A multi-feature fusion based confidence estimation model for Chinese end-to-end speech recognition

Lijuan Shi*, Zixiao Wang, Yang Du, Yang Wang, Ying Chen, Jizhuang Zhao, Kun Wang
AI Research Center, China Telecom Research Institute, Beijing, China

ABSTRACT

Speech recognition has shown excellent performance in recent years, but as we known, this kind of AI task is still supervised and needs labelled data to feed models so that it can achieve a better effect and to inference to get model's metrics. In actual application scenarios, it always needs many people to check speech recognition's correctness because of lacking labelled data, it is not realistic to label massive online audio. In addition, this method can't guarantee comprehensiveness of checking, so confidence estimation algorithm is proposed, which can evaluate speech recognition model's results and predict error transcriptions automatically. This paper proposed a confidence estimation model based on a multi-feature fusion mechanism and it mainly focus on Chinese end-to-end speech recognition tasks in complex application scenarios. Experiments in Aishell-1 dataset and China telecom's internal dataset have proved that this model can obtain a good performance.

Keywords: Chinese end-to-end automatic speech recognition, confidence estimation, multi-feature fusion, pretrained language model

1. INTRODUCTION

Recently, end-to-end deep neural networks have achieved excellent results in speech recognition tasks, and have enabled various applications in the industry. Based on this development, the voice technology team of China Telecom Research Institute uses speech recognition technology to provide smart service in the customer service scenario. In the process of research and development, it was found that, although the accuracy of current SOTA speech recognition models have reached more than 95% accuracy[1], in a complex circumstance like customer scenario who have more user groups and diverse environment, the speech recognition model may face various problems such as multiple dialects, serious colloquialism, and special domain vocabulary, etc., so that an inevitable difficulty arises: the recognition results of massive and diverse online unlabelled audio is not evaluable and the actual accuracy can be far less than experimental accuracy. Poor recognition result may bring irreparable losses to downstream tasks and worse user experience, so the detection of incorrectly recognized texts has an important meaning. To solve the problem above, the confidence estimation algorithm was proposed, who is an algorithm that evaluates the effect of speech recognition model in each token of transcripts and predicts the location of error tokens which are the tokens speech recognition models are less confident in.

Confidence estimation models can be divided into three categories, firstly, some studies have shown that the posterior probability output of ASR (Automatic Speech Recognition) model can be used directly as the confidence evaluation score[2], but the effect of this kind of method mainly depends on the ASR model itself, which is not suitable for complex application scenarios. Secondly, based on recognized transcriptions, the ASR model's confidence estimation problem can be treated as a natural language processing problem, so that we can use language model to learn semantic scores of transcripts as confidence scores. As shown in Reference[3], researchers proposed a method that uses the language pretrained model ELECTRA as a discriminator to calculate the confidence score of ASR model's hypothesis. Thirdly, this issue can be thought as a binary classification problem, by collecting various kind of features, feature engineering and modeling can be used to obtain confidence scores[4,5]. These works mainly presented to store information during the process of ASR model's recognition as features, and then use these features to obtain confidence scores. the above two papers mainly focus on ASR model's acoustic information which is insufficient for industry's complex scenarios.

Since the third kind of methods have better effects than others, and considering about application scenarios, this paper follows the third technical route and mainly focus on the confidence estimation problem of current mainstream CTC (Connectionist Temporal Classification) based end-to-end ASR models, who are a series of models that use the CTC

* shilj2@chinatelecom.cn

Third International Conference on Computer Science and Communication Technology (ICCSCT 2022)
edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12506, 1250626
 · 0277-786X · doi: 10.1117/12.2662485

algorithm to convert audios to texts directly. It presents innovatively a confidence estimation model based on fusing multi-features, which extracts speech audio quality features, linguistic and semantic features, and acoustic features of ASR model. This confidence estimation model is mainly used to adapt to complex online speech recognition scenarios.

The paper is organized as follows. Section 2 discuss the confidence estimation model's architecture and in Section 3, it introduces different kind of features' extraction methods, Section 4 presents the encoder and decoder of confidence estimation model. The followed Section 5 describes the experiment and final results. Finally, in Section 6, we made a conclusion and discussed our further work in the future.

2. MODEL ARCHITECTURE

This section describes a confidence estimation model based on a multi-feature fusion mechanism, and according to these features, a neural network binary classification model is built to get the confidence score of ASR model and to predict the error tokens in each recognized hypothesis. Our architecture is composed of three parts: automatic speech recognition, multi-feature extractor, confidence score's encoder and decoder.

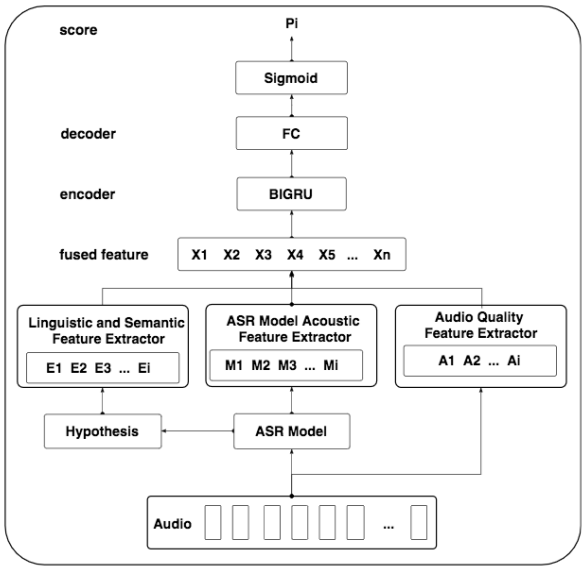

Figure 1. Model architecture.

As shown in the figure 1, for each piece of audio, it will be transcribed to its corresponding recognized text utterance by ASR model, the ASR model is one of current mainstream CTC-based deep neural network like CTC-Conformer[1]. Since this paper mainly discuss the confidence estimation algorithm of speech recognition model, it just needs to use the trained frozen speech recognition neural network. Based on the transcribed texts, the linguistic features can be extracted, and the audio quality features can be obtained simultaneously from audios. In the ASR model acoustic feature extractor, it merges ASR model's acoustic features stored during speech recognition process, which can present the recognize effect in the view of model itself. Additionally, in the linguistic and semantic feature extractor, this paper innovatively proposes the use of BERT (Bidirectional Encoder Representations from Transformers)[6] model to extract general semantic features of hypothesis and n-gram language model[7] to calculate domain linguistic features. After fusing these features, a feature matrix X is obtained who can better present the information in a complex scenario. In order to get a better feature representation, an encoder is built whose input is the matrix X and finally a decoder to get the final confidence score who is also the possibility of error recognition for each token in every hypothesis. Here, we use the Bi-GRU network[8] as the encoder, and a simple feedforward network as the decoder.

3. MULTI FEATURE EXTRACTOR AND FEATURE FUSION

In this part, we describe the detail of feature extractors. We propose three kinds of extractors to obtain separately audio quality features, linguistic and semantic features of hypothesis, and ASR model's acoustic features.

3.1 Audio quality features

Although the front-end signal processing process can guarantee the quality of speech recognition, the ASR model faces a great uncertainty in a variety of complex application scenarios. The effect of speech recognition may be affected by the audio quality related information. In this part, we propose to use the following audio quality related features, which are shown in Table 1.

Table 1. Audio quality features.

Feature type	Feature description
Transcripts' length	Transcribed text length of speech segment recognized by ASR model. Recognition's effect may vary with different length of transcripts because of training data's distribution and longer sentences have higher error probability of having error tokens in common sense.
Audio duration	The audio segment duration to be recognized. Recognition's effect may vary because of different user's pause preference and longer audio may have higher probability to having error tokens and user's voice quality may decrease with long duration audio.
Speaking speed	Speech speed of the input audio. The quicker user's speaking speed, the more speech's intelligibility drops.

3.2 Linguistic and semantic features

With the rapid development of NLP (Natural Language Processing) technology, sentence's fluency, grammatical errors can be evaluated accurately with a series of language models. The confidence and correctness of ASR transcriptions can be evaluated through linguistic and semantic features of ASR transcribed text. But there is still a difference between confidence estimation problem and the traditional text error detection task of natural language understanding. NLP tasks mainly focus on finding the spelling and grammatical errors, but confidence estimation tasks need to find less confident recognized tokens of ASR model, which means there is possibility that a less confident token can be totally correct in terms of spelling or grammar, but not in terms of domain situation. So not only do we need to assess the grammatical correctness of the transcribed text, but also how well it fits the context. This paper innovatively proposes here a method of combining pre-trained language model's information with domain semantic model knowledge, just like human beings may be confused by homophones or homographs in the process of dictation, but with the accumulation of knowledge and the combination of current contextual knowledge, the expression or determination of dictation vocabulary will gradually become more and more accurate. As shown in Figure 2, here we use a large pretrained model embedding to present human accumulated knowledge and n-gram's information[7] as context knowledge. Advantages for not finetuning the pretrained model here are that not only can we avoid the knowledge bias problem, but also the large computing cost of model training. Statistical n-gram model is adapted to focus on domain knowledge and having a quicker computing speed than neural models.

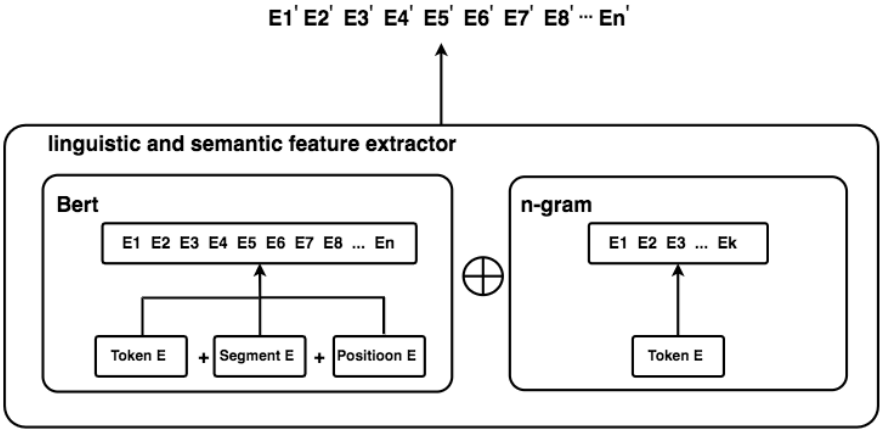

Figure 2. Linguistic and semantic features.

In this paper, we use the large pretrained Chinese language model BERT[6], and each input text sentence will be converted into embedding matrix E. with a n-gram extractor and based on a windowing mechanism, for each token we design here three kinds of features: current state probability, previous state probability and next state probability. For Chinese, token is usually in char-level, so we can define the states of each token as shown in Figure 3.

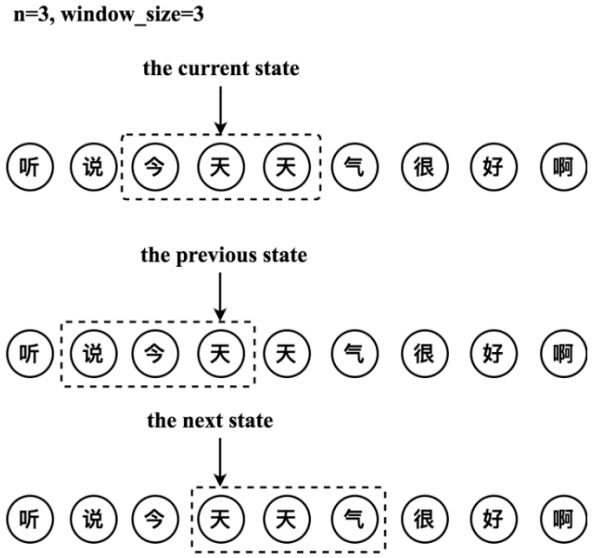

Figure 3. An example of n-gram feature extractor.

Finally, the feature of each token in the hypothesis can be composed of BERT's token embedding, segment embedding, position embedding, and three-dimensional features of n-gram, in general, we choose n=3 or n=5.

3.3 ASR model acoustic features

ASR model's recognition effect can be reflected from various model's internal factors. We selected the following information in Table 2 to be stored in the process of recognition and to represent ASR model's acoustic features:

Table 2. ASR model acoustic features.

Feature type	Feature description
Maximum posterior probability	The maximum posterior probability presents the most likely token in the vocabulary for each piece of audio, which can be store in the process of ASR model's decoding stage, it represents the confidence extent of ASR model itself intuitively.
Second largest posterior probability	The second largest probability in the decoding candidates who can reflect ASR model's confusion level.
Frame duration	The number of frames that each token has, which is designed to reflect the impact of CTC decoding algorithm.
Frame spacing	The number of frames between two decoding tokens, who is designed to measure the impact of <BLANK> in CTC decoding algorithm.

3.4 Feature fusion

After the above feature extraction steps, we can have the audio quality feature matrix A, the linguistic and semantic feature matrix E' and the ASR model acoustic feature matrix M. Based on these feature matrices, we can calculate the final fusion feature matrix by equation (1).

$$X_i = Concat\left(A_i, E_i, M_i\right) \tag{1}$$

For the *i*-th token in each recognized hypothesis, the fused feature matrix X_i is the input of encoder.

4. ENCODER AND DECODER

In this work, in order to further fuse and encode feature X, we use a bidirectional GRU network[8] as the encoder and a feedforward network as the decoder, finally, to obtain the final confidence score or token error probability, a sigmoid activation function is used. For each token of the transcription utterance, the probability can be defined as:

$$P_i = P\left(y_i = 1 | X\right) = \sigma\left(Wh_i + b\right) \tag{2}$$

where P_i denotes the conditional probability given by the network which present the error probability of transcribed token and also our ASR model's confidence score here. And σ denotes the sigmoid function, h_i denotes the hidden state of Bi-GRU, W and b are parameters. Furthermore, the hidden state is defined as:

$$\vec{h}_i = GRU(\vec{h}_{i-1}, e_i) \tag{3}$$

$$\overleftarrow{h}_i = GRU(\overleftarrow{h}_{i+1}, e_i) \tag{4}$$

$$h_i = [\vec{h}_i; \overleftarrow{h}_i] \tag{5}$$

where e_i denotes the embedding of input sequence X_i, $[\vec{h}_i; \overleftarrow{h}_i]$ denotes the concatenation of GRU hidden states from the two directions and GRU is the GRU function.

5. EXPERIMENT AND RESULTS

5.1 Datasets

The Aishell-1 dataset[9] is an open-source mandarin speech recognition corpus, which is a 500 hours multi-channel mandarin speech corpus designed for various speech/speaker processing tasks. The corpus includes training set, development set and test sets. Training set contains 120,098 utterances from 340 speakers, development set contains 14,326 utterance from the 40 speakers, and test set contains 7,176 utterances from 20 speakers. Since our purpose is to experiment with confidence estimation for speech recognition models, we use the speech recognition model trained on the Aishell-1 training set, and the Aishell-1 test set to create the train and test sets of our confidence model. For every utterance in the test set, we use the trained frozen ASR model to recognize it and obtain the corresponding transcript sentence, and, after comparing with the labelled text, we can get the train and test set of confidence estimation model (CEM). The dataset is split in an 8/1/1 ratio, so finally we can have about 5740 utterances of training set and 718 utterances for dev set and test set.

Table 3. An example of CEM's data creation method.

Transcription	听	说	今	天	-	气	很	好	好	啊
Reference	听	说	今	天	天	气	很	好	-	啊
Error type	C	C	C	C	D	C	C	C	I	C
Label of CEM	0	0	0	0	1	0	0	0	1	0

We present an example of the CEM's data creation method in Table 3, among them, there are three types of token alignment error: substitution, insertion and deletion. But here we do not need to consider them, for all the error tokens, we make the label of CEM to 1, and 0 to the correct tokens. In addition to the Aishell-1 dataset, we have also conducted experiments on our internal smart customer service scenario dataset. Because of the sensitivity of the data and the confidentiality for users, we didn't present the detail of our data in this paper. We used 18714 utterances to be test set of ASR model and also to be the experiment dataset of confidence estimation task.

5.2 ASR models

We choose the frozen Conformer model[1] trained on the AISHELL-1 dataset and based on the WENET toolkit[10]. The Conformer model is a convolution neural network and Transformer based ASR model who have achieved a new SOTA result recently. The model features that we use are showed in Table 4.

Table 4. Settings of frozen conformer ASR model.

	Conformer Aishell-1	Conformer Internal
Feature	Fbank-80	Fbank-80
Vocabulary size	4233 (characters)	3160(characters)
Attention heads	4	4
Linear units	2048	2048
Optimizer	Adam	Adam
Learning rate	0.002	0.002
CER Aishell-1	4.97	10.82

5.3 Model parameters

We use the normal Chinese pretrained BERT model who is trained on the Wikipedia corpus, having attention-head=12, hidden-size=768, and hidden-layer=12. In the encoder layer block, we choose the bidirectional GRU with hidden size 768 likewise. In the training process, we choose an Adam optimizer, a binary cross entropy loss, and initial learning rate 1e-3.

5.4 Evaluation metrics

As the purpose of confidence estimation problem is to evaluate ASR model's effect and to predict error tokens, it can be thought as a detection task in one aspect and evaluate the experiment's effect through precision, recall and F1-score. The formulas of these three metrics are:

$$\text{Precision}=\text{TP} / (\text{TP}+\text{FP}) \tag{6}$$

$$\text{Recall} = \text{TP} / (\text{TP}+\text{FN}) \tag{7}$$

$$\text{F1} = (2*\text{Precision}*\text{Recall}) \div (\text{Precision}+\text{Recall}) \tag{8}$$

Among them, TP represents the number of correctly detected misrecognized transcripts, FP represents the number of incorrectly detected misrecognized transcripts, and FN represents the number of undetected misrecognized transcripts.

5.5 Results and analysis

After 50 epoch's training, our confidence estimation model got results as showed in Table 5. We performed experiments in two datasets. One is in the Aishell-1 dataset, and it is used here to prove the feasibility of our experiment. The other is our internal dataset which is selected from different cities of online customer service scenario. Compared with the Aishell-1 dataset, it is actually more representative to analyse the effect of our model as it has the characteristics of including multiple dialects, multiple user groups, more environmental influence factors, so it is normal that the final results perform better in the Aishell-1 dataset in general. Besides, the output of CEM for each hypothesis is an error probability matrix, so in order to determine whether each token is misrecognized or not, we have to choose a confidence probability threshold, and here we choose 0.7, 0.8 and 0.9, different threshold brings different final result. When we choose a stricter threshold, we notice that the recalls of final results are generally smaller than precisions in our internal dataset but there is little difference in the Aishell-1 dataset, it is because that our data's diversity and complexity make it more difficult to detect error tokens than detect correctly. And if we choose a smaller threshold, the recall and the precision can have a trade-off.

Table 5. Experiment results.

	Threshold	Precision	Recall	F1
Conformer Aishell-1	0.7	0.858	0.838	0.848
	0.8	0.852	0.817	0.834
	0.9	0.846	0.795	0.820
Conformer internal	0.7	0.797	0.744	0.769
	0.8	0.814	0.682	0.742
	0.9	0.817	0.692	0.749

6. CONCLUSION

This paper focuses on the Chinese end-to-end speech recognition model's hypothesis evaluation problem, and proposes a confidence estimation model based on a multi-feature fusion mechanism which combines various information in complex application scenario. Audio quality information, linguistic and semantic information, ASR model's semantic information are used in the CEM model to evaluate ASR model's effect and to predict error transcriptions, among of which a pretrained model is also used to further help select linguistic information. The experimental results have shown that this CEM model can bring a good result but there is always a trade-off between recall and precision. In the further work we may try different kind of pretrained models or find more influential features.

REFERENCES

[1] Gulati, A., Qin, J., Chiu, C. C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y. and Pang, R., "Conformer: Convolution-augmented transformer for speech recognition," Interspeech, (2020). https://doi.org/10.21437/interspeech.2020-3015

[2] Oneata, D., Caranica, A., Stan, A. and Cucu, H., "An evaluation of word-level confidence estimation for end-to-end automatic speech recognition," 2021 IEEE Spoken Language Technology Workshop (SLT), (2021). https://doi.org/10.1109/slt48900.2021.9383570

[3] Futami, H., Inaguma, H., Mimura, M., Sakai, S. and Kawahara, T., "ASR rescoring and confidence estimation with electra," 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), ((2021). https://doi.org/10.1109/asru51503.2021.9688175

[4] Ogawa, A., Tawara, N., Kano, T. and Delcroix, M., "BLSTM-based confidence estimation for end-to-end speech recognition," 2021 IEEE Inter. Conf. on Acoustics, Speech and Signal Processing (ICASSP), (2021). https://doi.org/10.1109/icassp39728.2021.9414977

[5] Ragni, A., Li, Q., Gales, M. J. and Wang, Y., "Confidence estimation and deletion prediction using bidirectional recurrent neural networks," 2018 IEEE Spoken Language Technology Workshop (SLT), (2018). https://doi.org/10.1109/slt.2018.8639678

[6] Kamath, U., Graham, K. L. and Emara, W., "Bidirectional encoder representations from Transformers (Bert)," Transformers for Machine Learning, 43-70 (2022). https://doi.org/10.1201/9781003170082-3

[7] Brown, P. F., Della Pietra, V. J., de Souza, P. V., Lai, J. C. and Mercer, R. L., "Class-based n-gram models of natural language," Computational Linguistics 18(4), 467-480 (1992).

[8] Bahdanau, D., Cho, K. H. and Bengio, Y., "Neural machine translation by jointly learning to align and translate," 3rd Inter. Conf. on Learning Representations, (2015).

[9] Bu, H., Du, J., Na, X., Wu, B. and Zheng, H., "AISHELL-1: An open-source Mandarin speech corpus and a speech recognition baseline," 2017 20th Conf. of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA), (2017). https://doi.org/10.1109/icsda.2017.8384449

[10] Yao, Z., Wu, D., Wang, X., Zhang, B., Yu, F., Yang, C., Peng, Z., Chen, X., Xie, L. and Lei, X., "WeNet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit," Interspeech, (2021). https://doi.org/10.21437/interspeech.2021-1983