Multimedia Tools and Applications
https://doi.org/10.1007/s11042-022-12136-3

A hybrid CTC+Attention model based on end-to-end framework for multilingual speech recognition

Sendong Liang[1] · Wei Qi Yan[1]

Received: 16 May 2021 / Revised: 2 August 2021 / Accepted: 3 January 2022

Abstract

Speech recognition is an important field in natural language processing. In this paper, the end-to-end framework for speech recognition with multilingual datasets is proposed. The end-to-end methods do not require complicated alignment and construction of the pronunciation dictionary, which show a promising prospect. In this paper, we implement a hybrid model of CTC and attention (CTC+Attention) model based on PyTorch. In order to compare speech recognition methods for multiple languages, we design and create three datasets: Chinese, English, and Code-Switch. We evaluate the proposed hybrid CTC+Attention model in multilingual environment. Throughout our experiments, we find that the proposed hybrid CTC+Attention model based on end-to-end framework achieves better performance compared with the HMM-DNN model in a single language and Code-Switch speaking environment. Moreover, the results of speech recognition with regard to different languages are compared in this paper. The CER(i.e., Character Error Rate) of the proposed hybrid CTC+Attention model based on the Chinese dataset defeated the traditional model and reached 10.22%.

Keywords Speech recognition · End-to-end framework · Attention model · CTC model · Code-Switch

1 Introduction

Spoken language is essential to modern human cultures. Speech is a way of social communications amongst people through languages. In the past decades, with the development of intelligent devices, the developed applications began to enter people's daily life. While

✉ Wei Qi Yan
weiqi.yan@aut.ac.nz

Sendong Liang
ghg0412@autuni.ac.nz

[1] School of Engineering, Computer & Mathematics, Auckland University of Technology,
No. 31 Symonds Street, Auckland 1010, New Zealand

Published online: 20 May 2022

people communicate with intelligent devices through languages, automatic speech recognition(ASR) plays pivotal role. ASR aims to convert audio data into corresponding text, the text is further processed through human-computer interaction, such as multilingual translations and hand talk, etc.

In the early stage, because it was impossible to directly model the audio-to-text conversion, Bayes' theorem was implemented to convert human speech into text so as to calculate the corresponding audio features. Accordingly, the probability of audio feature sequences is decomposed to the product of conditional probabilities of corresponding audio features.

With the development of HMM, speech recognition has transited from seperated words of a small system for speech recognition to a large vocabulary continuous system nowadays [40]. The framework based on HMM model for speech recognition shows its excellence and reliable stability, which was the mainstream speech recognition model.

Classic HMM speech recognition model assumes that the model state transferring has the homogeneous Markov property. The HMM needs to be trained based on a set of speech sequences and requires a larger dataset. It is said that smaller models are easier to understand, but larger models can fit the data [31]. By considering the non-stationary process with the range of frequency and time of the speech signals, HMM models have weak robustness performance because they focus on the temporal analysis.

Deep neural network has contributed to acoustic models and formed the HMM-DNN speech recognition framework [18]. The output of objective function of the DNN-based acoustic model is the probability of a HMM state given a sequence of audio features. With more and more research work dedicated to deep learning models, the ability of acoustic models becomes stronger and stronger. A classification network has been directly created from the HMM-DNN acoustic model [8].

Owing to the soaring deep learning, the end-to-end frameworks for speech recognition have shown exemplary performance in high-resource languages. However, it is hard to perform well in low-resource datasets for speech recognition, such as Chinese-English and Code-Switch environment [41]. In this paper, we will devote to bridge the gap and provide our solution for resolving this problem.

Up to date, our daily communications often are surrounded by mixed languages, which is academically named as Code-Switch model. For example, a lot of Chinese people mingle English words in a Chinese sentence. The speech blended with the words from multilingual is one of the critical challenges in speech recognition. The main technical difficulty also includes the non-native accents, the composition brings difficulties to model the mixed acoustics, meanwhile, the labelled datasets for the mixed speech recognition are extremely scarce.

Traditional phonetic framework is based on basic acoustic unit for language recognition. The linguistic information is various for different languages, such as Chinese phonetic consonants and English phonemes. The framework relies on specific linguistic knowledge and is tough to be expanded to multilingual speech recognition. The end-to-end frameworks employ a unified network for modelling and are dependent more on datasets than linguistic information. Accordingly, we have a great interest in using the end-to-end framework to resolve this emerging speech recognition problem.

The focus of this paper is on speech recognition using our labeled datasets and multilanguage environment in the end-to-end framework. On the basis of related work, four experiments will be implemented with the datasets and multiple models. The corresponding results will also be analysed. The contributions of this paper include: (1) Investigating the performance of a hybrid model for speech recognition based on our labelled datasets; (2) Exploring an end-to-end framework for speech recognition with multi-language datasets;

(3) Comparing the end-to-end framework based on hybrid CTC+Attention model with the traditional speech recognition model. Moreover, the experimental results will be compared and analysed.

The structure of the proposed CTC+Attention hybrid model is split into three main parts. The first part is pre-net, which is composed of Deep CNN inspired by using the VGG structure. The second part is the encoder shared by CTC and Attention. The last part is a joint decoder, which is composed of a CTC decoder, an attention decoder, and a language model. The architecture of the CTC/Attention hybrid model is shown in Fig. 1.

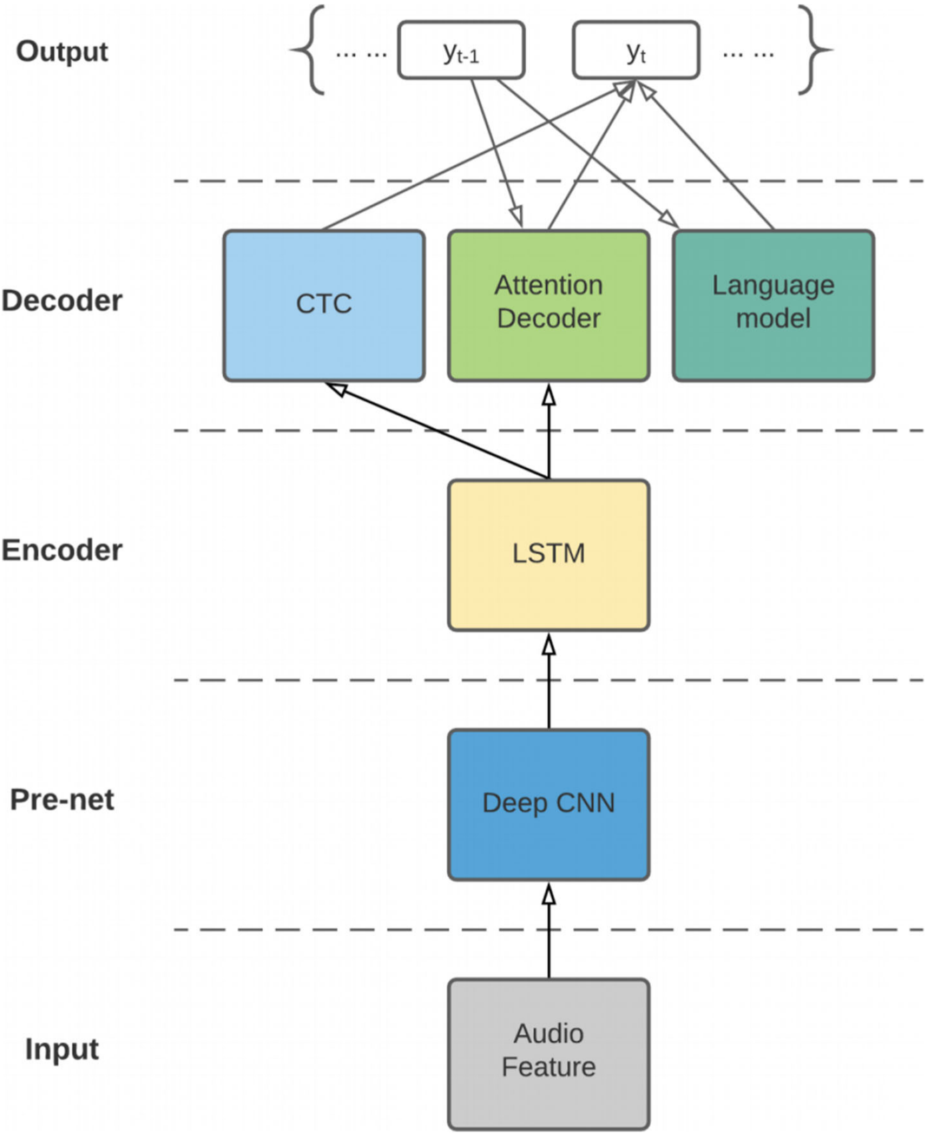


Fig. 1 The structure of hybrid model

The remaining part of this paper is organized as follows. We have our literature review in Section 2, our method is presented in Section 3, our result analysis is shown in Section 4, our conclusion is drawn in Section 5.

2 Literature review

Speech recognition models are grouped into three categories. The first category includes rule-based models, such as Shoebox and Harpy created by IBM and CMU in 1962 and 1976, respectively [11]. The second group encompasses statistical models such as Large Vocabulary Continuous Speech Recognition (LVCSR) and HMM. HMM-Gaussian Mixture Model (GMM) has been the dominant framework in the field of speech recognition till the emerge of deep learning [37]. The third one is deep-learning-based models such as Deep Neural Networks (DNNs), Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM), and Bidirectional LSTM (BiLSTM) [26, 32]. LSTM is sensitive to the static data, which may lead to delays with respect to features, BiLSTM appears as a special architecture operating the input sequence in both directions [17].

In 2006, an unsupervised method was employed to pretrain the Deep Belief Networks (DBNs), which solved the problem that gradient descent is sensitive to the initial value [9]. DNNs have been applied to acoustic modelling to form the HMM-DNN framework in speech recognition [18, 32]. The objective function of DNN-based acoustic models is the probability of a HMM state given a sequence of audio features. In 2009, DNNs were applied to the TIMIT phoneme recognition and achieved excellent performance [23]. In 2012, CNNs were applied to the LVCSR, which normalized the data and obtained a higher performance in speech recognition [1].

Moreover, an end-to-end framework for speech recognition based on RNNs and Weighted Finite State Transducer (WFST) decoding method was proposed in 2015. This end-to-end framework directly utilises analogue signals as its input, which improves the recognition rate [21] instantaneously. In 2017, a new method of end-to-end speech recognition was proffered, which employs CTC model in a multitask framework to improve robustness of the system and achieves quick convergence, therefore alleviates the problem of data alignment [12].

However, most of the previous research work in speech recognition focuses on better performance of speech recognition for a single language or a single task. Pertaining to comparison of speech recognition between multiple languages, the publications of speech recognition in a multilingual or Code-Switch environment are relatively rare.

Feedforward Neural Networks (FFNNs) are dependent on the preceding words, which cannot learn the dependent information of long sentences. With development of deep learning, RNNs turn up, which aims at solving the issue of long sequence dependence. RNNs are defined as a type of ANNs that make classification and predictions for various data, such as text, audio, video, genomes, etc. [19].

Compared with traditional FFNNs, such as multilayer perception, using static classifiers and only considering fixed-size input windows which are irrespective of surrounding context, RNNs are more effective and suitable to transcribe time series such as speech transcription because of its hidden network layers [6]. Different from FFNNs using fixed-length context, RNNs do not take use of a limited size of context, which contain cache models that encode temporal information implicitly with arbitrary lengths [22]. The recurrent connections allow information to cycle inside networks for a long time adapting to the past inputs [2]. A continuous vector space best suits for word representation, meanwhile, deep learning methods reflect the relationship between continuous words. Accordingly, compared with FFNNs, RNNs best solve the contextual dependency problem which spans over a fixed number of predecessor words [33].

The end-to-end framework for speech recognition refers to directly transduce the input sequence of acoustic feature vectors to the output sequence of token such as phonemes, characters, or words [14]. The end-to-end model is split into three categories based on the alignment methods: Connectionist Temporal Classification (CTC), Attention Encoder-Decoder (AED), and RNN Transducer (RNN-T), which have been widely utilized in large-scale speech recognition [14]. The end-to-end models are more suitable for on-device applications than conventional speech recognition because there are fewer parameters by folding the acoustics, pronunciation, and language models into one neural network [13].

Attention-based encoder-decoder model such as LAS (Listen, Attend, and Spell) contains three main components: Encoder as an acoustic model, attender as an alignment model, and decoder as a language model [4], which subsumes the components of acoustics, pronunciation and language models into a single neural network without a lexicon or a text normalization component [5].

Applied to Google voice search, the proposed model achieves a WER of 5.6%, while the hybrid HMM-LSTM model attains 6.7% WER. Throughout testing the same models based on dictation, the proposed model reaches up to 4.1%, the HMM-LSTM model gets 5% WER. The decoding process of sequence-to-sequence (S2S) models with soft attention incurs a quadratic time cost, which is regarded as a challenge for online sequence transduction [5].

In order to address the online streaming challenge of the attention-based model, monotonic-chunkwise attention was put forward, which splits an input sequence into a number of small chunks [5]. Triggered attention equipped with the CTC-based classifier performs well to control the activation of the attention-based decoder [24].

3 Methods

3.1 Data preparation

3.1.1 Language features

By considering English and Mandarin are worldwide widely used languages in the low-resource environment of bilingual Code-Switch corpus, it is important to select appropriate speech units for acoustic modelling, which convert a speech unit to a corresponding feature vector sequence [30].

English is an Indo-European language, while Mandarin is a Sino-Tibetan language [3]. Based on the Oxford Dictionary, English is written in Latin alphabets, namely, the Roman alphabets, which contains 26 letters and nearly 170,000 words. The general modelling units in English include phone, subword, and character [38].

In the field of Chinese speech recognition, there are various available acoustic modelling units, including Chinese characters (word), syllable (syllable), semi-syllable (initial/final), phoneme (phone), which is generally based on phonetic knowledge or data-driven generation [43]. Mandarin contains more than 6,000 characters, 60 phonemes, 408 atonal syllables, and 1,302 toned syllables [16]. Each syllable includes initials, finals and tones. In total, Chinese (Mandarin) has 22 initials and 39 finals of syllables.

Meanwhile, there are a plenty of homophones and polyphones in Chinese [44], which may need high-level non-acoustic context knowledge for speech recognition. A small flexible unit may lead to the difficulties in calibrating the training dataset. By contrast, the limitation of flexibility as well as the high requirement of lexicon and out of vocabulary (OOV) [35] are significant challenges albeit the large unit with high recognition performance. The syllable unit meets the requirement and flexibility [29]. Moreover, the utilization of syllables with tones effectively increases the recognition accuracy compared with Chinese characters and syllable initial/final with tones [7].

3.2 Corpus design

Based on the language features in Section 3.1.1, three speech datasets are applied to our experiments, which include dataset Alpha (Mandarin), dataset Beta (English), and dataset Gamma (Mandarin-English).

We create the three datasets through the text-to-speech iFLYTEK InterPhonic toolkit that is a software developed by iFLYTEK [10], which converts text into male or female voices. The toolkit is based on an advanced large corpus and a phonetic prosody description, the quality of the synthesized voices in $.wav$ format files is comparable to that of a real person.

Converting the transcript text to synthesised acoustic speeches controls the variables. For example, speaker's accents, emotions, and environmental noises are consistent. The focus of our experiments by using the synthesised acoustic audio is on the recognition rates within the multilingual environment without considering useless variables or factors.

3.2.1 Dataset alpha (mandarin)

Alpha is a Mandarin speech dataset. We created this dataset through the iFlytek toolkit based on the transcripts of the THCHS-30 corpus. In the THCHS-30 corpus, there are numerous transcript files corresponding to the acoustic files. Each file contains three rows, which is the correctly-labeled transcript of the corresponding acoustic sentence. The first row is the Chinese characters of the corresponding acoustic sentence. The second row is the Chinese Pinyin with a tone, which corresponds to the Chinese characters. The third row is the syllable initial-final tone.

As shown in Fig. 2, we selected 8,000 files of the THCHS-30 corpus. The first row (Chinese characters) of each selected file are collected as the input of the iFlytek InterPhonic

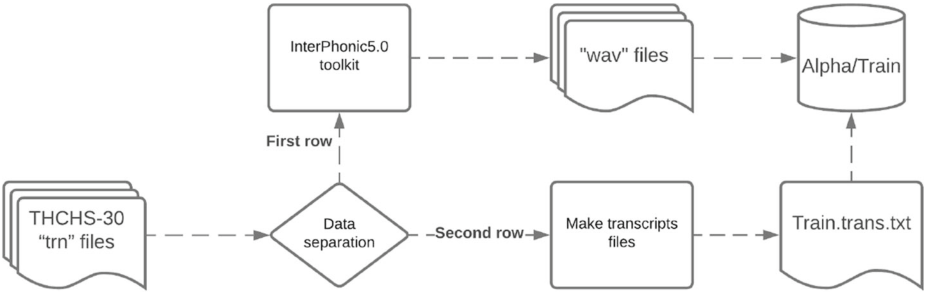


Fig. 2 The process of creating alpha-train

toolkit. The selected Chinese characters are synthesised into 8,000 files sampled at 16 kHz. The second row (Chinese Pinyin with tones) of each selected file is collected together as one text file as our labeled transcript for training. Throughout this operation process, we created the training subset based on Alpha (Mandarin).

Similarly, we created the "Dev" (development) subset and the "Test" subset from Alpha (Mandarin) through the same process. The "Dev" subset is synthesised from other 1,000 files of the THCHS-30 corpus except for the selected 8,000 files. The "Test" subset is synthesised from other 1,000 transcript files except for the selected 9,000 files.

The synthesised acoustic files in the dataset Alpha (i.e., Mandarin) are split into three groups as shown in Table 1: One-hour acoustic files as the test subset (i.e., Alpha-Test); One-hour acoustic files as the development subset (Alpha-Dev) to train the rapid reaction and performance evaluation, the other 10-hour acoustic files as the training subset (i.e., Alpha-Train).

3.2.2 Dataset beta (english)

Dataset Beta (English) is an English speech dataset. We created this dataset by using the iFlytek InterPhonic toolkit based on the transcript files of the LibriSpeech corpus [25]. The production method is similar to Alpha as shown in Fig. 3. The structure of the Beta dataset is shown in Table 2: 10-hour acoustic files synthesised as the training subset (i.e., Beta-Train), one-hour audio files as the development subset (i.e., Beta-Dev) to train the rapid reaction and performance evaluation, and one-hour audio files as the testing subset (i.e., Beta-Test).

3.2.3 Dataset Gamma (mandarin-english)

Dataset Gamma (Mandarin-English) is a mixed Mandarin-English bilingual speech database. We created this dataset through the iFlytek InterPhonic toolkit based on the transcript files of the TAL-CSASR courpus [34]. The original TAL-CSASR corpus is the audio captured in English class teaching environment. The production method and folder structure of the Gamma dataset are similar to Alpha and Beta. The production method is shown in Fig. 4. The folder structure of the Gamma dataset is as same as Alpha and Beta. The details of each sub-dataset are shown in Table 3.

3.3 Our experiments

3.3.1 Experiment environment and setup

We implement Kaldi [28] in the experiment of traditional speech recognition. The end to end framework takes use of PyTorch and ESPnet [39] frameworks based on LAS [4] in the RNN language model. Based on the inspiration [36], we superimposed mutilhead and location

Table 1 The statistics of dataset alpha (mandarin)

Datasets	Time(Hours)	Speakers	Male	Female	Percentages of THCHS-30
Alpha-Train	10	2	1	1	29.878%
Alpha-Dev	1	2	1	1	2.988%
Alpha-Test	1	2	1	1	2.988%

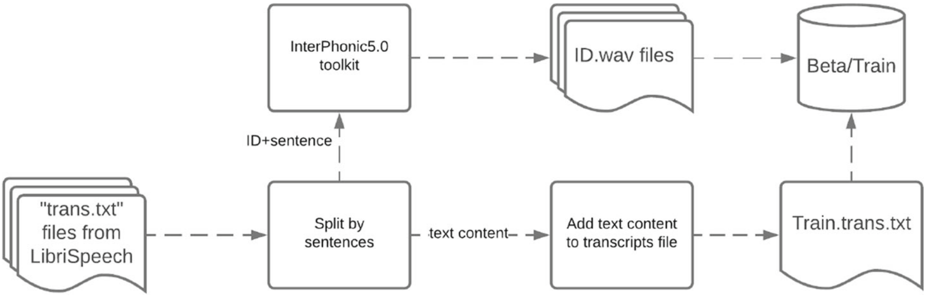


Fig. 3 The process of creating dataset beta-train

Table 2 The dataset beta (english)

Dataset	Time(hour)	Speakers	Male	Female	Percentage of LibriSpeech-clean
Train	10	2	1	1	8.977%
Test	1	2	1	1	0.898%
Dev	1	2	1	1	0.898%

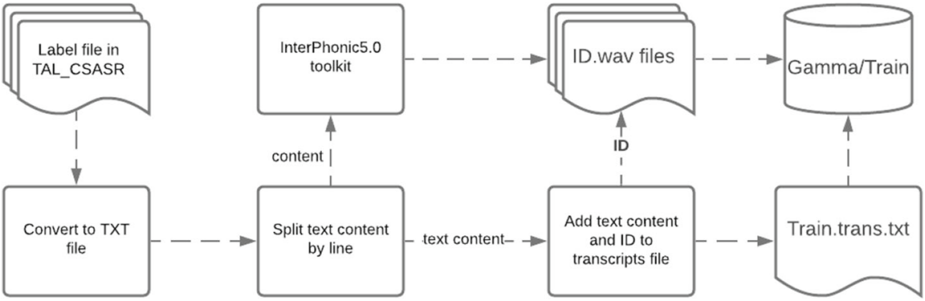


Fig. 4 The process of creating dataset gamma

Table 3 The dataset gamma (mandarin-english)

Datasets	Time(hours)	Speakers	Male	Female
Train	10	2	1	1
Dev	1	2	1	1
Test	1	2	1	1

attention in the hybrid framework. Table 4 shows the details of hardware configuration and operating system.

3.3.2 Evaluation methods

The most straightforward approach to evaluate the performance of speech recognition is to calculate the word error rate (WER) [20]. WER is employed for the English dataset, which is computed through (1).

$$R = \frac{I_E + D_E + S_E}{N} \times 100\% \tag{1}$$

where I_E refers to the insertion number of English words, D_E means the deletion number of English words, S_E stands for the substitution number of English words, and N is the total number of English words in the correct sentence. However, characters are considered instead of words in the Chinese THCHS-30 corpus. Similarly, the character error rate (CER) is calculated through (2), in which the suffixes ended with M take the place of E.

$$R = \frac{I_M + D_M + S_M}{N} \times 100\% \tag{2}$$

3.3.3 Parameter optimization

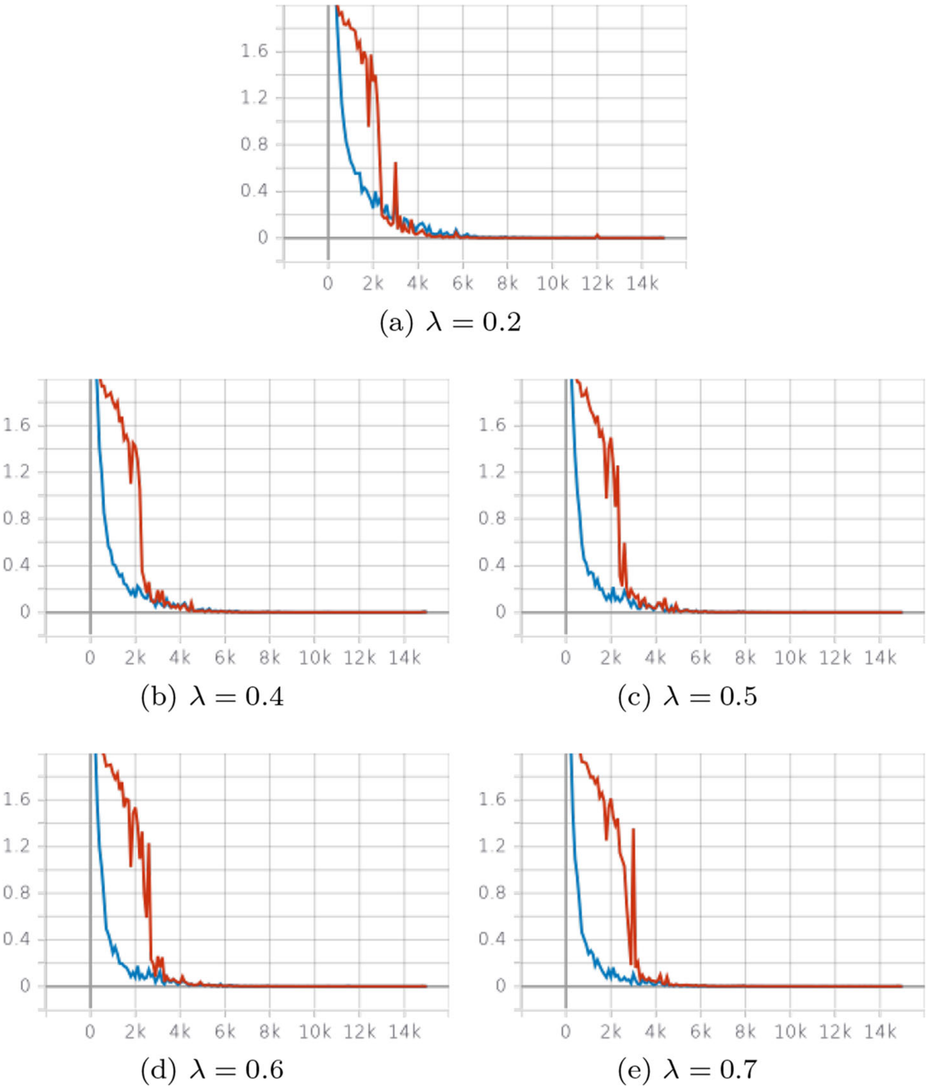
The main idea of the CTC+Attention hybrid model is to utilise CTC to force the alignment of the eigenvectors of the audio frames to reduce a single tag [27]. At the same time, CTC does not allow skipping label output under the same audio characteristics to avoid the frame skipping mentioned in the attention subsection. The score function of the hybrid model is described by using (3) based on the schematic diagram of the combination of CTC and attention formulas. Among them, O_{hybrid} is the prediction result of this model, Y is the text label sequence, X is the feature vector sequence corresponding to the audio frame, λ is the evaluation weight of the CTC model. $\log P_c(Y \mid X)$ is the score function of CTC, and $\log P_a(Y \mid X)$ is the score function of the attention model. Therefore, the optimization process is regarded as finding the optimal solution of λ.

$$O_{hybrid} = (1 - \lambda) \log P_a(Y \mid X) + \lambda \log P_c(Y \mid X) \tag{3}$$

Five model trainings were carried out for λ from 0.2 to 0.7 based on Alpha dataset. The losses with different λ are shown in Fig. 5. The red line is the loss of the attention model, the blue line is the loss of CTC. Judged from the loss of CTC, the convergence speed of the five experiments has been improved. While CTC improves the learning efficiency, the convergence of attention does show relatively large fluctuations. If λ equals to 0.7, the loss of attention has a huge fluctuation in a short period.

Table 4 The hardware equipment of our experiments

OS Name	Ubuntu20.04
OS Type	64-bit
Memory	32 GiB
Processor	Inter(R)Core i9-9900k CPU @ 3.60GHZ x 16
GPU	GeForce RTX 2080Ti 11GiB x 2
Disk Capactiy	1.5T



(a) $\lambda = 0.2$

(b) $\lambda = 0.4$

(c) $\lambda = 0.5$

(d) $\lambda = 0.6$

(e) $\lambda = 0.7$

Fig. 5 The comparisons of loss with various CTC weights

On the other hand, the CERs of our five experiments based on the Alpha-Dev subset are listed in Table 5. If λ is less than 0.6, the CER obtained with the increase of λ becomes smaller, from 8.85% to 8.32%. However, if λ is equal to 0.6, CER starts increasing. Unlike the slight decrease in attention model, the performance of CTC regarding CER dropped from 14.28% to 11.04% as λ on the raise. The minimum is reached if λ is equal to 0.6. Although λ is equal to 0.5, attention model has achieved the best results. Nevertheless, compared with the slight difference in attention, if λ is equal to 0.6, the CTC improvement is much remarkable. Combined the experimental results in Fig. 5 and Table 5, we see that λ equals 0.6 in the mixed model is the optimal solution.

Table 5 The comparisons of CERs based on “Dev” subset with various CTC weights

λ	CER(Attention)	CER(CTC)
0.2	8.85%	14.28%
0.4	8.52%	12.84%
0.5	8.24%	12.02%
0.6	8.32%	11.04%
0.7	8.51%	11.17%

3.4 Experimental results

The evaluations were carried out based on three datasets: Alpha, Beta, and Gamma, respectively. The final experimental results and the previous three experimental results are compared in Table 6.

4 Result analysis

The first row in Table 6 shows that the traditional HMM-based speech recognition model achieved 10.91% error rate in Chinese and 18.23% in English. These two items are much higher than only implementing CTC or attention model in speech recognition. However, the CTC+Attention hybrid model obtained better results based on the Chinese dataset, with the error rate 0.069% less than that of the traditional model. Based on the Code-Switch dataset, the result of the traditional speech recognition model regarding word error is the best of the three models, but the difference with CTC+Attention is not obvious, and the outcome is improved by 3%. The evaluation results of all models in the Gamma dataset are the worst one compared with other two datasets. That is due to the complexity of mixed languages. Moreover, through the experimental results, we see that the accuracy of CTC+Attention hybrid model is the highest one, an error rate 10.22% was achieved based on the Chinese dataset.

From the language perspective, the experimental results show that the English dataset is more challenging to be used for model training than the Chinese one. Compared with the three languages, Chinese labels with Pinyin performs better than English in speech recognition. The convergence in the training phase and the WERs in the verification phase consistently reflect that the Chinese dataset Alpha labeled with Pinyin is easier to be trained, the recognition accuracy is greater than English. The utilization of Pinyin effectively avoids the situation of one sound and multiple characters in Chinese.

On the other hand, the WER of the HMM-TDNN-F model in the dataset Gamma is 25.62%. In contrast, the CTC+Attention model has a 37% gap between the two datasets. The results show that the difference of recognition accuracy of the CTC+Attention model by

Table 6 The comparisons of experimental results(CERs/WERs)

Models	Alpha	Beta	Gamma
HMM-TDNN-F	10.91%	18.23%	25.62%
CTC+Attention	10.22%	19.05%	26.11%

using different language datasets is the smallest one. The difference of multilingual speech recognition is better than the traditional framework.

In this paper, our experimental results are analyzed from two aspects. In contrast, CTC+ Attention model performs better in the Chinese dataset, while the traditional model performs better in English and mixed language. In terms of language, Pinyin labeled datasets are easier to be applied to model training. CTC + Attention has less difference by taking different datasets into account in terms of the accuracy of different language recognition.

5 Conclusion

Throughout this paper, we provide a research foundation for future exploration of speech recognition with mixed languages. The comparison of actual outcomes in multiple languages provides a future research direction based on the characteristics of speech recognition.

In this paper, we investigate speech recognition performance based on the CTC and attention hybrid model by using our three labeled datasets. The model attains 10.91%, 18.23%, and 25.62% error rates based on the Chinese, English, and Chinese-English Code-Switch datasets. The CTC+Attention model adopts the optimal solution to complete the model evaluations and achieves similar performance to the traditional model. The evaluation based on the Chinese dataset defeated the traditional model and reached 10.22% CER. Albeit the performance based on the English and Code-Switch datasets was not as good as the traditional model, the gap remained within 3%.

In future, more attention model will be added to our project to replace the current model [42]. We will implement the predetermined algorithms to improve the training results. In addition, we will explore multiple languages for speech recognition by using the methods of creating datasets in this paper [15].

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

References

1. Abdel-Hamid O, Mohamed AR, Jiang H, Penn G (2012) Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition. In: IEEE International conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 4277–4280
2. Boden M (2002) A guide to recurrent neural networks and backpropagation. The Dallas project: SICS technical report
3. Chan JY, Ching P, Lee T, Meng HM (2004) Detection of language boundary in code-switching utterances by bi-phone probabilities. In: International symposium on chinese spoken language processing. IEEE, pp. 293–296
4. Chan W, Jaitly N, Le QV, Vinyals O (2015) Listen, attend and spell. arXiv:1508.01211
5. Chiu CC, Raffel C (2017) Monotonic chunkwise attention. arXiv:1712.05382

6. Eyben F, Wöllmer M, Schuller B, Graves A (2009) From speech to letters-using a novel neural network architecture for grapheme based ASR. In: IEEE Workshop on automatic speech recognition & understanding. IEEE, pp 376–380
7. Fu L, Li X, Zi L (2020)
8. Georgescu AL, Cucu H, Burileanu C (2019) Kaldi-based DNN architectures for speech recognition in Romanian. In: International conference on speech technology and human-computer dialogue (sped). IEEE, pp 1–6
9. Hinton GE, Osindero S, Teh YW (2006) A fast learning algorithm for deep belief nets. Neural Comput 18(7):1527–1554
10. iFLYTEK Co., Ltd: Online TTS WebAPI. Website (2020). https://global.xfyun.cn/products/online_tts
11. Jason CA, Kumar S (2020) An appraisal on speech and emotion recognition technologies based on machine learning. Language 67:68
12. Kim S, Hori T, Watanabe S (2017) Joint CTC-attention based end-to-end speech recognition using multi-task learning. In: IEEE International conference on acoustics, speech and signal processing. IEEE, pp 4835–4839
13. Li B, Chang Sy, Sainath TN, Pang R, He Y, Strohman T, Wu Y (2020) Towards fast and accurate streaming end-to-end ASR. In: IEEE International conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 6069–6073
14. Li J, Zhao R, Hu H, Gong Y (2019) Improving RNN transducer modeling for end-to-end speech recognition. In: IEEE Automatic speech recognition and understanding workshop. IEEE, pp 114–121
15. Liang S (2021) Multilingual speech recognition based on the end-to-end framework (Master's Thesis). Auckland University of Technology, New Zealand
16. Lin CH, Lee LS, Ting PY (1993) A new framework for recognition of Mandarin syllables with tones using sub-syllabic units. In: IEEE International conference on acoustics, speech, and signal processing, vol. 2. IEEE, pp 227–230
17. Liu Z, Chen Q, Hu H, Tang H, Zou Y (2019) Teacher-student learning and post-processing for robust biLSTM mask-based acoustic beamforming. In: International conference on neural information processing. Springer, pp. 522–533
18. Maas AL, Qi P, Xie Z, Hannun AY, Lengerich CT, Jurafsky D, Ng AY (2017) Building DNN acoustic models for large vocabulary speech recognition. Comput Speech Lang 41:195–213
19. Manaswi NK, Manaswi NK, John S (2018) Deep learning with applications using python. Springer
20. Mansikkaniemi A (2010) Acoustic model and language model adaptation for a mobile dictation service (master's thesis). Aalto university
21. Miao Y, Gowayyed M, Metze F (2015) EESEN: End-to-end Speech recognition using deep RNN models and WFST-based decoding. In: IEEE Workshop on automatic speech recognition and understanding (ASRU). IEEE, pp 167–174
22. Mikolov T, Karafiát M, Burget L, Černockỳ J, Khudanpur S (2010) Recurrent neural network based language model. In: Annual conference of the International speech communication association
23. Mohamed Ar, Dahl G, Hinton G (2009) Deep belief networks for phone recognition. In: NIPS Workshop on deep learning for speech recognition and related applications, vol. 1. Vancouver, Canada, p 39
24. Moritz N, Hori T, Le Roux J (2019) Triggered attention for end-to-end speech recognition. In: IEEE International conference on acoustics, speech and signal processing. IEEE, pp 5666–5670
25. Panayotov V, Chen G, Povey D, Khudanpur S (2015) Librispeech: an ASR corpus based on public domain audio books. In: IEEE International conference on acoustics, speech and signal processing. IEEE, pp 5206–5210
26. Passricha V, Aggarwal RK (2020) A hybrid of deep cnn and bidirectional LSTM for automatic speech recognition. J Intell Syst 29(1):1261–1274
27. Petridis S, Stafylakis T, Ma P, Tzimiropoulos G, Pantic M (2018) Audio-visual speech recognition with a hybrid CTC/attention architecture. In: IEEE Spoken language technology workshop. IEEE, pp 513–520
28. Povey D, Ghoshal A, Boulianne G, Burget L, Glembek O, Goel N, Hannemann M, Motlicek P, Qian Y, Schwarz P et al (2011) The Kaldi speech recognition toolkit. In: IEEE Workshop on automatic speech recognition and understanding. IEEE signal processing society
29. Qu Z, Haghani P, Weinstein E, Moreno P (2017) Syllable-based acoustic modeling with CTC-SMBR-LSTM. In: IEEE Automatic speech recognition and understanding workshop. IEEE, pp 173–177
30. Senior A, Sak H, Shafran I (2015) Context dependent phone models for LSTM RNN acoustic modelling. In: IEEE International conference on acoustics, speech and signal processing. IEEE, pp 4585–4589
31. Shi F, Cheng X, Chen X (2012) The summarize of improved HMM Model. In: International conference on computer and information application, pp. 627–630
32. Smit P, Virpioja S, Kurimo M (2021) Advances in subword-based HMM-DNN speech recognition across languages. Comput Speech Lang 66:101158

33. Sundermeyer M, Oparin I, Gauvain JL, Freiberg B, Schlüter R, Ney H (2013) Comparison of feedforward and recurrent neural network language models. In: IEEE International conference on acoustics, speech and signal processing. IEEE, pp 8430–8434
34. TAL Education Group: TAL CS Auto Speech Recognition Data set. Website (2019). https://ai.100tal.com/dataset
35. Ueno S, Inaguma H, Mimura M, Kawahara T (2018) Acoustic-to-word attention-based model complemented with character-level CTC-based model. In: IEEE International conference on acoustics, speech and signal processing. IEEE, pp 5804–5808
36. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. arXiv:1706.03762
37. Wang D, Wang X, Lv S (2019) An overview of end-to-end automatic speech recognition. Symmetry 11(8):1018
38. Wang W, Wang G, Bhatnagar A, Zhou Y, Xiong C, Socher R (2020) An investigation of phone-based subword units for end-to-end speech recognition. arXiv:2004.04290
39. Watanabe S, Hori T, Karita S, Hayashi T, Nishitoba J, Unno Y, Soplin NEY, Heymann J, Wiesner M, Chen N et al (2018) ESPnet: End-to-end speech processing toolkit. arXiv:1804.00015
40. Woodland PC, Odell JJ, Valtchev V, Young SJ (1994) Large vocabulary continuous speech recognition using HTK. In: IEEE International conference on acoustics, speech and signal processing, vol. 2. IEEE, pp II–125
41. Wu CH, Shen HP, Yang YT (2014) Chinese-english phone set construction for code-switching ASR using acoustic and DNN-extracted articulatory features. IEEE/ACM Transactions on Audio, Speech, and Language Processing 22(4):858–862
42. Yan WQ (2021) Computational methods for deep learning. Springer
43. Zenkel T, Sanabria R, Metze F, Waibel A (2017) Subword and crossword units for CTC acoustic models. arXiv:1712.06855
44. Zheng Y, Yang X, Dang X (2020) Homophone-based label smoothing in end-to-end automatic speech recognition. arXiv:2004.03437

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.