Literature Review: Speech Enhancement for ASR

Duff Bastasa, Mohammad Jameel Jibreel Mamogkat, et al.

October 11, 2025

1 The Core Problem: Noise Degradation in ASR

The central challenge this project addresses is the degradation of Automatic Speech Recognition (ASR) performance in noisy environments. Unwanted noise can "alter the main characteristic features of voice signals," which corrupts the quality of the speech and the information it contains [1]. This mismatch between the noisy, real-world audio and the typically cleaner data ASR models are trained on is a primary cause of recognition errors.

2 The Solution: Speech Enhancement as an ASR Front-End

A common and effective strategy is to use a DSP-based speech enhancement (SE) module as a front-end pre-processor to clean the audio before it reaches the ASR engine.

- **Validation:** Modern research confirms this is a viable approach. Papers by Pandey et al. [2] and Kinoshita et al. [3] show that a high-quality SE front-end can lead to significant, measurable improvements in ASR accuracy. In fact, Kinoshita et al. reported a relative word error reduction of over 30% on a challenging dataset [3].

3 The Central Challenge: Optimizing for ASR, Not Human Listening

A crucial finding in the literature is that the goal of speech enhancement for ASR is different from enhancement for human perception.

- **Guiding Principle:** The paper by Kawase et al. [4] explicitly states that traditional enhancement techniques designed to maximize metrics for human listening (like Signal-to-Distortion Ratio) "cannot always maximize ASR accuracy."

- **The Trade-Off:** An aggressive cleaning algorithm might remove background noise, but it could also introduce subtle "processing artifacts" that distort the very speech features the ASR model relies on [2, 3]. Pandey et al. also note that these artifacts can make the enhanced speech "suboptimal" for ASR modeling [2].
- **Our Goal:** Therefore, the success of our project's DSP module must be measured by the final ASR accuracy (e.g., Word Error Rate), not by how "clean" the audio sounds subjectively.

4 Analysis of Potential DSP Techniques

The literature discusses several single-channel noise reduction techniques. Below is a comparative analysis to inform our choice of implementation.

4.1 Spectral Subtraction

- **Principle:** This is a foundational DSP method that operates by estimating the noise spectrum from a silent portion of the audio and then subtracting this estimate from the entire signal's spectrum [1, 5].
- **Pros:** It is a simple and computationally efficient algorithm, making it straightforward to implement. Studies have shown it can produce "promising results" for improving ASR performance, particularly in very noisy (low SNR) conditions [1].
- **Cons:** Its primary drawback is the tendency to introduce an artifact known as "musical noise" [5, 6]. The algorithm is also based on a mathematically convenient but false assumption that "cross-terms" between the speech and noise signals are zero, which is a source of error [6].

4.2 Wiener Filtering

- **Principle:** This is a more advanced statistical technique. Instead of direct subtraction, it constructs an optimal filter that aims to minimize the mean square error between the estimated clean signal and the true (unknown) clean signal.
- **Pros:** It is generally considered more robust and produces fewer artifacts than basic Spectral Subtraction. It has been shown to perform well across a wide range of noise conditions [1].
- **Cons:** It is more complex to implement correctly, as it requires estimating the power spectra of both the noise and the desired clean signal [1, 4].

4.3 Deep Learning (Neural Network) Methods

- **Principle:** This is the state-of-the-art approach where a deep neural network is trained on pairs of noisy and clean audio to learn a complex mapping function for noise reduction [2, 3].
- **Pros:** These methods can significantly outperform traditional DSP techniques in improving ASR accuracy [3].
- **Cons:** The complexity is very high, requiring large datasets and GPU training, effectively a full machine learning project in itself.

References

[1] K. Garg and G. Jain, "A comparative study of noise reduction techniques for automatic speech recognition systems," in *2016 Intl. Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2016, pp. 2098–2103.

[2] A. Pandey, C. Liu, Y. Wang, and Y. Saraf, "Dual application of speech enhancement for automatic speech recognition," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 223–228.

[3] K. Kinoshita, T. Ochiai, M. Delcroix, and T. Nakatani, "Improving noise robust automatic speech recognition with single-channel time-domain enhancement network," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7009–7013.

[4] T. Kawase, M. Okamoto, T. Fukutomi, and Y. Takahashi, "Speech enhancement parameter adjustment to maximize accuracy of automatic speech recognition," *IEEE Transactions on Consumer Electronics*, vol. 66, no. 2, pp. 125–134, May 2020.

[5] M. Gupta, R. K. Singh, and S. Singh, "Analysis of optimized spectral subtraction method for single channel speech enhancement," *Wireless Personal Communications*, Sep. 2022.

[6] T. Yadava G, N. B. G, and J. H. S, "A spatial procedure to spectral subtraction for speech enhancement," *Multimedia Tools and Applications*, vol. 81, p. 23633–23647, Mar. 2022.