Intelligent Vocal Training Assistant System

Yihong Li(✉) and Chengzhe Luo

College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China
1691135092@qq.com

Abstract. In professional vocal training, a way to evaluate the quality of vocalization is often needed. In order to solve various problems caused by the lack of professional instructors, a vocal training system for detecting the closed state of the vocal cords is proposed. We have proposed a robust vocal training system for the detection of the closed state of the vocal cords on the mobile terminal, which can analyze the closure of the vocal cords of the human body during vocalization, so as to evaluate the vocal cord ability without a professional teacher or professional equipment. In this system, we can use two wearable sensors, vibrating plate and headset, to collect the signals of the human vocal cords, or directly use the microphone of the mobile device to collect. And we use the convolutional neural network to analyze the signals and classify the closed state of the vocal cords. In order to build this system, we constructed a vowel data set classified by degree vocal cords closure, including the wearable sensors and mobile phone microphones we used. We compared the performance of the traditional vocalization pattern classification method and the convolutional neural network method we used to classify the vocal cord closure types on the data set we constructed and the public data set, and finally tested under two noise environments, and the preliminary results proved the usability of our system.

Keywords: Wearable devices · Vocal cord closed state detection · Convolutional neural network

1 Introduction

Traditional vocal music training generally requires one-to-one guidance from a professional teacher, and the place of teaching is also in a quiet and fixed place such as the piano room. The existing mobile vocal music system focuses on pitch and rhythm, almost no assessment of vocal quality, and cannot meet the basic vocal learning needs. If students do not focus on maintaining good sound quality at the beginning, they will develop wrong vocal habits, which will affect the sound effect and the health of the vocal cords.

In the quality of vocalization, the closed state of the vocal cords is a very basic indicator. In many vocal training theories, including EVT (Estill Voice Training), the vocal cords are divided into three basic states: normally closed, not tightly closed, and excessive closure. We use the terminology proposed by Sandberg to describe these three

K. Wu et al. (Eds.): ICECI 2021, LNICST 437, pp. 120–127, 2022.
https://doi.org/10.1007/978-3-031-04231-7_11

modes. The sound mode in which the vocal cords are not tightly closed is called breathy, because the sound is like the airflow noise during breathing. The sound mode in which the vocal cords are over-closed is called pressed, which means that the vocal cords are squeezed and over-closed and for normal vocal cords closed, it is called model.

Based on this, we have carried out research work on a vocal training system aimed at detecting the closed state of the vocal cords.

2 Related Work

The closed state of the vocal cord detection has always been a hot research issue in the field of vocal theory research. Its multi-faceted application value and academic value have caused extensive research by scholars and companies at home and abroad. It is mainly divided into the following research directions:

2.1 Detection of the Closed State of the Vocal Cords Based on Electroglottograph

The principle of detecting the closed state of the vocal cords based on the electroglottograph is to capture the bioelectric signals on the skin surface close to the vocal cords. The technique mentioned is to create a database consisting of sound booth recordings from participants, whom are required to sustain vowels (/a/, /e/, /i/, /o/, /u/) in their typical speaking voice at a comfortable pitch and loudness. Then, they are asked to produce the same vowel set again, while mimicking three non-modal voice qualities (Fig. 1).

Fig. 1. Electronic glottis

This method requires professional medical testing equipment, and has poor universality and no mobility, which cannot satisfy the need of general user.

2.2 Detection of the Closed State of the Vocal Cords Based on Accelerometer

It is a non-invasive attempt to indirectly measure the subglottal sound pressure, which can efficiently resist the external noise interference.

This method used a three axis acceleration sensor to press against the skin of neck to record the sensor signals. It identified a specific tissue around the larynx to access the subglottal pressure (Fig. 2).

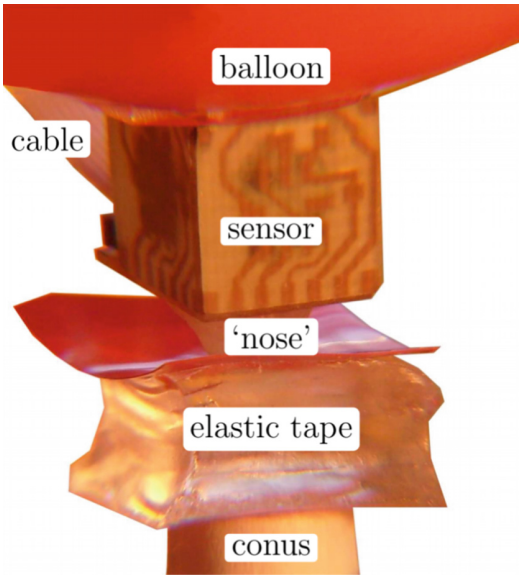

Fig. 2. Sensor in test environment-pressed on compressible elastic tape

On the other hand, this system requires user to firmly attach the accelerometer to their skin, which may cause skin discomfort. Furthermore, due to the low-resolution characteristics of the accelerometer itself, it is easily affected by body disturbances, which resulting in low recognition accuracy.

2.3 Detection of the Closed State of the Vocal Cords Based on Audio Signal Analysis

This method is a non-invasive detection method, and can be combined with existing commercial equipment (such as earphones, smart phones), and has universal applicability.

This detection method has a significant effect on distinguishing the voice type of voice and singing voice. Classification with support vector machine classifiers indicated that the proposed features and their combinations showed improved accuracy compared to usually employed glottal source features and mel-frequency cepstral coefficients (MFCCs).

However, the current technology can be divided into three categories: excessive closure, normal closure, and lax closure, and the accuracy of the results is greatly affected by changes in personnel, and also by changes in sound pitch, and they cannot achieve real-time detection.

2.4 Some Problems Existed in Related Research

The equipment is inconvenient to carry, and it is difficult to achieve real-time monitoring.

If user need to measure more accurate physiological signals, such as brain waves and electro-skin signals, you need to use large machines and oscilloscopes, which are difficult to carry at any time.

Some methods are inconvenient for people to actually use, and easily cause skin discomfort.

Some methods are based on accelerometers to detect the closed state of the vocal cords. But the accelerometer needs to be close to the skin close to the vocal cords, which will cause skin discomfort.

Monitoring results are not accurate enough.

Some low-resolution methods, such as accelerometers, may also be interfered by body motions, and the recognition accuracy is low. Moreover, the accuracy of the results is greatly affected by changes in personnel, as well as changes in voice pitch.

3 System Design

This system is an auxiliary vocal training system based on smart wearable devices. The overall design is based on skin vibration signals and sound sensing signals, it detects the closed state of the vocal cords when the user is speaking to guide the user's vocal learning in real time.

The provided system generally includes: a wearable vocal cord vibration sensor and a smart terminal, where the wearable vocal cord vibration sensor is used to collect skin vibration signals near the vocal cords and ear canal sound signals.

The wearable vocal cord vibration sensor includes an audio signal collector, a vibration signal collector, an audio processing module, a control unit, and a communication unit. The audio signal collector is used to collect the sound signal generated by the user's voice. The vibration signal collector is used to collect the skin at the position of the vocal cord. The skin vibration signal caused by the vibration. The audio processing module is used to process the collected sound signal and the skin vibration signal, and transmit it to the smart device via the communication unit. The control unit is used to coordinate the information interaction process between other modules or units, such as controlling the start and end of sound signal and skin vibration signal collection, and controlling the data transmission between the wearable vocal cord vibration sensor and the smart terminal. The smart device processes the received sound signal and skin vibration signal into a time-frequency graph, and inputs the pre-trained machine learning classifier to obtain the classification result of the closed state of the vocal cords (Fig. 3).

The following takes a smart phone with an Android operating system as an example to illustrate the detection process of the closed state of the vocal cords, including the following steps:

Start the smart vocal cord vibration sensor, the ear canal microphone collects the sound of the user's ear canal, and the original vibration signal of the skin near the vocal cord of the piezoelectric ceramic sheet. Filter and amplify the original vibration signal through the band-pass filter amplifying circuit of the intelligent vocal cord vibration sensor. The collected ear canal sound signal and the amplified and filtered vibration signal are transmitted to the mobile terminal in real time through the communication unit.

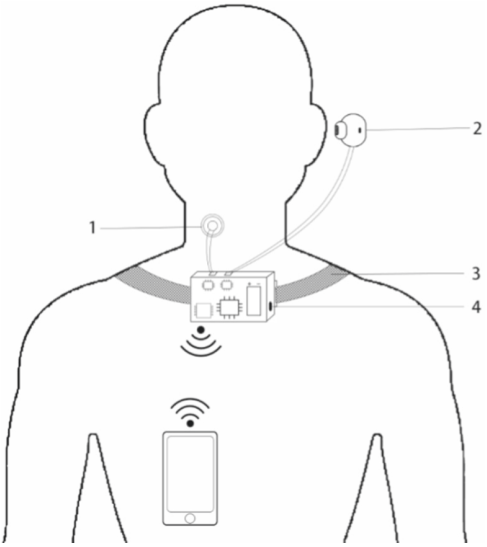

Fig. 3. An in-ear microphone[1]; A piezoelectric ceramic sheet[2]; An adjustable neckband[3]; A type-c charging port[4].

The mobile terminal performs processing on the received sound data and vibration signal data, and obtains the judgment result of the closed state of the vocal cords and vocal practice suggestions, and displays them in the APP. Finally, feed the result back to the user, and generate a related report.

Particularly, we design a system to analyze the sound data and vibration data. The step specifically includes the following steps:

The original ear canal sound data and skin vibration signal data are divided into frames, and divided into multiple windows for processing. First, the amplified and filtered skin vibration signal and the ear canal sound signal collected by the microphone are received, and the signal is divided into windows to process the data of each window.

Use a voice time detection algorithm (VAD) to detect voice data, and extract a data frame corresponding to the user's voice.

Convert the data frame into a time-frequency image through a short-time Fourier transform (SFFT). Specifically, the incoming data stream is buffered, the buffered data is retrieved after accumulation for a certain period of time, and both the sound data and the skin vibration signal data are converted into a time-frequency map through a short-time Fourier transform algorithm.

Through machine learning, the deep learning model analyzes the time-frequency image to recognize the closed state of the vocal cords.

4 Data Set

We invited 20 students major in vocal performance to assist us in the collection of data sets. In the process of collecting data, the students were asked to make sounds with

different vocal cords closed by referring to the sample audio in a natural or simulated state (Fig. 4).

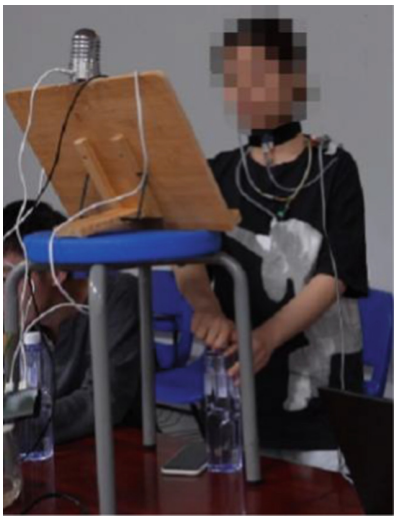

Fig. 4. Students in test

We asked the students to try a combination of 4 sounding scenes and 3 types of vocal cord closure situations. The 4 sounding scenes includes long vowel, short vowel, staccato[1], and fade.[2] The 3 types of vocal cord closure situations had been mentioned above, which are breathy, neutral, and press. We use a variety of data-collecting devices: mobile phone microphones, in-ear microphone, piezoelectric ceramic, and high-fidelity microphone.

5 Experiment and Result

After collecting the data, we use the current convolutional neural network model for feature extraction. After trying various convolutional neural network models, we got the following results.

The graph depicts that resnet18 method performs better than other methods and reach a considerable performance (Fig. 5).

On this basis, we tried to use different equipment to compare the better-performing methods in machine learning with resnet18 for cross-person[3] and non-cross-person experiments (Figs. 6 and 7).

[1] Tone changing sound.

[2] Volume changing sound.

[3] Non-cross-person refers to the mixture of all people's data, and then divides the training set and test set in proportion. Cross-person refers to the test on one person's data alone, and the data of other people is used to train the model.

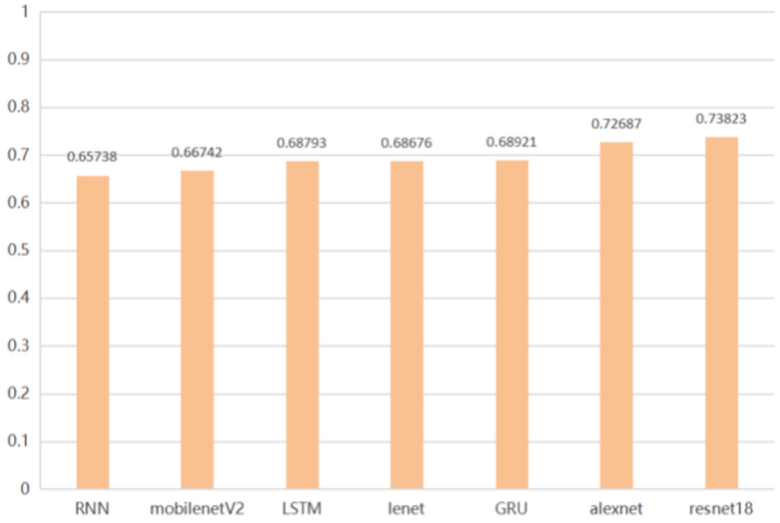

Fig. 5. Different performance of multiple Convolutional Neural Network models

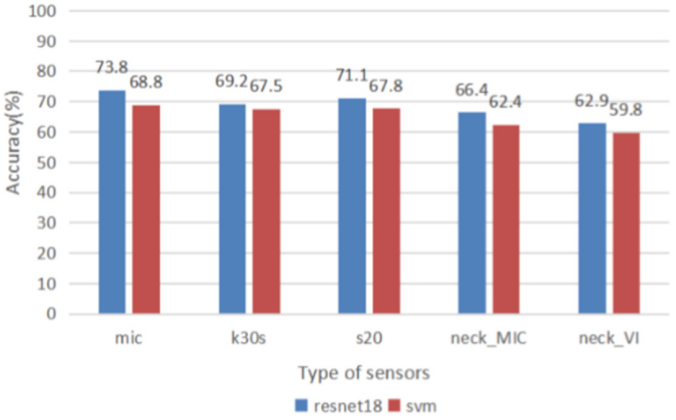

Fig. 6. Experimental results of different devices without cross-person

As is demonstrated in the graph, resnet18 is better in performance than traditional machine learning methods in many places.

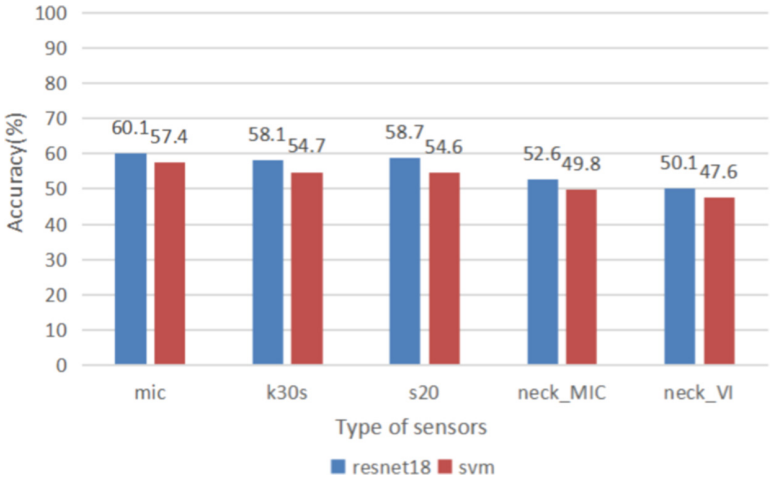

Fig. 7. Experimental results of different devices with cross-person

6 Conclusion

We propose a robust vocal music training system for vocal cord closure detection at the mobile end, which can analyze the vocal cord closure of human body when speaking. After comparing a variety of convolutional neural network models with traditional machine learning methods, we find that resnet18 has better performance. The accuracy of the test on the headset and vibrator is significantly lower than that of other devices. The possible reason for this phenomenon is the deviation of the data quality collected by the two sensors, such as the inherent noise of the hardware and the lack of high-frequency information. To solve this problem, we will improve the existing hardware equipment to improve the quality of collected data and improve the performance of the system.

References

1. Sundberg, J. The Science of the Singing Voice. Illinois University Press (1987)
2. Murphy, P.J.: Temporal measures of the initial phase of vocal fold opening across different phonation types. In: Models and Analysis of Vocal Emissions for Biomedical Applications: 6th International workshop, 14–16 December 2009, Firenze, Italy. Proceedings e report, p. 54. Firenze University Press, Firenze (2009)
3. Wokurek, W., Pützer, M.: Accelleration sensor measurements of subglottal sound pressure for modal and breathy phonation quality. In: Models and Analysis of Vocal Emissions for Biomedical Applications: 6th International Workshop, 14–16 December 2009, Firenze, Italy. Proceedings e report, p. 54. Firenze University Press, Firenze (2009)
4. Kadiri, S.R., Alku, P.: Mel-frequency cepstral coefficients of voice source waveforms for classification of phonation types in speech. In: Interspeech (2019)
5. Kadiri, S.R., Alku, P., Yegnanarayana, B.: Analysis and classification of phonation types in speech and singing voice. Speech Commun. **118**, 33–47 (2020)