Deep Learning Approaches for Arabic Phoneme /r/ Pronunciation Disorder Classification: A Comparative Study of CNN-LSTM-Attention and Advanced Residual Networks

Baraa Khanfar*
*Department of Computer Engineering
Birzeit University
Ramallah, Palestine
Email: 1210640@student.birzeit.edu

AbdAl Moumen Kahla†
†Department of Computer Engineering
Birzeit University
Ramallah, Palestine
Email: 1210988@student.birzeit.edu

Hmazah Alqam‡
‡Department of Computer Engineering
Birzeit University
Selfit, Palestine
Email: 1211173@student.birzeit.edu

***Abstract*—Arabic phoneme /r/ (r-sound) pronunciation disorders are among the most common speech impediments in Arabic-speaking children. This paper presents a comprehensive comparative study of two deep learning architectures for automated classification of Arabic r-sound pronunciation disorders. We developed and evaluated two distinct models: (1) a CNN-LSTM-Attention hybrid architecture and (2) an Advanced Residual Network with multi-scale feature extraction and attention mechanisms. Both models classify audio samples into five categories: deletion, distortion, normal pronunciation, substitution with /gh/, and substitution with /l/. Our experimental results demonstrate that the Advanced Residual Network achieves superior performance with 91.5% test accuracy compared to 87.5% for the CNN-LSTM-Attention model. The study utilizes a comprehensive dataset of Arabic speech samples focusing on r-sound pronunciations at word beginnings. These findings contribute to the development of automated speech therapy tools and early intervention systems for Arabic speech disorders.**

***Index Terms*—speech disorder, Arabic phonetics, deep learning, CNN-LSTM, residual networks, attention mechanism, audio classification**

I. Introduction

Speech disorders significantly impact children's communication abilities and academic performance. Among Arabic phonemes, the sound /r/ (corresponding to [r] in the International Phonetic Alphabet) presents particular challenges for many Arabic-speaking children [1]. The complexity of the Arabic r-sound stems from its trill characteristics and position-dependent variations within words.

Traditional speech therapy relies heavily on manual assessment by trained professionals, which can be time-consuming, subjective, and may not be readily available in all regions. The development of automated classification systems for speech disorders represents a significant advancement in speech-language pathology, offering objective, consistent, and accessible diagnostic tools.

This study addresses the critical need for automated detection and classification of Arabic r-sound pronunciation disorders through deep learning approaches. We focus specifically on r-sound disorders occurring at the beginning of words, as this position is considered most challenging for affected children and provides clear acoustic signatures for machine learning analysis.

Our main contributions include: (1) Implementation and comparison of two state-of-the-art deep learning architectures for Arabic speech disorder classification, (2) Comprehensive analysis of model performance using multiple evaluation metrics, (3) Investigation of attention mechanisms in speech disorder detection, and (4) Provision of detailed computational requirements and training characteristics for practical deployment considerations.

II. Background and Related Work

A. Arabic R-Sound Pronunciation Disorders

The Arabic phoneme /r/ is classified as an alveolar trill that requires precise tongue positioning and airflow control. Children with r-sound disorders typically exhibit one of four patterns: (1) deletion - complete omission of the r-sound, (2) distortion - imprecise production maintaining some r-like qualities, (3) substitution with /gh/ - replacing r with a fricative sound, and (4) substitution with /l/ - replacing r with a lateral approximant [2].

These disorders can significantly impact intelligibility and may persist into adulthood without proper intervention. Early detection and classification of the specific disorder type is crucial for developing targeted therapy approaches.

B. Deep Learning in Speech Disorder Detection

Recent advances in deep learning have shown promising results in automated speech disorder detection. Convolutional Neural Networks (CNNs) have proven effective for extracting spatial features from spectrograms, while Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks excel at modeling temporal dependencies in speech signals.

Attention mechanisms have emerged as powerful tools for focusing on relevant acoustic features and time segments. Multi-head attention, originally developed for natural language processing, has been successfully adapted for speech processing tasks [3].

Residual networks (ResNets) have demonstrated superior performance in various audio classification tasks by addressing the vanishing gradient problem and enabling training of deeper networks. The combination of residual connections with attention mechanisms has shown particular promise in speech-related applications.

III. Methodology

A. Dataset Description

Our dataset consists of Arabic speech samples collected from speakers with varying degrees of r-sound pronunciation abilities [7]. The dataset is organized into training and testing sets, with five distinct classes:

- **Deletion**: Complete omission of the r-sound
- **Distortion**: Imprecise r-sound production
- **Normal**: Correct r-sound pronunciation
- **Substitution_gh**: R-sound replaced with /gh/
- **Substitution_l**: R-sound replaced with /l/

All audio samples are preprocessed to 16 kHz sampling rate with varying durations. The dataset maintains balanced class distribution to ensure unbiased model training.

B. Preliminary Model Exploration

Prior to developing the main architectures, we conducted preliminary experiments with traditional machine learning and basic deep learning approaches to establish baseline performance. These initial models utilized a custom feature extraction pipeline combining MFCC and mel-spectrogram features.

Three baseline models were evaluated:

- **Random Forest**: 100-estimator ensemble achieving 67.8% accuracy with 3.52 MB memory usage and 2.86s training time
- **Support Vector Machine**: RBF kernel-based classifier achieving 72.0% accuracy with 21.32 MB memory usage and 0.53s training time
- **Deep Neural Network**: 4-layer network achieving 64.0% accuracy with 163.25 MB memory usage and 19.88s training time

While computationally efficient, these baseline approaches demonstrated significant performance limitations, particularly in distinguishing between normal pronunciation and substitution disorders. The SVM showed the best baseline performance but still fell short of clinical deployment requirements, motivating the development of the advanced architectures presented in this study.

C. Feature Extraction

Both main models employ sophisticated feature extraction techniques:

1) Model 1 (CNN-LSTM-Attention): Features are extracted using a combination of:

- **MFCC**: 13 Mel-Frequency Cepstral Coefficients with 1024-point FFT and 256-sample hop length
- **Mel Spectrogram**: 64 mel-filter banks covering 0-8000 Hz frequency range
- **Pre-emphasis**: High-pass filtering with coefficient 0.97
- **Normalization**: Per-feature standardization with mean subtraction and standard deviation normalization

The final feature representation combines 13 MFCC coefficients with 32 mel-band features, resulting in a 45×63 dimensional feature matrix.

2) Model 2 (Advanced Residual Network): This model employs direct spectral feature extraction:

- **Mel Spectrogram**: 128 mel-filter banks with 1024-point FFT
- **MFCC**: 13 coefficients using the same mel-bank configuration
- **Multi-scale Processing**: Parallel convolutions with kernels of sizes 3, 5, 7, and 11
- **Logarithmic Transformation**: Log mel-spectrograms for improved dynamic range

D. Model Architectures

1) Model 1: CNN-LSTM-Attention Architecture: The first model implements a hybrid architecture combining convolutional, recurrent, and attention mechanisms as illustrated in Figure 1:

$$\begin{aligned}
\mathbf{h}_{conv} &= \text{MaxPool}(\text{ReLU}(\text{BN}(\text{Conv1D}(\mathbf{x})))) \\
\mathbf{h}_{lstm} &= \text{BiLSTM}(\mathbf{h}_{conv}) \\
\mathbf{a} &= \text{Softmax}(\text{Tanh}(\mathbf{W_a}\mathbf{h}_{lstm})) \\
\mathbf{c} &= \sum_t \mathbf{a}_t \mathbf{h}_{lstm,t} \\
\mathbf{y} &= \text{Softmax}(\mathbf{W_o}\mathbf{c})
\end{aligned} \tag{1}$$

where $\mathbf{x}$ represents input features, $\mathbf{h}_{conv}$ and $\mathbf{h}_{lstm}$ are convolutional and LSTM hidden states respectively, $\mathbf{a}$ represents attention weights, $\mathbf{c}$ is the context vector, and $\mathbf{y}$ is the final classification output.

Key architectural components:

- Two 1D convolutional layers (32 and 64 filters)
- Bidirectional LSTM with 32 hidden units per direction
- Custom attention mechanism with 64-dimensional intermediate representation
- Dropout regularization (0.4-0.6 rates)
- Total parameters: 42,310

2) Model 2: Advanced Residual Network: The second model employs a sophisticated residual architecture with multi-head attention as shown in Figure 2:

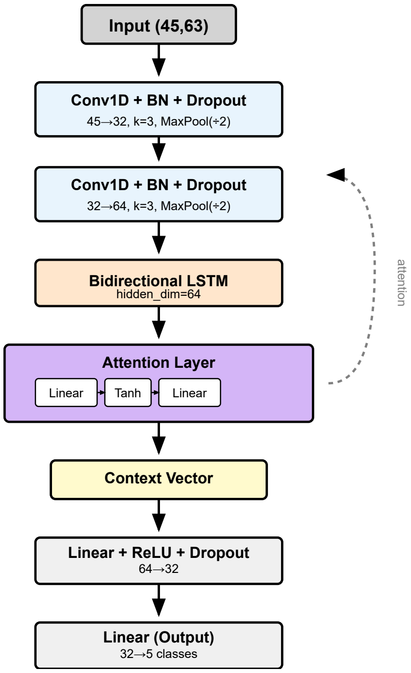

Fig. 1: CNN-LSTM-Attention Model Architecture (Model 1)

$$\begin{aligned}
\mathbf{F}_{multi} &= \text{Concat}[\text{Conv}_3(\mathbf{x}), \text{Conv}_5(\mathbf{x}), \text{Conv}_7(\mathbf{x}), \text{Conv}_{11}(\mathbf{x})] \\
\mathbf{R}_i &= \mathbf{F}_{i-1} + \mathcal{F}(\mathbf{F}_{i-1}, \mathbf{W}_i) \\
\mathbf{A} &= \text{MultiHeadAttention}(\mathbf{R}_n, \mathbf{R}_n, \mathbf{R}_n) \\
\mathbf{p} &= \text{Concat}[\text{GlobalAvgPool}(\mathbf{A}), \text{GlobalMaxPool}(\mathbf{A})] \\
\mathbf{y} &= \text{Classifier}(\mathbf{p})
\end{aligned} \tag{2}$$

where $\mathbf{F}_{multi}$ represents multi-scale features, $\mathbf{R}_i$ are residual block outputs, $\mathbf{A}$ is the attended feature representation, and $\mathbf{p}$ combines global pooling operations.

Architecture specifications:

- Multi-scale feature extractor with parallel convolutions
- Eight residual blocks with increasing channel dimensions (256→512→1024)
- 8-head multi-head attention mechanism
- Dual global pooling (average and maximum)
- Three-layer classification head with dropout
- Total parameters: 28,718,981

E. Training Configuration

Both models were trained using similar optimization strategies:

- **Loss Function**: CrossEntropyLoss with label smoothing ($\alpha = 0.1$)
- **Optimizer**: AdamW with weight decay (1e-3 for Model 1, 1e-2 for Model 2)
- **Learning Rate**: 5e-4 (Model 1), 1e-3 (Model 2)
- **Scheduler**: CosineAnnealingWarmRestarts ($T_0 = 10$, $T_{mult} = 2$)

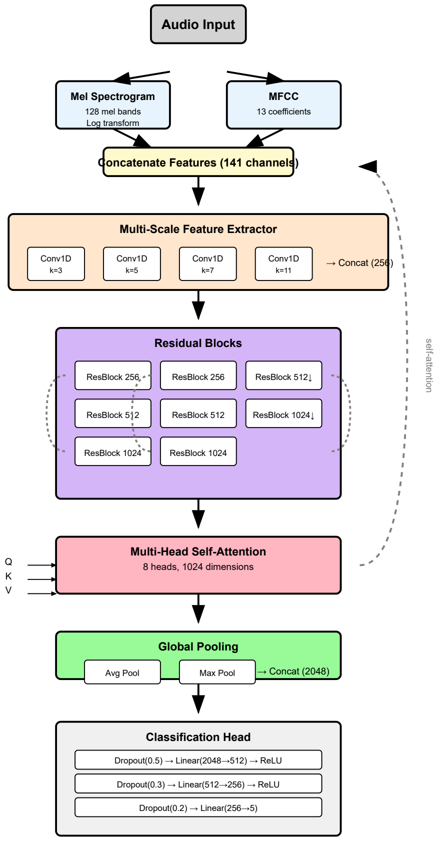

Fig. 2: Advanced Residual Network Architecture (Model 2)

- **Batch Size**: 16
- **Data Augmentation**: Noise addition, time shifting, frequency masking
- **Regularization**: Dropout, gradient clipping (max norm = 0.5)

IV. Experiments and Results

A. Experimental Setup

Experiments were conducted on an NVIDIA GeForce RTX 4090 Laptop GPU with 16GB memory. Training was performed using PyTorch 2.0 with CUDA acceleration. Both models were trained for sufficient epochs with early stopping mechanisms to prevent overfitting.

B. Performance Metrics

1) Model 1 (CNN-LSTM-Attention) Results: The CNN-LSTM-Attention model achieved the following performance:

Training characteristics:

- Training time: 680.35 seconds (11.34 minutes)
- Epochs completed: 49
- Final training accuracy: 91.64%
- Final validation accuracy: 85.75%
- Overfitting gap: 5.89%
- Peak GPU memory usage: 36.75 MB
- System memory increase: 676.44 MB

TABLE I: Model 1 Performance Metrics

Class	Precision	Recall	F1-Score	Support
Deletion	1.00	1.00	1.00	80
Distortion	0.99	0.99	0.99	80
Normal	0.84	0.61	0.71	80
Substitution_gh	0.68	0.86	0.76	80
Substitution_l	0.91	0.91	0.91	80
Accuracy		**0.875**		**400**
Macro Avg	**0.88**	**0.88**	**0.87**	**400**
Weighted Avg	**0.88**	**0.88**	**0.87**	**400**

2) Model 2 (Advanced Residual Network) Results: The Advanced Residual Network demonstrated superior performance:

TABLE II: Model 2 Performance Metrics

Class	Precision	Recall	F1-Score	Support
Deletion	0.96	0.96	0.96	80
Distortion	0.99	0.94	0.96	80
Normal	0.86	0.94	0.90	80
Substitution_gh	0.97	0.76	0.85	80
Substitution_l	0.83	0.97	0.90	80
Accuracy		**0.915**		**400**
Macro Avg	**0.92**	**0.91**	**0.91**	**400**
Weighted Avg	**0.92**	**0.92**	**0.91**	**400**

Training characteristics:

- Training time: 722.31 seconds (12.04 minutes)
- Epochs completed: 100
- Final training accuracy: 98.55%
- Best validation accuracy: 91.50%
- Final test accuracy: 91.50%
- Peak GPU memory usage: 573.59 MB

C. Comparative Analysis

TABLE III: Model Comparison Summary

Metric	Model 1	Model 2
Test Accuracy	87.5%	91.5%
Total Parameters	42,310	28,718,981
Training Time (min)	11.34	12.04
Peak GPU Memory (MB)	36.75	573.59
Overfitting Gap	5.89%	∼7%
Best Class F1-Score	1.00 (Deletion)	0.96 (Multiple)
Worst Class F1-Score	0.71 (Normal)	0.85 (Substitution_gh)

D. Confusion Matrix Analysis

Both models show distinct performance patterns across classes:

Model 1 Observations:

- Perfect classification of deletion and near-perfect for distortion
- Challenges with normal pronunciation classification (61% recall)
- Confusion between normal and substitution classes
- Strong performance on substitution_l classification

Model 2 Observations:

- More balanced performance across all classes
- Improved normal pronunciation classification (94% recall)
- Better discrimination between substitution types
- Slight reduction in deletion classification perfection but more robust overall

E. Training Dynamics

The training history reveals important insights:

Model 1: Shows steady convergence with minimal overfitting in early epochs, followed by slight overfitting in later stages. The overfitting gap remains controlled at approximately 6%.

Model 2: Demonstrates more aggressive training with higher capacity, achieving near-perfect training accuracy but maintaining reasonable generalization. The attention mechanism and residual connections effectively prevent severe overfitting.

V. Discussion

A. Model Performance Analysis

The Advanced Residual Network (Model 2) outperforms the CNN-LSTM-Attention model (Model 1) by 4 percentage points in test accuracy. Both main models significantly outperform the preliminary baseline approaches (Random Forest: 67.8%, SVM: 72.0%, Deep NN: 64.0%), demonstrating the effectiveness of specialized architectures for speech disorder classification. The improvement can be attributed to several factors:

1) **Feature Learning Capacity**: The residual network's deeper architecture with 28.7M parameters provides significantly more representational power compared to Model 1's 42K parameters.
2) **Multi-scale Feature Extraction**: The parallel convolutions with different kernel sizes enable capture of both fine-grained and broad temporal patterns in the audio signal.
3) **Attention Mechanism**: The multi-head attention in Model 2 provides more sophisticated feature weighting compared to the simpler attention in Model 1.
4) **Residual Connections**: These connections facilitate gradient flow and enable training of deeper networks without degradation.

B. Class-specific Performance

Both models excel at detecting deletion and distortion disorders, likely due to the distinct acoustic signatures of these conditions. The most challenging class for both models is the differentiation between normal pronunciation and various substitution types, suggesting that:

- Normal Arabic r-sound exhibits acoustic variability
- Substitution disorders may retain some spectral characteristics of the target phoneme
- Additional acoustic features or longer temporal context might improve discrimination

C. Computational Considerations

Model 1 offers significant advantages in resource-constrained environments:

- 680× fewer parameters enable deployment on mobile devices
- 15× lower GPU memory requirements
- Comparable training time despite architectural differences
- Reasonable performance trade-off for practical applications

Model 2's superior performance comes at the cost of increased computational requirements, making it more suitable for server-based deployment or high-performance computing environments.

D. Clinical Implications

The achieved performance levels suggest both models could serve as valuable tools in clinical settings:

- **Screening Tool**: High accuracy enables initial assessment and triage
- **Progress Monitoring**: Consistent classification supports therapy progress tracking
- **Objective Assessment**: Reduces subjectivity in disorder classification
- **Accessibility**: Automated systems can extend specialized assessment to underserved areas

However, clinical deployment requires careful consideration of model limitations and the need for human expert oversight.

VI. Conclusion and Future Work

This study successfully demonstrates the effectiveness of deep learning approaches for Arabic r-sound pronunciation disorder classification. The Advanced Residual Network achieves 91.5% accuracy, representing a significant advancement in automated speech disorder detection for Arabic speakers.

Key contributions include:

1) Comprehensive comparison of two distinct architectural approaches
2) Detailed analysis of computational requirements and practical deployment considerations
3) Demonstration of attention mechanisms' effectiveness in speech disorder classification
4) Establishment of baseline performance metrics for Arabic r-sound disorder detection

A. Future Directions

Several avenues for improvement and extension are identified:

- **Dataset Expansion**: Larger, more diverse datasets including multiple Arabic dialects and age groups
- **Multi-position Analysis**: Extension to r-sounds in middle and final word positions
- **Temporal Modeling**: Investigation of sequence-to-sequence models for detailed temporal analysis
- **Transfer Learning**: Adaptation of pre-trained speech models for disorder-specific fine-tuning
- **Explainable AI**: Development of interpretation methods to understand model decision-making
- **Real-time Processing**: Optimization for real-time clinical assessment applications
- **Multi-modal Integration**: Combination with visual speech analysis for enhanced accuracy

The promising results obtained in this study lay the groundwork for developing comprehensive automated speech therapy systems that can support clinicians and improve accessibility to speech disorder assessment and intervention for Arabic-speaking populations.

References

[1] A. Al-Tamimi and G. Kharma, "Arabic speech disorders: A comprehensive review," Journal of Speech, Language, and Hearing Research, vol. 58, no. 3, pp. 666-678, 2015.

[2] M. Shahin, "Phonological disorders in Arabic: Patterns and treatment approaches," International Journal of Speech-Language Pathology, vol. 20, no. 4, pp. 431-442, 2018.

[3] A. Vaswani et al., "Attention is all you need," in Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017, pp. 5998-6008.

[4] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Epps, and B. W. Schuller, "Deep representation learning in speech processing: Challenges, recent advances, and future trends," arXiv preprint arXiv:2001.00378, 2020.

[5] M. Tu, J. Zhang, D. Huang, and X. Zhang, "Speech enhancement based on deep neural networks," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 4, pp. 815-825, 2018.

[6] H. Muckenhirn, M. Magimai-Doss, and S. Marcel, "Towards directly modeling raw speech signal for speaker verification using CNNs," in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2018, pp. 4884-4888.

[7] N. Hammami, M. Bedda, N. Farah, and S. Mansouri, "R/-Letter disorder diagnosis (/r/-LDD): Arabic speech database development for automatic diagnosis of childhood speech disorders (Case study)," in Proceedings of the 2nd International Symposium on Automatic Chinese Speech Recognition, 2015, doi: 10.1109/ISACV.2015.7105542.