# REVERB AND NOISE AS REAL-WORLD EFFECTS IN SPEECH RECOGNITION MODELS

*G. Costantini[1], V. Cesarini[1]*

[1] Department of Electronic Engineering, University of Rome Tor Vergata
Via del Politecnico 1, 00133 - Rome, Italy

The study focused on understanding and addressing the impact of real-world audio distortions, specifically reverberation and background noise, on speaker recognition (SR) models. We recognized that while many SR systems perform well in controlled, clean environments, their performance tends to degrade sharply when exposed to acoustic disturbances that are common in practical applications. Most SR models traditionally rely on mel-frequency cepstral coefficients (MFCCs) for feature extraction, which are excellent at capturing speech characteristics in ideal conditions but highly sensitive to noise and reverberation.

To quantify this issue and explore ways to improve robustness, we conducted a series of experiments based on the DEMoS dataset [1], which contains recordings from 56 speakers of varied age and gender groups. We augmented this clean dataset by adding two types of disturbances: reverberation, simulated through convolution with real room impulse responses, and pink noise, added at several signal-to-noise ratios to cover different degrees of interference.

We evaluated three feature extraction strategies: using Mel-frequency Cepstral Coefficients (MFCC [2]) alone, applying RASTA filtering [3], and combining MFCCs with RASTA features into a hybrid set. RASTA filtering was of particular interest to us because it enhances the temporal dynamics of speech while attenuating slow and constant noise variations obtained through a frequency-domain high-pass filtering, theoretically offering a level of robustness absent in standard MFCCs. Our models were trained exclusively on the clean recordings, allowing us to assess generalization abilities by testing them on the distorted versions without retraining or adaptation. Through our experiments, we observed that MFCCs alone, although effective in pristine conditions with around 93% accuracy, suffered dramatic drops under heavy noise and reverberation, falling to about 62%, equivalent to a performance loss of over 30%. RASTA features alone improved resilience, achieving about 90% in clean conditions and maintaining approximately 68% accuracy under strong distortions, thus showing a clear improvement of around 6% over MFCCs in challenging conditions but a slight drop when the audio was clean.

The hybrid MFCC-RASTA feature set consistently outperformed the individual approaches. It achieved around 92% accuracy in clean conditions, which is almost as good as pure MFCCs, but crucially it preserved about 75% accuracy even under the most difficult combinations of noise and reverberation. This means that the hybrid method reduced the loss from over 30% to about 17%, representing an improvement of nearly 13% compared to MFCCs alone in harsh conditions. Additionally, across a wide range of noise levels, the hybrid model showed a consistent advantage of 8–15% over the other configurations, suggesting it adapts better across varying degrees of disturbance. We also noticed that while RASTA alone could resist noise better than MFCCs, it struggled in cases where fine spectral details were crucial for distinguishing between similar speaker voices, likely due to its temporal smoothing effect.
Based on our results, we concluded that combining MFCCs and RASTA features offers a powerful synergy: MFCCs capture the fine-grained spectral structure necessary for speaker discrimination, while RASTA processing introduces a robustness against environmental factors that degrade traditional feature sets [4]. Therefore, the hybrid approach offers a practical and efficient solution for building SR systems capable of maintaining reliability even when faced with complex acoustic

environments. Our study [5] highlights the importance of designing feature extraction pipelines that reflect real-world conditions rather than relying solely on laboratory-ideal datasets. We believe that further research could explore expanding the hybrid approach by integrating additional noise-robust features or by training models specifically on augmented data, but even without domain adaptation, the feature combination we proposed already leads to a significant step forward in making speaker recognition systems more practical for everyday use.

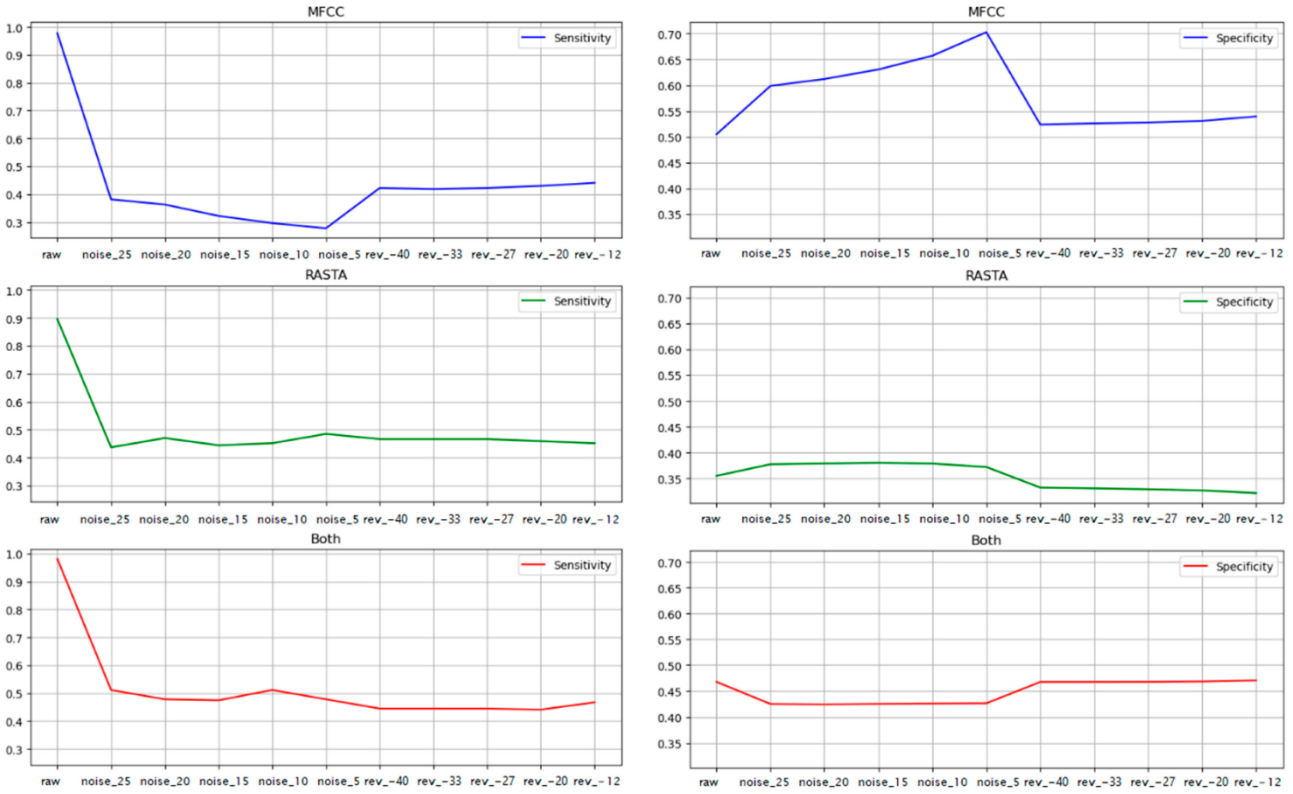


Fig. 1 - Sensitivity (class 1 recall) on the left, Specificity (class 2 recall) on the right. MFCC is worse than RASTA for Sensitivity, but better for Specificity. The classifier using both is balanced in terms of Sensitivity and Specificity thus resulting in better generalization.

## References

[1] E. Parada-Cabaleiro, G. Costantini, A. Batliner, M. Schmitt, and B. Schuller, 'DEMoS: an Italian emotional speech corpus: Elicitation methods, machine learning, and perception', *Language Resources and Evaluation*, vol. 54, Feb. 2019, doi: 10.1007/s10579-019-09450-y.

[2] B. P. Bogert, 'The quefrency alanysis of time series for echoes ; Cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking', *Time Series Analysis*, pp. 209–243, 1963.

[3] H. Hermansky and N. Morgan, 'RASTA processing of speech', *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, Oct. 1994, doi: 10.1109/89.326616.

[4] S. Selva Nidhyananthan, R. Shantha Selva Kumari, and T. Senthur Selvi, 'Noise Robust Speaker Identification Using RASTA–MFCC Feature with Quadrilateral Filter Bank Structure', *Wireless Pers Commun*, vol. 91, no. 3, Dec. 2016, doi: 10.1007/s11277-016-3530-3.

[5] V. Cesarini, G. Costantini, 'Reverb and Noise as Real-World Effects in Speech Recognition Models: A Study and a Proposal of a Feature Set', *Applied Sciences*, MDPI, 2024, *14*(23), 11446; https://doi.org/10.3390/app142311446