COMPARATIVE ANALYSIS OF END-TO-END NEURAL SPEECH RECOGNITION FRAMEWORKS

Hasan Gyulyustan, Pavel Kyurkchiev

Abstract. *This paper presents a comprehensive comparative analysis of modern end-to-end automatic speech recognition (ASR) frameworks, emphasizing architectural diversity, performance trade-offs, and real-world deployment considerations. The study examines representative toolkits such as WeNet, Wav2-Letter++, ESPnet, Fast Conformer, and PyTorch-Kaldi across multiple evaluation metrics, including word and character error rates, training and inference efficiency, and hardware requirements. Beyond controlled benchmark results, the paper investigates robustness under noisy and demographically diverse conditions, highlighting persistent biases in commercial ASR systems. Furthermore, it explores the integration of neural language models and multi-task learning architectures, which enhance alignment, adaptability, and contextual understanding. Emerging domains such as cross-lingual ASR, speech emotion recognition, and silent speech interfaces are analyzed to reveal limitations of traditional evaluation frameworks when applied to multi-domain and non-auditory data. The findings underscore the necessity of holistic performance metrics that incorporate fairness, robustness, and resource efficiency for next-generation ASR system development.*

Key words: End-to-End Speech Recognition; Neural Architectures; WeNet; ESPnet; Fast Conformer; Wav2Letter++; PyTorch-Kaldi; Automatic Speech Recognition (ASR); Multi-Task Learning; Language Model Integration; Noise Robustness; Demographic Bias; Cross-Domain Generalization; Speech Emotion Recognition; Silent Speech Recognition.

Hasan Gyulyustan[1], Pavel Kyurkchiev[1]
[1] Paisii Hilendarski University of Plovdiv,
Faculty of Mathematics and Informatics,
236 Bulgaria Blvd., 4003 Plovdiv, Bulgaria
Corresponding author: `hasan@uni-plovdiv.bg`