Research on Modeling of Vocal State Duration Based on Spectrogram Analysis

Xiaoyan Zhang¹

¹School of Preschool and Elementary Education of CWNNU
CWNNU (China West Normal University), Nanchong, Sichuan, China

Abstract—In the early stage of vocal music education, students generally do not understand the structure of the human body, and have doubts about how to pronounce their voices scientifically. However, with the continuous development of computers, computer technology has become more and more developed, and computer processing speed has been greatly increased, which provides favorable conditions for the development of the application of vocal spectrum analysis technology in vocal music teaching. In this paper, we first study the GMM-SVM and DBN, and combine them to extract the deep Gaussian super vector DGS, and further construct the feature DGCS on the basis of DGS; then we study the convolutional neural network (CNN), which has achieved great success in the image recognition task in recent years, and design a CNN model to extract the deep fusion features of vocal music. The experimental simulations show that the CNN fusion-based speaker recognition system achieves very good results in terms of recognition rate.

1 INTRODUCTION

The development of integrated technology has driven the development of semiconductor chips, each chip is able to store tens of thousands of transistors, allowing calculators and controllers to be concentrated on a single chip, which led to the emergence of the microprocessor, combined with large-scale, very large-scale integrated circuits to form the microcomputer, which is now a notebook and desktop computer [1]. In general, PPT is used in vocal music teaching, which not only ensures the quality of the classroom, but also enhances the entertainment of the classroom and achieves a good teaching effect [2]. In vocal music teaching, spectral analysis technology is widely used in vocal music teaching through computers [3].

The rapid conversion of analog signals to digital signals using computers is a way for students to visualize their voices on a computer screen. The computer displays the time, frequency, and intensity of sound occurring from the computer screen primarily through three-dimensional images; the y-axis represents the frequency spectrum, and the darker the grayscale, the stronger the intensity, and the lighter the grayscale, the weaker the intensity [4]. The spectrum is a two-dimensional image, usually with curves and data to represent the frequency and intensity of the sound. The use of computer technology in the vocal music teaching process has significantly improved the teaching quality of vocal music education [5]. Through the use of computers in the classroom, the organic unification of science and technology with school education can be fully reflected, which can effectively mobilize students'

initiative in learning. The use of advanced spectral analysis technology, the usual abstract concepts of sound into computer graphics, so that students can be more vivid and concrete understanding of their own voice, breaking through the traditional "one-to-one" teaching mode, to achieve the "mouth - ear - nose!" a hybrid teaching model [6]. Sound visualization focuses on the use of graphics to plot the different frequencies and amplitudes of collected sounds, representing the interaction between the frequency component and the overall sound. The use of advanced computer technology allows the use of text, video, images, and sound to add a variety of fun and excitement to the classroom [7]. As a result, advanced computerized spectral analysis techniques are widely used in vocal music classrooms, and they have achieved a very good result.

This paper provides a comprehensive introduction to the basics of speaker recognition, including the basic principles of speaker recognition, the feature extraction process, and the main recognition models. First, the extraction process of the speech feature parameters MFCC and LPCC is outlined in this paper. Then, the classical speaker recognition models, such as GMM, GMM-UBM general background model, support vector machine SVM, and deep neural network, are introduced in detail. According to the previous studies, these models have good performance in speaker recognition systems, and this thesis is also based on these models. Based on the superior performance of the fusion feature, this paper constructs a CNN fusion feature using a convolutional neural network. In this paper, the speech material is first converted into a speech map, and then the speech map is used as the input of the CNN to construct a speaker

* Corresponding author: 275809262@qq.com

recognition system. It is shown that the number of layers in the CNN network has an important impact on the system performance. In order to make better use of the features between different layers, the CNN features in two different layers with better recognition rate are fused in this paper. The experimental simulation shows that the CNN fusion-based speaker recognition system achieves a good recognition rate.

2 SPECTROGRAM-BASED VOCAL RECOGNITION

2.1 Associative Hyper-Vector-based Speaker Recognition System

The traditional GMM-SVM speaker recognition system is shown in Figure 1. In the training phase, the feature parameters are extracted from the training speech material after the preemphasis and endpoint detection processes. Then, the feature parameters are input into the GMM to extract the super vector, and the super vector is used to train the classifier SVM; for the test phase, the test feature parameters are extracted after the same pre-processing process, and then they are used to extract the super vector. Finally, the test super-vector is used for pattern recognition with the trained SVM to obtain the result.

In this chapter, MFCC is selected as the input eigenparameter of GMM because it better fits the auditory characteristics of human ear. The Gaussian mixture model is estimated by parameter λ (mixture weight, mean vector, and covariance), and then the appropriate M is chosen to characterize the phonetic personality. The vector m can be expressed by the following formula. Then the hypervector m can be expressed by:

$$m = \begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ m_i \end{bmatrix} \quad (1)$$

In a GMM-SVM-based speaker recognition system, the super vector can be seen as a refinement of the input feature parameters, which can not only effectively erase the text information but also highlight the speaker's identity and personality characteristics; the super vector is also a process of data downscaling, thus reducing the complexity of SVM processing to some extent.

Figure 1. Vocal recognition based on traditional GMM-SVM.

2.2 In-depth e-learning

RBMs are energy-based and can learn an unknown distribution of data. Each RBM is a Markov random field with a two-layer structure, i.e., a visible layer and a hidden layer. While the traditional Boltzmann machine layers are fully connected to each other, the RBMs have no connections between the visible layer and the nodes in the visible layer and between the hidden layer and the nodes in the hidden layer; only the hidden layer is connected to the nodes in the visible layer, and this connection is bidirectional. Given the statistics of the visible layer, then the RBM energy function for the hidden and visible unit states can be expressed by the following equation.

$$E(v, h) = -\sum_{i=1}^I a_i v_i - \sum_{j=1}^J b_j h_j \quad (2)$$

where ω denotes the weight matrix, a and b denote the thresholds for the visible and hidden layer units, respectively, and I and J are the number of visible and hidden layer neural units, respectively. From $E(v)$ in the above equation, a joint probability distribution can be obtained.

$$q = \frac{e^{-E(v, h)}}{Z} \quad (3)$$

where h is the normalization factor, which represents the sum of the energy values between the hidden layer and the visible layer cells. Further an independent distribution of v can be obtained, which can also be called a likelihood function.

$$q(v) = \frac{\sum_{v, h} e^{-E(v, h)}}{\sum_{v, h} e^{-E(v, h)}} \quad (4)$$

The units in the same visible or hidden layer of the RBM are not connected to each other, so that the units in the hidden layer are independent of each other given the parameter v for the visible layer. Given random h , one can obtain the probability of v .

$$q(v / h) = \prod \frac{\sum_{v, h} e^{-E(v, h)}}{\sum_{v, h} e^{-E(v, h)}} \quad (5)$$

Because of the lack of connectivity between units in the same layer of the RBM, the probability of activation is also conditionally independent in two cases.

$$q(v_i = 1 / h) = \sigma(a_i + \sum_{j=1}^J h w_{ij}) \quad (6)$$

$$q(h_i = 1 / v) = \sigma(b_i + \sum_{j=1}^J v w_{ji}) \quad (7)$$

where $\sigma(x) = 1 / (1 + e^{-x})$ is the sigmoid activation function. Combining the two equations above, that is, given v , first the hidden layer state p can be calculated by $v=1/h$, and then by v can be computed for the visual layer state reconstructions. The error between the visual layer units and the reconstructed visual layer units is minimized by

this quasi-side, and the hidden layer can be seen as an alternative representation of the visual layer. To conclude, the hidden layer units are the result of feature extraction from the visible layer input units, and thus the structure of the RBM can achieve the goal of feature extraction, as shown in Figure 2.

Figure 2. The relationship between hidden and visible layers.

3 EXPERIMENTAL SIMULATION

This thesis uses speech material recorded by the team under anechoic chamber conditions with 210 speakers, where the number of statements per speaker is 180 and the average duration of each statement is 3 seconds. In addition, the signal is sampled at 16k Hz and quantized at 16bit, and the frame length N is set to 256 before extracting the parameters. To analyze the results of the comparative experiments, 10 speakers are selected in all experiments and 80 statements are randomly selected from each speaker's speech material. 60 statements are used for training and 20 statements are used for testing. In this thesis, the initial input features of the speaker recognition system are all MFCCs, and it is shown that the original MFCC contains only static information about the speech, but in order to make the features more dynamic, the first order reciprocal of the MFCC can be obtained and spliced with the original MFCC.

3.1 Experimental Simulation of Associated Supervectors

In order to verify the validity of the correlation concept proposed in this chapter, an experimental simulation of a GMM-SVM system based on the correlation super vector is performed. The Gaussian mixed number M is a power of 2 (1, 2, 4, 8, 16,...). The choice of M has a great impact on the performance of the GMM. If M is too small, the GMM fits the training data poorly and the recognition rate is unsatisfactory; if M is too large, the model is difficult to converge and the parameter errors are large. The type of kernel function greatly affects the performance of the SVM classifier, however, there is no specific theory guiding which kernel function should be chosen in which case. Therefore, the experiments in this section simulate all four basic types of kernel functions. The effect of the Gaussian correlation number on the speaker recognition rate is shown in Figure 3.

Figure 3 Experimental simulation results based on correlation super vector.

As with M , the Gaussian correlation number q takes the power of 2. It can be seen that the radial basis kernel function performs best under the four kernel function conditions and the polynomial kernel function performs worst. Based on the line graph of the radial basis kernel function, it can be seen that the lowest recognition rate is obtained when $q=1$, i.e., the original hypervector is the worst case. M is inversely related to p . As q increases, the number of samples in the mean vector decreases, but the number of dimensions per sample increases significantly. The increase in the number of dimensions leads to redundancy and introduces noise interference, so that q is too large and the recognition rate decreases. It can be concluded from the line graph that the recognition rate of systems with $i > 1$ is higher than $i = 1$, i.e., the performance of the associated super-vector is better than the original super-vector for any Gaussian number of correlations.

3.2 Vocal duration analysis

In a deep learning model, if the problem to be solved is uncorrelated with the variation of the sample mean, the mean of the features should be zeroed to reduce the effect of the features on the DBN network model and reduce the correlation. In speech processing, the use of inverted spectral mean normalization can reduce the undesirable effects of channel distortion by subtracting the mean of the MFCC parameters from the statement. Global feature normalization scales the feature data in each dimension so that the feature vectors lie within a similar dynamic range. In speakers, global transformations are typically used to normalize the feature parameters to zero mean and unit variance. Given a training data set, one can compute the mean and standard for each dimensional feature i of it.

$$\mu = \frac{\sum_{m=1}^M o_i}{M} \quad (8)$$

$$\mu = \sqrt{\frac{\sum_{m=1}^M (o_i - \mu)^2}{M}} \quad (9)$$

Next, standardize using the following formula.

$$\tilde{o}^m = \frac{o^m - \mu}{\sigma} \quad (10)$$

For DBN models, increasing the number of hidden layers can reduce the network error, but the training time of the network increases and is accompanied by a tendency to overfit. So in order to find the most appropriate number of hidden layers, fix the number of nodes for all hidden layers (including bottlenecks) and change the number of hidden layers. In the case of two hidden layers, the first hidden layer is set as the bottleneck layer; in the case of 3, 4, and 5 hidden layers, the second layer is set as the bottleneck layer. The experimental results are shown in Figure 4.

Figure 4. Number of nodes in the bottleneck layer.

3.3 Application of Audio Frequency Analysis Techniques

By configuring the computer with a corresponding sound frequency analysis software, it is possible to systematically evaluate and test the singer's voice through this software. By analyzing each frequency component and intensity, the sound frequency analysis software shows the singer in the form of an image of the sound spectrum. In a two-dimensional image of a sound spectrum, the x-axis represents the frequency and the y-axis represents the amplitude, and the spectral lines vary in length in the graph. The longer the line length, the greater the amplitude of the corresponding frequency, and the shorter the line length, the smaller the amplitude of the corresponding frequency. The frequency of the fundamental tone is called the fundamental frequency represented by the first spectral line, also called the first harmonic, and so on after the first harmonic is called the second harmonic, the third harmonic... The frequency of the second harmonic is twice the frequency of the first harmonic, the third harmonic is three times the frequency of the first harmonic, and the frequency of the harmonic

remains multiplicative with the fundamental tone. The frequency of the second harmonic is twice the frequency of a harmonic, and the third harmonic is three times the frequency of a harmonic, and the frequency of the harmonic maintains a multiple relationship with the fundamental tone. We can use the spectral software on the computer to display the sound graphically, as shown in Figure 5.

Figure 5. Sound picturization changes

The use of computerized noise acoustics detection systems is the main embodiment of the use of sound spectral analysis technology in computers. The system is able to identify the spectral differences between artistic and non-artistic vocalizations, where the bandwidth of the artistic vocalization increases with the increase of the resonant peak energy and vice versa, and can effectively identify the differences between professional and amateur vocalists using resonant vocalizations. The detection system analyzed the vowel signals of 100 professional vocalists and analyzed the spectral characteristics of the spectra, which were mainly influenced by harmonics, resonance peaks, and noise components.

4 CONCLUSION

This paper provides a comprehensive introduction to the basics of speaker recognition, including the basic principles of speaker recognition, the feature extraction process, and the main recognition models. First, the extraction process of the speech feature parameters MFCC and LPCC is outlined in this paper. Then, the classical speaker recognition models, such as GMM, GMM-UBM general background model, support vector machine SVM, and deep neural network, are introduced in detail. According to the previous studies, these models have good performance in speaker recognition systems, and this thesis is also based on these models. Based on the superior performance of the fusion feature, this paper constructs a CNN fusion feature using a convolutional neural network. In this paper, the speaker speech material is converted into a speech spectrogram, and then the spectrogram is used as the input of CNN to construct a speaker recognition system.

REFERENCES

1. Alzamendi, Gabriel A., and Gastón Schlotthauer. "Modeling and Joint Estimation of Glottal Source and Vocal Tract Filter by State-Space Methods." *Biomedical Signal Processing and Control*, vol. 37, April 2017, pp. 5-15.
2. Galindo, G. E., Peterson, S. D., Erath, B. D., Castro, C., Hillman, R. E., & Zañartu, M. "Modeling the Pathophysiology of Phonotraumatic Vocal Hyperfunction with a Triangular Glottal Model of the Vocal Folds." *Journal of Speech Language and Hearing Research*, vol. 60, no. 9, March 2017, pp. 2452-2471.
3. Yoshinaga, T., Hirtum, A. V., & Wada, S. "Multimodal Modeling and Validation of Simplified Vocal Tract Acoustics for Sibilant /S/." *Journal of Sound and Vibration*, vol. 411, June 2017, pp. 247-259.
4. Gamba, M., Favaro, L., Araldi, A., Matteucci, V., Giacoma, C., & Friard, O. "Modeling Individual Vocal Differences in Group-Living Lemurs Using Vocal Tract Morphology." *Current Zoology*, vol. 63, no. 4, July 2017, pp. 467-475.
5. Zhang, J., Honda, K., & Wei, J. "Tooth Visualization in Vowel Production MR Images for Three-Dimensional Vocal Tract Modeling." *Speech Communication*, vol. 96, July 2018, pp. 37-48.
6. Speed, M., Murphy, D., & Howard, D. "Modeling the Vocal Tract Transfer Function Using a 3D Digital Waveguide Mesh." *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 2, March 2014, pp. 453-464.
7. Nasir, Baucom, B., Bryan, C., Narayanan, S., & Georgiou, P. "Modeling Vocal Entrainment in Conversational Speech Using Deep Unsupervised Learning." *IEEE Transactions on Affective Computing*, no. 1, May 2020, pp. 11-21.