A brief literature review on the use of "machine learning for speech signal classification and recognition."

Author: Shazia Parween **shaziaparween900@gmail.com**

Introduction:

Speech recognition, a domain dedicated to the study of analyzing, capturing and manipulating speech information that essentially exists as an acoustic form of energy (Natarajan et al., 2025) and generating application-based outputs such as speech detection, text-to-speech conversion and speech synthesis, is one of key research areas that was revolutionized with the emergence of machine learning, notably deep learning techniques. The trends in research paradigm are clearly evident with the emergence of E2E (End-to-end) models substituting speech recognition traditional models (Natarajan et al., 2025). ASR models in the last decade of Automatic Speech Recognition (ASR) research reflect the apparent shift from the classical Hidden Markov Model (HMM) based ASR architectures to the integration of deep learning techniques like, Deep Neural Networks (DNNs), Convolutional Neural Networks (CNNs), and Recurrent Neural Networks (RNNs) into their architectures. (Hajin, 2024)

Overview of the classical speech recognition systems:

Traditional speech recognition models utilize basic nonparametric models for extracting features from speech signals. These models rely heavily on meaningful feature extractions from raw audio input signals through mathematical operations like Fourier Transform, wavelets transforms and linear predictive coding for enabling conventional regression or classification-based models to predict accurately (Mehrish et al., 2023). The classical approach essentially decomposes an ASR system into components: acoustic feature extraction from speech or audio signals (several mathematical functioning and computing is required in this phase as mentioned earlier), acoustic modeling, language modeling and search which is based on Bayes decision. The classical acoustic modeling is based on hidden Markov models (HMM) which are used to model the probability distribution of speech signals (Mehrish et al., 2023) while language modeling relies on count-based approaches. GMMs are another generative model used to model the acoustic properties of speech signals. For classification and recognition of speech patterns SVMs are used.

End to end architectures:

Prabhavalkar et al. (2023) document the shifts in the speech recognition research with the introduction of end-to-end (E2E) learning and categorises E2E systems into three groups:

Connectionist Temporal Classification (CTC), Recurrent Neural Network Transducer (RNN-T) and attention-based encoder–decoder models. Deep learning was introduced into acoustic modeling, replacing its predecessor Gaussian mixture distributions (hybrid HMM) as well as language modeling substituting the other traditional approaches. According to Prabhavalkar et al. (2023) these were "*ASR model that enables joint training and recognition consistently minimizing expected word error rate, avoiding separately obtained knowledge sources*." thus solidifying the understanding that end to end architecture of the ASR systems are seen as the unification of all sub-tasks into a single differentiable objective that are highly integrated and completely neural, simplifying engineering and enabling joint optimisation at a large scale.

Deep neural networks for speech enhancement and recognition:

A neural network consists of a set of connected *neurons*: the building block of artificial knowledge, with Deep neural networks containing advanced multi-layer hidden networks made of such neurons (Jawad, 2023), that are capable of handling extremely large datasets and allowing for complex signal processing with minimal to low human effort. DNN architectures involve networks like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), that could capture complicated speech patterns with improved speed and accuracy, excelling in precision even with the presence of noise in speech signals which would have taken explicit feature extraction in the classical approach allowing for improved speech recognition with multi-task learning frameworks (Natarajan et al., 2025). This revolution of speech recognition gave rise to the integration of DNN into E2E ASR models with intricate focus on a "joint modeling approach," replacing traditional models with neural networks in the ASR architecture. (Prabhavalkar et al., 2023)

Self-supervised representation learning:

With the emergence of deep learning-based ASR models one of the more fundamental challenges it the field of speech recognition emerged, the scarcity of labelled data. Deep ASR models still depend on thousands of hours of labeled speech, but self-supervised learning (SSL) addresses this data scarcity issue in speech processing by introducing pre-trained self-supervised models, which when fine-tuned on downstream tasks, often outperform supervised baselines while requiring significantly less labeled data. Kheddar et al. (2024) survey this advancement in deep learning approaches for automatic speech recognition (ASR), with particular emphasis on transformer-based architectures and their variants, highlighting the that self-attention mechanisms in transformers enabling the modeling of long-range dependencies in speech sequence compared to traditional RNN-based models. Thus, although the advanced E2E ASR technologies failed to make use of raw unlabeled audio and text data, other prominent SSL models such as wav2vec 2.0, HuBERT, and WavLM, came out as relevant by being able to leverage unlabeled speech data to learn universal speech representations, democratizing access to high-performing models in low-resource languages (Mohamed et al., 2022). The Conformer architectures, which combine convolutional layers with self-attention to capture

both local and global speech patterns, have achieved state-of-the-art results on benchmark datasets. (Kheddar et al.,2024)

Multimodal (audio-visual) speech recognition:

The speech signals involve not just acoustic but also visual information, the two sources compliment the information obtained by each other and thus with the advancements in the integration of these visual information with acoustic signals, human perception through these models can be improved on a significant basis (Ivanko, Ryumin, & Karpov, 2023). Audio-visual speech recognition (AVSR) techniques essentially integrate lip-reading with acoustic signals, and although this area demands significant improvements as of yet it is still progressively enhancing the human speech recognition that still only relied on audio samples. Early-fusion architectures in the ASVR models concatenate visual and acoustic features prior to encoding, while late-fusion models merge modality-specific logits. Attention-based fusion dynamically re-weights modalities, delivering the best trade-off: a 35% relative WER reduction in −5 dB SNR compared with audio-only baselines (Ivanko, Ryumin, & Karpov, 2023).

Challenges and future research:

This literature review points to several emerging trends and improvements in the speed recognition research with the integration of machine learning and its advancements but several key challenges and drawbacks of this approach also needs to be addressed. While movement toward more efficient architectures that maintain high accuracy while reducing computational requirements it also poses challenges for global speech technology adoption.

SSL alleviates labelling costs but does not solve bias towards major languages indicating that the improvement of fairness metrics is important. While compression helps, state-of-the-art models still demand billions of operations, and the research area still demands improvement in the training models in order to make it energy efficient as well (Kheddar et al., 2024). There are privacy concerns that many have admitted to, and these need to be addressed as well when dealing with advanced models that contain large amount of private user data.

Conclusions:

The integration of machine learning into speech signal classification and recognition has led to significant advancements without doubt. It has transformed classical ASR and led to the expansions of models that encompass advanced deep learning architectures, and self-supervised learning models built around deep-learning core, progressing from CNN/RNN hybrids to Transformer-based end-to-end multimodal systems, in the domain of speech recognition. The scalability of ML models has made it possible to integrate even the visual data with the acoustic data signals, introducing another layer into the depths of speech recognition,

the ASVR systems. While challenges such as data scarcity and computational demands still exist the ongoing research and innovations shine a hopeful light into the bright direction of the future that hopefully contains more sophisticated, accessible, and efficient speech technologies.

References:

Hajin, S. (2024). Voice recognition based on machine learning classification algorithms: A review. *Indonesian Journal of Computer Science, 13*. https://www.researchgate.net/publication/382769642_Voice_Recognition_Based_on_Machine_Learning_Classification_Algorithms_A_Review

Ivanko, D., Ryumin, D., & Karpov, A. (2023). A review of recent advances on deep learning methods for audio-visual speech recognition. *Mathematics, 11*(12), 2665. https://doi.org/10.3390/math11122665

Jawad, E. (2023). The deep neural network – A review. *IJRDO - Journal of Mathematics, 9*, 1–5. https://www.researchgate.net/publication/374151186_THE_DEEP_NEURAL_NETWORK-A_REVIEW

Kheddar, H., Hemis, M., & Himeur, Y. (2024). Automatic speech recognition using advanced deep learning approaches: A survey. *Information Fusion*, 102422. https://arxiv.org/abs/2403.01255

Mehrish, A., Majumder, N., Bharadwaj, R., Mihalcea, R., & Poria, S. (2023). A review of deep learning techniques for speech processing. *Information Fusion, 99*, 101869. https://arxiv.org/abs/2305.00359

Mohamed, A., Lee, H. Y., Borgholt, L., Havtorn, J. D., Edin, J., Igel, C., ... & Watanabe, S. (2022). Self-supervised speech representation learning: A review. *IEEE Journal of Selected Topics in Signal Processing, 16*(6), 1179–1210. https://arxiv.org/abs/2205.10643

Natarajan, S., Al-Haddad, S. A. R., Ahmad, F. A., Kamil, R., Hassan, M. K., Azrad, S., ... & Dauitbayeva, A. (2025). Deep neural networks for speech enhancement and speech recognition: A systematic review. *Ain Shams Engineering Journal, 16*(7), 103405. https://doi.org/10.1016/j.asej.2025.103405

Prabhavalkar, R., Hori, T., Sainath, T. N., Schlüter, R., & Watanabe, S. (2023). End-to-end speech recognition: A survey. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 32*, 325–351. https://arxiv.org/abs/2303.03329