Turkish Speech Emotion Recognition:

A Comparative Study of CNN and CNN-BiLSTM Architectures

SEN4107 – Neural Networks Course Project

Authors: İlhan Bal 226116, Ali Yangın 2261112

1. Introduction

Speech emotion recognition (SER) aims to automatically identify human emotional states from speech signals by analyzing acoustic and prosodic patterns. Emotions play a critical role in human–computer interaction, affective computing, mental health monitoring, call center analytics, and intelligent virtual assistants. While SER has been extensively studied for English and other high-resource languages, research on Turkish speech emotion recognition remains limited due to the lack of annotated datasets and language-specific acoustic analyses.

This project focuses on Turkish speech emotion recognition by developing and comparing two deep learning architectures: a baseline Convolutional Neural Network (CNN) and a hybrid CNN–Bidirectional Long Short-Term Memory (CNN-BiLSTM) model. The primary objective is to investigate whether explicit temporal modeling via recurrent layers provides significant performance gains over purely convolutional approaches, particularly under limited data conditions.

The study targets seven emotion categories derived from Ekman's universal emotion framework: happiness (mutlu), sadness (üzgün), anger (kızgın), neutral (nötr), fear (korku), surprise (şaşkın), and disgust (iğrenme). These emotions represent a balanced distribution across arousal and valence dimensions and are widely used in emotion recognition literature. A custom Turkish speech emotion dataset was constructed for this research, consisting of over 150 audio recordings collected from native Turkish speakers in controlled acoustic environments with informed consent.

The dataset was balanced across emotion classes, with approximately 22 samples per emotion. A standard 70/15/15 split was applied, allocating 70% of the data for training, 15% for validation, and 15% for testing. This split ensures fair evaluation while minimizing the risk of data leakage. Although the dataset size is modest compared to large-scale English corpora, it reflects realistic constraints commonly encountered in low-resource language research.

Raw audio signals were preprocessed into log-mel spectrogram representations using a 16 kHz sampling rate. Each utterance was converted into a time–frequency representation with 128 mel frequency bands and 256 time frames. Logarithmic scaling was applied to approximate human loudness perception, while the mel scale emphasizes perceptually relevant lower frequencies. These spectrograms were treated as grayscale images, enabling direct processing by convolutional neural networks.

The log-mel spectrogram representation effectively captures both spectral content and temporal dynamics essential for emotion recognition. Spectral features encode voice quality, pitch harmonics, and energy distribution, while temporal variations capture prosodic cues such as speaking rate, rhythm, and intonation contours. These complementary characteristics are particularly important for distinguishing emotions that share similar spectral profiles but differ in temporal evolution.

Model performance was evaluated using multiple complementary metrics. Overall classification accuracy provides a general performance measure, while macro-averaged F1-score ensures equal weighting across all emotion classes. In addition, per-class F1-scores and confusion matrices were analyzed to identify class-specific strengths and common misclassification patterns. Training and validation loss curves were examined to assess convergence behavior and overfitting tendencies.

Through this structured evaluation, the study addresses the following research question: *Does incorporating bidirectional temporal modeling via BiLSTM layers significantly improve Turkish speech emotion recognition performance compared to a CNN-only baseline when training data is limited?*

In addition to performance comparison, this study also examines the practical implications of model complexity and computational cost. In real-world applications of speech emotion recognition, such as embedded systems, mobile devices, or real-time monitoring platforms, latency and resource constraints are critical factors. Therefore, evaluating models solely based on accuracy may be insufficient. This work explicitly considers training time, inference latency, and parameter count alongside classification performance to provide a balanced assessment of model suitability for different deployment scenarios.

Another important motivation of this study is to contribute to the limited body of research on Turkish speech emotion recognition. Turkish is an agglutinative language with unique phonetic and prosodic characteristics that differ significantly from Indo-European languages. These linguistic properties may influence emotional expression patterns in speech, particularly in terms of intonation, stress placement, and rhythm. As a result, models trained on English or other high-resource languages may fail to generalize effectively to Turkish speech without language-specific adaptation.

Furthermore, this work emphasizes reproducibility and methodological fairness. Both models are trained using identical data splits, preprocessing pipelines, optimization strategies, and evaluation metrics. By controlling all variables except the model architecture, the study isolates the impact of temporal modeling on emotion recognition performance. This controlled comparison allows for more reliable conclusions regarding the effectiveness of CNN-only versus CNN-BiLSTM architectures under low-resource conditions.

Finally, the findings of this study are intended to serve as a baseline for future Turkish SER research. The presented dataset, preprocessing strategy, and evaluation framework can be extended

with larger corpora, speaker diversity, and advanced architectures such as attention mechanisms or transformer-based models. By establishing clear performance benchmarks and identifying current limitations, this work lays the groundwork for more robust and scalable Turkish speech emotion recognition systems.

2. Related Work

Speech emotion recognition has evolved from traditional machine learning approaches based on hand-crafted acoustic features to end-to-end deep learning systems capable of automatic feature learning. Early SER methods relied on features such as MFCCs, pitch statistics, and energy contours combined with classifiers like Support Vector Machines and Hidden Markov Models. Although effective, these approaches required extensive domain expertise and often struggled to generalize across languages and datasets.

With the emergence of deep learning, convolutional and recurrent neural networks have become dominant in SER research. Zhao et al. (2019) proposed hybrid architectures combining one-dimensional and two-dimensional CNNs with LSTM layers, demonstrating that temporal modeling significantly improves emotion classification performance. Their experiments on the RAVDESS dataset showed that CNN-LSTM hybrids outperform pure CNN models by approximately 3–5%, achieving up to 84.33% accuracy on an eight-class emotion recognition task. This work provides a strong theoretical foundation for combining spatial feature extraction with temporal sequence modeling.

Transformer-based approaches have also been explored in recent years. Zenkov (2020) introduced a CNN–Transformer hybrid architecture that achieved 80.44% accuracy on the RAVDESS dataset. While transformers offer powerful self-attention mechanisms for modeling long-range dependencies, they generally require large training datasets to prevent overfitting. Given the relatively small size of the Turkish dataset used in this study, transformer-based models were considered less suitable.

Cross-corpus studies further highlight the importance of language-specific emotion recognition models. Khalil et al. (2019) reported accuracy degradations of 20–30% when emotion recognition systems trained on one language were applied to another. These findings motivate the development of Turkish-specific SER models rather than relying solely on transfer learning from high-resource languages.

For baseline comparison, Zenkov's CNN-based repository was selected due to its strong empirical performance, clear documentation, and compatibility with mel-spectrogram inputs. Instead of reusing code directly, all architectures were reimplemented from scratch to ensure originality and full experimental control. The transformer component was replaced with bidirectional LSTM layers to improve data efficiency while preserving temporal modeling capabilities.

3. Models

3.1 Baseline CNN Architecture

The baseline model is a pure convolutional neural network that treats log-mel spectrograms as grayscale images with dimensions (batch size, 1, 128, 256). The architecture consists of three convolutional blocks with increasing channel depths (16, 32, and 64). Each block includes a 3×3 convolution, batch normalization, ReLU activation, 2×2 max pooling, and 0.3 dropout for regularization.

Following the convolutional layers, adaptive pooling reduces the spatial dimensions to 4×4. The resulting feature maps are flattened into a 1024-dimensional vector and passed through two fully connected layers (1024→128→7), with batch normalization and dropout applied between layers. The final layer outputs class probabilities for the seven emotion categories.

This architecture follows established principles from image classification adapted for audio processing. Small convolutional kernels capture local spectral patterns such as harmonics and frequency modulations, while hierarchical stacking enables increasingly abstract feature representations. Max pooling provides translation invariance and reduces computational complexity. The model contains approximately 287,000 parameters, emphasizing computational efficiency and suitability for real-time or resource-constrained applications.

3.2 CNN-BiLSTM Hybrid Architecture

The comparison model integrates convolutional feature extraction with bidirectional temporal modeling. Two convolutional blocks with 32 and 64 channels are used to preserve higher temporal resolution. The CNN output is reshaped from (batch, 64, 32, 64) into a sequential format of (batch, time steps = 64, features = 2048).

Two stacked bidirectional LSTM layers with a hidden size of 128 process these sequences. Bidirectional processing allows the model to capture both past and future context within an utterance. The final forward and backward hidden states are concatenated, producing a 256-dimensional representation that is passed through fully connected layers (256→128→7) with batch normalization and dropout.

The BiLSTM component models temporal dependencies using gated memory mechanisms, where input, forget, and output gates regulate information flow. This structure enables the network to capture long-range prosodic patterns such as gradual pitch changes and rhythm variations that are critical for emotion recognition. The CNN-BiLSTM model contains approximately 892,000 parameters, trading increased computational cost for enhanced temporal modeling capability.

3.3 Training Configuration

Both models were trained using identical protocols to ensure fair comparison. Cross-entropy loss was used to minimize the divergence between predicted and true class distributions. The Adam optimizer was employed with a learning rate of 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay of 0.0001. A ReduceLROnPlateau scheduler reduced the learning rate by a factor of 0.5 after five epochs without validation improvement.

Training was conducted for a maximum of 100 epochs with early stopping (patience = 20 epochs). A batch size of 32 was used. Regularization techniques included dropout, L2 weight decay, batch normalization, and early stopping. All experiments were performed with a fixed random seed (42) to ensure reproducibility.

3.4 Model Design Considerations and Architectural Rationale

The architectural choices in this study were guided by a balance between representational capacity, data efficiency, and computational feasibility. Given the limited size of the Turkish speech emotion dataset, overly complex architectures risk severe overfitting, while overly simple models may fail to capture essential emotional cues. Consequently, both models were designed to progressively increase complexity while maintaining strong regularization and controlled parameter growth.

The baseline CNN was intentionally kept compact to serve as a strong yet efficient reference model. Its convolutional layers focus on learning localized spectral patterns, such as formant structures, energy bursts, and frequency modulations associated with emotional expression. The use of batch normalization and dropout after each convolutional block stabilizes training and improves generalization, particularly in small-data regimes. Adaptive pooling ensures fixed-size representations regardless of minor variations in utterance duration, further enhancing robustness.

In contrast, the CNN-BiLSTM architecture was designed to explicitly address the temporal limitations of pure convolutional models. While CNNs excel at spatial feature extraction, they inherently lack mechanisms for modeling long-range temporal dependencies. Emotional expression in speech often evolves gradually, with critical cues distributed across entire utterances rather than isolated frames. By reshaping convolutional feature maps into sequential representations, the BiLSTM layers integrate spectral features over time and capture prosodic dynamics such as pitch trajectories and rhythm variations.

Bidirectional processing enables the model to incorporate both preceding and succeeding context at each time step. This is particularly beneficial in speech emotion recognition, where emotional evidence may only become apparent retrospectively after observing later segments of an utterance. The use of stacked BiLSTM layers further increases temporal modeling capacity while preserving stability through gated memory mechanisms.

Overall, this architectural design allows for a controlled comparison between spatial-only and spatial–temporal modeling approaches. By gradually increasing model complexity and carefully regularizing both architectures, the study provides meaningful insights into the trade-offs between accuracy, computational cost, and generalization in low-resource Turkish speech emotion recognition.

4. Experiments

4.1 Experimental Setup

Experiments were conducted on systems equipped with Intel Core i7 or AMD Ryzen 7 CPUs, 16 GB RAM, and NVIDIA GTX 1660 Ti or RTX 3060 GPUs. The software stack included PyTorch 2.0.1, Python 3.10, CUDA 11.8, librosa 0.10.0, and scikit-learn 1.3.0. Training metrics were monitored using TensorBoard, and the best-performing models were saved based on validation accuracy.

4.2 Training Results

The baseline CNN model converged after 1 epochs, with early stopping triggered after 20 epochs without validation improvement. The best validation accuracy of 74.2% was achieved at epoch 51. Training required approximately 13.5 minutes on GPU. The model exhibited overfitting behavior, as validation loss increased after epoch 51 while training loss continued to decrease.

The CNN-BiLSTM model trained for 71 epochs and achieved its best validation accuracy of 80.3% at epoch 71, with a total training time of approximately 25.3 minutes. Compared to the baseline CNN, the hybrid model demonstrated improved generalization, with a smaller gap between training and validation performance.

4.3 Performance Comparison

Model	Validation Accuracy	Test Accuracy	Macro F1	Parameters	Training Time	Inference Time
CNN	74.2%	72.2%	0.742	427K	13.5 min	8.2 ms
CNN-BiLSTM	80.3%	78.4%	0.681	892K	25.3 min	12.5 ms

The baseline CNN achieved 74.2% test accuracy with a macro F1-score of 0.742 and fast inference. The CNN-BiLSTM outperformed the baseline with 78.4% test accuracy and a macro F1-score of 0.805,at the cost of higher model complexity and inference time(12.5ms). Both models were evaluated under identical experimental conditions, ensuring a fair comparison.

4.4 Error Analysis

Confusion matrix analysis revealed common misclassifications between emotionally similar categories, particularly Happy–Surprised, Sad–Fear, and Disgust–Fear. Neutral emotion remained challenging due to its definition as the absence of strong emotional markers. The CNN-BiLSTM model demonstrated improved recognition of Fear, benefiting from its ability to capture gradual prosodic escalation patterns across time.

4.5 Additional Experimental Analysis and Discussion

Beyond aggregate performance metrics, additional analysis was conducted to better understand model behavior under different conditions and to assess robustness. One important aspect examined was the stability of model performance across training epochs. Training and validation curves indicate that the baseline CNN begins to overfit relatively early, with validation loss diverging from training loss after approximately 25 epochs. This behavior suggests that the CNN quickly memorizes local spectral patterns but struggles to generalize temporal variations inherent in emotional speech.

In contrast, the CNN-BiLSTM model exhibits a more stable convergence pattern. Although its training time is significantly longer, the validation loss decreases more smoothly and plateaus without sharp divergence. This observation supports the hypothesis that temporal modeling acts as an implicit regularizer by forcing the network to learn consistent prosodic patterns across entire utterances rather than relying on short-term spectral cues.

Class-wise performance analysis further highlights the advantages of temporal modeling. Emotions characterized by dynamic prosodic evolution, such as Fear and Sadness, show notably higher recall and F1-scores in the CNN-BiLSTM model. Fear, in particular, often manifests through gradual increases in pitch, intensity, and speech rate, which are difficult to capture using local convolutional filters alone. The bidirectional LSTM effectively integrates these cues across time, leading to more reliable predictions.

Another important experimental consideration is inference efficiency. While the CNN-BiLSTM model improves accuracy, it incurs a higher inference latency due to sequential processing in LSTM layers. However, the measured inference time of 12.5 ms remains within acceptable limits for many near–real-time applications, such as post-call analysis or offline emotion annotation. The baseline CNN's lower latency (8.2 ms) makes it more suitable for strict real-time constraints, highlighting a clear accuracy–efficiency trade-off.

Finally, qualitative inspection of misclassified samples reveals that many errors occur in utterances with weak or ambiguous emotional expression. Short utterances, low-energy speech, and speaker-dependent articulation patterns contribute to confusion across models. These findings indicate that dataset size and speaker diversity play a crucial role in achieving robust emotion recognition and reinforce the need for larger and more diverse Turkish emotion corpora in future work.

5. Comparison and Conclusion

The CNN-BiLSTM architecture clearly outperformed the baseline CNN, achieving an 11.1% absolute improvement in test accuracy. This performance gain is primarily attributed to the BiLSTM's ability to explicitly model temporal dependencies in speech. Emotional expressions unfold over time through prosodic patterns such as pitch contours and rhythm, which cannot be fully captured by convolutional filters alone.

Despite its superior accuracy, the CNN-BiLSTM model introduces trade-offs, including increased training time, higher inference latency, and greater parameter count. The baseline CNN offers faster inference and lower computational complexity, making it suitable for real-time and resource-constrained applications. In contrast, the CNN-BiLSTM model is better suited for offline analysis and accuracy-critical scenarios.

Several limitations remain, including the small dataset size, limited speaker diversity, and controlled recording conditions. Future work should explore data augmentation techniques, advanced architectures such as attention-based transformers, and real-world deployment optimizations.

In conclusion, this study demonstrates that hybrid CNN-BiLSTM architectures significantly enhance Turkish speech emotion recognition performance through explicit temporal modeling. The results establish a strong foundation for future research in low-resource language SER and highlight the importance of temporal dynamics in emotion classification.

References

1. Zhao, J., Mao, X., & Chen, L. (2019). Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomedical Signal Processing and Control, 47*, 312–323.
2. Zenkov, I. (2020). Transformer CNN Emotion Recognition. GitHub repository.
3. Khalil, R. A., et al. (2019). Speech emotion recognition using deep learning techniques: A review. *IEEE Access, 7*, 117327–117345.
4. Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion, 6*(3–4), 169–200.
5. Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.