HEARING AID SYSTEM BASED ON THE FUSION OF VISION AND HEARING

Field of the Invention

The present invention relates to the field of hearing aid technology, specifically a hearing aid system based on the fusion of vision and hearing.

Background to the Invention

Google Translate Glasses is an intelligent wearable device that looks similar to regular glasses and integrates Gemini artificial intelligence assistant. It has real-time translation of 24 languages and sign language functions, can scan books and display online reviews, and achieve interaction through microphones, cameras, etc. For business people, Google Translate Glasses can assist them in better understanding partners' viewpoints and accurately conveying their intentions in international conferences or cross-border negotiations. Whether it is English, French, Spanish, other major international languages, or even some niche languages, accurate translation can be achieved to avoid misunderstandings caused by language barriers. For hearing-impaired people, Google Translate Glass is of great significance in communication scenarios. It can convert sign language into text or text into sign language, allowing hearing-impaired individuals to communicate with others with normal hearing. For example, in family gatherings, school classrooms, or work meetings, people with hearing impairments can better integrate into the communication environment and express their ideas and needs with the help of this glasses.

Intelligent voice interaction is an important component of Google Translate Glass. For hearing-impaired individuals, the intelligent voice interaction system can accurately generate speech text for display after recognizing speech, to assist hearing-impaired individuals in recognizing speech interaction content. However, in complex environments, the intelligent voice interaction system is easily affected by environmental factors, which can affect the accuracy of speech recognition. Therefore, a hearing aid system based on

the fusion of vision and hearing is proposed to solve this problem.

Statement of Invention

（1） Technical problems solved

In response to the shortcomings of existing technology, the present invention provides a hearing aid system based on the fusion of vision and hearing, which combines speech recognition and lip recognition. The system can obtain more information to supplement and verify the speech recognition results, thereby improving the overall recognition accuracy, providing more accurate hearing aid assistance for hearing-impaired people, enabling them to clearly understand relevant information, and improving service quality and efficiency.

（2） Technical solution

To achieve the above objectives, the present invention provides the following technical solution: a hearing aid system based on the fusion of vision and hearing, comprising an environment monitoring module, a multidimensional recognition module, and an intelligent generation module;

The environmental monitoring module consists of a noise monitoring unit, an electromagnetic monitoring unit, and an environmental analysis unit, the noise monitoring unit is used to obtain noise data of the current environment and number the obtained current environmental noise data to form a noise dataset, the electromagnetic monitoring unit is used to obtain electromagnetic data of the current environment and number the obtained current environmental electromagnetic data to form an electromagnetic dataset, the noise monitoring unit and electromagnetic monitoring unit send the numbered noise dataset and electromagnetic dataset to the environmental analysis unit through the network, the environmental analysis unit calculates the environmental impact index based on the received dataset and determines whether the current environment will affect the speech recognition effect according to the environmental impact index, if it is determined that the current environment will affect the speech recognition effect, it sends an

environmental anomaly signal to the multi-dimensional recognition module;

The multidimensional recognition module includes a speech recognition unit and a lip recognition unit, the speech recognition unit is used to generate speech recognition text based on speech recognition, the lip recognition unit is used to generate fused recognition text based on lip recognition combined with speech recognition after receiving abnormal signals, the speech recognition unit and lip recognition unit send the generated recognition text to the intelligent generation module;

The intelligent generation module is used to display the final text.

Preferably, the noise monitoring unit is connected to a microphone through a network and obtains noise data through the microphone.

Preferably, the electromagnetic monitoring unit is connected to an electromagnetic environment monitoring instrument through a network and obtains electromagnetic data through the electromagnetic environment monitoring instrument.

Preferably, the expression for the number of the noise dataset is: Zs_1、Zs_2、Zs_3、…、Zs_n, where Zs_1 represents the first noise data obtained through the microphone, Zs_n represents the last noise data obtained through the microphone, and n represents the total number of data in the dataset.

Preferably, the expression for the number of the electromagnetic dataset is: Cs_1、Cs_2、Cs_3、…、Cs_n, where Cs_1 represents the first electromagnetic data obtained through an electromagnetic environment monitoring instrument, Cs_n represents the last electromagnetic data obtained through an electromagnetic environment monitoring instrument, n represents the total number of data in the dataset, and the data acquisition time of the noise dataset is consistent with that of the electromagnetic dataset.

Preferably, the calculation formula for the environmental impact index $HJyx$ is:

$$HJyx = \alpha_1 * \frac{Zs_i}{Zs_o} + \alpha_2 * \frac{Cs_i}{Cs_o}$$

In the calculation formula, $HJyx$ represents the environmental impact index, Zs_i represents the i-th noise data, Cs_i represents the i-th electromagnetic data,

Zs_orepresents the maximum noise interference data allowed by the speech recognition system, Cs_o represents the maximum electromagnetic interference data allowed by the speech recognition system, α_1 represents the weight of noise factors, and α_2 represents the weight of electromagnetic factors.

Preferably, when the calculated value of the environmental impact index is greater than the environmental impact threshold, it represents that the current environment will affect the speech recognition effect, and sends an abnormal signal to the multidimensional recognition module.

Preferably, the speech recognition text generation method is:

S1.1 the speech recognition unit processes the speech signal into frames based on the speech recognition system, and extracts D-dimensional acoustic features from each frame to obtain a feature sequence. The expression of the acoustic feature sequence is:

$$X = [X_1、X_2、X_3、\cdots、X_t]$$

In the expression, X represents the acoustic feature sequence, X_1 represents the D-dimensional feature vector of frame 1, X_t represents the D-dimensional feature vector of frame t, and t represents the number of time frames.

S1.2 the speech recognition unit calculates the conditional probability of the acoustic feature sequence on the candidate phoneme sequence through an acoustic model, and selects the text sequence with the highest probability as the generation result, the calculation formula for the probability of the acoustic feature is:

$$S_1 = \prod_{i=1}^{t} S(X_i|Q_i)$$

In the calculation formula, S_1 represents the probability of acoustic features, X_i represents the D-dimensional feature vector of the i-th frame, and Q_i represents the phoneme state corresponding to the i-th frame; $\prod_{i=1}^{t} S(X_i|Q_i)$ represents the multiplication of all probabilities from frame 1 to frame t.

Preferably, the method for generating fused recognition text is:

S2.1 the lip recognition unit uses face detection technology to locate the lip area, extracts the D-dimensional visual feature vector of each lip image frame, and obtains a visual feature sequence, the expression of the visual feature sequence is:

$$V = [V_1、V_2、V_3、\cdots、V_t]$$

In the expression, V represents the visual feature sequence, V_1 represents the D-dimensional visual feature vector of frame 1, $V_t$represents the D-dimensional visual feature vector of frame t, and t represents the number of time frames.

S2.2 the lip recognition unit calculates the conditional probability of visual features on candidate phoneme sequences through a visual acoustic model, and the probability calculation formula for the visual features is:

$$S_2 = \prod_{i=1}^{t} S(V_i|Q_i)$$

In the calculation formula, S_2 represents the probability of visual features, V_i represents the D-dimensional feature vector of the frame, and Q_i represents the phoneme state corresponding to the frame; $\prod_{i=1}^{t} S(V_i|Q_i)$represents the multiplication of all probabilities from frame 1 to frame t.

S2.3, the lip language recognition unit calculates the fusion probability between the probability of acoustic features and the probability of visual features, and selects the text sequence with the highest probability as the generated result, the calculation formula for the fusion probability is:;

$$S_r = \beta_1 \cdot S_1 + \beta_2 \cdot S_2$$

In the calculation formula, S_r represents the fusion probability, β_1 represents the weight of the probability of acoustic features, β_2 represents the weight of the probability of visual features, $\beta_1 + \beta_2 = 1$.

Preferably, the intelligent generation module is internally connected to a display device, and the final text is displayed through the display device.

Compared with existing technologies, the present invention provides a hearing aid system

based on the fusion of vision and hearing, which has the following beneficial effects:

The present invention comprehensively evaluates the environmental impact first, and in the case of environmental impact on speech recognition performance, combines speech recognition and lip recognition systems to obtain more information to supplement and verify the speech recognition results, thereby improving the overall recognition accuracy; At the same time, lip recognition is less likely to be affected by environmental factors compared to speech recognition systems. It can help the system observe changes in the speaker's mouth shape, adapt to different pronunciation characteristics, improve its ability to recognize various accents, and provide more accurate hearing aid assistance for hearing-impaired people, enabling them to clearly understand relevant information and improve service quality and efficiency.

Brief Description of the Drawings

Figure 1 is a schematic diagram of the system of the present invention.

Detailed Description

Below, the technical solutions in the embodiments of the present invention will be clearly and completely described in conjunction with the accompanying drawings. Obviously, the described embodiments are only a part of the embodiments of the present invention, not all of them. Based on the embodiments of the present invention, all other embodiments obtained by ordinary skilled persons in the art without creative labor are within the scope of protection of the present invention.

Please refer to Figure 1, a hearing aid system based on the fusion of vision and hearing, comprising an environment monitoring module, a multidimensional recognition module, and an intelligent generation module;

The environmental monitoring module consists of a noise monitoring unit, an electromagnetic monitoring unit, and an environmental analysis unit. The noise monitoring

unit is used to connect the microphone to the network to obtain noise data of the current environment, and to number the obtained current environmental noise data to form a noise dataset;

The expression for the number of the noise dataset is: Zs_1、Zs_2、Zs_3、…、Zs_n, where Zs_1 represents the first noise data obtained through the microphone, Zs_n represents the last noise data obtained through the microphone, and n represents the total number of data in the dataset.

The electromagnetic monitoring unit is used to connect electromagnetic environment monitoring instruments to the network to obtain electromagnetic data of the current environment, and to number the obtained electromagnetic data of the current environment to form an electromagnetic dataset;

The expression for the number of the electromagnetic dataset is: Cs_1、Cs_2、Cs_3、…、Cs_n, where Cs_1 represents the first electromagnetic data obtained through an electromagnetic environment monitoring instrument, Cs_n represents the last electromagnetic data obtained through an electromagnetic environment monitoring instrument, n represents the total number of data in the dataset, and the data acquisition time of the noise dataset is consistent with that of the electromagnetic dataset.

The noise monitoring unit and electromagnetic monitoring unit send the numbered noise dataset and electromagnetic dataset to the environmental analysis unit through the network. The environmental analysis unit calculates the environmental impact index based on the received dataset, and determines whether the current environment will affect the speech recognition effect according to the environmental impact index. If it is determined that it will affect the speech recognition effect, it sends an environmental anomaly signal to the multidimensional recognition module;

The calculation formula for the environmental impact index $HJyx$ is:

$$HJyx = \alpha_1 * \frac{Zs_i}{Zs_o} + \alpha_2 * \frac{Cs_i}{Cs_o}$$

In the calculation formula, $HJyx$ represents the environmental impact index, Zs_i

represents the i-th noise data, Cs_i represents the i-th electromagnetic data, Zs_orepresents the maximum noise interference data allowed by the speech recognition system, Cs_o represents the maximum electromagnetic interference data allowed by the speech recognition system, α_1 represents the weight of noise factors, and α_2 represents the weight of electromagnetic factors, $\alpha_1 + \alpha_2 = 1$;

When the calculated value of the environmental impact index is greater than the environmental impact threshold, it represents that the current environment will affect the speech recognition effect, and an abnormal signal is sent to the multidimensional recognition module;

Noise can directly affect the quality of speech signals, and electromagnetic interference may affect the normal operation of electronic devices in hearing aid systems. Combining noise and electromagnetic factors to comprehensively evaluate the environmental impact avoids the problem of a single factor being unable to accurately reflect the actual situation. It accurately determines whether the environment will increase the possibility of misidentification, thereby improving the accuracy and reliability of hearing aid systems;

The multidimensional recognition module includes a speech recognition unit and a lip recognition unit. The speech recognition unit is used to generate speech recognition text based on speech recognition;

The method for generating speech recognition text is:

S1.1、 The speech recognition unit processes the speech signal into frames based on the speech recognition system, and extracts D-dimensional acoustic features from each frame to obtain a feature sequence. The expression of the acoustic feature sequence is:

$$X = [X_1、 X_2、 X_3、 \cdots、 X_t]$$

In the expression, X represents the acoustic feature sequence, X_1 represents the D-dimensional feature vector of frame 1, X_t represents the D-dimensional feature vector of frame t, and t represents the number of time frames.

S1.2 the speech recognition unit calculates the conditional probability of the acoustic feature sequence on the candidate phoneme sequence through an acoustic model, and

selects the text sequence with the highest probability as the generation result, the calculation formula for the probability of the acoustic feature is:

$$S_1 = \prod_{i=1}^{t} S(X_i|Q_i)$$

In the calculation formula, S_1 represents the probability of acoustic features, X_i represents the D-dimensional feature vector of the i-th frame, and Q_i represents the phoneme state corresponding to the i-th frame; $\prod_{i=1}^{t} S(X_i|Q_i)$ represents the multiplication of all probabilities from frame 1 to frame t.

Speech recognition text generation reflects the process of directly recognizing speech through a speech recognition system without the influence of environmental factors. This includes extracting acoustic features, calculating the conditional probability of acoustic feature sequences on candidate phoneme sequences, and selecting the highest probability text sequence as the result for final text output, providing a solution for the difficulty of speech recognition for hearing-impaired individuals;

The lip recognition unit is used to generate fused recognition text based on lip language recognition combined with speech recognition after receiving abnormal signals;

The method for generating fused recognition text is:

S2.1 the lip recognition unit uses face detection technology to locate the lip area, extracts the D-dimensional visual feature vector of each lip image frame, and obtains a visual feature sequence, the expression of the visual feature sequence is:

$$V = [V_1、V_2、V_3、\cdots、V_t]$$

In the expression, V represents the visual feature sequence, V_1 represents the D-dimensional visual feature vector of frame 1, $V_t$represents the D-dimensional visual feature vector of frame t, and t represents the number of time frames.

S2.2 the lip recognition unit calculates the conditional probability of visual features on candidate phoneme sequences through a visual acoustic model, and the probability calculation formula for the visual features is:

$$S_2 = \prod_{i=1}^{t} S(V_i|Q_i)$$

In the calculation formula, S_2 represents the probability of visual features, V_i represents the D-dimensional feature vector of the frame, and Q_i represents the phoneme state corresponding to the frame; $\prod_{i=1}^{t} S(V_i|Q_i)$represents the multiplication of all probabilities from frame 1 to frame t.

S2.3, the lip language recognition unit calculates the fusion probability between the probability of acoustic features and the probability of visual features, and selects the text sequence with the highest probability as the generated result, the calculation formula for the fusion probability is:;

$$S_r = \beta_1 \cdot S_1 + \beta_2 \cdot S_2$$

In the calculation formula, S_r represents the fusion probability, β_1 represents the weight of the probability of acoustic features, β_2 represents the weight of the probability of visual features, $\beta_1 + \beta_2 = 1$.

The speech recognition unit and lip recognition unit send the generated recognition text to the intelligent generation module;

The fusion recognition of text generation reflects the use of speech recognition and lip recognition to solve the problem of hearing impairment for people with hearing impairment under the influence of environmental factors. The system can obtain more information to supplement and verify the speech recognition results, thereby improving the overall recognition accuracy. At the same time, lip language recognition is less likely to be affected by the environment compared to speech recognition systems, which can help the system observe changes in the speaker's mouth shape, adapt to different pronunciation characteristics, improve the recognition ability of various accents, provide more accurate hearing aid assistance for people with hearing impairment, enable people with hearing impairment to clearly understand relevant information, and improve service quality and efficiency;

The intelligent generation module is used to connect the display device and display the

final text.

Although the embodiments of the present invention have been shown and described, it will be understood by those skilled in the art that various changes, modifications, substitutions, and variations can be made to these embodiments without departing from the principles and spirit of the present invention. The scope of the present invention is limited by the appended claims and their equivalents.