applied sciences

MDPI

Article

Tongan Speech Recognition Based on Layer-Wise Fine-Tuning Transfer Learning and Lexicon Parameter Enhancement

Junhao Geng [1], Dongyao Jia [2,*], Ziqi Li [2], Zihao He [2], Nengkai Wu [2], Weijia Zhang [2] and Rongtao Cui [2]

[1] Beijing Research Institute of Automation for Machinery Industry Co., Ltd., No.1 Jiaochangkou Deshengmenwai, Xicheng District, Beijing 100120, China; gqgjh@126.com
[2] School of Automation and Intelligence, Beijing Jiaotong University, No.3 Shangyuancun, Haidian District, Beijing 100044, China; 21111057@bjtu.edu.cn (Z.L.); 20125029@bjtu.edu.cn (Z.H.); 19111054@bjtu.edu.cn (N.W.); zhangweijia1996yx@163.com (W.Z.); 23140002@bjtu.edu.cn (R.C.)
* Correspondence: donyaojia1974@163.com

Abstract

Speech recognition, as a key driver of artificial intelligence and global communication, has advanced rapidly in major languages, while studies on low-resource languages remain limited. Tongan, a representative Polynesian language, carries significant cultural value. However, Tongan speech recognition faces three main challenges: data scarcity, limited adaptability of transfer learning, and weak dictionary modeling. This study proposes improvements in adaptive transfer learning and NBPE-based dictionary modeling to address these issues. An adaptive transfer learning strategy with layer-wise unfreezing and dynamic learning rate adjustment is introduced, enabling effective adaptation of pretrained models to the target language while improving accuracy and efficiency. In addition, the MEA-AGA is developed by combining the Mind Evolutionary Algorithm (MEA) with the Adaptive Genetic Algorithm (AGA) to optimize the number of byte-pair encoding (NBPE) parameters, thereby enhancing recognition accuracy and speed. The collected Tongan speech data were expanded and preprocessed, after which the experiments were conducted on an NVIDIA RTX 4070 GPU (16 GB) using CUDA 11.8 under the Ubuntu 18.04 operating system. Experimental results show that the proposed method achieved a word error rate (WER) of 26.18% and a word-per-second (WPS) rate of 68, demonstrating clear advantages over baseline methods and confirming its effectiveness for low-resource language applications. Although the proposed approach demonstrates promising performance, this study is still limited by the relatively small corpus size and the early stage of research exploration. Future work will focus on expanding the dataset, refining adaptive transfer strategies, and enhancing cross-lingual generalization to further improve the robustness and scalability of the model.

Keywords: Tongan; low-resource speech recognition; transfer learning; layer-wise fine-tuning; dictionary optimization

check for updates

Academic Editors: Rajvardhan Patil and Venkat N. Gudivada

Received: 2 October 2025
Revised: 21 October 2025
Accepted: 21 October 2025
Published: 24 October 2025

Citation: Geng, J.; Jia, D.; Li, Z.; He, Z.; Wu, N.; Zhang, W.; Cui, R. Tongan Speech Recognition Based on Layer-Wise Fine-Tuning Transfer Learning and Lexicon Parameter Enhancement. *Appl. Sci.* **2025**, *15*, 11412. https://doi.org/10.3390/app152111412

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Speech recognition, as a core technology of artificial intelligence, is widely applied in security, education, healthcare, and defense, enhancing work efficiency and facilitating global communication. Within the global linguistic landscape, Tongan is a representative low-resource language. Although spoken by a relatively small population, it embodies rich cultural traditions and historical heritage, giving it significant value for cultural preservation and transmission. As vital channels for intercultural communication, low-resource

languages have gained growing importance in global research and cultural exchange. Tongan, which integrates indigenous traditions with Western influences, represents a unique linguistic system that deserves greater attention in computational linguistics. Consequently, advancing research on Tongan speech recognition is not only academically significant but also contributes to the digital preservation and international dissemination of its linguistic resources.

In recent years, deep learning has greatly advanced speech recognition for resource-rich languages such as English and Chinese, thereby facilitating international linguistic communication [1]. In contrast, low-resource languages like Tongan still encounter major challenges, including data scarcity, complex dialectal variation, and significant cultural-contextual differences. Globally, more than 7000 languages exist, the majority of which are low-resource [2], and they generally lack sufficient annotated corpora as well as lexical and grammatical resources. For example, while high-resource languages such as English or Mandarin typically have access to over 10,000 h of labeled speech data, many low-resource languages possess fewer than 100 h of transcribed material, which severely limits model training and generalization. This deficiency severely constrains the effectiveness of conventional speech recognition systems. Furthermore, low-resource languages often encode unique cultural knowledge, while their dialectal diversity and expressive variability increase the difficulty of model training. Traditional methods based on Gaussian Mixture Models (GMMs), Hidden Markov Models (HMMs), or Convolutional Neural Networks (CNNs) provide limited modeling capacity for such languages and struggle to capture complex acoustic features and fine-grained semantic patterns. In addition, extreme data sparsity makes it difficult to train deep models directly, as this frequently leads to overfitting and hinders further improvements in recognition performance.

Against this backdrop, transfer learning and end-to-end modeling emerge as promising strategies for low-resource language speech recognition. Transfer learning leverages pretrained models on high-resource languages and applies adaptive fine-tuning to capture phonetic features of the target language, thereby improving modeling efficiency and recognition performance [3]. This approach not only mitigates the shortage of annotated data but also enhances adaptability to linguistic variation. End-to-end systems, in turn, adopt unified neural architectures to directly map speech signals to text, simplifying the traditional pipeline of feature extraction, acoustic modeling, and language modeling [4]. Such frameworks achieve higher recognition accuracy and exhibit strong robustness to dialectal and pronunciation variations common in low-resource languages. Recent advances in neural architectures, such as Transformer and DeepSpeech, have further expanded opportunities in this field [5]. Despite these advances, models are prone to overfitting during transfer due to the scarcity of labeled data, while existing strategies often rely on fixed, coarse-grained parameter adjustments that lack fine-grained and dynamic adaptation across layers and linguistic features. Moreover, within end-to-end frameworks, the design of sub-word units and construction of efficient, expressive vocabularies remain critical obstacles—particularly for languages with complex structures and heterogeneous pronunciations. Therefore, developing targeted optimizations in transfer mechanisms and dictionary modeling is essential to further enhance the accuracy and robustness of speech recognition systems for Tongan and other low-resource languages.

To address these challenges, this study introduces three key innovations, focusing on data augmentation and sampling, cross-lingual transfer strategy, and dictionary parameter optimization for low-resource languages. First, to mitigate the issues of limited corpus size and imbalanced class distribution, a Supervised Random Augmentation with Deep Random Forest Filtering (SRA-DRF) algorithm is developed. This method integrates generative adversarial networks (GAN) with signal processing techniques to generate synthetic speech

data and applies similarity-based filtering to retain high-quality samples. In addition, a weighted K-means stratified sampling strategy ensures balanced class representation, thereby enhancing feature learning and model generalization. Second, in the transfer learning stage, an adaptive layer-wise unfreezing strategy is proposed. This approach dynamically adjusts learning rates and performs top-down fine-tuning, preserving low-level general features while improving adaptability to target language characteristics. The effectiveness of this strategy is further validated through Centered Kernel Alignment (CKA) visualization. Finally, for dictionary modeling, a hybrid MEA-AGA—combining the Mind Evolutionary Algorithm with the Adaptive Genetic Algorithm—is employed to automatically search for the optimal number of byte-pair encoding (NBPE) units, thereby improving recognition accuracy and decoding efficiency.

This paper is organized into seven sections as follows: Section 2 presents related work, reviewing the current research status of low-resource speech recognition, with a focus on challenges in data scarcity, transfer learning optimization, and dictionary parameter tuning, and briefly outlines the proposed strategies in this study. Section 3 details the proposed methods, including the design of the layer-wise adaptive transfer learning strategy and the MEA-AGA for dictionary parameter optimization. Section 4 describes the experimental setup, covering dataset sources, model configurations, and evaluation metrics. Section 5 reports the experimental results, including data augmentation, transfer learning, and dictionary optimization experiments, along with comparative analysis against mainstream methods to validate the effectiveness of the proposed approaches. Section 6 summarizes the overall contributions of this work in light of the experimental findings. Section 7 outlines future directions, addressing the current limitations and proposing potential areas for further exploration.

2. Related Work

2.1. Research Challenges

While deep learning has led to remarkable achievements in speech recognition for high-resource languages, its effectiveness in low-resource languages remains limited. These limitations mainly arise from three key aspects: data scarcity and corpus imbalance, model training strategy, and dictionary-based modeling optimization.

(1) Data scarcity and corpus imbalance

Low-resource languages often suffer from limited annotated data and a lack of publicly available corpora, resulting in a notable performance gap between their speech recognition systems and those of high-resource languages [6]. This scarcity poses significant challenges for effective modeling. In recent years, end-to-end speech recognition models have gained popularity due to their streamlined architectures and strong modeling capabilities. Researchers have begun applying these models to low-resource language tasks and have achieved preliminary progress.

Mukhamadiyev et al. [7] introduced a hybrid DNN-HMM and CTC-attention model for Uzbek speech recognition, achieving a WER of 14.3% on a medium-scale dataset. Kamper et al. [8] proposed a CNN-DTW-based keyword spotting system for Luganda, leveraging CAE features and multilingual bottleneck features, yielding over 27% performance gains under low-resource settings. Mamyrbayev et al. [9] developed a real-time recognition system for Kazakh using RNN-T, achieving a CER of 10.6% without relying on a language model, outperforming existing end-to-end approaches. Zhi et al. [10] conducted preliminary research on Mongolian using the ESPnet framework. Oukas et al. [11] built an end-to-end Arabic speech recognition model based on the Common Voice dataset, showing better performance with DeepSpeech compared to Wav2Vec. Changrampadi et al. [12] built a Tamil recognition system using DeepSpeech, expanding training data

with online resources and semi-supervised techniques, and achieved a WER of 24.7%, surpassing Google's results. Shetty et al. [13] incorporated language identity tokens and acoustic embeddings into a Transformer for Indian languages, achieving a 6–11% relative CER reduction compared to monolingual baselines. Pan et al. [14] introduced subword modeling and initialization strategies to improve recognition efficiency and convergence for Tibetan. Anoop et al. [15] developed the first Sanskrit speech processing system, using Mel-spectrogram augmentation and WFST decoding to achieve a 7.64% WER under ultra-low-resource conditions, demonstrating the feasibility of end-to-end architectures in such challenging scenarios.

Despite the promising potential of end-to-end approaches for various low-resource languages, several challenges remain. The scarcity of training data leads to overfitting and poor generalization while the large parameterization of end-to-end models makes convergence difficult and susceptible to local optima. Therefore, enhancing training efficiency and robustness remains a key focus in low-resource speech recognition research.

(2) Model training strategy

To address these challenges, transfer learning offers an effective solution. By leveraging knowledge acquired from source tasks, it improves learning efficiency and performance in target domains, substantially reducing dependence on large-scale labeled data. Transfer learning has shown notable advantages in low-resource settings and has achieved success in diverse applications such as medical diagnosis, image recognition, and industrial inspection [16–19]. As a result, it has become a widely adopted strategy in low-resource speech recognition research.

In low-resource speech recognition, fine-tuning pre-trained models remains the most commonly adopted transfer strategy. Mainzinger et al. [20] proposed adapter-based fine-tuning on Mvskoke, which significantly improved training efficiency and recognition accuracy, demonstrating strong parameter efficiency under limited data. Abdullah et al. [21] fine-tuned the Whisper model on Northern Kurdish using various parameter adaptation techniques, showing that adding auxiliary modules led to a notable WER reduction to 10.5%. Cui et al. [22] applied a multilingual, multi-task weakly supervised Whisper decoder to Burmese ASR, achieving a 5% absolute reduction in character error rate. Zeng et al. [23] employed transfer learning using a hybrid Transformer-LSTM structure for Malay ASR and introduced text augmentation strategies, achieving over 60% relative reduction in WER under low-resource conditions. Hjortnaes et al. [24] adopted transfer learning with the DeepSpeech model for Komi recognition and integrated a KenLM-based language model for output correction, resulting in more than 25% and 20% relative improvements in CER and WER, respectively. Bansal et al. [25] improved Mboshi speech-to-text translation using high-resource pretraining, achieving promising results with just four hours of data. Lee et al. [26] investigated cross-lingual transfer to predict phrase breaks in multilingual models, aiming for cost-effective adaptation. Baller et al. [27] enhanced RNN robustness to disfluent speech in low-resource languages such as Tunjia via transfer learning. Anoop et al. [28] introduced a MAML-based pretraining framework that outperformed multilingual joint training in rapidly adapting to Indian low-resource languages.

In recent years, research on transfer learning has further expanded toward multilingual and zero-shot paradigms, both of which aim to enhance model generalization under data-scarce conditions. In the multilingual direction, Yadav et al. [29] surveyed multilingual ASR models focusing on cross-lingual transfer and self-supervised learning, demonstrating that joint multilingual training can effectively improve recognition for low-resource languages. Yu et al. [30] proposed Master-ASR, a modular multilingual ASR framework that achieves scalable cross-lingual adaptation and robust low-resource performance through a modularize-then-assemble learning strategy, reducing inference overhead by 30% while

improving character error rate by up to 2.41 over state-of-the-art models. For zero-shot learning, Yan et al. [31] introduced a transliteration-based framework for zero-shot code-switched ASR, which eliminates explicit language segmentation and leverages bilingual decoding with external language models to achieve effective recognition on Mandarin–English SEAME datasets. Another study [32] proposed a universal phonetic model that decomposes phonemes into phonetic attributes (e.g., vowels and consonants), enabling recognition for languages with no audio data and achieving an average 9.9% improvement in phone error rate across 20 unseen languages.

Transfer learning has shown considerable promise in mitigating the performance limitations of low-resource speech recognition. Nevertheless, most existing methods still rely on static fine-tuning strategies, which lack dynamic perception and adaptive adjustment mechanisms across different model layers and linguistic structures. In addition, the limited interpretability of current transfer mechanisms further constrains their optimization and wider applicability. These issues remain critical challenges in advancing low-resource speech recognition research.

(3) dictionary-based modeling optimization

In addition, dictionary-based modeling plays an essential role in improving recognition accuracy and has demonstrated strong adaptability across diverse languages. By optimizing the NBPE, recognition accuracy and output efficiency can be effectively balanced.

Hsiao et al. [33] explored the impact of sub-word vocabulary size on recognition performance using a VQ-BPE modeling approach for English and Chinese speech recognition. Their findings showed that setting the NBPE to 8000 led to accuracy improvements of 5.8% and 3.7% for English and Chinese, respectively, validating the effectiveness of dictionary granularity optimization. Deng et al. [34] proposed a Byte-Level BPE method that incorporates length and character penalty mechanisms. Trained on a 5k-hour bilingual corpus, the model achieved performance comparable to character-level models while reducing parameter size. Ma Jian et al. [35] introduced a BPE-dropout approach with sub-word length thresholds, which significantly reduced WER on THCHS30 and ST-CMDS datasets and improved recognition of rare and out-of-vocabulary words. In low-resource language studies, sub-word modeling strategies have also shown strong potential. Cai Yuqing et al. [36] addressed unit inconsistency in Tibetan ASR by combining BPE with a hybrid CTC-Attention framework. By adjusting the NBPE, the system achieved a WER of 9.0%, outperforming models based on syllables or components. Gaurav et al. [37] demonstrated that NBPE size significantly impacts recognition accuracy in Indian multilingual tasks, with a 7.5 k vocabulary yielding the best results. Turi et al. [38] developed an end-to-end ASR system for the Oromo language using a BPE tokenizer. Experiments showed that setting the NBPE to 500 achieved optimal performance, highlighting the value of sub-word unit tuning. Patel et al. [39] examined telephone speech recognition for Indian languages using a Conformer-based model, comparing character-level and various BPE granularities. The results indicated that a properly selected NBPE significantly improves accuracy, achieving a WER of 30.3% without external language model support, surpassing baseline methods.

Overall, NBPE, as a key parameter in sub-word modeling, exerts a significant influence on both the accuracy and efficiency of speech recognition systems. Although prior studies have demonstrated the benefits of NBPE optimization, most methods still depend on manual, experience-driven configurations. Systematic approaches for automated tuning and cross-lingual adaptation remain insufficiently explored and merit further investigation.

2.2. Proposed Method

In summary, low-resource speech recognition continues to face major challenges, including limited training data, insufficient adaptability in transfer learning, and the lack

of systematic optimization in dictionary parameter tuning. To address these issues, this paper proposes the following improvements:

Data Augmentation: To mitigate the limitations of small-scale corpora, SRA-DRF algorithm is introduced. This method integrates GAN-based generation with signal processing techniques to produce high-quality synthetic speech, while similarity-based filtering ensures the reliability of augmented samples. In addition, to alleviate training bias caused by class imbalance, a weighted K-Means stratified sampling strategy is applied. This guarantees balanced class distributions across training, validation, and test sets, thereby improving model robustness and representation of diverse speech patterns.

Transfer Learning Optimization: To enhance the adaptability of pretrained models for low-resource tasks, a layer-wise adaptive transfer learning strategy is developed. This approach preserves general low-level acoustic features while progressively fine-tuning higher layers, guided by both the loss function trajectory and phonetic similarity between source and target languages. Moreover, Centered Kernel Alignment (CKA) is employed to visualize feature alignment across layers, providing intuitive evidence of effective knowledge transfer.

Dictionary Parameter: Optimization: To overcome the lack of systematic sub-word granularity tuning, a novel MEA-AGA hybrid optimization algorithm is proposed, combining the Mind Evolutionary Algorithm with the Adaptive Genetic Algorithm. This approach automatically searches for the optimal NBPE, improving sub-word modeling accuracy without compromising decoding efficiency, and offering a more adaptive and stable strategy for low-resource language modeling.

3. Methodology

To provide a clear overview of the proposed framework, Figure 1 presents the overall model pipeline designed for low-resource speech recognition. The process begins with raw audio waveforms, which are subjected to data augmentation and preprocessing to expand the effective training corpus. The augmented data are then utilized within an adaptive transfer learning framework, where the pretrained Mixformer model is progressively fine-tuned to capture language-specific phonetic representations. Subsequently, the MEA-AGA is applied to optimize the NBPE parameters for dictionary modeling, enabling efficient subword decoding and accurate text generation.

3.1. Data Augmentation and Partitioning

In low-resource speech recognition tasks, data scarcity and class imbalance remain major factors that hinder model performance [40]. To improve training effectiveness and recognition accuracy, this section focuses on preprocessing and augmenting the original Tongan speech corpus, with the goal of constructing a balanced, diverse, and high-quality training dataset.

Data augmentation techniques are generally divided into signal processing-based and deep learning-based approaches. The former, such as speed perturbation, pitch shifting, and noise addition [41], are easy to implement but often introduce pronunciation distortions and fail to preserve language-specific features in low-resource settings. The latter, represented by Generative Adversarial Networks [42–44], generate higher-quality data but demand large datasets and significant computational resources, limiting their use in low-resource languages. To overcome these limitations, this study proposes an SRA-DRF algorithm that integrates signal processing techniques with deep learning methods. This hybrid strategy not only increases data diversity but also reduces reliance on large datasets and computational resources, thereby enabling effective augmentation of Tongan speech data. The overall workflow is illustrated in Figure 2.

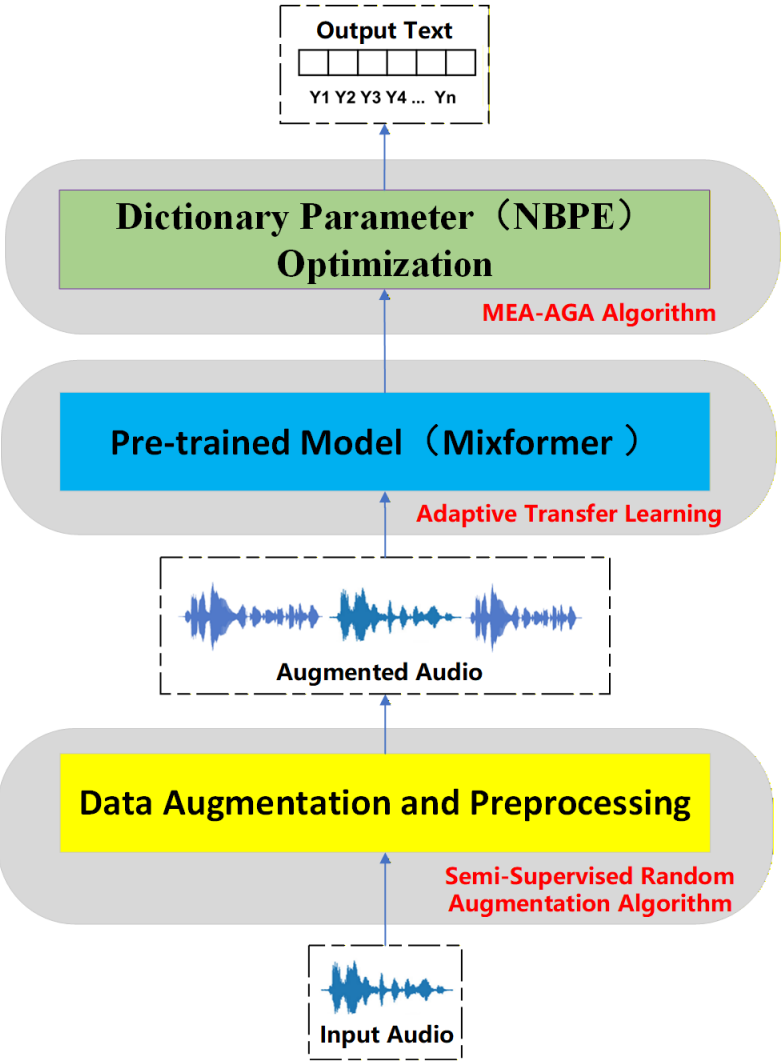


Figure 1. The model pipeline of the proposed low-resource speech recognition system.

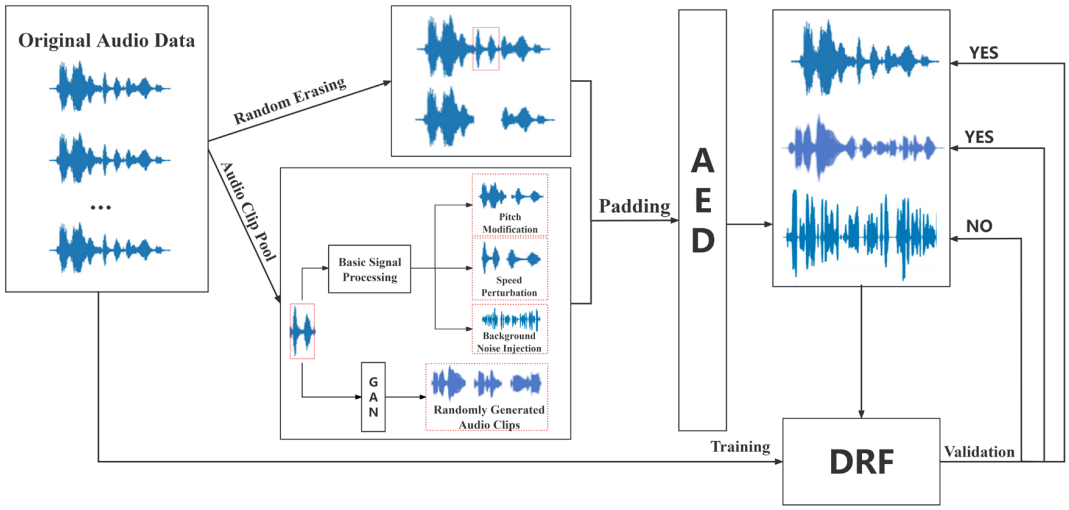


Figure 2. Framework of the SRA-DRF (Supervised Random Augmentation with Deep Random Forest Filtering) algorithm.

Let C denote the original Tongan audio feature matrix. Random segments are removed from C, and the erased time spans are stored in a separate matrix referred to as the audio

segment pool. The residual features, after the erasure process, are preserved in a new matrix denoted as C_R.

$$Y = C \odot M_{erase} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1N} \\ y_{21} & y_{21} & \cdots & y_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ y_{M1} & y_{M2} & \cdots & y_{MN} \end{bmatrix}, Y \in \mathbb{R}^{M \times N} \tag{1}$$

$$y_{ij} = \begin{cases} x_{ij} & m_{ij} = 1 \\ 0 & m_{ij} = 0 \end{cases} \tag{2}$$

$$C_R = C \odot (1 - M_{erase}),\ C_R \in \mathbb{R}^{M \times N} \tag{3}$$

M_{erase} is a masking matrix composed of 0 and 1, having the same dimensionality as Y. A value of 0 indicates that the corresponding data point is masked, while 1 means it is retained. y_{ij} denotes the amplitude of the $j-th$ frame of the $i-th$ utterance after the original data has been masked.

In the audio segment pool, signal processing and deep learning methods are jointly employed. On the one hand, the extracted audio segments are augmented through speed perturbation, pitch shifting, and noise addition. On the other hand, a GAN-based network is applied to generate additional realistic audio segments. The corresponding formulation is defined as follows:

$$Y_{PRO} = \begin{cases} \mathcal{F}(\mathcal{F}^{-1}(Y) \cdot a), & \text{if speed perturbation} \\ Y \cdot e^{2\pi i b f}, & \text{if pitch shifting} \\ Y + \gamma N, & \text{if noise addition} \end{cases} \tag{4}$$

$$Y_{GAN} = G(Z; \theta; Y), Y_{GAN} \in \mathbb{R}^{M \times N} \tag{5}$$

where a is the speed perturbation factor applied to modify speaking rate, b is the pitch shift factor for adjusting vocal pitch. N denotes the noise matrix that follows a Gaussian distribution, and γ is the noise strength coefficient. G is the generator in the adversarial network; Z represents the noise input vector, which is randomly sampled and fed into the generator; θ denotes the generator's parameters which control the properties of the generated audio. Y_{PRO} is the processed audio segment using signal-based operations on C_R, while Y_{GAN} represents the audio segment generated by the GAN. These two types of augmented data are then combined into a unified audio segment pool Y_{POOL}:

$$Y_{POOL} = Y_{PRO} + Y_{GAN}, Y_{POOL} \in \mathbb{R}^{M \times N} \tag{6}$$

Next, several candidate segments from the audio segment pool Y_{POOL}, which stores the erased portions of the original audio, are aligned with the residual speech features C_R. An attention-based encoder–decoder (AED) module is then applied to integrate these segments, producing a supplementary data matrix C':

$$C' = C_R + \text{softmax}(C_R, Y_{POOL}) = C_R + \text{softmax}\left(\frac{C_R + Y_{POOL}^T}{\sqrt{N}}\right) \tag{7}$$

Finally, the original Tongan corpus is utilized to train a DRF model. The augmented dataset is subsequently input to the model for evaluation, with the metric defined in Equation (8) adopted as the evaluation criterion.

$$score = \text{sim}(C, C') = \frac{C \cdot C'}{\|C\|\|C'\|} \tag{8}$$

Audio segments with higher scores are classified as Yes and retained as valid augmentations, while the rest are discarded.

After augmenting the speech data, the dataset must be partitioned in an appropriate manner to enhance the model's generalization capability. A commonly adopted approach is K-Means clustering, which facilitates balanced class distribution [45–47]. Its mathematical formulation is defined as follows:

$$SSE = \sum_{i=1}^{k} \sum_{x_j \in C_i} \|x_j - \mu_i\|^2 \tag{9}$$

Here, C_i denotes the $i-th$ cluster, x_j represents the sample in the cluster, and μ_i is the cluster centroid.

This method performs well when clustering data with balanced class distributions. However, for tasks involving low-resource languages such as Tongan, the small dataset size and class imbalance may result in biased cluster centers. Direct application of traditional clustering often leads to convergence issues, misalignment between training and test sets, and reduced model stability and generalization capability [48].

To overcome these limitations, we modify the objective function by introducing a class-aware weighting term, which balances the influence of minority-class samples during clustering. The revised formulation is defined as:

$$SSE' = \sum_{i=1}^{k} \sum_{x_j \in C_i} \omega_j \|x_j - \mu_i\|^2 \tag{10}$$

The weight ω_j is used to adjust the influence of each sample in the cluster based on its class distribution, and is defined as:

$$\omega_j = \frac{1}{frequency(c(x_j))} \tag{11}$$

where $frequency(c(x_j))$ denotes the frequency of the class label $c(x_j)$, to which sample x_j belongs in the dataset. The detailed definition is:

$$frequency(c(x_j)) = \frac{n_{c(x_j)}}{N} \tag{12}$$

Here, N is the total number of samples in the dataset, and $n_{c(x_j)}$ is the number of samples belonging to class $c(x_j)$, which is calculated as:

$$n_{c(x_j)} = \sum_{x_i \in D} \mathbb{I}(c(x_i) = c(x_j)) \tag{13}$$

$\mathbb{I}(\cdot)$ is the indicator function, which equals 1 if the condition is satisfied, and 0 otherwise. D denotes the full dataset. Finally, by substituting the above expressions, the improved K-Means clustering objective function is given as:

$$SSE' = \sum_{i=1}^{k} \sum_{x_j \in C_i} \frac{N}{\sum_{x_i \in D} \mathbb{I}(c(x_i) = c(x_j))} \left\| x_j - \mu_i \right\|^2 \tag{14}$$

To achieve more balanced data partitioning, this study integrates stratified sampling with the improved weighted K-Means clustering method. Specifically, samples are first stratified according to their distances from the cluster centroids, and proportional sampling is then applied within each cluster to ensure balanced class representation across the training, validation, and test sets. The overall procedure of this data partitioning method is illustrated in Figure 3.

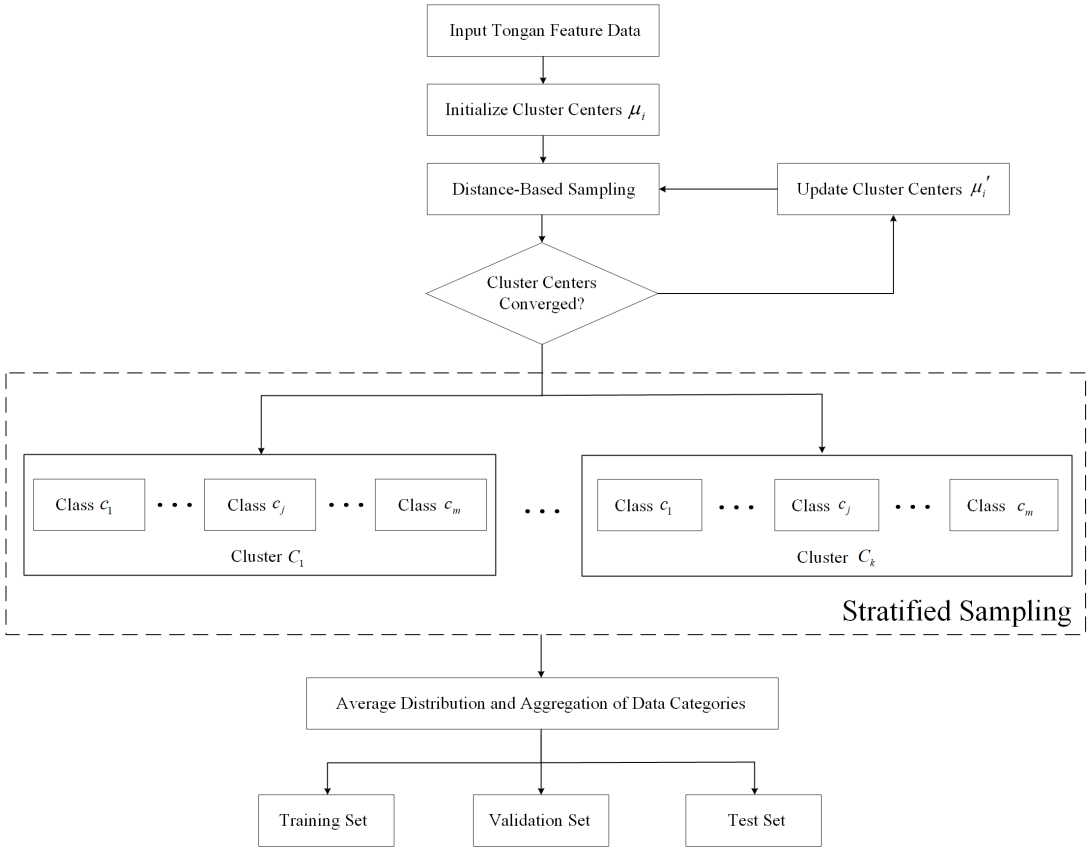


Figure 3. Tongan data partitioning flowchart based on weighted stratified sampling.

The basic steps of the algorithm are as follows:

(1) Based on the features extracted from the input Tongan speech data, perform statistical analysis to initialize k cluster centers μ_i (where $i = 1, 2, \ldots, k$ and k is the number of clusters).

(2) Calculate the weighted distance from each sample to each cluster center and assign the sample to the nearest cluster. For a given sample x_j and cluster center C_i, the distance is computed according to Equation (14).

(3) Update each cluster center based on the current cluster members and their weights. The new cluster center μ_i' is computed as:

$$\mu_i' = \frac{\sum_{x_j \in C_i} \omega_j x_j}{\sum_{x_j \in C_i} \omega_j} \tag{15}$$

(4) Check whether the new cluster centers meet the convergence condition. If convergence is achieved, the algorithm terminates; otherwise, return to step (2) for further updates.

(5) After clustering, compute the proportion of each class within each cluster:

$$p_{ij} = \frac{n_{ij}}{n_i} \tag{16}$$

Here, p_{ij}, n_{ij} and n_i represent, respectively, the proportion, the number of samples, and the total number of samples of class c_j within cluster C_i.

(6) Finally, stratified sampling is performed within each cluster based on the class distribution and the specified data split ratios. The total number of samples selected from each cluster is then used to construct the training, validation, and test sets. The formulas are defined as follows:

$$N^T = \sum_{i=1}^{k} \sum_{j=1}^{m} p_{ij} \times n_{ij}^T \tag{17}$$

$$N^V = \sum_{i=1}^{k} \sum_{j=1}^{m} p_{ij} \times n_{ij}^V \tag{18}$$

$$N^T = \sum_{i=1}^{k} \sum_{j=1}^{m} p_{ij} \times n_{ij}^S \tag{19}$$

Here, N^T, N^V and N^S represent the total number of training, validation, and test samples, respectively. k is the number of clusters, m is the number of classes, and n_{ij}^T, n_{ij}^V and n_{ij}^S represent the number of training, validation, and test samples from class c_j within cluster C_i.

3.2. Layer-Wise Unfreezing for Adaptive Transfer Learning

Transfer learning, which adapts knowledge from a source domain to a target task, has proven particularly effective in low-resource or highly specialized scenarios and has therefore attracted considerable academic attention [49]. However, for large-scale models such as Transformer, Wav2Vec, and GPT, significant challenges remain, including the risk of negative transfer and high computational costs. To address these issues, this study proposes an adaptive transfer learning approach based on layer-wise unfreezing. The overall network architecture and algorithmic workflow are illustrated in the figure below.

The model architecture adopted in this study incorporates a fine-grained feedback mechanism with a clear distinction between frozen and tunable layers, thereby improving the precision and efficiency of transfer learning. The unfreezing process follows a top-down strategy, where the number of trainable layers and the corresponding learning rates are initialized according to cross-lingual similarity. This design preserves pre-trained knowledge while gradually adapting the model to the target task. The overall architecture of the proposed adaptive transfer learning network is illustrated in Figure 4. As training progresses, a loss-driven adjustment strategy is applied to dynamically refine the unfreezing process, enabling more adaptive and task-specific optimization. The detailed training procedure is summarized in Algorithm 1.

Algorithm 1: Adaptive layer-wise unfreezing and learning rate scheduling for transfer learning

Require: M_{pre}, n, η_0, α_0, β_0, L
Ensure: M_{trans}

$M_{trans} \Leftarrow M_{pre}$; $\eta \Leftarrow \eta_0$; $L \Leftarrow 0$;
for t = 1 to n **do**
    /* Stage 1: Similarity and Loss Calculation */
    Compute sim_{score} and training loss L_{total} (t);
    /* Stage 2: Layer Unfreezing Decision */
    if t == 1 **then**
        $L \Leftarrow \text{floor}[N \cdot (1 - sim_{score})]$;
    else
        L $\Leftarrow$ L + sign(L_{total} $(t) - L_{total}$ $(t\text{-}1)) \times \Delta L$;
    end if
    /* Stage 3: Learning Rate Adjustment */
    if unfreeze **then**
        $\eta \Leftarrow \eta_0 \cdot \exp(-\alpha(t) \cdot sim_{score} - \beta(t) \cdot L_{total}(t))$;
    end if
    $\alpha(t) \Leftarrow \alpha_0 \cdot \exp(-\gamma \cdot t)$; $\beta(t) \Leftarrow \beta_0 \cdot (1 - \exp(-\gamma \cdot t))$;
    /* Stage 4: Layer-Wise Training */
    Freeze layers < L; unfreeze layers $\geq L$;
    Train M_{trans} with η;
    /* Stage 5: Termination Check */
    if stopping criterion met **then**
        break;
    end if
end for

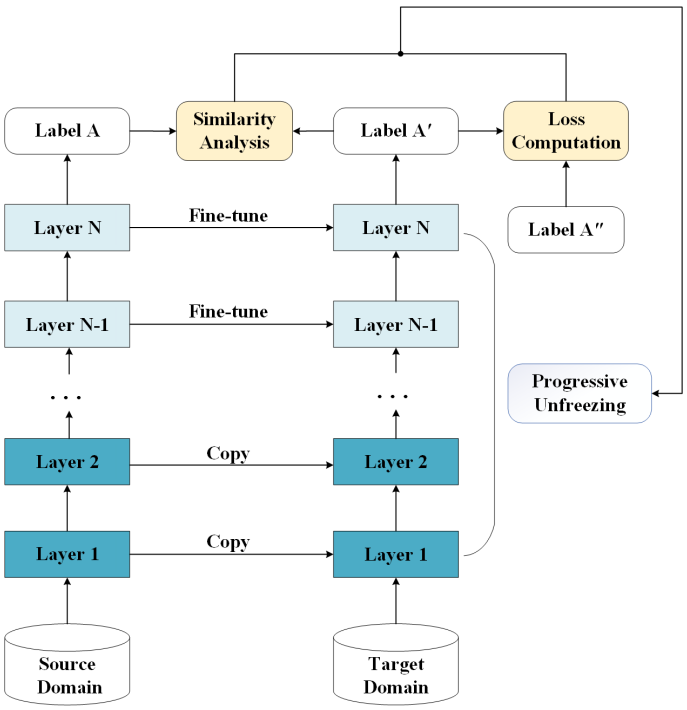


Figure 4. Layer-wise unfreezing adaptive transfer learning network architecture.

The specific steps of the proposed algorithm are as follows:

(1) Model Initialization: This includes setting the maximum number of iterations n, the pretrained model M_{pre}, the transfer model M_{trans}, the number of model layers N, and the initial learning rate η_0.

(2) Similarity and Loss Evaluation: The similarity between source and target languages is calculated using Cosine similarity, denoted as sim_{score}. Simultaneously, the overall training loss L_{total} is computed to guide the subsequent unfreezing strategy.

(3) Determining the Number of Unfrozen Layers: The initial number of unfrozen layers L_0 is determined based on the computed similarity score. The definition and iterative update of the unfreezing depth are formulated as follows:

$$L_0 = \lceil N \cdot (1 - sim_{score}) \rceil \tag{20}$$

$$L_{t+1} = L_t + sign(\Delta L_{total}(t) - \Delta L_{total}(t-1)) \times \Delta L \tag{21}$$

where $\Delta L_{total}(t)$ and $\Delta L_{total}(t-1)$ represent the change in loss values at iterations t and $t-1$, respectively, and the step size ΔL is set to 1.

(4) Learning Rate Adjustment: The learning rate η is dynamically adjusted based on similarity and loss:

$$\eta = \begin{cases} \eta_0 \cdot e^{-\alpha(t) \cdot sim_{score} - \beta(t) \cdot L_{total}}, & if\ unfrozen\ layer \\ 0, & if\ frozen\ layer \end{cases} \tag{22}$$

where $\alpha(t)$ and $\beta(t)$ are time-dependent weighting coefficients that control the influence of similarity and loss, respectively. At the early stage of training, $\alpha(t)$ is set relatively high and decreases gradually, while $\beta(t)$ starts low and increases gradually as training progresses.

$$\alpha(t) = \alpha_0 \cdot e^{-\gamma \cdot t}, \beta(t) = \beta_0 \cdot (1 - e^{-\gamma \cdot t}) \tag{23}$$

where γ is a positive constant (set to 0.1) to control the rate of decay and growth. The initial values are $\alpha_0 = 0.9$ and $\beta_0 = 0.1$.

(5) Unfreezing Execution: Based on the determined unfreezing depth, the model is incrementally unfrozen from the top layer downward. The learning rate is applied accordingly to fine-tune each layer.

(6) Convergence Check: The process checks whether convergence criteria are met. If not, the algorithm returns to step (2) for further updates.

(7) Final Output: Once convergence is achieved, transfer training is completed, and the final target language recognition model is obtained.

To further assess the feature changes during the transfer process, this study employs the CKA similarity matrix to quantitatively analyze key layers of the network before and after fine-tuning. CKA measures the alignment of representations in kernel space across different layers or stages and is widely used in neural network visualization and interpretability research [50]. By comparing similarity scores across layers, the analysis reveals how hierarchical features are adjusted during transfer, providing theoretical support for the effectiveness of the layer-wise unfreezing strategy.

3.3. Dictionary Parameter Optimization Driven by MEA-AGA

In automatic speech recognition, dictionary construction is a critical component, typically achieved through sub-word unit segmentation to standardize output. Among these, NBPE serves as a core parameter that directly impacts vocabulary coverage, sequence length, and model complexity. A small NBPE can increase coverage and reduce annotation

cost, but it leads to longer sequences and greater training difficulty. Conversely, a large NBPE may simplify training but could weaken the model's linguistic expressiveness [36]. In low-resource scenarios such as Tongan, optimizing NBPE is crucial to improving recognition performance. To this end, this study proposes the MEA-AGA method to determine the optimal dictionary configuration for Tongan.

The genetic algorithm (GA) is an optimization method inspired by natural evolution, which searches for optimal solutions through selection, crossover, and mutation operations. However, traditional GA uses fixed parameter settings, making it prone to local optima and lacking adaptability. The adaptive genetic algorithm (AGA) addresses this by introducing dynamic parameter adjustment mechanisms, allowing the crossover and mutation rates to evolve throughout the search process, thereby enhancing global exploration and algorithm robustness. The functions are defined as follows:

$$P_c = \begin{cases} \frac{k_1(f_{\max} - f')}{f_{\max} - f_{avg}}, & f' \geq f_{avg} \\ k_2, & f' < f_{avg} \end{cases} \tag{24}$$

$$P_m = \begin{cases} \frac{k_3(f_{\max} - f)}{f_{\max} - f_{\text{avg}}}, & f \geq f_{\text{avg}} \\ k_4, & f < f_{\text{avg}} \end{cases} \tag{25}$$

In the equations, k_1, k_2, k_3 and k_4 are constants between 0 and 1. $f_{\max}$ and f_{avg} denote the maximum and average fitness of the current population, respectively; f' represents the fitness of the better individual among the two selected for crossover; f corresponds to the fitness of the individual undergoing mutation.

However, Equations (24) and (25) consider only individual fitness and neglect the overall evolution of the population, making it difficult to capture population-wide trends and potentially leading to local optima. To address this limitation, this study introduces improvements to both the crossover and mutation probability functions:

$$p_c = \begin{cases} p_{c\max} * e^{-\frac{f' - f_{\text{avg}}}{f_{\max} - f_{\text{avg}}}}, & f' \geq f_{\text{avg}} \\ p_{c\max}, & f' < f_{\text{avg}} \end{cases} \tag{26}$$

$$p_m = \begin{cases} p_{m\max} * e^{-\frac{f - f_{\text{avg}}}{f_{\max} - f_{\text{avg}}}}, & f \geq f_{\text{avg}} \\ p_{m\max}, & f < f_{\text{avg}} \end{cases} \tag{27}$$

$P_{c\max}$ represents the maximum crossover probability, and $P_{m\max}$ is the maximum mutation probability. As shown in the equations, regardless of an individual's fitness value, both the crossover and mutation probabilities are prevented from dropping to zero. This not only effectively preserves high-performing individuals but also avoids premature convergence to local optima.

However, the exponential terms in the equations tend to magnify differences in fitness values. When an individual's fitness deviates substantially from the population mean, the resulting rapid growth or decay of the exponential component may induce excessive fluctuations in crossover and mutation probabilities. Such instability can lead to over-exploration and undermine the robustness of the algorithm. To mitigate this issue, fitness variance is introduced to further refine and stabilize the equations:

$$p_c = \begin{cases} p_{c\max} * e^{-\frac{f' - f_{\text{avg}}}{\sqrt{Var(f')}}}, & f' \geq f_{\text{avg}} \\ p_{c\max} \quad , & f' < f_{\text{avg}} \end{cases} \tag{28}$$

$$p_m = \begin{cases} p_{m\max} * e^{-\frac{f-f_{\text{avg}}}{\sqrt{Var(f)}}}, & f \geq f_{\text{avg}} \\ p_{m\max}, & f < f_{\text{avg}} \end{cases} \tag{29}$$

$Var(f)$ is a statistical measure used to quantify the dispersion of fitness values within the population. A larger variance indicates greater differences in individual fitness, while a smaller one implies a more uniform population. The calculation formula is as follows:

$$Var(f) = \frac{1}{N}\sum_{i=1}^{N}(f_i - f_{avg})^2 \tag{30}$$

By incorporating fitness variance, the crossover and mutation probabilities are adjusted more smoothly, thereby suppressing drastic fluctuations induced by extreme fitness values. This adjustment enhances the stability of the algorithm and strengthens its capability to explore the solution space effectively.

Although this strategy increases parameter flexibility, it is still susceptible to local optima in complex search spaces, particularly when fitness differences among individuals are minimal. This limitation reduces adaptability and hinders global exploration. Moreover, AGA encounters challenges in balancing evolutionary speed and search quality: overly aggressive adjustment rates may lead to divergence and inefficiency, whereas overly conservative rates restrict the search range and degrade performance. To address these limitations, this study integrates the MEA framework to further enhance AGA. MEA simulates human-like rapid evolution through learning and innovation, thereby improving both adaptability and global search capability. In each iteration, a randomly initialized population undergoes fitness-based selection to generate a central individual, which partitions the population into elite and temporary subpopulations. The elite group preserves high-quality solutions, while the temporary group performs exploratory search, together enabling cooperative optimization. The procedure is illustrated in Figure 5.

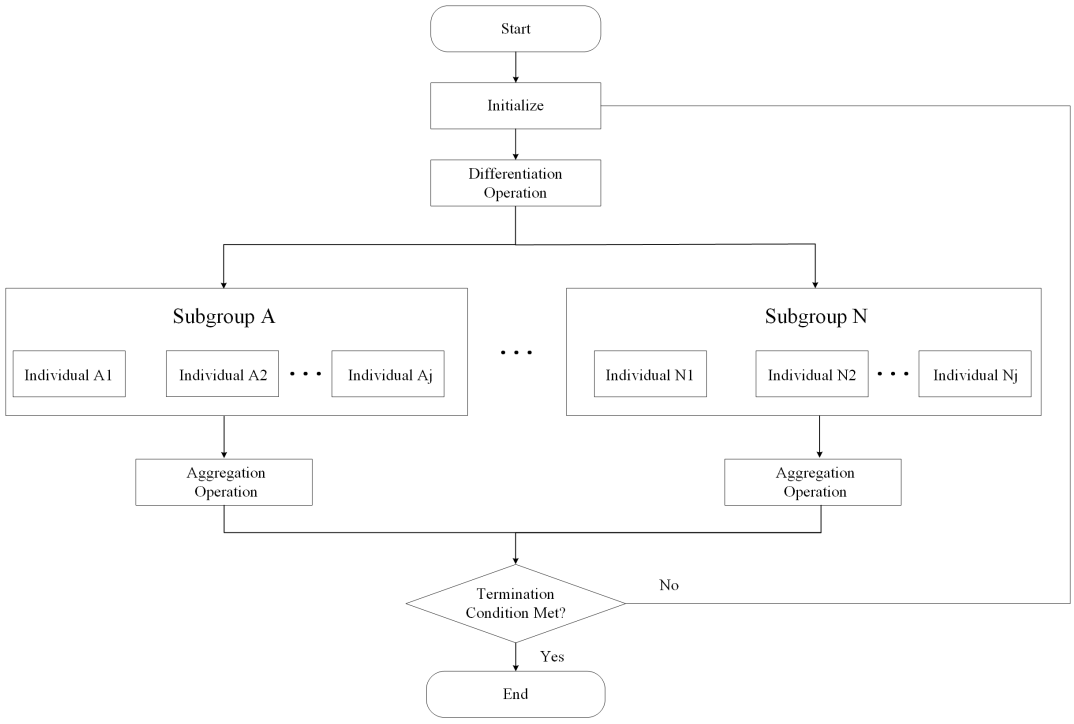


Figure 5. MEA flowchart.

As illustrated in the figure, the algorithm achieves a balance between global exploration and local optimization through a two-phase "diversification–convergence" strategy. During the diversification phase, population diversity is increased to promote broad exploration of the search space and identify multiple promising candidate solutions. In the subsequent convergence phase, these candidate solutions are refined to enhance solution quality and improve convergence efficiency. The detailed procedures of the diversification and convergence phases are depicted in Figures 6 and 7, respectively.

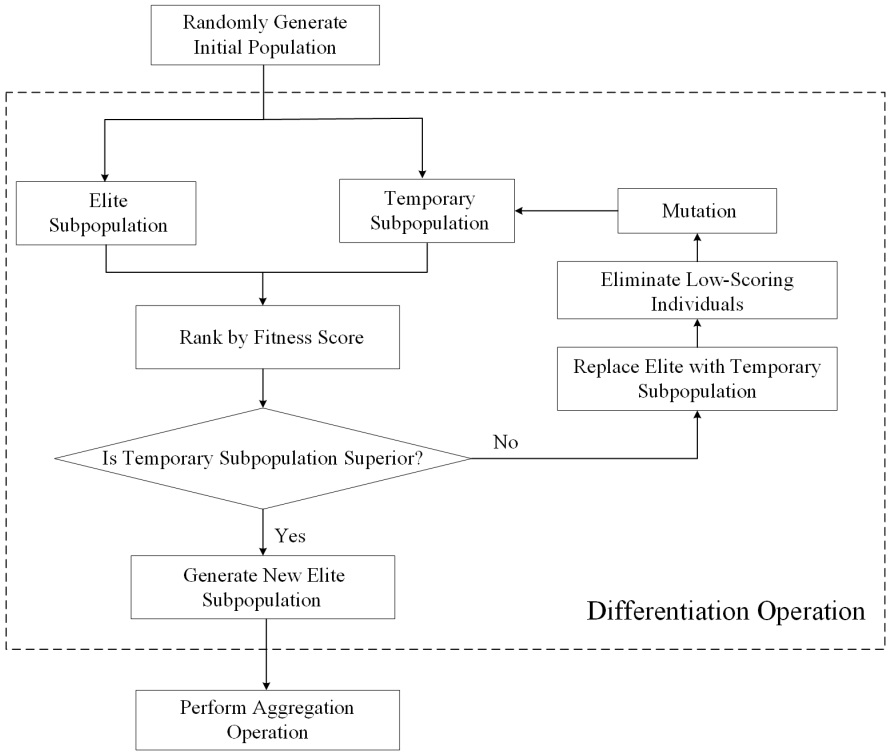


Figure 6. Differentiation operation flowchart.

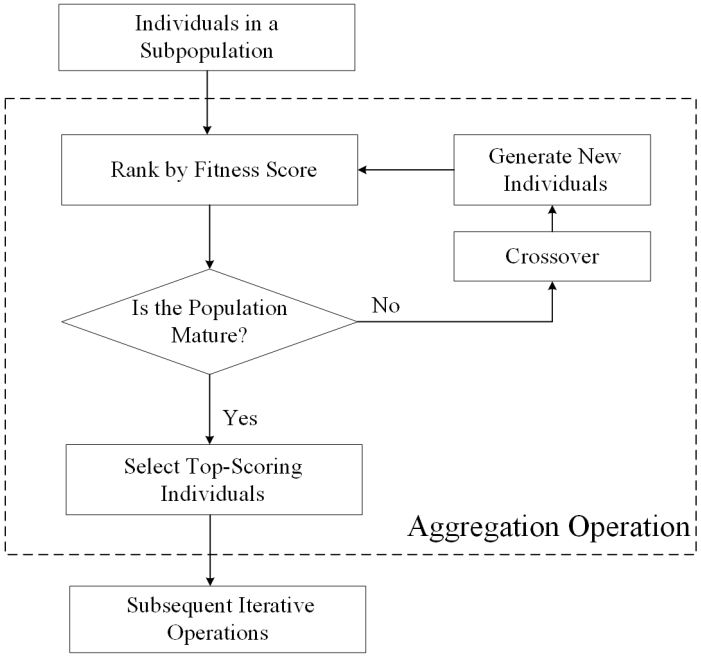


Figure 7. Aggregation operation flowchart.

In the global search phase, a diversification operation is first conducted. The population is divided into elite and temporary subpopulations according to fitness values, facilitating competitive interactions that help identify multiple promising global optima. If a temporary subpopulation outperforms the current elite subpopulation in terms of fitness, it replaces the elite group and is incorporated into the set of optimal solutions. Meanwhile, the temporary subpopulation with the lowest fitness is discarded and reinitialized across the entire solution space. Since this phase emphasizes search breadth and allows for larger parameter perturbations, the mutation mechanism from (29) is adopted to expand the search range.

In the local search phase, a convergence operation is carried out. This operation involves a two-stage refinement of the elite subpopulation obtained during the diversification phase. Guided by the fitness function, it seeks to further enhance solution quality and approximate the global optimum. To prevent excessive parameter perturbation, the crossover mechanism is applied, with the crossover rate calculated using (28), thereby improving the precision of local search. When the optimal individual remains stable across successive iterations and no longer changes, the subpopulation is considered matured, marking the completion of the convergence process.

In summary, the MEA-AGA is employed to optimize the NBPE parameter in Tongan speech recognition, with the objective of balancing recognition accuracy, decoding speed, and training cost. Prior to the optimization process, a fitness function is defined to comprehensively evaluate multiple factors in the recognition process, including model accuracy, decoding efficiency, and training time. The calculation formula is as follows:

$$\text{fitness} = \alpha(1 - WER_{dev}) + \beta(1 - WER_{test}) + \lambda \frac{WPS}{10^{\lceil \log_{10}(WPS) \rceil}} + \mu \frac{10^{\lfloor \log_{10}(t_{train}) \rfloor}}{t_{train}} \quad (31)$$

WER_{dev}, WER_{test}, WPS, and t_{train} represent the word error rate on the validation set, the word error rate on the test set, the number of words recognized per second, and the model training time, respectively. The corresponding weight coefficients α, β, λ, and μ are set to 0.3, 0.3, 0.3, and 0.1, reflecting the equal importance of recognition accuracy and decoding speed, while giving relatively less emphasis to training time. In addition, to eliminate the influence of differing units of measurement, WPS and training time are normalized to ensure that all evaluation indicators fall within a comparable range.

The overall procedure of the proposed optimization algorithm is summarized in Algorithm 2, and its detailed steps are described as follows:

(1) Parameter Initialization: Set the maximum number of iterations, population size N, crossover probability P_c, and mutation probability P_m.

(2) Population Initialization: Randomly generate 10 initial individuals $pop_{1n} \sim pop_{10n}$.

(3) Fitness Evaluation: Calculate the fitness scores of all individuals in the population using (30).

(4) Diversification Operation: Sort individuals based on their fitness scores. The top five individuals form the elite subpopulation $pop_{best1n} \sim pop_{best5n}$, and the bottom five form the temporary subpopulation $pop_{temp1n} \sim pop_{temp5n}$. Determine whether the fitness score of the newly generated temporary subpopulation is higher than that of the elite subpopulation. If so, apply mutation (29) to the lowest-scoring temporary individual to generate a new temporary subpopulation $pop'_{\text{temp}1n} \sim pop'_{temp5n}$, and return to step 3; otherwise, the diversification phase is complete, and the elite subpopulation is finalized.

(5) Fitness Calculation of the New Elite Subpopulation: Recalculate the fitness scores of all individuals in the new elite group using (30).

(6) Convergence Operation: Sort individuals in the elite subpopulation by fitness; the highest-scoring individual is identified as the winner. Determine whether the population

has matured. If the winner changes, perform crossover (28) on the remaining individuals to generate new individuals, and return to step 5; otherwise, the convergence phase is complete, and the global optimal individual is obtained.

(7) Termination Check: Determine whether the stopping condition is met. If the maximum number of generations has been reached, the process ends and the current optimal NBPE value is output; otherwise, return to step 3 to continue iteration.

(8) Dictionary Construction: Based on the optimal NBPE value, segment Tongan words and construct the Tongan dictionary, which serves as a standard for both training and inference, ultimately enabling accurate and efficient Tongan speech recognition.

Algorithm 2: MEA-AGA for Dictionary Parameter Optimization

Require: N, n_{max}, P_c, P_m, pop_1~pop_{10}
Ensure: Optimal *NBPE*

```
/* Stage 1: Initialization */
Initialize parameters; generate population pop1~pop10;
for iteration = 1 to n_max do
    /* Stage 2: Fitness Evaluation */
    Compute fitness for all individuals;
    /* Stage 3: Differentiation */
    Rank population; select top 5 as pop_elite, others as pop_temporary;
    if pop_temporary better than pop_elite then
        Replace pop_elite; mutate poor individuals (see Equations (3)–(32));
    else
        Regenerate pop_elite;
        continue;
    end if
    /* Stage 4: Aggregation */
    Optimal NBPE ⇐ NBPE of best individual
    if not converged then
        crossover individuals (see Equations (3)–(31));
        continue;
    end if
    /* Stage 5: Termination */
    if termination met then
        break;
      end if
end for
```

4. Experimental Setup

4.1. Dataset Configuration

In transfer learning, the choice of source language plays a crucial role in determining model performance. This is especially relevant for low-resource speech recognition, where selecting a source language with abundant resources and phonetic structures similar to the target language can significantly improve transfer effectiveness and generalization. Tongan, the official language of the Kingdom of Tonga, belongs to the Polynesian language family and represents a typical low-resource case [51]. Due to historical colonization and the influence of modern education systems, English has gradually become one of the official languages and is widely taught as a second language in schools [52,53]. Prolonged language contact has resulted in extensive borrowing from English, both in vocabulary and

phonological systems, creating notable similarities in phoneme distribution and speech structure. Furthermore, English, as the most resource-rich language, offers large-scale annotated corpora and well-developed pretrained models, providing a strong foundation for transfer learning. Therefore, this study selects English as the source language for building pretrained models to facilitate Tongan speech recognition.

This study employs two datasets. For English, the publicly available LibriSpeech corpus (960 h) is used for pretraining. For Tongan, the speech data were recorded in our laboratory by several professional researchers to ensure both speaker diversity and consistent recording quality. All recordings were conducted in a quiet indoor environment using a high-quality condenser microphone positioned approximately 20 cm from the speaker's mouth. Each utterance was recorded by a single speaker without overlap to avoid multi-speaker interference. All recordings were captured in mono format, with no speaker diarization applied, as each utterance was produced by a single speaker. The sampling rate was set to 16 kHz. The transcription process was manually performed using Notepad++ and strictly aligned with the reference text to guarantee complete labeling accuracy. The recordings feature clear and standardized pronunciation, making them suitable for evaluating model performance and recognition robustness under low-resource conditions. In total, 1.44 h of Tongan speech data were obtained and divided into training, validation, and test sets. Detailed statistics are provided in Table 1.

Table 1. Tongan corpus configuration (raw data).

Dataset	Duration (h)	Sampling Rate	Format
Training	1.13	16 kHz	FLAC
Validation	0.15		
Test	0.15		
Total	1.43		

As shown in the table, the Tongan dataset is relatively small in scale. Direct training on such limited data is likely to cause overfitting and reduce the model's generalization capability. Therefore, in the subsequent experiments, the data will be appropriately augmented and expanded to reach a more reasonable scale.

4.2. Experimental Design

This study adopts the Mixformer network as the base architecture for transfer learning. The pre-trained weights are obtained from previous English speech recognition experiments that were trained on the LibriSpeech corpus [54]. The Mixformer model, containing approximately 90 million parameters, performs loss-based weighted fusion of the Conformer, Unified Conformer, and U2++ Conformer architectures to improve overall robustness. During decoding, it combines both CTC and attention mechanisms, incorporating a penalty function to dynamically optimize decoding paths, thereby achieving a balance between recognition accuracy and inference efficiency. Building on this foundation, the proposed layer-wise adaptive transfer strategy is applied to fine-tune the model, improving its adaptability and training efficiency for the Tongan language task.

To provide a clearer view of the model's architecture and performance, Tables 2 and 3 present the Mixformer's core configurations, along with its training settings and recognition results on the English corpus.

Table 2. Model architecture.

Layer Name	Output Size	Input Size/Patch Size, Stride
Input	80 d	1000 × 80
Conv2D (input layer)	250 × 20 × 256	80 × T, 4 × 4, stride 2
Positional encoding	250 × 256	250 × 256
Conformer block × 12	250 × 256	250 × 256
Feed-forward layer	250 × 256	250 × 256
Self-attention layer	250 × 256	250 × 256
CNN module	250 × 256	250 × 256
Fully connected layer	5002 d	250 × 256
Output	5002 d	250 × 5002

Table 3. Parameter configuration and recognition results in English pretraining stage.

Parameter	Value
Learning rate	0.001
Decoder layers	6
Maximum sequence length	2000
Batch size	12
Number of training epochs	100
CTC loss weight	0.3
Decoding strategy	Adaptive Decoding with Dynamic Rescoring
Recognition performance (WER)	3.92%

Experiments were conducted on an NVIDIA RTX 4070 GPU (16 GB, NVIDIA Corporation, Santa Clara, CA, USA) with CUDA 11.8, running on 64-bit Ubuntu 18.04. The environment was configured with Python 3.8 and PyTorch 2.1.2. Detailed parameter settings for the Tongan training stage are summarized in Table 4.

Table 4. Parameter configuration in Tongan training stage.

Parameter	Value
Learning rate	0.001
Decoder layers	6
Maximum sequence length	2000
Batch size	12
Number of training epochs	25
CTC loss weight	0.3
Optimizer	Adam
Dropout rate	0.1

In addition, standard training management techniques were employed, including early stopping based on validation loss and automatic checkpoint saving after each epoch. Furthermore, a warm-up followed by cosine-annealing learning rate scheduling was utilized to stabilize the training process and promote smoother convergence.

4.3. Performance Metrics

In speech recognition, evaluating model performance is critical. This study adopts Word Error Rate (WER) as the primary evaluation metric. WER is calculated by measuring the minimum edit distance—comprising substitutions, insertions, and deletions—between the recognized text and the reference text and dividing it by the total number of words in the reference. The formula is as follows:

$$WER = \frac{S_w + D_w + I_w}{N_w} \tag{32}$$

Here, S_w, D_w, and I_w represent the number of substitutions, deletions, and insertions, respectively, while N_w denotes the total number of words. A lower WER indicates higher recognition accuracy.

In addition to recognition accuracy, decoding speed is also an important evaluation criterion. This study uses Words Per Second (WPS) to reflect the model's real-time processing capability. The calculation formula is as follows:

$$WPS = \frac{N_w}{T_I} \tag{33}$$

T_I represents the time required to complete an inference of speech recognition.

In addition, to systematically evaluate the effectiveness of the data augmentation and partitioning strategies, this study conducts quantitative analysis from two perspectives: the similarity quality of augmented data and the class balance of the partitioned subsets.

For similarity evaluation, the proposed augmentation algorithm incorporates a DRF module to assess feature similarity between synthetic pseudo-samples and original samples. DRF filters and retains high-quality augmented samples based on a scoring mechanism, as detailed in (8).

For class balance evaluation, given that audio signal features are high-dimensional and structurally complex, making them difficult to interpret directly, this study employs the t-SNE dimensionality reduction algorithm to project the features into a two-dimensional space for intuitive visualization of data clustering. To quantitatively assess distributional divergence between subsets, a weighted total variance metric is introduced as follows:

$$\sigma^2_{total} = \frac{n_{train} \cdot \sigma^2_{train} + n_{val} \cdot \sigma^2_{val} + n_{test} \cdot \sigma^2_{test}}{n_{train} + n_{val} + n_{test}} \tag{34}$$

σ^2_{train}, σ^2_{val} and σ^2_{test} denote the variances of the training, validation and test sets, respectively; n_{train}, n_{val} and n_{test} represent the number of samples in the training, validation and test sets.

5. Experimental Results

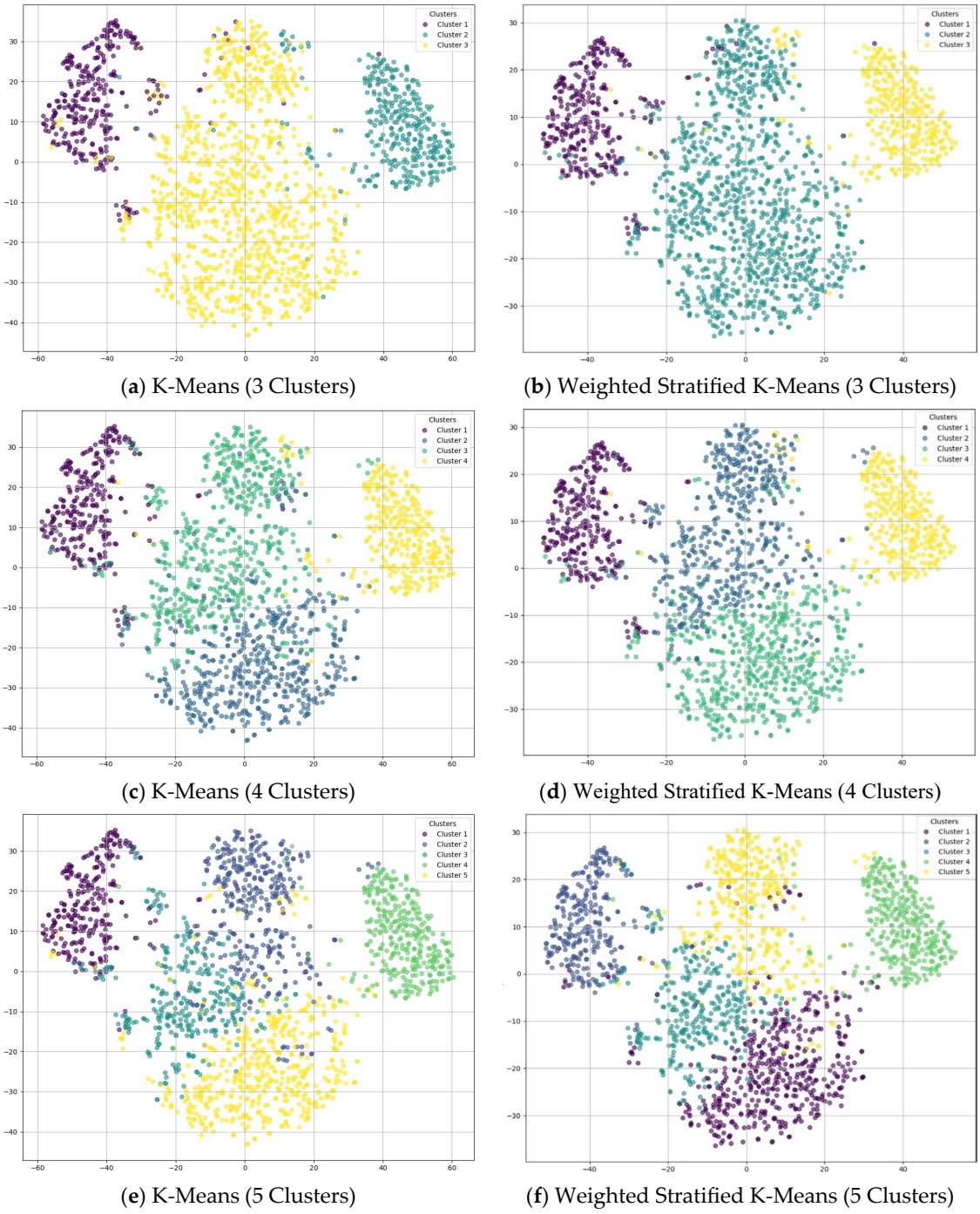
5.1. Data Augmentation Experiments

To address data scarcity and class imbalance in Tongan speech recognition, this study augments the original dataset using SRA-DRF combined with weighted stratified sampling. The effectiveness is evaluated from two perspectives: data partitioning and augmentation similarity.

For the data partitioning experiment, given the large volume of augmented samples, qualitative analysis uses only a subset for t-SNE visualization to avoid visual clutter. It is important to note that the x- and y-axes in the t-SNE plots reflect only relative sample distribution without physical meaning; therefore, axis ticks are omitted in the corresponding figures. In contrast, quantitative analysis is performed on the full dataset to compute the total variance, ensuring comprehensive and representative evaluation. The data are clustered into 3 to 7 groups, and samples are further categorized into near, medium, and far subgroups according to their distances from cluster centers. Stratified sampling is then applied to construct the training, validation, and test sets. Finally, the proposed method is compared with the baseline approach both qualitatively and quantitatively, as shown below.

Figure 8 presents the clustering results of the traditional algorithm and the proposed method under different numbers of clusters (3–7). When the number of clusters is set to 3,

the clustering is relatively coarse, making it difficult to distinguish the data effectively. As the number increases to 4 or 5, the data distribution becomes clearer, and clustering quality improves markedly. However, with 6 or 7 clusters, the boundaries become blurred and class overlap intensifies, leading to a decline in clustering performance. Comparing performance across different cluster settings, the traditional K-Means performs reasonably well with fewer clusters but exhibits severe class mixing as the number increases. In contrast, the proposed method maintains clearer separation even with higher cluster counts, alleviating the overlap issue to some extent and consistently outperforming the baseline.



(**a**) K-Means (3 Clusters)

(**b**) Weighted Stratified K-Means (3 Clusters)

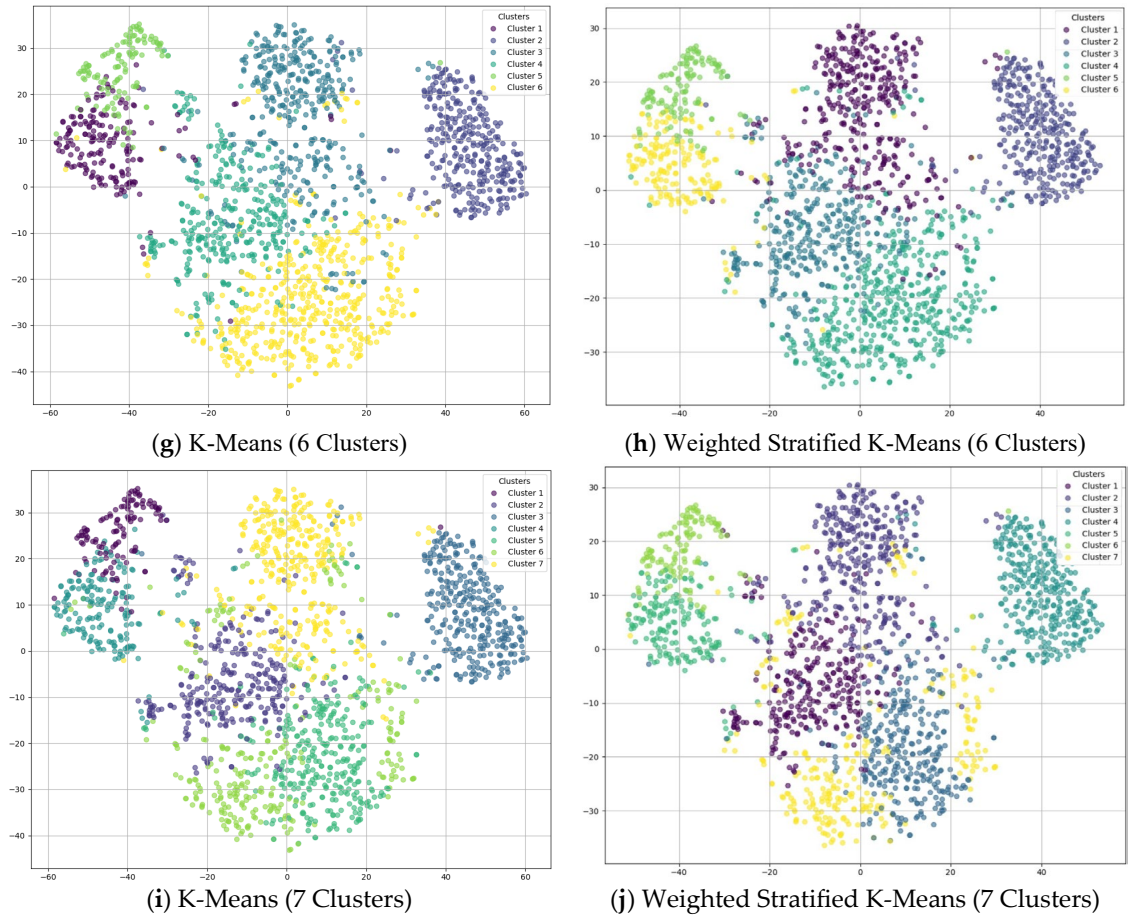
(**c**) K-Means (4 Clusters)

(**d**) Weighted Stratified K-Means (4 Clusters)

(**e**) K-Means (5 Clusters)

(**f**) Weighted Stratified K-Means (5 Clusters)

Figure 8. *Cont.*



(**g**) K-Means (6 Clusters) (**h**) Weighted Stratified K-Means (6 Clusters)

(**i**) K-Means (7 Clusters) (**j**) Weighted Stratified K-Means (7 Clusters)

Figure 8. Comparison of clustering results.

Table 5 presents the dataset partitioning results obtained using K-Means and weighted stratified K-Means under different cluster numbers (three to seven). The experimental results indicate that although the total variance of K-Means generally decreases as the number of clusters increases, its performance becomes unstable at higher cluster counts (six or seven), particularly on the test set. In contrast, the proposed method maintains better balance even with fewer clusters and exhibits more stable variance trends as the cluster number increases, demonstrating improved consistency and uniformity in partitioning. Based on the overall evaluation, the configuration with five clusters—yielding the most balanced results—is selected for subsequent data partitioning. The final statistics are reported in Tables 6 and 7.

Table 5. Quantitative comparison of dataset partitioning results.

Method	No. of Clusters	Variance (Train)	Variance (Val)	Variance (Test)	Total Variance
K-means	3	1.0002	1.0166	0.9859	1.0004
	4	1.0025	1.0165	1.0130	1.0050
	5	0.9998	0.9467	0.9649	0.9910
	6	0.9907	0.9754	0.9904	0.9891
	7	1.0112	0.9856	1.0409	1.01116
Weighted Stratified K-Means	3	0.9598	0.9833	1.0230	0.9685
	4	1.0008	0.4244	0.6290	0.9059
	5	0.9110	0.6476	0.7583	0.8694
	6	1.0060	0.6299	0.7208	0.9398
	7	1.0751	0.7032	0.5919	0.9894

Table 6. Dataset partition statistics.

Dataset	Number of Samples
Training Set	1449
Validation Set	181
Test Set	182
Total	1812

Table 7. Final sample classification statistics.

Cluster	Subclass
Cluster 1: 267	Class 1: 88 Class 2: 89 Class 3: 91
Cluster 2: 328	Class 1: 126 Class 2: 125 Class 3: 131
Cluster 3: 333	Class 1: 110 Class 2: 112 Class 3: 111
Cluster 4: 324	Class 1: 107 Class 2: 107 Class 3: 110
Cluster 5: 506	Class 1: 167 Class 2: 172 Class 3: 167
Total:	1812

To validate the effectiveness of the augmentation method, eight controlled experiments were conducted to examine the impact of different augmentation techniques and class sample distributions on classification accuracy. The experimental configurations and results are summarized in Table 8.

Table 8. Data augmentation effectiveness experiment.

Experiment	1	2	3	4	5	6	7	8
A	1804	1830	1946	400	1946	1946	1946	973
B	1766	1772	1820	1820	400	1820	1820	1820
C	1796	1804	1870	1870	1870	400	1870	1870
D	1726	1768	1804	1804	1804	1804	400	1804
E	1816	1828	1846	1846	1846	1846	1846	400
Similarity	79.60%	88.62%	90.63%	73.50%	72.13%	72.09%	72.90%	70.30%

In this study, the dataset consists of five categories (A, B, C, D, E), clustered by the weighted stratified sampling algorithm. In Experiment 1, data augmentation was performed using a Generative Adversarial Network (GAN), which achieved a similarity of 79.60%. Experiment 2 applied traditional signal processing techniques such as speed and pitch perturbation, achieving a similarity of 88.62%. Experiment 3 integrated both approaches with the proposed augmentation algorithm, achieving the highest similarity of 90.63%. This result outperformed Experiments 1 and 2, demonstrating that the proposed method effectively enhances data quality and improves model robustness. Experiments 4–8

involved removing one category at a time from the Tongan dataset. The results show that removing any category led to a significant drop in accuracy, indicating that maintaining class balance plays a crucial role in data augmentation. Therefore, the dataset generated in Experiment 3, which achieved the highest classification accuracy, was selected for subsequent transfer learning experiments.

After data augmentation and partitioning, the final Tongan speech dataset consists of 8686 audio samples, divided into five categories: A (1946 samples), B (1820), C (1870), D (1804), and E (1846), totaling approximately 11.44 h. All audio files are stored in FLAC format and sampled at 16 kHz. The dataset distribution is shown in Table 9.

Table 9. Configuration of the expanded data sample set.

Sample Set	Category	Sampling Rate	Format
Training Set: 6950 (9.15 h)	A: 1946 (2.56 h) B: 1820 (2.40 h) C: 1870 (2.46 h)	16 kHz	FLAC
Validation Set: 866 (1.14 h)	D: 1804 (1.59 h)		
Test Set: 870 (1.15 h)	E: 1846 (2.43 h)		
Total: 8686 (11.44 h)			

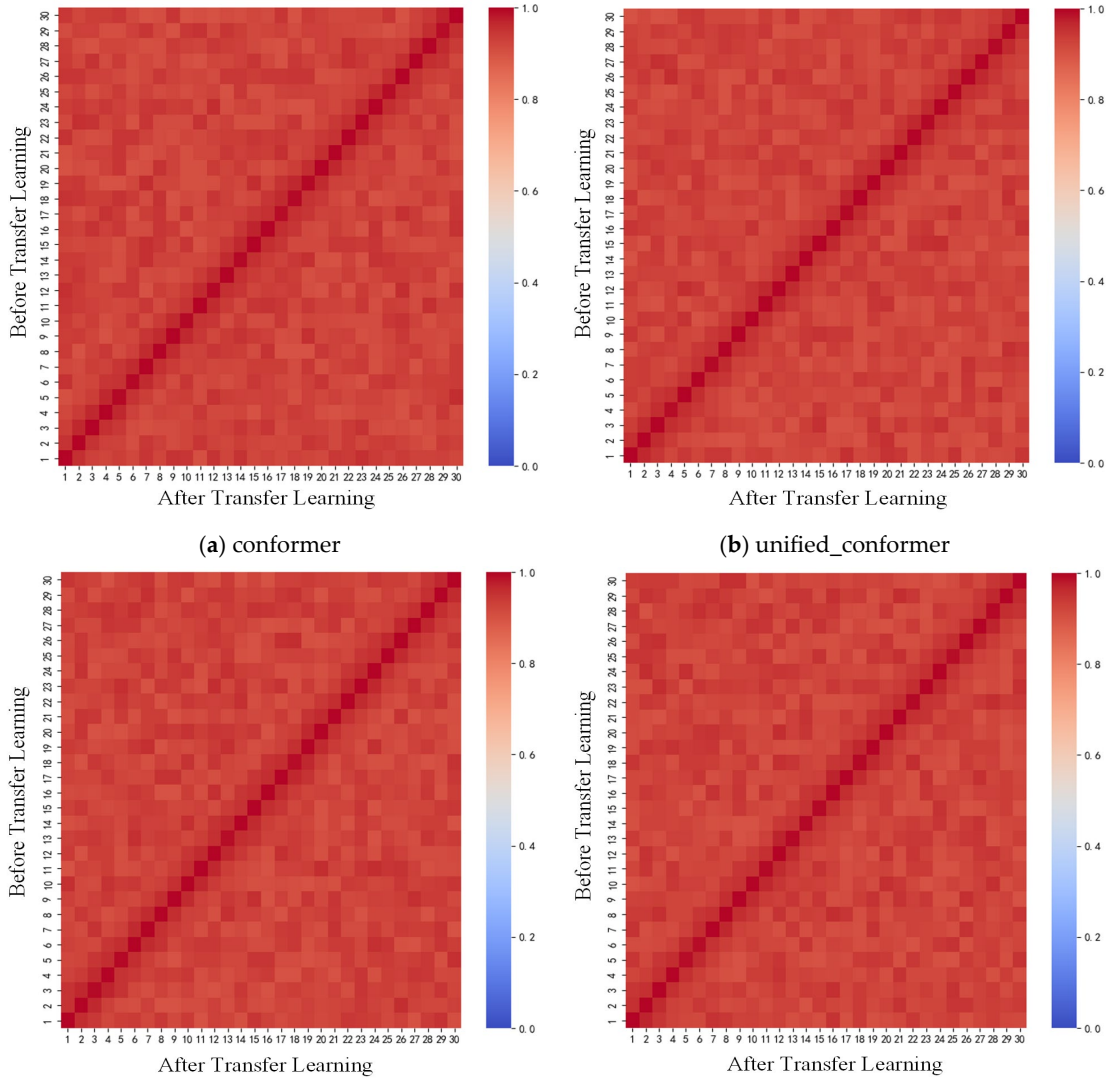
In summary, the proposed data augmentation method not only expanded the corpus size but also preserved a high similarity in feature distributions between the augmented samples and the original data, confirming the effectiveness and practicality of the augmented corpus. The weighted stratified sampling strategy effectively improved class balance during dataset partitioning and enhanced the consistency across the training, validation, and test sets. Together, these two strategies enabled the development of a stable, high-quality low-resource speech dataset, providing a solid foundation for subsequent model training.

5.2. Transfer Learning Experiments

To comprehensively evaluate the effectiveness of the proposed transfer learning strategy, this section conducts experimental validation from two perspectives: a visual analysis of the model migration process and a comparison of recognition performance after transfer.

The experiment begins with a visualization-based analysis to examine structural changes in the model before and after fine-tuning. Given the complexity of the adopted architecture, it is impractical to cover all layers. To address this, a representative selection strategy is adopted: the lower layers are sampled using stratified sampling, and the last 30 upper layers are selected to align with the layer-wise unfreezing process and to observe the evolution of feature representations. The corresponding experimental results are presented below.

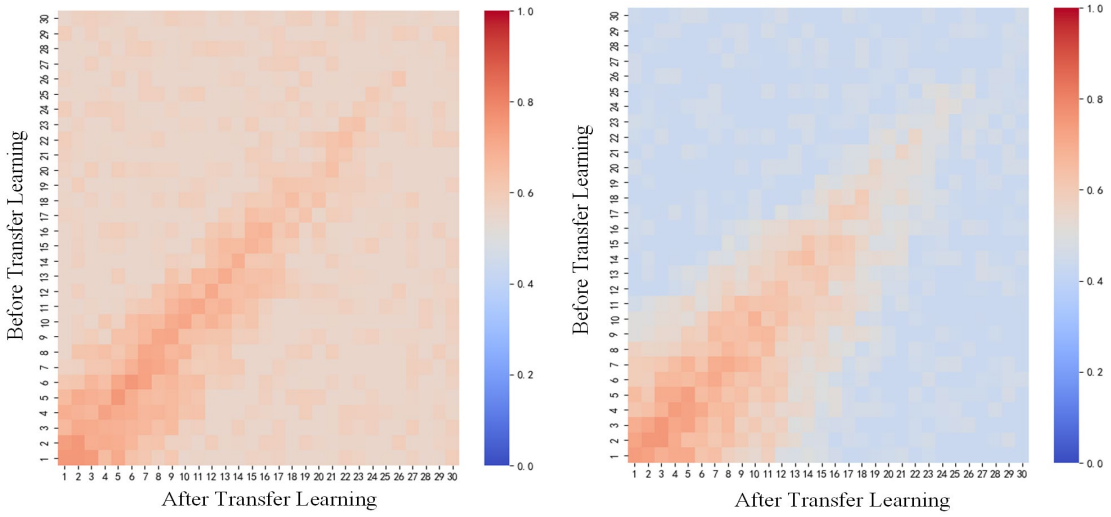
Figures 9 and 10 present the CKA similarity matrices of four models before and after training, focusing on the lower and upper layers. The results show that the similarity patterns in the lower layers are highly consistent across all models, with values along the main diagonal close to 1. This suggests that lower-level features remain largely unchanged during fine-tuning and are minimally affected by parameter updates. In contrast, the upper-layer similarities gradually decrease with increasing depth, as reflected by the fading of the main diagonal. This trend suggests that parameter adjustments intensify progressively in higher layers. Such a pattern aligns well with the layer-wise unfreezing strategy discussed in Section 3.2, further confirming that higher-level representations gradually adapt to the target task, whereas lower-level structures remain stable. This demonstrates the effective implementation of the proposed transfer strategy.



(**a**) conformer (**b**) unified_conformer

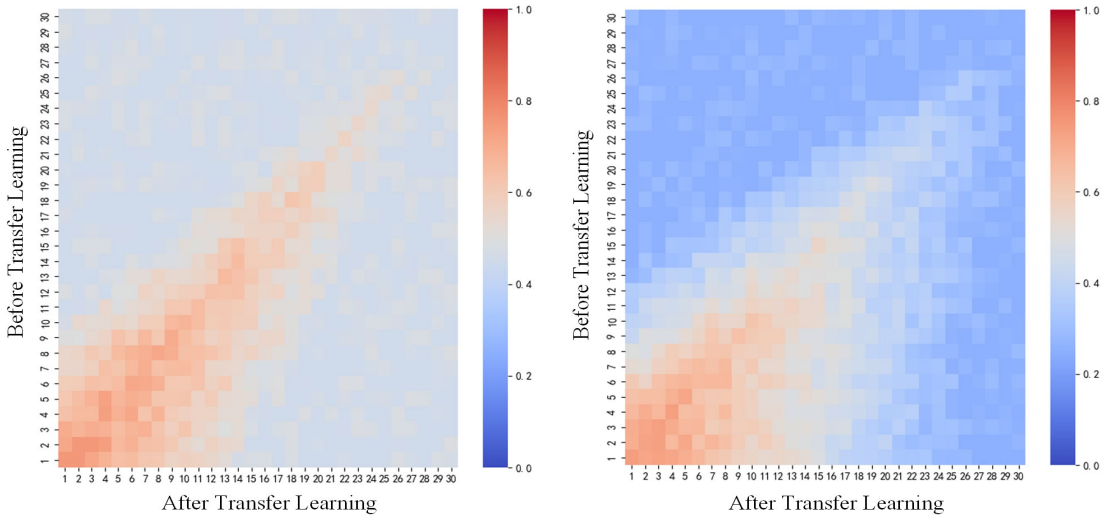
(**c**) u2++_conformer (**d**) Mixformer

Figure 9. CKA similarity of low-level structures before and after transfer across models.



(**a**) conformer (**b**) unified_conformer

Figure 10. *Cont.*



(**c**) u2++_conformer (**d**) Mixformer

Figure 10. CKA similarity of high-level structures before and after transfer across models.

Figure 11 quantitatively illustrates the changes in feature similarity across different layers. The similarity scores in the lower layers remain close to 1, indicating that these features remain stable during fine-tuning. In contrast, similarity in the upper layers decreases progressively with increasing depth, with the largest drop observed in the top layers, thereby confirming the effectiveness of the layer-wise unfreezing strategy. The extent of high-layer variation differs among models: Conformer exhibits the smallest change, while Unified Conformer and U2++ Conformer show more noticeable decreases. Mixformer demonstrates the most substantial drop, suggesting a more thorough adaptation of high-level features, which may enhance its potential for Tongan speech recognition tasks.

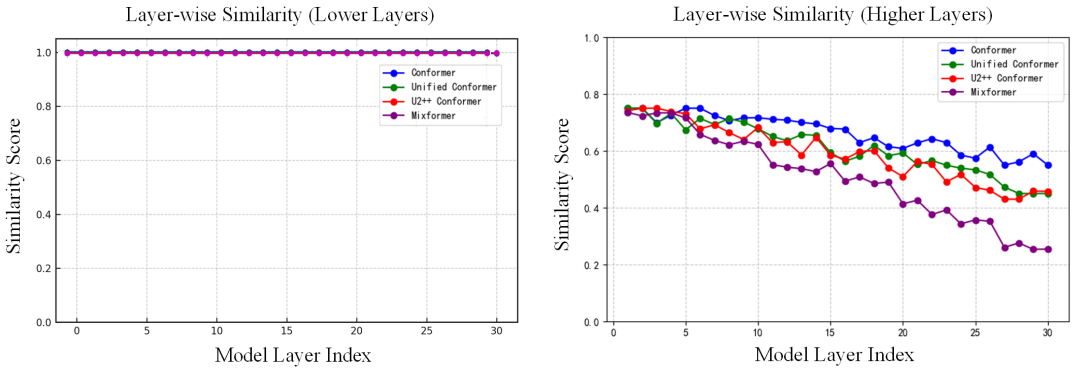


Figure 11. Layer-wise similarity variation before and after transfer in different models.

Building on the preceding visualization analysis, this section further evaluates recognition performance after transfer learning. Table 10, Figures 12 and 13 summarizes the experimental results for the combinations of three data augmentation strategies (without augmentation, random augmentation, and the proposed method) and two training approaches (direct training and layer-wise transfer learning).

Table 10. Comparison of data augmentation methods.

Augmentation Method	Training Strategy	WER(%)		Epoch
		Dev	Test	
Without Augmentation	Direct Training	53.77%	54.93%	50
	Transfer Learning	42.31%	43.43%	25
Random Augmentation	Direct Training	44.01%	46.01%	50
	Transfer Learning	36.80%	36.09%	25
SRA-DRF Augmentation	Direct Training	35.05%	36.91%	50
	Transfer Learning	28.56%	30.42%	25

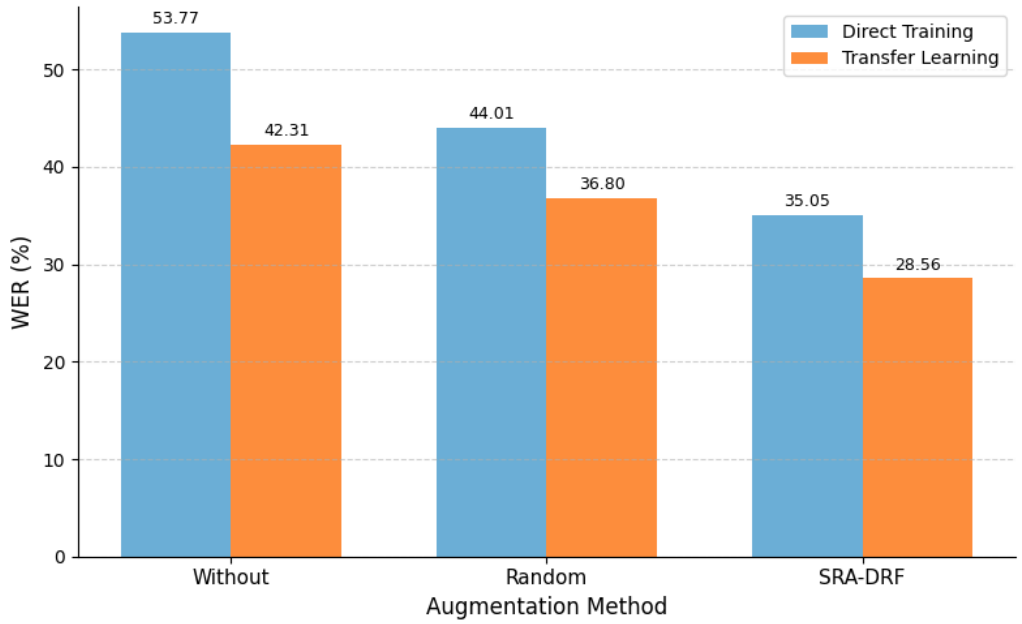


Figure 12. Development set comparison of data augmentation methods.

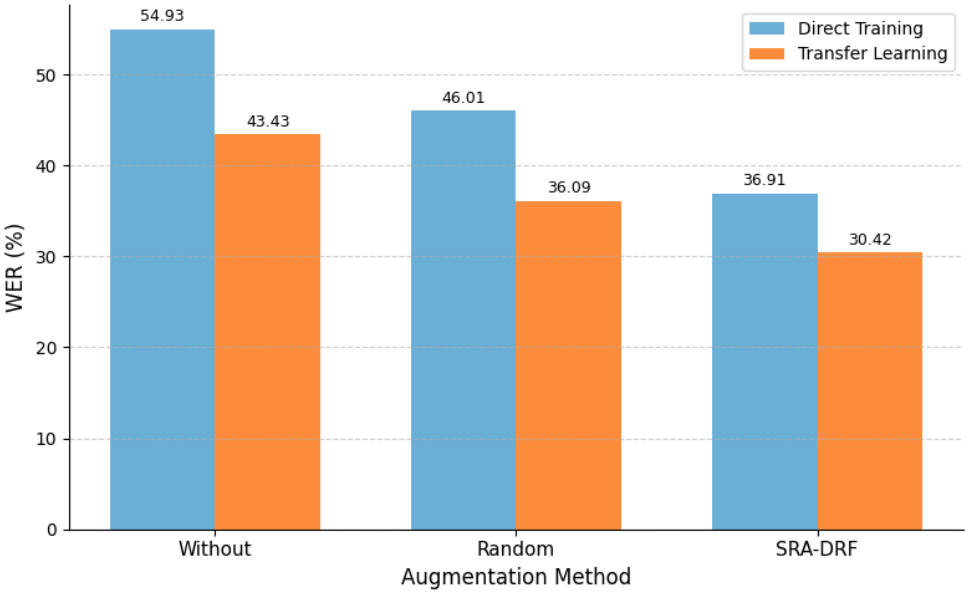


Figure 13. Test set comparison of data augmentation methods.

The results show that direct training without augmentation results in poor recognition accuracy. Random augmentation provides only modest and limited improvements. In contrast, the proposed method combined with layer-wise transfer markedly improves recognition accuracy while halving the required training epochs. A plausible explanation for this improvement lies in two main aspects. First, the SRA-DRF augmentation expands the training data with higher quantity and quality, avoiding invalid or overly dissimilar samples that may arise in random augmentation. Its selective mechanism guarantees that the generated audio remains acoustically consistent with the original recordings, thereby enhancing the model's generalization ability and reducing learning bias. Second, the adopted transfer learning strategy enables the model to leverage the pretrained English representations, which share similar phonetic and acoustic characteristics with Tongan. This facilitates more efficient feature adaptation, leading to faster convergence and higher recognition accuracy. These findings highlight both the effectiveness and efficiency of the proposed method in addressing data scarcity in low-resource languages, providing a promising modeling approach for Tongan speech recognition.

5.3. Dictionary Parameter Optimization Experiments

This section performs automatic optimization of the core dictionary parameter, NBPE, using the MEA-AGA, in order to determine the optimal dictionary size. The process begins with the differentiation phase to generate the initial population. The results are presented below.

Table 11 lists ten sets of NBPE parameters generated during the initial Differentiation phase, along with their corresponding WER, training time, and WPS performance metrics. The fitness scores, calculated according to (31), are used to rank these candidates The top five individuals are designated as the elite subpopulation, providing the foundation for subsequent optimization of dictionary parameters, while the remaining five are classified as temporary individuals, forming the temporary subpopulation.

Table 11. Results of differentiation operation.

NBPE	Dev-WER	Test-WER	Training Time	WPS	Fitness
96	27.75%	30.56%	8633.962	35.32	0.646
197	28.57%	30.44%	4769.055	63.40	0.823
248	28.44%	30.38%	4912.812	60.98	0.810
304	28.49%	30.49%	4733.265	63.10	0.824
346	28.56%	30.49%	4621.743	67.50	0.842
398	28.4%	30.59%	4501.978	66.90	0.846
504	29.55%	31.29%	7031.657	63.80	0.750
627	29.05%	30.88%	5442.410	63.50	0.800
715	30.1%	31.85%	9789.546	61.50	0.701
824	27.84%	30.3%	7421.311	34.59	0.662

Building on this foundation, an Aggregation operation is applied to both the elite and temporary subpopulations to enable fine-grained search and improve overall optimization performance. Figure 14 illustrates the variation in fitness scores of the two subpopulations throughout the Aggregation process.

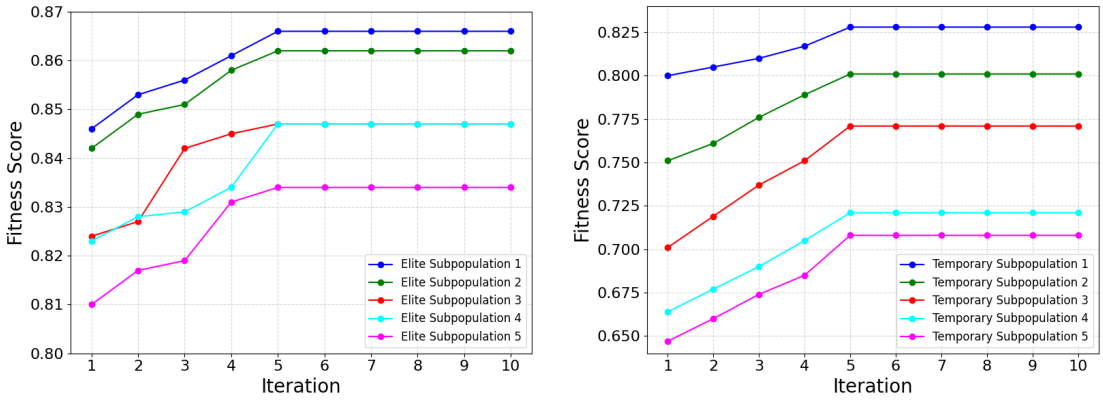


Figure 14. Iteration process of the dominant and temporary subpopulation.

The results indicate that the MEA-AGA achieves stable performance in dictionary parameter optimization, with fitness scores steadily improving during the early stages and gradually converging in later iterations. The elite subpopulation consistently outperforms the temporary subpopulation, and the optimal solution stabilizes at a fitness value of 0.865. These findings validate the effectiveness of the algorithm. Detailed iterative results for each subpopulation are provided in Tables 12 and 13.

Table 12. Results of aggregation operation (dominant subpopulation).

Elite Subpopulation	NBPE	Dev-WER	Test-WER	Training Time	WPS	Fitness
1	398	28.4%	30.59%	4501.978	66.9	0.846
	402	27.58%	29.53%	4467.74	67.1	0.853
	404	27.35%	29.36%	4454.104	67.5	0.856
	396	26.58%	29.12%	4453.273	67.8	0.861
	401	26.18%	28.64%	4397.113	68	0.866
2	346	28.56%	30.49%	4621.743	67.5	0.842
	349	27.82%	30.3%	4518.454	67.3	0.849
	353	27.49%	30.53%	4495.042	67.5	0.851
	347	27.42%	29.82%	4405.517	67.6	0.858
	352	27.54%	29.51%	4337.336	67.8	0.862
3	304	28.49%	30.49%	4733.265	63.1	0.824
	298	28.4%	30.46%	4797.737	65	0.827
	305	28.54%	30.38%	4563.825	66.8	0.842
	301	28.6%	30.4%	4596.438	68.4	0.845
	302	28.47%	30.46%	4571.535	68.3	0.847
4	197	28.57%	30.44%	4769.055	63.4	0.823
	200	28.49%	30.38%	4730.056	64.35	0.828
	199	28.56%	30.51%	4719.169	64.72	0.829
	198	28.4%	30.49%	4625.665	64.85	0.834
	196	28.42%	30.4%	4520.073	67.36	0.847
5	248	28.44%	30.38%	4912.812	60.98	0.810
	253	28.38%	30.38%	4965.732	63.5	0.817
	250	28.48%	30.42%	4945.569	64.45	0.819
	249	28.54%	30.69%	4843.549	67.36	0.831
	255	28.42%	30.4%	4618.713	64.6	0.834

Table 13. Results of aggregation operation (temporary subpopulation).

Temporary Subpopulation	NBPE	Dev-WER	Test-WER	Training Time	WPS	Fitness
1	627	29.05%	30.88%	5442.41	65.3	0.800
	628	28.48%	30.34%	5016.44	60.69	0.805
	631	28.54%	30.28%	4974.087	61.8	0.810
	630	28.54%	30.40%	4786.692	61.6	0.817
	629	28.51%	30.69%	4726.863	64.67	0.828
2	504	29.55%	31.29%	7031.657	63.8	0.751
	508	29.32%	31.05%	6474.046	62.5	0.761
	503	29.01%	30.74%	5882.4	61.8	0.776
	502	28.79%	30.48%	5435.942	60.9	0.789
	506	28.42%	30.36%	4882.962	57.5	0.801
3	715	30.10%	31.85%	9789.546	61.5	0.701
	713	29.80%	31.60%	8216.906	60.5	0.719
	715	29.51%	31.29%	7150.462	59.8	0.737
	715	29.25%	31.00%	6446.574	58.9	0.751
	711	28.99%	30.80%	5692.336	58.2	0.771
4	824	27.84%	30.30%	7421.311	34.59	0.664
	826	28.02%	30.45%	6677.349	34.2	0.677
	828	28.15%	30.60%	6071.508	33.8	0.69
	819	28.30%	30.80%	5505.059	33.5	0.705
	820	28.44%	31.00%	5006.507	33.2	0.721
5	96	27.75%	30.56%	8633.962	35.32	0.647
	98	27.90%	30.70%	7619.316	34.9	0.660
	99	28.05%	30.85%	6797.429	34.5	0.674
	94	28.20%	31.00%	6240.127	34.1	0.685
	95	28.35%	31.20%	5402.427	33.8	0.708

Finally, the NBPE parameter was optimized using the MEA-AGA strategy. The experimental results indicate that when NBPE is set to 401, the model achieves WERs of 26.18% on the Dev set and 28.64% on the Test set, with a decoding speed of 68 WPS. These outcomes demonstrate a favorable balance between recognition accuracy and decoding efficiency, thereby validating the practical value of the proposed optimization strategy for low-resource language recognition tasks.

5.4. Comparative Analysis of Model Performance

To systematically evaluate the effectiveness of the proposed optimization strategies, a series of comparative experiments were conducted. The evaluation metrics included Dev-WER, Test-WER, and WPS. The results are summarized in Table 14, where "√" indicates that the corresponding strategy was applied, and "— —" denotes that it was not.

Table 14. Comparative analysis of model performance under different optimization strategy.

Experiment	Data Augmentation	Transfer Learning Optimization	Parameter Optimization	Dev-WER	Test-WER	WPS
1	— —	— —	— —	53.77%	54.93%	49.22
2	√	— —	— —	44.01%	46.01%	49.83
3	√	√	— —	28.56%	30.42%	63.75
4	√	√	√	26.18%	28.64%	68.00

The experimental results are further visualized in Figure 15, which illustrates the performance variations under different optimization strategy combinations.

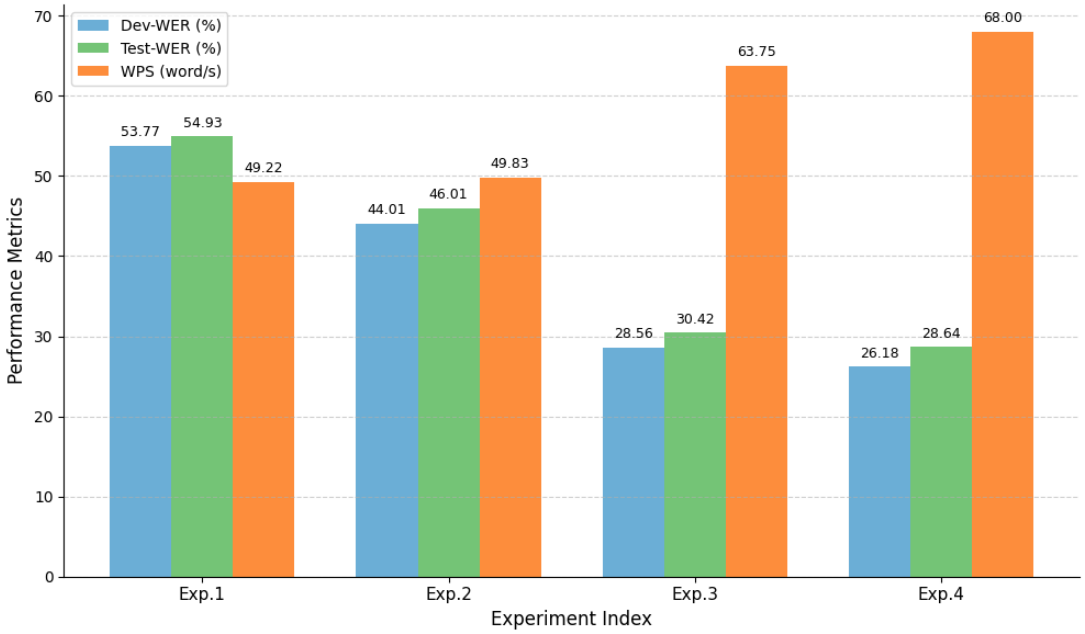


Figure 15. Comparative analysis of model performance under different optimization strategy.

Experiment 1 serves as the baseline model without any optimization strategies. In Experiment 2, the proposed data augmentation and partitioning strategy is applied, which ensures both the quantity and quality of the expanded data while also addressing class imbalance. As a result, both dev-WER and test-WER are significantly reduced. This demonstrates its effectiveness in improving recognition accuracy and generalization, particularly for low-resource languages like Tongan. Experiment 3 additionally incorporates the layer-wise adaptive transfer learning strategy. By leveraging the similarity between English (source language) and Tongan (target language) in phonetic structures, the model achieves faster convergence and a further reduction in WER. These results highlight the adaptability and practicality of the proposed method for low-resource tasks. In Experiment 4, the NBPE value is optimized to refine vocabulary granularity, thereby enhancing the model's capacity to represent Tongan linguistic features and improving both recognition accuracy and decoding speed.

In addition, Table 15 compares the performance of several mainstream models on the Tongan dataset. The results indicate that the proposed method surpasses the others in both recognition accuracy and inference efficiency, thereby demonstrating superior overall performance.

Table 15. Comparative analysis of speech recognition performance of state-of-the-art models.

Method	WER (%)		WPS (Word/s)
	Dev	Test	
DeepSpeech	31.50%	32.39%	48
Kaldi	31.54%	32.81%	55
HuBERT	27.50%	28.98%	62
Wav2Vec 2.0	27.20%	29.10%	65
XLS-R	26.80%	28.86%	64
Whisper	26.90%	29.02%	60
Mixformer (Our method)	26.18%	28.64%	68

To further evaluate the overall performance of the proposed model, a horizontal comparison was conducted between the recognition results for Tongan and those reported for other low-resource languages. As shown in Table 16 and Figure 16, although the available Tongan corpus is relatively limited (11.44 h), the achieved recognition accuracy remains within a reasonable range, demonstrating performance that consistent with the dataset scale. This finding further underscores the practicality and adaptability of the proposed method in low-resource scenarios.

Table 16. Performance comparison of speech recognition on other low-resource languages.

Language	Model Method	Duration (Hours)	Test-WER(%)
Quechua [55]	wav2letter++	198	22.75%
Māori [56]	Wav2Vec2	4	23%
Arabic [11]	Deepspeech	139	24.3%
Sesotho [57]	Kaldi	53.8	45.01%
Turkish [58]	CTC/Attention	16.36	48.01%
Tonga	Mixformer	11.44	28.64%

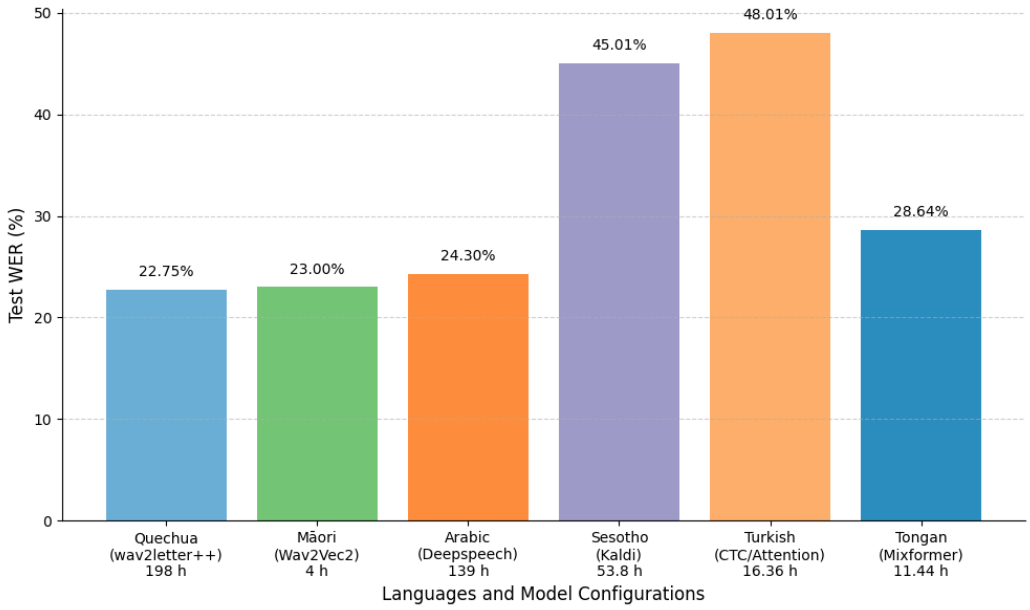


Figure 16. Performance comparison of speech recognition on low-resource languages.

6. Conclusions

With the rapid development of deep learning and artificial intelligence, speech recognition technology has been widely adopted worldwide. However, most existing research has focused on resource-rich languages such as English and Chinese, while studies on low-resource languages like Tongan remain scarce. These languages typically lack systematic corpora and effective modeling approaches, leading to challenges such as data scarcity, limited model transferability, and inadequate dictionary modeling mechanisms.

To overcome these limitations, this study first constructs a pretrained model using resource-rich English corpora and then fine-tunes it on the target language, Tongan, through a layer-wise adaptive transfer learning strategy. This approach enables both efficient and accurate speech recognition. Moreover, it provides valuable theoretical and technical support for the preservation of Tongan linguistic and cultural resources, while also promoting international cultural exchange. The main contributions of this study are summarized as follows:

(1) To address the issue of limited Tongan language data, this study proposes an SRA-DRF algorithm. By combining GAN-based synthetic data generation with traditional signal processing techniques, high-quality audio samples are produced. The effectiveness of the augmented data is validated through similarity comparison, demonstrating that the dataset size increases from 1.43 h to 11.44 h, with a similarity score of 90.63% between the augmented and original data. This effectively alleviates the problem of data scarcity. Furthermore, a weighted stratified sampling strategy is employed to achieve class-balanced partitioning, ensuring that the training, validation, and test sets maintain a complete and balanced sample distribution, thereby enabling the model to fully learn the features of each category.

(2) In the transfer learning phase, this study introduces a layer-wise adaptive strategy that preserves the low-level general features of the pretrained model while dynamically adjusting the learning rates of the higher layers according to loss values and source–target language similarity. Fine-tuning is applied primarily to the higher layers. The effectiveness and rationality of this strategy are further demonstrated through CKA similarity matrix analysis, which reveals distinct patterns of hierarchical feature adaptation.

(3) For optimizing the critical NBPE parameter in Tongan dictionary construction, this study proposes the MEA-AGA optimization algorithm. The NBPE value is optimized from an initial setting of 300 to 401. With this configuration, the optimized model achieves a Dev-WER of 26.18% and a Test-WER of 28.64% on the Tongan dataset, along with a decoding speed (WPS) of 68. Compared with the baseline model (Dev-WER 53.77%, Test-WER 54.93%, WPS 49.22), these results represent approximately 51.3% and 47.9% relative reductions in WER, and a 38.2% increase in decoding speed. These results demonstrate substantial improvements in both recognition accuracy and inference efficiency. Compared with mainstream approaches, the proposed method exhibits clear advantages and achieves performance within a reasonable accuracy range for low-resource language speech recognition.

Beyond its technical contributions, the development of speech recognition for low-resource languages such as Tongan also carries important ethical and societal significance. From a cultural perspective, it supports linguistic diversity by promoting the digital preservation and accessibility of endangered languages, helping safeguard intangible cultural heritage and enabling inclusive global communication. From an ethical standpoint, enhancing AI inclusivity helps narrow the digital divide and prevent the marginalization of smaller linguistic groups. At the societal level, improving recognition for minority languages contributes to cultural sustainability and supports education and social inclusion in multilingual regions.

7. Future Work

Although the proposed transfer strategy and parameter optimization algorithm have achieved promising results in Tongan speech recognition, several limitations remain. Future research can therefore be pursued in the following directions:

(1) Dataset Expansion: The current Tongan speech corpus remains relatively small. Although the proposed data augmentation and balanced partitioning strategies have partially alleviated data scarcity and mitigated imbalance to some extent, the corpus size and distribution are still limited compared with mainstream languages. This constraint may affect the model's generalization performance across phonetic categories. Future work should prioritize the collection of larger and more balanced Tongan speech datasets with higher quality to further enhance model robustness and generalizability.

(2) Optimization of Transfer Strategies: Although the proposed framework effectively adapts pretrained models to low-resource speech recognition, it still relies on su-

pervised fine-tuning. This dependency may limit adaptability when labeled data are scarce or domain variations occur. Future research could explore more flexible transfer learning paradigms to overcome these limitations. For instance, unsupervised or semi-supervised adaptation strategies could reduce reliance on annotated data, while zero-shot or few-shot learning mechanisms may enhance model generalization under extremely low-resource conditions.

(3) Broader Source Language Selection: In this study, English was selected as the source language due to its phonetic similarity to Tongan. Future research could investigate multilingual joint transfer strategies to enhance cross-lingual generalization. Moreover, given the demonstrated effectiveness of the proposed method for Tongan, its applicability to other low-resource languages deserves systematic evaluation to validate its universality and scalability.

Author Contributions: Conceptualization, J.G. and D.J.; methodology, J.G. and Z.L.; software, Z.L. and W.Z.; validation, J.G. and D.J.; formal analysis, J.G.; investigation, Z.H. and W.Z.; resources, N.W. and R.C.; data curation, N.W. and W.Z.; writing—original draft preparation, J.G.; writing—review and editing, D.J.; visualization, J.G.; supervision, D.J.; project administration, D.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key R&D Program (Project No. 2023YFF0612100).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets generated and analyzed during the current study are not publicly available because the research is still ongoing, and early disclosure may affect subsequent work. However, the data can be made available from the corresponding author upon reasonable request.

Acknowledgments: The authors express their sincere gratitude to Jia for his invaluable guidance and support throughout this research. We also thank the students who participated in this project for their efforts and dedication, which were essential to the success of this study. Generative AI tools were used exclusively to improve the language and grammar of this manuscript. All scientific content, analyses, interpretations, and conclusions were entirely conceived, written, and verified by the authors.

Conflicts of Interest: Author Junhao Geng was employed by the company Beijing Research Institute of Automation for Machinery Industry Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Deng, L.; Hinton, G.; Kingsbury, B. New types of deep neural network learning for speech recognition and related applications: An overview. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, BC, Canada, 26–31 May 2013; pp. 8599–8603. [CrossRef]
2. Haddow, B.; Bawden, R.; Barone, A.V.M.; Helcl, J.; Birch, A. Survey of Low-Resource Machine Translation. *Comput. Linguist.* **2022**, *48*, 673–732. [CrossRef]
3. Pan, S.J.; Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **2009**, *22*, 1345–1359. [CrossRef]
4. Graves, A.; Mohamed, A.; Hinton, G. Speech recognition with deep recurrent neural networks. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, BC, Canada, 26–31 May 2013; pp. 6645–6649. [CrossRef]
5. Cho, J.; Baskar, M.K.; Li, R.; Wiesner, M.; Mallidi, S.H.; Yalta, N. Multilingual Sequence-to-Sequence Speech Recognition: Architecture, Transfer Learning, and Language Modeling. In Proceedings of the IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 18–21 December 2018; pp. 521–527. [CrossRef]
6. Singh, S.; Hou, F.; Wang, R. A Novel Self-Training Approach for Low-Resource Speech Recognition. *arXiv* **2023**, arXiv:2308.05269.

7. Mukhamadiyev, A.; Khujayarov, I.; Djuraev, O.; Cho, J. Automatic Speech Recognition Method Based on Deep Learning Approaches for Uzbek Language. *Sensors* **2022**, *22*, 3683. [CrossRef] [PubMed]
8. van der Westhuizen, E.; Kamper, H.; Menon, R.; Quinn, J.; Niesler, T. Feature Learning for Efficient ASR-Free Keyword Spotting in Low-Resource Languages. *Comput. Speech Lang.* **2022**, *71*, 101275. [CrossRef]
9. Mamyrbayev, O.; Oralbekova, D.; Kydyrbekova, A.; Turdalykyzy, T.; Bekarystankyzy, A. End-to-End Model Based on RNN-T for Kazakh Speech Recognition. In Proceedings of the 2021 3rd International Conference on Computer Communication and the Internet (ICCCI), Nagoya, Japan, 25–27 June 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 163–167. [CrossRef]
10. Zhi, T.; Shi, Y.; Du, W.; Li, G.; Wang, D. M2ASR-MONGO: A Free Mongolian Speech Database and Accompanied Baselines. In Proceedings of the 2021 24th Conference of the Oriental COCOSDA International Committee for the Co-Ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA), Singapore, 18–20 November 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 140–145. [CrossRef]
11. Oukas, N.; Zerrouki, T.; Haboussi, S.; Djettou, H. Arabic Speech Recognition Using Deep Learning and Common Voice Dataset. In Proceedings of the 2022 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT), Sakheer, Bahrain, 20–21 November 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 642–647. [CrossRef]
12. Changrampadi, M.H.; Shahina, A.; Narayanan, M.B.; Khan, A.N. End-to-End Speech Recognition of Tamil Language. *Intell. Autom. Soft Comput.* **2022**, *32*, 1049–1064. [CrossRef]
13. Shetty, V.M.; Mysore Sathyendra, N.J. Improving the Performance of Transformer-Based Low-Resource Speech Recognition for Indian Languages. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 8279–8283. [CrossRef]
14. Pan, L.; Li, S.; Wang, L.; Dang, J. Effective Training of End-to-End ASR Systems for Low-Resource Lhasa Dialect of Tibetan Language. In Proceedings of the 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Lanzhou, China, 18–21 November 2019; pp. 1152–1156. [CrossRef]
15. Anoop, C.S.; Ramakrishnan, A.G. CTC-Based End-to-End ASR for the Low-Resource Sanskrit Language with Spectrogram Augmentation. In Proceedings of the 2021 National Conference on Communications (NCC), Bangalore, India, 27–30 July 2021; pp. 1–6. [CrossRef]
16. Taylor, M.E.; Stone, P. Transfer Learning for Reinforcement Learning Domains: A Survey. *J. Mach. Learn. Res.* **2009**, *10*, 1633–1685.
17. Raghu, M.; Zhang, C.; Kleinberg, J.; Bengio, S. Transfusion: Understanding Transfer Learning for Medical Imaging. In *Advances in Neural Information Processing Systems 32*; Neural Information Processing Systems Foundation, Inc.: La Jolla, CA, USA, 2019.
18. Hussain, M.; Bird, J.J.; Faria, D.R. A Study on CNN Transfer Learning for Image Classification. In Proceedings of the 18th UK Workshop on Computational Intelligence (UKCI 2018), Nottingham, UK, 5–7 September 2018; pp. 191–202. [CrossRef]
19. Li, W.; Huang, R.; Li, J.; Liao, Y.; Chen, Z.; He, G.; Yan, R.; Gryllias, K. A Perspective Survey on Deep Transfer Learning for Fault Diagnosis in Industrial Scenarios: Theories, Applications and Challenges. *Mech. Syst. Signal Process.* **2022**, *167*, 108487. [CrossRef]
20. Mainzinger, J.; Levow, G.-A. Fine-Tuning ASR Models for Very Low-Resource Languages: A Study on Mvskoke. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), Bangkok, Thailand, 11–16 August 2024; pp. 76–82. [CrossRef]
21. Abdullah, A.A.; Tabibian, S.; Veisi, H.; Mahmudi, A.; Rashid, T. End-to-End Transformer-Based Automatic Speech Recognition for Northern Kurdish: A Pioneering Approach. *arXiv* **2024**, arXiv:2410.16330.
22. Cui, J.; Yang, J. Front-End Fusion and Large-Scale Weakly Supervised Decoding Module Based Myanmar Speech Recognition. In Proceedings of the 2023 International Conference on Asian Language Processing (IALP), Singapore, 17–19 November 2023; pp. 348–352. [CrossRef]
23. Zeng, Z.; Pham, V.T.; Xu, H.; Khassanov, Y.; Chng, E.S.; Ni, C. Leveraging Text Data Using Hybrid Transformer-LSTM Based End-to-End ASR in Transfer Learning. In Proceedings of the 2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP), Hong Kong, China, 24–27 January 2021; pp. 1–5. [CrossRef]
24. Hjortnaes, N.; Arkhangelskiy, T.; Partanen, N.; Rießler, M.; Tyers, F. Improving the Language Model for Low-Resource ASR with Online Text Corpora. In Proceedings of the 1st Joint SLTU and CCURL Workshop (SLTU-CCURL 2020), Marseille, France, 11–12 May 2020; pp. 1–6.
25. Bansal, S.; Kamper, H.; Livescu, K.; Lopez, A.; Goldwater, S. Pre-training on High-Resource Speech Recognition Improves Low-Resource Speech-to-Text Translation. *arXiv* **2018**, arXiv:1809.01431.
26. Lee, H.; Yoon, H.-W.; Kim, J.-H.; Kim, J.-M. Cross-Lingual Transfer Learning for Phrase Break Prediction with Multilingual Language Model. *arXiv* **2023**, arXiv:2306.02579. [CrossRef]
27. Baller, T.; Bennett, K.; Hamilton, H.J. Transfer Learning and Language Model Adaption for Low Resource Speech Recognition. In Proceedings of the 34th Canadian Conference on Artificial Intelligence, Vancouver, BC, Canada, 25–28 May 2021. [CrossRef]
28. Anoop, C.S.; Ramakrishnan, A.G. Meta-Learning for Indian Languages: Performance Analysis and Improvements with Linguistic Similarity Measures. *IEEE Access* **2023**, *11*, 82050–82064. [CrossRef]

29. Yadav, H.; Sitaram, S. A Survey of Multilingual Models for Automatic Speech Recognition. In Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2022), Paris, France, 20–25 June 2022; pp. 5071–5079.
30. Yu, Z.; Zhang, Y.; Qian, K.; Fu, Y.; Lin, Y. Master-ASR: Achieving Multilingual Scalability and Low-Resource Adaptation in ASR with Modular Learning. *arXiv* **2023**, arXiv:2306.15686.
31. Yan, B.; Wiesner, M.; Klejch, O.; Jyothi, P.; Watanabe, S. Towards Zero-Shot Code-Switched Speech Recognition. *arXiv* **2022**, arXiv:2211.01458.
32. Li, X.; Dalmia, S.; Mortensen, D.R.; Metze, F.; Black, A.W. Zero-Shot Learning for Speech Recognition with Universal Phonetic Model. In Proceedings of the International Conference on Learning Representations (ICLR 2019), New Orleans, LA, USA, 6–9 May 2019.
33. Hsiao, R.; Deng, L.; McDermott, E.; Travadi, R.; Zhuang, X. Optimizing Byte-Level Representation for End-to-End ASR. In Proceedings of the 2024 IEEE Spoken Language Technology Workshop (SLT), Macao, China, 2–5 December 2024; pp. 462–467. [CrossRef]
34. Deng, L.; Hsiao, R.; Ghoshal, A. Bilingual End-to-End ASR with Byte-Level Subwords. In Proceedings of the ICASSP 2022—IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 6417–6421. [CrossRef]
35. Ma, J.; Duo, L.; Wei, G.; Tang, J. End-to-End Speech Recognition Based on Thresholded BPE-Dropout Multi-Task Learning. *J. Jilin Univ. (Sci. Ed.)* **2024**, *62*, 674–682. [CrossRef]
36. Cai, Y.; Wang, C.; Ren, D.; Zhu, Y.; Zhang, J.; Nima, Z. End-to-End Tibetan Speech Recognition Study Based on Byte Pair Coding. In Proceedings of the 23rd Chinese National Conference on Computational Linguistics (CCL 2024), Taiyuan, China, 25–28 July 2024; Volume 1, pp. 305–313.
37. Gaurav, M.S.A.I.S. Investigating Transformer Architecture for Automatic Speech Recognition and Machine Translation in Indian Languages. Ph.D. Thesis, Indian Institute of Technology Madras, Chennai, India, 2021.
38. Abu, T.; Shi, Y.; Zheng, T.F.; Wang, D. Sagalee: An Open Source Automatic Speech Recognition Dataset for Oromo Language. In Proceedings of the 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 6–11 April 2025. [CrossRef]
39. Patel, T.; Scharenborg, O. Using Cross-Model Learnings for the Gram Vaani ASR Challenge 2022. In Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH 2022), Incheon, Republic of Korea, 18–22 September 2022; pp. 4880–4884.
40. Abayomi-Alli, O.O.; Damaševičius, R.; Qazi, A.; Adedoyin-Olowe, M.; Misra, S. Data Augmentation and Deep Learning Methods in Sound Classification: A Systematic Review. *Electronics* **2022**, *11*, 3795. [CrossRef]
41. Nugroho, K.; Noersasongko, E.; Purwanto; Muljono; Setiadi, D.R.I.M. Enhanced Indonesian Ethnic Speaker Recognition Using Data Augmentation Deep Neural Network. *J. King Saud Univ.—Comput. Inf. Sci.* **2022**, *34*, 4375–4384. [CrossRef]
42. Ramesh, V.; Vatanparvar, K.; Nemati, E.; Nathan, V.; Rahman, M.M.; Kuang, J. CoughGAN: Generating Synthetic Coughs That Improve Respiratory Disease Classification. In Proceedings of the 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Montréal, QC, Canada, 20–24 July 2020; pp. 5682–5688. [CrossRef]
43. Esmaeilpour, M.; Cardinal, P.; Koerich, A.L. Unsupervised Feature Learning for Environmental Sound Classification Using Weighted Cycle-Consistent Generative Adversarial Network. *Appl. Soft Comput.* **2020**, *86*, 105912. [CrossRef]
44. Yella, N.; Rajan, B. Data Augmentation Using GAN for Sound-Based COVID-19 Diagnosis. In Proceedings of the 2021 11th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), Cracow, Poland, 22–25 September 2021; Volume 2, pp. 606–609. [CrossRef]
45. Raj, S.; Anant, A.; Suryadevara, H.; Kumar, R.K. Hybrid Deep Learning Architecture with K-means Clustering for Weapon Detection in CCTV Surveillance. In Proceedings of the 2024 IEEE International Conference on Computer Vision and Machine Intelligence (CVMI), Prayagraj, India, 19–20 October 2024; pp. 1–6. [CrossRef]
46. Goel, A.; Majumdar, A.; Chouzenoux, E.; Chierchia, G. Deep Convolutional K-means Clustering. In Proceedings of the 2022 IEEE International Conference on Image Processing (ICIP), Bordeaux, France, 16–19 October 2022; pp. 211–215. [CrossRef]
47. Guo, W.; Lin, K.; Ye, W. Deep Embedded K-means Clustering. In Proceedings of the 2021 International Conference on Data Mining Workshops (ICDMW), Auckland, New Zealand, 7–10 December 2021; pp. 686–694. [CrossRef]
48. Nam, Y.; Han, S. Random Forest Variable Importance-Based Selection Algorithm in Class Imbalance Problem. *J. Classif.* **2025**, 1–14. [CrossRef]
49. Kheddar, H.; Himeur, Y.; Al-Maadeed, S.; Amira, A.; Bensaali, F. Deep Transfer Learning for Automatic Speech Recognition: Towards Better Generalization. *Knowl.-Based Syst.* **2023**, *277*, 110851. [CrossRef]
50. Davari, M.R.; Horoi, S.; Natik, A.; Lajoie, G.; Wolf, G.; Belilovsky, E. Reliability of CKA as a Similarity Measure in Deep Learning. *arXiv* **2022**, arXiv:2210.16156. [CrossRef]
51. Stalker, P. *A Guide to Countries of the World*; Oxford University Press: Oxford, UK, 2010.
52. Moala, K. *Search of the Friendly Islands*; Pasifika Foundation Press: Suva, Fiji, 2009.

53. Tod, D. Rhythm in the Kingdom: A Variationist Analysis of Speech Rhythm in Tongan English. *Engl. Lang. Linguist.* **2024**, *28*, 203–225. [CrossRef]
54. Geng, J.; Jia, D.; He, Z.; Wu, N.; Li, Z. Enhanced Conformer-Based Speech Recognition via Model Fusion and Adaptive Decoding with Dynamic Rescoring. *Appl. Sci.* **2024**, *14*, 11583. [CrossRef]
55. Zevallos, R.; Bel, N.; Cámbara, G.; Farrús, M.; Luque, J. Data Augmentation for Low-Resource Quechua ASR Improvement. *arXiv* **2022**, arXiv:2207.06872. [CrossRef]
56. Coto-Solano, R.; Nicholas, S.A.; Datta, S.; Quint, V.; Wills, P.; Powell, E.N.; Koka'ua, L.; Tanveer, S.; Feldman, I. Development of Automatic Speech Recognition for the Documentation of Cook Islands Māori. In Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2022), Marseille, France, 20–25 June 2022; pp. 3872–3882.
57. Mak, F.; Govender, A.; Badenhorst, J. Exploring ASR Fine-Tuning on Limited Domain-Specific Data for Low-Resource Languages. *J. Digit. Humanit. Assoc. S. Afr.* **2024**, *5*. [CrossRef]
58. Ren, Z.; Yolwas, N.; Slamu, W.; Cao, R.; Wang, H. Improving Hybrid CTC/Attention Architecture for Agglutinative Language Speech Recognition. *Sensors* **2022**, *22*, 7319. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.