Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 2, No. 3, 2022

Research on Algorithm of Video Analysis System Based on Text Error Correction

Jinjin Wang*, Yang Qin, Jiahao Shi, Jiachen Luo, Guo Huang, Jiaqi Lu

School of Information Engineering, Yancheng Teachers University, Yancheng, 224000, China
* **Corresponding author**: Jinjin Wang (Email: wangjj@yctu.edu.cn)

Abstract: When making a video, if the video has a language organization error, it needs to be re-recorded. It is not possible to remove inappropriate or unnatural pronunciation parts of the recording more effectively. In response to this problem, this paper studies the speech extraction, error correction and synthesis of video, which is divided into three parts: (1) Speech segmentation and speech-to-text of video; (2) Text recognition error correction; (3) Text-to-speech and video speech synthesis. For the first part, we applied the staged and efficient algorithm based on (Bayesian Information Criterion) BIC & (Statistical Mean Euclidean Distance) MEdist to segment the video voice, and then, the segmented audio is subtracted to reduce noise, and finally converted to text using the iFLYTEK interface. For the second part, we apply the (Double Automatic Error Correction) DAEC algorithm to text error correction. For the third part, we use the (Improved Chinese Realtime Voice Cloning) I-Zhrtvc for text-to-speech. Then merge the voice into the video. The simulation result shows that the staged and efficient algorithm based on BIC & MEdist, which accurately segmented by sentences, can identify audio with dialect accents, and has high accuracy in translating to text, up to an average of 95.8%. DAEC algorithm has a high error correction rate. The audio prosody accuracy after synthesis is high. ZVTOW text-to-speech (Mean Opinion Score) MOS up to 4.5.

Keywords: Speech segmentation; Text recognition error correction; Text-to-speech; Video speech synthesis.

1. Introduction and Related Work

At present, the courseware content and artificial voice in the online recording of educational videos often appear as a whole, and when the video needs to be processed, more professional multimedia tools and teams are often required to complete the processing. What knowledge points need to be explained for how long, how long do you need to code demonstration cases, whether the language organization is reasonable, there is no more effective tool to assist the analysis, and often the recorder often re-records after the language organization error, and cannot more effectively remove the inappropriate or unnatural pronunciation part of the recording.

Therefore, video analysis schemes based on text error correction have become a research hotspot.Chen Lijiang et al. [1] studied the synthesis of speech with corresponding text and EGG signals, and obtained a Mel cepstral distortion (MCD) of 5.877 and an average opinion score (MOS) of 3.87 based on the improved Tacotron-2 model, and achieved an improvement of 0.42 with a relatively small model size. A fine-grained fundamental frequency modification method is proposed, which adjusts the fundamental frequency according to the EGG signal. Compared with the unmodified method, the MCD is lower at 5.781 and the MOS is higher at 3.94.Jin-song Zhang et al. [2] studied Robust tone recognition based on multi-level core framework; The contribution of tone information to pinyin-to-text conversion, based on mutual information to find tones depends on information phonemes. These studies and their results are applicable and instructive for the development of tone processing techniques in Chinese CAPT techniques. A. Mouchtaris et al. [3] address the more general problem of multichannel audio synthesis by extending the model used for the recomposition problem, namely how to fully synthesize a multichannel recording from a particular stereo or mono recording, by adapting the recomposition transformation parameters to the statistical properties of the record that we wish to enhance. When a particular model is applied to a different context (speaker, language, or channel), this parameter adaptation is similar to task adaptation used in speech recognition. Mahmut Emilian-Erman et al. [4] proposes a method of audio file segmentation that attempts to alleviate the problem of different individuals having different durations for the same utterance. A method for determining the maximum cross-correlation value between two audio files and a subsequent automatic segmentation method are described in order to extract two valid sound samples of target consonants, with the aim of preprocessing the audio file and feeding the evenly trimmed audio samples to a computerized SSD screening system. Dabbabi Karim et al. [5] propose an optimized audio classification and segmentation algorithm for dividing stacked audio streams into 10 main audio types based on their content. They tested KNN, SVM and GASOM algorithms on two audio classification systems. In the first system using GASOM algorithm and leave-one verification technique, the average accuracy of music/ambient sound discrimination reached 99.17%. Compared with KNN and SVM algorithms, GASOM algorithm always achieves the best performance results.

2. Video Analysis

2.1. Speech segmentation and speech-to-text of video

In terms of audio segmentation, we applied the staged and efficient algorithm based on BIC & MEdist to segment the video voice. firstly, the audio separated from the video is subjected to spectral reduction and noise reduction processing; then, the noise-reduced speech is segmented in stages; finally, call iFLYTEK Interface, complete voice-to-text.

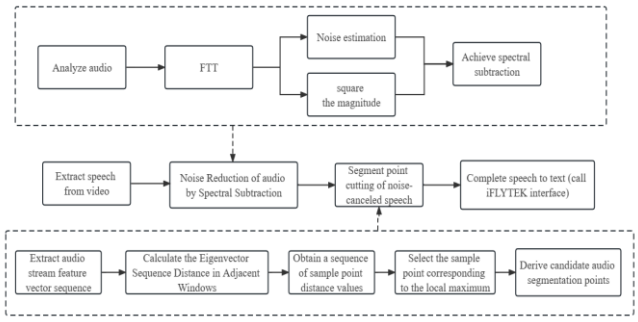

Figure 1. The process of staged efficient algorithms based on BIC and MEdist

2.2. Text recognition error correction

In this part, we use the DAEC algorithm to correct the text. first, automatic error correction is performed on the converted text (manual error correction is also supported); Then, the corrected text is audited, including hot word frequency audit, video (and corresponding voice and text) modification (editing) times and location audit, video upload and video editing number proportion audit, etc. Finally, the audit results are visualized in the developed teaching video voice extraction and audit system. As shown in Figure 2.

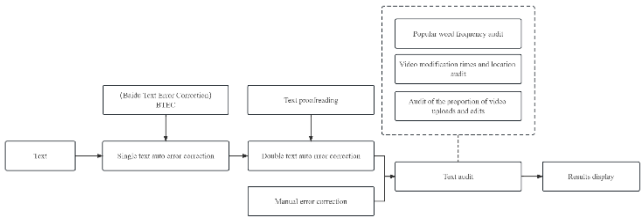

Figure 2. The DAEC Algorithm Video Voice Text Audit Solution

Model input: the input embeddings of the entire model framework are composed of the addition and embedding of word embedding, position embedding, and segment embedding of each character in the text sentence. Therefore, it can be seen that the input of the model framework of the whole method is actually the same as the general input form of the DAEC algorithm.

$$e_i = word\ embedding(e_i) + position\ embedding(x_i) + segment\ embedding(x_i) \quad (1)$$

Detection Network: The Detection Network in the DAEC algorithm framework is essentially a bidirectional GRU model (Bi-GRU). The Bi-GRU model encodes each text sequence forward and backward, and then merges the hidden state of the forward coding of the last hidden layer Chinese the sequence horizontally with the hidden state of reverse encoding, and the calculation process of the Bi-GRU model is as follows:

$$\overrightarrow{h_i^d} = GRU(\overrightarrow{h_{i-1}^d}, e_i) \quad (2)$$

$$\overleftarrow{h_i^d} = \mathrm{GRU}(\overleftarrow{h_{i+1}^d}, e_i) \quad (3)$$

$$h_i^d = [\overrightarrow{h_i^d}\ ;\ \overleftarrow{h_i^d}] \quad (4)$$

h_i^d is the hidden state of the embedded ei of the character i in the text sequence in the last hidden layer after being calculated by the Bi-GRU model? The number of hidden layer dimensions of the reference Bi-GRU model is set to 256, and the number of hidden layer output dimensions after bidirectional encoding is 512. The hid calculated by the Bi-GRU model is then fed into two fully connected layers.

Loss function: As shown in the following schema, $\mathcal{L}_d$ represents the cross-entropy loss calculated by the final output value of the Detection Network of the error detection network; $\mathcal{L}_c$ represents the cross-entropy loss calculated for the output value after the error correction network.

$$\mathcal{L}_d = -\sum_{i=1}^{n} \log Pd(g_i|X) \quad (5)$$

$$\mathcal{L}_c = -\sum_{i=1}^{n} \log P_c(y_i|X) \quad (6)$$

The loss function of the whole DAEC algorithm is composed of the loss function of the Detection Network of the error detection network and the loss function of the error correction network Correction Network:

$$\mathcal{L} = \lambda \cdot \mathcal{L}_c + (1-\lambda) \cdot \mathcal{L}_d \quad (7)$$

The above equation is the loss function representation of DAEC algorithm. In the formula, 1−λ and λ are the linear combination coefficients of the loss function of the Detection Network and the loss function of the Correction Network, that is, the linear combination of the respective loss functions of the two networks is the total loss function of the final Soft-Masked BERT model.

2.3. Text-to-speech and video speech synthesis

In this part, we use the I-Zhrtvc for text-to-speech. Then merge the voice into the video.

The prototype is zhrtvc, using machine learning algorithms, zhvoice is the Chinese speech corpus, the voice is clearer and more natural, including 8 open source datasets, 3200 speakers, 900 hours of speech, 13 million words. Large corpus splicing speech synthesis system based on statistical rules ultra-large-scale sound library production: corpus design; Soundbank recording; fine segmentation; prosodic callouts; The sound quality is good, the difference in sound quality of recording synthesis is small, and the naturalness of normal sentences is good. The zhvoice corpus can be used to train the base model of speech cloning. ge2e_pretrained_iwater.pt: A speech encoder model trained with Chinese open-source speech corpus. Mel-Spectrogram: Vocoder

The process of I-Zhrtvc for text-to-speech and merge the voice into the video as shown in Figure 3. first obtain the text that needs to be replaced; Then, the text-to-speech encoder retrained based on the fusion Data_aishell dataset automatically converts text into speech (and also supports manual recording of speech). Then, the converted voice is normalized and integrated with the original voice intensity, prosody and duration. Finally, the fused audio is merged with the silent video to get the corrected video.

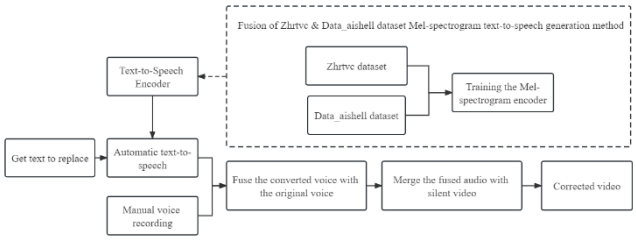

Figure 3. The process of I-Zhrtvc for text-to-speech and merge the voice into the video

3. Experiment and Result Analysis

3.1. Experimental results of the staged and efficient algorithm based on BIC & MEdist

Compared with traditional BIC audio segmentation

algorithm,the staged and efficient algorithm based on BIC & MEdist is Proposed in this paper. which discards the GLR (General Likelihood) Comparable distance calculation formulas such as (Kullback-Leibler)KL, etc., use the statistical mean Euclidean distance MEdist, combined with the local maximum selection and Significance detection. Then use the BIC method to confirm the candidate segmentation points. The test results show that this method not only improves the overall audio segmentation speed to a large extent, it is 400 times higher than the traditional BIC method, and the deletion error rate MDR is reduced by 15.1%. The insertion error rate FAR is slightly reduced, which greatly improves the accuracy of text-to-text conversion after speech extraction.

Audio split point insertion error (FA) means that a split point that does not actually exist is detected; audio split point deletion error (MD) means that a split point that actually exists is not detected. FA (False Alarms Rate) and MDR (Missed Detection Rate) are defined as follows:

$$FAR = \frac{FA}{ACP+FA} \times 100\% \quad (8)$$

$$MDR = \frac{MD}{ACP} \times 100\% \quad (9)$$

Among them, ACP (Actual Changing Points) represents the real audio split point. Similarly, we can define the total error score ES (Error Score)

$$ES = FAR + MDR \quad (10)$$

The specific algorithm comparison chart is shown in the figure 4.

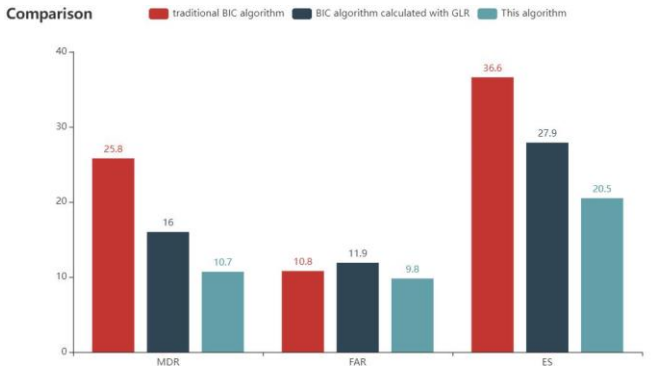

Figure 4. The specific algorithm comparison

Table 1. Accuracy comparison of speech extraction algorithms

	Total number of test cases	Speech-to-speech error rate <10%	Speech-to-speech error rate <5%	Speech-to-speech error rate <3%
New Design Speech Extraction Algorithm	28	100%	92.80%	89.20%
Traditional Speech Extraction Algorithms	29	36%	10.70%	0.00%

In terms of speech-to-text noise reduction in speech extraction, by analyzing the audio and processing the audio through FFT (Fast Fourier Transform), noise estimation and amplitude quadratic calculation are performed to achieve spectral reduction, which can not only Noise reduction, it also has a certain ability to recognize some audio with dialect accents except Mandarin Chinese. After 20,000 audio tests, we concluded that the audio after noise reduction can have an average recognition accuracy of 95.8%, which can improve the accuracy of text conversion.

Combining the above two points and testing a set of Java videos (28), compared with the traditional algorithm, the new algorithm can extract language with an accuracy of 95.8%. There is no error in converting most audio segments to speech, and the error rate in a small number of use cases can also be controlled below 20%. However, based on the traditional speech extraction algorithm, there are large errors in the speech extraction of audio segments with high noise or audio segments with low segmentation accuracy, and the correct rate is only 35.3%. The test results are shown in Table 1.

3.2. Experimental results of the DAEC algorithm

The video voice text audit scheme corrected by DAEC algorithm has a high error correction rate. In 20000 text error correction tests, the correct rate of normal text and spoken text of the new double error correction method has increased by 10.8% and 15.3% respectively compared with BERT, and by 0.7% and 2.0% respectively compared with Baidu, As shown in Table 2.

Table 2. Comparison of algorithms

	Normal Text Error Correction Rate	Correctness Rate of More Spoken Texts
DAEC	98.30%	96.60%
Baidu	97.60%	94.60%
BERT	77.50%	81.30%

3.3. Experimental results of I-Zhrtvc

Based on the innovative zhrtvc algorithm, this paer retrained with Chinese data from 400 people in Data_aishell, and in Vocoder, Mel-spectrogram generated by Synthesis-trained models was used instead of the original Griffin-Lim. In addition to the realization of automatic text-to-speech (TTS), the improved algorithm also improves the prosody accuracy by 15% compared with the previous Pyttsx, and the MOS value reaches 4.5. The MOS calculation formula is as follows:

$$\text{MOS} = \frac{\sum_{j=1}^{M}\sum_{i=1}^{N} s_{ij}}{M} \quad (11)$$

By collecting the 100 use cases outlined, compared with the traditional Pyttsx speech-to-speech algorithm, the newly designed text-to-speech pronunciation accuracy can reach 95%, most of the use cases pronunciation standards are coordinated with prosody, a small number of prosody and pronunciation errors, and a few use cases have large deviations. Based on the traditional Pyttsx to speech algorithm, the algorithm has a large error in the prosody results of long or complex text-to-speech. As shown in Table 3.

Table 3. I-Zhrtvc vs. Pyttsx pronunciation, prosody, dictionary accuracy comparison results

	Total number of test cases	pronunciation accuracy	Prosodic accuracy	dictionary accuracy
I-Zhrtvc	100	95%	85%	83%
Pyttsx	100	87%	70%	80%

In addition, our team's I-Zhrtvc algorithm with reference to Zhrtvc can achieve a Mos value (voice quality evaluation index) of 4.5, as shown in Table 4. By properly training the Gmw vocoder, it sounds very natural, and the Dbplus

algorithm is used to level the average db of the audio, and the audio time is used a. Speed is the same duration as the original audio, sounds complete without obvious abnormal rhythmic ups and downs, is relatively clear and smooth, and is easier to understand, reaching the quality of people's ordinary conversations, and listeners are willing to accept it.

Table 4. Comparison table of MOS values of various mainstream TTS algorithms

Name	I-Zhrtvc	gTTS	Pyttsx	Win32com
MOS	4.5	4.2	3.8	4.0

4. Conclusion

This paper studies video analysis based on text error correction from three aspects: speech segmentation and speech to text, text recognition and error correction, text to speech and video speech synthesis. Bayesian Information Criterion BIC and Statistical Mean Euclidean Distance will be used. MEdist's phased efficient algorithm, DAEC algorithm and I-Zhrtvc algorithm are applied to video analysis research based on text error correction. Experiments show that the video analysis scheme based on text error correction proposed and designed in this paper has high efficiency and accuracy.

References

[1] Chen Lijiang,Ren Jie,Chen Pengfei,Mao Xia,Zhao Qi. Limited text speech synthesis with electroglottograph based on Bi-LSTM and modified Tacotron-2[J]. Applied Intelligence,2022,52(13).

[2] Jin-song Zhang & Wen Cao Center for studies of Chinese As a Second Language College of Information Sciences Beijing Language University No. 15, Road Xueyuan, Haidian, Beijing 100083, P. R. China. Tone Information Processing for Chinese Automatic Speech Recognition and A Discussion of Its Application to Computer Aided Pronunciation Training.

[3] A. Mouchtaris,S.S. Narayanan,C. Kyriakakis. Multichannel audio synthesis by subband-based spectral conversion and parameter adaptation[J]. IEEE/ACM Transactions on Audio Speech and Language Processing,2005,13(2).

[4] Mahmut Emilian-Erman,Nicola Stelian,Stoicu-Tivadar Vasile. Cross-Correlation Based Automated Segmentation of Audio Samples.[J]. Studies in health technology and informatics,2020,272.

[5] Dabbabi Karim,Cherif Adnen,Hajji Salah. An Optimization of Audio Classification and Segmentation using GASOM Algorithm[J]. International Journal of Advanced Computer Science and Applications (IJACSA),2018,9.