Systematic Review of Speech Signal Processing: Advances, Challenges, and Future Directions

Anas Fouad Ahmed
Al-Iraqia University,
Electrical Engineering Dept
Baghdad 6029, Iraq
anas.ahmed@aliraqia.edu.iq

Ikhlas M. Farhan
University of Technology
College of Electrical Engineering
Baghdad 19006, Iraq
ikhlas.m.farhan@uotechnology.edu.iq

Zaid Abdulkareem Hussein
TheSunni Endowment
Baghdad 6029, Iraq
lhashimi200@gmail.com

Abstract. Speech signal Processing has led to significant progress in recent years, driven by deep learning, self-perpetuated learning, and end-to-end models. This systematic review examines traditional and modern techniques for speech recognition, growth, and synthesis, and evaluates benefits, boundaries, and potential future research directions. We compare the most important algorithms, evaluate their performance in the dataset, and discuss the strengthening of new trends such as self-employed learning and adversity. This article presents a broad comparative analysis using wide tables and ensures clarity for researchers in the field.

Keywords: Speech signal processing, automatic speech recognition, feature extraction, deep learning, speech enhancement, speaker verification.

1.Introduction

The speech signal Processing is fundamental to the processing of artificial intelligence (AI), which affects applications such as speech recognition, speaker identity, and noise shortage. Over the past two decades, research has been transferred from hidden Markov models (HMM) and Gaussian-Mixture Models [1], [2] to Deep Neural Networks (DNN) and Transformers [3], [4]. This trend has improved striking accuracy, strength and efficiency.

It offers a systematic review of paper speech improvement, automatic voice recognition (ASR) and speech synthesis, which highlights benchmark models, datasets and comparison of performance. The contribution from this review includes:

A historical perspective on speech signal Processing.

Detailed comparison of speech Processing techniques (HMM-GMM, CNN, LSTM, transformer, self-preserved learning).

Intensive discussion about challenges and future research directions.

2. Speech Enhancement Techniques

2.1. Traditional Speech Enhancement

Traditional speech growth techniques include spectral subtraction, veneer filtration and MMSE estimates [5], [6]. The purpose of these methods is to reduce the noise in the background by preserving speech information.

Table 1. Traditional Speech techniques

Model	Architecture	limitation	References
Spectral Subtraction	Simple, low computational cost	Distortion in residual noise	[5]
Wiener Filter	Adaptive filtering	Requires accurate noise estimation	[6]
MMSE Estimator	Minimizes speech distortion	Computationally expensive	[7]

Recent progress includes DNN, Gans and self-preserved learning methods such as intensive learning-based denoising [8]-[10].

.

3. Wavelet Packet Transform (WPT) for Feature Extraction

3.1. Hidden Markov Models & Gaussian Mixture Models (HMM-GMM)

The HMM-GMM framework was prominent in ASR before deep learning. The HMMS model is a stochastic sequence, whereas GMM's approach the phonemes [11], [12]. Despite the success, these models are struggling with reference-dependent learning and large vocabulary tasks [13].

3.2. Deep Learning-Based ASR

Contemporary ASR systems are based on convolutional neural networks (CNNs). recurrent neural networks (RNNs), and transformers [14] [16]. These models have greater recognition accuracy and less feature. engineering.

Table 2: Deep Learning approach

Model	Architecture	Advantages	limitation	References
CNNs	Convolutional layers for feature extraction	Effective for short-term features	Limited long-range dependencies	[14]
RNNs	Sequential processing of speech signals	Good for temporal dependencies	Vanishing gradient problem	[15]
Transformers	Decomposes the EEG into frequency sub-bands	Handles long-range dependencies	High computational cost	[16]

4. Speech Synthesis

4.1. Statistical Parametric and Concatenative Synthesis

Extremist synthesis techniques comprise formant synthesis, concatenative. synthesis, hidden Markov model-based synthesis (HTS) [17]-[19]. These models have artificial prosody and need large databases of phonetic rules.

4.2. Deep Learning-Based Speech Synthesis

Contemporary methods, like WaveNet, Tacotron and FastSpeech have. naturalness and prosody [20]-[22] were considerably enhanced.

Table 3: Comparison of Deep Learning Preprocessing Methods

Model	Key feature	Strength	Weakness	References
WaveNet	Autoregressive model	High-quality synthesis	Slow inference time	[20]
Tacotron	Sequence-to-sequence model	Natural prosody	Requires large datasets	[21]
FastSpeech	Non-autoregressive	Fast inference	Reduced prosody variation	[22]

5. Challenges and Future Directions

5.1. Robustness to Noise and Adversarial Attacks

The performance of Automatic Speech Recognition (ASR) systems can be seriously affected by adversarial examples that are able to manipulate the input waveforms in a way that is not detectable to the human ear [23], [24]. Adversarial defense research is a current subject, and methods including spectral domain filtering and self-supervised adversarial training [25]–[27].

5.2. Data Efficiency & Low-Resource ASR

The wav2vec 2.0 and HuBERT are self-supervised learning methods that lessen the reliance on labelled datasets and enhance ASR on the low-resource languages [28]–[30].

5.3. Real-Time and Edge Deployment

Models built for speech recognition are very heavy computationally and that makes it difficult to deploy on the edge. Pruning, quantization and knowledge distillation are needed to make real-time applications [31]–[33].

5.4. Explain ability & Interpretability

For bias reduction and fairness, it is essential to understand how deep models process speech. The new approaches are layer-wise relevance propagation, and attention-based visualization [34]–[36].

6. Conclusion

The paper is a review of the speech signal processing including speech enhancement, speech recognition, and speech synthesis. We have touched on conventional and contemporary methods, contrasted key models, and pointed out future research problems. Advancements in self-supervised learning, adversarial robustness, and real-time processing are crucial for the future.

References:

[1] Gao, M., Wu, S., Chen, H., Du, J., Lee, C.H., Watanabe, S., Chen, J., Marco, S.S. and Scharenborg, O., "The multimodal information-based speech processing (MISP) 2025 challenge: Audio-visual diarization and recognition". arXiv preprint arXiv:2505.13971. 2025

[2] Grigoryan, L., Karpov, N., Albasiri, E., Lavrukhin, V. and Ginsburg, B., "Open Automatic Speech Recognition Models for Classical and Modern Standard Arabic." In ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1-5. IEEE, 2025.

[3] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modelling in speech recognition: The shared views of four research groups," IEEE Signal Process. Mag., vol. 29, no. 6, pp. 82–97, Nov. 2012.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Proc. NeurIPS, 2017.

[5] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoust., Speech, Signal Process., vol. 27, no. 2, pp. 113–120, 1979.

[6] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," Proc. IEEE, vol. 67, no. 12, pp. 1586–1604, 1979.

[7] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," IEEE Trans. Acoust., Speech, Signal Process., vol. 32, no. 6, pp. 1109–1121, 1984.

[8] D. Wang and J. S. Lim, "The unimportance of phase in speech enhancement," IEEE Trans. Acoust., Speech, Signal Process., vol. 30, no. 4, pp. 679–681, 1982.

[9] X. Xu, R. H. M. Chan, and K. Yao, "A deep learning approach for speech enhancement using a complex spectrogram," IEEE/ACM Trans. Audio, Speech, Lang. Process, vol. 26, no. 11, pp. 2136–2145, 2018.

[10] A. Pandey and D. Wang, "Densely connected neural network with dilated convolutions for real-time speech enhancement," in Proc. IEEE ICASSP, 2019.

[11] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proc. IEEE, vol. 77, no. 2, pp. 257–286, 1989.

[12] S. J. Young, "The HTK hidden Markov model toolkit: Design and philosophy," Tech. Rep. Cambridge Univ., 1993.

[13] T. Hain, "Implicit pronunciation modelling in ASR," in Proc. IEEE ICASSP, 2002.

[14] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in Proc. ICML, 2006.

[15] T. N. Sainath and B. Kingsbury, "Deep convolutional neural networks for LVCSR," in Proc. IEEE ICASSP, 2013.

[16] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in Proc. Interspeech, 2020.

[17] P. Taylor, "Text-to-speech synthesis," Cambridge Univ. Press, 2009.

[18] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," Speech Commun., vol. 51, no. 11, pp. 1039–1064, 2009.

[19] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden Markov models," Proc. IEEE, vol. 101, no. 5, pp. 1234–1252, 2013.

[20] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," in Proc. Interspeech, 2016.

[21] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. V. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in Proc. Interspeech, 2017.

[22] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, and T.-Y. Liu, "FastSpeech: Fast, robust and controllable text to speech," in Proc. NeurIPS, 2019.

[23] N. Carlini and D. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," in Proc. IEEE SPW, 2018.

[24] H. Qin, P. Mittal, and R. B. Lee, "Imperceptible and robust adversarial examples for automatic speech recognition," in Proc. Interspeech, 2019.

[25] A. Esmaeilpour, X. Liu, and A. M. Eskandarian, "Detection and mitigation of adversarial attacks in ASR systems," in Proc. IEEE ICASSP, 2020.

[26] Z. Chen, S. K. Sharma, P. K. Varshney, and S. K. Das, "Adversarial robustness of deep speech recognition: A survey," IEEE Commun. Surv. Tutorials, vol. 23, no. 4, pp. 2179–2203, 2021.

[27] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in Proc. NeurIPS, 2020.

[28] K. Hsu, A. Baevski, and M. Auli, "Hubert: Self-supervised speech representation learning by masked prediction," in Proc. IEEE ICASSP, 2021.

[29] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in Proc. IEEE ICASSP, 2013.

[30] J. Villalba and N. Dehak, "X-vector: Robust deep neural network embeddings for speaker recognition," in Proc. IEEE ICASSP, 2018.

[31] J. Villalba and N. Dehak, "X-vector: Robust deep neural network embeddings for speaker recognition," in Proc. IEEE ICASSP, 2018.

[32] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with SincNet," in Proc. IEEE SLT, 2018.

[33] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation," IEEE/ACM Trans. Audio, Speech, Lang. Process, vol. 27, no. 8, pp. 1256–1266, 2019.

[34] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in Proc. Interspeech, 2020.

[35] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, "Listen, attend and spell," in Proc. IEEE ICASSP, 2016.

[36] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in Proc. EMNLP, 2014.

[37] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997.

[38] T. Mikolov, M. Karafiát, L. Burget, J. Černocký, and S. Khudanpur, "Recurrent neural network based language model," in Proc. Interspeech, 2010.

[39] A. Graves, "Sequence transduction with recurrent neural networks," in Proc. ICML Workshop on Representation Learning, 2012.

[40] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in Proc. IEEE ICASSP, 2013.

[41] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," IEEE Trans. Signal Process., vol. 45, no. 11, pp. 2673–2681, 1997.

[42] R. J. Williams and D. Zipser, "A learning algorithm for continually running fully recurrent neural networks," Neural Comput., vol. 1, no. 2, pp. 270–280, 1989.

[43] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in Proc. ICLR, 2015.

[44] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019.

[45] Y. Zhang, W. Chan, and N. Jaitly, "Very deep convolutional networks for end-to-end speech recognition," in Proc. IEEE ICASSP, 2017.

[46] P. J. Liu, M. Saleh, E. Pot, B. Goodrich, R. Sepassi, L. Kaiser, and N. Shazeer, "Generating Wikipedia by summarizing long sequences," in Proc. ICLR, 2018.

[47] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in Proc. NeurIPS, 2015.

[48] T. N. Sainath, R. J. Weiss, A. Senior, K. Rao, and C. J. McLean, "Learning the speech front-end with raw waveform CLDNNs," in Proc. Interspeech, 2015.

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE CVPR, 2016.
[50] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in Proc. NeurIPS, 2012.