Hindawi
Journal of Sensors
Volume 2021, Article ID 3633896, 11 pages
https://doi.org/10.1155/2021/3633896

Research Article

English Pronunciation Standards Based on Multimodal Acoustic Sensors

Lingyi Zhu [1,2]

[1]*School of Foreign Languages, Xinyang Agriculture and Forestry University, Xinyang, Henan 464000, China*
[2]*Office of International Exchange & Cooperation, Xinyang Agriculture and Forestry University, Xinyang, Henan 464000, China*

Correspondence should be addressed to Lingyi Zhu; shuixing521@xyafu.edu.cn

Received 17 August 2021; Revised 27 August 2021; Accepted 28 August 2021; Published 17 September 2021

Academic Editor: Guolong Shi

In recent years, economic globalization is the trend, and communication between countries is getting closer and closer; more and more people begin to pay attention to learning spoken English. The development of computer-aided language learning makes it more convenient for people to learn spoken English; however, the detection and correction of incorrect English pronunciation, which is its core, are still inadequate. In this paper, we propose a multimodal end-to-end English pronunciation error detection and correction model based on audio and video, which does not require phoneme forced alignment of the English pronunciation video signal to be processed, and uses rich audio and video features for English pronunciation error detection, which improves the error detection accuracy to a great extent especially in noisy environments. To address the shortcomings of the current lip feature extraction algorithm which is too complicated and not enough characterization ability, a feature extraction scheme based on the lip opening and closing angle is proposed. The lip syllable frames are obtained by video frame splitting, the syllables are denoised, the key point information of the lips is obtained using a gradient enhancement-based regression tree algorithm, the effects of speaker tilt and movement are removed by scale normalization, and finally, the lip opening and closing angles are calculated using mathematical geometry, and the lip feature values are generated by combining the angle changes.

1. Introduction

Economic globalization is the trend, and the relationship between countries is getting closer and closer, and all countries in the world have in effect become a "global village." English, as the most widely spoken language in the world, is crucial to economic and cultural exchanges between countries, so it is widely learned and used in all countries. The accuracy of English pronunciation determines the smoothness of communication, and incorrect pronunciation of spoken English can easily lead to misunderstandings and unnecessary misunderstandings among interlocutors. The best way to teach spoken English pronunciation is for teachers to teach students English pronunciation and correct their English pronunciation mistakes. However, the economic development of different regions varies, and many regions have a lack of teachers' resources, which leads to the focus of English teaching all on words and grammar, not being able to also teach oral English pronunciation, much less to correct students' wrong English pronunciation one by one [1]. Nowadays, we have entered the information age, and technology has advanced significantly. Computers have entered people's lives and led to the boom of online language teaching, but most online teaching only provides videos of correct spoken English pronunciation but does not check students' English pronunciation and point out their English pronunciation errors. Many phonemes in English have different facial visual characteristics, especially vowels, which can be distinguished by the roundness and tenseness of the lips, so it is of great significance to integrate speech signals and video signals to realize end-to-end multimodal English pronunciation. Therefore, it is of great significance to integrate speech and video signals to achieve end-to-end multimodal English pronunciation error detection.

The multimodal acoustic sensor can accurately extract the characteristic information of English pronunciation

syllables from different modalities and combine the characteristic information of each modality to obtain information that is beneficial for utterance confirmation, which can obtain better recognition performance and improve accuracy, and it has been proved that combining information from multiple modalities can greatly improve recognition performance [2]. Due to the different recognition mechanisms of different acoustic sensors, different syllable modalities are obtained and contain a wide variety of contents. When applied to recognition tasks, multimodal syllables need to be fused to extract features that are common to the same observed object present in multiple modalities, which can improve the correctness of recognition results. The multimodal syllable fusion technique can take multiple information about the same object acquired by multiple sensors and process it in different ways (null-frequency domain conversion, feature extraction, or decision-level determination) to obtain unified information, thus obtaining richer, more accurate, and more reliable information [3]. In this paper, we propose a multimodal end-to-end English pronunciation error detection and correction model based on audio and video, and we propose to analyze the causes of errors in the perception process to reveal the adaptability of different perception methods to the dynamic environment and to build a domain knowledge base for a cognitive dynamic environment. We propose an autonomous perception model based on the listening mechanism and research autonomous English pronunciation recognition methods for intelligent machines.

2. Related Work

The literature [4] investigated the characteristics of common error types in the learning of Dutch by Germans, distinguishing English pronunciation duration, resonance peaks, and fundamental frequency features for different rhymes. Among them, English pronunciation duration is the most characteristic English pronunciation feature. However, for vowels, these features are not well distinguished. Therefore, in the literature [5], ROR correlation features, energy amplitude, overzero rate, and English articulation duration were used as distinguishing features for vowels, and the phonemes were classified by decision trees and linear discriminant analysis. The energy amplitude at a specific position was found to be the most discriminative feature. In [6], a study was conducted on the process of learning English pronunciation of Chinese consonants in Japanese, and the experimental results showed that the delivery of air could be used as a distinguishing feature of consonant English pronunciation. In [7], a study was conducted on the flat and warble consonants in Chinese. It was found that the energy of the flat-tongue was generally concentrated in the higher frequency band, while the energy band of the warble was relatively low. Therefore, the energy band can be used as a distinguishing feature of flat and warble consonants. The literature [8] combined the time, frequency, and location information extracted from EEG signals to classify the features extracted by CNN through Stacked Autoencoder (SAE) of deep network and obtained 90% accuracy. In the literature [9], a 7-layer Long Short-Term Memory Network (LSTM) model was proposed to perform intent recognition on the raw data of EEG signals, and its accuracy could reach 95.53%. The literature [10] improved the LSTM model and then introduced a cascaded and parallel convolutional recurrent neural network model to accurately identify the intended behavior of the human body by efficiently learning the spatio-temporal representation of the raw EEG stream and to capture the spatial correlation between physically adjacent EEG signals by converting the chained EEG sequences into a two-dimensional mesh hierarchy.

The literature [11] treats each dimension of the accelerometer as a channel, just like the RGB channel of an image, and then performs convolution and pooling, respectively. The literature [12] further proposes the problem of sharing the weights of CNNs in multiple sensors by using one-dimensional convolution in the same time window and then adjusting the size of the convolution kernel to obtain the best convolution kernel applicable to the sensor data. A more complex algorithm is designed in the literature [13], which converts the original sensor signal into a two-dimensional signal image using a specific permutation algorithm and Discrete Cosine Transformation (DCT) technique, and then puts the two-dimensional signal image into a two-layer two-dimensional convolution for classification. Converting time series into images, this method exploits the temporal relationship of the sensor data. The literature [14] proposed a network structure consisting of convolutional and LSTM recurrent units that can be applied to multimodal sensors and naturally perform sensor fusion with better performance than CNN. In the same year, the literature [15] designed the Lasagna system that combines CNN and SAE to manage and search motion data in a semantic manner, addressing the deep understanding of arbitrary behavior and the problem of semantic search of arbitrary activities for large amounts of mobile sensor data. The literature [16] proposed a deep sense deep model integrating CNN and recurrent neural network (RNN) to merge local interactions of different sensor modalities into global interactions and extract temporal relationships to model signals for smartphones and embedded devices. The literature [17] optimized the inception structure by incorporating LSTM and proposed OI-LSTM model, which has excellent recognition effect and also the model has good fault tolerance. The literature [18] proposed the recursive multilevel fusion network (RMFN), which decomposes the spatio-temporal fusion problem into multiple stages, each focusing on a subset of multimodal signals for specialized and efficient fusion. The literature [19] proposes a multimodal information fuzzy fusion algorithm. The identification of gait phases was accomplished using multimodal sensory data using a fuzzy fusion algorithm. The literature [20] proposed a 3D point cloud classification method based on multimodal feature fusion, which starts from two perspectives: point cloud features and image features, introduce projection maps as information supplements, and weighted fusion of point cloud features extracted by point CNN and image features extracted by image CNN models to improve the overall classification accuracy of the model. A discriminative multimodal dimensionality reduction method is

proposed in the literature [21], which can seamlessly fuse multimodal data and exploit the potential correlations between different patterns to achieve robust representation learning.

3. English Pronunciation Standards Based on Multimodal Acoustic Sensors

3.1. Multimodal Acoustic Sensor Model. AttnSence is a deep neural network model incorporating a self-attentive mechanism for multimodal human activity recognition and consists of four parts. The first part is a separate feature extraction network for each modality, consisting of a multilayer convolutional network, the second part is a multimodal fusion subnet combined with a self-attentive mechanism, the third part is a bidirectional GRU subnet with a layer of self-attentive network added on top of the hidden layer of GRU states, and the last part is an output layer implemented by a fully connected network and SoftMax. As mentioned before, the input to Aptness is a 3-dimensional tensor X, as shown in Figure 1.

Next, we introduce DT sampling (DT, dense trajectories) methods. Spatio-temporal points of interest have achieved better results in some pedestrian detection and human pose estimation tasks, and dense trajectory sampling (DT) methods were known as the best feature extraction methods before human action recognition methods entered the deep learning field in a big way. The method differs from spatio-temporal interest points in that the DT algorithm uses a strategy of dense extraction of video blocks. As shown in Figure 2, the main steps include a dense sampling of feature points, feature point trajectory tracking, and trajectory-based feature extraction. The DT algorithm densely samples feature points at multiple scales of the image separately using grid partitioning. Sampling on multiple spatial scales ensures that the sampled feature points cover all spatial locations and scales, and usually 8 spatial scales are set. And then, the IDT algorithm was improved by adding a method to eliminate the background light flow, which makes the trajectory tracking more accurate.

In the recognition of multimodal sensor signals, convolutional neural networks have huge applications, and convolutional networks are usually composed of stacked convolutional layers and pooling layers (pooling); the pooling layer is also known as down sampling or under sampling; its main function is to down sample the features and compress the data and the number of parameters to avoid overfitting. Generally speaking, the deeper the convolutional network, the more powerful the model, but the too deep convolutional network is also prone to overfitting, gradient disappearance, and gradient explosion problems, so the choice of the appropriate convolutional network structure has a huge impact on the model performance; in later sections, we will analyze the impact of different convolutional network structure and depth on the performance of the model. In HAR signals, not all modalities can contribute considerably to the activity recognition task; irrelevant modalities are often useless or even hindering to the activity recognition, so a mechanism is needed to prioritize the modalities with important information. To this end, we introduce a multimodal fusion network based on a self-attentive mechanism, as described in the previous subsection, which takes as input the feature representation characteristic matrix $_{Vt}$ of multimodal signals learned by a convolutional neural network and can generate weights corresponding to each modality based on its corresponding input signal quality, which indicates the importance of different sensor modal signals in the HAR task, and then, the network learns by the learned weights are then fed into a weighted average of the feature matrix $_{Vt}$ to compute a uniform feature representation vector ct. The more important modalities will have a greater impact on this feature representation vector, while useless modalities, because of their small weights, will have little impact on the value of the feature vector. The multimodal fusion network can be represented by the following equation.

$$K_v^t = \ln\left(\kappa + {\mu_k}^{\alpha}\right) + l^{\gamma}, \tag{1}$$

$$\mu_k = \lim_{m,n\longrightarrow\infty} \sum_{i,j}^{i=m,j=n} l_i^{\beta} \nu_j^{\varphi} + C, \tag{2}$$

$$R = \int \left(\frac{\arcsin\theta}{\sin^{-1}\theta} \cdot \gamma\right) d\theta. \tag{3}$$

Here, we first transform the K_v^t MLP (multilayer perceptron) to obtain its implicit representation μ_k and then compute the similarity (vector inner product) l with the context vector ν_j (model parameter, representing an abstract concept, obtained by gradient descent learning) and use the computed result as the weight of the first k sensor modality. However, the range of the weights calculated in this way is from negative infinity to positive infinity, so finally, we need to normalize the weights by the SoftMax function to obtain the μ_k range [0, 1]. R represents the fused feature vector of all the modal information, which is obtained by weighting and averaging the eigenvectors and learning weights of each mode. γ and θ are the network parameters which will be initialized during training and learned by gradient descent.

With the multimodal fusion network, we combine all sensor features into feature vectors c_i, and we combine the feature vectors for all moments into a feature matrix, where the C is a time series of.

$$C = (c_1, c_2, c_3, \cdots, c_i). \tag{4}$$

After the feature recognition, filtering, transformation, and fusion of the sensor signal, we can get the time series vector. Although the previous network has extracted the salient features between different modes, it has not captured the temporal features embedded in the time series, so we need to use recurrent neural nets to learn the temporal features. The recurrent neural network is one of the neural networks, and the idea behind it is to make full use of the sequence information. It is called a recurrent neural network because it performs the same computational task cyclically for all the inputs in the sequence, and at each time point,

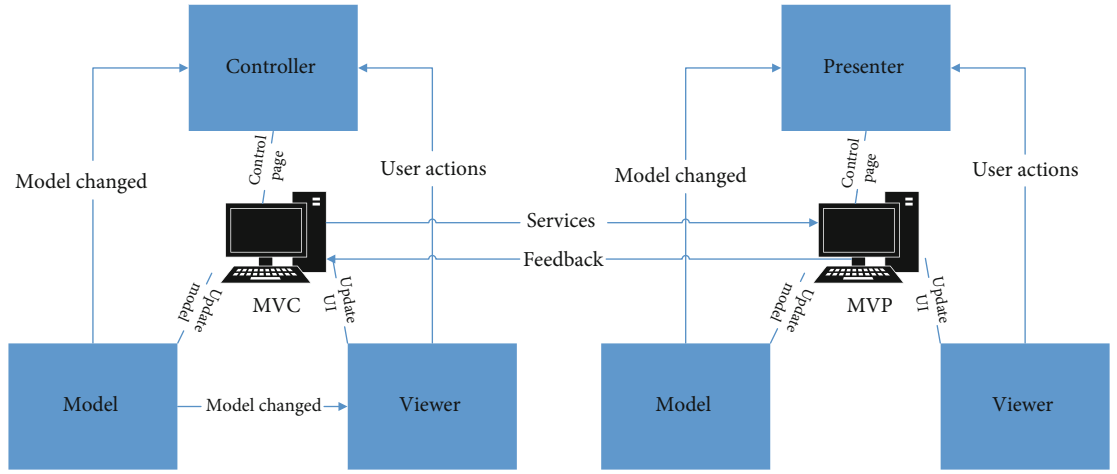

Figure 1: AttnSence model architecture.

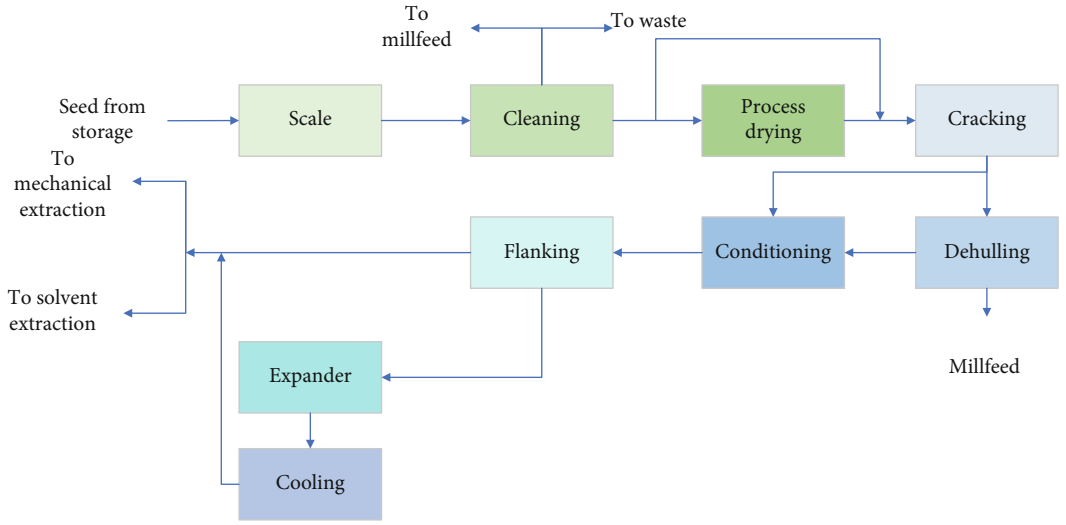

Figure 2: DT algorithm extraction process.

the previously saved hidden layer information and the current input are computed to obtain a piece of new hidden layer information to be used in the subsequent computation, as shown in Figure 3.

4. Design of Multimodal Acoustic Sensors in English Pronunciation

To solve the problem of English pronunciation prediction, a neural network structure was designed to perform the function of English pronunciation prediction by lip movements. The name of this neural network is Gait Trajectory Prediction Network (GTPN), which means English pronunciation prediction network. The main function of GTPN is to compress and extract the key spatio-temporal information from a large number of parameters of lips (such as mouth shape, tongue shape, and movement changes) and finally reconstruct the English pronunciation of this subject from this information; when the English and when the pronunciation state are good, this gesture is also the gestural representation of English pronunciation we need, as shown in Figure 4.

The design scheme of the actuator as the power source of the multimodal acoustic sensor is that to ensure the overall structural strength while reasonably selecting the fabrication materials and structural weight reduction, which has been achieved to reduce the weight of the actuator, and this design causes the actuator to become the main source of system weight at the same time. Excessive system weight will burden the human body and thus reduce the efficiency of recognition. According to the characteristics of human English pronunciation, the power demand of the system is calculated, and a suitable power source method is selected. The reasonably designed Bowden line's retracting and releasing mechanism ensures the efficiency of power transfer and the reliability of the system. The actuator is highly integrated with the following components, frame, protective housing, bobbin cover, control module, and drive module. The drive module in turn consists of a bobbin, flange, motor, motor bracket, and driver.

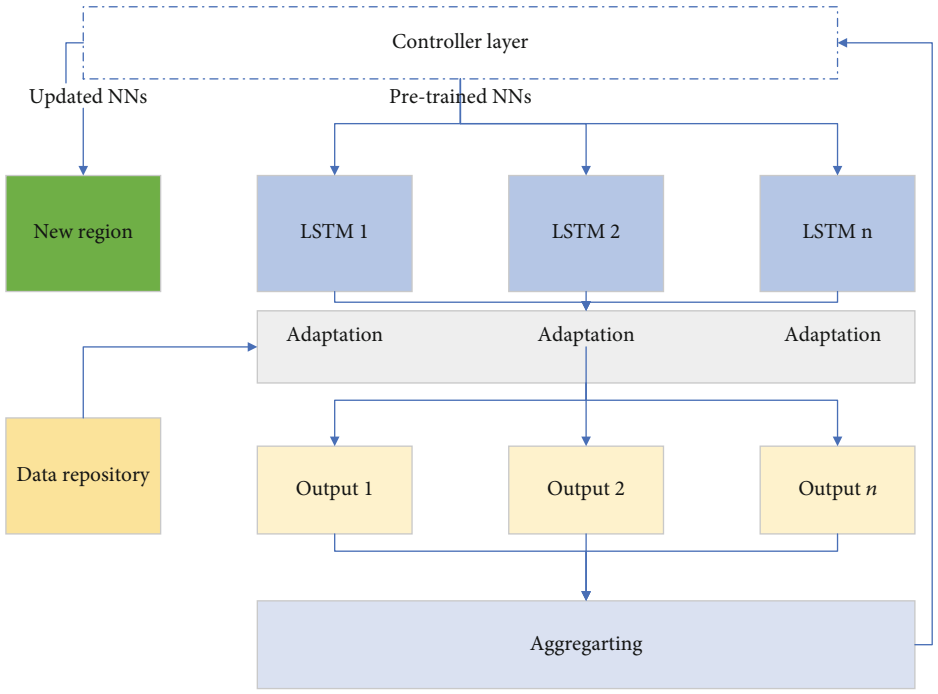

Figure 3: Recursive neural network structure.

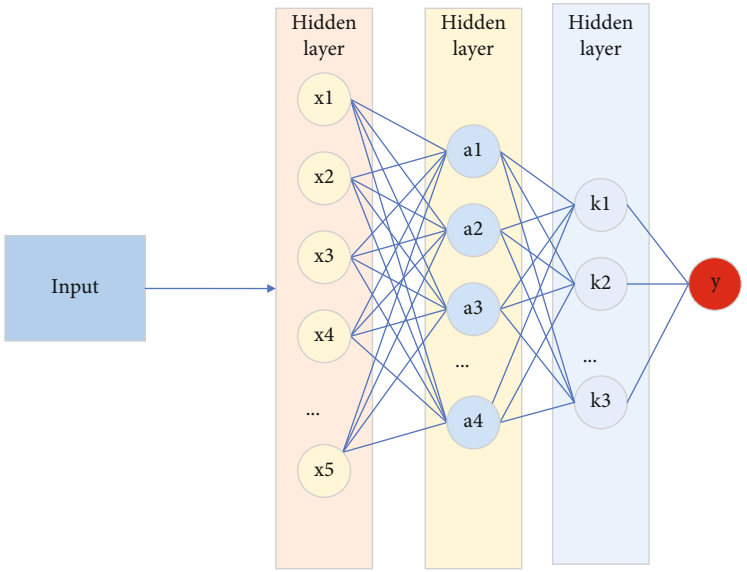

Figure 4: English pronunciation presolver neural network.

The drive module is the core component of the actuator and is the power source of the overall system. To efficiently integrate the Bowden wire with the motor, a bobbin is designed with a 72 mm diameter bobbin reservoir, which is fixed by a flange to the output shaft of the gearbox, for driving and storing the Bowden wire. The bobbin storage tank is used to store the bobbin wound on the bobbin, and the bobbin stopper is used to prevent the bobbins in the storage tank from sliding out of the bobbin. On the other side of the drive module, a motor bracket is designed to connect the motor and the driver, which is a more compact structure of the drive module, and the whole drive module is fixed to the frame by double-end fixing, which disperses the self-weight of the drive module and the reaction force of the booster, avoids the cantilever fixing of the drive module, and increases the stability of the overall mechanism.

To filter out the characteristic factors in the massive information, we can comprehensively process multimodal information from various sources and various forms through information fusion. It is usually divided into three levels: data-level fusion, feature-level fusion, and decision-level fusion. The fusion algorithms include classical

algorithms including Bayesian estimation, great likelihood estimation, D-S evidence theory, Kalman filter, and modern algorithms developed based on artificial intelligence such as cluster analysis, rough set theory, neural networks, support vector machines, and hidden Markov chains. The application of multisensor multimodal information fusion can be used to improve the universality and accuracy of the perception system and enhance the performance of the perception system.

To achieve good recognition results, multimodal English pronunciation detection models must choose suitable audio-video datasets, and the quality of audio-video datasets is decisive for the recognition accuracy. Common audio-video datasets include AVLetters dataset based on alphabetic words, BANCA dataset based on number sequences, GRID dataset based on phrases, and OuluVS dataset based on everyday utterances. The audio-video corpus is scarcer than the unimodal corpus, and most of them are not publicly available to the outside world. The GRID corpus is a sentence-level audio-video corpus that is rarely publicly available and widely used in the field of lip recognition.

We can obtain the corpus in two ways; the first is to download it manually by clicking on the web page, which is a time-consuming and laborious operation. The second method is to use a Python crawler, which makes it easier and faster to download and categorize the corpus files needed for preservation. In this paper, the corpus website is simple and uses Xpath, a language that can query information in HTML files, to obtain information about the target tags. To import request and tree package, the first step is to find all the links, request the page using request, get the source code of the page, and build the object that Xpath can parse using etree. The HTML functions use Xpath technology to filter the tags and eliminate file names that do not need to be downloaded, thus obtaining audio and video file names that need to be downloaded and saving them in the document.

5. Experimental Design and Analysis

The experiments in this chapter mainly use two datasets, WISDM and UCIHAR, where WISDM is an act tracker dataset publicly available from Wireless Sensor Data Mining Lab, which contains 1,098,213 original samples collected using a triaxial accelerometer with a sampling frequency of 20 Hz (20 data points per second), and a total of 36 volunteers were recruited to perform 6 different English pronunciation behaviors. The distribution of behaviors performed by the 36 volunteers is shown in Figure 5, where the horizontal coordinates indicate the volunteer number and the vertical coordinates indicate the number of samples of different behaviors performed by each volunteer. The dataset was operated using the ascending sensor, and the English pronunciation samples were collected by the system. The scenario of the experiment was that the volunteers communicated casually and were identified based on distance, speed of speech, noise, etc. Before each volunteer starts the collection, we will show him an animated demonstration video once, and while he is watching it, we will explain to him the considerations of the collection at the same time. Our staff will make a point to emphasize that the formal collection does not require strict adherence to the steps in the demonstration video. We conduct the first collection by letting the subjects face the camera. However, from the second acquisition onwards, we will tell volunteers not to intentionally face any camera but to start a new movement from the direction where the movement begins. For example, for the second time, you can consider walking while talking on the phone, and for the third time, you can consider sitting on a chair. This is the best way to ensure that our volunteers can interpret the real action. It is important to note that our acquisition instructions are all through the computer playing the voice, and then, the volunteers follow the voice instructions to do the corresponding actions. Six data modalities will be released in this dataset: depth map sequences, infrared images, sound data, RGB video, inertial data, and Wi-Fi-CSI data. The Wi-Fi data can be used for tasks such as identity recognition and human posture estimation.

The labeling of the dataset contains two parts; one is the semantic labeling of the actions, and the other is the timeline segmentation labeling. To ensure the accuracy of the labels, our dataset labeling consists of three stages. Firstly, in the first stage, our staff has recorded the action labels and preliminary timeline segmentation labels when playing voice commands during the volunteer data collection. The second phase is dedicated to labeling the timeline segmentation labels, and we outsource the labeling task to university students for remote labeling. The third stage was to ensure the accuracy of the outsourced annotation, and we checked and corrected all the annotations. Finally, when cleaning the data, we performed another sampling and correction work on the final labels. The Gaussian kernel function was chosen for the SVM. The preprocessed multimodal data are put into the SVM model for recognition so that the multimodal information is fused at the data level, and the intent recognition results are obtained by the multimodal information. Since the preprocessing method of this paper is used, the type of dual operation can be distinguished.

The UCIHAR dataset is a record of X, Y, and Z-axis accelerometer data (linear acceleration) and three-axis gyroscope data (angular velocity) collected from a Samsung Galaxy S I phone, where the acceleration data is broken down into gravity and physical activity components. The dataset consisted of 470528 raw samples, sampled at 50 Hz (50 data points per second), and 30 volunteers between the ages of 19 and 48 years were recruited. The experimental scenario was to place the smartphone at the waist, with each volunteer placing the device on the left- and right-hand side, and to perform two sequential activity sequences consisting of six different behaviors, namely, walking, walking up, walking down, sitting, standing, and lying down (laying). The data distribution of the raw data is shown in Figure 6, from top to bottom, body X, Y, and Z acceleration, body X, Y, and Z angular velocities, and total X, Y, and Z acceleration, with the horizontal coordinates indicating the velocity values and the vertical coordinates indicating the corresponding sample numbers, and it can be seen from the distribution that the

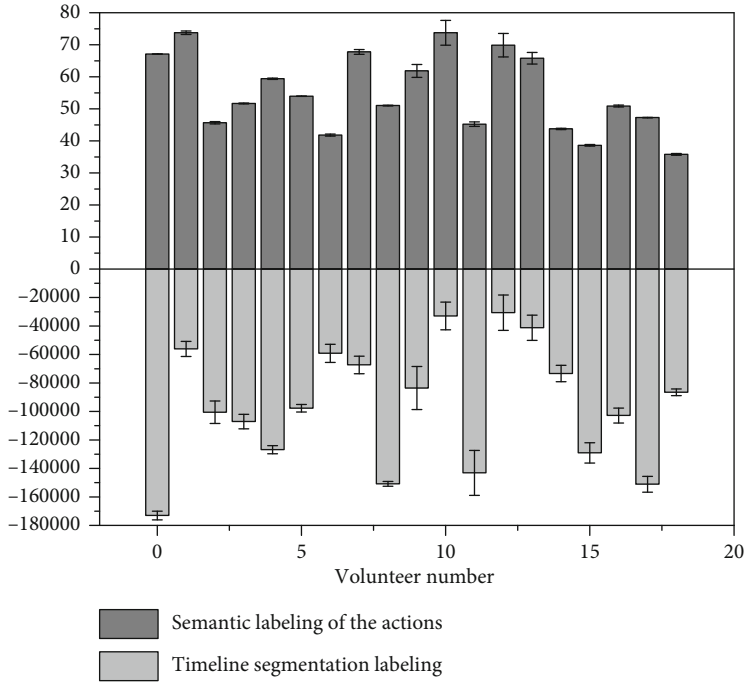

Figure 5: Distribution of sound dataset of WISDM.

features obey a Gaussian distribution except for the total X acceleration.

The experimental comparison results of the CASIA-HFB sound dataset are shown in Figure 7. From Figure 7, it can be obtained that the algorithm in this paper can achieve a higher recognition rate in the NIR and visible datasets compared with other algorithms. Except for the classical PCA algorithm which has a relatively low recognition rate, the recognition rates of SRRS and CLRS are not very good for the NIR and visible datasets, because SRRS and CLRS algorithms are dedicated to solving the situation that the perspective information of the training and test samples is unknown. The other four algorithms based on coupled dictionary learning can achieve recognition rates of more than 98%, which also shows that dictionary learning plays a favorable role in the extraction of recognition features. All the comparison algorithms designed in this paper can handle data from both modalities, and the figure presents the Rank-1 recognition rates of all the algorithms regarding the two perspective combination schemes (Case 1, Case 2, Case 3, Case 4, Case 5, Case 6). As can be seen in Figure 7, the proposed method in this paper outperforms the other nine algorithms for four of the six combination schemes in the case of both modalities. Among all the algorithms compared, the first four algorithms learn the dictionaries corresponding to the two modalities by coupled learning, so they can only process the data of the two modalities. The latter four and the method proposed in this paper can process data of multiple modalities. The data in the figure shows that the recognition results of the first four algorithms are overall higher compared to the latter four algorithms because the dictionaries learned with coupling have greater recognition power for data of two modalities, so the latter four algorithms based on subspace learning will be slightly inferior. The algorithm in this paper uses dictionary learning and subspace projection, so it has stable play in all the combined schemes of the two perspectives. The algorithm in this paper reduces the modal differences through the common subspace before the dictionary learning and feature fusion steps, so it can get a high recognition rate on the scenarios with large viewpoint differences (Case 1 combination scenario). Since the algorithm in this paper mainly solves the problem in the scenario with large modal differences, it has no significant advantage in the scenarios with similar modalities (Case 2 and Case 5). In addition, this experiment also reflects the effect of viewpoint difference on recognition rate. According to the different viewpoint combination schemes and the experimental results, it can be concluded that the recognition rate of the schemes with larger viewpoint differences is also lower, while the combination schemes with smaller viewpoint differences have a higher recognition rate for all methods. This conclusion also confirms that the assumptions made for multimodal data are realistic.

In terms of English pronunciation, PPS may be far more reliable than sensors such as IMU. Although IMU is far more accurate and continuous than PPS for articulatory gesture detection, IMU detects angles with few features, and the program can only detect special events at curve turning points or locations with large change rates, while other locations are relatively stable and have few feature changes, making it difficult to detect critical articulatory gesture events. For PPS, there are abundant and obvious force interaction events in each gait cycle, and these time points generally correspond to some key articulatory gestures that are very important for the booster-type flexible lip form. The detection of these two events, HS and TO, is difficult to achieve

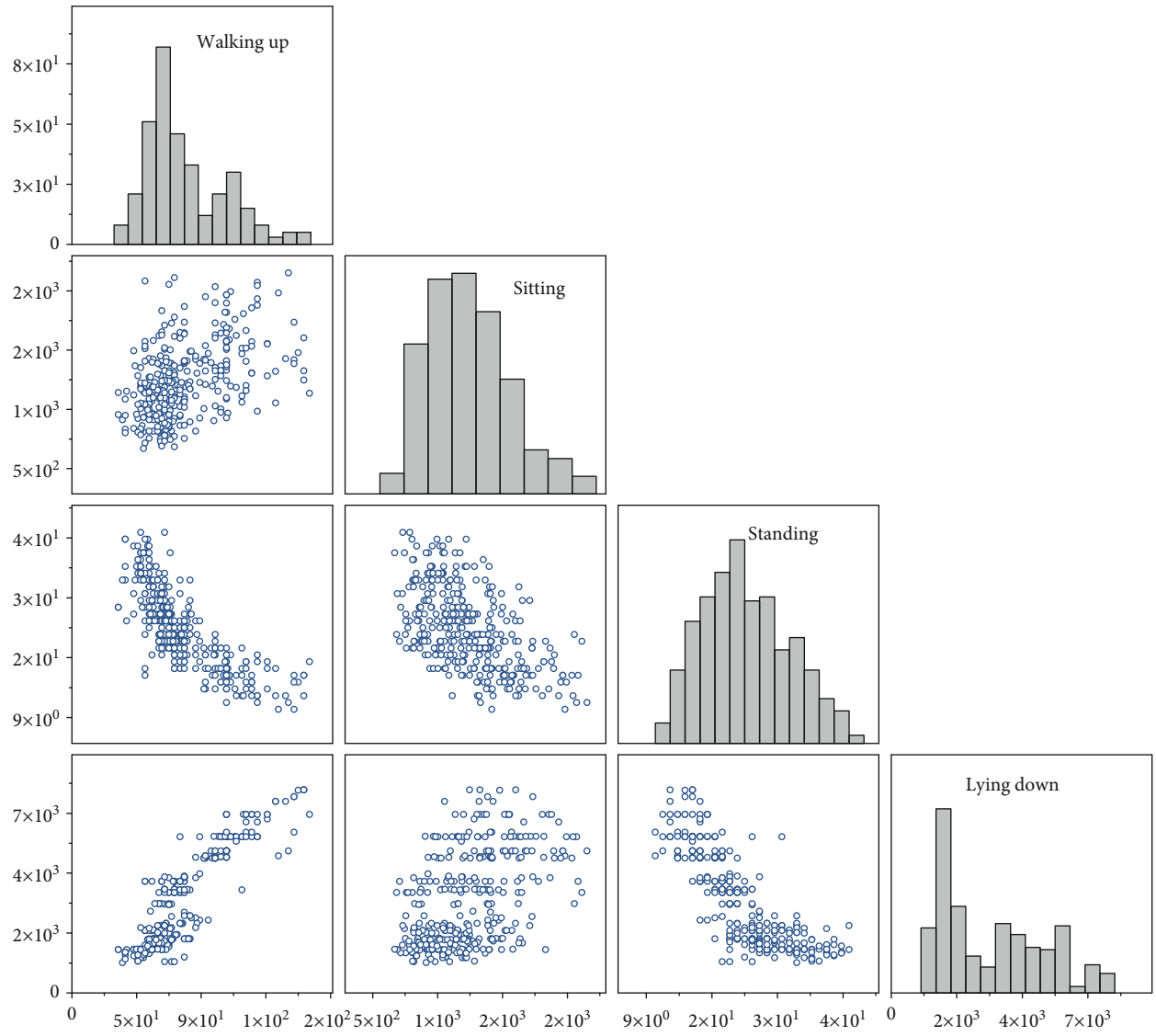

Figure 6: Distribution of behavioral data of UCIHAR.

because the IMU readings do not have very obvious detectable features [21]. Instead, we used PPS sensors encapsulated in a mask that the wearer wears on the face to detect pressure changes. Each mask contains eight force-sensitive units (FSRs), and we use three in the front and three in the back for sensing, with the three FSRs in the front placed below the lips and the three FSRs in the back placed below the lips to allow more complete sensing of contact force changes. The first used here have a high precision thin film pressure-sensitive material, so the pressure detection time delay is only about 1 ms, and the real-time performance is very high. The two lip angle data obtained by the IMU are difficult to obtain trigger judgments for events such as TO and HS, where the points before and after these events are very similar and the curves are relatively smooth, making it difficult to determine the characteristics of the trigger points and therefore to detect these events by the IMU. At the same time, by comparing the results of IMU and PPS, we can also find that the rate of change of the lip angle before and after the HS event is large, indicating that the motion state before and after the gesture is not stable, and there is a change in the direction of motion and other motion that may lead to drastic changes in acceleration, and the estimation of the angle by the accelerometer is relatively poor at this time. The rate of change of the lip angle before and after the TO event is relatively stable, which indicates that the motion is less violent and the acceleration of the motion is smaller, and the estimation of the lip angle by the accelerometer should be relatively more accurate at this time.

The training process of speech recognition for speech modality, multimodality based on keypoint location fusion, and multimodality based on angular feature fusion under noiseless conditions is shown in Figure 8. As the number of model iterations increases, the loss of the training set keeps decreasing, and the training accuracy keeps improving. Among these three schemes, the angular feature fusion and speech unimodal speech recognition converge faster and converge at about the 150th round. The multimodal

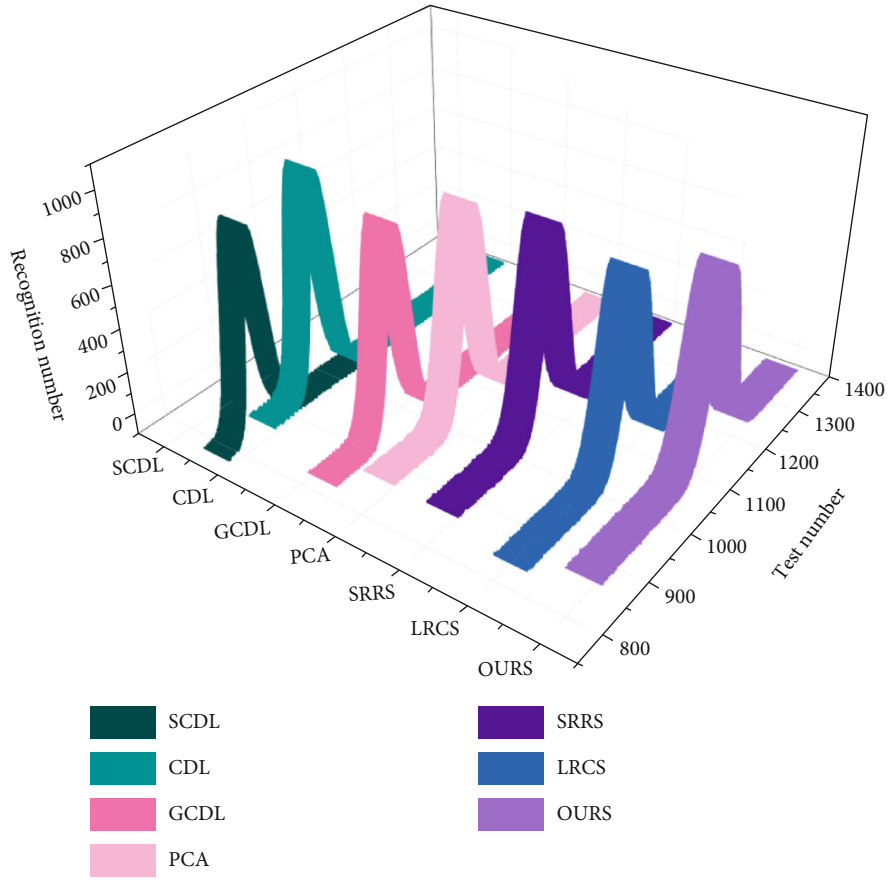

Figure 7: Rank-1 recognition rate (%) of CASIA-HFB sound dataset.

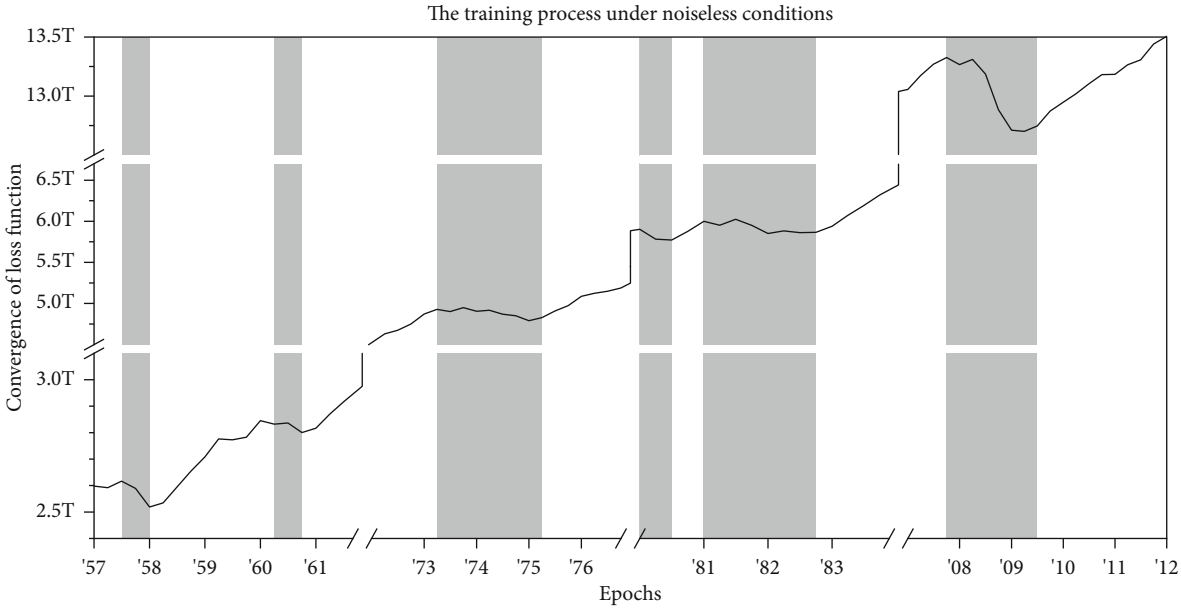

Figure 8: The training process of recognition of multimodal acoustic sensors under noiseless conditions.

fusion based on keypoint features converges more slowly and converges gradually only in the 250th round. In the noise-free case, there is little difference in the correct recognition rate between the three, and the recognition effect of speech unimodal and angular feature fusion is slightly better than that of keypoint location fusion. In terms of convergence speed, the keypoint location-based feature fusion is slightly inferior to the other two approaches. The analysis

shows that the feature dimension of the key point is too large and adds too much redundant information, while the speech modality plays a greater role in recognition in the noiseless case. Therefore, in the noiseless case, both the unimodal recognition rate and the angular feature fusion with lower dimensionality have better recognition rates. After the model training is mature, this paper brings the audio-video corpus of mispronunciations into the above-trained model to further measure the error detection accuracy of pronunciation error detection and correction.

To avoid the temporal desynchronization of audio and video information, a decision-based fusion BiLSTM-CTC pronunciation error detection model is constructed, which performs a decision-level fusion of audio and video features through the recognition results of the BiLSTM-CTC model and finally outputs the phoneme sequences, eliminating the steps of audio and video feature alignment and phoneme forced alignment.

6. Conclusion

Pronunciation check and error correction (MDD) are the cores of multimodal acoustic sensor English speech recognition. With the widespread application of deep learning technology, English spoken pronunciation error detection and correction has made great progress, but there are still some limitations. Many MDD studies focus on the results of mispronunciation and ignore the causes of speakers' mispronunciation, which makes it impossible to provide corrective advice from pronunciation actions; MDD studies are conducted based on voice unimodality, ignoring the importance of lip features during pronunciation; most pronunciation detection and error correction studies ignore the influence of noise on detection results.

The end-to-end English spoken pronunciation error detection algorithm based on multimodal acoustic sensors proposed in this paper is effective in both phoneme sequence recognition and error detection rate and possesses a certain degree of noise immunity. The previous pronunciation error detection and correction problems are investigated in-depth, and we seek to solve the related problems. The pronunciation mode and position of vowel consonants are summarized, and the correspondence between vowel and lip rounding is proposed to provide theoretical support for correcting the pronunciation mode, and common audio-video extraction methods and deep learning algorithms are studied. This paper also constructs a GRID corpus based on phoneme annotation and records an error detection test set for lip vowels by itself, analyzes common rounded-lip and spread-lip word pronunciation errors, preprocesses audio and video separately, and compares video feature extraction methods to select suitable extraction methods, gives a detailed introduction to the extraction process of two modal information, and finally introduces the feature-level fusion and decision fusion features and principles. Based on the multimodal feature fusion model of lip angle features, a video feature extraction method with lower dimensionality but accurate characterization of lip information is proposed, the interpolation and alignment work before the modal cascade is introduced, the principle of model training is presented, and the experiments verify that the model has higher recognition and error detection rate for phoneme sequences and possesses certain noise immunity. To address the drawback that feature-level fusion requires audio-video timing alignment, we construct a decision fusion model that does not require audio-video timing alignment, introduce the principles of weighted coefficient fusion and Dempster-Shafer decision theory, compare the two decision fusion models, and finally suggest corrections for mispronounced phonemes based on articulation mode and articulation position.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The author declares that there are no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] R. Wielgat, R. Jędryka, A. Lorenc, Ł. Mik, and D. Król, "POLEMAD-a database for the multimodal analysis of Polish pronunciation," *Speech Communication*, vol. 127, pp. 29–42, 2021.

[2] S. U. Maheswari, A. Shahina, R. Rishickesh, and A. N. Khan, "A study on the impact of Lombard effect on recognition of Hindi syllabic units using CNN based multimodal ASR systems," *Archives of Acoustics*, vol. 45, no. 3, pp. 419–431, 2020.

[3] L. Wu, J. Yang, M. Zhou, Y. Chen, and Q. Wang, "LVID: a multimodal biometrics authentication system on smartphones," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1572–1585, 2020.

[4] D. Psaltos, K. Chappie, F. I. Karahanoglu et al., "Multimodal wearable sensors to measure gait and voice," *Digital biomarkers*, vol. 3, no. 3, pp. 133–144, 2020.

[5] M. Abuhamad, A. Abusnaina, D. Nyang, and D. Mohaisen, "Sensor-based continuous authentication of smartphones' users using behavioral biometrics: a contemporary survey," *IEEE Internet of Things Journal*, vol. 8, no. 1, pp. 65–84, 2021.

[6] W. Pouw, J. P. Trujillo, and J. A. Dixon, "The quantification of gesture–speech synchrony: a tutorial and validation of multimodal data acquisition using device-based and video-based motion tracking," *Behavior Research Methods*, vol. 52, no. 2, pp. 723–740, 2020.

[7] J. Trouvain and B. Möbius, "Speech synthesis: text-to-speech conversion and artificial voices," in *Handbook of the Changing World Language Map*, pp. 3837–3851, Springer, 2020.

[8] L. D. Rosenblum and J. Dorsi, "Primacy of multimodal speech perception for the brain and science," in *The Handbook of Speech Perception*, pp. 28–57, Wiley, 2021.

[9] M. Jarosz, P. Nawrocki, B. Śnieżyński, and B. Indurkhya, "Multi-platform intelligent system for multimodal human-computer interaction," *Computing and Informatics*, vol. 40, no. 1, pp. 83–103, 2021.

[10] N. Elouali, "Time well spent with multimodal mobile interactions," *Journal on Multimodal User Interfaces*, vol. 13, no. 4, pp. 395–404, 2019.

[11] R. Han, Z. Q. Feng, J. L. Tian, X. Fan, X. H. Yang, and Q. B. Guo, "An intelligent navigation experimental system based on multi-mode fusion," *Virtual Reality & Intelligent Hardware*, vol. 2, no. 4, pp. 345–353, 2020.

[12] K. J. Heaton, J. R. Williamson, A. C. Lammert et al., "Predicting changes in performance due to cognitive fatigue: a multimodal approach based on speech motor coordination and electrodermal activity," *The Clinical Neuropsychologist*, vol. 34, no. 6, pp. 1190–1214, 2020.

[13] Z. Huang, J. Epps, D. Joachim, and V. Sethu, "Natural language processing methods for acoustic and landmark event-based features in speech-based depression detection," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 435–448, 2020.

[14] M. Cohen and W. L. Martens, "Spatial soundscape superposition, part II: signals and systems," *Acoustical Science and Technology*, vol. 41, no. 1, pp. 297–307, 2020.

[15] T. Yoshinaga, K. Nozaki, and S. Wada, "Aeroacoustic analysis on individual characteristics in sibilant fricative production," *The Journal of the Acoustical Society of America*, vol. 146, no. 2, pp. 1239–1251, 2019.

[16] V. S. Nallanthighal, Z. Mostaani, A. Härmä, H. Strik, and M. Magimai-Doss, "Deep learning architectures for estimating breathing signal and respiratory parameters from speech recordings," *Neural Networks*, vol. 141, pp. 211–224, 2021.

[17] J. L. Saunders and M. Wehr, "Mice can learn phonetic categories," *The Journal of the Acoustical Society of America*, vol. 145, no. 3, pp. 1168–1177, 2019.

[18] U. Kumaran, S. Radha Rammohan, S. M. Nagarajan, and A. Prathik, "Fusion of mel and gammatone frequency cepstral coefficients for speech emotion recognition using deep C-RNN," *International Journal of Speech Technology*, vol. 24, no. 2, pp. 303–314, 2021.

[19] S. Li, W. Gu, L. Liu, and P. Tang, "The role of voice quality in Mandarin sarcastic speech: an acoustic and electroglottographic study," *Journal of Speech, Language, and Hearing Research*, vol. 63, no. 8, pp. 2578–2588, 2020.

[20] R. Haeb-Umbach, J. Heymann, L. Drude, S. Watanabe, M. Delcroix, and T. Nakatani, "Far-field automatic speech recognition," *Proceedings of the IEEE*, vol. 109, no. 2, pp. 124–148, 2021.

[21] K. Grabowski, A. Rynkiewicz, A. Lassalle et al., "Emotional expression in psychiatric conditions: new technology for clinicians," *Psychiatry and Clinical Neurosciences*, vol. 73, no. 2, pp. 50–62, 2019.