symmetry

MDPI

Article

Cross-Language End-to-End Speech Recognition Research Based on Transfer Learning for the Low-Resource Tujia Language

Chongchong Yu [1,*], Yunbing Chen [1], Yueqiao Li [1], Meng Kang [1], Shixuan Xu [2] and Xueer Liu [1]

[1] College of Computer & Information Engineering, Beijing Technology and Business University, Beijing 100048, China; 10011316215@st.btbu.edu.cn (Y.C.); 1604010512@st.btbu.edu.cn (Y.L.); 1830401006@st.btbu.edu.cn (M.K.); 1604010513@st.btbu.edu.cn (X.L.)

[2] Institute of Ethnology & Anthropology, Chinese Academy of Social Sciences, Beijing 100081, China; xusx@cass.org.cn

* Correspondence: yucc@btbu.edu.cn; Tel.: +86-139-1111-9035

Received: 17 December 2018; Accepted: 29 January 2019; Published: 2 February 2019

Abstract: To rescue and preserve an endangered language, this paper studied an end-to-end speech recognition model based on sample transfer learning for the low-resource Tujia language. From the perspective of the Tujia language international phonetic alphabet (IPA) label layer, using Chinese corpus as an extension of the Tujia language can effectively solve the problem of an insufficient corpus in the Tujia language, constructing a cross-language corpus and an IPA dictionary that is unified between the Chinese and Tujia languages. The convolutional neural network (CNN) and bi-directional long short-term memory (BiLSTM) network were used to extract the cross-language acoustic features and train shared hidden layer weights for the Tujia language and Chinese phonetic corpus. In addition, the automatic speech recognition function of the Tujia language was realized using the end-to-end method that consists of symmetric encoding and decoding. Furthermore, transfer learning was used to establish the model of the cross-language end-to-end Tujia language recognition system. The experimental results showed that the recognition error rate of the proposed model is 46.19%, which is 2.11% lower than the that of the model that only used the Tujia language data for training. Therefore, this approach is feasible and effective.

Keywords: low-resource speech recognition; Tujia language; cross-language end-to-end; transfer learning

1. Introduction

Endangered languages are non-renewable intangible cultural resources. The core task of salvaging and preserving endangered languages is the mechanism of recording speech, processing corpus, and preserving language information. In China, the dialects of many ethnic minorities have no texts and exist only in the form of spoken language. The number of native speakers is so small that the recordings of long natural speech are extremely limited. There are more than 130 dialects of Chinese ethnic minorities, nearly half of which are in a state of recession and dozens of which are endangered. This trend is continuing and even worsening [1]. Therefore, it is an imperative duty to protect endangered languages, maintain the diversity of language culture, and retain valuable historical cultural heritage. Automatic speech recognition of endangered languages is a new and effective way to rescue and preserve endangered languages. Because endangered languages have a low-resource attribute, speech recognition for endangered languages requires low-resource speech recognition. The Tujia language, one of the endangered languages, poses a great challenge to automatic speech recognition.

Speech signal is a non-stationary timing signal whose formation and sensing are a complex signal process. Speech recognition can be regarded as a sequence-to-sequence classification problem, in which the acoustic observation sequence $X = (x_1, x_2, \cdots, x_T)$ is mapped to the character sequence $W = (w_1, w_2, \cdots, w_N)$ and calculates the probability $P(W|X)$, where T is the time and N is the number of characters [2]. In the timing classification task, the commonly used method is to train by frame using the hidden Markov model (HMM) under the condition that the input data and the given label must be frame-level aligned in time. However, the frame-by-frame training output is the single frame probability. For timing problems, the probability of an output sequence is much more important than the probability of outputting a single frame. At present, most of the automatic speech recognition technologies at home and abroad rely on many data resources. The training of the entire model is divided into multiple stages and there are multiple optimization goals. Thus, it is difficult to find the global optimal solution. In response to this problem, recent research internationally in the field of speech recognition based on deep learning has partly focused on end-to-end speech recognition technology [3–7]. This method directly models between the phoneme sequence or context-dependent phone (CD-phone) sequence and the corresponding phonetic feature sequence that does not need constraint alignment to obtain frame-level annotation with HMM. Compared with the traditional acoustic models, it can provide a better performance.

Currently, few researchers are applying the end-to-end model to low-resource speech recognition. Previous research methods for low-resource speech recognition primarily included cross-language, transfer learning, and semi-supervising. The structure of a deep neural network (DNN) for cross-language speech recognition was generally that the input layer and the hidden layer were shared by all languages, whereas the output layer was not shared. Moreover, each language had its own softmax layer to estimate the posterior probability after clustering [8,9]. Transfer learning is an emerging research field in machine learning that aims to use the knowledge learned from other tasks to help solve the target tasks. Transfer learning reduced the dependence on target training data by finding the common knowledge between existing tasks and target tasks, which can help the model to learn target tasks better [10,11]. According to [12], the accuracy rate of the cross-language speech recognition model based on sequence was 6% higher than that of the speech recognition model using only a single language, but the model still needed to be improved. The attention-based connectionist temporal classification (CTC) model in [13], effectively completing the keyword search and speech recognition of low resource languages, had an insufficient effect of speech recognition. It achieved topic and keyword identification that the cross-language transfer learning method in [14] used to learn the characteristics of low-resourced languages from rich-resourced languages achieves, but it only used unsupervised learning to generate valid data due to a lack of native transcriptions. The speech emotion classification model of deep belief networks (DBNs) was trained in five types of speech databases in [15], and the transfer learning method was used to improve the speech emotion classification accuracy. The experimental results showed that the accuracy rate improved significantly.

The preservation of an endangered language corpus requires text processing, such as labelling and translation, for the recording of natural language discourses. At present, it has become a bottleneck in the protection of the Tujia language. First, much manpower and time is needed for a language protection project. Generally, at least an hour of text processing is needed for a minute of recording. Speech data without text processing has no intelligibility and preservation value. Second, few people use the Tujia language, and professionals who can process text corpus are scarce. Therefore, the achievement of the speech recognition system in this paper can help linguists to complete the work of transcribing the Tujia language, which can greatly reduce labor and time cost and has important theoretical significance and applied value.

There are two main contributions from this paper. First, speech features are extracted using a convolutional neural network. In addition, cross-language hidden layer weights are shared using a bi-directional long short-term memory network. Using Chinese speech datasets as extended datasets can solve the insufficiency of the Tujia language. The second point is to use the method of transfer

learning. The initial model is built by training the Tujia language and Chinese corpus by means of modifying the softmax layer of the initial model and proceeding with sample transfer learning. Then, continuing to train with the Tujia corpus obtains the final model.

The remainder of this paper is organized as follows. Section 1 presents reviews of related deep learning algorithms. Section 2 introduces the proposed method and model. The experimental results and model parameters are presented in Section 3, and conclusions are drawn in Section 4.

2. Review of Related Work

Deep learning uses multi-layered nonlinear structures to transform low-level features into more abstract high-level features and transforms input features in a supervised or unsupervised method, thereby improving the accuracy of classification or prediction [16,17]. Compared with the traditional shallow model, the deep learning model is more efficient in expression and modelling. Moreover, it has more advantages in the processing of complex signals. Deep learning was first applied to speech recognition in 2009 [18]. It provides a 20% improvement over the traditional Gaussian mixture model-hidden Markov model (GMM-HMM) speech recognition system. Since then, the acoustic models based on the deep neural network have gradually replaced the GMM as the mainstream acoustic model in speech recognition, which has greatly promoted the development of speech recognition technology, breaking through the bottleneck of speech recognition requirements in some practical application scenarios. In the past two years, the end-to-end model based on deep learning, such as using a CNN or CLDNN to implement an end-to-end model in the CTC framework or the recently proposed low frame rate and chain model, which are based on coarse-grained modelling unit technology [19,20], has enabled progress to be made in recognition performance and has become a research direction.

2.1. Feature Extraction Based on CNN

The characteristics of a speech signal are primarily in the time and frequency domains. Time domain characteristics include the short-term average energy, short-term average zero-crossing rate, formant, and pitch period. The frequency domain features include the linear prediction coefficient (LPC), LP cepstral coefficient (LPCC), line spectrum pair parameter (LSP), short-term spectrum, and Mel frequency cepstral coefficient (MFCC). All these features only include some features of speech signals. To fully characterize speech signals, the transformation and selection of features and the use of feature timing information have become important research topics [21–23]. In References [24–26] proposed a bottleneck (BN) deep neural network with a narrow intermediate layer, extracting the bottleneck features in the middle of the network to replace traditional MFCC features. These bottleneck features not only have a long-term correlation of speech but can also extract the signal. Compared with traditional speech features, the CNN can obtain more robust features using local filtering and maximum pooling techniques [27,28]. The CNN was originally designed to solve problems such as computer vision. However, the spectral characteristics of speech signals can be regarded as an image.

For example, everyone's pronunciation is distinct, so the frequency bands of the formant are different on spectrograms. A CNN's special structure with local weights sharing not only reduces the complexity of the network but can also learn complex information well. Therefore, the CNN that can eliminate difference effectively will facilitate acoustic model building. Abdel-Hamid et al. [29] gave the first demonstration that a CNN can normalize speaker differences on the frequency axis and reduce the phoneme error rate from 20.7% to 20.0% in the TIMIT phoneme recognition task. Subsequently, a CNN also achieved a relative improvement of 3%~5% in the continuous speech recognition task of a large vocabulary [30–32]. It is observed that a CNN applied to the feature extraction of speech recognition can overcome the diversity of speech signals using the translation invariance of convolution in time and space.

2.2. End-to-End Speech Recognition Based on LSTM-CTC

In the time series model, the most basic model is the recurrent neural network (RNN) [33], which has a wide range of application scenarios. The RNN can be used in fields such as speech recognition, machine translation, picture-taking, and question-and-answer systems. The RNN network structure improves on exploiting the information of sequence data because of its memory and powerful modelling ability in time series data learning. The background information of the data can be combined in a flexible manner, and the learning task can be effectively performed even for data in which a local distortion occurs. In practice, if the memory window is too long, the RNN will have problems, such as unstable training and gradient disappearance or explosion. Hochreiter and Schmidhuber et al. proposed a long short-term memory (LSTM) [34] to overcome the memory defects of the RNN. In the LSTM network, each neuron is a memory cell with an input gate, a forget gate, and an output gate to selectively remember historical information. The input gate determines when the input can enter the cell unit, the forget gate determines when the memory of the previous moment should be remembered, and the output gate decides when to let the memory flow to the next moment. The gated recurrent unit (GRU) model [35] can be understood as a simplified version of the LSTM network, but it retains the long-term memory function of the LSTM model. The input gate, forget gate, and output gate in the LSTM network are replaced with an update gate and a reset gate, and the two states of the cell state and the output are combined into one, so the GRU model has a strong contrast with the LSTM model.

For an acoustic input of T frames, the probability that the CTC network [36] learns to obtain label sequence π of length T is defined as

$$P(\pi|x) = \prod_{t=1}^{T} P(\pi_t|t, x) \tag{1}$$

Given a sequence of labels μ, input sequence $X = (x_1, x_2, \cdots, x_T)$ and time from 1 to T, the best path for the CTC network decoding is to find an output sequence with the highest probability.

$$\mu \approx B(\pi^*), \pi^* = \underset{\pi}{\operatorname{argmax}}\, P(\pi|x) \tag{2}$$

where π^* is the label sequence corresponding to the maximum posterior probability of T frames input sequence.

The aim of maximum likelihood training is to simultaneously maximize the log probabilities of all the correct classifications in the training set. This means minimizing the following objective function:

$$L_{CTC}(D) = -\ln \prod_{(x,y)\in D} p(\pi|x) = -\sum_{(x,y)\in D} \ln p(\pi|x) \tag{3}$$

where $(x, y) \in D$ denotes the training samples. The training is conducted using the back propagation through time (BPTT) algorithm.

The CTC algorithm differs from the traditional method in that it does not require the label to be aligned at the frame level in time to be trained. It does not care much about the predictions made at any point in the input data, and its focus is on whether the output is consistent with the label. The output of the CTC network is the probability of the overall sequence, which in turn reduces the tedious task of label pre-delineation. The CTC network output layer also contains a blank node, which is primarily used to model silences, pauses between words, and confusion between words. Therefore, the CTC is very good at addressing timing classification problems. In addition, the cascading structure of the BiLSTM and CTC networks has become a new standard combination in the field of speech recognition [37–39].

3. Proposed Method

In this paper, we propose to use cross-lingual speech recognition and transfer learning to establish a Tujia language speech recognition model. The model scheme is shown in Figure 1.

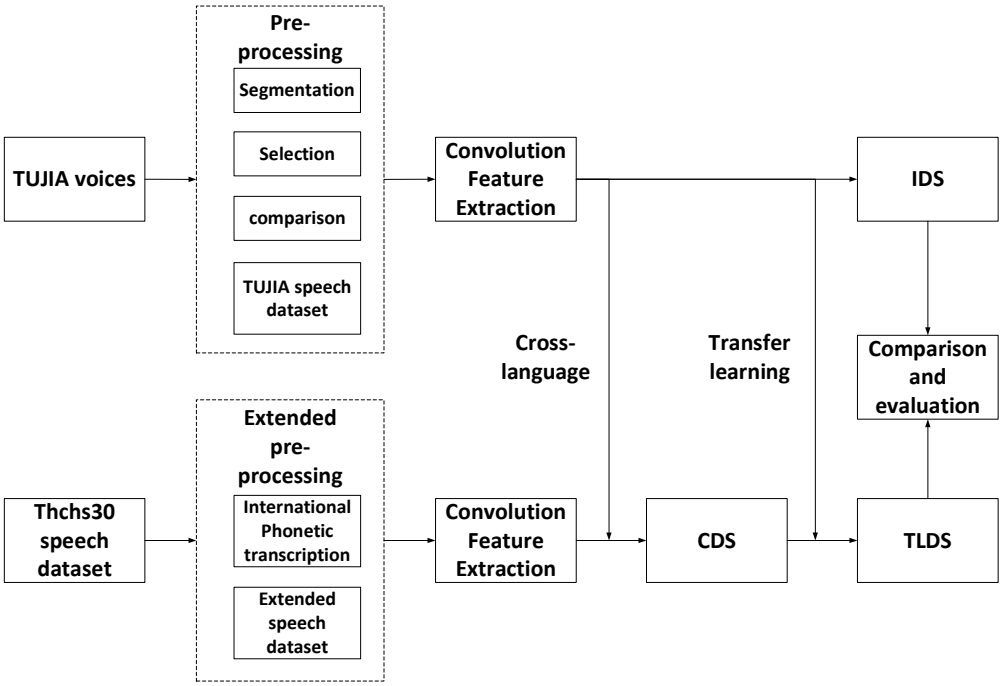


Figure 1. Model scheme.

First, the Tujia language database, the extended Chinese corpus and the cross-language corpus are established successively through data pre-processing. Then, convolution features are extracted from speech data. The speech recognition model based on the BiLSTM and CTC is constructed. Hereinto, Improved Deep Speech 2 (IDS) is a model obtained using the Tujia language corpus as training data. Cross-language Deep Speech 2 (CDS) is a model obtained using the Tujia language and Chinese corpus as training data. The sample transfer learning is performed on the initial model, CDS, and the CDS-based Transfer Learning Deep Speech 2 (TLDS) is obtained from the Tujia corpus as training data. Finally, according to the experimental results, the models are compared and evaluated.

3.1. Cross-Language Data Pre-processing and Feature Extraction

3.1.1. Tujia Language Corpus

In this study, the Tujia language corpus used in the experimental part includes 19 oral corpora. There is a total of 3716 sentences, with a total duration of 2 h, 54 min, and 4 s. The manual labelling of the Tujia language is completed using the Elan tool, and its contents include spoken to broad IPA, spoken to Chinese one-to-one translation, and spoken to Chinese translation. The manual annotation content of the Tujia language is shown in Table 1.

Table 1. The manual annotation content of the Tujia language.

Label Type	Label Content
Broad IPA	lai^{55} $xuã^{55}$ $lã^{55}$ ti^{21} xua^{21}, $mɬe^{35}$ su^{21} le^{53}
Chinese one-to-one translation	今天早晨（的话）天亮（过）
Chinese translation	今天早上天亮以后

Table 1 shows that the text processing of the Tujia language recording materials generally includes multiple levels. First, the voice is recorded with broad IPA or Tujia dialect symbols created, and then Tujia language words are translated one by one to Chinese. Finally, the sentences of the Tujia language are translated to Chinese. For the national language rather than the pure-speech language, it is necessary to write down the voice in the national language, which is called transliteration. The meaning of the label is more general and can be used for text processing including all layers. The narrowly defined labels refer to the grammatical labelling layer including the wording, prefix, suffix, and so on. The phoneme system of the Tujia language is composed of initials and finals. The 21 initials include two semi-vowel initials. The finals consist of 6 monophthongs, 11 diphthongs, and 8 orinasals [40,41].

3.1.2. Extended Speech Corpus

The thchs30 Chinese corpus [42] was recorded by 25 people. The text labels include Chinese characters and pinyin with tones. There are 13,395 sentences in total, the total recording time is 30 h, the sampling frequency is 16 kHz, and the sampling size is 16 bits. This paper needs to convert Chinese characters into broad IPA. The IPA transcription process of the thchs30 corpus is show in Table 2.

Table 2. Extended speech corpus data processing.

Label Type	Label Content
Chinese Character	菜 做好 了 一碗 清蒸 武昌鱼 一碗 蕃茄 炒鸡蛋 一碗 榨菜 干 子 炒肉丝
Chinese Pinyin	cai4 zuo4 hao3 le5 yi4 wan3 qing1 zheng1 wu3 chang1 yu2 yi4 wan3 fan1 qie2 chao3 ji1 dan4 yi4 wan3 zha4 cai4 gan1 zi3 chao3 rou4 si1
Narrow IPA	ts^hai(51) tsuo(51) xɑu(214) lɣi(51) uan(214) $tɕ^h$iŋ(55) tʂəŋ(55) u(214) $tʂ^h$ɑŋ(55) y(35) i(51) uan(214) fan(55) $tɕ^h$iɛ(35) $tʂ^h$ɑu(214) tɕi(55) tan(51) i(51) uan(214) tʂa(51) ts^hai(51) kan(55) tsɻ(214) $tʂ^h$ɑu(214) rou(51) sɻ(55)
Broad IPA	ts^hai(51) tsuo(51) xau(214) lɣi(51) uan(214) $tɕ^h$iŋ(55) tʂəŋ(55) u(214) $tʂ^h$aŋ(55) y(35) i(51) uan(214) fan(55) $tɕ^h$iɛ(35) $tʂ^h$au(214) tɕi(55) tan(51) i(51) uan(214) tʂa(51) ts^hai(51) kan(55) tsi(214) $tʂ^h$au(214) rou(51) si(55)

First, from Chinese characters to pinyin, then to IPA. The Broad IPA are used in the process of recording the Tujia language, and the Narrow IPA are used in Chinese. Therefore, it is necessary to convert narrow IPA into broad IPA, which is consistent with the Tujia language IPA. Its conversion rules are shown in Table 3.

Table 3. The rules for converting Chinese IPA into Tujia language IPA.

Narrow IPA	Broad IPA
ɑ	a
iou	iu
uei	ui
iɛn	ian
ɿ	i
ɻ	i

The Chinese speech data set is used as the extended data of the Tujia language, which solves the problem of the insufficient voice data for the Tujia language. At the same time, a cross-language corpus is constructed and a unified IPA code dictionary for the Chinese and Tujia languages is established. This paper compares the similarities and differences between the Tujia and Chinese languages according to the Jaccard similarity coefficient [43]. The definition of the Jaccard index is as follows:

$$J(A_{Tujia}, B_{Chinese}) = \frac{|A_{Tujia} \cap B_{Chinese}|}{|A_{Tujia} \cup B_{Chinese}|} \quad (4)$$

where A_{Tujia} is the IPA transcription of the Tujia language, $B_{Chinese}$ is the Chinese IPA transcription, the denominator represents the intersection of the Tujia language IPA and the Chinese IPA, and the numerator represents the union of the Tujia language IPA and the Chinese IPA. The statistical results show that the similarity between the IPA dictionaries of the Tujia and Chinese languages is 53.33%. Therefore, according to the definition of transfer learning, the sample transfer method can be used to model the Tujia language data. In fact, the thchs30 data set is large. To ensure the normal fitting of the CDS model to the Tujia language, the experimental comparison results show that when the Tujia language and Chinese corpus are 1:1.2 in the number of sentences, the model recognition effect is optimal, so for this paper only 4460 Chinese data were used when training CDS.

3.1.3. Feature Extraction

In general, speech recognition is based on the speech spectrum after the time-frequency analysis, in which the speech time spectrum is structurally characterized. To improve the speech recognition rate it is necessary to overcome the diversity of speech signals, including the diversity of speakers (between the speaker and the speaker), and the diversity of the environment. For modelling capabilities, a CNN can effectively reduce the frequency domain changes [44]. As shown in the following figure, the 768-dimensional convolution feature of Tujia and Chinese speech is compared with the traditional 13-dimensional MFCC speech features.

In Figures 2 and 3, the horizontal axis is the IPA tag instead of the time, and the vertical axis is the frequency. The different colors indicate the amount of energy. As marked by a red rectangle in the above figure, the energy change is particularly noticeable after a period of silence. In addition, after the high-dimensional feature extraction of the convolutional neural network, because the feature dimension increases substantially, the frequency domain energy change is smaller than the MFCC feature, which is more conducive for the subsequent LSTM network learning.

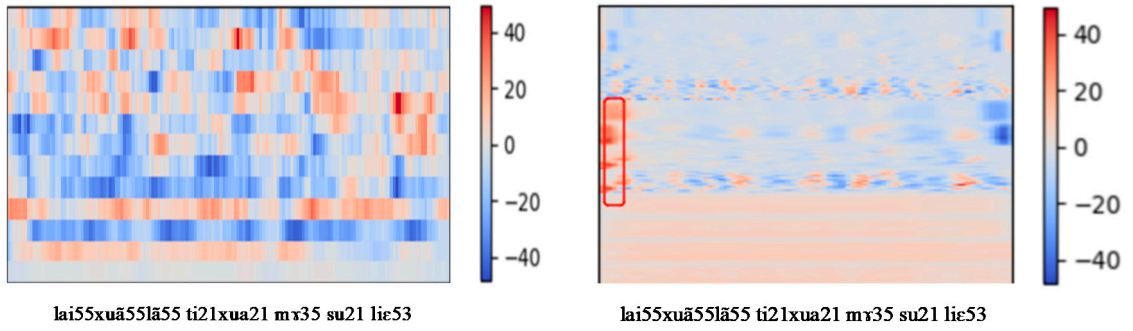


Figure 2. Tujia MFCC feature (**left**) and Tujia convolution feature (**right**).

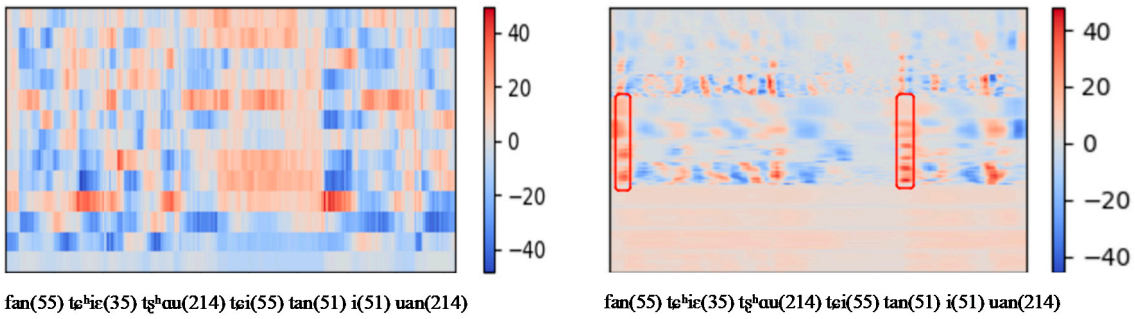


Figure 3. Chinese MFCC feature (**left**) and Chinese convolution feature (**right**).

3.2. End-to-End Speech Recognition Model Based on Transfer Learning

To solve the problem of the low resources of the Tujia language and establish a better speech recognition model, this paper adopts a baseline model based on Deep Speech 2 [45]. The experiments

of Deep Speech 2 using English and Mandarin Chinese data show that the recognition error rate is reduced by more than 10%. In this paper, the CNN, BiLSTM, and CTC networks are also used in combination to construct a cross-language end-to-end speech recognition system. For the target domain Tujia language voice data set D_t, the Chinese voice data set D_s of related but different fields is used as the source domain voice data set. In the target domain speech data set, given the input sequence $X_t = (x_1, x_2, \cdots, x_n)$, its corresponding output sequence is $Y_t = (y_1, y_2, \cdots, y_m)$; in the source domain speech data set, given the input sequence $X_s = (x_1, x_2, \cdots, x_{n'})$, its corresponding output sequence is $Y_s = (y_1, y_2, \cdots, y_{m'})$.

According to the coding dictionary of the Tujia language and Chinese IPA, that is, the category of model output, the weight of the model softmax layer can be divided into $W = \{w_t, w_s, w_c\}$, which corresponds to the weight of the intersection of the Tujia language IPA, the Chinese IPA, and the IPA dictionary of the two languages. The output of the model's softmax layer is defined as

$$P(k|t, x) = \frac{\exp(y_t^k)}{\sum_{k=1}^{K} \exp(y_t^{k\prime})} \tag{5}$$

where K is the total number of tags, x is the feature input at time t, and y is the output at time t. In this work, the model parameters are updated immediately following the SortaGrad batch, as elaborated in Algorithm 1.

Algorithm 1 Training End-to-end Speech Recognition Model Based on Transfer Learning

Input: $D_t = \{x_n, y_m\}, D_s = \{x_{n'}, y_{m'}\}$, training set
$x_n, x_{n'}$, input features
$y_m, y_{m'}$, output labels
μ, learning rate
Output: θ_1, CDS parameters
θ_2, TLDS parameters
Initialize CDS parameters θ_1
While model does not converge θ_1 **do**
Read a SortaGrad batch $S = \left\{x_i, y_j\right\}_{b=1}^{B}$ from D_t, D_s
Train model using S,
$\theta_1 \leftarrow \theta_1 - \frac{\mu}{B}\sum_{b=1}^{B} L_{CTC}(x_i, y_j, \theta_1)$
Calculate $\{w_t, w_s, w_c\}$ of the softmax layer using equation (5)
Calculate L_{CTC} using Equation (3) for S
End while
Initialize TLDS parameters θ_2
While model does not converge θ_2 **do**
Read a SortaGrad batch $S' = \{x_n, y_m\}_{b'=1}^{B'}$ from D_t
Train model using S',
$\theta_2 \leftarrow \theta_2 - \frac{\mu}{B'}\sum_{b'=1}^{B'} L_{CTC}(x_n, y_m, \theta_2)$
Calculate $\{w_t', w_c'\}$ of the softmax layer using Equation (5)
Calculate L_{CTC}' using Equation (3) for S'
End while

The cross-language end-to-end speech recognition model structure based on transfer learning is shown in Figure 4. In the training phase of CDS, the input of the model is the spectrum of the Tujia language and Chinese phonetics. After the multi-layer convolutional layer is extracted, it enters the multi-layer BiLSTM, which completes the cross-language acoustic feature extraction and the shared hidden layer weight learning. At this time, in the softmax layer of the model, the weights of the IPA of the Tujia language, the Chinese IPA, and the IPA dictionary of the two languages are $\{w_t, w_s, w_c\}$, respectively, and finally the CTC model is used instead of the traditional HMM to calculate the transition probability between states. The recognition result is decoded by the Tujia language

code dictionary, and the output of the model is the Tujia language IPA. However, sample transfer is performed on the model CDS. First, the weight w_{s-c} corresponding to the difference set of the D_s dictionary with respect to the D_t dictionary is set to 0. At this point, the output of the model is only the Tujia language IPA sequence, and then the model is continued using the Tujia language data. A few iterative trainings are performed to improve the generalization and stability of the model to the Tujia language data, and finally the model TLDS is obtained.

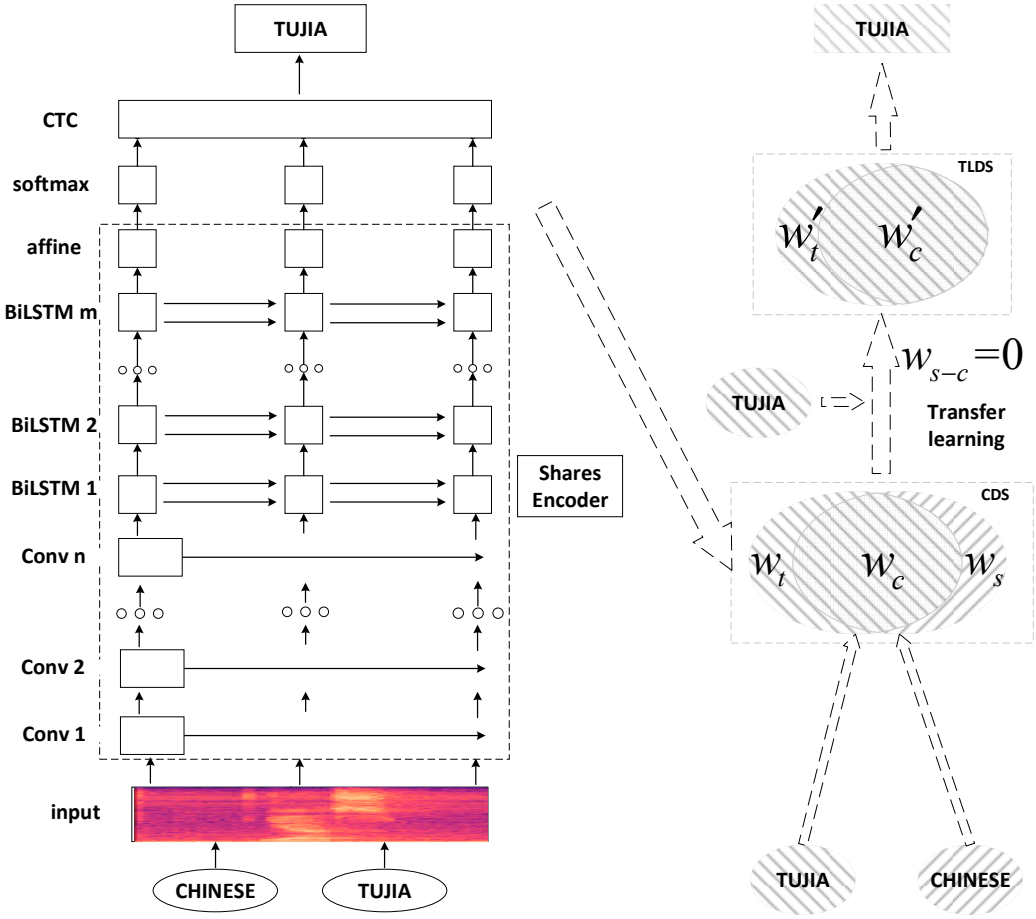


Figure 4. Cross-language end-to-end speech recognition model based on transfer learning for the Tujia language.

In the decoding phase, the CTC can be viewed as an objective function that directly optimizes the likelihood of the input sequence and the output target sequence. For this objective function, the CTC automatically learns and optimizes the correspondence between input and output sequences during the training process. The input layer of the CTC network is the softmax layer, and the number of nodes and label sequences are identical. In addition, the blank node plays an important role in solving the problem of overlays. It is difficult to give a label in one frame of data in speech recognition, but it is easy to determine the corresponding pronunciation label in tens of frames. In the CTC network, it is the existence of the blank node that can take the method of frame skipping. The CTC output and label sequence satisfy an equivalence relationship similar to the following equation [36]:

$$F(a-an-) = F(-aa--an) = aan \tag{6}$$

where "*a*" and "*n*" are the IPA of the Tujia language, and "-" is the blank. As observed from Equation (4), multiple input sequences can be mapped to one output. Therefore, the CTC can not only increase

the decoding speed but also automatically optimize the correspondence between input and output sequences during the training process.

4. Experiments and Result

4.1. Experimental Environment

In this study, the Dell PowerEdge R730 server device is used, in which the processor is Intel(R) Xeon(R) CPU E5-2643 v3 @3.40 GHz, the memory size is 64 G, the GPU is NVIDIA Tesla K40 m × 2, and the memory size is 12 GB × 2. The experimental environment for the deep learning framework installed on the Ubuntu 16.04 system is the GPU version of paddlepaddle 0.12.0.

4.2. Parameters of the Models

The output of the model is the Tujia language IPA, and the input of the model is the spectrogram of the Tujia language and Chinese phonetics. First, the spectrogram is subjected to a layer of convolution and a layer of maximum pooling, and the feature data obtained by the maximum pooling is more sensitive to the information of the texture features. Then, the output of the previous layer enters the second layer of convolution, where the input of each node is a small block of the last largest pooling layer, and each small block is further analyzed to extract more abstract features, that is, high-dimensional features. The structure of the convolutional network is shown in Table 4.

Table 4. CNN parameters.

Convolutional Layer / Parameters	First Convolutional Layer	Second Convolutional Layer
Filter size	11 × 41	11 × 21
Number of input channels	1	1
Number of output channels	32	32
Stride size	3 × 2	1 × 2
Padding size	5 × 20	5 × 10

Second, the high-dimensional features extracted by the CNN enter the 3-layer BiLSTM. The parameters are shown in Table 5. The number of hidden layer nodes is 512 in each layer, the learning rate is 0.001, and the minibatch size is 16.

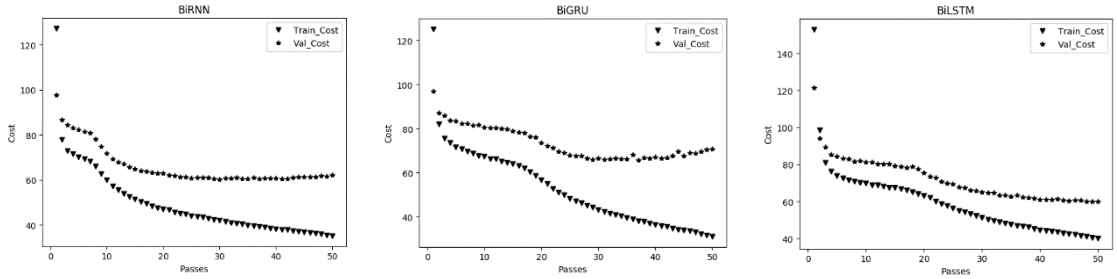
Table 5. BiLSTM (Bi-directional long short-term memory)-CTC (connectionist temporal classification) parameters.

Network	Parameter Type	Parameter Content
BiLSTM	minibatch size	16
	LSTM layers	3
	cells of per layers	512
	learning rate	0.001
CTC	Decoder	beam search

After the spectrogram passes through the CNN and BiLSTM networks, it indicates that the cross-language sharing acoustic feature extraction is completed, and the context information is fully mined and learned. The CNN trains the speech frame by frame and moves the BiLSTM to the frame-by-frame training so that the GPU can perform parallel computation at high speed. Because the Tujia language corpus has many short sentences, the CTC's choice decoding method is beam search. At the same time, the experimental results show that the accuracy of the beam search decoding method is 3.2% higher than the best path decoding method in the Tujia language data set.

In the model training process, we use batch normalization (BN) [46] and SortaGrad [45] to optimize the model. BN is a standardized process for the input data of each layer during the training of the neural network. The traditional neural network only normalizes the sample (such as the subtraction mean) before the sample enters the input layer to reduce the difference between the samples. In BN, not only is the input data of the input layer standardized but the input for each hidden layer is also normalized. SortaGrad gradually increases the length of the sentence in the batch according to the duration of the voice data. This not only accelerates model convergence but also makes the model more stable.

In the Tujia language corpus, when the other parameters are consistent, we modify the RNN cell state to compare the change in the loss function value during the training of the model. Finally, the result of the character error rate (CER) over the test set is shown in Figure 5.



(**a**) the cost of BiRNN (**b**) the cost of BiGRU (**c**) the cost of BiLSTM

Figure 5. The change in the cost of BiRNN, BiGRU, and BiLSTM's as the number of training passes increases.

In Figure 5, the abscissa represents the number of iterations of the training model, and the ordinate represents the cost value of the training and validation sets.

In Figure 6, the abscissa represents the number of iterations of the training model, and the ordinate represents the time of training different RNN cell states.

Figures 5 and 6 and Table 6 show that although the BiRNN network training speed is faster than that of BiGRU and BiLSTM; the final recognition accuracy rate is highest for BiLSTM. According to our actual needs, we choose to use the BiLSTM network.

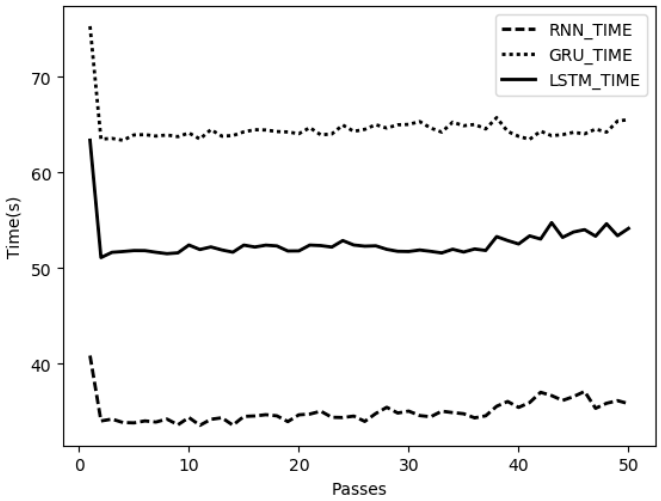


Figure 6. Time to train different RNN (Recurrent neural network) cell states.

Table 6. Character error rate of different RNN cells over the development and test set.

RNN Cell	Dev	Test
BiRNN (Bi-directional recurrent neural network)	42.37%	53.37%
BiGRU (Bi-directional long short-term memory)	37.09%	51.95%
BiLSTM (Bi-directional gated recurrent unit)	35.82%	48.30%

4.3. Experimental Results

The experimental results in Table 7 show that the IDS model obtained using only the Tujia corpus as the training data is better than the CDS model obtained using the Tujia language and the Chinese corpus as the training data, and the sample transfer learning is performed on the initial model CDS. The CDS-based TLDS is obtained from the Tujia corpus as training data, and its performance is better than that of IDS. Therefore, the scheme of this paper is feasible and effective for speech recognition of the Tujia language with less data.

Table 7. Tujia language recognition rate of different models.

Model	Dev	Test
IDS (Improved Deep Speech 2)	35.82%	48.30%
CDS (Cross-language Deep Speech 2)	40.11%	50.26%
TLDS (Transfer Learning Deep Speech 2)	31.11%	46.19%

5. Conclusions

In this paper, we combined the CNN, BiLSTM, and CTC algorithms to construct a cross-language end-to-end speech recognition system. For the low resource problem of the Tujia language, the Chinese data were used to expand the Tujia language data, the cross-language recognition method was used to share acoustic features, and the sample transfer learning method was used to optimize the model. The final recognition accuracy was increased by 4.08%. However, the Tujia language is only a spoken language, and there is no text or strict grammatical structure, so it is difficult to establish a language model. In subsequent work, we will attempt to build a dictionary model based on common spoken vocabulary in the Tujia language.

Author Contributions: Conceptualization, C.Y.; methodology, C.Y. and Y.C.; software, Y.C. and S.X. and M.K.; validation, Y.C., Y.L. and M.K.; formal analysis, C.Y. and Y.C.; investigation, Y.L. and X.L.; resources, S.X. and C.Y.; data curation, S.X.; writing—original draft preparation, Y.C.; writing—review and editing, C.Y. and Y.C.; visualization, Y.C.; supervision, C.Y.; funding acquisition, C.Y.

Funding: This research was funded by Ministry of Education Humanities and Social Sciences Research Planning Fund Project, grant number 16YJAZH072, and Major projects of the National Social Science Fund, grant number 14ZDB156.

Acknowledgments: We would like to acknowledge Beijing Key Laboratory of Big Data Technology for Food Safety and Key Laboratory of Resources utilization and Environmental Remediation for providing a research grant to conduct this work. We express gratitude to the editors for the editing assistance. Lastly, we would like to thank the reviewers for their valuable comments and suggestions on our paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Xu, S. The course and prospect of endangered language studies in China. *J. Northwest Univ. Natl. Philos. Soc. Sci.* **2015**, 83–90. [CrossRef]
2. Rosenberg, A.; Audhkhasi, K.; Sethy, A.; Ramabhadran, B.; Picheny, M. End-to-end speech recognition and keyword search on low-resource languages. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, New Orleans, LA, USA, 5–9 March 2017; pp. 5280–5284.
3. Hannun, A.; Case, C.; Casper, J.; Catanzaro, B.; Diamos, G.; Elsen, E.; Prenger, R.; Satheesh, S.; Sengupta, S.; Coates, A.; et al. Deep Speech: Scaling up end-to-end speech recognition. *arXiv* **2014**, arXiv:1412.5567.

4. Zhang, Y.; Chan, W.; Jaitly, N. Very Deep Convolutional Networks for End-to-End Speech Recognition. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 4845–4849.
5. Lu, L.; Kong, L.; Dyer, C.; Smith, N.A. Multitask Learning with CTC and Segmental CRF for Speech Recognition. In Proceedings of the INTERSPEECH 2017, Stockholm, Sweden, 20–24 August 2017; pp. 954–958.
6. Ochiai, T.; Watanabe, S.; Hori, T.; Hershey, J.R. Multichannel End-to-end Speech Recognition. *arXiv* **2017**, arXiv:1703.04783.
7. Parcollet, T.; Zhang, Y.; Morchid, M.; Trabelsi, C.; Linarès, G.; De Mori, R.; Bengio, Y. Quaternion Convolutional Neural Networks for End-to-End Automatic Speech Recognition. *arXiv* **2018**, arXiv:1806.07789.
8. Ghoshal, A.; Swietojanski, P.; Renals, S. Multilingual training of deep neural networks. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 7319–7323.
9. Heigold, G.; Vanhoucke, V.; Senior, A.; Nguyen, P.; Ranzato, M.A.; Devin, M.; Dean, J. Multilingual acoustic models using distributed deep neural networks. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 8619–8623.
10. Evgeniou, T.; Pontil, M. Regularized multi-task learning. In Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 22–25 August 2004; pp. 109–117.
11. Pan, S.J.; Yang, Q. A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.* **2010**, *22*, 1345–1359. [CrossRef]
12. Dalmia, S.; Sanabria, R.; Metze, F.; Black, A.W. Sequence-based Multi-lingual Low Resource Speech Recognition. *arXiv* **2018**, arXiv:1802.07420.
13. Kim, S.; Hori, T.; Watanabe, S. Joint CTC-Attention based End-to-End Speech Recognition using Multi-task Learning. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 4835–4839.
14. Chen, W.; Hasegawa-Johnson, M.; Chen, N.F. Topic and Keyword Identification for Low-resourced Speech Using Cross-Language Transfer Learning. In Proceedings of the INTERSPEECH, Hyderabad, India, 2–6 September 2018; pp. 2047–2051.
15. Latif, S.; Rana, R.; Younis, S.; Qadir, J.; Epps, J. Transfer Learning for Improving Speech Emotion Classification Accuracy. *arXiv* **2018**, arXiv:1801.06353.
16. Deng, L. An Overview of Deep-Structured Learning for Information Processing. In Proceedings of the Asian-Pacific Signal and Information Processing-Annual Summit and Conference (APSIPA-ASC), Xi'an, China, 19–22 October 2011.
17. Bengio, Y. Learning Deep Architectures for AI. *Found. Trends Mach. Learn.* **2009**, *2*, 1–127. [CrossRef]
18. Mohamed, A.R.; Dahl, G.; Hinton, G. Deep Belief Networks for phone recognition. In *Nips Workshop on Deep Learning for Speech Recognition and Related Applications*; MIT Press: Whister, BC, Canada, 2009; Volume 39.
19. Lin, H.; Ou, Z. Partial-tied-mixture Auxiliary Chain Models for Speech Recognition Based on Dynamic Bayesian Networks. In Proceedings of the IEEE International Conference on Systems, Taipei, Taiwan, 8–11 October 2006.
20. Pundak, G.; Sainath, T.N. Lower Frame Rate Neural Network Acoustic Models. In Proceedings of the INTERSPEECH, San Francisc, CA, USA, 8–12 September 2016; pp. 22–26.
21. WÖLfel, M.; Mcdonough, J. Speech Feature Extraction. In *Distant Speech Recognition*; John Wiley & Sons, Ltd.: Hoboken, NJ, USA, 2009.
22. Lee, J.H.; Jung, H.Y.; Lee, T.W. Speech feature extraction using independent component analysis. In Proceedings of the IEEE International Conference on Acoustics, Istanbul, Turkey, 5–9 June 2000.
23. Li, H.; Xu, X.; Wu, G.; Ding, C.; Zhao, X. Research on speech emotion feature extraction based on MFCC. *J. Electron. Meas. Instrum.* **2017**. (In Chinese) [CrossRef]
24. Yu, D.; Seltzer, M. Improved Bottleneck Features Using Pretrained Deep Neural Networks. In Proceedings of the Conference of the International Speech Communication Association, Florence, Italy, 27–31 August 2011; pp. 237–240.
25. Maimaitiaili, T.; Dai, L. Deep Neural Network based Uyghur Large Vocabulary Continuous Speech Recognition. *J. Data Acquis. Process.* **2015**, 365–371. [CrossRef]
26. Liu, X.; Wang, N.; Guo, W. Keyword Spotting Based on Deep Neural Networks Bottleneck Feature. *J. Chin. Comput. Syst.* **2015**, *36*, 1540–1544.

27. Ozeki, M.; Okatani, T. Understanding Convolutional Neural Networks in Terms of Category-Level Attributes. In *Computer Vision—ACCV 2014*; Springer International Publishing: Cham, Switzerland, 2014; pp. 362–375.
28. Zeiler, M.D.; Fergus, R. Visualizing and Understanding Convolutional Networks. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2013; Volume 8689, pp. 818–833.
29. Abdel-Hamid, O.; Mohamed, A.R.; Jiang, H.; Penn, G. Applying Convolutional Neural Networks concepts to hybrid NN-HMM model for speech recognition. In Proceedings of the IEEE International Conference on Acoustics, Kyoto, Japan, 25–30 March 2012.
30. Abdelhamid O Deng, L.; Yu, D. Exploring Convolutional Neural Network Structures and Optimization Techniques for Speech Recognition. In Proceedings of the INTERSPEECH, Lyon, France, 25–29 August 2013; pp. 3366–3370.
31. Sercu, T.; Puhrsch, C.; Kingsbury, B.; LeCun, Y. Very deep multilingual convolutional neural networks for LVCSR. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Shanghai, China, 20–25 March 2016; pp. 4955–4959.
32. Sercu, T.; Goel, V. Advances in Very Deep Convolutional Neural Networks for LVCSR. *arXiv* **2016**, arXiv:1604.01792.
33. Graves, A.; Mohamed, A.R.; Hinton, G. Speech recognition with deep recurrent neural networks. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 March 2013; pp. 6645–6649.
34. Graves, A. *Supervised Sequence Labelling with Recurrent Neural Networks*; Springer: Berlin/Heidelberg, Germany, 2012.
35. Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *arXiv* **2014**, arXiv:1412.3555.
36. Graves, A.; Gomez, F. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the International Conference on Machine Learning, Pittsburgh, PA, USA, 25–29 June 2006; pp. 369–376.
37. Miao, Y.; Gowayyed, M.; Metze, F. EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding. In Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, Scottsdale, AZ, USA, 13–17 December 2016; pp. 167–174.
38. Zeghidour, N.; Usunier, N.; Synnaeve, G.; Collobert, R.; Dupoux, E. End-to-End Speech Recognition from the Raw Waveform. *arXiv* **2018**, arXiv:1806.07098.
39. Watanabe, S.; Hori, T.; Karita, S.; Hayashi, T.; Nishitoba, J.; Unno, Y.; Soplin, N.E.Y.; Heymann, J.; Wiesner, M.; Chen, N.; et al. ESPnet: End-to-End Speech Processing Toolkit. *arXiv* **2018**, arXiv:1804.00015.
40. Xu, S. Grammatical and semantic representation of spatial concepts in the Tujia language. *J. Minor. Lang. China* **2013**, *1*, 35–45.
41. Xu, S. Features of Change in the Structure of Endangered Languages: A Case Study of the South Tujia Language. *J. Yunnan Natl. Univ. (Soc. Sci.)* **2012**, *29*. (In Chinese) [CrossRef]
42. Wang, D.; Zhang, X. THCHS-30: A Free Chinese Speech Corpus. *arXiv* **2015**, arXiv:1512.01882.
43. Wu, C.; Wang, B. Extracting Topics Based on Word2Vec and Improved Jaccard Similarity Coefficient. In Proceedings of the IEEE Second International Conference on Data Science in Cyberspace, Shenzhen, China, 26–29 June 2017.
44. Inc, G. Convolutional, long short-term memory, fully connected deep neural networks. In Proceedings of the IEEE International Conference on Acoustics, Brisbane, Australia, 19–24 April 2015.
45. Amodei, D.; Ananthanarayanan, S.; Anubhai, R.; Bai, J.; Battenberg, E.; Case, C.; Casper, J.; Catanzaro, B.; Cheng, Q.; Chen, G.; et al. Deep Speech 2: End-to-End Speech Recognition in English and Mandarin. *arXiv* **2015**, arXiv:1512.02595.
46. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 448–456.