Audio-Visual Fusion Framework for Low-Resource Language Speech Recognition Based on Progressive Down-sampling and Grouped Multi-Heads Attention Mechanism

ChongChong Yu*, Xiaolong Xu, Zhaopeng Qian, Kejing Xiao, Yuchen Tan
Beijing Technological and Business University, China
Beijing Institute of Graphic Communication, China
* Corresponding author: chongzhy@vip.sina.com

***Abstract*—The model of automatic speech recognition (ASR) has an overfitting problem in the task of low-resource speech recognition, due to the lack of training data. Different from conventional methodologies that rely on signal acoustic modality, we proposed to design a novel ASR model for the low-resource speech recognition task by fusing acoustic and visual modalities, standing from the perspective of the physiological mechanism of human pronunciation. In this paper, the progressive down-sampling (PD) and grouped self-attention mechanism (GASM) are applied to design the audio-visual speech recognition (AVSR) framework, called PD-GASM, for the low-resource speech recognition task. In particular, The PD algorithm reduces computational resource consumption during the training process. The GASM algorithm is used to help the trained model handle both short and long sequences; therefore, the proposed model can be more suitable for processing spoken dialogue sequences. Our proposed method has been evaluated based on the Tujia language audio-visual data corpus, miniLRW and miniLRS2 audio-visual data corpus. The experimental results show that the best character error rates can achieve 35.15%, 14.89%, and 16.33%, respectively. Our research has significant implications for expanding the application of AVSR in the field of low-resource speech recognition.**

I. Introduction

Collecting annotated data for low-resource languages is extremely difficult. In [1], the authors argue that despite the existence of nearly 7,000 languages worldwide, the majority lack sufficient labeled data for effective Natural Language Processing (NLP) modeling. Data scarcity in low-resource application scenarios severely limits the performance of models. To address the problem of data scarcity, some researchers have explored developing automatic speech recognition (ASR) for low-resource language speech.

The statistical language model for low-resource languages faces a great challenge due to the scarcity of training data. In [2], authors utilize abundant training data from high-resource language speech to improve the performance of ASR for low-resource language speech. However, multi-linguistic and cross-linguistic modeling encounter severe mismatching problems in the acoustic feature domain. Moreover, in [3] to address the problem of data scarcity in low-resource tasks, transfer learning is used to pre-train the model based on a large-scale data corpus and then fine-tune the upper layer of the target language speech model using the target language speech corpus. In [4], data augmentation enriches the amount of training data by generating pseudo-labeled data. Generally, a text-to-speech (TTS) system is used to synthesize the target speech according to the input text. In [5], meta-learning is employed to address the issue of the model adapting to unseen data for training. In [6] multi-task learning can enhance the generalization performance of the acoustic model by jointly learning multiple related tasks.

Traditionally, researchers have overlooked the biophysical knowledge of speech when developing ASR systems for low-resource language speech. The acoustic modality is used as a clue to recognize speech. However, a single acoustic modality cannot meet the requirements in low-resource language speech recognition. Audio-visual fusion technology can significantly enhance the robustness of the speech recognition model. In [7], audio-visual speech recognition (AVSR) effectively improve the performance of the model for low-resource language speech. Therefore, differing from traditional methodologies that rely on a single acoustic modality, we propose to integrate the visual lip movements with the acoustic information of speech to further enhance the model's performance in low-resource tasks. AVSR aims to mitigate the overfitting issue in low-resource language speech recognition.

We previous study [8], explored using the AVSR for low-resource speech recognition. Our experimental results show that the best character error rate (CER) of AVSR decreased by 16.9% compared with the ASR using single acoustic modality and decreased by 11.8% compared with the model using single visual lip movements. However, the performance of AVSR still limits the accuracy of low-resource language speech recognition. Therefore, in this paper, we propose to improve the performance of AVSR by incorporating the progressive down-sampling (PD) and grouped self-attention mechanism (GSAM) for low-resource language speech.

Our main contributions in this paper are as follows:

Humanities and Social Sciences Research Planning Fund of the Ministry of Education of China (No. 21YJAZH107)

979-8-3315-7206-8/25/$31.00

1) GSAM and PD are incorporated to design the AVSR framework based on PD-GSAM for low-resource language speech.

2) The PD algorithm in the framework is used to reduce the cost of training the conformer.

2) GSAM is used to address the problem that the complexity of the self-attention mechanism simultaneously handling both long and short sequences is too high.

4) We designed data increment experiments, modality comparison experiments, and ablation experiments to evaluate the proposed method.

II. RELATED WORKS

In tackling the challenge of data scarcity for speech in low-resource languages, scholars have introduced a myriad of innovative approaches. This paper provides a comprehensive overview of prior studies, structuring them into seven distinct categories. These include the ”data corpus of low-resource language speech,” which addresses the fundamental issue of data availability. The ”multi-linguistic and cross-linguistic model” part discusses methods that transcend language-specific limitations. The ”data augmentation of speech” section explores techniques to enhance the existing data. The ”transfer-learning model” part delves into the application of knowledge transfer between models. The ”meta-learning model” section examines how models can learn to adapt quickly to new tasks with limited data. The ”multi-task learning model” part discusses simultaneous learning of multiple tasks to improve efficiency and performance. Lastly, the ”improvement of deep learning model” category focuses on enhancements directly to the deep learning architectures employed in ASR systems. Each of these categories plays a critical role in advancing the field of low-resource ASR, offering unique solutions to the overarching challenge of data scarcity.

A. Low-Resource Language Speech Databases

The availability of data corpora plays a pivotal role in the training of deep learning models for low-resource language speech tasks. Despite its importance, research on data corpora for such tasks is still relatively scarce.For example in [9], authors contributed to the field with the Tujia language speech corpus, which includes 2105 audio-visual pair utterances from 4 native Tujia language speakers. These endeavors collectively underscore the progressive efforts to enrich the data landscape for low-resource speech recognition tasks.

B. Cross-Lingual Model and Multi-Lingual Model

In response to the challenge posed by the scarcity of data for low-resource languages, interpolation techniques have been frequently employed to transition from models trained on high-resource language speech to those trained on low-resource language speech. In [10], authors put forward the concept of utilizing cross-lingual language modeling with syntactic information for the recognition of low-resource speech.

In [11], authors incorporated the adaptive activation network into the domain of low-resource multilingual ASR. This integration saw the adaptive activation function replacing traditional activation functions within the layers of recurrent neural networks (RNNs) and deep neural networks (DNNs). Moreover, they championed a dual-learning strategy paradigm to actualize this linguistic adaptability. The first strategy is cross-lingual learning, which involves the substitution of language-adaptive activation from a source to a target language. The second strategy is multilingual learning, whereby all target languages are trained in parallel, utilizing the Connectionist Temporal Classification (CTC) loss for each respective language. This process is further enhanced by incorporating the trace-norm loss among different languages. Embracing multilingualism, the multilingual model emerges as an efficacious strategy to augment the performance of ASR for speech with limited resources. A compelling method to mitigate the dearth of data in low-resource languages is through the integration of multilingual knowledge into a cohesive multilingual end-to-end model.

C. Transfer-Learning Model and Pre-training Model

Transfer learning is a prevalent method for addressing the challenges of low-resource language speech recognition. In [12], data from high-resource languages is pre-trained as the initialization and then fine-tuned on the target language in a low-resource setting.

The paucity of training data for low-resource languages often results in the overfitting problem when this limited dataset is used to retrain models. This issue reduces the efficiency of the transfer-learning method. In response, recent advances have introduced adapt and adjust techniques designed to fine-tune the parameters effectively in multilingual or cross-lingual contexts. These techniques aim to mitigate the overfitting challenges. Furthermore, speech data corpora are employed to pre-train acoustic models. Nonetheless, pre-training necessitates an extensive search space to enhance recognition outcomes, which slows down the decoding speed. To address this, [13] proposed a method that integrates self-training and pre-training, achieving state-of-the-art (SOTA) performance. However, it’s important to note that the fine-tuning, initialized by pre-training, still depends on the language model.

D. Data Augmentation

Data augmentation is a widely adopted strategy to enhance the performance of ASR systems for languages with limited resources. The primary goal of data augmentation is to increase the quantity of training data by generating pseudo labels. For example, in [14] used both unpaired speech and text to train a general ASR model, employing data pairs by synthesizing the missing components before model training. They proposed an alternative training method using prepared speech-PseudoLabel and Synthesized Audio-text pairs, highlighting the complementary nature of both acoustic and linguistic features. Experimental results on Libri-light demonstrated the

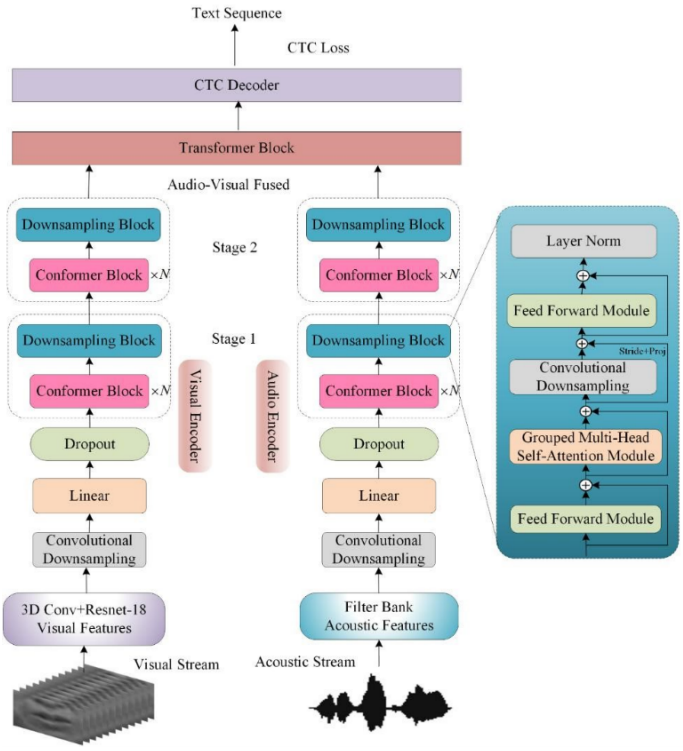

Fig. 1. Overview of AVSR framework (PD-GSAM) for low-resource language speech.

efficacy of joint training as well as two second-round training strategies, validating its superiority over recent models, especially in extreme low-resource cases.

E. Meta-Learning Model

Meta-learning addresses the challenge of rapid adaptation to unseen data, particularly domain mismatch problems in low-resource language speech recognition. In [15] advanced low-resource speech ASR with a Model-Agnostic Meta-Learning method. In [16] noted that multilingual meta-learning offers superior model initialization from various sources, aiding fast target language adaptation. Yet, significant disparities in data scales and learning challenges across languages mean models often favor larger, simpler source languages. Moreover, learning a shared semantic space among languages proves difficult without multilingual pre-training constraints. In summary, prior studies have investigated methods to minimize the dependence on labeled data, including unsupervised and semi-supervised pre-training approaches. However, these research efforts necessitate ample unlabeled data and computational resources. In certain scenarios, access to extensive unlabeled data and computational resources may be limited. Alternative approaches, including transfer learning, multilingual transfer learning, and multilingual meta-learning, aim to acquire general knowledge from high-resource language speech and then adapt it to target low-resource language speech. All conventional techniques rely solely on the acoustic modality. Deviating from this traditional perspective, in this paper, we propose the integration of audio and visual modalities as semantic cues to enhance the model's performance in low-resource language speech recognition tasks.

III. PROPOSED WORK

In this paper, the PD-GSAM framework is proposed for low-resource language speech recognition. PD-GSAM achieves near-linear attention complexity via local group-wise attention with global cross-group fusion, while simultaneously reducing computational costs and accelerating ASR training/inference through strided depthwise convolution for temporal down-sampling in encoders. PD-GSAM is a deep learning block that includes two parts: progressive down-sampling and a grouped self-attention mechanism. The overview of the AVSR framework based on PD-GSAM is shown in Figure 1.

The proposed AVSR framework in Figure 1 includes the input feature extraction of both acoustic and visual streams; acoustic and visual encoders that utilize the same deep learning block (PD-GSAM); a fusion module based on Transformer, and a decoder based on CTC. In this paper, Filter Bank is used as the acoustic feature parameter. The visual features are extracted using 3D ResNet-10. The encoders for both the acoustic and visual streams employ the same deep learning block (PD-GSAM), which consists of a convolutional down-sampling module, linear transformation module, dropout module, Conformer module, and PD module, as shown in Figure 1.

A. Progressive Down-sampling

This paper proposes integrating the PD algorithm into a Conformer-based encoder to reduce Convolutional Neural Network (CNN) computational costs and accelerate ASR-inspired training/inference. The PD-based deep learning block architecture applies acoustic feature down-sampling via a 3×3 convolutional stem (stride=2), followed by two encoder stages with conformer blocks (uniform feature dimensions). Each conformer block combines multi-head self-attention and a convolution module sandwiched between two Feedforward Neural Network (FFN) layers, with post-layer normalization after each block. Sequence down-sampling is performed in the final blocks of both encoder stages, replacing the standard convolution module with a convolutional down-sampling module. This module uses strided depthwise convolution for down-sampling and pointwise convolution (with expansion factor $2 \times d_{out}/d_{in}$ and gated linear unit) for channel adjustment.

PD is designed on the basis of the Conformer architecture. The stride of the strided depthwise convolution module is set to > 1, aiming to achieve down-sampling along the temporal dimension. The encoded sequence is progressively projected to a wider feature dimension, ensuring that the complexity of the hidden layers remains the same for each encoder stage.

B. Grouped Self-Attention Mechanism

While changing the feature dimension of each encoder stage can achieve a similar hidden layer complexity across different encoder stages, the attention complexity is quadratic in sequence length, which introduces computational asymmetry in the network where early attention layers require more multiply-add operations than later layers. To address this problem, in this paper, we use a group self-attention mechanism to calculate the local attention results of each group after transforming the input sequence into metrics. Subsequently, all the results of the grouped attention are linearly transformed and concatenated. Finally, the results of global attention are calculated to capture the relationships among the sequences.

Attention queries, keys, values, and relative position embeddings are reshaped from $Q,\ K,\ V \in \mathbb{R}^{n\times d}$ and $E \in \mathbb{R}^{(2n-g)\times d}$ to $Q^{grp},\ K^{grp},\ V^{grp} \in \mathbb{R}^{\kappa'\times'}$ and $E^{grp} \in \mathbb{R}^{(2n'-1)\times d'}$ where $n' = n/g$ and $d' = d/g$. The output of a head j can be calculated using (1).

$$O_j^{grp} = softmax\left(\frac{O_j^{grp}K_j^{grp^T} + S_j^{rel}}{\sqrt{d'_j}}\right)V_j^{grp} \quad (1)$$

where the grouped attention output $O_j^{grp} \in \mathbb{R}^{n'\times d'}$ is reshaped to $O \in \mathbb{R}^{n\times d}$ before the output projection layer. $S_j^{rel} \in \mathbb{R}^{n\times n}$ is a relative position score matrix that satisfies $S^{rel}[l,m] = Q_l E^T_{m-i}$ with relative position embedding $E = RW^E$. $O \in \mathbb{R}^{n\times d}$ represents a sinusoidal matrix with positions ranging from $-(n_{max}-1)$ to $(n_{max}-1)$. W^E represents the parameter matrix of position embedding.

GSA module is a type of local attention mechanism that can also reflect global attention features. First, the input nodes are divided into several groups and attention is calculated only within that local block. Second, for each group, the summarized queries $Q \in \mathbb{R}^{n\times d}$, keys $K \in \mathbb{R}^{n\times d}$, and values $V \in \mathbb{R}^{n\times d}$ are generated by linear layers to form summarized nodes that represent the summarized information of the nodes within the group. Finally, the global attention output feature O is calculated using (2).

$$O = \alpha(O_j^{grp})_{j=1}^{m} + \beta\overline{O^S} \quad (2)$$

where m represents the number of groups. $\overline{O^S} = \sum_{k=1}^{ls}(O^S)_k$ is concatenated by $\overline{O_j^S} \in \mathbb{R}^{1\times d}$, where O_j^S is pooled by average pooling to form $\overline{O_j^S}$. α and β are learnable parameters for each group that determine the reflection of global output into local attention.

TABLE I
PARAMETERS OF ARCHITECTURE FOR AUDIO STREAM.

Stage	Layers
Fourier Transfer	STFT: 400 window length 160 hop length, 512ffts
Mel Scale	80 mels
Stem	Conv2d: 3×3, 180filters, 2×2 stride
Proj	Linear, 180 units

TABLE II
PARAMETERS OF 3D RESNET-10 ARCHITECTURE.

Stage	Layers
Stem	Conv3d: kernel size $5\times7\times7$,64, $1\times2\times2$ stride MaxPoo3d: $1\times3\times3$, $1\times2\times2$ stride
Res1	$2\times$ [$Conv2d:\ 3\times3,\ 64$filters $Conv2d:\ 3\times3,\ 64$filters]
Res2	$2\times$ [$Conv2d:\ 3\times3,\ 128$filters $Conv2d:\ 3\times3,\ 128$filters]
Res3	$2\times$ [$Conv2d:\ 3\times3,\ 128$filters $Conv2d:\ 3\times3,\ 128$filters]
Res4	$2\times$ [$Conv2d:\ 3\times3,\ 512$filters $Conv2d:\ 3\times3,\ 512$filters]
Pool	Global Average Pooling
Proj	Linear, 256 units

IV. EXPERIMENTAL SETUP

A. Experimental Conditions

Inspired by [17], the main parameters of architecture for dealing with the audio stream are shown in Table I. The main parameters of 3D Resnet-10 for dealing with the visual stream are shown in Table II. Other parameters relate to the GSAM model follow [18].

In addition, the baselines in this paper include deep bidirectional long short-term memory (DBLSTM), Transformer, audio enhanced multi-modality speech recognition (AE-MSR) and Conformer.

B. Data Preparation

In this paper, the International Phonetic Alphabet (IPA) is employed to annotate the pronunciation of the Tujia language speech. This approach is necessitated by the lack of character symbols in the Tujia language. Specifically, the Longshan Tujia language speech comprises 21 initials, which include 6 stop

initials, 4 affricate initials, 5 fricative initials, 3 nasal initials, 1 lateral initial, and 2 zero initials. Additionally, the Longshan Tujia language speech features 25 finals, including 6 unitary finals, 11 complex vowels, and 8 nasalized vowels. The Tujia language corpus comprises a total of 45,541 finals.

In this paper, our self-made Tujia language corpus contains 55 pieces of long text stories, including 5 categories: ethnic origin history, traditional customs, ethnic craftsmanship, village introduction and folk legends. The corpus is recorded by 16 native speakers with a total duration of 10 hours and 16 seconds. The self-made corpus is a typical multimodal corpus with 7,739 pairs of audio-video-text files. The content of Tujia language speech corpus includes 298 core vocabulary words, 2,169 commonly used Tujia words and 7,102 Tujia phrases. We partition the speakers into three subsets: training (70%), validation (15%), and test (15%). Specifically, the speakers was first randomly shuffled, then allocated to each subset according to the predetermined proportions while ensuring no speaker appeared in more than one subset.

Moreover, we randomly chose a part of the LRW and LRS2 corpuses to form the miniLRW and miniLRS2. In particular, both LRW and LRS2 are English audio-visual-text data corpuses. Finally, the miniLRS2 data corpus contains 9 hours, 59 minutes, and 9 seconds, with a total of 10,180 utterances. The miniLRW data corpus contains 9 hours, 59 minutes, and 43 seconds, with a total of 10,254 utterances.

C. Evaluation metrics

The evaluation index is the Character Error Rate (CER) per IPA character, calculated using (3).

$$CER = \frac{S + D + I}{len\,(label)} \tag{3}$$

where S shows the number of characters to be replaced; D shows the number of characters to be deleted; I shows the number of characters to be inserted. The lower the CER, the better the recognition results. In Tujia language evaluation, IPA are used as the primary unit to calculate CER instead of characters.

V. EXPERIMENTAL RESULTS

A. Training Data Increment Results

To evaluate the performance of our proposed method in dealing with low-resource language speech, we designed an incremental training data experiment. In this experiment, we set 1,000, 3,000, 5,000, and 7,000 as the amounts of training data. The experimental results of the approaches areshown in Table III.

Experimental results indicate that all methods improve with increased training data, yet Conformer outperforms AE-MSR due to its superior handling of visual information—critical for Tujia language ASR, wild-environment recordings. Our proposed PD-GSAM further enhances visual processing by replacing the traditional convolutional module with a convolutional down-sampling module, enabling a reduction in computational resource consumption. Notably, PD-GSAM achieves 43.26% CER with only 1,000 training utterances, surpassing the best CER of DBLSTM, Transformer, and AE-MSR (all trained on 7,000 utterances). This highlights PD-GSAM's strong performance in low-resource scenarios with limited data.

TABLE III
CER (%) OF TRAINING DATA INCREMENT IN TUJIA LANGUAGE SPEECH DATA CORPUS.

Approaches	1000	3000	5000	7000
BDLSTM	52.69	52.17	51.58	50.75
Transformer	50.42	49.38	48.25	47.16
AE-MSR	48.53	47.69	46.84	44.11
Conformer	44.17	43.77	42.39	40.14
PD-GSAM	43.26	41.88	39.52	37.93

B. Comparison and Analysis of Different Modalities

To explore how different modalities influence the performance of approaches, we conducted experiments. The experimental results on the miniLRW, miniLRS2, and Tujia language speech corpus are shown in Tables IV. Moreover, since the Tujia Language lacks of character symbols, most evaluation metrics are not applicable to it. Given that we employs IPA to annotate the pronunciation of Tujia language, we adopt the CER as the evaluation metric for this experiment. Additionally, IPA is used as the primary unit for calculating CER instead of characters.

In Tables IV, AO represents audio-only modality; VO represents visual-only modality; AV represents audio-visual fusion modality. The experimental results shows that the CERs of AV are lower than those of AO and VO. In particular, the CERs of AO are lower than those of VO. This result illustrates that although audio-visual fusion-based approaches perform better than single modality-based approaches, acoustic information still contributes more than visual information in low-resource speech recognition tasks. Visual information can be used to enhance acoustic information but should not replace it.

Moreover, the CERs on the Tujia language speech are all higher than those on miniLRW and miniLRS2. The task of low-resource language speech recognition is more difficult than the task of low-resource speech recognition. Besides, visual information has great significance for improving PD-GSAM in the Tujia language speech recognition task.

C. Experimental Results of Ablation Studies

In this paper, we propose PD-GSAM to improve the Conformer encoder for low-resource language speech recognition. To evaluate which part contributes the most in the whole framework, we design an ablation experiment. The details of the ablation experimental results are shown in Table V.

The results in Table V show that the CERs of Conformer+GSAM are lower than those of Conformer+PD. However, the differences between Conformer+PD and Conformer+GSAM are not significant. Therefore, we cannot conclude which part plays a more important role than the other. Nevertheless, Conformer+PD+GSAM performs the best in this experiment.

TABLE IV
CER (%) OF DIFFERENT MODALITY ON DATA CORPUS.

Approaches	miniLRW			miniLRS2			Tujia Language Speech		
	AO	VO	AV	AO	VO	AV	AO	VO	AV
BDLSTM	30.54	66.33	25.51	31.33	66.57	27.21	52.83	79.27	49.58
Transformer	23.22	55.49	20.37	26.34	57.89	24.10	50.67	74.69	45.94
AE-MSR	21.56	52.28	18.57	25.26	54.83	23.06	48.72	70.11	42.12
Conformer	20.25	46.28	16.44	23.59	48.30	18.87	44.23	63.85	38.36
PD-GSAM	18.56	41.12	14.89	19.45	45.91	16.33	43.44	58.65	35.15

TABLE V
ABLATION EXPERIMENTAL RESULTS ON MINILRW, MINILRS2 AND TUJIA LANGUAGE SPEECH.

Approaches	CER(%)		
	mini_LRW	mini_LRS2	Tujia Language Speech
Conformer+PD	19.45	20.03	38.69
Conformer+GSAM	18.55	19.72	38.27
Conformer+PD+GSAM	14.89	16.33	35.15

VI. SUMMARY AND CONCLUSIONS

In this paper, We propose the PD-GSAM AVSR framework to address the overfitting problem in low-resource language speech recognition due to the scarcity of training data. Our proposed method can effectively reduce computational resource consumption with a simpler architecture, which can alleviate the overfitting issue of the AVSR model. However, the performance of our proposed method is still limited by the scarcity of training data. In the future, we plan to explore further ways to improve the performance of AVSR for low-resource language speech recognition using pre-training and fine-tuning frameworks, and exploring the sensitivity of the GSAM to group number, and of PD to its down-sampling strategy. In summary, our research has significance in expanding the application field of AVSR, such as low-resource language speech recognition.

REFERENCES

[1] Qian, Y., Zhou, Z., 2022. Optimizing data usage for low-resource speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30, 394-403. doi: 10.1109/TASLP.2022.3140552.

[2] Toshniwal, S., Sainath, T. N., Weiss, R. J., Li, B., Moreno, P., Weinstein, E., 2018. Multilingual speech recognition with a single end-to-end model. 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 15-20 April, Calgary, AB, Canada. pp. 4904-4908. doi: 10.1109/ICASSP.2018.8461972.

[3] Hu, K., Bruguier, A., Sainath, T.N., Prabhavalkar, R., Pundak, G. 2019. Phoneme-based contextualization for cross-lingual speech recognition in end-to-end models. Interspeech 2019, September 15–19, 2019, Graz, Austria, pp. 2155-2159. doi: 10.21437/Interspeech.2019-1868.

[4] Thomas, S., Audhkhasi, K., Kingsbury, B., 2020. Transliteration Based Data Augmentation for Training Multilingual ASR Acoustic Models in Low Resource Settings. Interspeech 2020, October 25–29, Shanghai, China. pp. 4736-4740. doi: 10.21437/Interspeech.2020-2593.

[5] Hsu, J. Y., Chen, Y. J., Lee, H. 2020. Meta learning for end-to-end low-resource speech recognition. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 204-08 May, Barcelona, Spain. pp. 7844-7848. doi: 10.1109/ICASSP40776.2020.9053112.

[6] Chen, D., Mak, B. K. W., 2015. Multitask learning of deep neural networks for low-resource speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23(7), 1172-1183. doi: 10.1109/TASLP.2015.2422573.

[7] Yu, C., Su, X., Qian, Z., 2023a. Multi-stage audio-visual fusion for Dysarthric speech recognition with pre-trained models. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 31, 1912-1921. doi: 10.1109/TNSRE.2023.3262001.

[8] Yu, C., Yu, J., Qian, Z., Tan, Y., 2023b. Improvement of Acoustic Models Fused with Lip Visual Information for Low-Resource Speech. Sensors, 23(4), 2071. doi: 10.3390/s23042071.

[9] Yu, C., Yu, J., Qian, Z., Tan, Y., 2022. Endangered Tujia language Speech Recognition Research based on Audio-Visual Fusion. Proceedings of the 2022 5th Artificial Intelligence and Cloud Computing Conference. December 17-19, Osaka Japan. pp. 190-195. doi: 10.1145/3582099.358212.

[10] Xu, P., Fung, P., 2013. Cross-lingual language modeling for low-resource speech recognition. IEEE transactions on audio, speech, and language processing, 21(6), 1134-1144. doi: 10.1109/TASL.2013.2244088.

[11] Luo, J., Wang, J., Cheng, N., Zheng, Z., Xiao, J., 2022. Adaptive activation network for low resource multilingual speech recognition. 2022 International Joint Conference on Neural Networks (IJCNN). IEEE, 18-23 July, Padua, Italy. pp. 1-7. doi: 10.1109/IJCNN55064.2022.9892396.

[12] Tong, S., Garner, P. N., Bourlard, H., 2017a. An investigation of deep neural networks for multilingual speech recognition training and adaptation. Interspeech 2017, August 20–24, Stockholm, Sweden. pp. 714-718. doi: 10.21437/Interspeech.2017-1242.

[13] Xu, Q., Baevski, A., Likhomanenko, T., Tomasello, P., Conneau, A., Collobert, R., Synnaeve, G., Auli, M., 2021b. Self-training and pre-training are complementary for speech recognition. 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 06-11 June, Toronto, 10.1109/ICASSP39728.2021.9414641. ON, Canada. pp. 3030-3034. doi:

[14] Du, Y. Q., Zhang, J., Fang, X., Wu, M., Yang, Z., 2023. A semi-supervised complementary joint training approach for low-resource speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31, 3908-3921. doi: 10.1109/TASLP.2023.3313434.

[15] Singh, S., Wang, R., Hou, F., 2022. Improved meta learning for low resource speech recognition. ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 23-27 May, Singapore, Singapore. pp. 4798-4802. doi: 10.1109/ICASSP43922.2022.9746899.

[16] Chen, Y., Yang, X., Zhang, H., Zhang, W., Qu, D., Chen, C., 2024. Meta adversarial learning improves low-resource speech recognition. Computer Speech Language, 84, 101576. doi: 10.1016/j.csl.2023.101576.

[17] Burchi M, Timofte R. Audio-visual efficient conformer for robust speech recognition[C]//Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2023: 2258-2267.

[18] Jung B, Mukuta Y, Harada T. Grouped self-attention mechanism for a memory-efficient Transformer[J]. arXiv preprint arXiv:2210.00440, 2022.