International Journal of Web-Based Learning and Teaching Technologies
Volume 18 • Issue 2

The Reform of Pronunciation Teaching in Colleges and Universities by Praat Software From the Perspective of Deep Learning

Khuselt It, Inner Mongolia Minzu University, China*

ABSTRACT

Due to the difficulties of speech signal processing, there is still a considerable gap between the ability of machines to correctly process and that of human beings. In order to overcome the defects of isolated learning and noise sensitivity of SOM, this paper proposes a new time self-organization model (TSOM) from the perspective of deep learning. On the basis of self-organizing mapping network, time enhancement mechanism is introduced to improve the system performance. This method makes up for the fixed spatial topology of the original self-organizing mapping network and the neglect of the time factor, which is crucial to the voice signal. At the same time, this paper makes full use of computer-aided technology and rich network resources to provide a comprehensive and systematic English pronunciation learning database and establish learners' pronunciation files. Once learners understand and master the operation of voice analysis software, they can conduct self-assessment and judgment to find out their blind spots and weaknesses in voice acquisition.

KEYWORDS

College Pronunciation Teaching, Deep Learning, English, Praat Software, TSOM Model, Visualization

INTRODUCTION

English as an international language plays a pivotal role in today's international communication process. Clear and understandable speech is the foundation of people's communication. Whether the pronunciation is correct or not directly affects the expression of meaning. Voice can be divided into vowels and consonants, among which vowels are the core and backbone and are the basis of pronunciation. Due to factors that are difficult to grasp, such as the opening degree and tongue position, it is not easy to master the single tone. Traditional English pronunciation teaching mainly relies on teachers' demonstration of the pronunciation process and explanation of pronunciation methods. Learners rely on their own auditory perception to listen, imitate, and mechanically practice the teachers' pronunciation, audio or video, and practice repeatedly after teachers' error correction. Traditional phonetic teaching lacks objective perceptual evaluation criteria and cannot achieve satisfactory teaching results.

DOI: 10.4018/IJWLTT.325225 *Corresponding Author

For unit phonetic teaching, current research mainly focuses on the comparison of vowels between English and Chinese, the mechanism of second language acquisition, negative transfer of mother tongue, and interlingual research. Through the comparison of English and Chinese vowels, many scholars have discovered the similarities and differences between the two, and pointed out the essentials of pronunciation. They are helpful and enlightening to Chinese English learners, but the method and correctness of pronunciation are only based on subjective perception, not objective enough. The development of global informatization has provided new means for education. The birth and development of the Internet has provided a wider development space for education. The organic combination of modern information technology and education, that is, the application of artificial intelligence technology in modern education, makes learning an ongoing and lifelong process. Nowadays, the construction of networked courses is on the rise. How to reflect intelligence in the teaching platform is a hot topic of various networked courses. Most of the existing intelligent products are in the stage of experimental and theoretical research, and their degree of intelligence is limited. Therefore, it is of great practical significance to design a network-based artificial intelligence-assisted teaching system. At present, the existing voice analysis software is generally not specially designed for English voice teaching. Teachers need to constantly learn modern educational technology, combine the advantages of existing voice software, establish a humanized visual voice teaching platform, improve their own software operation level and ability, and provide effective guidance and suggestions for students. Although the waveform and spectrum speech analysis software can visually and vividly present speech information, teachers and learners need special knowledge and training to understand and master this kind of visual information, which is difficult. In view of the current problems in English phonetics teaching and the inspiration of multimodal discourse theory, many researchers have been devoted to the research and development of speech visualization software and the research on the practical application effect of speech software. Among them, the most widely used is Praat voice processing software. Praat phonetics software, formerly known as "Praat: doing phonetics by computer," is a cross-platform multifunctional phonetics professional software, which is mainly used for the analysis, annotation, processing, synthesis, and other experiments of digital speech signals. In addition, it generates various language diagrams and text reports at the same time.

After the computer converts the sound into text through processing, it still cannot understand the specific meaning of the converted language. For the feature extraction ability developed automatically by the machine, it is hoped that the machine can adjust its sensory equipment online after being in contact with the outside world for a period of time, making the machine more sensitive to those external stimuli that change significantly. This process can also be considered a tutorless learning process. In this regard, from the perspective of neural networks, unsupervised learning models can also be used to achieve dimensionality reduction (Yang & Zhao, 2021). In general, a self-organizing mapping network maps a high-dimensional vector to a specific position of a low-dimensional matrix. This low-dimensional matrix is constructed according to the principle that input vectors that are very similar will be placed in the same position in the matrix, and vectors that are closer will be placed in adjacent positions (Li & Xiong, 2021). There is a representative vector at each position of the matrix, and the similarity is calculated by the input vector and the representative vector at some positions. When the position closest to the current input vector is found, the representative vectors for this position and its neighbors are recalculated. In this way, the self-organizing map network is updated, and, from the perspective of memory, it summarizes all the input information in the past (Jia & Zhang, 2021).

The research on speech synthesis and speech recognition is to realize human voice generation and perception functions through computers (Liang, 2021). However, the principles and methods of their application are far from the human mechanism. In this research, the author focused on the study of the mechanism of speech generation and perception, especially the information transmission and processing mode of speech generation and perception in the brain. The researcher used the advantages of existing technologies in speech analysis to propose a new speech recognition method that describes

the parameters of speech in the brain perception system and displays them in a graphical form. The author created and described a voice readable mode based on time self-organizing mapping network. On the basis of self-organizing mapping network, the author introduced one kind of time enhancement mechanism to improve the system performance. This method makes up for the fixed spatial topology of the original self-organizing mapping network and the neglect of the time factor which is crucial to the voice signal. Further research is necessary on the effectiveness of applying Praat software to assist English pronunciation teaching from the perspective of contrastive English and Chinese monophones.

LITERATURE REVIEW

Overview of Voice Recognition

Language is an important tool for human beings to convey ideas and communicate information. Voice recognition technology transmits voice data to a machine; the machine automatically translates and encodes the voice data, and then converts it into a corresponding digital signal (Zhou et al., 2021). Sound recognition covers a wide range of disciplines, including many basic disciplines, such as acoustics, mathematical statistics, and artificial intelligence (Chang, 2022). A voice recognition system can improve efficiency by replacing much of humans' similar and repetitive work with machines, such as automatic voice alarm system, voiceprint authentication system, and fault recognition system (Liang, 2021). In some conditions beyond human limits, such as volcanoes, abyss or the bottom of the sea, it is possible to complete some specific tasks through voice recognition systems (Yang et al.,2018; Zhang, J, 2021).

In the 1970s, Dr. Kai-Fu Lee applied the hidden Markov model (HMM) in speech recognition technology; the SphinxE system was born and had a huge impact on the subsequent changes in related technologies. In the 1980s, the research direction of sound recognition technology shifted from isolated word recognition to continuous word recognition, and its key research method also developed from pattern matching to building statistical models (Yan, 2021).

In the early 1990s, many famous large companies such as BM, Apple, AT&T, and NT invested heavily in the practical research of speech recognition system. Speech recognition technology has a good evaluation mechanism, that is, the accuracy of recognition, which was continuously improved in the laboratory research in the middle and late 1990s (Zhang, Y, 2021).

Before 2010, the GMM-HMM (where GMM stands for Gaussian mixture model) was called the best speech recognition model at that time, and Mel frequency cepstral coefficient algorithm and fBank feature extraction algorithm were the most commonly used extraction algorithms for such models. In the study of how to imitate human auditory processes, this article replaces the function of GMM with the function of DNN automatic learning features. The application of deep learning related technology in the HMM can provide the most distinguishing features for it. At the same time, combining DNN- and HMM- related technologies into speech recognition can reduce the recognition error rate of speech recognition system (Zhu, 2021).

Recognition technology is found in various industries in real life, such as security, communication, and mechanical maintenance and other scenarios. Software such as mobile phone voice assistants or automatic voice translations commonly used in people's terminals is to apply voice recognition technology to today's widely used smart phones through technical means, so as to realize the function of mutual communication between humans and machines. The products include Siri developed by Apple and Xiaodu developed by Baidu.

In recent years, the rapid development of artificial intelligence technology has brought speech recognition technology to the focus of researchers again. They are committed to making machines understand human voice commands, and hope to use voice to control the operation of machines. As the key technology of human-computer interaction, voice recognition technology has been fully developed. People have carried out various transformations and optimizations in its entire process, hoping to improve each step, so as to improve the sound recognition in different applications. The

research starts with speech recognition of the simplest small-batch vocabulary, and then gradually moves in-depth to progressively more complex problems.

The Audry system is the earliest speech recognition system that can recognize 10 English digits. By the end of the 1980s, with the improvement of software and hardware configuration, these could provide better basic support for voice recognition technology, and voice recognition technology has gradually become commercialized. Businesses brought more efficient ways, and voice recognition technology was recognized as one of the most far-reaching technologies in the decade after the beginning of the 2000s (Shang & Liu, 2021).

The development of voice recognition tends to be far-field and fusion, and there are still many problems in far-field reliability. For example, in the case of noise interference, it is necessary to rely on technologies such as voice separation (Liu & Qin, 2021; Lu, 2021). New technologies or processing methods are needed to completely solve these problems, so that machines can have a perception ability far beyond that of humans. To achieve this state, related algorithms and common technological advancements in the entire industry chain of sound recognition are required, such as developing more sensitive sensors and creating more powerful chips. Although some progress and breakthroughs have been made in the relative accuracy of the current voice recognition technology, there is still much room for improvement (Nguyen & Hung, 2021).

The rapid development of speech recognition has also led to a series of practical problems. Changes in the environment and noise interference have led to a rapid decline in the recognition rate, which is also the bottleneck for which speech recognition has not really entered the practical application from the laboratory. At present, most speech recognition systems adopt the template matching method, that is, first divide the speech into segments of one frame, then calculate the characteristic parameters of each frame one by one, and then train these parameters with multiple sets of data to form a template for recognition. This method does not make use of the intrinsic characteristics necessary for each classification, but blindly uses the pattern matching of hard quantification of voice. Besides, the processing of pattern matching often ignores the coexistence of voice information and speaker's personal characteristics in the pattern information, so it is difficult to achieve success when the number of speakers is large. Therefore, in this paper the author proposes a new time self-organization model (TSOM) from the perspective of deep learning. On the basis of maintaining the self-organization model (SOM) topology map, the model improves the neighbourhood adjustment rules and adjusts the weights of neurons and their neighbourhoods in time and space. With the development of deep learning, voice recognition technology will be further improved.

Research on Praat-Assisted Phonics Teaching

At present, many scholars have conducted research on Praat software, applied it to assist English supersegmental phonetics teaching, and achieved good teaching results. Li (2020) inspected the application of the voice visualization software Praat in the hyper-segmentation teaching, and explained how to further improve the teaching effect through visual signal feedback. On the premise of analyzing the English pronunciation teaching mode, Zhang (2016) briefly introduced Praat software, the preliminary preparation of applying Praat software to assist English pronunciation teaching, the method of using the software, and the feasibility and advantages of applying the software to error correction and evaluation in English pronunciation teaching. Shang (2016) selected 32 freshmen majoring in English, recorded them with Audit V 1.5, then compared them with standard pronunciation, and made a speech map using Praat voice analysis software. She analyzed the vowel bias of students' pronunciation by analyzing the trend of broadband spectrogram and the resonance peak value, analyzed the consonant bias of students' pronunciation by analyzing the impulse bar on the narrow-band spectrogram, and pointed out the stress bias of students' pronunciation by comparing the narrow-band spectrogram and broadband spectrogram. The purpose iswas to prove the ability of Praat software in English phonetic error analysis.

The research of many scholars shows that, no matter what kind of language, when the Praat software is applied to the phonetic teaching, it will optimize its teaching effect. However, recently, most researchers on English phonetics teaching have applied Praat software to the study of suprasegmental features, and few have applied it to the study of segmental features. Segmental features are also an important part of phonetics teaching, though, and vowels are the basis, key, and difficult points of phonetics. Although there are theories to guide practice, the teaching of monophonic phonetics is still abstract, subjective, and lacks timely feedback. Therefore, it is necessary to carry out further research on the effectiveness of Praat software to assist English pronunciation teaching from the perspective of contrastive English and Chinese monophones.

METHODS

Network Structure of the Time Self-Organization Model

In 1981, Finnish scientist Kohonin put forward the SOM. This neural network reflects a series of characteristics of the biological nervous system, such as the memory mode of brain nerve cells and the excitation law of nerve cells when stimulated.

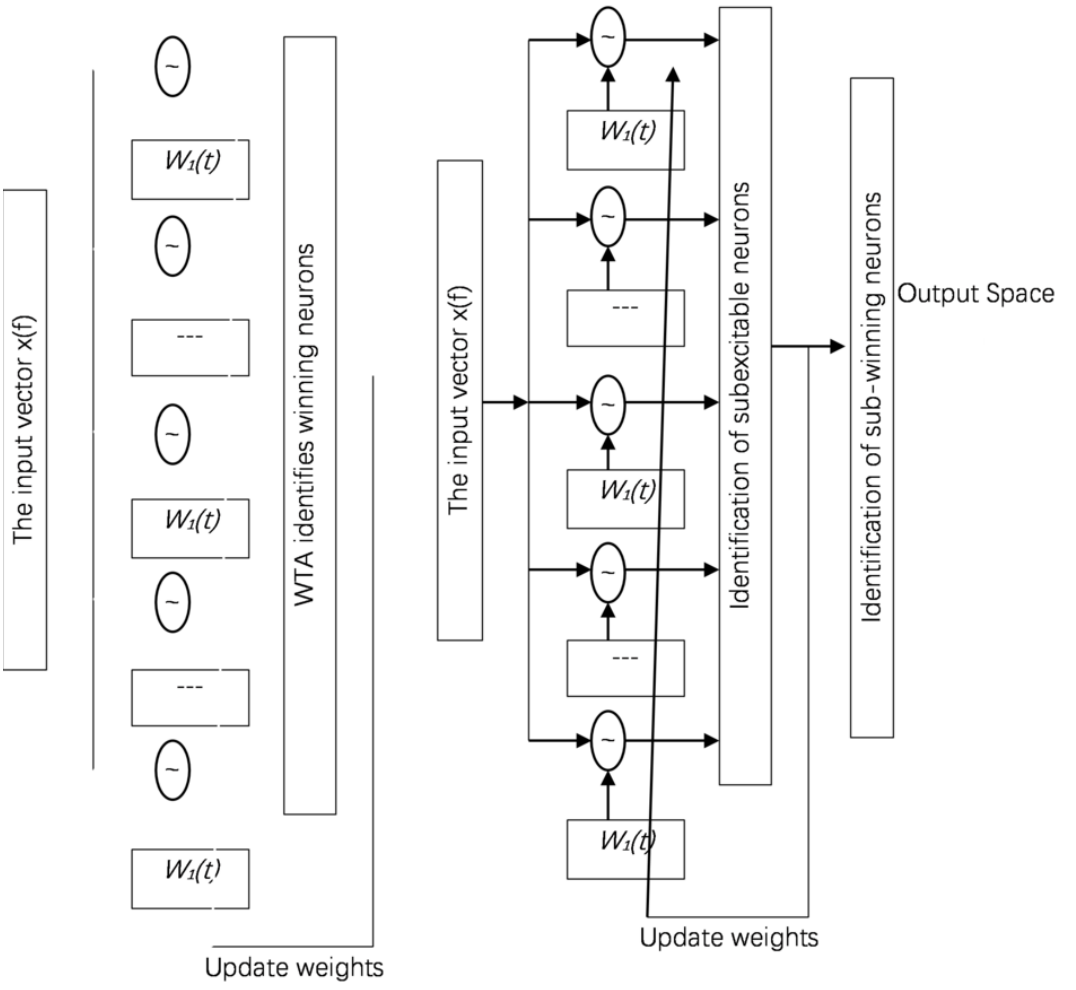
The SOM network mainly has two layers: The input layer and the competition layer. The competition layer can be a two-dimensional plane array composed of M=m2 neurons. Full interconnection is implemented between the neurons in the input layer and the competing layer. Sometimes, there are lateral inhibitory connections between neurons in the competition layer. The SOM algorithm is an unsupervised clustering method, which can map an arbitrary-dimensional input pattern into a one-dimensional or two-dimensional discrete graph at the output layer, and keep its topological structure unchanged (Galante & Piccardo, 2022). In the competition layer, the neurons that win the competition have the highest level of excitability, and the neurons in the surrounding Ng area have different levels of excitability, while the neurons outside the ng area are stimulated to different degrees, forming an excitatory area in the shape of Mexico cap function (Xiao & Park, 2021). The Ng region can also be any other shape, but it is usually uniform and symmetrical, such as a square or hexagon. Ng is a function of time and can also be expressed as *Ng (t)*. The range of *Ng (t)* decreases with the increase of *t* (Liu & Chen, 2021).

In order to overcome the shortcomings of isolated learning and noise sensitivity, the author introduced the time enhancement mechanism based on the space adjustment of self-organizing map, and proposed a new TSOM (Kang & Kermad, 2017). On the basis of preserving the topological mapping, the model improves the neighbourhood adjustment rules, adjusts the weights of neurons and their neighbourhood in time and space, and obtains ideal results. After obtaining the sample input, the model no longer determines a single neuron as the winning neuron according to the winner-take-all (WTA) criterion of SOM, but determines a group of winning neurons in the competition layer with certain rules (evaluation function), called subexcitatory neuron cluster. The evaluation function selects the Euclidean distance and takes the neurons whose evaluation value is higher than a certain threshold as subexcitatory neurons, or takes the first few neurons that are the closest to the current input mode (Wu, 2019). Figure 1 shows the overall structure of the model.

Algorithm Process of the Self-Organization Model and Time Self-Organization Model

Figure 2 illustrates a typical connection between two-dimensional competition layer neurons and input layer neurons. Let the input mode of the network be analog vector $U_k = \left[U_1^k, U_2^k, \cdots, U_n^k\right]^T, k = 1,2,\cdots,p$. The output response of the corresponding competition layer neuron *j* is a digital quantity $V_j, j = 1,2,\cdots,M$. The connection weight vector between the competition layer neuron *j* and the input layer neuron is $W_j = \left[W_{j1}^k, W_{j2}^k, \cdots, W_{jn}^k\right]^T, j = 1,2,\cdots,M$.

Figure 1. Overall structure of SOM and TSOM models

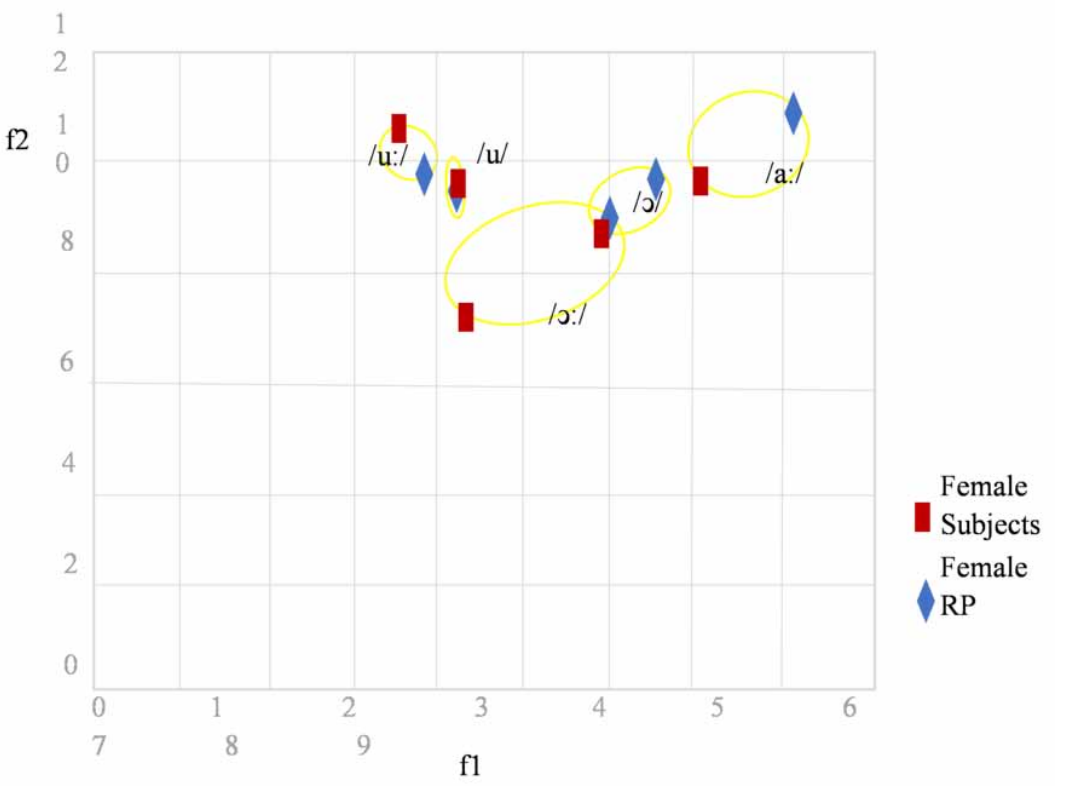


This connection weight vector uses the inner product of the input vector and the connection weight as an evaluation method to distinguish the input vector U and which neuron in the competition layer matches the best. For all $j=1,2, ..., M$, comparing each inner product, the neuron corresponding to the largest value is the winning neuron g in the competition layer. According to the above analysis, the learning work rules of the SOM network can be obtained as follows:

1. Initialize and assign $[W_{ji},]$ to a random value in the [0, 1] interval.
2. Normalize the input mode as follows:

$$U_k = \left[U_1^k, U_2^k, \cdots, U_n^k\right]^T = \frac{U_k}{U^k} \tag{1}$$

Figure 2. The SOM network structure model



$$U^k = \left[(U_1^k)^2, (U_2^k)^2, \cdots, (U_n^k)^2\right]^{\frac{1}{2}} \tag{2}$$

3. Calculate the Euclid distance between the normalized connection weight vector as follows:

$$W_j^{'} = \left[W_{j1}^{'}, W_{j2}^{'}, \cdots, W_{jn}^{'}\right]^T$$

$$d_j = \left[\sum_{i=1}^{n} (U_i^{k'} - W_{jn}^{'})^2\right]^{\frac{1}{2}} \left(j = 1, 2, \cdots, M\right) \tag{3}$$

4. Find the smallest Euclidean distance *dg* and determine the winning neuron *g*:

$$d_g = \min d_j \left(j = 1, 2, \cdots, M\right) \tag{4}$$

5. Mediate the connection weight, trimming the connection weight between all neurons in the competition layer neighborhood *Ng*(*t*) and the input layer as follows:

$$W_{jt}^{'}\left(t+1\right)=\begin{cases} W_{jt}^{'}\left(t\right)+\eta\left(t\right)\left[U_{i}^{k'}-W_{jt}^{'}\left(t\right)\right] j\in Ng\left(t\right) \\ W_{jt}^{'}\left(t\right) \quad \text{other} \end{cases} \tag{5}$$

6. Normalize Equation 5 As follows:

$$W_{j}^{'}\left(t+1\right)=\frac{W_{j}\left(t+1\right)}{W_{j}\left(t+1\right)} \tag{6}$$

7. Add the next mode to the input layer and return to step 2, until all P learning modes are learned once.
8. Update the learning rate $\eta(t)$ as follows:

$$\eta\left(t\right)=\eta_{0}\left(1-\frac{t}{T}\right) \tag{7}$$

$$Ng\left(t\right)=INT\left[Ng\left(0\right)\left(1-\frac{t}{T}\right)\right] \tag{8}$$

9. Let *t=t+1* and return to step 2 until *t=T*. Through the above learning process, the information that needs to be remembered can be saved in the connection right. When another pattern that is the same or similar to the learning pattern is added to the input of the network, the network will find out the output of the competition layer corresponding to the input pattern by "recalling." The process of network recall is to find the neuron *g* of the competition layer. The Euclidean distance between the weight vector connected to the neuron *g* and the input pattern vector should be the shortest. The specific calculation method is as follows:

$$V_{g}=1, if d_{g}=\left[d_{j}\right], j=1,2,L,M \tag{9}$$

$$V_{j}=0,\left(j=1,2,L,M;j\neq g\right) \tag{10}$$

$$d_{j}=\left[\sum\nolimits_{i=1}^{n}U_{i}-W_{ji})^{2}\right]^{\frac{1}{2}} \tag{11}$$

Here, the output of the winning neuron *g* in the competition layer is actually set to 1, and the output of the other competition layer neurons is 0. Neuron *g* represents the classification of its input pattern.

Although the signal of the input pattern is stored in multiple neurons in the *Ng* area during learning, only the neuron with the strongest "response" is found during recall. This does not mean that other neurons in ng do not work for recognition; however, when the input pattern is not the original training pattern, but a pattern similar to the training pattern, other neurons in ng will be found. Especially when this area and one neuron are damaged, other neurons take over its role. Therefore, this network has high stability.

The learning algorithm of TSOM is as follows:

1. Initialize the representative vector matrix *W*, initialize the learning parameter *U*, define the nearest neighbor function *N*, and set *k=0*.
2. Check whether it meets the requirements of the end of the training; if yes, exit, if not, continue.

3. For each input vector *X* used for training, perform steps 4—8.
4. Determine a group of winning neurons (subexcitatory neuron clusters) in the competition layer and set the weight threshold function *Td*; when *di<Td*, the neuron *i* is a subexcitatory neuron, or the first few neurons are taken after sorting *di*.
5. Find the representative vector that best matches the current input and make the $X-\left(W_i k\right)^2$ the smallest.

Update all the nearest neighbor representative vectors within the definition of its neighbor function *Ni(k)* (*i* is the winning representative vector); if *j* is within the definition of *Ni(k)*:

$$W_j\left(k+1\right)=W_i\left(k\right)+U\left(k\right)\left[X\left(k\right)-W_j\left(k\right)\right] \tag{12}$$

If *j* is not within the definition of *Ni(k)*:

$$W_j\left(k+1\right)=W_i\left(k\right) \tag{13}$$

7. Appropriately reduce the learning parameter *U(k)*.
8. Appropriately reduce the range of the nearest neighbor function *N(k)*.

k=k+1; go back to step 2. The network selects the neuron node with the smallest distance from the input signal *x(t)* in the weight vector *wi(t)* as the winning node, and the weight is updated as follows:

$$w_i\left(t+1\right)=w_i\left(t\right)+\varepsilon h_{i,r}(x\left(t\right)-w_i\left(t\right),\forall i \tag{14}$$

where *r* is the neuron that competes to win under the input *x(t)*; ε is the learning gain coefficient, which decreases exponentially with the number of learnings. In other words, under the action of the WTA criterion, the SOM determines the only winning neuron each time, and the adjustment process is also within a certain topological range of the winning neuron. The Euclidean distance between *wi(t)* and *x(t)* after initializing the weights is calculated as follows:

$$d_i=\left[\sum\nolimits_{j=1}^{N}x_j-wi_j)^2\right]^{\frac{1}{2}},i=1,2,\cdots,M \tag{15}$$

The TSOM model introduces a time dynamic process in the space. When the node is in a subexcited state in the output space, the neurons will modulate the connection weight of the surrounding nodes. If a subexcited neuron has not won in the future training, the excitability of the neuron will gradually decline until it disappears. However, if the neuron is reexcited in the subsequent training, the subexcited neuron sequence will be added again and the network will be modulated. Because of this mechanism, it is possible to determine the direction and range of the next winning neuron according to the intensity of neuron excitation. The time dynamic process is introduced into the space to make the model have certain memory function, which can shorten the self-organization process and improve the self-organization effect, especially for the ordered vector input (speech signal processing and recognition). In TSOM, a temporal enhancement mechanism is introduced while adjusting the space. This makes the nodes have a certain memory, which not only has a strong topological mapping effect on the continuous input sequence, but can also fully suppress the interference of noise on training. Human learning also has similar characteristics, and it has the effect of masking noise for continuous

input sequences and enhancing learning. In conclusion, the TSOM uses a new temporally-enhanced kernel function as the neighbourhood adjustment rule for self-organizing learning by introducing a temporally-enhanced mechanism, and the temporally-enhanced neuron nodes have short-term memory. This modifies the WTA competition rules, that is, each learning generates a set of subexcitatory neurons, and the final winning neuron is determined by the subexcitatory neurons.

The computer operating system was Windows 10, the simulation software was MATLAB2020, and the recording software was Adobe Audition1.5. Speech signal acquisition was the first step of speech signal processing. The author collected voice samples using Adobe Audit 1.5 audio editing software. In this test, the researcher applied only the convenient multichannel recording function and flexible waveform browsing and editing function of the software (Yan et al., 2019).

RESULT ANALYSIS AND DISCUSSION

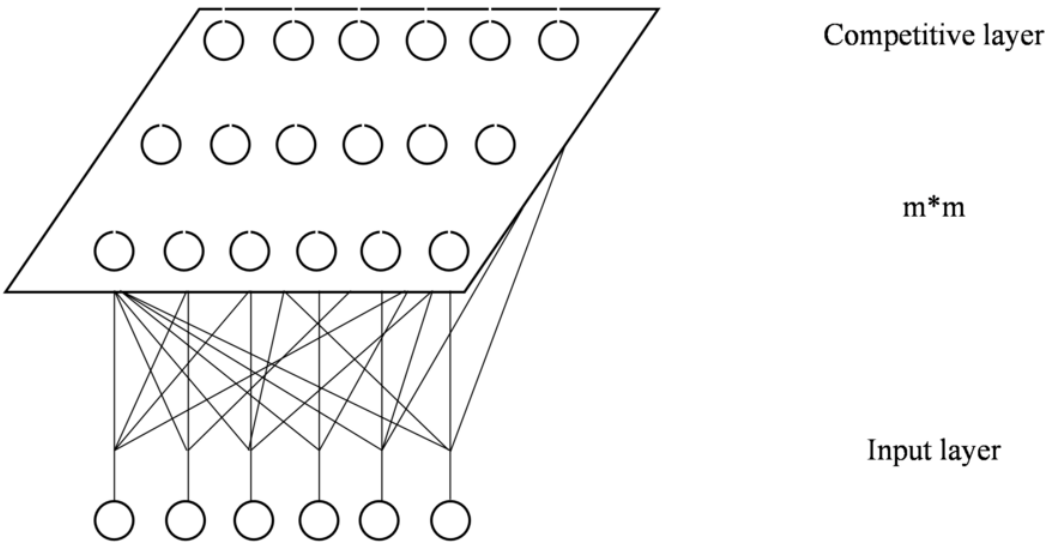
Using Praat to Extract Speech Feature Map

The main function of Praat is to collect, analyse, and label the speech signal of natural language, and perform various processing tasks including transformation and filtering. As the analysis result, Figure 3 shows the text report and language map.

Voice Map of the Time Self-Organization Model

All pattern vectors at the beginning of TSOM have a random value. First, the pattern vector closest to the input vector is found according to the Euclidean distance; then, the best matching pattern vector and the pattern vector of the surrounding position are corrected in the direction of the input vector value; the order of magnitude of the correction gradually decreases after successive input samples. On consecutive inputs, the number of response pattern vectors also decreases intensively around the best matching pattern vector. At the beginning of TSOM, the number of correction pattern vectors around the best matching pattern vector is large, and, as the TSOM process ends, only the nearest neighbors of the best matching pattern vector remain to be corrected. The pattern vector reaches a steady state if the statistical distribution of the input vector no longer changes with time. Clusters

Figure 3. Speech /a/ feature map extracted by Praat

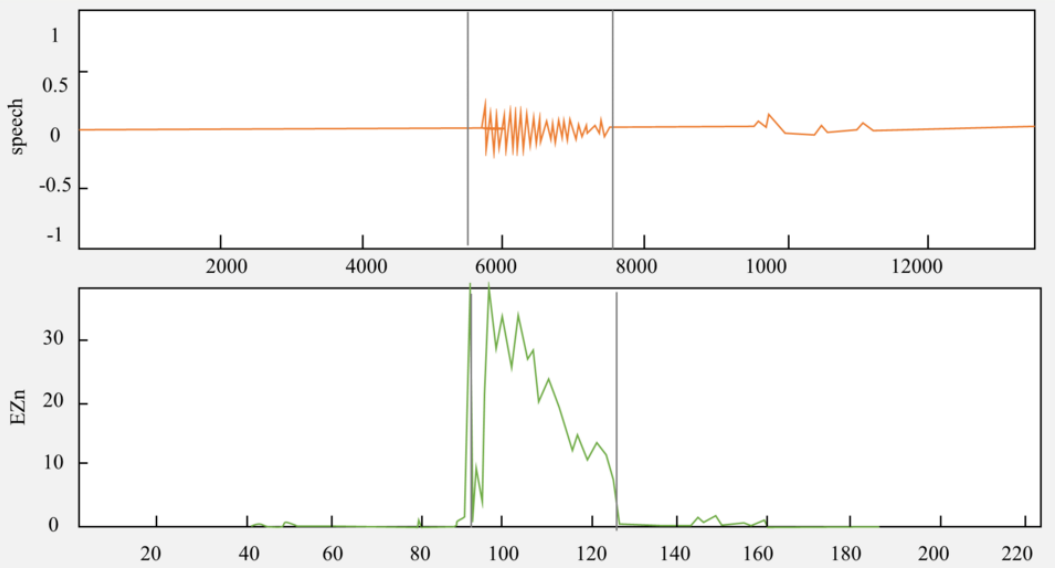


of acoustically similar input pattern vectors are then close to each other, and the centers of these clusters can be marked by phonemes during the TSOM process. Figure 4 shows the TSOM-based voice mapping trajectory diagram.

As for the voice signal itself, the fundamental frequency of women's voice is generally in the range of 100-500Hz, and that of men's voice is in the range of 50-250Hz. According to Nyquist's sampling law and in order to take into account the highest frequency of men, women, and both, the general voice acquisition frequency is 8kHz. However, considering that it is impossible to obtain a pure voice environment, noise and voice will be superimposed in the noise environment, so the sampling frequency of the voice signal used is 11025Hz (Xiwen, 2022). During the test, both the experimental group and the control group were required to put 11 words in the bearing sentences respectively. During the test, both the experimental group and the control group were required to read the load sentences three times. During the recording, if the subjects feel that they have mispronounced a word and need to be stressed, the recorder will give them the opportunity to stress it.

A phonetic mapping trajectory is calculated separately for males and females, and the male and female maps show the projected trajectory of vowels. Acoustically similar vowels, the projections are close to each other and the projections are far away. Vowel projection maps are somewhat similar to F1/F2 formant maps; however, formant maps cannot convey information from the continuum, which is what TSOM voice maps can do. There are also differences between male and female projections because of differences in speech. However, the most obvious difference is that [a] is at a higher position in the female projection and [a] is at a lower position in the male projection, which is a random feature, mainly determined by the random state of the initial pattern vector. At the beginning of TSOM, although males and females looked like mirror images, there was no exact correspondence between the positions of the respective maps, and the two maps represented their respective sets of spectral subspaces. The input vector is compared to the pattern vector, and the best matching pattern vector determines the location of the sample on the map. The mapping of a single spectral vector is shown by a small dot; when several samples map to the same position, the position is marked by a phoneme. Line segments on the map reveal the sequence of consecutive data points, that is, the temporal patterns of the speech signal. The mapping positions are not marked in this paper, and the speech subspace defined by each mapping position depends on the specific instance in the TSOM

Figure 4. TSOM-based phonetic mapping trajectory



process. When the mappings are agreed on instances, the speech perception subspace for each mapping location can be qualitatively identified. In this study, the mapping position in the rectangular grid is given with reference to an arbitrarily chosen position.

In order to make the experimental results clearer, only the length of the speech segment in the map is displayed in the results section. After the pronunciation reaches the field and continues above, the researcher will calculate the following parameters. The track length (L) indicates that there are large changes in the continuous samples, and the number of continuous samples with the same representation (N) indicates the small-scale stability. During 150ms, the average length (S) of displacement and the position of [a] on the map are related to the result of perceptual judgment, and the regularity of speech is reflected through the track map.

Comparison Between Chinese English Learners' Monophonic Pronunciation and British Pronunciation

This experiment takes two parallel classes taught by the same English teacher in the first grade of senior high school as the research object. Through screening, 30 people were selected from each class of the two classes as the data collection objects of the control group and the experimental group, including 15 boys and 15 girls. The English vowels in this study are mainly referred to the sixth edition of Gibson English Phonics Course. The classification standard of Chinese vowel system mainly refers to Lin Tao and Wang Lijia and Wang Lijia (Pan, 2023). The English used for standard reference values refers to the British Received Pronunciation (RP) in the southern part of the United Kingdom. In this experiment, the vowel formants F1 and F2 of RP were used as the standard values for comparison and analysis with the experimental group and the control group. In order to show the differences between Chinese English learners and RP English standard values more comprehensively and find out the problems, and taking into account the acoustic differences between male and female pronunciation, the researcher transformed the data of 30 males and 30 females from all 60 subjects, and drew the vowel acoustic map of male and female subjects compared with RP standard characters.

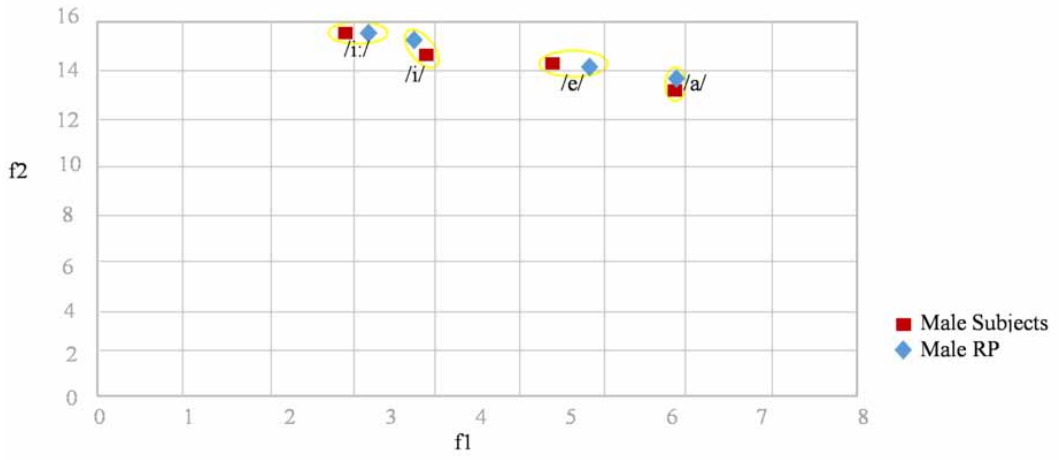
Figure 5 and Figure 6 show that male subjects did not acquire the pronunciation of the four front vowels very correctly. Although the tongue position is very close to the RP pronunciation when pronouncing /i:/, male subjects tended to have a larger mouth opening than RP when they sent /i:/ and /e/, and their tongue position was lower. At /i/, the mouth opening was smaller than that of RP, and the tongue position was higher. In terms of the front and rear of the tongue, the tongue position of male subjects was forward when sending /i/ and tongue, and the tongue position was backward when sending /e/. While female subjects were pronouncing, f2 of /i:/ was almost on the same horizontal line as RP, and f1 at the same time was almost on the same vertical line, indicating that female subjects could better learn /i:/ The tongue position when /a/ is pronounced before and after and the tongue position when /a/ is pronounced. However, it is not ideal in other respects either.

Figure 7 is the acoustic diagram of the central vowel of Pronunciation of male and female subjects. When male subjects pronounce the vowel /ə:/, there is little difference between f1 and RP, and they are almost on the same vertical line. The tongue position is still quite different from RP in front and back. The situation with vowels is even less satisfactory. The pronunciation of central vowels in female subjects showed a similar trend to that in male subjects, but there was a greater gap between female subjects and RP. Specifically, the female subjects had too large mouth opening and too low and forward tongue position during hair, which was much worse than the average of RP. Female subjects also tended to place their tongues further back than RP when issuing /ə:/.

In Figure 9 and Figure 10, the pronunciation of /u/ in male subjects is almost the same as *female*, but the remaining four vowels have major problems, and even they are mistakenly pronounced as RP. Male subjects tended to place their tongues too low when pronouncing back vowels. As for the front and back of the tongue, the male subjects were in front of the tongue when they pronounced /a:/, and three posterior vowels, and the tongue was in the back when they pronounced /u:/. Female subjects were close to RP in acquisition of /u/, but, unlike male subjects, the distribution of /u/ and

Figure 5. Acoustic diagram of the front vowels of male and female: Front vowel acoustic map of male subjects and male RP

Figure 6. Acoustic diagram of the front vowels of male and female: Front vowel acoustic map of female subjects and female RP



RP/u/ almost completely overlapped. Specifically, female subjects were too low on the tongue when pronouncing the four rear vowels except /u/. Except for the tongue position when /u:/ is pronounced, the F2 value of the remaining three vowels is greater than the RP standard value, that is, the tongue position is forward.

The above analysis evidences the comparison between male and female subjects' unit pronunciation and RP English. Male and female subjects have overlapping vowel/u/, central vowel/3:/ and RP pronunciation, and the pronunciation is good. Although the overlapping area of the first vowel/æ/is not as large as that of/u/, /3:/, there is still a small part of the overlap, indicating that Chinese English learners are not poor in pronunciation. However, there is a big gap between the pronunciation of other vowels and that of RP. The first resonance peak F1 is related to the opening degree and the tongue position, that is, the lower the frequency of F1, the lower the opening degree and the higher the tongue position, and vice versa. The second resonance peak F2 is related to the front and back of the tongue position and the round spread of the lip shape, that is, the lower the frequency of F2, the

Figure 7. Post vowel acoustic of male and female pronunciation: Vowel acoustic map of male subjects and male RP

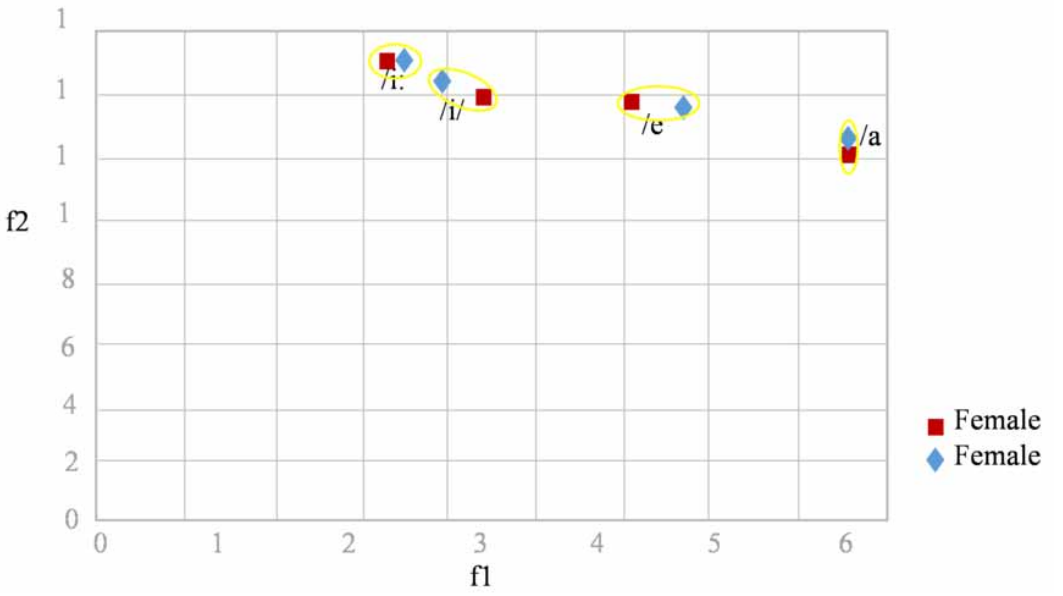
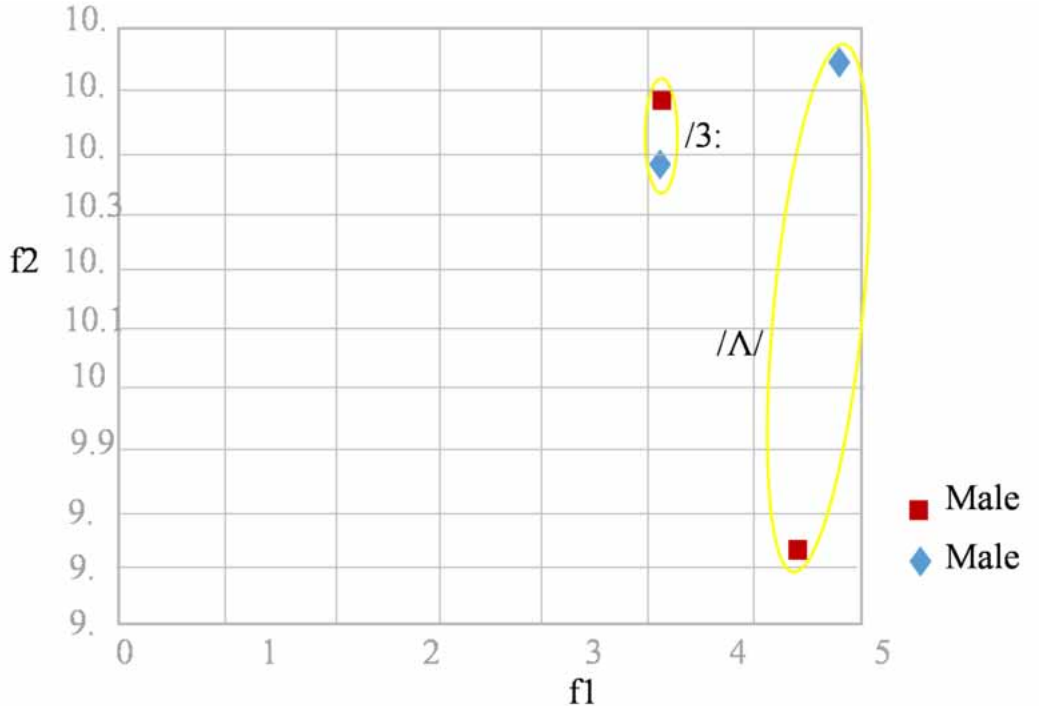


Figure 8. Post vowel acoustic of male and female pronunciation: Vowel acoustic map of female subjects and female RP



later the tongue position, the flatter the lip shape, and vice versa. From the perspective of quantitative analysis, the greater number of people to sample, the more credible the experimental results will be. However, subject to the constraints of objective conditions, the sample selection of this study is

Figure 9. The back vowel acoustics of male and female pronunciation figure: Vowel acoustic map of male subjects and male RP

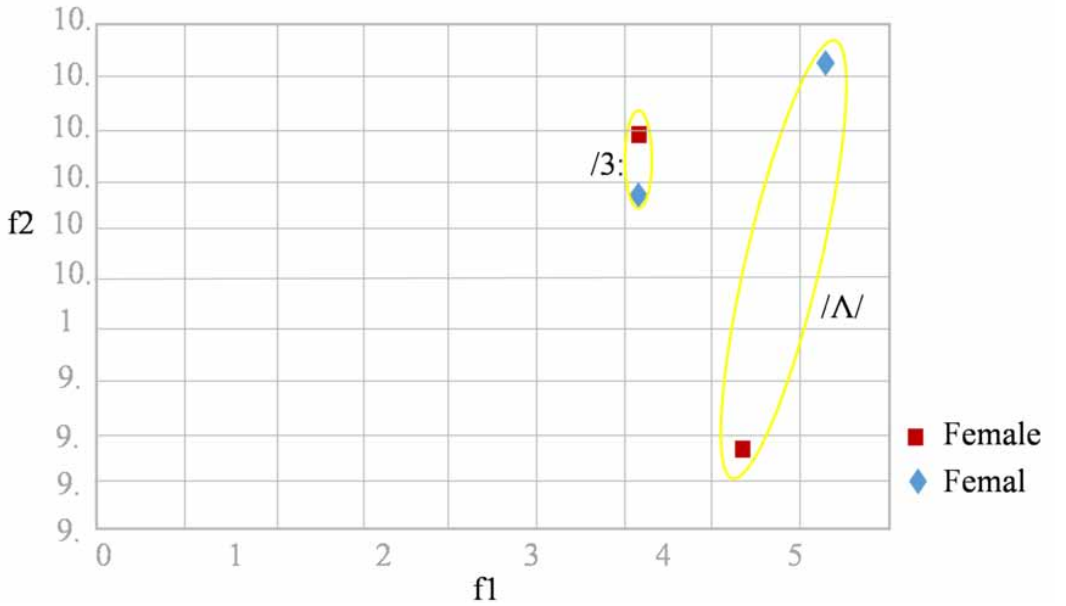
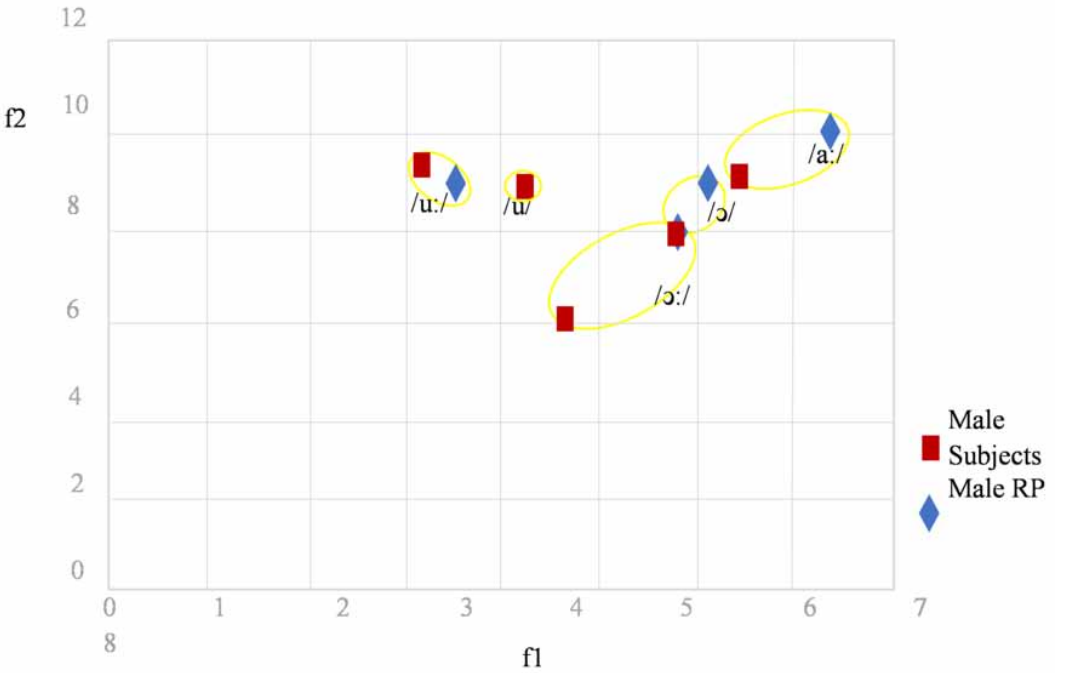


Figure 10. The back vowel acoustics of male and female RP: Vowel acoustic map of female subjects and female RP



only from two classes in a high school, with a small sample size. The experimental group and the control group have only 30 people each. In addition, the sample selection was screened to exclude the regional and dialect interference factors.

CONCLUSION

The purpose of this research was to apply the phonetic visualization Praat software and TSOM model to the study of English monosyllabic teaching from the perspective of English-Chinese monosyllabic comparison. The author established and described a speech recognition model based on time self-organization mapping network, introduced time enhancement mechanism based on self-organization mapping, and combined speech visualization Praat software to improve speech recognition performance, thus improving the effect of English speech teaching.

This study can provide the following suggestions for the practitioners of English pronunciation teaching: First, making reasonable use of the role of mother tongue transfer in effective English teaching, and making full use of the similarities between the two languages to enhance the positive transfer of mother tongue. Second, pronunciation teaching should run through the whole process of English teaching. Learners can use software to assist voice learning, such as Praat software. After the special pronunciation class, pronunciation training should run through the whole English teaching process.

The author plans to further explore two aspects in future work:

1. The essential principle of speech feature extraction and the construction of human auditory organs: On this basis, the author aims to propose an automatic feature extraction algorithm that is closer to the classification hearing process. Of course, it is also possible to consider some manual feature extraction to simulate the evolved feature extraction function of the ear.
2. The development mechanism of human voice and language in the process of growth: The research in this field may involve more fields with which the author is unfamiliar, such as neurobiology, cognitive science, and linguistics.

DATA AVAILABILITY

The figures used to support the findings of this study are included in the article.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

FUNDING STATEMENT

This work was supported by the National Social Science Fund Project: “Research on intelligent protection and development of dong xiang language based on multimodal corpus”, Project No: 20CYY043; and the Doctor initiated fund project: “Research on dong xiang language based on speech acoustic parameter database”, Project No: BS458.

ACKNOWLEDGMENT

The authors would like to show sincere thanks to those techniques who have contributed to this research.

REFERENCES

Chang, Y. (2022). A research proposal on applying Chinese phonetic system in teaching pronunciation of English words to older Chinese EFL adult learners. *Journal of Higher Education Research*, *3*(1), 21–25. doi:10.32629/jher.v3i1.631

Galante, A., & Piccardo, E. (2022). Teaching pronunciation: Toward intelligibility and comprehensibility. *ELT Journal*, *76*(3), 375–386. doi:10.1093/elt/ccab060

Jia, S., & Zhang, X. (2021). Teaching mode of psychology and pedagogy in colleges and universities based on artificial intelligence technology. *Journal of Physics: Conference Series*, *1852*(3), 032033. doi:10.1088/1742-6596/1852/3/032033

Kang, O., & Kermad, A. (2017). Assessment in second language pronunciation. In *The Routledge handbook of contemporary English pronunciation* (pp. 511–526). Routledge. doi:10.4324/9781315145006-32

Li, K. (2020). The application of Praat in English pronunciation teaching. In *Proceedings of the 6th International Conference on Education, Language, Art, and Intercultural Communication (ICELAIC 2019)* (pp. 374—376). Atlantis Press.

Li, P., & Xiong, J. (2021). Research on English teaching process supported by network multimedia technology in higher vocational colleges. *Journal of Physics: Conference Series*, *1915*(4), 042052. doi:10.1088/1742-6596/1915/4/042052

Liang, Y. (2021). Design and implementation of practical teaching management system based on Web in higher vocational colleges. *Journal of Physics: Conference Series*, *1852*(2), 022078. doi:10.1088/1742-6596/1852/2/022078

Liu, C., & Chen, Q. (2021). Current situation and suggested measures of Japanese teaching in colleges and universities based on computer aid. *Journal of Physics: Conference Series*, *1744*(3), 032051. doi:10.1088/1742-6596/1744/3/032051

Liu, Y., & Qin, Y. (2021). The innovation research and practice of the hybrid teaching mode in colleges and universities based on computer technology. *Journal of Physics: Conference Series*, *1744*(4), 042055. doi:10.1088/1742-6596/1744/4/042055

Nguyen, L. T., & Hung, B. P. (2021). Communicative pronunciation teaching: Insights from the Vietnamese tertiary EFL classroom. *System*, *101*, 102573. doi:10.1016/j.system.2021.102573

Pan, H., Li, Z., Tian, C., Wang, L., Fu, Y., Qin, X., & Liu, F. (2023). The Light GBM-based classification algorithm for Chinese characters speech imagery BCI system. *Cognitive Neurodynamics*, *17*(2), 373–384. doi:10.1007/s11571-022-09819-w PMID:37007202

Shang, C. Y. (2016). Research on the application of Praat in English pronunciation class. *Journal of Mudanjiang University*, *25*(4), 4.

Shang, Q., & Liu, Y. (2021). Research on the mode of short video project teaching and media talent training in colleges and universities under 5G industry chain based on multimedia technology. *Journal of Physics: Conference Series*, *1992*(2), 022057. doi:10.1088/1742-6596/1992/2/022057

Wu, Y. (2019). Review of Chinese English learners' prosodic acquisition. *English Language Teaching*, *12*(8), 89–94. doi:10.5539/elt.v12n8p89

Xiao, W., & Park, M. (2021). Using automatic speech recognition to facilitate English pronunciation assessment and learning in an EFL context: Pronunciation error diagnosis and pedagogical implications. *International Journal of Computer-Assisted Language Learning and Teaching*, *11*(3), 74–91. doi:10.4018/IJCALLT.2021070105

Xiwen, Y. (2022). Design of voice recognition acoustic compression system based on neural network. *Wireless Personal Communications*, *127*(3), 2121–2139. doi:10.1007/s11277-021-08773-w

Yan, C. (2021). Innovative research on German education teaching mode in colleges and universities from the perspective of BD. *Journal of Physics: Conference Series*, *1852*(3), 032006. doi:10.1088/1742-6596/1852/3/032006

Yan, Q., Yang, R., & Huang, J. (2019). Detection of speech smoothing on very short clips. *IEEE Transactions on Information Forensics and Security*, *14*(9), 2441–2453. doi:10.1109/TIFS.2019.2900935

Yang, C., & Jin, W. (2018). Chinese as a second language pronunciation teaching survey. *Journal of the National Council of Less Commonly Taught Languages*, *23*, 153–189.

Yang, W., & Zhao, X. (2021). Research on the function of visual phonetic software Praat in vocational English phonetics teaching. *Journal of Physics: Conference Series*, *1856*(1), 012057. doi:10.1088/1742-6596/1856/1/012057

Zhang, J. (2021). Application of big data collection-analysis-visualization in the teaching process of colleges and universities under the background of the epidemic. *Journal of Physics: Conference Series*, *1800*(1), 012009. doi:10.1088/1742-6596/1800/1/012009

Zhang, X. L. (2016). The auxiliary function of Praat in English pronunciation teaching. *Science and Technology Economic Guide*, *2016*(23), 24–25.

Zhang, Y. (2021). Application of computer information processing technology in teaching management information system of colleges and universities. *Journal of Physics: Conference Series*, *1852*(4), 042089. doi:10.1088/1742-6596/1852/4/042089

Zhou, F., Guan, Y., Hu, H., Deng, Y., & Wei, S. (2021). Teaching practice and exploration of cloud class combined with mind map design in biochemistry course in colleges and universities of traditional Chinese medicine in the Internet era. *Journal of Physics: Conference Series*, *1852*(44), 042056. doi:10.1088/1742-6596/1852/4/042056

Zhu, W. (2021). Research on computer assisted English teaching in foreign language teaching in colleges. *Journal of Physics: Conference Series*, *1802*(3), 032069. doi:10.1088/1742-6596/1802/3/032069