Evaluation of Custom Masking-Based Speech Enhancement and Pretrained NVIDIA BNR Model for Slovenian Language Applications

Abstract

This report presents a detailed evaluation of speech enhancement approaches for Slovenian speech. Two configurations were investigated: an offline masking model trained on 100 hours of pre-mixed noisy-clean pairs, and an online mixing pipeline using Lhotse with randomised noise levels between 0 and 20 dB. Room Impulse Response (RIR) augmentation was applied with probability 1.0 during training to simulate reverberant environments. While this improved realism, it also introduced excessive reverberation that reduced intelligibility. In parallel, NVIDIA's Maxine Background Noise Removal (BNR) pretrained model was evaluated and demonstrated superior noise suppression, consistent quality across languages, and real-time performance. Based on these findings, we recommend adopting the NVIDIA BNR model for deployment while continuing targeted experiments to optimise augmentation parameters, particularly RIR probability.

1. Introduction

Speech enhancement (SE) is a crucial pre-processing step for automatic speech recognition (ASR), improving intelligibility and overall system robustness. For Slovenian-language applications, the challenge is to develop or select an SE model that generalises well across diverse noise conditions. Given that SE models are fundamentally language-independent, it is important to determine whether a custom-trained model offers advantages over state-of-the-art pretrained solutions.

2. Methodology

2.1 Offline Mixing Configuration

The offline training configuration used approximately 100 hours of curated, clean Slovenian speech combined with noise from DEMAND and locally recorded sources. Each utterance was pre-mixed at fixed SNR levels of 0, 5, 10, 15, and 20 dB, creating multiple noisy-clean pairs per file. The model was trained using 4-second segments selected with random offsets. This approach provided initial improvements but lacked robustness to unseen noise.

2.2 Online Mixing with Lhotse

To overcome storage and diversity limitations, we implemented an online augmentation pipeline using Lhotse. This pipeline mixes noise dynamically during training with SNR values uniformly sampled between 0 and 20 dB, thereby exposing the model to a wider variety of noise conditions. Room Impulse Response (RIR) augmentation was applied with probability 1.0 to simulate reverberant conditions. Validation and test sets were pre-mixed once with fixed noise conditions to ensure reproducibility.

2.3 Configuration Comparison

Aspect	Offline Config	Online Config
Noise Mixing	Fixed SNR levels: 0, 5, 10, 15, 20 dB	Randomised SNR between 0 and 20 dB
RIR Augmentation	Not applied	Applied with probability = 1.0
Data Storage	High (pre-generated noisy files stored)	Low (on-the-fly mixing, no pre-storage)

3. Experimental Results

The offline configuration produced moderate PESQ and STOI improvements but exhibited limited generalisation to unseen noise. The online configuration improved data diversity and eliminated storage bottlenecks, but RIR augmentation with probability 1.0 introduced excessive reverberation, reducing intelligibility for ASR. NVIDIA BNR achieved high perceptual quality, natural-sounding output, and consistent results across Slovenian, confirming its language independence.

4. Discussion

Our findings indicate that speech enhancement generalises well across languages and does not require language-specific training. The custom-trained model was limited by augmentation settings, particularly the heavy use of RIR, which negatively impacted intelligibility. In contrast, NVIDIA BNR provided robust noise suppression without requiring additional training or parameter tuning, making it a practical choice.

5. Recommendations

1. Adopt NVIDIA Maxine BNR as the primary speech enhancement solution for Slovenian applications.
2. Continue controlled experiments with lower RIR probabilities (e.g., 0.0 or 0.1) to improve model generalisation and speech clarity.
3. Use custom model development as a research pathway, focusing on augmentation tuning and evaluating performance against BNR.

6. Conclusion

This study confirms that a pretrained, language-independent speech enhancement solution such as NVIDIA BNR currently offers the best balance of performance, generality, and deployment readiness. The online augmentation pipeline is a valuable research tool, and future work should explore the effect of reduced RIR probability to further improve intelligibility and ASR compatibility.