encountered by existing ENRs-based methods [9]. In the paper, it is proposed that the Acoustic-Graphemic-Phonemic Model (AGPM) using multi-distributed DNNs to receive the input acoustic features, and the corresponding normative transcription vectors, can implicitly model phonemes with similar pronunciation, and experiments show many improvements in the error rejection rate and phoneme error detection rate. Kang et al. studied the use of deep learning constructs for automatic speech scoring, using a large English corpus of approximately 800 hours of non-native vocabulary to build an ASR system [10]. Deep Neural Networks (DNN ASR) and Gaussian Mixture Model (GMM ASR) were compared, and the results showed that the deep learning ASR system was significantly better than the Gaussian ASR system. The ASRs using deep learning architecture improved in scoring accuracy, with scores closer to human expert scores, and always better than GMM. At the same time, Yarra et al. studied the use of DNN-based speech feature modeling to improve the accuracy of error detection in pronunciation detection [11].

As mentioned above, learners usually prefer to know directly what is wrong with their pronunciation and how to adjust their pronunciation to improve it, because the types of mispronunciation are often related to the mispronunciation of the learners' pronunciation-related organs or actions. Considering that the types of pronunciation errors mainly targeted in previous studies are typical phoneme mispronunciations, omissions, and insertions, this paper investigates the types of phoneme-level pronunciation errors that learners make due to the non-standard pronunciation actions. The MFCC has good acoustic properties and is widely used in natural language processing speech recognition and other techniques. It is used as an input feature for machine learning algorithms. This model is a good way to improve the error detection rate of pronunciation classification compared with the traditional techniques, and the MFCC feature dimensions are free to choose. In our experiments, we found that most of the learners' pronunciation errors are concentrated on common error types, which have a large sample size, while some error types are only present in a small number of learners' pronunciations. Not all the pronunciation error types have a balanced sample size but are normally distributed, thus limiting the coverage of error types based on MFCC-RF pronunciation classification error detection.

2 OPTIMIZED DESIGN OF COMPUTER-AIDED ENGLISH PRONUNCIATION TRAINING DATA ANALYSIS SYSTEM

2.1 Design of Computer-Aided English Pronunciation Training Data Analysis System

Traditional language teaching uses classroom learning, in which even the smallest class has dozens of students, and it is impossible to teach pronunciation on a one-to-one basis, but only vocabulary, grammar, and semantics. Therefore, the lack of instructional time for oral learning is often a weakness of classroom learning. The inability to get corrective feedback on pronunciation exercises in class is a difficulty for learners to learn English pronunciation. This is also the reason "dumb English" has plagued public English teaching in China for many years. E-learning is one of the most popular learning methods, such as live lectures and online quizzes, which allow learners to learn by themselves regardless of time and place. However, spoken pronunciation training is not very successful, and there are even fewer platforms for correcting mistakes. Only a few of them can detect oral pronunciation problems and providing feedback, but the feedback function is not enough to solve the learners' problems, as shown in Figure 1.

Since the multi-level adaptive deep network is essentially a classification network, the target domain-related features contain only the most distinguishing information related to pronunciation, while discarding some information related to visual feature generation. Simply using Target Domain-Specific Tandem Feature (TDSTF) instead of the original acoustic features as input to the BLSTM-RNN generated network will greatly reduce the visual feature association by stitching them together to form new features, which are recorded as Target Domain-Specific Tandem Feature (TDSTF), and use this feature to As input to the BLSTM-RNN generated network.

Speech signals in the time-frequency domain are not easy to model, so the signal is often time-frequency decomposed to transform a one-dimensional speech signal into a two-dimensional time-frequency domain signal. Speech is sparsely distributed in the time-frequency domain, and distinct harmonic structures can be observed, all of which are favorable for subsequent speech processing. In speech enhancement, the short-time Fourier transform is a common method for time-frequency decomposition.

Reviewing the development of supervised speech enhancement research, many time-frequency masking class targets have been proposed. A comparison and analysis of these time-frequency masking reveal that most of the masking class goals can be viewed as approximations to the optimal masking cIRM under certain assumptions. If IRM is decomposed in a Cartesian coordinate system, the projection of cIRM on the real axis is PSM. We can compare the reconstruction effects of different ideal time-frequency masks. The CIRM can achieve the almost lossless reconstruction of pure speech, while the other masks all suffer some degree of performance loss due to certain specific assumptions. Although cIRM is the optimal mask, the use of other simplified masking targets can reduce the difficulty of prediction.

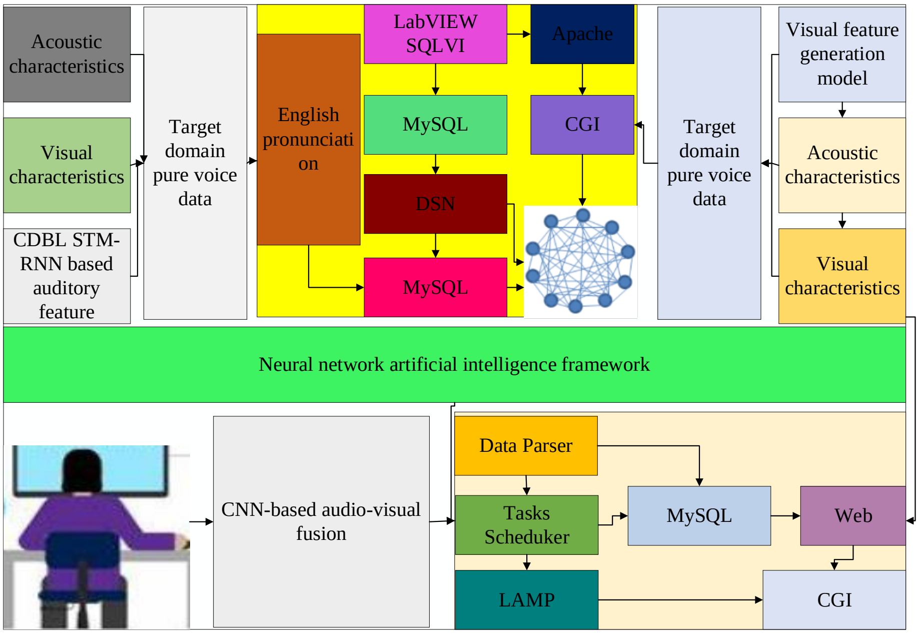
Acoustic characteristics
Visual characteristics
CDBL STM-RNN based auditory feature
Target domain pure voice data
English pronunciation
LabVIEW SQLVI
Apache
MySQL
CGI
DSN
MySQL
Target domain pure voice data
Visual feature generation model
Acoustic characteristics
Visual characteristics
Neural network artificial intelligence framework
CNN-based audio-visual fusion
Data Parser
Tasks Scheduker
MySQL
Web
LAMP
CGI

Figure 1: Application of cross-domain adaptive visual feature generation-based approach to bimodal modeling.

Therefore, early speech enhancement studies chose to use simple masking targets such as IBM or IRM. In cases where model capacity is limited, cIRM is often not the best choice, and optimal enhancement performance is obtained by choosing targets that match the modeling capabilities of the model.

2.2 Optimal Design Analysis of the System

Different roles such as teachers and students have different access and operation rights. The automatic pronunciation correction system should be able to run stably. The automatic pronunciation correction system is designed for learners. The system should be developed in MVC mode with the principle of low coupling and high cohesion, and each functional unit should be divided into modules. In the design of the Web system, scalability is an important aspect to determine whether the system is mature or not. The English pronunciation correction system should be extensible to add other functions later [12].

As shown in Figure 2, the roles of the user management module are divided into three categories, according to different roles have different access and operating privileges, in addition to the basic user registration and login, the main operation of the students only self-learning self-assessment. The teacher's operation for all students, also, the teacher can check for errors in words or statements to add or delete a series of operations, while the system administrator has the highest authority, can manage the system. The pronunciation data collection module collects student's pronunciation data during the student self-testing. The pronunciation data check module checks the collected student's pronunciation by classification through the embedded pronunciation check model, and the pronunciation data correction module gives the result of the pronunciation check and corrective comments. The Pronunciation Data Correction module gives the results of the student's pronunciation check and the corrections. The history data display module can view the previous checks. The automatic pronunciation correction system is based on the advantage of the Internet. Learners access the system through a computer or mobile device, upload their pronunciations, and the system processes them to give feedback on the pronunciation corrections. Learners will be able to recognize their pronunciation problems and improve their spoken English pronunciation through repeated practice. The design of the automatic pronunciation correction system is divided into a general system design, a business design, and a database system design.

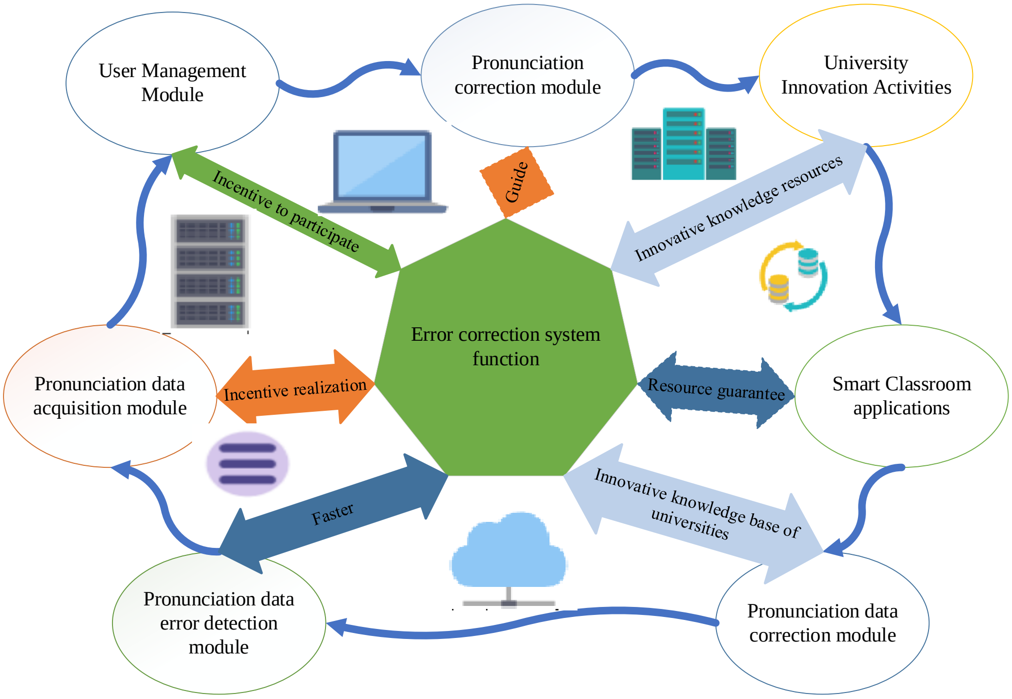
User Management Module
Pronunciation correction module
University Innovation Activities
Guide
Incentive to participate
Innovative knowledge resources
Error correction system function
Pronunciation data acquisition module
Incentive realization
Resource guarantee
Smart Classroom applications
Faster
Innovative knowledge base of universities
Pronunciation data error detection module
Pronunciation data correction module

Figure 2: Diagram of Automatic Pronunciation and Error Correction System.

The system development is divided into three layers, the web layer (also called View), which is responsible for interacting with the client browser to receive requests, and the logical layer (also called Service), which is responsible for handing over the request processing to the salesman. After the database operation, the result is returned to the business logic layer for encapsulation, and the response is sent to the view layer, which is parsed by the view layer and displayed on the client browser page. Java-based Spring is the latest popular lightweight Web MVC pattern framework, through the request-response driver model to simplify our development.

2.3 Analysis of Performance Indicators

In the adjustment stage of the model, the author, by reviewing relevant literature, believes that the initial model design is still inadequate. According to the definition of blended learning, the design of the teaching model uses a combination of online student self-learning and offline teacher classroom teaching; at the same time, the teaching stages are divided into before, during, and after class, as shown in Figure 3.

In the pre-lesson phase, the focus is on promoting student-directed learning, using problem-based learning, where students are given tasks to carry out. The teacher will set the teaching task before the lesson, which will make students think actively and motivate them to learn. However, the setting of the task should not be too complicated to avoid making students lose interest in learning and causing unsatisfactory teaching results. At the same time, teachers look for suitable teaching resources on the appropriate online resource sites before the lesson, push these teaching resources to students in time, and collect students' learning problems at the same time.

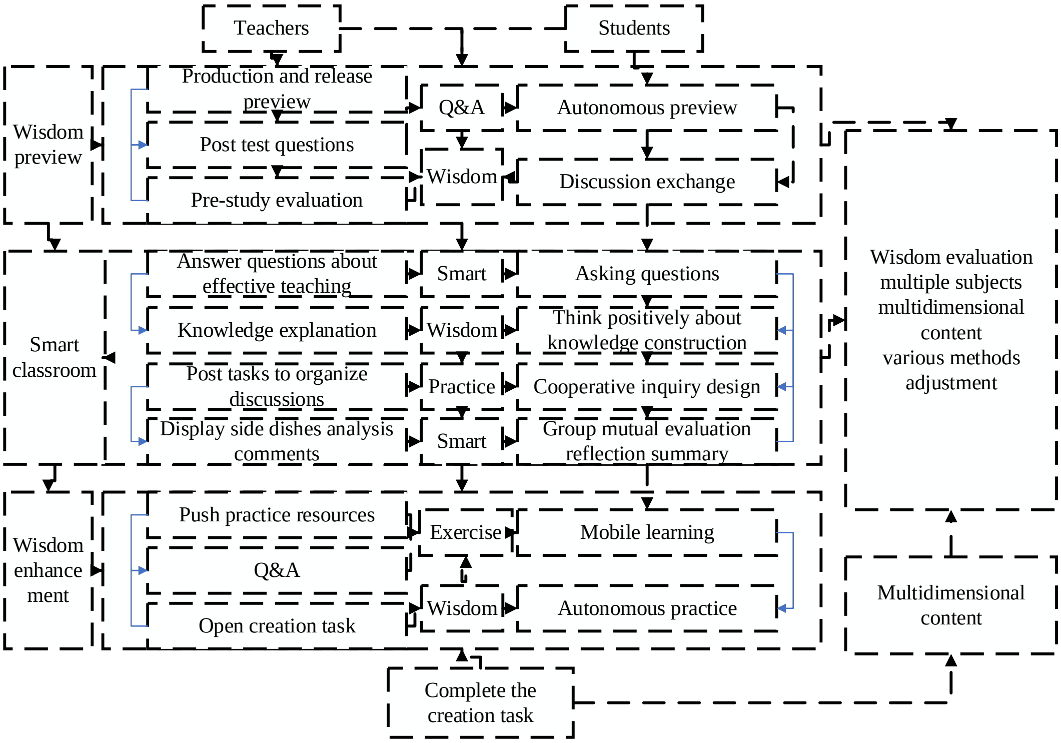
Teachers
Students
Wisdom preview
Production and release preview
Post test questions
Pre-study evaluation
Q&A
Wisdom
Autonomous preview
Discussion exchange
Smart classroom
Answer questions about effective teaching
Knowledge explanation
Post tasks to organize discussions
Display side dishes analysis comments
Smart
Wisdom
Practice
Smart
Asking questions
Think positively about knowledge construction
Cooperative inquiry design
Group mutual evaluation reflection summary
Wisdom evaluation multiple subjects multidimensional content various methods adjustment
Wisdom enhance ment
Push practice resources
Q&A
Open creation task
Exercise
Wisdom
Mobile learning
Autonomous practice
Multidimensional content
Complete the creation task

Figure 3: Performance index test design.

The main content of student learning is based on online learning with instructional videos chosen by the teacher, and they need to complete the corresponding online quizzes at the end of the learning process. In the learning process, we focus on students' ability to grasp and remember difficult words and phrases to form their knowledge structure system. According to Gardner's theory of multiple intelligences, each student has different intellectual strengths and therefore different levels of proficiency in what they have learned. In this model, the teacher can guide students to develop their own learning goals and learning plans according to their different bases, provided that the students do not exceed the existing teaching objectives. The student's tasks in the pre-class period are carried out entirely by the student, and the teacher only provides learning resources and collects the difficulties encountered by the student in watching the video, so the student can choose the place of learning and just watch the video.

In the classroom, the teacher understands the results of students' independent learning before the lesson, and at the same time, students encounter some learning questions before the lesson to provide answers. For students who do not understand the question, the teacher for student feedback to do detailed answers. The teacher will adjust the teaching content according to the questions raised by students. The teacher firstly determines the teaching objectives and then explains the basic knowledge points of the lesson in a way that is easy to understand and deep enough to ensure that students at different levels can understand. Then the teacher puts forward further tasks that students need to complete according to the content of the lesson, grouping students in the class, taking into account the different characteristics of different students, students who do not want to study in groups can also complete the learning tasks alone. Ultimately, students will present their learning outcomes, in the form of specific presentations including sending representatives to the stage to express themselves and group members engaging in situational dialogue presentations to deepen their understanding of the knowledge. Also, the classroom should consider teaching evaluation, which takes the form of not only a single focus on student learning results, but more will be the participation of students in the classroom, the changes in student learning attitudes, and the application of learning outcomes, such as the combination of the promotion of evaluation forms of diversification, evaluation of personalized content. The student-oriented concept is always implemented in the assessment process, combining formative and summative assessments to cultivate high-level talents. At this stage, teaching activities mostly take place inside the classroom, and teachers are the instructors of teaching activities.

3 ANALYSIS OF RESULTS

3.1 Comparison of Optimization Results

For real-time speech recognition, the external radio device selects a headset with two plugs. Of the two plugs, one transmits the voice signal and pushes the headset to sound; the other transmits the voice signal received by the headset microphone to the workstation and stores the voice signal as a wav audio file. The audio recorded in real-time can be recognized directly or by processing it and then recognizing it. Pre-processing, on the other hand, can be done with the endpoint detection and denoising algorithms described above. Endpoint detection detects the entire audio segment and picks out the audio segment and saves it as a new file, while denoising processes the entire audio and generates a new file. The newly generated file can be identified using the model. The results for the noise matching condition are first compared, and the detailed data are shown in Figure 4. The performance difference between the DNN and the capsule network on the noise matching test set is very small, and it can identify and process the same type of noise seen in the training phase very well.

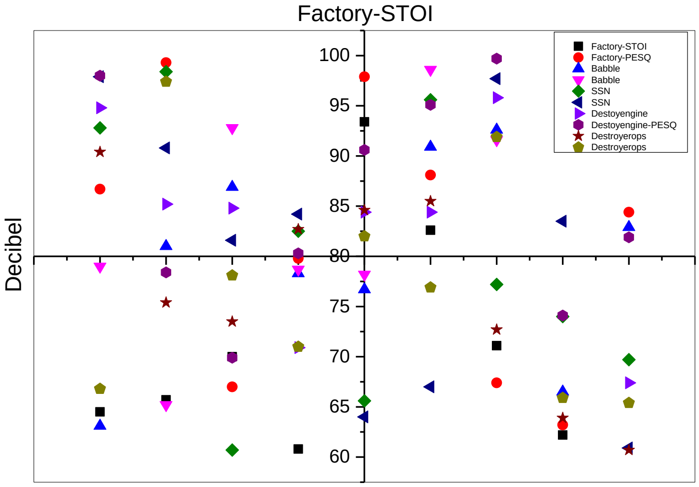

Figure 4: Comparison between capsule network and DNN under noise matching condition.

The baseline system uses a DNN model with four hidden layers, each with 1024 nodes, and a linear rectification unit (ReLU) as the activation function. The output layer of the DNN is activated using a Sigmoid function, and the result is used as the predicted IRM. To prevent overfitting, the model is trained using the Dropout technique. The discard probability of the nodes is set to 0.2. the convolutional kernel size in the capsule network using one layer is set to 9, the number of convolutional kernels is set to 256 for generating 32 sets of 8-dimensional capsules, and the number of dynamic routing iterations is set to 3. The loss function of the model is the mean square between the predicted IRM and the true IRM Error. All models were optimized using the Adaptive Moment Estimation (Adam) optimizer with the learning rate set to 0.001 and the Batch Size set to 32. The results of the tests performed under noise mismatch conditions are shown in Figure 5. Since there is a gap between these noise types and the noise used in training, it reflects the generalization performance of the model to noise. It can be found that the DNN has very little index improvement on these untrained noises, indicating that the DNN's generalization performance to noise is poor. The capsule network, however, is still able to detect and suppress these noises well, which indicates that the capsule network learns more essential discriminative patterns of speech and noise from limited data.

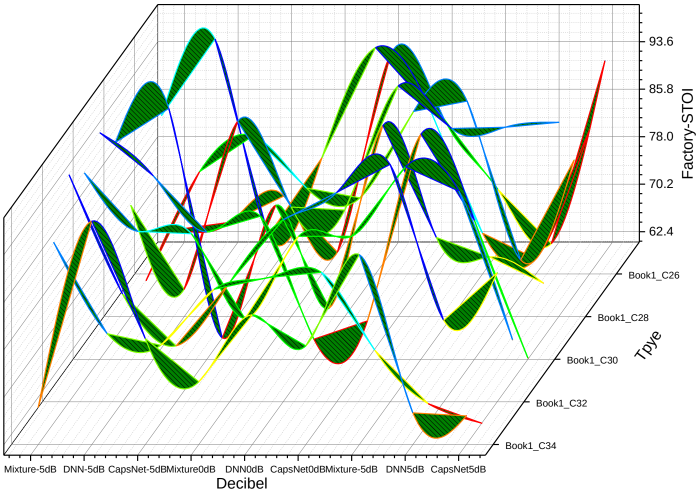

Figure 5: Comparison of capsule network and DNN under noise mismatch condition.

The temporal convolution contained in the TCRN can be thought of as a time-frequency decomposition of the signal using a set of filters whose parameters can be learned. After the model training is complete, we can analyze the learned set of filters. To facilitate the analysis, a Fourier transform is first performed on each filter, after which the filters are rearranged according to their center frequencies, which yields the results shown in Figure 5. The horizontal axis in the figure represents the number of filters after rearrangement, and the vertical axis represents the frequency. Looking at the distribution of the filters, one can see that more filters are concentrated in the low-frequency region, while the high frequencies are more sparsely distributed. This phenomenon may be related to the fact that the energy of the human voice is mainly distributed in the middle and low frequencies, and the signal in the middle and low-frequency parts needs to be more finely depicted.

3.2 Analysis of Performance Indicator Results

Figure 6 shows a comparison of the performance of the BLSTM-RNN generated model when trained with different amounts of audiovisual parallel data. Left ventricular diastolic is a complex process involving many factors, including the active bradycardia and the passive filling of the heart cavity, and it is affected by many factors, such as the rigidity of the wall, the relaxation function of the left ventricle, the coordination of the myocardium, the load before and after the heart. The developing central nervous system is most susceptible to environmental factors, and long-term, low-dose environmental physical and chemical factors damage the nervous system's functions often before organic damage, which is specifically manifested as changes in behavioral functions. The behavioral test method can detect the damage to the body from environmental harmful factors early and sensitively. The water maze was designed by British psychologists in the twentieth century.

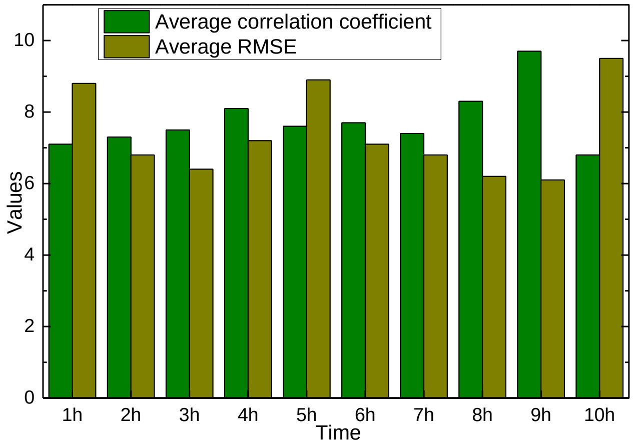

Figure 6: Performance of different near-field visual feature generation models based on two-way long and short-term memory networks.

By comparing the RMSE results in Figure 6, the visual feature trajectories generated in the far-field conditions are closer to the visual feature trajectories generated in the near-field conditions than those generated in the far-field conditions. Spatial cognition is formed by processing external clues of spatial information. According to the encoding and processing methods of information, memory can be divided into declarative memory and procedural memory. In the water maze experiment, the position of the platform has nothing to do with the position and state of the rats themselves. It is a kind of reference cognition with reference to heterosexuality, and the memory formed is a kind of spatial reference memory. Judging from the information processing and extraction methods, this spatial memory enters the consciousness system, and its mechanism mainly involves the limbic system such as the hippocampus and the cerebral cortex-related brain regions, which are declarative memories. Also, while the poorer far-field speech signals degrade the quality of the generated visual information, this has a limited impact on the performance of the AVSR system obtained from the final training.

Also, to analyze the performance of the BLSTM-RNN generation model more directly, the best BLSTM-RNN generation model under far-field conditions is shown in Figure 7, and an "invisible" sentence is randomly selected from the test set, and then the corresponding generated visual feature trajectory is compared with the BLSTM-RNN model. The first three components of the real trajectory were compared, and the results are shown in Figure 7. as seen in Figure 7, the difference between the real and generated trajectories is very small. Given the results in Figure 7, and to further explore the processing of the fall-based speaker adaptive training method for invisible speakers, the following section provides a detailed analysis of the results for the fMLLR-based speaker adaptive training on the auditory channel. Figure 8 shows the word error rates of the recognition system performance on the "visible set" and "invisible set" using the fMLLR-based speaker adaptation training technique in both near-field and far-field test environments, respectively. The word error rate of the recognition system using the fMLLR-based adaptive speaker training technique decreases more in the "invisible set" than in the "visible set", both in the near-field and far-field conditions. For example, Figure 8 shows that the word error rate of the traditional AVSR system "CNN-av" dropped by 34.45% on the "invisible set" after adaptive training in the far-field test condition, but the word error rate of "CNN-av" dropped by 34.45% on the "visible set" after adaptive training in the far-field test condition. The word error rate on the "visible set" was only reduced by a relative 5.02%. This indicates that the improved performance of the fall-based speaker adaptive training technique is mainly due to the reduction of variability caused by the "invisible" speaker.

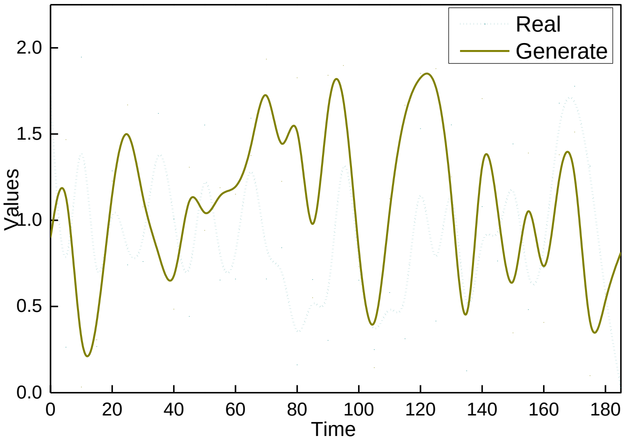

Figure 7: Comparison of the first three components of the real trajectory and the generated visual feature trajectory.

The BLM-RNN generative model training phase and the AAVSR acoustic model training phase. Numerous experiments have shown that in the BLSTM-RNN generative model training phase. The experiments show that by using this BLSTM-RNN-based visual feature generation method, only a small amount of audio-visual parallel data is required, and then a large amount of pure speech data and additional generated visual data can be used to construct the AVSR system, thus improving the performance of the original ASR baseline system; especially in the complex far-field environment, the proposed method brings more obvious recognition system performance improvement. It should be noted that this paper makes some limitations in this part of the experiment: the speech part of the audiovisual parallel data is the same as the acoustic acquisition environment for pure speech data. However, the pure speech data from real life is often susceptible to various environmental factors, so its characterization in the acoustic domain generally differs greatly from the speech part of the audiovisual parallel data from the laboratory environment, which is the so-called acoustic domain mismatch problem. The next chapter will explore this problem in-depth, based on a bimodal modeling approach based on visual feature generation.

4 CONCLUSION

To improve learners' pronunciation, this paper builds a pronunciation error detection model with machine learning algorithms to fill the gap in the detection of errors in pronunciation. With the experimental model of pronunciation error detection and the need of most English learners to improve their oral language, we are committed to pushing the experimental model into industrial applications, so in today's rapid development of Internet technology, we combine the Internet online learning methods.

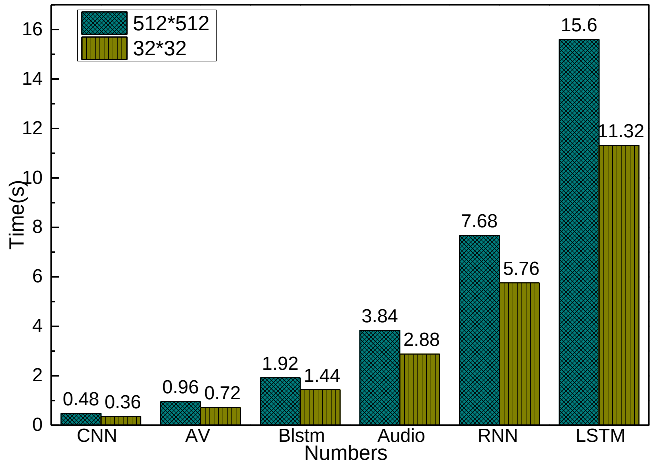

Figure 8: Bimodal modeling with speaker adaptive training analysis based on visual feature generation method.

We embed the built model into a web system, aiming to build a real-time online pronunciation correction system that everyone can use. After studying the corpus and acoustic features, we select the acoustic features and corpus suitable for this paper and propose a pronunciation classification error checking model method based on MFCC-RF. The random forest classifier is trained on the acoustic information carried by the acoustic features as a discriminating feature to classify and detect most of the common pronunciation errors, and the experimental results show that the model can accurately identify the categories of phoneme pronunciation errors. The experimental results show that the model can accurately identify the categories of phoneme mispronunciation. This provides a new method for automatic mispronunciation detection.

Chaohui Liang, https://orcid.org/0000-0001-7072-5128
Jiling Shang, https://orcid.org/0000-0002-0519-5578

REFERENCES

[1] Krecichwost, M.; Miodonska, Z.; Badura, P.; Trzaskalik, J.; Mocko, N.: Multi-channel acoustic analysis of phoneme/s/mispronunciation for lateral sigmatism detection, Biocybernetics and Biomedical Engineering, 39(1), 2019, 246-255. https://doi.org/10.1016/j.bbe.2018.11.005

[2] Qian, X.; Meng, H.; Soong, F.: A two-pass framework of mispronunciation detection and diagnosis for computer-aided pronunciation training, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24(6), 2016, 1020-1028. https://doi.org/10.1109/TASLP.2016.2526782

[3] Farouk, A.; Zhen, D.: Big data analysis techniques for intelligent systems, Journal of Intelligent & Fuzzy Systems, 37(3), 2019, 3067-3071. https://doi.org/10.3233/JIFS-179109

[4] Therese, S.-S.; Lingam, C.: Optimisation of training samples in recognition of overlapping speech and identification of speaker in a two speakers situation, International Journal of Advanced Intelligence Paradigms, 17(1-2), 2020, 159-177. https://doi.org/10.1504/IJAIP.2020.108773

[5] O'Brien, M.-G.; Derwing, T.-M.; Cucchiarini, C.; Hardison, D. -M.; Mixdorff, H.; Thomson, R. -I.; Levis, G. -M: Directions for the future of technology in pronunciation research and teaching, Journal of Second Language Pronunciation, 4(2), 2018, 182-207. https://doi.org/10.1075/jslp.17001.obr

[6] Sanchez-Lara, A.; Chochlidakis, K. -M.; Lampraki, E.; Molinelli, R.; Molinelli, F.; Ercoli, C.: Comprehensive digital approach with the Digital Smile System: A clinical report, The Journal of prosthetic dentistry, 121(6), 2019, 871-875. https://doi.org/10.1016/j.prosdent.2018.10.012

[7] Ho C. -W. -L.; Soon, D.; Caals, K.; Kapur, J.: Governance of automated image analysis and artificial intelligence analytics in healthcare, Clinical radiology, 74(5), 2019, 329-337. https://doi.org/10.1016/j.crad.2019.02.005

[8] Huang, H.; Xu, H.; Hu, Y.; Zhou, G.: A transfer learning approach to goodness of pronunciation based automatic mispronunciation detection, The Journal of the Acoustical Society of America, 142(5), 2017, 3165-3177. https://doi.org/10.1121/1.5011159

[9] Pouyanfar, S.; Sadiq, S.; Yan, Y.; Tian, H.; Tao, Y.; Reyes, M. -P.; Iyengar, S. -S.: A survey on deep learning: Algorithms, techniques, and applications, ACM Computing Surveys (CSUR), 51(5), 2018, 1-36. https://doi.org/10.1145/3234150

[10] Chien, J. -T.; Mak, M. -W.: Guest Editorial: Modern Speech Processing and Learning, Journal of Signal Processing Systems, 92(8), 2020, 775-776. https://doi.org/10.1007/s11265-020-01577-4

[11] Kang, O.; Johnson, D.: The roles of suprasegmental features in predicting English oral proficiency with an automated system, Language Assessment Quarterly, 15(2), 2018, 150-168. https://doi.org/10.1080/15434303.2018.1451531

[12] Yarra, C.; Ghosh, P. -K.: Automatic intonation classification using temporal patterns in utterance-level pitch contour and perceptually motivated pitch transformation, The Journal of the Acoustical Society of America, 144(5), 2018, EL471-EL476. https://doi.org/10.1109/SACI.2018.8440938