Abstract: - With the rapid development of artificial intelligence, and people are not only seeking material needs, but also pursuing happiness and spiritual aspirations. Especially some singing talent shows appear, but there are many controversial results. In order to solve the problem of singing with note-name evaluation and evaluate the singer's singing level in the professional music education of solfeggio and ear training, the paper uses artificial intelligence algorithms to recognize the note-name. Firstly, preprocess the singing sound data, including pre emphasis, framing, and windowing to reduce the impact of noise, and extract the acoustic features of the singing sound, the acoustic speech features are organized into a form suitable for DCNN input and a DCNN model is designed to obtain the pinyin form of the singing sound, the softmax layer adopts an end-to-end CTC structure, which classifies the input singing speech, finds corresponding phoneme sequences, and obtains the output results. The end-to-end CTC structure is used to classify and optimize the recognition process, and finally the state of the obtained features is output. And then singing pronunciation dictionary generates candidate word sequences based on the mapping relationship between phonemes and notes. Finally, based on the acoustic model score, the candidate note sequence with the highest score is obtained through decoder processing, and obtained the singing with note-name result of the singing sound. In order to further improve recognition performance, the paper introduces a new CTC-DCNN acoustic model. In this model, residual blocks can transfer input features to the output part through shortcuts, allowing the multi-layer convolutional speech features to be preserved as much as possible. At the same time, the deep structure can also better achieve the extraction and analysis of speech features. A new and improved CTC-DCNN acoustic model is obtained by optimizing the maxout function. The algorithm proposed in this paper is fair and equal in obtaining information, and no one participates in the entire process of obtaining scores. It is believed that the scoring results obtained in this way should be more objective.

Keywords: Convolutional Neural Network, End-to-End Model, Singing with Note-Name, Acoustic Model, CTC, Maxout Function

I. INTRODUCTION

With the rapid development of artificial intelligence, it has involved all aspects of people's lives [1-3]. With the continuous development of our country's economy, the living standards of our citizens have greatly improved. People's needs for life are no longer just reflected in material things. Numerous entertainment programs have also received public attention. At this time, some talent shows also emerged. Especially in some singing talent shows, which boast singing skills as a benchmark, there are always many controversial results. For example, there are often insider information leaked online, or the evaluation results of the judges differ significantly from the subjective feelings of the majority of the audience, causing great controversy. The fundamental reasons for these phenomena are believed by researchers to be the lack of a unified standard or the failure to refine each standard to a specific level of quantity. Simply put, there is only qualitative analysis, not quantitative analysis. The 2011 "Curriculum Standards for Primary and Secondary School Students" pointed out that attention should be paid to the requirements for singing posture, breathing methods, intonation, and other aspects in music curriculum content. The practice of singing skills should focus on combining singing practice activities. Based on the drawbacks of the current music teaching situation, which cannot meet the requirements for practical music learning, people have proposed the use of signal processing technology, digital storage, analysis of music and singing, integration of information technology into music teaching, development of music grading and feedback systems, and implementation of one-on-one singing guidance. Therefore, the study of music grading methods, as a comprehensive study in audio signal processing, not only has theoretical value, but also has significant practical significance. The timbre is defined by the harmonic spectrum system, and its physical basis is resonance. However, different timbres can bring different emotional experiences to people. Kazazis S et al. established a timbre space

1 Bintao Deng
2 Shimin Wang
3 Xiaodan Ren
4 Haiyang Ni
5 Ning Wang
6 Ming Zhao

1 College of Education and music, Sanming University, Sanming, 365004, China
2 School of Computer Information and Engineering, Jiangxi Normal University, Nanchang, Jiangxi 330000, China
3 Faculty of Education, Languages, Psychology and Music, SEGI University, Kuala Lumpur, 47810, Malaysia
4 Faculty of Education, Languages, Psychology and Music, SEGI University, Kuala Lumpur, 47810, Malaysia
5 Faculty of Education, Languages, Psychology and Music, SEGI University, Kuala Lumpur, 47810, Malaysia
6 School of Education, Jiangxi Institute of Applied Science and Technology, Nanchang, Jiangxi 330000, China

*Corresponding author: Ming Zhao
Copyright © JES 2024 on-line : journal.esrgroups.org
by extracting the objective features of the underlying timbre, comprehensively describing the connection between instrument timbre and human emotions [4]. At present, the spectral based audio features mainly include linear predictive cepstrum coefficient (LPCC) [5], Mel frequency cepstrum coefficient [6], constant Q transform [7], discrete wavelet transform, and discrete cosine transform. The characteristics of music emotions are usually composed of the following: for music information, the more complete the information contained in the original information, and the transformation of information generally accompanies a decrease in the amount of information, so choosing the appropriate audio features has become the top priority. Researchers usually choose MFCC coefficients and PLP coefficients because their unique cepstrum based extraction method is more in line with human auditory principles, and is also the most common and effective audio feature extraction algorithm; The PLP coefficient is due to its better robustness against MFCC coefficients. In 2021, Vasishtha Srishti et al. creatively proposed a keyword based method for audio sentiment detection [8]. Ryczkowska Alicja et al. used song intensity, timbre, and rhythm as emotional features of music to detect and track song emotions [9]. Trohidis et al. used the rhythm and timbre of music as features for multi label classification of audio [10]. Therefore, it is necessary to combine the music theory knowledge provided by music itself, namely the MFCC and PLP features in note information and audio information, to describe the various information provided by music itself more comprehensively from two dimensions. Yoon et al. [11] propose TutorNet not only distills the knowledge between networks with different topologies but also significantly contributes to improving the performance of the distilled student. Yoon et al. [12] also propose a simple yet effective knowledge distillation (KD) for the CTC framework, Panchapagesan et al. [13] distillation loss is simple and efficient, and uses only the “y” and “blank” posterior probabilities from the RNN-T output probability lattice. Kim et al. [14] propose a two-stage textual knowledge distillation method that matches utterance-level representations and predicted logits of two modalities during pre-training and fine-tuning, Vignesh et al. [15] introduce an auxiliary loss to distill the encoder logits from a teacher transducer's encoder, and explore training strategies where this encoder distillation works effectively, Rathod et al. [16] proposed a multi-level asymptotic compression method for ConformTransformer. Wu et al. [17] propose a Mixup-based Knowledge Distillation (MKD) method which combines Mixup, a data-agnostic data augmentation method, with softmax-level knowledge distillation.

In professional music education of solfeggio and ear training, a crucial issue is the evaluation of singing with note-names. Singing is the most important medium for solfeggio and ear training teaching. Without singing, it is impossible to carry out this teaching activity. This shows that the singing method serves the teaching of solfeggio and ear training, and it plays an irreplaceable role. In order to solve the controversial issue of measuring the singing performance of contestants, this paper provides a specific evaluation value from the aspect of note recognition in singing. Based on the specific evaluation value, a comprehensive score is given according to the predetermined scoring method. The significance is that all scoring rules are fair and equal for singers in terms of information acquisition, and no one participates in the entire process of obtaining the score, That is to say, there are no subjective factors of mixed people involved, and I believe that the scoring results obtained in this way should be more objective.

II. PRINCIPLES AND METHODS

A. Singing with Note-name

Due to the uncertainty of the pitch relationship that has already been demonstrated among various singing with note-names, it is impossible for people to understand the accurate interval relationship between different musical notes through vocals, and to intuitively determine the position and significance of each note in a specific mode. In addition, although the fixed roll method is more suitable for the reading of scores by performers of instrumental music (especially keyboard instruments with fixed pitch), it is not very suitable for the sight singing of vocal singers using human voice as an instrument. For a beginner who uses fixed solofing or someone who primarily focuses on vocal music and does not understand any musical instruments, it is difficult to use fixed solofing, even for music that sings a natural tone system, because the mode status of each note and the natural full and half pitch positions between each note cannot be distinguished from the fixed notation. To address this issue, European musicians have conducted long-term explorations. Until the early 19th century, Swiss British female teacher Sarah A Glover (1785-1867) founded the Tonic sol fa method based on her musical practice. After being revised by Pastor John Curwen (1816-1880), it quickly spread to some European countries [18].

In the first tone note method, the notes of the seven fundamental levels in the natural major mode scale are represented by the do, re, mi, fa, so, la, ti. In addition, if the fundamental level of the mode is raised by a half tone, the final vowel of the fundamental level roll is changed to "i". For those who experience a half tone decrease in the
fundamental level of a mode, the final consonant of the fundamental level roll scale shall be changed to "e". To avoid the similarity between the basic roll "re" and its flat half roll, the flat half roll of "re" is replaced by the roll "reh". Since the emergence of the first tone singing method, the culture represented by the United Kingdom and France, as well as the fixed singing method, has been widely spread around the world with the influence of their respective cultures. Nowadays, a situation has emerged worldwide where both fixed and first tone singing methods coexist and are used equally.

B. **Main Features of Sound Signals**

The features parameters will definitely vary among different singers, and this difference is called the difference between singers. The differences between singers are caused by their different vocal characteristics, and it is this difference that distinguishes different singers. Another difference is called the difference of the singers themselves. People cannot repeat a lyric or tone exactly twice; there is always a difference between the two. This difference is mainly caused by factors such as the singer's speed, the singer's emotions, the environment around the singer, the distortion of recording equipment and transmission channels.

According to the stability of parameters, singer characteristic parameters can be roughly divided into two types. The first type is the inherent characteristics determined by the singer's physiology (such as personality differences in vocal tract construction, etc.), mainly manifested in the frequency structure of the singer's sound signal. Representative characteristic parameters include pitch frequency and resonance peak of the sound. This type of feature is not easily imitated, but is easily influenced by health. The other type is the dynamic characteristics of vocal tract movement, namely the way and habit of vocalization. The representative feature parameters are the linear regression coefficients of the cepstrum and pitch. The cepstrum coefficient reflects the resonance performance of the vocal tract and is currently widely used as a speaker feature parameter. Commonly used cepstrum coefficients include linear prediction cepstrum coefficient (LPCC) and Mel cepstrum coefficient (MFCC).

C. **End-to-End CTC Model**

Speech recognition refers to the correct conversion of a given speech signal into corresponding text content output. Traditional speech recognition system [19] mainly includes four models: acoustic model, language model, pronunciation dictionary, and decoder. The specific process of speech recognition is as follows in the figure 1. Firstly, pre-processing the speech data, including pre emphasis, framing, and windowing to reduce the impact of noise, and convert the speech signal from time domain to frequency domain to obtain acoustic features. Acoustic models are used to obtain state sequences corresponding to acoustic features and generate corresponding acoustic model scores. A pronunciation dictionary generates candidate word sequences based on the mapping relationship between phonemes and words. The language model is used to evaluate the probability of the occurrence of candidate word sequences and generate corresponding language model scores. Finally, based on the acoustic model score and language model score, the candidate word sequence with the highest score is obtained through decoder processing, which is the recognition output text.

Since the adoption of hybrid modeling, the accuracy of ASR has significantly improved. This breakthrough replaces the traditional Gaussian mixture model with DNN for acoustic likelihood assessment, while still retaining all components such as acoustic model, language model, and dictionary model as a hybrid ASR system. Recently, there has been a new breakthrough in speech recognition, transitioning from hybrid modeling to end-to-end, where the latter directly converts input speech sequences into output token sequences using a single network. This breakthrough has overturned all the modeling components that have been used in traditional ASR systems for decades. Next, we will provide a detailed introduction to the mainstream end-to-end speech recognition model architecture: the CTC model architecture [20].
CTC introduces frame by frame potential alignment to represent the alignment between input speech frames and output markers, addressing the need for pre-alignment operations in traditional speech recognition models. The key idea of CTC is to use an intermediate label representation that allows for duplicate and blank labels to identify non-output labels. Due to the fact that the length of output labels in speech is generally smaller than the length of input speech frames, it is necessary to solve the problem of obtaining corresponding labels for each frame. Therefore, in the CTC model, the input audio frame and output labels are in a many-to-one relationship. CTC constructs a CTC path with the same length as the input speech frame by allowing duplicate output labels and inserting a Blank Label marker, such as "_", between duplicate labels. Blank labels are used to prevent false removal of identical labels when removing duplicate labels, such as "xx_xxxk", which ultimately outputs "xxk" instead of "xk". We represent the speech input sequence as x, the original output label as y, and set all CTC paths that can be mapped to y as Q(y). Encoder network maps input acoustic features to higher-level representations.

CTC loss can be effectively calculated using a forward-backward algorithm, but it still predicts the target for each frame and assumes that the target is conditionally independent. Therefore, the loss function of CTC is defined as the sum of all negative logarithmic probabilities that can be mapped to the correct label in the following formulas.

\[
\text{Loss}_{\text{CTC}} = -\ln P(y \mid x) \\
P(y \mid x) = \sum_{\pi \in Q(y)} P(\pi \mid x)
\]

Where \(\pi\) represents a CTC path, assuming conditional independence, the probability of obtaining a path in the following formula:

\[
P(\pi \mid x) = \prod_{i=1}^{T} P(\pi_i \mid x)
\]

Where T is the input length of the speech sequence, CTC is the first widely applied end-to-end algorithm for speech recognition, avoiding the complex forced alignment operations in traditional speech recognition.

III. SINGING WITH NOTE-NAME RECOGNITION ALGORITHM BASED ON END-TO-END MODEL

Traditional speech recognition systems usually combine acoustic and language models for decoding during the recognition stage to fully utilize the linguistic knowledge of external language models. Due to the independent and irrelevant assumption output by the CTC model, it is believed that the predicted samples at each time are independent and ignore the correlation between speech information. Therefore, if a language model can be added to the CTC, the impact of this unreasonable assumption can be improved, thus, collaborative training of language models and acoustic models was achieved. This section first introduces the overall framework of the model, and then elaborates on modules such as the acoustic model and the note pronunciation dictionary model in detail.

![Figure 2: The Model Framework for Singing with Note-name Recognition](image)

A. Modeling Framework

The specific process of singing with note-name recognition is as follows in the figure 2. Firstly, preprocess the singing sound data, including pre-emphasis, framing, and windowing to reduce the impact of noise, and extract the acoustic features of the singing speech signal. The acoustic model of singing speech is used to obtain the state sequence corresponding to acoustic features and generate corresponding acoustic model scores. Singing pronunciation dictionary generates candidate word sequences based on the mapping relationship between phonemes and notes. Finally, based on the acoustic model score, the candidate note sequence with the highest score is obtained through decoder processing, which is the recognition output note sequence. As shown in the figure 2.
B. Acoustical Model of Singing with Note-name

The processing process of traditional speech recognition models is complex. In recent years, with the improvement of computing power and the expansion of data resources, end-to-end (CTC) speech recognition systems have become a current research hotspot. The model based on CTC regards speech recognition as a classification problem, where each acoustic input frame corresponds to an output label. Using duplicate labels and blank labels to identify acoustic frames without output labels can effectively solve the alignment problem.

The CTC-CNN acoustic model belongs to the shallow layer model and has limited effectiveness in speech recognition. In order to further improve recognition performance, residual structures are applied to CTC-CNN. This paper introduces a new CTC-DCNN acoustic model. In this model, residual blocks can transfer input features to the output part through shortcuts, allowing the multi-layer convolutional speech features to be preserved as much as possible. At the same time, the deep structure can also better achieve the extraction and analysis of speech features. A new and improved CTC-DCNN acoustic model is obtained by optimizing the maxout function. The speech recognition flowchart is shown in the figure 3.

![Flow Chart](image)

Figure 3: Flow Chart

In the singing with note-name recognition system of this paper, the training process of the designed CTC-DCNN model is as follows:

1. After preprocessing and feature extraction, the corresponding feature vectors are obtained for the singing speech.
2. Input the feature vectors of singing speech into a deep convolutional neural network, and the convolutional layer adopts a residual structure. After six layers of convolutional operations, the error of the original features is small, fully extracting local information of Chinese speech features. Among them, all steps in the convolutional layer are set to 1 × 1; Conv1 convolution kernel size 1 × 1. Implement dimensionality reduction of speech signals.
3. Input the extracted features into the pooling layer for maximum pooling sampling.
4. In the input 2-layer fully connected layer, each layer has 1024 nodes, and the convolution operation obtains the corresponding posterior probability.
5. The softmax layer adopts an end-to-end CTC structure, which classifies the input singing speech, finds corresponding phoneme sequences, and obtains the output results. The end-to-end CTC structure is used to classify and optimize the recognition process, and finally the state of the obtained features is output. CTC can further improve the robustness of the model and accelerate decoding speed.

IV. CONCLUSION

The paper aims to address the issue of singing with note-name evaluation in professional music education for solfeggio and ear training. Artificial intelligence algorithms are used to identify the roll of the singing with note-names and evaluate the singer's singing level. The resulting scoring results should be more objective. The specific research work of this paper is as follows:
Construction of end-to-end acoustic model based on CTC-DCNN. The CTC algorithm was used to achieve end-to-end processing of speech recognition acoustic models, and DCNN was used as the input network for research. In terms of system input, traditional speech features have been organized into a form suitable for DCNN. At the same time, in order to fully utilize the feature extraction ability of DCNN and avoid the defect of traditional features overly relying on prior knowledge, spectrogram features containing more original speech information have been introduced.

The paper conducts in-depth research on end-to-end speech recognition based on DCNN and CTC, and provides improved methods for input network, which has improved the performance of singing with note-name recognition to a certain extent. However, there are still several aspects that need further research:

1. In practical environments, the signals received by speech recognition systems may contain a lot of noise, and the noise and background noise in the sound can lead to a sharp deterioration of the system's recognition performance. In order to complete the recognition task, it is usually necessary to use front-end technology of speech signals for noise reduction and other processing. Therefore, how to improve the robustness of the model and enable speech recognition systems to achieve better recognition performance in noisy environments is also a key direction that needs to be broken through in the future.

2. With the development of internet information technology, the amount of available singing data is increasing day by day. In order to fully utilize this data, the scale of singing with note-name recognition systems is becoming increasingly large. Although the introduction of end-to-end mechanisms solves some of the shortcomings of traditional methods, the use of large granularity modeling units will further exacerbate the requirements for data volume and model size. The increase in model size means that stronger computing power is needed to ensure recognition efficiency, which is difficult to meet in places with limited computing resources such as mobile devices. Therefore, more in-depth research is needed on compression algorithms for acoustic models.

ACKNOWLEDGMENT

This work was supported by projects grant from Jiangxi education science and technology research (Grant No.GJJ20203304 and GJJ160324).

REFERENCES


