Towards multilingual end-to-end speech recognition for air traffic control

Yi Lin | Bo Yang | Dongyue Guo | Peng Fan

College of Computer Science, Sichuan University, Chengdu, Sichuan, China

Correspondence
Bo Yang, No.24 South Section 1, Yihuan Road, Chengdu, 610065 China.
Email: boyang@scu.edu.cn

Funding information
National Natural Science Foundation of China, Grant/Award Numbers: 62001315, U20A20161; Sichuan Provincial S&T Projects, Grant/Award Numbers: 2020YFG0327, 2020YFG0306

Abstract
In this work, an end-to-end framework is proposed to achieve multilingual automatic speech recognition (ASR) in air traffic control (ATC) systems. Considering the standard ATC procedure, a recurrent neural network (RNN) based framework is selected to mine the temporal dependencies among speech frames. Facing the distributed feature space caused by the radio transmission, a hybrid feature embedding block is designed to extract high-level representations, in which multiple convolutional neural networks are designed to accommodate different frequency and temporal resolutions. The residual mechanism is performed on the RNN layers to improve the trainability and the convergence. To integrate the multilingual ASR into a single model and relieve the class imbalance, a special vocabulary is designed to unify the pronunciation of the vocabulary in Chinese and English, i.e., pronunciation-oriented vocabulary. The proposed model is optimized by the connectionist temporal classification loss and is validated on a real-world speech corpus (ATC-Speech). A character error rate of 4.4% and 5.9% is achieved for Chinese and English speech, respectively, which outperforms other popular approaches. Most importantly, the proposed approach achieves the multilingual ASR task in an end-to-end manner with considerable high performance.

1 | INTRODUCTION

Air traffic control (ATC) is an essential service provided by ground-based air traffic controllers (ATCOs) to guide the flight to be operated in a safe manner (i.e. prevent conflict), and further to organize and expedite the traffic flow. As the primary communication way between the ATCO and the aircrew, the spoken instruction through the very high frequency (VHF) radio transmission implies a wealth of contextualized situational information, which is important to the real-time ATC decision-making. In the current ATC management system, the ATC is a non-automatic procedure (human-in-the-loop) and is always regarded as a potential risk for the air traffic operation [1]. Numerous studies have demonstrated that monitoring the control conversation is a promising way to obtain real-time traffic dynamics [2, 3, 4], which benefits to formulate a closed-loop ATC management. To this end, the automatic speech recognition (ASR) technique, with the purpose of building the bridge between the human (ATCO and pilot) and machine (ATC systems), has attracted significant attention worldwide in the ATC domain. Recently, several existing air traffic issues, such as monitoring the flight safety [1, 5], reducing the controller’s workload [6, 7], and robot pilot for training simulator [8, 9], have been studied by understanding the spoken instructions based on the speech recognition technique.

For common applications, the ASR is a well-studied research topic that hence generates many promising outcomes. However, when it comes to the ATC domain, the ASR research is facing the following technical difficulties due to the domain-specific characteristics.

- Speech quality: On the one hand, the ATC speech is transmitted through radio communication, which is always an obstacle to collecting high-quality speeches. On the other hand, due to the limited communication resources, multiple pilots in a control sector usually communicate with the ATCO by...
sharing the same frequency. That is to say, the transmission condition, the equipment and the system error at a single frequency are changing as the speaker is switched, which further causes a time-varied noise model for the ATC speech. In summary, compared to the common ASR applications, the speech features are broadly distributed in a latent space, which is a key challenge for improving the final recognition accuracy.

- **Multilingual recognition**: As is well-known, English is the universal language for ATC all over the world. In practice, ATCOs are accustomed to communicating with domestic flights in local languages, while English is for international flights. For example, Chinese is used for the ATC commutation of domestic flights in China mainland. Furthermore, the ATC resources, such as runway number, waypoint name, are named with English letters. Consequently, the ATC speech is spoken in multiple languages in a single utterance, which is the most distinguishing characteristic from that of in common ASR applications. In any case, multilingual recognition is an inevitable problem and thus is in need of addressing in the ATC domain.

- **Imbalanced vocabulary**: In general, the International Civil Aviation Organization (ICAO) published the standard ATC rules, concerning the procedure, terminology, pronunciation, etc. This leads to the fact that the vocabulary of the ATC speech is a special sub-set of the common vocabulary in daily life. In practice, since speakers may not strictly comply with the rules, there are still some out-of-vocabulary (OOV) words that appear at a low frequency in the vocabulary, such as the modal particle. What is similar is the vocabulary for the location-dependent waypoint name. Therefore, the frequency of different words in the ATC corpus is imbalanced, which is supposed to reduce the classification accuracy between speech frames and text labels.

Although enormous efforts have been made to achieve the ASR task in the ATC domain, such as the independent system for a single language [1, 10, 11], the cascaded pipeline for multilingual recognition [2, 12], it is also believed that an end-to-end multilingual framework with the ability to address the above technical problems is the ultimate solution for the ASR research in the ATC domain. To this end, an improved end-to-end ASR model is proposed to address the multilingual ASR task in this work, which is able to transcribe the speech signals into human-readable texts in Chinese characters and English words.

In this work, an end-to-end ASR model is formulated by combining the convolutional neural networks (CNN), recurrent neural networks (RNN) with the Connectionist Temporal Classification (CTC) loss function [13], which is capable of mapping the variable-length speech features to the variable-length text labels automatically. Facing the sparse data distribution of the ATC speech, a hybrid feature embedding (HFE) block is designed to extract discriminative representations from different raw speech features. Following the HFE block, the improved RNN block, i.e., long short-term memory (LSTM), is designed to learn the temporal correlations, wherein the residual mechanism is performed to relieve the training deficiency and further improve the final performance. In general, the HFE and LSTM blocks serve as the feature extractor on the raw inputs to support the final recognition task. A prediction layer is finally appended to classify the extracted high-level features into the vocabulary label, which represents the label probability conditional on the input features. To achieve a multilingual end-to-end recognition paradigm, a dedicated vocabulary, called pronunciation-oriented vocabulary (PoV), is designed for the acoustic model, whose modelling units are the sub-words in multiple languages. A real-world ATC speech corpus is applied to validate the proposed approach, in which different experiments are conducted to prove the efficiency and effectiveness of certain technical improvements. Experimental results show that the proposed end-to-end multilingual ASR model outperforms the baselines, achieving a 4.4% and 5.9% character error rate for the Chinese and English speech, respectively, with language model (LM) decoding. In addition, all the proposed techniques are confirmed to be helpful to improve the final performance. All in all, this work contributes to ASR research in the ATC domain in the following ways:

- An end-to-end framework is proposed to transcribe the ATC speech into human-readable text, without any lexicon, which is able to integrate the multilingual speech recognition into a single model. Considering the structured ATC speech, the CNN and LSTM combined neural network model is applied to achieve the end-to-end ASR task.

- A hybrid feature embedding block is designed to extract discriminative features from raw waves to support the subsequent acoustic modelling. The issue of the sparse distribution of the speech features is addressed by designing multiple convolutional kernels from hybrid input features.

- To overcome the training deficiency of the LSTM layer, the residual mechanism is applied to improve the trainability, which allows the model to obtain the desired prediction accuracy with fewer training epochs.

- A highly efficient vocabulary is built to achieve the multilingual ASR task in the ATC domain, i.e., PoV, which focuses on unifying the pronunciation in multiple languages. In addition, the data imbalance can also be relieved by splitting the words into PoV units, which benefits to improve the training efficiency and advance the final performance.

The rest of this paper is organized as follows. Existing related works are briefly reviewed in Section 2. In Section 3, the implementation details of the proposed approach are subsequently introduced. The experimental configurations are listed in Section 4, where the experimental results are also reported and discussed. Finally, the paper is concluded in Section 5.

## 2 RELATED WORKS

Automatic speech recognition is a hot research topic related to interdisciplinary knowledge, including computer science,
linguistics, signal processing, pattern recognition, etc. [14]. The ASR technique allows the machine to automatically transcribe the speech into human-readable texts, which can be traced back to the 1950s [15]. The ASR performance was significantly improved by applying the hidden Markov model (HMM) to build the state transitions of the acoustic unit [16], where the Gaussian mixture model (GMM) was introduced to capture the data distribution between the speech and label, i.e., the HMM/GMM framework. Lately, the introduction of the n-gram language model [17] made great contributions to improve the ASR performance from the perspective of linguistics and scenario-related semantics. With the development of the deep learning techniques [18, 19, 20], the deep neural network (DNN) was applied to solve the tasks in the ASR research. The DNN was proposed to build the data distribution (as GMM does) [14], which formulated the HMM/DNN framework. Considering the spatial characteristics of the speech feature, the CNN layer was naturally applied to the ASR model [21, 22], in which the convolution operation is able to overcome the diversity of the speech signal. Similarly, the RNN block and its variations were studied to build the temporal dependencies among the speech frames [23–26], which captures the core patterns of the speech signal in the ASR task. As with the HMM-based methods, the alignment between the speech frames and the text labels is an indispensable step to build the data distribution, which aggravates the cost of the sample annotation. In consideration of this technical difficulty, Graves et al. proposed the CTC loss function [13] to align the variable-length speech sequence with the variable-length text sequence by inserting blanks, which is known as the end-to-end ASR framework. The end-to-end ASR framework not only considerably advanced the final performance but also simplified the system architecture. Enormous works have been achieved based on this framework, such as deep speech 2 (DS2) [27], Jasper [28], CLDNN [21], etc. Following the end-to-end idea, more neural network-based architectures were also explored to complete the ASR task, such as LAS [29], transformer [30]. The binary neural network was also proposed to achieve the sequential classification task for the ASR model. The proposed framework is sketched in Figure 1, from which we can see that the whole model comprises four parts: the feature engineering, the hybrid feature embedding block, the residual bidirectional long-short term memory (RBLSTM) layers, and a dedicated vocabulary based on the specificities of the ATC speech. The NV and NV are the number of the CNN and LSTM layers, respectively. Basically, the HFE and RBLSTM blocks are the feature extractors, while a dense layer serves as the classifier to achieve the sequential classification task for the ASR model.

In general, the ATC communication procedure is subject to predefined rules, and its vocabulary contains domain-dependent terms, such as airline names, waypoint names, etc. For instance, an airline’s name is usually followed by digits or letters to formulate a unique flight identification. Therefore, the RNN-based ASR model is a highly preferred option for mining the strong correlations between vocabulary words [9, 38]. Considering the representations of the speech signal, three types of speech features are selected to achieve the proposed HFE block, in which the CNN layer is applied to mine the high-level spatial dependencies and compress the data dimension. The residual and the bidirectional mechanism are performed on the LSTM layers to improve the trainability and the modelling ability of the temporal sequence. By referring to [39], the English word is split into sub-words to ensure that the pronunciation of their sub-word is compatible with that of the Chinese character. In this way, the class imbalance problem can also be addressed to some extent, and further to improve the training convergence.

After determining the modelling unit, a prediction module is designed to classify the high-level representations into the vocabulary by a time-distributed fully connected layer. Generally speaking, both the HFE and the residual LSTMs can be regarded as the feature extractor to support the final sequential classification task. The training error between the predicted labels and the truth labels is evaluated by the CTC loss function, which is further back-propagated to previous layers to upgrade their trainable parameters [40]. In this way, an end-to-end multilingual ASR model is formulated by cascading the HFE, RBLSTMs, prediction layer, and the CTC evaluation.

3.1 Hybrid feature embedding

Feature engineering is an elementary component of the machine learning approaches. By analyzing the unique characters of the raw data, the feature engineering attempts to
extract task-oriented principal patterns (with high distinguished features), which is expected to provide discriminative parametric representations to support the following learning tasks, such as classification, regression. As is well known, the waveform is a type of one-dimensional (1D) temporal signal, which has only minor discriminations among different frames and utterances from the perspective of the data feature. A common solution is to transfer the temporal domain signal into a frequency domain by the signal processing techniques, from 1D to 2D (temporal and frequency) [41]. Based on the activations of human hearing, various algorithms have been proposed to represent the speech features in the frequency domain, such as the spectrogram, log filter-bank (log-fbank), spectrum, and mel frequency cepstrum coefficient (MFCC). As shown in Figure 2, the left is the raw waveform, while the right parts (from top to down) are the spectrogram, log-fbank, and MFCC features, respectively. For the speech features, the horizontal and vertical axes denote the temporal and frequency dimensions, respectively. It can be seen that the extracted speech features obtained by different methods present distinguished patterns that focus on different underlying causes or patterns, with separate features or directions in feature space corresponding to different causes. Different speech features also allow the model to learn specific high-level representations, and further support the speech recognition task.

In general, the feature engineering of the ASR research is a corpus-specific option, and there is no analytic advantage and disadvantage to be referred to. In existing works, different types of speech features were selected to achieve the ASR task according to the intrinsic nature of the training samples, the capacity of the approach or model, etc. In this work, facing the inferior speech quality and sparse feature distribution, a hybrid speech feature is designed to serve as the input of the proposed multilingual end-to-end ASR model, which learns the underlying patterns from different types of feature engineering. A hybrid feature embedding block, i.e., HFE, is designed to further mine decisive patterns from the heterogeneous inputs. The HFE block is based on the CNN layers that are also able to build spatial correlations and compress the data dimension.

The architecture of the proposed HFE block is illustrated in Figure 3. The three Conv1D layers, called feature embedding, are designed to map different speech features into the same dimension, i.e., 256 in this work. In succession, taking each embedding as a channel, the three embeddings are concatenated into a single tensor, which is the input of the spatial modelling.

FIGURE 1  The framework of the proposed model

FIGURE 2  The features of the waveform
The size of the concatenated tensor is \((T, 256, 3)\), where \(T\) is the length of the frames. To address the sparse distribution of speech features, the architecture in Inception [42] is applied to the spatial modelling block by designing multiple Conv2D kernels. As for the convolutional operation, different kernels take different local receptive fields as a single filter object which corresponds to the intensity of the feature map. In general, the multiple kernels on the temporal dimension are capable of improving the robustness for coping with different speech rates by taking different frames as a single acoustic unit, while the multiple kernels on frequency dimension are to make up the impact of the distributed vocalization features on the recognition task. The stride of the Conv2D is set to \((2, 2)\) to reduce the data size on both the temporal and frequency dimensions, which allows the model to discard disturbing features and reserve discriminative features. The padding mode of the Conv2D operation is set to ‘same’ to support the concatenation of the feature maps. Finally, the Conv2D operation with \((1, 1)\) kernel is followed to improve the ability to model the nonlinear features, which aims to learn more robust and discriminative features for the subsequent temporal modelling.

### 3.2 Residual LSTMs

In the ASR research, a long-standing idea is that the temporal correlations among speech frames are particularly important to improve the final performance. On the one hand, based on the real fact that the waveform is a natural temporal signal, temporal modelling is beneficial to improve the performance of the sequential classification task for the research of continuous speech recognition. On the other hand, thanks to the ATC procedure and terminology, the temporal correlations among ATC words are tighter than that of speech in our daily life. For instance, an aircraft identification, ‘air china one seven eight four’, always consists of an airline company name (terminology) and several followed digits. This procedure in the ATC domain enables us to promote the final performance by improving the accuracy of the temporal modelling.

Therefore, the RNN blocks, specifically LSTM, are applied to serve as the temporal feature extractor in existing works [23]. The LSTM was first proposed to address the modelling for long-term temporal dependencies [43]. Four modules, including forget gate, input gate, output gate, and cell, are designed to learn the transmission weights, which are able to control the information flow among different modules. The complicated architecture allows the LSTM layer to remember the long-term important information and forget the useless information. The inference rules of the LSTM block can be shown below:

\[
\begin{align*}
    I^t &= f(W_{ix}x^t + W_{ih}h^{t-1} + W_{ic}C^{t-1} + b_i) \\
    F^t &= f(W_{fx}x^t + W_{fh}h^{t-1} + W_{fc}C^{t-1} + b_f) \\
    C^t &= F^t \odot C^{t-1} + I^t \odot g(W_{cx}x^t + W_{ch}h^{t-1} + b_c) \\
    O^t &= f(W_{ox}x^t + W_{oh}h^{t-1} + W_{oc}C^t + b_o) \\
    h^t &= O^t \odot g(C^t)
\end{align*}
\]

where \(I^t,F^t,C^t,O^t\) are the activations of the input gate, forget gate, cell, and output gate at time instant \(t\), respectively. \(h^t\) is the hidden representation. \(W\) is the trainable weight matrix whose subscripts indicate the direction of the information flow. For example, the subscript /fh shows that the information is from the hidden unit of last time instant to the forget gate at the current time. \(\odot\) is the dot product operation. \(f\) and \(g\) are the sigmoid and tanh nonlinear functions, respectively.

In general, a deeper architecture is designed to learn the high-level representations to support the ASR task, in which the number of the LSTM layers is up to 7. In this way, the complicated architecture is supposed to impose training burdens and lead to the gradient vanishing problem. In this work, the residual mechanism is applied to the LSTM layers to improve its
trainability, and further to obtain better model convergence and final performance. In addition, since the ASR is a sequential classification task, the bidirectional mechanism is also performed on the LSTM layers to formulate the BLSTM layer, which benefits to improve the modelling accuracy from the past and future direction simultaneously. All in all, the scheme of the proposed LSTM layers, called RBLSTM, is described in Figure 4.

The LSTM layers follow the HFE block, which takes the CNN output as the input. The skip connection between the continuous layers is the residual connection. The RBLSTM can be represented as in (6), where \( \Gamma \) is the LSTM inference rules and \( \oplus \) denotes the elementwise adding operation:

\[
LSTM_i = \Gamma(i-1) \oplus \Gamma(i-2), \quad \forall i > 2
\]  

(6)

3.3 Pronunciation-oriented vocabulary

The main purpose of the ASR technique is to classify the input speech into a certain text class frame-wisely and search for a global optimal sequence for the whole utterance. The final performance highly depends on the vocabulary that can fully represent the mappings of the speech features. Existing works mainly focused on recognizing the speech in a single language, where the vocabulary can be phoneme, syllable, grapheme, word, or their hybrid combinations [33]. With the application of the deep learning technique, grapheme-based vocabulary has become the preferred option for the ASR task [27].

As described before, multilingual recognition is the crucial problem that needs to be solved for the ASR research in the ATC domain. In our previous works, the phoneme-based vocabulary was proposed to solve the multilingual ASR task in a non-end-to-end manner [12]. To achieve the end-to-end multilingual ASR task, an intuitive solution is to build the vocabulary with the raw grapheme for each language, i.e., Chinese character and English letter. However, the following issues in the ATC domain cause additional technical difficulties:

- From the linguistics perspective, the English letter is not in an identical granularity level with the Chinese character, correspondingly, the English word. If taking English word as the modelling unit, the data imbalance is another dilemma, which will impose unexpected burdens on the modelling training and convergence.
- With the raw graphemes, there will be a big gap between the length of the text sequence for Chinese and English speech. In general, a Chinese utterance usually contains only 8–12 Chinese characters, while it takes up to 30–50 English letters to represent an English ATC command [38]. It means that excessive blanks are needed to be padded to implement parallel training, which severely increases the consumption of the computational resource and training time and degrades the training efficiency and the final performance.

Therefore, it can be seen that the raw grapheme is not an optimal solution for the multilingual speech recognition task in this work. As well-known, Chinese is a single syllable language, while English is a multiple-syllable one [32]. It is believed that unifying the syllable scale is particularly important to achieve the multilingual recognition task in the end-to-end model. Following this idea, a pronunciation-oriented vocabulary (PoV) is designed to divide the English word into sub-words to ensure that all the units in the vocabulary are in a single syllable. For instance, the word 'maintain' is divided into 'main' and 'tain'. In this work, the PoV has the following advantages for addressing the multilingual ASR task:

- The PoV units are on an identical syllable scale with the Chinese character, which unifies the pronunciation scale for multilingual ASR task and further improves the model trainability and convergence.
- The PoV unit is between letter and word from the linguistics perspective. Unlike our previous work [12], all English words can be directly combined from PoV units, which are capable of generating human-readable texts and further formulating an end-to-end ASR paradigm.
- By dividing the English word into PoV unit, the class imbalance can be relieved to some extent, as demonstrated in Figure 5. It can be seen that almost 50% of English words appear less than 10 times in the ATC corpus, while the frequency of the PoV unit centres at a high value, more than 1500.
- By dividing the English word into PoV unit, the difference in the text sequence length is reduced. It further reduces the
sequence of padding for parallel training to improve the training efficiency.

Based on the above descriptions, unifying the pronunciation in multiple languages is the primary principle to generate the PoV for the ASR research in the ATC domain. The PoV implementation is based on the CMUDict tool. For the corpus in this work, a total of 305 sub-words are designed for the English speeches in this work, i.e., half of the words (584). The English sub-words are combined with the Chinese character to formulate the final vocabulary of the proposed ASR model. Note that the PoV is corpus-dependent, which is required to be optimized based on the training corpus.

4 | EXPERIMENTS

4.1 | Experimental configurations

In this work, a real-world ATC speech corpus is applied to validate the proposed approach, i.e., ATCSpeech. The details of the corpus can be found in our other paper [38]. The total data size is about 58 hours, including a 39-hour Chinese speech and a 19-hour English speech. The vocabulary consists of 698 Chinese characters and 584 English words. In this work, all the models are trained on the training set and evaluated on the joint set of the validation and test in the ATCSpeech corpus. Based on the technical properties of the proposed model, the following models are selected to serve as the baselines in this work:

- **DS2**: The architecture is constructed with two Conv1D layers (kernel size 11, and 1000 filters), seven BLSTM layers (512 nodes in each direction), and a fully connected layer. The details can be found in [27].

- **SPT**: The SPT model is a cascaded pipeline with the paradigm of ‘speech-phoneme-text’. The architecture is the same as that of in the [12], where the implementation details can also be found.

In the proposed model, the HFE configuration is first reported, while the BLSTM configuration is the same as that of the DS2 model (with a residual mechanism). Note that the DS2 model may be applied to several comparative experiments, which achieve the ASR task in single or multiple languages based on the baseline requirements. The modelling unit for the Chinese speech is the Chinese character (CC), while it is the English word (EW) or English letter (EL) for the English speech. Furthermore, the RNN-based LM [44, 45] is separately trained to correct the prediction errors from the perspective of the ATC semantic, where the n-best decoding strategy is applied to search the optimal text sequence.

In this work, all the deep learning models are implemented based on the open framework Keras with a TensorFlow backend. The training server is configured as follows: 2*Intel Core i7-6800K, 2*NVIDIA GeForce RTX 2080Ti, and 64 GB memory with operation system Ubuntu 16.04.

The Adam optimizer is selected to optimize the trainable parameters of the proposed model, whose initial learning rate is set to $10^{-4}$. The learning rate will be halved if the validation accuracy does not decrease for 10 consecutive tests (500 iterations per test). The batch size is set to 160. To reduce the training loss to a certain level as soon as possible, all training samples are sorted by their durations in the first training epoch. In the ATC domain, the speeches with a similar duration may imply that their texts belong to the same instruction and share a higher similarity. From the second epoch, the training samples are randomly shuffled to improve the robustness of the model. An early stopping strategy based on the validation loss is applied to check the training progress. For the LM decoding, the order of the LM is 5 and the beam width is 10.

The character error rate (CER), i.e., word error rate based on the Chinese character and English letter, is applied to evaluate
the proposed ASR model, as shown below:

\[
CER = \sum_{i=1}^{M} \frac{O_{i}}{N_{i}} \times 100\%
\]  

(7)

where \(M\) is the number of the test samples. For the \(i\)th sample, \(N_{i}\) is the length of the truth label, while \(O_{i}\) is the number of the operations that need to be performed to convert the prediction label into the ground truth, including insertion, deletion, and substitution. The CER evaluation is designed as a dataset-level metric to clarify the validation on the whole dataset.

In general, four modules, including the HFE block, residual LSTMs, PoV, and multilingual ASR paradigm, are proposed to address the speech recognition issue in the ATC domain. Therefore, the experimental design also follows the hint and different types of experiments are conducted to validate the proposed improvements, correspondingly.

### 4.2 Experiments for validating the proposed HFE block

The experiments in this section mainly focus on validating the HFE block and the multilingual ASR scheme for the ASR research in the ATC domain. To validate the HFE block, the LSTM configuration in the proposed model is the same as that of the DS2 model, i.e., without residual connection. The baseline is the DS2 model with the aforementioned configurations, in which different methods are applied to extract the speech features. Both the models are trained to transcribe the speech with a single language (Chinese or English) and multiple languages. The S and M in the ‘language’ column indicate that a certain model is applied to achieve the independent ASR task and multilingual ASR task, respectively. In this section, the modelling unit is the CC and EL for the Chinese and English speech, respectively. The experimental results are reported in Table 1, in which the listed CERs are obtained with the LM decoding strategy.

From the experimental results we can come to the following conclusions:

a. With the same configuration of the model architecture and feature engineering method, the multilingual ASR systems are able to obtain better performance compared to that of the independent ASR systems, as demonstrated in the experiment A1 vs. A2, A3 vs. A4, A5 vs. A6, and A7 vs. A8. It can be attributed that the multilingual speech samples can provide more distinctive distributions between the speech features and the text labels, which also proves that the multilingual ASR model is a practical solution in the ATC domain.

b. Regardless of the feature engineering method, the proposed HFE block achieves higher prediction accuracy than that of the baseline model (DS2), about 9.1% and 18.7% relatively CER improvements for the Chinese and English speech, respectively. The performance improvements result from the following two ways: the hybrid features to provide more data patterns for model learning, and multiple CNN kernels to fit sparse data distribution from the temporal and frequency dimensions.

c. For the baseline model DS2, the performance gaps for different speech features are marginal, while the hybrid feature input significantly promotes the final performance. This results further validates the proposed hybrid input and its feature embedding blocks.

In summary, by the comparative experiments, it is believed that the proposed HFE block with hybrid feature input and multiple CNN kernels advances the final performance for the ASR research in the ATC domain.

### 4.3 Experiments for validating the proposed residual LSTMs

In general, the efficiency and effectiveness of the HFE block and multilingual ASR are confirmed in the last section. Thus, the above improvements are directly applied to the experiments in this section to validate the proposed residual BLSTM architecture. In addition, the residual mechanism is also applied to the baseline DS2 model to confirm its applicability and generalization. The modelling unit for the multilingual ASR task is the joint set of the CC and EL, and the input of the DS2 model is the log-fbank feature.

Experimental results are reported in Table 2. In general, applying the residual mechanism on the BLSTM layers is capable of improving the final performance for both the proposed model and the baseline model. As a comparison, both the experiment B1 (vs. A2) and the B2 (vs. A6) achieve about 0.5% absolutely CER reduction.

As mentioned before, the most straightforward motivation for applying the residual mechanism on the BLSTM layer is to improve the training efficiency. To confirm this point, we also consider the change of the training loss during the training process, as described in Figure 5, in which the loss values are smoothed by averaging them with a sliding window (100).
TABLE 2  Experimental results for validating the residual BLSTMs

| No. | Architecture         | Chinese (%) | English (%) |
|-----|----------------------|-------------|-------------|
| A2  | HFE+BLSTM            | 5.5         | 7.0         |
| B1  | HFE+RBLSTM           | 5.1         | 6.5         |
| A6  | DS2                  | 6.0         | 8.4         |
| B2  | DS2 with RBLSTM      | 5.5         | 7.8         |

FIGURE 6  The training loss of the proposed approach with or without residual mechanism

The training loss of experiment B2 is not reported in the figure to clarify the comparison, which also shows a similar trend with that of experiment B1. It can be seen that the training loss generated with the RBLSTM layers can be reduced at a faster rate compared to that without the residual connections. It also indicates that we can achieve a preferred accuracy with less training time. Furthermore, the convergence loss generated with the RBLSTM layer is smaller than that without the residual mechanism, which allows the model to yield better performance. In summary, the residual mechanism on the BLSTM layers enable the ASR model to achieve higher accuracy with less training time.

4.4  Experiments for validating the proposed PoV

By the above experiments, the proposed HFE and residual BLSTMs are validated, which are applied to the experiments in this section. Three experiments are conducted to consider the performance improvements generated by the proposed PoV. In this section, three types of vocabulary serve as the modelling unit for the proposed ASR model, including CC+EL, CC+EW, and the proposed PoV. The model architecture is based on the HFE, RBLSTMs, and FC layer. The baseline model DS2 is also applied to train a multilingual ASR model to prove the generalization of the PoV.

The experimental results are reported in Table 3, which can be concluded as follows:

- From the perspective of the modelling unit, the proposed PoV demonstrates the preferred performance superiority compared to that obtained with the CC+EW and CC+EL on both two approaches. Taking the English word as the modelling unit, both the two models suffer the largest prediction error among the three options. It can be attributed that the frequency of the English word in the corpus is extremely imbalanced. In other words, there are no sufficient training samples between the speech feature and the vocabulary words that appear minor times, which results in the problems of model training and convergence and further degenerates the final performance.
- From the perspective of the model architecture, the proposed approach achieves higher performance compared to the baseline DS2 model regardless of the modelling unit. The results also indicate that the proposed architecture has the ability to accommodate different modelling units by the proposed HFE and residual BLSTMs architecture.

TABLE 3  Experimental results for validating the PoV

| No. | Architecture | Vocabulary | Chinese (%) | English (%) |
|-----|--------------|------------|-------------|-------------|
| C1  | DS2          | CC+EW      | 6.1         | 10.5        |
| C2  | DS2          | CC+EL      | 5.4         | 7.8         |
| C3  | PoV          |            | 5.0         | 7.3         |
| C4  | Proposed     | CC+EW      | 5.5         | 9.1         |
| C5  | CC+EL        |            | 5.1         | 6.5         |
| C6  | PoV          |            | 4.4         | 5.9         |

4.5  Experiments for validating the proposed ASR model

In the last sections, we mainly focus on the ablation study to validate the proposed technical improvements. In this section, several experiments are conducted to confirm the overall performance improvement obtained by the proposed ASR model with all improvements, as summarized as follows:

- Independent ASR system (D1): two DS2-based models are trained to transcribe the Chinese and English speech, respectively. The model architecture is described before, while the modelling unit is the proposed PoV.
- End-to-end multilingual ASR system (D2): a DS2-based model is applied to transcribe the Chinese and English speech, simultaneously, whose configurations are the same as that of the C3.
In this work, to achieve the speech recognition task in air traffic control, an end-to-end framework is proposed to integrate multilingual (Chinese and English) speech into a single model. Analyzing the characteristics of the ATC speech, a recurrent neural network-based end-to-end framework is applied to consider the temporal correlations of the speech signal, which is trained by the CTC loss function. First, a hybrid feature embedding block, i.e., HFE, is designed to mine the spatial dependencies by taking different types of speech features as the input. In the HFE block, the convolutional neural network is applied to extract high-level representations, in which multiple CNN kernels with different filter configurations are used to address the distribution of the sparse feature space caused by the radio transmission in the ATC domain. In succession, to further improve the trainability and model convergence, the residual mechanism is applied to the RNN layers (specifically, LSTM), which also benefits to advance of the final performance. To address the issue of the class imbalance of the vocabulary, a pronunciation-oriented vocabulary is designed to unify the pronunciation of the modelling unit for both Chinese and English speeches, which also formulates an end-to-end multilingual ASR model. The real-world ATC speech corpus is applied to the proposed approach, and we achieve a 4.4%- and 5.9%-character error rate on Chinese and English speech, respectively, outperforming other popular methods, such as DS2, cascaded multilingual pipeline. Furthermore, all the proposed technical improvements are proved to be helpful to improve the final performance. In summary, the proposed approach achieves the multilingual ASR task with considerable high performance in an end-to-end manner, which can serve as strong support in the field of air traffic research.

In the future, to improve the performance of the proposed method, we plan to apply more efficient architecture to the ASR model, such as the ConvLSTM (learn the spatial and temporal dependencies in a simultaneous and high-efficient manner), Transformer (apply the attention mechanism to learn more discriminative and informative features to support the ASR task). Moreover, speech representation learning is also another topic that deserves to be studied.

ACKNOWLEDGEMENT
This study was co-supported by the National Natural Science Foundation of China (No: 62001315, U20A20161), and the Sichuan Provincial S&T Projects (No: 2020YFG0327, 2020YFG0306).

REFERENCES
1. Lin, Y., et al.: A real-time ATC safety monitoring framework using a deep learning approach. IEEE Trans. Intell. Transp. Syst. 21(11), 4572–4581 (2020) https://doi.org/10.1109/TITS.2019.2940992.
2. Lin, Y., et al.: Real-time controlling dynamics sensing in air traffic system.. Sensors 19( 3), 679, (2019) https://doi.org/10.3390/
s19030679.
3. Hawkins, F.H. Human factors in flight. Appl. Ergon. 19( 4), 337 (1988) https://doi.org/10.1016/0003-6870(88)90093-2.
4. Lin, Y., et al.: ATCSpeechNet: a multilingual end-to-end speech recognition framework for air traffic control systems. Preprint http://arxiv.org/ abs/2102.08535 (2021)
5. Ferreiros, J., et al.: A speech interface for air traffic control terminals. Aerospace. Sci. Technol. 21(1), 7–15 (2012) https://doi.org/10.1016/j.ast.2011.05.002.

6. Geacă, C.M.: Reducing pilot/ATC communication errors using voice recognition. In: 27th International Congress of the Aeronautical Sciences, (2010), pp. 1–7.

7. Helmke, H., et al.: Reducing controller workload with automatic speech recognition. In: 2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC), Sacramento, CA, USA, 25–29 Sept. 2016 https://doi.org/10.1109/DASC.2016.7778024.

8. Lancaster, J.A., Casali, J.G.: Investigating pilot performance using mixed-modality simulated data link. Hum. Factors 50, 183-193 (2008).

9. Lin, Y.: Spoken instruction understanding in air traffic control: challenge, technique, and application. Aerospace 8(3), 65 (2021) https://doi.org/10.3390/aerospace8030065.

10. Pellegrini, T., et al.: The airbus air traffic control speech recognition 2018 challenge towards ATC automatic transcription and call sign detection. In: Proceedings of the Annual Conference of the International Speech Communication Association. INTERSPEECH 2019, 2993–2997, https://doi.org/10.21437/Interspeech.2019-1962.

11. Lin, Y., et al.: Improving speech recognition models with small samples for air traffic control systems. Neurocomputing 445(1), 287–297 (2021) https://doi.org/10.1016/j.neucom.2020.08.092.

12. Lin, Y., et al.: A unified framework for multilingual speech recognition in air traffic control systems. IEEE Trans. Neural Networks Learn. Syst. 1–13 (2020) https://doi.org/10.1109/TNNLS.2020.3015830.

13. Graves, A., et al.: Connectionist temporal classification. In: Proceedings of the 23rd international conference on Machine learning - ICML 06, New York, US (2006), 369–376, https://doi.org/10.1145/1143844.1143891.

14. Hinton, G., et al.: Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Process. Mag. 29(6), 82–97 (2012) https://doi.org/10.1109/MSP.2012.2208597.

15. Jiang, B.-H., L.R. Rabbiner: Speech recognition, automatic history. In: Encyclopedia of Language and Linguistics, pp. 806–819, Elsevier (2006).

16. Rabbiner, L.R.: A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE 77(2), 257–286 (1989) https://doi.org/10.1109/5.18626.

17. Maríno, J.B., et al.: N-gram-based machine translation. Comput. Linguist. 32(4), 527–549 (2006) https://doi.org/10.1111/j.1064-1249.2006.29.115.s27.

18. LeCun, Y., et al.: Deep learning. Nature 521(7553), 436–444 (2015) https://doi.org/10.1038/nature14539.

19. Lin, Y., et al.: Automated traffic incident detection with a smaller dataset based on generative adversarial networks. Accid. Anal. Prev. 144, 105628 (2020) https://doi.org/10.1016/j.aap.2020.105628.

20. Li, L., et al.: Real-time traffic incident detection based on a hybrid deep learning model. Transp. A Transp. Sci. 1–21 (2020) https://doi.org/10.1080/23249935.2020.1813214.

21. Sainath, T.N.: Convolutional, long short-term memory, fully connected deep neural networks. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, 19–24 Apr. 2015 https://doi.org/10.1109/ICASSP.2015.7188838.

22. Zhang, Y., et al.: Towards end-to-end speech recognition with deep convolutional neural networks. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, (2016), 410–414, https://doi.org/10.21437/Interspeech.2016-1446.

23. Graves, A., et al.: Speech recognition with deep recurrent neural networks. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013 https://doi.org/10.1109/ICASSP.2013.6639344.

24. Deng, L., et al.: New types of deep neural network learning for speech recognition and related applications: an overview. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013 https://doi.org/10.1109/ICASSP.2013.6639344.
44. Mikolov, T., et al.: Recurrent neural network based language model. In: Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH, Makuhari, Chiba, Japan, 26–30 sept. 2010

45. Mikolov, T., et al.: Extensions of recurrent neural network language model. In: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic, 22–27 May 2011, https://doi.org/10.1109/ICASSP.2011.5947611

How to cite this article: Lin Yi, et al.: Towards multilingual end-to-end speech recognition for air traffic control. IET Intell. Transp. Syst., 1–12. (2021) https://doi.org/10.1049/itr2.12094.