IEICE TRANS. INF. & SYST., VOL.E106–D, NO.4 APRIL 2023

PAPER

Speech Recognition for Air Traffic Control via Feature Learning and End-to-End Training

Peng FAN[†], **Xiyao HUA**[††], **Yi LIN**[††], **Bo YANG**[††], **Jianwei ZHANG**[††], ***Nonmembers***, **Wenyi GE**[†††], ***Member***, ***and*** **Dongyue GUO**[†a)], ***Nonmember***

SUMMARY In this work, we propose a new automatic speech recognition (ASR) system based on feature learning and an end-to-end training procedure for air traffic control (ATC) systems. The proposed model integrates the feature learning block, recurrent neural network (RNN), and connectionist temporal classification loss to build an end-to-end ASR model. Facing the complex environments of ATC speech, instead of the handcrafted features, a learning block is designed to extract informative features from raw waveforms for acoustic modeling. Both the SincNet and 1D convolution blocks are applied to process the raw waveforms, whose outputs are concatenated to the RNN layers for the temporal modeling. Thanks to the ability to learn representations from raw waveforms, the proposed model can be optimized in a complete end-to-end manner, i.e., from waveform to text. Finally, the multilingual issue in the ATC domain is also considered to achieve the ASR task by constructing a combined vocabulary of Chinese characters and English letters. The proposed approach is validated on a multilingual real-world corpus (ATCSpeech), and the experimental results demonstrate that the proposed approach outperforms other baselines, achieving a 6.9% character error rate.

key words: *automatic speech recognition, feature learning, air traffic control, multilingual, end-to-end training*

1. Introduction

Automatic speech recognition (ASR) translates speech into human-readable texts and has been widely used in various scenarios [1]. Recently, introducing the ASR and spoken language understanding (SLU) techniques into air traffic communications to reduce the workload of the ATCo and ensure flight safety has gathered more attention from researchers [2]. In this procedure, the ASR is one of the fundamental components and a high-confidence ASR result is the key to supporting downstream applications.

However, compared to the common ASR research, the ATC has many new challenges and difficulties. In general, the ATCo and pilots' speeches are usually in English. However, in China, the ATCos and pilots communicate through Chinese for domestic flights more frequently. That is to say, speech on the same frequency usually in both Chinese and English, i.e., multilingual ASR is required for the ATC domain [3]. Our previous work introduced ASR into the ATC safety monitoring framework, and also converted ATCo and pilot speech into instructions for controlling intent inference [4].

Recently, the end-to-end speech recognition system has provided higher performance than traditional methods for common ASR tasks [6]. However, current end-to-end ASR systems usually use mel-frequency cepstral coefficients (MFCCs) or filter-bank (FBANK) to process the raw waveform speech instead of directly inputting the raw waveform. In this procedure, the raw speech is divided into frames with 25 ms frame length and 10 ms shift, and a series of signal processing transformations are applied to convert the 1D waveform into 2D feature map. After the raw waveform speech is processed by MFCC or FBANK, the extracted feature map is fed into the neural network for acoustic modeling. This method has achieved state-of-the-art results in many ASR tasks. However, the design of FBANK and MFCC is based on the human ear's response to audio, it may lose some of the raw waveform speech information. In addition, compared to the common domain, there are still many challenges to tackling the acoustic specificities of the ATC speech due to the complexity in the real environment [3], [7], including radio background noise, high speech rate, unstable speech rate, etc. Considering the complex ATC environment, handcrafted feature engineering may not be an optimal option for ASR tasks. Therefore, exploring more effective feature learning approaches has become a promising technique to boost ASR performance in the ATC domain. Furthermore, learning informative and discriminative features from raw waveforms by the neural network has achieved desired performance improvement in many previous works, such as SincNet [5], wav2vec [8].

In this work, an end-to-end neural network is designed to achieve the ASR task in the ATC domain, in which a novel dual paths feature learning block is proposed to extract high-level speech representations from raw waveforms. Since the raw waveform is a one-dimensional signal, it is natural that learn speech representation from the raw waveform using the 1D convolution mechanism. Motivated by wav2vec and SincNet, both the SincNet and 1D convolution blocks are used to build the dual paths learning block to learn features from raw waveforms. The backbone network is constructed by recurrent neural network (RNN) layers and is jointly optimized with features learning block by the connectionist temporal classification (CTC) loss function. Most impor-

Manuscript received August 25, 2022.
Manuscript revised December 8, 2022.
Manuscript publicized January 23, 2023.

[†]The authors are with the National Key Laboratory of Fundamental Science on Synthetic Vision, Sichuan University, China.
[††]The authors are with the College of Computer Science, Sichuan University, China.
[†††]The author is with the College of Computer Science, Chengdu University of Information Technology, China.
a) E-mail: dongyueguo@stu.scu.edu.cn (Corresponding author)

DOI: 10.1587/transinf.2022EDP7151

tantly, the proposed model leverages the dual paths feature learning block to implement the end-to-end training, which predicts the text sequence from raw waveform without any pretraining.

The model proposed in this paper introduces a feature learning approach for speech recognition tasks in the ATC domain, and the work in this paper is dedicated to solving the multilingual speech recognition problem in the ATC domain. The experimental results show that our approach outperforms other baselines on the ATCSpeech corpus, achieving a 6.9% character error rate (CER), i.e., 0.9% absolute CER reduction over other baselines.

2. Related Work

With the rapid development of deep learning techniques in the past decades, it has outperformed the conventional methods for the ASR task [9], [10]. A deep learning-based feature extracted method–SincNet, was proposed to deal with the ASR task and speaker recognition task, and the experimental results showed that the neural network based on SincNet achieved better performance for both two tasks. The SincNet can learn more informative and discriminative features from raw waveform [5], [11]. Other deep learning-based models, like wav2vec, were also proposed to extract speech features from raw waveforms. The wav2vec model explores unsupervised pre-training for speech recognition by learning representations from raw audio through several 1D convolution layers. The wav2vec model is trained on large amounts of unlabeled speech and the resulting representations serve as the input of the acoustic model for the ASR task [8].

For the common ASR applications, the SincNet was proposed to be combined with end-to-end architecture, which achieved higher accuracy on Wall Street Journal corpus [12]. In addition, the sinc-convolution was combined with the depthwise convolutions to construct the lightweight sinc-convolutions (LSC) model [13], which serves as a learnable feature extraction block for end-to-end ASR systems with minor trainable parameters. Furthermore, the wav2vec2.0 is combined with the BERT to construct an end-to-end ASR model, achieving higher performance by learning features from pre-trained models [14]. Additionally, the SincNet-based feature learning approaches were also applied in many applications and achieved desired performance improvements. In [15], the SincNet and LEAF [16] were employed as the learnable frontends to extract the time-frequency representations from the raw-waveform domain and show competitive performance in the speech command classification tasks. Similarly, a parameterized convolutional neural network (CNN) was used for acoustic modeling from raw waveform for the dysarthria speech recognition tasks [17]. In [18], a SincNet-based speech feature learning method was proposed to achieve automatic smoker identification tasks. It can be found that investigating learnable frontends has drawn a lot of attention from researchers and made significant progress in the field of speech processing.

For the ASR research in the ATC domain, several state-of-the-art ASR models were applied to build a benchmark, which was trained on more than 170 hours of ATC speech [19]. The contextual knowledge was integrated into the ASR model to achieve semi-supervised training, which provided better performance for recognizing callsign of the ATC instruction [20]. Our previous work built a unified framework for multilingual speech recognition in the ATC system to translate ATC and pilot multilingual speech to text [3]. To deal with the scare of high quality annotated training data in the ATC domain, a novel method was proposed to leverage pretraining and transfer learning [21]. In order to improve the call sign recognition rate in ATC speech, a context-aware language model is proposed [22]. In addition, a method based on deep learning is proposed to solve speaker role identification in the air traffic control domain [23].

3. Methodology

In this work, a complete end-to-end architecture is proposed to achieve the multilingual ASR task in the ATC domain by cascading a dual paths feature learning block and backbone network. Both the SincNet and convolutional layers are applied to formulate a dual paths feature learning block to extract high-level representations from raw waveforms. By combining with the backbone network (referred to the Deep Speech 2 model [24]), the proposed model is finally optimized by the CTC loss function. In addition, to consider the multilingual ASR issue in the ATC domain, a special vocabulary is built based on Chinese characters and English letters. The architecture of the proposed model is illustrated in Fig. 1. Thanks to the design of the dual paths feature learning block, the proposed model is able to directly predict the text sequence from the waveforms and can be optimized in an end-to-end manner with the goal of an ASR task, instead of a pretraining task.

3.1 Dual Paths Feature Learning

In the ATC domain, the ASR task faces the challenges of multilingual language, complex background noise, etc, which results in the fact that handcrafted feature engineering may not be an optimal option for the ASR task. Therefore, the dual paths feature learning block is applied to extract features from raw waveforms in a learnable way, which further supports the acoustic modeling of the ASR task.

In this work, a novel dual paths feature learning block is proposed to learn informative features from raw waveforms, whose outputs are generated by concatenating the feature maps of different learning paths. For ASR, the wav2vec model learning speech representations from the raw waveforms by 1D convolution beyond the traditional handcrafted feature [8]. Motivated by wav2vec, the 1D convolution be chosen as one of the dual paths. In general, the dual paths feature learning block consists of dual paths:

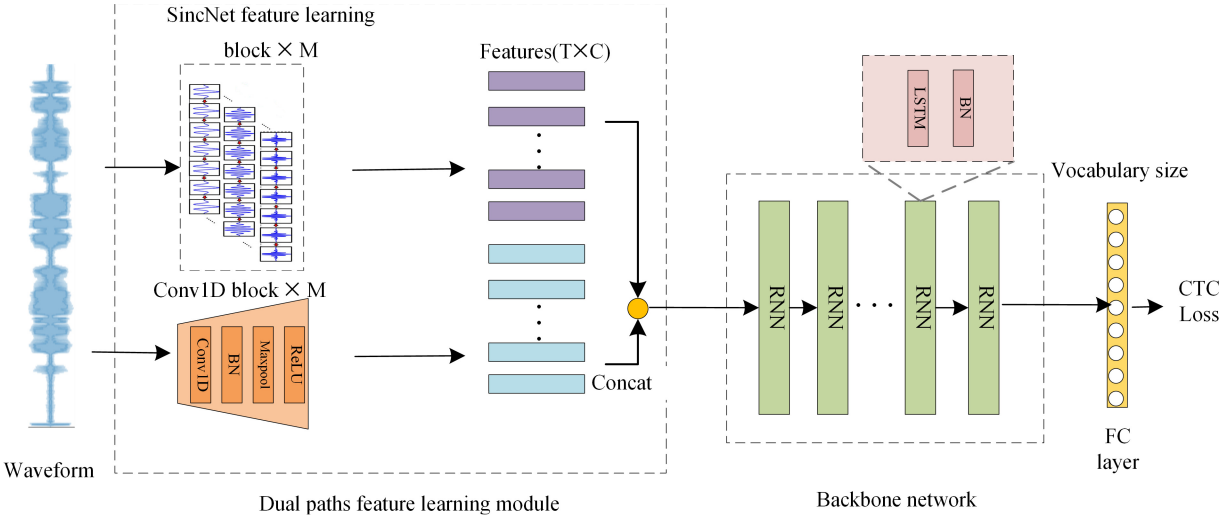

Fig. 1 The architecture of the proposed end-to-end automatic speech recognition system. The model consists of the dual paths feature learning block and the backbone network. A dual paths feature learning block is used to learn features from the raw waveforms. The drawn of SincNet components are referred from [5].

a) SincNet path: including a sinc-convolutional layer and four convolutional layers. The configurations are as follows: filter-sizes [129, 3, 3, 3, 3], strides [1, 1, 1, 1, 1], max pooling size [3, 3, 3, 3, 3], and ReLU activations. b) Conv1D block path: including five Conv1D blocks, whose configurations are the same as that of the SincNet path to keep the same feature map size, supporting the concatenation operation of the dual paths. Each Conv1D block is cascaded by a 1D convolutional layer, batch normalization layer, max-pooling layer, and ReLU layers.

SincNet is a novel convolutional neural network (CNN), which performs convolutional perception using the specific Sinc function. The layer is an improved neural network based on the parametrized Sinc functions, which has the ability to learn task-oriented features from speech signals directly. In general, the convolution operation is defined as follows:

$$y[n] = x[n] * h[n] = \sum_{l=0}^{L-1} x[l] \cdot h[n-l]. \tag{1}$$

$$y[n] = x[n] * g[n, \theta], \tag{2}$$

where $*$ is convolution operation, $x[n]$ is a chunk of the speech signal, $h[n]$ is a filter with length L, and $y[n]$ is final filtered result. Compared to a large number of parameters in the common CNN block, the sinc-convolution operator is able to achieve the signal processing with much fewer trainable parameters by defining the filter function g. In this work, the rectangular bandpass filter is applied to serve as the filter-bank to implement the speech processing, this function g can be written in the time domain as follow:

$$g[n, f_1, f_2] = 2f_2 \operatorname{sinc}(2\pi f_2 n) - 2f_1 \operatorname{sinc}(2\pi f_1 n). \tag{3}$$

In the above function, the sinc function is defined as $\operatorname{sinc}(x) = \sin(x)/x$, where f_1 and f_2 are the learned low and high cutoff frequency of the bandpass filter, respectively. The parameter f is randomly initialized in the range of $[0, f_s/2]$, and f_s is the sample rate of the speech signal.

$$g_w[n, f_1, f_2] = g[n, f_1, f_2] \cdot w[n]. \tag{4}$$

$$w[n] = 0.54 - 0.46\cos(2\pi n/L). \tag{5}$$

Furthermore, to smooth out the abrupt discontinuities at the end of g, an available option is used to make g multiplied by a window function w, such as the popular Hamming window [5].

3.2 The Backbone Network

Our work is motivated by the end-to-end ASR system Deep Speech 2 [24]. In this work, we explore architecture with a dual paths feature learning block and backbone network. The backbone network consists of 7 RNN blocks and a fully connected layer, which is further optimized with the CTC loss. The RNN block includes bidirectional long short-term memory (Bi-LSTM) layers and a batch normalization layer. The batch normalization is applied to speed up the model convergence, while the dropout layer is used to prevent the overfitting problem. The ReLU is selected as the activation function for the proposed model.

3.3 The CTC Loss

In the proposed end-to-end ASR model, the goal is to predict the text sequence $S = \{s_1, \ldots, s_m\}$ from the input speech signal $X = \{x_1, \ldots, x_o\}$, in which s_i is from a special vocabulary based on Chinese characters and English letters.

In general, multiple frames in X correspond to a token of S. The length of speech frames is usually much longer than the label length. To address this issue, the CTC loss

function was designed to automatically achieve the alignment between the speech and label sequence. The tth frame corresponds to the output label k and its probability is denoted $z^t_{\pi_t}$. Given the speech input X, the probability of the output sequence π is shown in (6). Therefore, the probability of the final sequence can be obtained by (7), in which v is the set of all possible sequences and A denote the length T sequences over the vocabulary. For example, by using '_' to denote a blank, both the outputs "X_YY_Z" and "$\_XY\_Z\_$" correspond to the final output "XYZ" [25].

$$p(\pi|X) = \prod_{t=1}^{T} z^t_{\pi_t},\ \pi \in A. \tag{6}$$

$$p(S|X) = \sum_{\pi \in v^{-1}(s)} p(\pi|X). \tag{7}$$

4. Experiments

4.1 ATC Corpus

In this work, the training data of the proposed model is the ATCSpeech corpus, which is collected from the real-world ATC environment and manually annotated [26]. All the utterances of the ATCSpeech corpus are with the 8000 Hz sample rate. The ATCSpeech corpus is a multilingual corpus containing Chinese and English speeches. There are 16939 transcribed English utterances (about 18.69-hour) in the corpus, while 45586 (about 39.83-hour) for Chinese speech. The division for the train, validation and test set can be found in Table 1.

4.2 Experimental Configurations

In this work, the proposed model is constructed based on the open framework PyTorch 1.7.0. The training server was equipped with an Intel i7-9700 processor, a single NVIDIA TITAN RTX GPU, 32-GB memory, and an Ubuntu 18.04 operating system.

During the model training, the Adam optimizer is used to optimize the trainable parameters. The initial learning rate is 0.0001. The batch size is set to 32. In the first epoch, the speech samples are sorted in reverse order (based on speech duration) to detect the overflow of GPU memory as early as possible. In the following epochs, the training samples are shuffled to improve the model's robustness. The vocabulary is built on Chinese characters and English letters, and also with some special tokens (<UNK>, <$SPACE$>). Finally, a total of 705 tokens in the vocabulary. As in [26], the Deep Speech 2 [24], Jasper [27] and Wav2Letter++ [28], which are applied to achieve the monolingual and multilingual ASR task in this work. In addition, a competitive ASR architecture, Conformer [29] is also selected as the baseline. In order to ensure the fairness of the experiment, all those models are trained on the same dataset (ATCSpeech) without extra training data.

To confirm the effectiveness of the acoustic model, no language model is integrated into the ASR decoding procedure, i.e., greedy strategy.

4.3 Overall Results

In this section, all the models are applied to achieve both monolingual and multilingual ASR tasks. To confirm the efficacy of the dual paths feature learning block, handcrafted feature engineering is performed to generate the input for baseline models. The experimental results for all the models are reported in Tables 2–4.

In general, the proposed approach yields the highest performance among all the models for both monolingual and multilingual ASR tasks. The Deep Speech 2 baseline obtains better performance among the four baselines while the conformer model does not perform best due to limited dataset size and batch size. We also observed that the loss of the Deep Speech 2 model decreases fastest and the network converges earliest than other approaches. This can be attributed to that temporal modeling provides sig-

Table 1 Data size of the corpus. "#U" denotes the speeches utterances and "#H" denotes the speeches hours.

Language	Train		Dev		Test	
	#U	#H	#U	#H	#U	#H
Chinese	43186	37.77	1200	1.04	1200	1.03
English	15282	16.84	850	0.95	807	0.89

Table 2 The result of Chinese speech. "Fea." details the type of input features employed, and "Training" denotes the loss and epoch of the training. "Dev" denotes the validation dataset. Results are expressed in CER.

Models	Fea.	Training		Dev	Test
		Loss	Epoch	CER%	CER%
Deep Speech 2 [26]	FBANK	0.53	33	8.1	8.1
Jasper10*3 [26]	FBANK	2.45	101	11.2	11.3
Wav2letter++ [26]	FBANK	2.37	136	14.2	14.3
Conformer [29]	FBANK	0.31	91	8.9	8.9
Ours	RAW	0.26	105	7.6	7.6
Ours (multilingual model)	RAW	0.26	102	7.3	7.3

Table 3 The result of English speech.

Models	Fea.	Training		Dev	Test
		Loss	Epoch	CER%	CER%
Deep Speech 2 [26]	FBANK	0.54	107	10.4	10.4
Jasper10*3 [26]	FBANK	0.91	200	9.2	9.3
Wav2letter++ [26]	FBANK	1.06	307	11.3	11.4
Conformer [29]	FBANK	0.22	178	10.9	11.0
Ours	RAW	0.21	150	8.9	8.9
Ours (multilingual model)	RAW	0.26	102	6.3	6.3

Table 4 The result of the multilingual speech.

Models	Fea.	Training		Dev	Test
		Loss	Epoch	CER%	CER%
Deep Speech 2	FBANK	0.45	30	7.8	7.8
Jasper10*3	FBANK	2.35	97	10.0	10.1
Wav2letter++	FBANK	2.29	115	12.3	12.4
Conformer [29]	FBANK	0.28	88	8.3	8.4
Ours	RAW	0.26	102	6.9	6.9

Table 5 The result of the model with different SincNet kernel sizes. "Kernel Size" details the kernel size of the proposed method first convolution layer.

Kernel Size	Training		Dev	Test
	Loss	Epoch	CER%	CER%
251	0.31	90	7.3	7.4
129	0.26	102	6.9	6.9
65	0.25	95	7.0	7.1

Table 6 The results of different feature learning networks structure. "Models" is the backbone network with different feature learning blocks.

Models	Training		Dev	Test
	Loss	Epoch	CER%	CER%
1D CNN*1 (a)	0.27	93	7.4	7.5
SincNet*1 (b)	0.26	95	7.1	7.2
SincNet*2 (c)	0.25	97	7.1	7.1
ours (d)	0.26	102	6.9	6.9

nificant performance improvement for the ASR task, which is also the motivation of our backbone network. In addition, the multilingual ASR system is able to obtain higher accuracy than that of the monolingual ASR system, benefiting from the larger dataset and prominent discrimination between Chinese and English. However, since the proposed approach learns speech features directly from the raw waveforms, there are more parameters that need to be learned which makes the loss decrease and the network converge slower than the Deep Speech 2 baseline in training. Specifically, for the Chinese speech, the proposed model achieves 7.6% CER, i.e., 0.5% absolute CER reduction over the Deep Speech 2 model. Furthermore, for the proposed model, it achieves 7.3% CER, i.e., 0.3% absolute CER reduction train on the multilingual data and is superior to that of only training on Chinese speech. For English speech, the proposed model obtains 0.4% absolute CER reduction over the Jasper model (highest baseline accuracy), reaching 8.9% CER. The proposed model training on multilingual speech achieves 6.3% CER, i.e., 2.6% absolute CER reduction over that of training on English speech. To be specific, for the multilingual ASR task, the proposed model achieves higher performance than that of the monolingual ASR system, achieving 6.9% CER. The results also confirm the effectiveness of the dual paths feature learning block in the proposed model, which learns more informative and discriminative features for the ASR task.

4.4 Ablation Study

In this section, we explore different convolution configurations of the SincNet by ablations to find an optimal model architecture. As illustrated in [5], the size of the convolution kernel and the size of max pooling will affect the risk of aliasing in the filtered signal. The kernel size of 251 (as in [5]) is selected as the baseline for the proposed model in this work.

Considering the higher speech rate in the ATC environment [3], a smaller kernel size is designed for the proposed dual paths feature learning block (for both the SincNet and common CNN paths), including 129 and 65. For the convolutional operation, a smaller kernel takes fewer frames as a single phoneme state, corresponding to a higher speech rate. The experimental results are listed in Table 5. It can be seen that the proposed model achieves the highest accuracy with the kernel size of 129, which is the optimal option for the ATCSpeech corpus.

In addition, the contribution of the different paths in the dual paths feature learning block is also considered to improve the final performance. As listed in Table 6, a total of 4 configurations are designed in this ablative study, including the single path (a and b) and dual paths (c and d). As can be seen from the experimental results, both the four configurations achieve desired performance improvement baseline models, i.e., 7.4% CER v.s. 7.8% CER, which confirms the effectiveness of the learning mechanism for the ASR task in the ATC domain. In general, the dual paths configuration achieves higher performance than that of the single path, which benefits from the model capacity for learning diverse task-oriented features. In addition, the single path SincNet (a) is able to obtain comparable performance with that of the dual paths SincNet (c), which indicates that only simply accumulating the same architecture fails to learn informative features for improving the ASR performance. Finally, the proposed model achieves the most significant performance improvement (6.9% CER) by concatenating the SincNet and CNN block, which supports the motivation of model design in this work. In conclusion, both the model capacity and the feature diversity are indispensable to tackling the ASR specificities in the ATC domain.

5. Conclusion

In this work, we present an innovative fully end-to-end ASR model using the proposed dual paths feature learning block in the ATC domain, which results in better recognition performances without extra training data. For the ASR in the ATC domain, the proposed dual paths feature learning block can learn more meaningful information from raw waveforms, and the feature learning-based ASR system performance beyond the handcrafted feature learning-based ASR system. Moreover, the dual paths feature learning block can learn more diverse features than the single path feature learning ASR system in the field of ATC. In addition, it is also found that adding SincNet paths cannot further improve the final performance, while better results can be obtained by cascading SincNet and Conv1D blocks. Thanks to the increased capacity of the proposed model, more diverse features can be learned in ASR tasks in the ATC domain. The experimental results have demonstrated the proposed method outperforms other baselines in multilingual ASR tasks on the ATCSpeech corpus.

Acknowledgments

This work was supported by the National Natural Sci-

ence Foundation of China under Grants 62001315 and U20A20161, the Open Fund of Key Laboratory of Flight Techniques and Flight Safety, Civil Aviation Administration of China (CAAC) under Grant No. FZ2021KF04, and Fundamental Research Funds for the Central Universities under Grant No. 2021SCU12050.

References

[1] M. El Ayadi, M.S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," Pattern Recognition, vol.44, no.3, pp.572–587, 2011.
[2] C.M. Geacăr, "Reducing pilot/ATC communication errors using voice recognition," Proc. ICAS, 2010.
[3] Y. Lin, D. Guo, J. Zhang, Z. Chen, and B. Yang, "A unified framework for multilingual speech recognition in air traffic control systems," IEEE Trans. Neural Netw. Learn. Syst., vol.32, no.8, pp.3608–3620, 2021.
[4] Y. Lin, L. Deng, Z. Chen, X. Wu, J. Zhang, and B. Yang, "A real-time ATC safety monitoring framework using a deep learning approach," IEEE Trans. Intell. Transp. Syst., vol.21, no.11, pp.4572–4581, 2020.
[5] M. Ravanelli and Y. Bengio, "Interpretable convolutional filters with SincNet," arXiv preprint arXiv:1811.09725, 2018.
[6] S. Kim, T. Hori, and S. Watanabe, "Joint CTC-attention based end-to-end speech recognition using multi-task learning," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.4835–4839, IEEE, 2017.
[7] Y. Lin, "Spoken instruction understanding in air traffic control: Challenge, technique, and application," Aerospace, vol.8, no.3, 65, 2021.
[8] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," arXiv preprint arXiv:1904.05862, 2019.
[9] T.N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.4580–4584, IEEE, 2015.
[10] H. Soltau, H. Liao, and H. Sak, "Neural speech recognizer: Acoustic-to-word LSTM model for large vocabulary speech recognition," arXiv preprint arXiv:1610.09975, 2016.
[11] M. Ravanelli and Y. Bengio, "Speech and speaker recognition from raw waveform with SincNet," arXiv preprint arXiv:1812.05920, 2018.
[12] T. Parcollet, M. Morchid, and G. Linarès, "E2E-SincNet: Toward fully end-to-end speech recognition," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.7714–7718, IEEE, 2020.
[13] L. Kürzinger, N. Lindae, P. Klewitz, and G. Rigoll, "Lightweight end-to-end speech recognition from raw audio data using sinc-convolutions," arXiv preprint arXiv:2010.07597, 2020.
[14] C. Yi, S. Zhou, and B. Xu, "Efficiently fusing pretrained acoustic and linguistic encoders for low-resource speech recognition," IEEE Signal Process. Lett., vol.28, pp.788–792, 2021.
[15] S. Yadav and N. Zeghidour, "Learning neural audio features without supervision," Proc. Interspeech 2022, pp.396–400, 2022.
[16] N. Zeghidour, O. Teboul, F. de Chaumont Quitry, and M. Tagliasacchi, "LEAF: A learnable frontend for audio classification," arXiv preprint arXiv:2101.08596, 2021.
[17] Z. Yue, E. Loweimi, H. Christensen, J. Barker, and Z. Cvetkovic, "Dysarthric speech recognition from raw waveform with parametric CNNs," Proc. Interspeech 2022, pp.31–35, 2022.
[18] Z. Ma, Y. Qiu, F. Hou, R. Wang, J.T.W. Chu, and C. Bullen, "Determining the best acoustic features for smoker identification," ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.8177–8181, 2022.
[19] Z.G. Juan, P. Motlicek, Q. Zhan, R. Braun, and K. Vesely, "Automatic speech recognition benchmark for air-traffic communications," Tech. Rep., ISCA, 2020.
[20] J. Zuluaga-Gomez, I. Nigmatulina, A. Prasad, P. Motlicek, K. Veselỳ, M. Kocour, and I. Szöke, "Contextual semi-supervised learning: An approach to leverage air-surveillance and untranscribed ATC data in ASR systems," arXiv preprint arXiv:2104.03643, 2021.
[21] Y. Lin, Q. Li, B. Yang, Z. Yan, H. Tan, and Z. Chen, "Improving speech recognition models with small samples for air traffic control systems," Neurocomputing, vol.445, pp.287–297, 2021.
[22] D. Guo, Z. Zhang, P. Fan, J. Zhang, and B. Yang, "A context-aware language model to improve the speech recognition in air traffic control," Aerospace, vol.8, no.11, 348, 2021.
[23] D. Guo, J. Zhang, B. Yang, and Y. Lin, "A comparative study of speaker role identification in air traffic communication using deep learning approaches," ACM Trans. Asian Low-Resour. Lang. Inf. Process., accepted.
[24] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen, et al., "Deep speech 2: End-to-end speech recognition in English and Mandarin," International Conference on Machine Learning, pp.173–182, PMLR, 2016.
[25] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," Proc. 23rd International Conference on Machine Learning, pp.369–376, 2006.
[26] B. Yang, X. Tan, Z. Chen, B. Wang, D. Li, Z. Yang, X. Wu, and Y. Lin, "ATCspeech: A multilingual pilot-controller speech corpus from real air traffic control environment," arXiv preprint arXiv:1911.11365, 2019.
[27] J. Li, V. Lavrukhin, B. Ginsburg, R. Leary, O. Kuchaiev, J.M. Cohen, H. Nguyen, and R.T. Gadde, "Jasper: An end-to-end convolutional neural acoustic model," arXiv preprint arXiv:1904.03288, 2019.
[28] V. Pratap, A. Hannun, Q. Xu, J. Cai, J. Kahn, G. Synnaeve, V. Liptchinsky, and R. Collobert, "Wav2letter++: A fast open-source speech recognition system," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.6460–6464, IEEE, 2019.
[29] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," arXiv preprint arXiv:2005.08100, 2020.

Peng Fan received the B.E. degree in mathematics from Heilongjiang Bayi Agricultural University, Daqing, China, and the M.S. degree in software engineering from Chengdu University of Technology, Chengdu, China. He is currently pursuing the Ph.D degree at the National Key Laboratory of Fundamental Science on Synthetic Vision, Sichuan University, Chengdu, China. His current research interests include pattern recognition and automatic speech recognition.

Xiyao Hua received the B.E. degree in electronic and information engineering from Northwest Normal University, Lanzhou, China, and the M.S. degree in computer science from Jilin University, Jilin, China. He is currently pursuing the Ph.D degree with the College of Computer Science, Sichuan University, Chengdu, China. His research interests include light field image processing and deep learning.

Yi Lin received the Ph.D. degree from Sichuan University, Chengdu, China, in 2019. He currently works as an Associate Professor with the College of Computer Science, Sichuan University. He is also a Visiting Scholar with the Department of Civil and Environmental Engineering, University of Wisconsin-Madison, Madison, WI, USA.

Bo Yang received the Ph.D. degree from Sichuan University, Chengdu, China, in 2012. He has taught and conducted research work with Sichuan University since 1996. His research interest includes deep-learning application for air traffic management. Dr. Yang was the recipient of the National Science and Technology Progress Award twice in China.

Jianwei Zhang received the Ph.D. degree from Sichuan University, Chengdu, China, in 2008. He has taught and conducted research at Sichuan University since 1993. He has published more than 30 articles. His research interests include air traffic management, and intelligent image analysis and processing.

Wenyi Ge received the Ph.D. degree from Sichuan University, Chengdu, China, in 2020. He currently works as an Assistant Professor with the College of Computer Science, Chengdu University of Information Technology, Chengdu, China. His research interests include air traffic flow management and related machine learning applications.

Dongyue Guo received the M.E. degree in 2020 from Sichuan University, Chengdu, China, where he is currently working toward the Ph.D. degree with the National Key Laboratory of Fundamental Science on Synthetic Vision, College of Computer Science. His research interest is in automatic speech recognition and flight trajectory prediction.