IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 2, 2020, accepted August 20, 2020, date of publication August 31, 2020, date of current version September 18, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3020421

Acoustic Modeling Based on Deep Learning for Low-Resource Speech Recognition: An Overview

CHONGCHONG YU[1], MENG KANG[1], YUNBING CHEN[2], JIAJIA WU[1], AND XIA ZHAO[1]
[1]Key Laboratory of Industrial Internet and Big Data, Beijing Technology and Business University, Beijing 100048, China
[2]Putian Information Technology Company, Ltd., Beijing 100080, China

Corresponding author: Chongchong Yu (yucc@btbu.edu.cn)

This work was supported in part by the Ministry of Education Humanities and Social Sciences Research Planning Fund Project under Grant 16YJAZH072, in part by the Major projects of the National Social Science Fund under Grant 14ZDB156, and in part by the Graduate Student Research Capacity Improvement Program of Beijing Technology and Business University in 2020.

ABSTRACT The polarization of world languages is becoming more and more obvious. Many languages, mainly endangered languages, are of low-resource attribute due to lack of information. Both language conservation and cultural heritage face important challenges. Therefore, speech recognition for low- resource scenario has become a hot topic in the field of speech. Based on the complex network structures and huge model parameters, deep learning has become a powerful science in the process of speech recognition, which has a broad and far-reaching significance for the study of low-resource speech recognition. Aiming at the characteristic of low resource, this article reviews the history and research status of two kinds of acoustic models of deep learning neural networks and acoustic end-to-end structures. We further elaborate on several key techniques for improving performance in the two aspects of data and model training. There are two projects for low-resource languages introduced in this article. The possible future developments are finally pointed out. These works provide some reference for computer speech and language processing.

INDEX TERMS Low-resource languages, automatic speech recognition, acoustic model, data augmentation, multitask learning, transfer learning, meta learning.

I. INTRODUCTION

Speech is the most simple and smooth way of communication in human interaction, which can quickly and accurately convey effective information. Nowadays, people are devoted to studying how to communicate with various smart devices through the medium of speech while using them. At present, there are a variety of voice assistants that understand human voice information through interactive real-time intelligent dialogue and realize automatic operation according to the content, such as Apple's Siri and Google's assistant. Therefore, Automatic Speech Recognition (ASR) is the key technology throughout the human-computer interaction processing. The purpose of ASR is to transform speech signals into textual information, thus providing a strong foundation for further semantic understanding. ASR is an interdisciplinary and comprehensive technology, including computer technology, acoustics, digital signal processing, statistics, linguistics, artificial intelligence, and so on. Thanks to the rapid development of related disciplines, the performance of ASR system has been greatly improved and widely used in various scenarios, including military, medical, and service industries, which greatly saves human resources and improves work efficiency.

The rapid development of ASR relies more on the support of a large amount of speech data and annotated text. The recognition of majority languages such as Chinese and English has achieved very mature performance, but it is difficult to be applied in many dialects and minority languages. There are many kinds of these languages that lack resources in terms of transcribed speech data, pronunciation dictionaries, language knowledge, text annotation, and others, which are defined as low-resource languages [1]. It is estimated that there are more than 7,000 languages in the world, among which at least 40% are endangered languages, and about half of them have no written forms [2]. From the perspective of

The associate editor coordinating the review of this manuscript and approving it for publication was Javier Medina.

natural ecology, the language pattern of the contemporary world is polarized between the major languages and the endangered languages.

As a carrier and an important part of culture, the extinction of language is an irreparable loss to rich language resources of the world and greatly damage the diversity of culture. Faced with the impact of majority languages, globalization, and internet, the international linguistic community has gradually attached importance to the theoretical creation and technological development of language protection. It has adopted high-tech digital means to collect all kinds of endangered languages in the world and established audio information database of digital languages [3]. However, the processing of speech in the construction of corpora is extremely difficult. All the speech in the audio and video materials obtained through field investigations must be manually annotated to make these corpora more widely understandable. Manual annotation requires a lot of manpower. Moreover, there is a shortage of native speakers or professionals who can perform corpus annotation processing. As a result, a lot of original audio or video materials of many languages are piled up and cannot be processed. Efforts to preserve linguistic diversity and sustainability remain to be explored.

Although ordinary ASR systems are not ideal for low-resource languages that are difficult to provide large amounts of training data, the use of speech technology is still the most direct and effective way to conduct research, as demonstrated by unwritten languages. More and more scholars have begun to conduct in-depth research and continuous improvement on speech recognition technology under the condition of limited language resources. This work is called low-resource speech recognition, which has become one of the hot issues and important challenges in the field of ASR. Figure 1 is a statistical graph of the quantity of published articles on the Web of Science website, Engineering Village and Scopus about the keywords Low Resource and Speech Recognition in the past decade. On the whole, the number of papers is increasing year by year, which reflects the international attention on low-resource speech recognition research. Low-resource speech recognition has important research significance, especially playing an irreplaceable role in the protection and promotion of linguistic diversity, cultural inheritance, and human communication in the world.

The rest of this article is organized as follows. In Section 2, we briefly review the principle and history of ASR. In Section 3, we introduce the application and improvement of acoustic models based on neural networks and end-to-end structure in low-resource scenario. Section 4 discusses how to improve the performance of low-resource speech recognition from data-wise and model-wise. Section 5 introduces two projects for low-resource languages. The possible future works of the low-resource speech recognition are given in Section 6.

II. BACKGROUND OF AUTOMATIC SPEECH RECOGNITION

The traditional architecture of speech recognition system is shown in the Figure 2, which consists of four parts: feature extraction, acoustic model, language model, and decoder. The signal preprocessing and feature extraction take speech signal as input, enhance speech quality by eliminating noise and channel distortion and transform signal from time domain to frequency domain for extracting feature vector. After the acoustic model and language model, the best sequence W^* corresponding to the speech is output using the maximum posterior probability criterion and related decoding algorithm. W^* can be calculated by Bayesian formula. The specific calculation method of W^* is shown in the (1):

$$W^* = \arg\max_W P(W|O) = \arg\max_W \frac{P(O|W) \cdot P(W)}{P(O)} = \arg\max_W P(O|W) \cdot P(W) \quad (1)$$

where W is text label sequence, O is feature vector, $P(W)$ is the probability of language model, and $P(O|W)$ is the probability of acoustic model. In other words, on the premise of a given text label sequence W, the probability of acoustic feature sequence O is obtained, which can measure the matching degree of the speech feature sequence O and the text label sequence W to construct the model. $P(W|O)$ represents the posterior probability of W, which can be used as the output text sequence of speech recognition with the idea of posterior probability maximization. For a given label sequence, the feature vector is fixed, which does not affect the recognition result, so it can be ignored. The acoustic model plays the most important role in ASR, which is the focus in this article.

The most popular ASR systems in recent years usually use spectrogram, Linear Prediction Coefficient (LPC), Mel Frequency Cepstrum Coefficient (MFCC), Perceptual Linear Predictive (PLP) and feature transformation methods

including Linear Discriminant Analysis (LDA), Maximum Likelihood Linear Transformation (MLLT), i-Vector, feature Maximum Likelihood Linear Regression (fMLLR) and others as speech feature extraction. The breakthrough period of acoustic model was in the 1980s. The method based on statistical models which were represented by the Gaussian Mixture Model-Hidden Markov model (GMM-HMM)

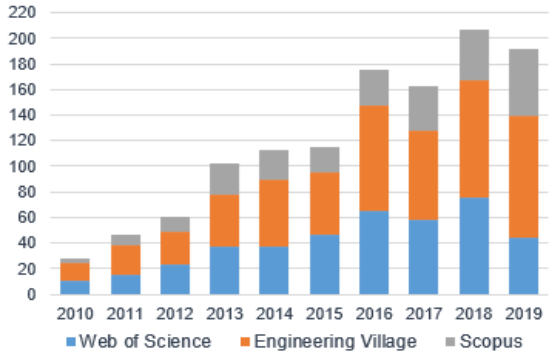


FIGURE 1. Statistics on published papers according to year retrieved in Web of Science, Engineering Village and Scopus with the subject terms of low resource and speech recognition. The horizontal axis represents the year. The vertical axis represents the quantity.

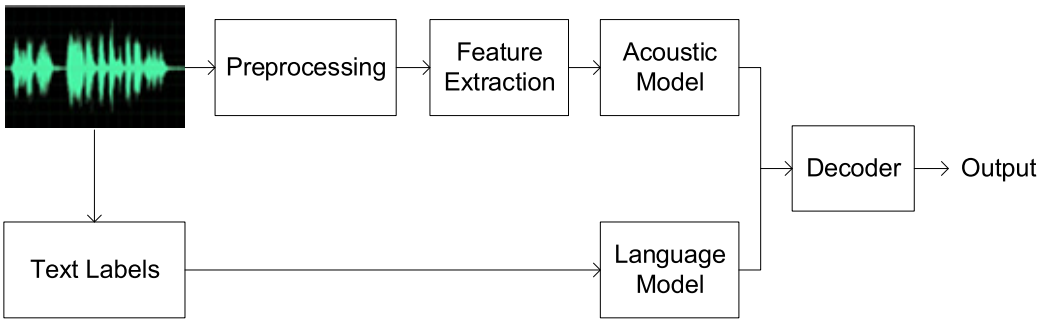


FIGURE 2. The traditional architecture of speech recognition system.

method [4] gradually became dominant in speech recognition research. A series of other related technologies based on the HMM method have also been derived, such as the use of Maximum Likelihood Linear Regression (MLLR) [5] and the maximum posterior probability criterion to overcome the problem of parameter adaptation in the HMM training process. Furthermore, the idea of merging states was used to achieve decision tree state tying when there are many training parameters in the case of less training data [6]. Artificial Neural Network (ANN) also provided a new research idea for speech recognition afterwards. In 2006, Hinton *et al.* [7] used Restricted Boltzmann Machine (RBM) to initialize the nodes of the neural network and the Deep Belief Network (DBN) came into being. The network used a non-supervised greedy layer-by-layer method to keep the weight of the modeled object as much as possible, and continuously fitted to obtain the weight. Since then, the combination of deep learning and traditional methods has occupied the mainstream, and Deep Neural Network (DNN) has shown a trend of surpassing the GMM model, instead of using the traditional GMM method to HMM state modeling. The first breakthrough was the DNN-HMM acoustic model, which greatly promoted the application of deep learning in speech recognition [8]. This is enough to demonstrate the power of deep learning.

Most importantly, the powerful feature extraction ability of Convolutional Neural Network (CNN) can better understand complex speech features. Recurrent Neural Network (RNN), which is suitable for sequence modeling, can make better use of the characteristics of time series relationships to establish context-dependent models. In particular, the latest end-to-end speech recognition system overcomes the problem of forced alignment of traditional HMM and realizes the overall optimization of sentence sequence. The two most mainstream end-to-end models are the Connectionist Temporal Classification (CTC) and the encoder-decoder model based on attention mechanism.

Deep learning uses the multi-layer nonlinear structure to transform the low-level features into more abstract high-level features, and transforms the input features with or without supervision, thereby improving the accuracy of classification or prediction [9]. Deep learning models generally refer to deeper structural models, which have more layers of nonlinear transformations than traditional shallow models. They are more powerful in expression and modeling [10]. They also have advantages in complex signal processing such as non-stationary and random speech signals.

III. LOW-RESOURCE SPEECH RECOGNITION ACOUSTIC MODELS

A. ACOUSTIC MODELS WITH NEURAL NETWORKS

In the current speech recognition system based on neural networks, the common hybrid DNN acoustic model has been gradually replaced by more accurate RNN or CNN. These are the two best options for effectively using variable-length contextual information [11]. This section introduces some applications and improvements of these two structures in low-resource speech recognition.

1) RECURRENT NEURAL NETWORK

RNN has the ability to remember and have strong modeling capabilities in time series data learning, which solves the problem of not modeling the dynamic characteristics of speech in DNN-HMM and has become the most widely used neural network structure in the field of ASR. In fact, if the memory window of the basic RNN is too long, there will be problems with unstable training, gradient disappearance or explosion, and it is difficult to deal with the problem of long-term dependence. Therefore, Long Short-Term Memory (LSTM) structure [12] is now commonly used to replace traditional RNN. LSTM is considered to be a complex and delicate network unit which have memory function so that it can store information for a long time. The LSTM structure which can selectively remember historical information contains three types of gates: input gate, forget gate and output gate. The input gate decides when to let the input into the cell unit, the forget gate decides when to remember the memory of the previous moment, and the output gate decides when to let the memory flow to the next moment. LSTM is calculated at time step t according to the following equations:

$$i_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1} + W_{ic}c_{t-1} + b_i) \quad (2)$$

$$f_t = \sigma(W_{fx}x_t + W_{fh}h_{t-1} + W_{fc}c_{t-1} + b_f) \quad (3)$$

$$c_t = f_t c_{t-1} + i_t \phi(W_{cx}x_t + W_{ch}h_{t-1} + b_c) \quad (4)$$

$$o_t = \sigma(W_{ox}x_t + W_{oh}h_{t-1} + W_{oc}c_t + b_o) \quad (5)$$

$$h_t = o_t \phi(c_t) \tag{6}$$
$$y_t = W_{yh} h_t + b_y \tag{7}$$

where $i_t, f_t, c_t, o_t, h_t, x_t$, and y_t respectively represent the input gate, forget gate, memory unit, output gate and hidden layer, input and output state at time step t. Where W denotes the weight matrix of each part, such as W_{ix} denotes the weight matrix between the input gate and the input layer. Where b denotes the bias matrix. Where σ denotes the sigmoid and ϕ denotes the neuron activation function.

The method based on RNN has shown excellent performance due to its powerful modeling capability and deeper architecture. For example, deep acoustic model with five layers of LSTM proposed by Google has achieved impressive improvements for large vocabulary speech recognition task [13]. It is well known that the depth of neural networks is critical for acoustic modeling. However, the stack of multilayer LSTMs in low-resource scenario makes the model more difficult to train, because the performance tends to saturate and decline with the increase of depth. The improvement of ASR performance by LSTM structure with residual learning was studied in [14], which introduced cross-layer quick connection in multilayer LSTMs instead of simply stacking layers. These shortcut connections represented feature mapping between shallow and high layers. Not only could they ensure the information flow forward across several layers, but error could pass back across several layers without attenuation. Zhou *et al*. [15] further proved the effectiveness of the Shared Hidden Layer (SHL) LSTMs with residual learning for multilingual low-resource speech recognition, which alleviated the degradation problem without adding additional parameters and computational complexity.

Actually, in order to make full use of the subsequent context information, most of speech recognition systems adopt the BLSTM (Bidirectional LSTM) structure, which is composed of two unidirectional LSTMs superimposed each other. Its output is determined jointly by the state of these two LSTMs and can provide the output layer with complete past and future context information. The calculation processes are as follows:

$$\overrightarrow{h_t} = \overrightarrow{LSTM}(x_t, h_{t-1}) \tag{8}$$
$$\overleftarrow{h_t} = \overleftarrow{LSTM}(x_t, h_{t+1}) \tag{9}$$

where $h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}]$. BLSTM processes data in both directions by using two separate parameter sets, forward parameters and backward parameters.

Graves [16] first tried to use BLSTM for acoustic modeling of speech recognition and achieved the best recognition performance at that time on the TIMIT corpus. Subsequently, many researchers have studied the low- resource speech acoustic modeling of BLSTM [17]–[20]. Although the BLSTM has achieved good results, the structure has a large number of parameters and some complex training mechanisms. In order to solve this problem, several layers of Bidirectional Gated Recurrent Unit (BGRU) can be added to the model, which are used in combination with BLSTM and usually does not completely replace BLSTM [21]. Therein GRU [22] can be understood as a simplified version of LSTM, which not only retains the long-term memory function of LSTM, but also replaces the input gate, forget gate, and output gate in LSTM with update gate and reset gate. In addition, GRU combines the two vectors of cellular state and output. Better expressiveness is shown in low-resource scenario on the condition that GRU has fewer parameters. Kang *et al*. [23] proposed local BGRU with residual learning. All time-dependency relationships were considered in a fixed local window.

2) CONVOLUTIONAL NEURAL NETWORK

The spectral characteristics of speech signal can be regarded as an image with two dimensions of time and frequency information. Therefore, the extensive application of CNN [24] in the image field also provides ideas for speech processing. For example, the pronunciation of each person is very different, and the frequency band of the formant is different on the spectrogram. CNN can effectively remove this difference, which is conducive to acoustic modeling in low-resource scenario. Typical CNN is usually divided into two parts: convolutional filter and max-pooling. The convolutional filter captures the local structural characteristics, and each node of its feature map is only convolved with f nodes of the local frequency band of the previous layer. Local convolution has two advantages: (1) the clean spectrum can be used to calculate the characteristics with excellent performance, and only a few of the characteristics will be affected by the noise part, so the robustness of the model is improved; (2) the higher layers of the network combine the calculated values of each frequency band to balance the speech information of adjacent frequency bands. After the convolution operation, max-pooling is conducted to provide additional translation and rotation invariance [25].

In the acoustic model based on CNN, the input feature vectors are divided into N non-overlapping frequency bands $\{v_i | i = 0, \ldots, N-1\}$. Then each s adjacent bands are treated as a band group $\{v_{i+r} | r \in 0, \ldots, s-1\}$. These band groups pass through weights matrix W, bias vector b and nonlinear transformation function θ of the convolution layer to produce the output h_i. The calculation process is as follows:

$$h_i = \theta\left(\sum_{r=0}^{s-1} W_r^T \cdot v_{i+r} + b\right) \tag{10}$$

During the convolution operation, the weight matrix W and the bias vector b shift d frequency bands every time, thereby generating a set of convolutional layer frequency band outputs $\{h_j | j \in 0, \ldots, M-1\}$. The number s of frequency bands input by the convolutional layer is called band width, and the number d of frequency bands shifted by the convolutional layer every time is called band shift. Max-pooling follows the convolution operation to achieve the purpose of dimensionality reduction of hidden nodes. The

formula is as follows:

$$p_i^m = \max_{j \in i \cdot k, \ldots, (i+1) \cdot k-1} h_j^m \tag{11}$$

where p_i^m stands for the output of the max-pooling operation, i and j are the indexes of the bands, m is the index of the neurons in a band and k is the pooling size.

Time Delay Neural Network (TDNN) is the first proposed simple one-dimension CNN applied to speech recognition tasks without pooling or subsampling. It can be used to efficiently build long term time-dependency relationships whether in small or large data scenarios. TDNN has been shown to be effective in learning the temporal dynamics information of signals even from short term feature representations [26]. The TDNN system for multilingual training was further established in [27]–[29]. Moreover, chain model training can significantly improve the speed and Word Error Ratio (WER), which is characterized by Lattice-Free Maximum Mutual Information (LF-MMI) as the training criterion without frame level cross entropy pre-training.

Initially, CNN was only used as a tool for robust feature extraction, so generally only one or two layers of layers were added at the bottom and then the upper layers were modeled using other neural network structures. For example, a CNN layer was added at the lowest of the hybrid neural network and HMM model on a small vocabulary task to normalize the spectral change of speech signal [30]. Two-layer CNNs were used for high-dimensional speech feature extraction for low-resource speech recognition. Compared with MFCC features, their frequency domain energy changes are smaller, which is beneficial to the learning of high-level networks [31]. Then inspired by VGGNet [32], a very deep convolutional network architecture with up to 14 weight layers was applied to low-resource speech recognition in [33]. A deep structure based on Gated Convolutional Network (GCN) was proposed in [34]. GCN is more suitable for the combination of gate mechanism and convolutional operation of sequential tasks. It can better learn the acoustic feature representation, in which gates are used to control the information passed in the hierarchy. It combines the advantages of RNN and CNN on low-resource task to improve training speed and robustness.

However, the problem of gradient disappearance and overfitting in the optimization process of CNN are still two factors that affect the performance of the model. Convolutional Maxout Neural Network (CMNN) which uses maxout neuron and dropout training is an effective way to solve this problem [35], [36]. The nonlinear function of the original network is changed from sigmoid to maxout. It selects the maximum value for the output of neuron nodes at adjacent positions within the same frequency band, making the model easy to optimize. Dropout discards the neurons in the network with a certain probability during each training, reducing the network parameters to be adjusted to prevent overfitting.

Compared to computer vision, the behavior of the two dimensions of time and spectrum of speech signals may be quite different. Therefore, the two-dimension convolution structure was proposed in [37], which emphasized the importance of considering both time and spectrum in a convolutional filter. The experimental result showed that the two-dimension convolution network was superior to the fully connected DNN in only 10 hours of training data. Over time, spectral convolution became much less important.

B. ACOUSTIC MODELS WITH END-TO-END STRUCTURE

Deep learning algorithms still play a limited role in speech recognition systems in the form of traditional pipeline. The end-to-end model integrates multiple modules in traditional speech recognition such as acoustic model, pronunciation lexicon, and language model into a network for joint training [38]. The end-to-end model realizes the direct mapping of the input sound sequence to the label sequence without carefully designing the intermediate state, which greatly simplifies the training process and significantly reduces the calculation complexity. This integration reduces the dependence on prior expert knowledge and avoids the obstacle that the traditional speech recognition framework cannot overcome for the scarcity of effective data, so it is also applied in the low-resource speech recognition. End-to-end learning allows us to process a wide variety of sounds, including noisy environments, accents, and different languages [39].

1) CONNECTIONIST TEMPORAL CLASSIFICATION

In ASR task, the length of feature sequence of input speech frame is usually greater than that of output label sequence, and the corresponding label is required for each speech frame for effective training. Hence, the speech needs to be preprocessed with frame-by-frame alignment marks before training, which requires to be iterated repeatedly. The proposal of CTC perfectly solves this problem [40]. It focuses on whether the output is consistent with the label as a whole. It can be trained without the frame level alignment of the label in time and does not care about the prediction of the input data at any time. The output of CTC is the probability of the overall sequence, thereby reducing the tedious work of forcing alignment to get frame-level annotations. In addition, CTC adds an additional blank label to the target label set and uses the blank label to indicate the probability of not issuing any labels at a specific time step.

Given an input sequence $x = \{x_1, \ldots, x_T\}$ of length T and a target output sequence $y = \{y_1, \ldots, y_N\}$ of length N, we define a set of target labels $\Omega(y_n \in \Omega)$ and an extended set of CTC target output labels $\Omega^* = \Omega \cup \{-\}$. The CTC first maps x to the path $\pi = \{\pi_1, \ldots, \pi_T\}$ with the same length T, $\pi_t \in \Omega$. The conditional probability of any path π is calculated as follows:

$$P(\pi|x) = \prod_{t=1}^{T} P(\pi_t|x) \tag{12}$$

The set of all paths π of length T is recorded as $B^{-1}(y)$. Then the CTC needs to perform many-to-one, long-to-short mapping to aggregate multiple paths into a shorter label

sequence. The same labels that appear continuously in each path are merged into one and the blank labels are removed. The probability of the target label sequence is calculated as follows:

$$P(y|x) = \sum_{\pi \in B^{-1}(y)} P(\pi|x) \tag{13}$$

The loss function of CTC is defined as the sum of negative logarithmic probabilities of the correct label. This means minimizing the following objective function:

$$L_{CTC} = -\ln \prod_{(x,y)\in D} p(\pi|x) = -\sum_{(x,y)\in D} \ln p(\pi|x) \tag{14}$$

where $(x, y) \in D$ denotes training samples.

The essence of CTC is a special loss function or optimization criterion for sequence modeling. Its introduction effectively solves the classification problem of time series data. The combination of deep neural network and CTC makes deep learning more fully applied in speech recognition and becomes a hot spot in end-to-end speech recognition research. The output unit of CTC is very flexible, that can be phonemes, glyphs, syllables and other sub-word units, or even whole words [21].

Rosenberg *et al*. [41] explored the use of CTC in keyword search and speech recognition in low-resource languages, which did not exceed the result obtained by DNN-HMM but was also competitive. The result indicated the direction for the improvement of the end-to-end speech recognition system. The encoder in CTC model has great potential for improvement. A method of using segmentation to correct CTC loss was proposed in the training process, resulted in improvement while decoding with small beam size [21]. Vydana *et al*. [18] studied the Subspace Gaussian Mixture Model (SGMM) and the joint acoustic model based on RNN-CTC. Experimental results showed that the joint acoustic model trained with RNN-CTC performed better than the SGMM system on 120-hour Indian language data. Wang *et al*. [42] combined CTC with Tibetan linguistics knowledge and used bound triphones as a modeling unit to solve the problem of Tibetan acoustic modeling under resource constraints, which made the recognition rate based on the end-to-end acoustic model method exceed the speech recognition system based on BLSTM-HMM. Yu *et al*[20], [31] used the BLSTM-CTC joint model to achieve phonemic level speech recognition for a few hours of data.

However, CTC excludes the case where the output sequence is larger than the input sequence and makes independent assumptions between each time frame, without modeling the interdependence between outputs and ignoring the correlation between frames. RNN-Transducer (RNN-T) combines acoustics and language modeling on the basis of CTC model by adding a prediction network [43]. RNN-T regards the acoustic model as an encoder, the language model as a prediction network, and the joint network as a decoder. This model has been proven to be effective in speech recognition tasks [44], [45]. But RNN-T is more difficult to train, and it is necessary to pre-training the encoder and the prediction network separately to obtain better result. It has not been well applied in low-resource speech recognition.

2) SEQUENCE-TO-SEQUENCE MODELS

Attention-based model is an end-to-end model of encoder-decoder. The attention mechanism eliminates the need for pre-segment alignment of data and can be used with implicitly learn the soft alignment between input and output sequences, avoiding the conditional independence hypothesis problem in CTC [46]. The encoder in attention-based model converts the entire speech input sequence x to the high-level hidden vector sequence $h = \{h_1, \ldots, h_L\}$, and then the decoder uses the attention mechanism to select or assign different weights to the vectors in the hidden vector sequence h in each step of generating output label y, so that the most relevant vector is used for prediction.

Chorowski *et al*. [47] first used attention mechanism for the alignment of input and output sequences in speech recognition task. The encoder was bidirectional RNN. The decoder was the RNN that directly emitted the phoneme stream and sent each symbol based on the context created by a subset of input symbols selected using the attention mechanism. Later in [41] this model was used for low- resource speech recognition, but the encoder and decoder structure were replaced by GRU. This structure focuses the entire encoding sequence, which must wait until the encoding process is completely completed, thus increasing the delay. The introduction of the Listen, attention and Spell (LAS) [48] model solved this problem. The listener is a pyramid BLSTM that encodes the input sequence x into high-level feature sequence h. The speller is an attention-based decoder that generates characters y from h. The training difficulty and efficiency of the model can be optimized by scheduled sampling [49], label smoothing [50], and minimum word error rate training [51]. Moreover, joint training of attention and CTC can greatly improve the convergence of the model [52].

Transformer is a special encoder-decoder structure that avoids recursion and convolution and completely relies on the attention mechanism to describe the global dependency between input and output [53]. There are 6 layers in the encoder part, each of which contains two sublayers of multi-head mechanism and position-wise fully connected feed-forward network. A residual connection is used between two sublayers, and then layer normalization is used. Decoder has the same structure as encoder, with the difference that each layer contains three sublayers including two multi-head attention mechanisms and a fully connected layer. The first multi-head attention uses mask operation, and the second multi-head attention focuses on the encoding information of encoder. The transformer structure requires a fixed input length. If a sequence is shorter than this fixed length, the padding should be applied to the blank part behind. The function of mask operation is to keep the padding part out of the attention calculation and ensure that the predicted position can only depend on the position of less than known

output. In addition, positional encoding is added to input at the bottoms of these encoder and decoder stacks, which contributes to getting some information about the relative or absolute location of tokens in the sequence. The advantage of this change in the internal structure of encoder and decoder is that the model can parallelize training.

Transformer was originally used for Neural Machine Translation (NMT) tasks, both training speed and results far surpass other algorithms in [53]. Transformer was also proved to have good performer for other Natural Language Processing (NLP) tasks [54]. In the premise of the characteristic of the Transformer completing sequence-to-sequence transduction task, people began to try to introduce it into the ASR task.

ASR Transformer was proposed in [55], [56]. Its structure is basically the same as NMT Transformer. Only a linear transformation with layer normalization is added with the purpose of transforming the log-Mel filterbank feature into a dimension that matches the model input. Zhou *et al*. [57] studied using ASR Transformer to complete multilingual speech recognition on 6 low-resource languages. Another speech-Transformer model was put forward by [58], which took a two-dimension spectrogram with time axis and frequency axis as input and added two 3×3 CNN layers of 2 steps and M optional additional modules to the front of the encoder. Considering that the combination of time axis and frequency axis might be helpful in modeling of the temporal and spectral dynamics in a spectrogram, a two-dimension attention mechanism was proposed as an additional module to capture temporal and spectral dependencies. Mohamed *et al.* [59], [60] also combined the convolutional layer with the Transformer, but the positional encoding was eliminated. A two-dimension convolutional block with layer normalization and ReLU and a two-dimension max pooling layer were added at the bottom of encoder of the basic Transformer, also a one-dimension convolutional block with layer normalization and ReLU was added at the bottom of decoder [59]. The addition of convolutional layer in Transformer can better learn the long range acoustic characteristics of speech.

IV. HOW TO LEARN WITH LOW-RESOURCE LANGUAGES

In addition to the improvement of the basic acoustic model units mentioned above, some important technologies are applied in low-resource speech recognition, which is often the key to improving the performance, mainly including two aspects of data and model.

A. DATA AUGMENTATION

The amount of data directly affects the performance of deep learning. Deep learning on small data sets is prone to overfitting. Usually how to solve this problem from the data level is considered in the first place. However, collecting additional resources can be difficult for languages that are not widely used. Data augmentation, a technique designed to increase the amount of data needed to train speech recognition systems, has become a widely adopted approach in the field of low-resource speech recognition. Common data augmentation methods include semi-supervised training, multi-lingual processing, acoustic data perturbation, and speech synthesis [61]. Furthermore, multiple data augmentation methods are used in combination instead of a single method in many studies. For instance, semi-supervised training and acoustic data perturbation were combined in [61], [62]. Data from other languages were additionally used in [62]. Acoustic data perturbation and speech synthesis were combined in [63], resulting in a 14.8% relative WER improvement. Semi-supervised training refers to train the model with supervised data and unsupervised data through the confidence threshold. Multilingual processing refers to expand the resource-poor data with resource-rich data. The data obtained by these two data augmentation methods are natural. The two methods of acoustic data perturbation and speech synthesis can obtain artificial data, that is, interfere with the raw data in some way and generate the new data by speech synthesis technology.

1) ACOUSTIC DATA PERTURBATION

Acoustic data perturbation includes multiple types of data interference to achieve the purpose of expansion, such as speed and volume perturbation, noise injection, increasing reverberation, and so on. This method is easy to implement and has been widely used in low-resource speech recognition tasks, which effectively improves the robustness of the acoustic model. But the main disadvantage of such data is bad quality.

Vocal Tract Length Perturbation (VTLP) augments speech data by random linear distortion along the frequency dimension on the spectrogram. Different from Vocal Tract Length Normalization (VTLN) [64], it generates a random warp factor α for each utterance to warp the frequency axis and map the frequency to a new value instead of setting a warp factor for each training and testing speaker [65]. It laid the foundation for increasing data sets in the field of speech recognition without changing labels. It was later applied to several low-resource speech recognition tasks [61], [66]–[69]. Speed perturbation created two counterparts of the original training data by modifying the speed to 0.9 and 1.1 of the original rates. Tempo perturbation was additionally used to correct the rhythm of signal, in ensuring signal at the premise of pitch and spectrum unchanged. Due to the change of the signal length, the GMM-HMM system was used to realign the data after the speed perturbation [68]. Gokay *et al*. [63] used speed perturbation, volume perturbation and a combination of the two for data augmentation. Kanda *et al*. [70] studied three distortion methods of vocal tract length distortion, speech rate distortion, and frequency-axis random distortion. In a large vocabulary continuous speech recognition task with only 10 hours training sample, the relative WER with DNN-HMM training was reduced by 10.1%. Hsiao *et al*. [71] improved the robustness of speech recognition system by artificially adding noise and reverberation. Hartmann *et al.* [67] used fMLLR transformation of random speakers to enhance

bottleneck features. This approach combined with adding noise and speed perturbation.

Inspired by data augmentation in the image field, Google proposed an augmentation strategy for the log-Mel spectrum to help the expansion of speech recognition data called SpecAugment [72]. Its augmentation strategy includes distortion in the time direction and adding masking blocks to the frequency channel and time step. SpecAugment converts ASR from an over-fitting to an under-fitting problem. However, it can get better performance by using a larger network and longer training time. Wang *et al.* [60] continued to improve in SpecAugment and proposed a semantic mask based on regularization to train the end-to-end speech recognition model. It shielded all features corresponding to the output token during training, such as a word or a word fragment. The motivation is to encourage the model to fill in missing marks based on context information with fewer acoustic features, so that the model has stronger language modeling capabilities and stronger resistance to acoustic distortion.

2) SPEECH SYNTHESIS

Compared with the acoustic data perturbation method, speech synthesis is more flexible. A generation model is often used to obtain new speech data, which is similar to using the Generative Adversarial Networks (GAN) architecture [73] in the image field to generate synthetic new images that simulate the distribution of input data.

Voice Conversion (VC) is a technology that converts non-verbal information of a given voice while retaining language information. Kaneko *et al*. [74] proposed a non-parallel VC method that did not rely on parallel data called CycleGAN-VC. It used a Cycle-consistent Generative Adversarial Network (CycleGAN) with gated CNN and an identity-mapping loss. CycleGAN used both adversarial and cycle-consistent loss to learn forward and inverse mapping [75]. This made it possible to find the best pseudo pair from unpaired data. Gated CNN trained using identity-mapping loss allowed mapping functions to capture order and hierarchy while retaining linguistic information. Because it only learned one-to-one mapping. Kameoka *et al*. [76] proposed to use StarGAN [77] for non-parallel many-to-many VC. This method was an extension to CycleGAN-VC. It introduced a domain classifier to predict which class an input belongs to. Since StarGAN-VC only requires a few minutes of non-parallel and unmarked speech for each speaker, the architecture is very suitable for low-resource VC. Hsu *et al.* [78] proposed a non-parallel Variable Autoencoding Wasserstein Generative Adversarial Network (VAW-GAN) VC framework. The Variational Autoencoder (VAE) [79] simulated the speech characteristics of each speaker and the Wasserstein Generative Adversarial Network (W-GAN) [80] synthesized speech from different speakers. Thai *et al.* [81] used StarGAN-VC and VAW-GAN to perform VC in the Seneca language of 720 minutes. The experimental results showed that data augmentation helped to reduce the WER.

TABLE 1. Comparison of multitask learning, transfer learning, and meta learning.

Method	Source	Target
Multitask Learning	Task 1···N	Task 1···N
Transfer Learning	Task 1	Task 2
Meta Learning	Task 1···N	Task N+1

Another method of Speech synthesis is Text-to-Speech (TTS). Two TTS methods were used as strategies for Speech data augmentation in [63], namely the Google Translate Text to Speech (gTTS) and Deep Convolutional TTS (DCTTS) architectures [82]. About 10 hours of Turkish speech data were synthesized ultimately. In order to solve the problem that synthesized speech can only provide limited speaker diversity for data augmentation in low-resource tasks, Du *et al.* [83] proposed a speaker augmentation method that used VAE speaker representation to train the end-to-end TTS system so that TTS could synthesize sounds from unknown new speakers by sampling from the training potential distribution.

B. TRANSFER KNOWLEDGE FROM MODELS

When deep learning is used for small data, superior performance cannot be achieved generally for a single target task. At this point, the researchers come up with the idea that additional tasks could be learned to improve the performance of network learning. This idea is also widely used in the field of low-resource speech recognition, which provides an effective way to solve the problem of data sparsity. It mainly includes multitask learning, transfer learning, and meta learning. The intuitive comparison of the three is shown in Table 1.

1) MULTITASK LEARNING

Multitask Learning is a machine learning technology that aims to optimize the generalization performance of models by learning multiple related tasks in parallel. Especially for small data sets, multitask learning model can outperform the model which only optimize the performance for one task. If multiple tasks are related and share some internal representations, they can transfer knowledge to each other by learning them together. The general framework for multitask learning consists of three parts: (1) the task-specific input layer which is a feature transformation from domain-specific to domain-general representation; (2) SHL for task-independent feature extraction; (3) task-specific output layer which each task has a separate softmax used for estimating the posterior probability of the language senones information. But in practical applications it is necessary to adjust them according to specific tasks. Multitask learning is usually divided into monolingual multitask ASR and multilingual multitask ASR for low- resource scenario.

For monolingual multitask learning in ASR, it is usually required to find tasks related to the language of the main task

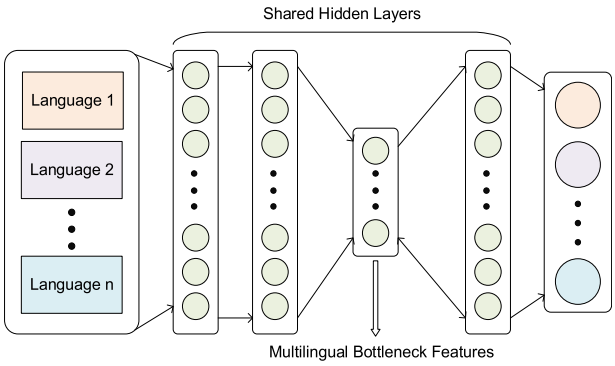


FIGURE 3. Illustration of the model based on multilingual bottleneck features.

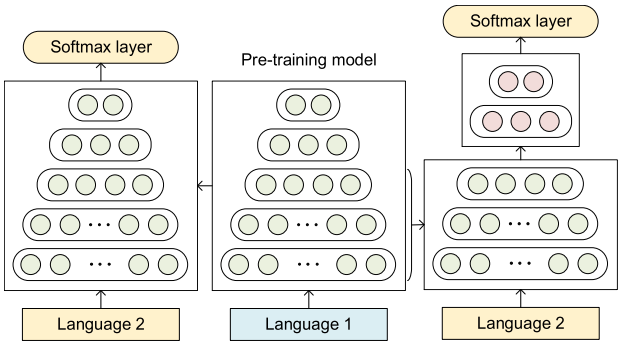


FIGURE 4. Illustration of two transfer learning methods for low- resource speech recognition.

without additional resources. They can be regarded as abstract phonetic categories and these category labels are used as auxiliary tasks for frame-level classification [84], [85]. For instance, triphone modeling and trigrapheme modeling are highly related learning tasks that can be estimated in parallel under the multitask learning framework [86]. Chen *et al*. [87] took grapheme modeling as an additional learning task and used multi-task learning DNN to learn the phone model of the target language. Fantaye *et al*. [88] studied to conduct joint training through multi-task learning for basic acoustic units of a low- resource language Amharic, including syllable, phone and rounded phone.

Multilingual multitask ASR usually embodies each language as a separate task and builds a model by bringing together multiple related languages. Taking multilingual correlation as a prerequisite, the structure that the input layer and the hidden layer are shared by all languages and the output layer is not shared is usually adopted, that is, the multilingual model of SHL model [15], [89]–[92]. These hidden layers are shared among different languages, so that the SHL encodes rich senones information that can be used to identify different languages and also makes the input features more discriminative to different languages after layer by layer abstraction [93]. Multilingual multitask ASR derives a Multilingual Bottleneck Features (MBNF) in order to better help acoustic modeling of low-resource languages in a multilingual environment. Unlike the SHL multilingual model, the MBNF-based model contains a bottleneck layer with only a few nodes. And it is usually a linear layer to retain as much multilingual information as possible [94]–[96]. The MBNF-based model is shown in the Figure 3.

However, some unnecessary language-specific information may be included in MBNF-based model. In order to ensure that the SHL can learn language-invariant features, adversarial training is introduced in multilingual learning. the language discriminator is added to the model to identify the language labels of each frame using shared features [97]–[99].

2) TRANSFER LEARNING

Transfer learning uses the similarities among data, tasks, or models to quickly and effectively develop a system with better performance for a new domain by using the knowledge learned from the source domain. Unlike multitask learning which focuses on improving the performance of all task, transfer learning emphasizes the improvement of the performance of the target task by transferring knowledge acquired on similar but different tasks. In transfer learning, the source domain and the target domain should be similar to each other in order to transfer knowledge smoothly. Transfer learning in ASR task is embodied in cross-linguistic acoustic modeling, aiming to transfer knowledge from one or more source language systems built with large amounts of training data to establish a target language system that provides only a limited amount of transcribed audio [100], [101]. Transfer learning is used in two ways for low-resource speech recognition.

The first method showed as the left of the Figure 4 is fine-tuning, a process of initializing the weight of network with the pre-training network weight instead of the original random initialization, and then readjustment for the task in the target domain. Because the pre-training model is usually not fully applicable to the target domain, fine-tuning is necessary. This method is suitable for tasks with high similarity between the source and target domains. The implementation of fine-tuning method is relatively simple. The input and hidden layers of the network remain common to the two domains, the model parameters are borrowed and then a new softmax output layer is created for the target low-resource language [29]. The goal of the output layer is to obtain the posterior information from the monolingual model of low-resource language.

The second transfer learning method is weight-based transfer showed as the right of the Figure 4, which is realized by transferring the hidden layers. First, an acoustic model is trained using high resource language, retaining the n hidden layers. Then the m randomly initialized hidden layers and a softmax layer are added on top of the n hidden layers. Finally, the transferred model is retrained using target low-resource language. In this case, the pre-training model can be regarded as a feature extractor [31]. In practice, the pre-training model can be used in combination with multilingual training.

3) META LEARNING

Meta learning is also called learning to learn. It acquires experience from previous learning in a systematic and data-driven

manner to speed up the learning process of new tasks. Generally, the following steps are required for Meta Learning: (1) metadata describing the previous learning task and the previous learning model needs to be collected, which is the algorithm configuration used to train the model, including hyperparameter settings, pipeline compositions and network architectures, the resulting model evaluations, such as accuracy and training time, the learned model parameters, such as the trained weights of a neural net, as well as measurable properties of the task itself; (2) we need to learn from the previous metadata to extract and transfer knowledge that guides the best model search for the new task [102]. Different from transfer learning, meta learning makes use of previous knowledge to make the model self-adjust and optimize according to new tasks.

Meta learning explores how to quickly adapt to unseen data. As meta learning has been widely used in the field of computer vision under the few-shot learning setting [103], [104]. Preliminary attempts have been made in language and speech processing in low-resource scenarios and good results have been achieved, such as machine translation [105], [106], dialogue generation [106] and speaker adaptation [107]. Klejch *et al*. [107] studied adapting a speaker-independent model in unseen speaker conditions using limited adaptation data, which could be regarded as a special case of few-shot learning. In this article, an acoustic model weight adaptive method based on meta learning was proposed. Treating speaker adaptation as a function *adapt*, its set of parameters Φ uses adaptive data to adjust a set of weights Θ of the acoustic model $f(x, \Theta)$ into a set of adaptive weights Θ^*. The 3-hour data divided into 18 speakers was used to train the meta-learner, and finally achieved a lower WER in DNN and TDNN models. Hsu *et al.* [91] proposed MetaASR, which was learned from six source tasks through the Model-Agnostic Meta Learning (MAML) algorithm to obtain a good initialization parameter of the shared encoder that performs quick fine-tuning for four target tasks. The results showed that MetaASR performed better than MultiASR in only four target languages directly. Meta learning is the key to the realization of general artificial intelligence. Although it is not widely used in the field of low-resource speech recognition, it is a direction worth further study.

V. PROJECTS FOR LOW-RESOURCE LANGUAGES

A. IARPA BABEL PROGRAM

Many speech transcription systems were originally developed for English and have significantly lower performance in non-English language. There is no existing speech technology for those minority languages. And it often takes years to develop and cover a small portion of the world's languages. The effective triage capabilities to assist those few analysts must be rapidly developed. The goal of The IARPA Babel program [108] is to develop agile and robust speech processing technology that can quickly adapt to any human language, so as to provide effective searching ability for analysts and process a mass of real-word recorded speech. The Babel's program worked with diverse languages from the outset and acquired speech data in-country for languages from a broad set of language families, such as Afro-Asiatic, Niger-Congo, Sino-Tibetan, Austronesian, Dravidian, Altaic. In Babel program, data from more than twenty low-resource languages are collected, which allows us to focus on multilingual experiments for feature extraction and acoustic modeling. The detailed information of 23 languages is shown in Table 2. They are available on Linguistic Data Consortium (LDC) website at present [109]. Audio data is presented as 8kHz 8-bit a-law encoded audio in sphere format and 48kHz 24-bit PCM encoded audio in wav format. Transcripts are encoded in UTF-8 in Latin script. Transcripts are included for approximately 80% of the speech. The gender distribution among speakers is approximately equal. speakers' ages range from 16 years to 70 years. Calls were made using different telephones from a variety of environments including the street, a home or office, a public place, and inside a vehicle. This allows us to deal with real recording conditions from the start.

Since 2013, the National Institute of Standards and Technology (NIST) has held an international keyword search (OpenKWS) evaluation every year. This evaluation is part of the Babel program. The goal of NIST OpenKWS evaluation is to establish a speech recognition system with limited training resources and perform keyword searh tasks within a limited time. In each evaluation, a surprise language is released whose language information is unknown beforehand. The primary measure of performance for NIST OpenKWS is Actual Term-Weight Value (ATWV) [110]. Its calculation formula is as follows:

$$ATWV(\theta) = 1 - \frac{1}{K}\sum_{kw=1}^{K}\left(\frac{N_{Miss}(kw,\theta)}{N_{Ture}(kw)} + \beta\frac{N_{FA}(kw,\theta)}{T - N_{Ture}(kw)}\right) \tag{15}$$

where θ is the uniform threshold used to determine whether each possible keyword was a real keyword. K is the number of different keywords. $N_{Miss}(kw, \theta)$ and $N_{FA}(kw, \theta)$ respectively represent the number of missed detection and false alarms of keyword kw for θ. $N_{True}(kw)$ is the number of reference occurrences of keyword kw. T stands for the size of test speech corpus. β is a constant, which is used to punish false alarms rate of the system. The higher the value of ATWV, the better the performance.

The OpenKWS evaluations for the Babel Program have base period, option period 1, option period 2 and option 3. The performance goals for every period is shown in Table 3.

B. DARPA LORELEI PROGRAM

Humanitarian Assistance Disaster Relief (HADR) crises can occur anywhere in the world. People in need often post real-time information online, and stakeholders involved in crisis management can use this information and analyze them. For example, understanding the comprehensive

TABLE 2. Presentation of language data in IARPA Babel program.

Language	Release ID	Hours	Regions
Turkish	105b-v0.5	213	spoken in seven dialect regions in Turkey.
Tagalog	106-v0.2g	213	North, Central and South dialect regions in Philippines.
Pashto	104b-v0.4bY	244	spoken in four dialect regions of Afghanistan and Pakistan.
Georgian	404b-v1.0a	190	Eastern and Western dialect regions in Georgia.
Cantoness	101b-v0.4c	215	Chinese provinces of Guangdong and Guangxi, and within those provinces, among five dialect groups.
Bengali	103b-v0.4b	215	spoken in India by native speakers of Bengali born in India.
Assamese	102b-v0.5a	205	three dialects spoken in Assam, a state in northeastern India.
Zulu	206b-v0.1e	211	KwaZulu-Natal-urban dialect region of South Africa.
Vietnamese	107b-v0.7	201	North, North-Central, Central and Southern dialect regions in Vietnam.
Tamil	204b-v1.1b	350	Northern, Central, Southern and Western dialect regions of the Indian state of Tamil Nadu.
Swahili	202b-v1.0d	350	Nairobi dialect region of Kenya.
Lao	203b-v3.1a	207	Vientiane dialect region in Laos.
Kurmanji Kurdish	205b-v1.0a	203	southeastern and eastern Anatolian regions of Turkey.
Haitian Creole	201b-v0.2b	203	Northern, Western and Southern dialect regions in Haiti.
Tok Pisin	207b-v1.0e	200	Papuan dialect region of Papua New Guinea.
Telugu	303b-v1.0a	201	Central, East, South and North Telugu dialect regions of India.
Kazakh	302b-v1.0a	203	Northeastern and Southern dialect regions of Kazakhstan.
Cebuano	301b-v2.0b	191	Cebu-North Kana, Sialo, and Mindanao dialect regions of Philippines.
Lithuanian	304b-v1.0b	210	Aukštaitian and Samogitian dialect regions of Lithuania.
Igbo	306b-v2.0c	207	Owerri, Onitsha, and Ngwa dialects spoken in Nigeria.
Guarani	305b-v1.0c	198	Paraguay.
Amharic	307b-v1.0b	204	Addis Ababa, Shewa, and Gondar dialect regions of Ethiopia.
Dholuo	403b-v1.0b	204	South Nyanza and Trans-Yala dialect regions of Kenya.

TABLE 3. The performance goals for the OpenKWS evaluation.

Period	Base	Option 1	Option 2	Option 3
Transcribed (%)	100%	≤75%	≤50%	≤50%
Pronunciation Lexicon	100%	≤75%	≤50%	≤50%
Channels	telephone	telephone and non-telephone	telephone and non-telephone	telephone and non-telephone
Languages Investigated Development + Surprise	4+1	5+1	6+1	7+1
Build Time for Surprise	4 weeks	3 weeks	2 weeks	1 weeks
ATWV	0.3	0.3	0.3	0.3

emotions of the affected population in a particular area may help inform decision makers on how to best allocate resources for effective disaster relief. However, these works can be severely limited by language barriers. The DARPA Low-Resource Languages for Emergent Incidents (LORELEI) program further developed language processing techniques for low-resource languages in the context of this humanitarian crisis. The goal of LORELEI program is to solve the problem that the existing methods cannot achieve the popularization of language technology through the research and development of language technology. It can eliminate the current dependence on huge, manually-translated, manually-transcribed or manually-annotated corpora and use the relevant language resources to fully develop low-resource languages. In this program, rather than translating foreign language materials into English, information elements based on local language materials, including situational descriptions, names, places, events, emotions and relationships, are identified [111].

The exploitation of LORELEI program can be divided into three stages. The first is the language analysis stage. In order to reduce the dependence on specific language information, it is necessary to analyze the common attributes and rules of the language from the known language data to establish a universal language technology model. The known resource-rich language information is mapped to the resource-poor language to establish a projection hypothesis relationship. Then the optimization algorithm for language- specific resource is also studied. The second stage is the language technology development stage. Driven by the knowledge fusion engines, run-time models are built and language processing tools are developed by combining the knowledge of linguistic experts

TABLE 4. The descriptions of the three tasks of LoReHLT 2019.

Task	Language	Input	Objective
MT	IL11, IL12	Text	Translate IL text documents to English.
SF	IL11, IL12, English	Text and Audio	Generate situation frames including sentiment emotion about the frame, link those situation frames into knowledge base level situations.
EDL	IL11, IL12, English	Text	Identify named mentions in both the ILs and in English, classify them into predefined entity types, link the mentions to a knowledge base or cluster them if they are not linkable to the knowledge base.

and scenario information. Third, to make it easier for analysts to use and analyze event-related data, the LORELEI program creates web service that integrates the Incident Language (IL) tools. Analysts can use these tools to convert low-resource ILs into English summary or other visual forms by accessing web services. The web service is constantly updated and improved as the data increases [112].

Unlike the Babel which focuses on speech processing, LORELEI focuses on situational awareness when emergencies occur, with an emphasis on text processing in low-resource languages. LDC is building text language packs for LORELEI, which includes data, annotations, NLP tools, lexicons and grammatical resources for 23 representative languages (Uzbek, Turkish, Hausa, Amharic, Arabic, Farsi, Hungarian, Mandarin, Russian, Somali, Spanish, Vietnamese, Yoruba, Akan, Bengali, Hindi, Indonesian, Swahili, Tagalog, Tamil, Thai, Wolof, Zulu) and 12 ILs (Uzbek, Mandarin, Oromo and other undisclosed languages) [113]. Representative languages packs which contain monolingual text, parallel text, annotation, text processing tools, segmentation, entity tagging, lexicons and gramma are selected to provide broad typological coverage. while ILs are selected to evaluate system performance on a language whose identity is disclosed at the start of the evaluation. There are two tools in the language pack, one to recreate original source data from the processed XML material and the other to condition text data users download from Twitter. Data were collected in discussion forums, news, reference, social network and weblog. All text data is encoded as UTF-8.

Low Resource Human Languages Technologies (LoReHLT) evaluation is designed in collaboration with NIST and LORELEI program. LoReHLT 2019 [114] included three tasks, Machine Translation (MT), Situation Frame (SF), and Entity Detection and Linking (EDL). The descriptions of the three tasks are shown in Table 4.

VI. CONCLUSION

In recent years, with the rise of deep learning, ASR technology has made great progress. Significant results have been achieved in low-resource scenarios. Data augmentation, multilingual and cross-lingual training have become the most widely used. But speech recognition systems still need to be designed with more sophisticated models to deal with speakers with accents or with higher levels of background noise. Low-resource speech recognition may also have the following improvement or breakthrough in the future.

First, the improvement of end-to-end system. The integration of additional language knowledge, learning with complex data such as noise and so on still need to be studied. The joint modeling of acoustic model and language model should be strengthened to better explore the correlation and complementarity between acoustics and language, so as to build a more thorough end-to-end speech recognition system to improve performance.

Second, excavating speech structure knowledge of multimodal information fusion. We can expand from a single mode only for speech to related modes such as images and videos. Because in the era of high popularity of multimedia technology, such data is easy to obtain and relatively easy to label. This is an important research direction for low-resource speech recognition with data scarcity.

REFERENCES

[1] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic speech recognition for under-resourced languages: A survey," *Speech Commun.*, vol. 56, pp. 85–100, Jan. 2014.
[2] *Ethnologue*. Accessed: Apr. 14, 2020. [Online]. Available: https://www.ethnologue.com/guides/continents-most-indigenous-languages
[3] S. Xu, "Research on the application of national languages from an international perspective," *Minority Translators J.*, vol. 1, pp. 11–17, Feb. 2020.
[4] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.
[5] A. Stolcke, L. Ferrer, S. Kajarekar, E. Shriberg, and A. Venkataraman, "MLLR transforms as features in speaker recognition," in *Proc. Interspeech*, Lisbon, Portugal, 2005, pp. 2425–2428.
[6] W. Reichl and W. Chou, "Robust decision tree state tying for continuous speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 5, pp. 555–566, Sep. 2000.
[7] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, Jul. 2006.
[8] A. Mohamed, T. N. Sainath, G. Dahl, B. Ramabhadran, G. E. Hinton, and M. A. Picheny, "Deep belief networks for phone recognition," in *Proc. ICASSP*, Prague, Czech Republic, 2011, pp. 5060–5063.

[9] L. Deng, "An overview of deep-structured learning for information processing," in *Proc. APSIPAASC*, Xi'an, China, 2011, pp. 1–14.

[10] Y. Bengio, "Learning deep architectures for AI," *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, Nov. 2009, doi: 10.1561/2200000006.

[11] D. Yu and J. Li, "Recent progresses in deep learning based acoustic models," *IEEE/CAA J. Automatica Sinica*, vol. 4, no. 3, pp. 396–409, Jul. 2017.

[12] D. Bukhari, Y. Wang, and H. Wang, "Multilingual convolutional, long short-term memory, deep neural networks for low resource speech recognition," *Procedia Comput. Sci.*, vol. 107, pp. 842–847, Apr. 2017.

[13] H. Sak, A. Senior, K. Rao, and F. Beaufays, "Fast and accurate recurrent neural network acoustic models for speech recognition," in *Proc. Interspeech*, Dresden, Germany, 2015, pp. 1468–1472.

[14] Y. Zhao, S. Xu, and B. Xu, "Multidimensional residual learning based on recurrent neural networks for acoustic modeling," in *Proc. Interspeech*, Sep. 2016, pp. 3419–3423.

[15] S. Zhou, Y. Zhao, S. Xu, and B. Xu, "Multilingual recurrent neural networks with residual learning for low-resource speech recognition," in *Proc. Interspeech*, Aug. 2017, pp. 704–708.

[16] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 6645–6649.

[17] M. Karafiát, M. K. Baskar, I. Szöke, V. Malenovský, K. Veselý, F. Grézl, L. Burget, and J. Černocký, "BUT OpenSAT 2017 speech recognition system," in *Proc. Interspeech*, Sep. 2018, pp. 2638–2642.

[18] H. Krishna, K. Gurugubelli, V. V. R. V, and A. K. Vuppala, "An exploration towards joint acoustic modeling for indian languages: IIIT-H submission for low resource speech recognition challenge for indian languages, INTERSPEECH 2018," in *Proc. Interspeech*, Sep. 2018, pp. 3192–3196.

[19] M. Karafidt, M. K. Baskar, K. Vesely, F. Grezl, L. Burget, and J. Cernocky, "Analysis of multilingual blstm acoustic model on low and high resource languages," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5789–5793.

[20] C. Yu, Y. Chen, Q. Sun, C. Liu, S. Xu, and W. Yin, "Research of endangered languages speech recognition based on dynamic BLSTM and CTC," *Appl. Res. Comput.*, vol. 36, no. 11, pp. 3334–3337, Nov. 2019.

[21] V. Bataev, M. Korenevsky, I. Medennikov, and A. Zatvornitskiy, "Exploring end-to-end techniques for low-resource speech recognition," in *Proc. SPECOM*, Leipzig, Germany, 2018, pp. 32–41.

[22] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014, *arXiv:1412.3555*. [Online]. Available: http://arxiv.org/abs/1412.3555

[23] J. Kang, W.-Q. Zhang, W.-W. Liu, J. Liu, and M. T. Johnson, "Advanced recurrent network-based hybrid acoustic models for low resource speech recognition," *EURASIP J. Audio, Speech, Music Process.*, vol. 2018, no. 1, p. 6, Dec. 2018, doi: 10.1186/s13636-018-0128-6.

[24] Y. Lecun and Y. Bengio, "Convolutional networks for images, speech, and time series," *Handbook Brain Theory Neural Netw.*, vol. 3361, no. 10, p. 1995, Nov. 1997.

[25] C. Qin and L. Zhang, "Acoustic modeling approach of multi-stream feature incorporated convolutional neural network for low-resource speech recognition," *J. Comput. Appl.*, vol. 36, no. 9, pp. 2609–2615, Sep. 2016.

[26] V. Peddinti, V. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proc. Interspeech*, Dresden, Germany, 2015, pp. 3214–3218.

[27] N. Fathima, T. Patel, M. C, and A. Iyengar, "TDNN-based multilingual speech recognition system for low resource indian languages," in *Proc. Interspeech*, Sep. 2018, pp. 3197–3201.

[28] F. Keith, W. Hartmann, M.-H. Siu, J. Ma, and O. Kimball, "Optimizing multilingual knowledge transfer for time-delay neural networks with low-rank factorization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 4924–4928.

[29] D. Bagchi and W. Hartmann, "Learning from the best: A teacher-student multilingual framework for low-resource languages," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6051–6055.

[30] O. Abdel-Hamid, A.-R. Mohamed, H. Jiang, and G. Penn, "Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2012, pp. 4277–4280.

[31] C. Yu, Y. Chen, Y. Li, M. Kang, S. Xu, and X. Liu, "Cross-language End-to-End speech recognition research based on transfer learning for the low-resource tujia language," *Symmetry*, vol. 11, no. 2, p. 179, Feb. 2019.

[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, San Diego, CA, USA, 2015, pp. 1–14.

[33] T. N. Sainath, A.-R. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for LVCSR," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 8614–8618.

[34] J. Kang, W.-Q. Zhang, and J. Liu, "Gated convolutional networks based hybrid acoustic models for low resource speech recognition," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Dec. 2017, pp. 157–164.

[35] M. Cai, Y. Shi, J. Kang, J. Liu, and T. Su, "Convolutional maxout neural networks for low-resource speech recognition," in *Proc. 9th Int. Symp. Chin. Spoken Lang. Process.*, Sep. 2014, pp. 133–137.

[36] J. Sun, J. Wushour, and T. Reyiman, "Research on CMN-based recognition of Kirgiz with less resources," *Modern Electron. Technique*, vol. 41, no. 24, pp. 132–136, Dec. 2018.

[37] W. Chan and I. Lane, "Deep convolutional neural networks for acoustic modeling in low resource languages," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 2056–2060.

[38] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng, "Deep speech: Scaling up end-to-end speech recognition," 2014, *arXiv:1412.5567*. [Online]. Available: http://arxiv.org/abs/1412.5567

[39] D. Amodei *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," 2015, *arXiv:1512.02595*. [Online]. Available: http://arxiv.org/abs/1512.02595

[40] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. 23rd Int. Conf. Mach. Learn. ICML*, 2006, pp. 369–376.

[41] A. Rosenberg, K. Audhkhasi, A. Sethy, B. Ramabhadran, and M. Picheny, "End-to-end speech recognition and keyword search on low-resource languages," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 5280–5284.

[42] Q. Wang, W. Guo, and C. Xie, "Towards end to end speech recognition system for tibetan," *Pattern Recognit. Artif. Intell.*, vol. 30, no. 4, pp. 359–364, Apr. 2017.

[43] A. Graves, "Sequence transduction with recurrent neural networks," 2012, *arXiv:1211.3711*. [Online]. Available: http://arxiv.org/abs/1211.3711

[44] E. Battenberg, J. Chen, R. Child, A. Coates, Y. G. Y. Li, H. Liu, S. Satheesh, A. Sriram, and Z. Zhu, "Exploring neural transducers for end-to-end speech recognition," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Dec. 2017, pp. 206–213.

[45] K. Rao, H. Sak, and R. Prabhavalkar, "Exploring architectures, data and units for streaming end-to-end speech recognition with RNN-transducer," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Dec. 2017, pp. 193–199.

[46] D. Wang, X. Wang, and S. Lv, "An overview of End-to-End automatic speech recognition," *Symmetry*, vol. 11, no. 8, p. 1018, Aug. 2019.

[47] J. Chorowski, D. Bahdanau, K. Cho, and Y. Bengio, "End-to-end continuous speech recognition using attention-based recurrent NN: First results," 2014, *arXiv:1412.1602*. [Online]. Available: http://arxiv.org/abs/1412.1602

[48] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 4960–4964.

[49] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer, "Scheduled sampling for sequence prediction with recurrent neural networks," in *Proc. NIPS*, Montreal, QC, Canada, 2015, pp. 1171–1179.

[50] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[51] R. Prabhavalkar, T. N. Sainath, Y. Wu, P. Nguyen, Z. Chen, C.-C. Chiu, and A. Kannan, "Minimum word error rate training for attention-based Sequence-to-Sequence models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 4839–4843.

[52] S. Kim, T. Hori, and S. Watanabe, "Joint CTC-attention based end-to-end speech recognition using multi-task learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 4835–4839.

[53] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS*, Long Beach, CA, USA, 2017, pp. 5999–6009.

[54] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*. [Online]. Available: http://arxiv.org/abs/1810.04805
[55] S. Zhou, L. Dong, S. Xu, and B. Xu, "Syllable-based sequence-to-sequence speech recognition with the transformer in mandarin Chinese," in *Proc. Interspeech*, Sep. 2018, pp. 791–795.
[56] S. Zhou, L. Dong, S. Xu, and B. Xu, "A comparison of modeling units in sequence-to-sequence speech recognition with the transformer on Mandarin Chinese," in *Proc. ICONIP*, Siem Reap, Cambodia, 2018, pp. 210-220.
[57] S. Zhou, S. Xu, and B. Xu, "Multilingual end-to-end speech recognition with a single transformer on low-resource languages," 2018, *arXiv:1806.05059*. [Online]. Available: http://arxiv.org/abs/1806.05059
[58] L. Dong, S. Xu, and B. Xu, "Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5884–5888.
[59] A. Mohamed, D. Okhonko, and L. Zettlemoyer, "Transformers with convolutional context for ASR," 2019, *arXiv:1904.11660*. [Online]. Available: http://arxiv.org/abs/1904.11660
[60] C. Wang, Y. Wu, Y. Du, J. Li, S. Liu, L. Lu, S. Ren, G. Ye, S. Zhao, and M. Zhou, "Semantic mask for transformer based end-to-end speech recognition," 2019, *arXiv:1912.03010*. [Online]. Available: http://arxiv.org/abs/1912.03010
[61] A. Ragni, K. M. Knill, S. P. Rath, and M. J. F. Gales, "Data augmentation for low resource languages," in *Proc. Interspeech*, Singapore, 2014, pp. 810–814.
[62] M. J. F. Gales, K. M. Knill, A. Ragni, and S. P. Rath, "Speech recognition and keyword spotting for low-resource languages: Babel project research at CUED," in *Proc. SLTU*, St. Petersburg, Russia, 2014, pp. 16–23.
[63] R. Gokay and H. Yalcin, "Improving low resource turkish speech recognition with data augmentation and TTS," in *Proc. 16th Int. Multi-Conf. Syst., Signals Devices (SSD)*, Mar. 2019, pp. 357–360.
[64] D. Qi, X. Wang, and W. Bingxi, "A frequency warping approach for vocal tract length normalization," in *Proc. 7th Int. Conf. Signal Process. ICSP*, Sep. 2004, pp. 691–694.
[65] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (VTLP) improves speech recognition," in *Proc. ICML*, Atlanta, Georgia, 2013, pp. 1–5.
[66] Z. Tüske, P. Golik, D. Nolden, R. Schlüter, and H. Ney, "Data augmentation, feature combination, and multilingual neural networks to improve ASR and KWS performance for low-resource languages," in *Proc. Interspeech*, Singapore, 2014, pp. 1420–1424.
[67] W. Hartmann, T. Ng, R. Hsiao, S. Tsakalidis, and R. Schwartz, "Two-stage data augmentation for low-resourced speech recognition," in *Proc. Interspeech*, Sep. 2016, pp. 2378–2382.
[68] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proc. Interspeech*, Dresden, Germany, 2015, pp. 3586–3589.
[69] X. Cui, V. Goel, and B. Kingsbury, "Data augmentation for deep neural network acoustic modeling," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 9, pp. 1469–1477, Sep. 2015.
[70] N. Kanda, R. Takeda, and Y. Obuchi, "Elastic spectral distortion for low resource speech recognition with deep neural networks," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand.*, Dec. 2013, pp. 309–314.
[71] R. Hsiao, J. Ma, W. Hartmann, M. Karafiat, F. Grezl, L. Burget, I. Szoke, J. H. Cernocky, S. Watanabe, Z. Chen, S. H. Mallidi, H. Hermansky, S. Tsakalidis, and R. Schwartz, "Robust speech recognition in unknown reverberant and noisy conditions," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand. (ASRU)*, Dec. 2015, pp. 533–538.
[72] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech*, Sep. 2019, pp. 2613–2617.
[73] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014, *arXiv:1406.2661*. [Online]. Available: http://arxiv.org/abs/1406.2661
[74] T. Kaneko and H. Kameoka, "Parallel-data-free voice conversion using cycle-consistent adversarial networks," 2017, *arXiv:1711.11293*. [Online]. Available: http://arxiv.org/abs/1711.11293
[75] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," 2017, *arXiv:1703.10593*. [Online]. Available: http://arxiv.org/abs/1703.10593
[76] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "StarGAN-VC: Non-parallel many-to-many voice conversion using star generative adversarial networks," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Dec. 2018, pp. 266–273.
[77] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain Image-to-Image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8789–8797.
[78] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks," in *Proc. Interspeech*, Aug. 2017, pp. 3364–3368.
[79] D. P Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*. [Online]. Available: http://arxiv.org/abs/1312.6114
[80] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," 2017, *arXiv:1701.07875*. [Online]. Available: http://arxiv.org/abs/1701.07875
[81] B. Thai, R. Jimerson, D. Arcoraci, E. Prud'hommeaux, and R. Ptucha, "Synthetic data augmentation for improving low-resource ASR," in *Proc. IEEE Western New York Image Signal Process. Workshop (WNYISPW)*, Oct. 2019, pp. 1–9.
[82] H. Tachibana, K. Uenoyama, and S. Aihara, "Efficiently trainable Text-to-Speech system based on deep convolutional networks with guided attention," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 4784–4788.
[83] C. Du and K. Yu, "Speaker augmentation for low resource speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 7719–7723.
[84] M. L. Seltzer and J. Droppo, "Multi-task learning in deep neural networks for improved phoneme recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 6965–6969.
[85] P. Swietojanski, P. Bell, and S. Renals, "Structured output layer with auxiliary targets for context-dependent acoustic modelling," in *Proc. Interspeech*, Dresden, Germany, 2015, pp. 3605–3609.
[86] D. Chen, B. Mak, C.-C. Leung, and S. Sivadas, "Joint acoustic modeling of triphones and trigraphemes by multi-task learning deep neural networks for low-resource speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 5592–5596.
[87] D. Chen and B. Kan-Wing Mak, "Multitask learning of deep neural networks for low-resource speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 7, pp. 1172–1183, Jul. 2015.
[88] T. G. Fantaye, J. Yu, and T. T. Hailu, "Investigation of various hybrid acoustic modeling units via a multitask learning and deep neural network technique for LVCSR of the low-resource language, amharic," *IEEE Access*, vol. 7, pp. 105593–105608, 2019.
[89] S. Tejaswi and S. Umesh, "Addressing data sparsity in DNN acoustic modeling," in *Proc. 23rd Nat. Conf. Commun. (NCC)*, Mar. 2017, pp. 1–5.
[90] A. Madhavaraj and A. G. Ramakrishnan, "Data-pooling and multi-task learning for enhanced performance of speech recognition systems in multiple low resourced languages," in *Proc. Nat. Conf. Commun. (NCC)*, Feb. 2019, pp. 1–5.
[91] J.-Y. Hsu, Y.-J. Chen, and H.-Y. Lee, "Meta learning for end-To-end low-resource speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 7844–7848.
[92] R. Sahraeian and D. Van Compernolle, "A study of rank-constrained multilingual DNNS for low-resource ASR," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 5420–5424.
[93] A. Zhang, "Research on low-resource mongolian speech recognition based on multilingual speech data selection," *Comput. Sci.*, vol. 45, no. 9, pp. 308–313, Sep. 2018.
[94] C. Ni, C.-C. Leung, L. Wang, N. F. Chen, and B. Ma, "Efficient methods to train multilingual bottleneck feature extractors for low resource keyword search," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 5650–5654.
[95] Y. Yuan, C.-C. Leung, L. Xie, H. Chen, B. Ma, and H. Li, "Pairwise learning using multi-lingual bottleneck features for low-resource query-by-example spoken term detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 5645–5649.

[96] R. Menon, H. Kamper, E. V. D. Westhuizen, J. Quinn, and T. Niesler, "Feature exploration for almost zero-resource ASR-free keyword spotting using a multilingual bottleneck extractor and correspondence autoencoders," in *Proc. Interspeech*, Sep. 2019, pp. 3475–3479.

[97] J. Yi, J. Tao, Z. Wen, and Y. Bai, "Adversarial multilingual training for low-resource speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 4899–4903.

[98] J. Yi, J. Tao, Z. Wen, and Y. Bai, "Language-adversarial transfer learning for low-resource speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 3, pp. 621–630, Mar. 2019.

[99] J. Yi, J. Tao, and Y. Bai, "Language-invariant bottleneck features from adversarial End-to-end acoustic models for low resource speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6071–6075.

[100] H. Xu, V. Do, H. X. Xiao, and E. S. Chng, "A comparative study of BNF and DNN multilingual training on cross-lingual low-resource speech recognition," in *Proc. Interspeech*, Dresden, Germany, 2015, pp. 2132–2136.

[101] S. Dalmia, X. Li, F. Metze, and A. W. Black, "Domain robust feature extraction for rapid low resource ASR development," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Dec. 2018, pp. 258–265.

[102] J. Vanschoren, "Meta-learning: A survey," 2018, *arXiv:1810.03548*. [Online]. Available: http://arxiv.org/abs/1810.03548

[103] J. Snell, K. Swersky, and R. S. Zemel, "Prototypical networks for few-shot learning," 2017, *arXiv:1703.05175*. [Online]. Available: http://arxiv.org/abs/1703.05175

[104] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," 2016, *arXiv:1606.04080*. [Online]. Available: http://arxiv.org/abs/1606.04080

[105] J. Gu, Y. Wang, Y. Chen, V. O. K. Li, and K. Cho, "Meta-learning for low-resource neural machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 3622–3631.

[106] F. Mi, M. Huang, J. Zhang, and B. Faltings, "Meta-learning for low-resource natural language generation in task-oriented dialogue systems," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 3131–3157.

[107] O. Klejch, J. Fainberg, and P. Bell, "Learning to adapt: A meta-learning approach for speaker adaptation," in *Proc. Interspeech*, Sep. 2018, pp. 867–871.

[108] *IARPA Babel*. Accessed: May 23, 2020. [Online]. Available: https://www.iarpa.gov/index.php/research-programs/babel

[109] *Linguistic Data Consortium*. Accessed: May 23, 2020. [Online]. Available: https://www.ldc.upenn.edu

[110] *National Institute of Standards and Technology*. Accessed: Jun. 10, 2020. [Online]. Available: https://www.nist.gov/system/files/documents/itl/iad/mig/KWS16-evalplan-v04.pdf

[111] V. R. Martinez, A. Ramakrishna, M.-C. Chiu, K. Singla, and S. Narayanan, "A system for the 2019 sentiment, emotion and cognitive state task of DARPA's LORELEI project," in *Proc. 8th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Sep. 2019, pp. 448–453.

[112] C. Christianson, J. Duncan, and B. Onyshkevych, "Overview of the DARPA LORELEI program," *Mach. Transl.*, vol. 32, nos. 1–2, pp. 3–9, Jun. 2018.

[113] S. Strassel and J. Tracey, "LORELEI language packs: Data, tools, and resources for technology development in low resource languages," in *Proc. LREC*, Portorož, Slovenia, 2016, pp. 3273–3280.

[114] *National Institute of Standards and Technology*. Accessed: Jun. 10, 2020. [Online]. Available: https://www.nist.gov/system/files/documents/2019/06/19/nist_lorehlt_2019_evaluation_plan_v1.0.pdf

CHONGCHONG YU received the Ph.D. degree in computer science from the University of Science and Technology Beijing, Beijing, China.

She is currently the Dean of the School of Artificial Intelligence, Beijing Technology and Business University. Her research interests include artificial intelligence, machine learning, and pattern recognition.

MENG KANG received the B.S. degree in information engineering from Beijing Technology and Business University, Beijing, China, in 2017, where she is currently pursuing the M.S. degree in control theory and control engineering.

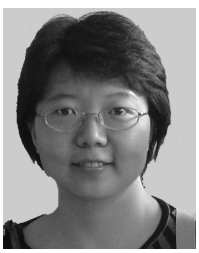
Her research interests include deep learning and speech recognition.

YUNBING CHEN received the M.S. degree in control theory and control engineering from Beijing Technology and Business University, Beijing, China.

He is currently with Putian Information Technology Company, Ltd. His research interests include deep learning and speech synthesis.

JIAJIA WU received the B.S. degree in information engineering from Beijing Technology and Business University, Beijing, China, where she is currently pursuing the M.S. degree in detection technology and automatic device.

Her research interests include speech recognition and signal processing.



XIA ZHAO received the Ph.D. degree in computer software and theory from Peking University, Beijing, China. She is currently an Associate Professor of computer science with Beijing Technology and Business University.

Her research interests include big data processing and intelligent terminal software design.