Exploring Speech Enhancement Models for Indic Language in Noisy Environments: A Benchmarking Study

Sagar Abhyankar[1]*, Aryan Dande*[1]*, Omkar Bhosale*[1]*, Ayush Gala*[1]*, Dr. Saswati Rabha*[2]*, Chintan Parikh*[2]*, Dr. Balwant Sonkamble*[1]

[1]Pune Institute of Computer Technology, [2]Reverie Language Technologies Pvt. Ltd., India

sagarabhyankar18@gmail.com, aryan.dande@gmail.com, omkarbhosale277@gmail.com. ayushgala2@gmail.com, saswati.rabha@reverieinc.com, chintan.parikh@reverieinc.com, basonkamble@pict.edu

Abstract

Speech enhancement for Indic languages, particularly in noisy environments, lacks extensive bench-marking of models. This study examines the effectiveness of a range of models and methodologies for enhancing speech in Indic languages. It involves exploring state-of-the-art models along with our proposed approach containing diverse input feature types and various phase reconstruction methods. The training is conducted utilizing a newly developed, comprehensive dataset that encompasses audio recordings from different Indic languages, along with a variety of noise samples. We investigate how models trained on English datasets perform differently from those trained on our Indic language dataset, aiming to understand the implications of this variation. Furthermore, we examine which models are most effective in handling different types of noise, providing guidance on selecting the appropriate model for specific noise conditions.

Index Terms: Speech enhancement, Indic languages, Phase reconstruction, Noise

1. Introduction

Speech enhancement, a vital area within the domain of signal processing, plays a pivotal role in various applications such as speech recognition, hearing aids, telecommunication, and audio restoration. The domain of Indic languages presents distinct problems because of their various phonetic structures, rich linguistic legacy, and frequently restricted linguistic resources for study and development, even though great progress has been made in improving speech quality for key languages. [1] Across the Indian subcontinent and the global diaspora, these languages are spoken by over a billion people, representing a diverse range of cultures, dialects, and linguistic subtleties. Nevertheless, Indic languages frequently lag behind their Western equivalents in terms of technological breakthroughs, especially in the area of speech processing and improvement, despite their widespread usage.

Speech enhancement models have generally performed extremely well on English languages, but in recent years, there has been a growing interest among researchers and practitioners to address the challenges of speech enhancement specifically tailored for Indic languages. Tahir et al. [2] released the IndicSUPERB dataset to standardize training and test datasets that are used to evaluate and compare the performance of two different models primarily for Indian languages. The Shrutilipi dataset scraped from All India Radio by AI4Bharat generated access to more than 2000 hours of clean and usable Indic speech data [3]. Other contributions include the subjective testing framework introduced by Chandan et al. [4] and the use of encoding techniques to further improve speech enhancement on Indic languages by Vinay et al. [5].

Speech Enhancement models trained on the English language or similar corpora have seen tremendous success and traction in the past few years. Xu et al. demonstrated the use of Deep Neural Networks for the problem of speech enhancement. [6] This laid the foundation for a variety of other neural network and deep learning approaches to be experimented on the problem. In recent years, Schröter et al. introduced a series of DeepFilterNet models that can enhance speech extremely well not just objectively but also demonstrate real-time noise filtering capabilities [7] [8] [9]. Other State-of-the-art models includes the MFNet model by Liu et al. [10] which won the Microsoft Deep Noise Suppression challenge in 2023 and SpeechBrain [11] which has been developed to function as a general purpose speech toolkit built in Pytorch. However the ability of these same models to perform on languages other than English has not been studied and we attempt to provide a comparative study for the same through our experiments.

In section 2, we describe our proposed method - a revised DNN-GRU model with audio reconstruction using the extracted phase information of the audio samples in the preprocessing phase. The decision to perform the study with DNN-GRU was made due to its comparatively good performance given its low computational complexity. Further in Section 3, we evaluate the performance of the models trained on either the English or Hindi speech data, under different noises and Signal-to-Noise Ratios (SNRs). The WB-PESQ and STOI metrics are calculated and the results of the study have been concluded in section 4.

2. Proposed Method

2.1. Preprocessing

We have used the architecture of DNN-GRU [12] for testing the performance of noise suppression models across different languages. However, we have used a different technique for preprocessing of sound files before they are fed into the model.

Preprocessing of audio samples is a key process for the noise suppression. For the model to learn the patterns for suppressing noise effectively, the audio sample should be passed in its most raw form. This can be possible by training neural networks on feature extraction methods like STFT, STDCT, DCT, .etc. Methods such as Short-Time Fourier Transform (STFT) represent the same audio sample in frequency domain containing complex as well as real valued data. This type of data needs a complex neural network with loss functions calculating differences between complex numbers which is challenging.

In our approach to overcome the challenges presented by complex valued STFT we have used Magnitude Spectrogram, which is derived from STFT by taking the absolute values of all

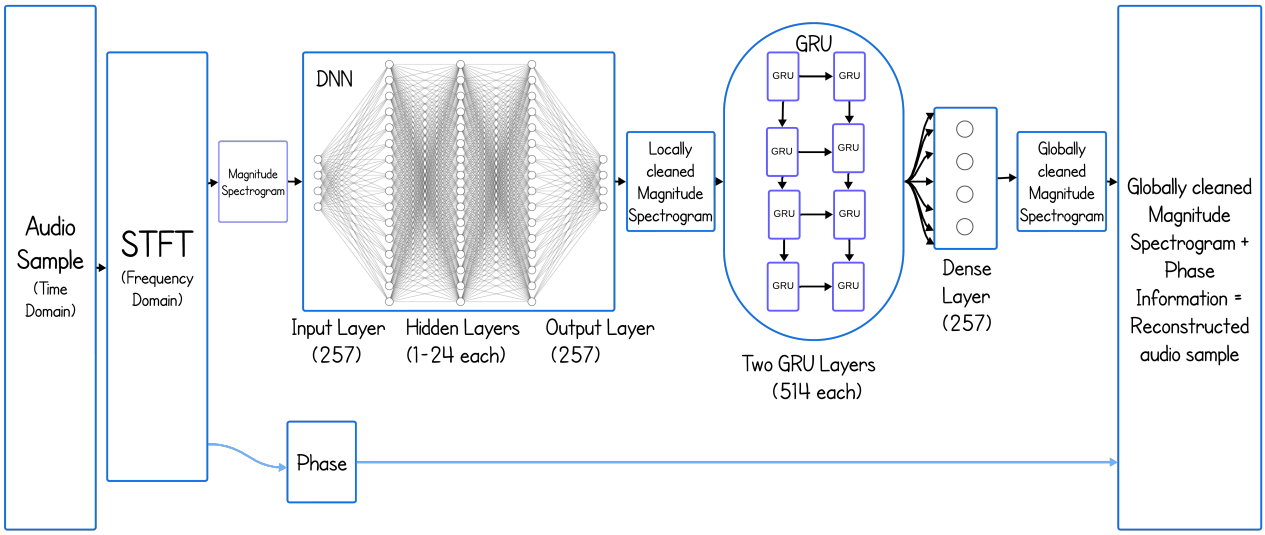

Figure 1: *Proposed DNN-GRU Architecture*

coefficient. Reconstructing the audio using magnitude spectrum does not yield a good result due to the loss of phase information while generating a magnitude spectrogram. To solve this issue we preserve the phase information while extracting magnitude spectrograms by storing phase angles of the complex-valued STFT coefficients.

$$\text{STFT}(x(t), \tau, f) = \int_{-\infty}^{\infty} x(t) w(t-\tau) e^{-j2\pi f t}\, dt \quad (1)$$

$$\text{Magnitude Spectrogram}(t, f) = |\text{STFT}(x(t), \tau, f)| \quad (2)$$

2.2. Model Architecture

The magnitude spectrogram of the audio file is passed through a deep neural network with three hidden layers each of 1024 neurons. The input layer has a shape of 257 units and the output layer is a Dense layer of 257 neurons with a linear activation function. Adam optimizer is chosen with a learning rate of 0.0001. Adam is a popular optimizer that adapts the learning rate during training and combines the advantages of two other extensions of stochastic gradient descent, AdaGrad and RMSProp. The learning rate is set relatively low to ensure stable training. Dropout is applied after the hidden layers with a dropout rate of 25%. Dropout is a regularisation technique that randomly sets a fraction of input units to zero during training, which helps prevent over-fitting by introducing noise and reducing interdependent learning among neurons. This cleans the magnitude spectrogram locally and prepares it for the GRU (Gated Recurrent Unit) layers.

The GRU architecture has 2 layers of 514 GRU units each. It cleans one audio sample at a time globally. Finally the globally cleaned magnitude spectrogram is combined with phase information to reconstruct the audio.

The mean squared error (MSE) is used as the loss function, which minimizes the error between the predicted and noisy speech features.

$$MSE = \sum_{i=1}^{D} (x_i - y_i)^2 \quad (3)$$

2.3. Post-Processing

As magnitude spectrogram results in loss of phase of information, the model recovers the phase for predicted sound by using the input noisy phase.

$$\text{STFT}_{\text{pred}}(t, f) = \text{Magnitude}_{\text{pred}}(t, f) \cdot e^{j \cdot \text{Phase}_{\text{inp}}(t, f)} \quad (4)$$

3. Experiments and Results

3.1. Datasets Training

The training for the proposed model was carried out by first creating a dataset for Hindi speech. The sources used to create this dataset include 60 hours of clean speech data from 'Shrutilipi' dataset developed by AI4Bharat along with an additional 40 hours of 'Tarini' dataset developed in-house from various media sources. The model was also separately trained on 20 hours of English speech audio from LibriSpeech ASR [13] corpus available publicly on OpenSLR. These audio samples were then mixed with different types of noises at a constant -5SNR and not at a variable SNR. Our tests found out that the model performed better when trained on -5SNR noisy audios than when trained on variable SNRs ranging from -5dB to 40dB. Alongside babble, a curated list of different noises which occur in everyday life such as traffic, crowded concert, airport announcements, air conditioners etc. have been included in the mixing process.

For testing, we wanted to create a comprehensive comparative analysis based on the noise and variable SNR performance for Hindi as well as English speech with the aim to find the best fit model for a given use case or environment. The test data for Hindi was obtained from IndicSUPERB [2] dataset developed by AI4Bharat, and for English it was obtained from test-clean LibriSpeech ASR corpus [13] available on OpenSLR. The unseen noises were taken from Microsoft Scalable Noisy Speech Dataset [14] (MS-SNSD) dataset.

With this goal in mind, we created a test set for Hindi and English with classes, namely Pitch/Frequency of the speaker, type of noise, SNR of the noisy audio, language of the audio and whether the noise was seen or unseen in training phase. With these classifications, we can evaluate the variation in per-

formance caused by language used to train the model along and the effectiveness of different models on different noises.

3.2. Evaluation Metrics

Perceptual Evaluation of Speech Quality (PESQ) and Short-Time Objective Intelligibility (STOI) are two widely used metrics in the field of speech signal processing to assess the quality and intelligibility of speech signals, respectively.[15] [16] PESQ measures the perceived quality of processed speech by comparing it to a reference signal, accounting for factors such as distortion, noise, and speech artifacts. It provides a single score indicating the perceived quality, with higher scores representing better quality. On the other hand, STOI evaluates the intelligibility of speech by comparing the short-term characteristics of the original and degraded speech signals, taking into account factors like background noise and reverberation.

The train test split used was 80 20. The DNN component of the model was trained on a total of 500 epochs and the GRU component was trained on 100 epochs.

3.3. Comparison with other State-of-the-Art methods

Table 1: *Experimental results on the prepared Datasets*

Model	WB-PESQ	STOI
Noisy (E)	1.945	0.877
Noisy (H)	2.079	0.795
DeepFilterNet (E)	2.617	0.925
DeepFilterNet (H)	1.967	0.801
SpeechBrain (E)	2.463	0.908
SpeechBrain (H)	1.755	0.783
DGDF (E)	1.447	0.845
DGDF (H)	1.294	0.649
DNN-GRU (Pro.) (E)	1.409	0.850
DNN-GRU (Pro.) (H)	1.346	0.687

In Table 1 we represent our model as DNN-GRU (Pro.) for proposed DNN-GRU, (E) stands for English test set performance and (H) for Hindi test set performance. The SNR values are averaged over all types of noises, each noise was mixed at -5, 0, 10, 20, 40 SNR levels. DGDF is short for DNN-GRU output passed to DeepFilterNet. Table 2 shows comparison for the

Table 2: *Comparison of DNN-GRU Model Performance*

Model	WB-PESQ	STOI
DNN-GRU (Pro.) (HoE)	1.239	0.77
DNN-GRU (Pro.) (EoH)	1.351	0.705
DNN-GRU (Pro.) (HoH)	1.34	0.687
DNN-GRU (Pro.) (EoE)	1.409	0.844

proposed model when trained on different language and tested on other. For example EoH stands for the model was trained on English data but tested on Hindi dataset.

Table 3 shows the performance of models at highly noisy, -5SNR level.

The Table 4 shows the performance of the models on specific Noises, + indicates the type of noise that was mixed. The result was averaged out over the test dataset.

Table 3: *Performance Comparison at -5SNR*

Model	WB-PESQ	STOI
DeepFilterNet (E)	1.319	0.790
DNN-GRU (Pro.) (E)	1.185	0.729
DGDF (E)	1.184	0.709
SpeechBrain (E)	1.157	0.734
DeepFilterNet (H)	1.281	0.618
DNN-GRU (H)	1.146	0.556
DGDF (H)	1.170	0.526
SpeechBrain (H)	1.107	0.525
Noisy (H)	1.104	0.543
Noisy (E)	1.063	0.682

Table 4: *Performance Based on Noise Type*

Model	WB-PESQ	STOI
DeepFilterNet (E) + Air conditioner	2.900	0.954
DeepFilterNet (E) + Miscellaneous	2.683	0.940
DeepFilterNet (E) + Airport	2.514	0.915
DeepFilterNet (E) + Babble	2.480	0.901
DeepFilterNet (H) + Air conditioner	2.173	0.845
DeepFilterNet (H) + Miscellaneous	2.064	0.828
DeepFilterNet (H) + Babble	1.868	0.775
DeepFilterNet (H) + Airport	1.860	0.779
SpeechBrain (H) + Air conditioner	1.750	0.833
SpeechBrain (H) + Miscellaneous	1.716	0.821
SpeechBrain (H) + Airport	1.696	0.796
SpeechBrain (E) + Airport	1.656	0.214
SpeechBrain (E) + Miscellaneous	1.647	0.193
SpeechBrain (E) + Babble	1.646	0.205
SpeechBrain (E) + Air conditioner	1.633	0.210
SpeechBrain (H) + Babble	1.632	0.767
DGDF (E) + Air conditioner	1.469	0.867

4. Conclusion

The Table 1 WB-PESQ values show that our proposed DNN-GRU model trained on just 20 hours of -5SNR data performed comparably well with state of the art models. Our proposed DNN-GRU English model was trained on a larger dataset compared to our proposed DNN-GRU Hindi model, which is a contributing reason to why it outperforms DNN-GRU Hindi model on Hindi test dataset. This demonstrates that English models tested on Hindi test dataset or vice versa don't show any significant advantage with less hours of training size. Table 3 demonstrates that training DNN-GRU on just 20 hours of -5SNR English data, performed comparable to DeepFilterNet which is trained on greater hours of English data of multiple SNR. It highlights the fact that the model generalised very well on all SNRs when trained solely on -5SNR.Therefore this model can perform well even with limited data resources which is a common problem among less popular Indic languages. It also shows that DeepFilterNet performs better on English than the Hindi test dataset, but the STOI values of Noisy English are higher than Noisy Hindi thus explaining the relative ranking.Table 4 highlights that DeepFilterNet and SpeechBrain performed significantly better than all other models on the all types of noises. In Future we wish to extend the analysis of the performances of the above stated model to other Indic languages like Tamil, Telugu, Marathi. We also aim to test the models on a larger dataset

for English and Hindi for a comprehensive analysis.

5. References

[1] S. Dey, M. Sahidullah, and G. Saha, "An overview of indian spoken language recognition from machine learning perspective," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 21, no. 6, nov 2022. [Online]. Available: https://doi.org/10.1145/3523179

[2] T. Javed, K. Bhogale, A. Raman, P. Kumar, A. Kunchukuttan, and M. M. Khapra, "Indicsuperb: A speech processing universal performance benchmark for indian languages," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 11, pp. 12 942–12 950, Jun. 2023. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/26521

[3] K. S. Bhogale, A. Raman, T. Javed, S. Doddapaneni, A. Kunchukuttan, P. Kumar, and M. M. Khapra, "Effectiveness of mining audio and text pairs from public data for improving asr systems for low-resource languages," 2022.

[4] C. K. A. Reddy, E. Beyrami, J. Pool, R. Cutler, S. Srinivasan, and J. Gehrke, "A scalable noisy speech dataset and online subjective test framework," 2019.

[5] H. Vinay, P. Lavanya, A. A. Hippargi, A. Purohith, and D. Lohith, "A comparative analysis on speech enhancement and coding techniques," in *2021 International Conference on Recent Trends on Electronics, Information, Communication Technology (RTEICT)*, 2021, pp. 543–549.

[6] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.

[7] H. Schröter, A. N. Escalante-B., T. Rosenkranz, and A. Maier, "Deepfilternet: A low complexity speech enhancement framework for full-band audio based on deep filtering," 2022.

[8] ——, "DeepFilterNet2: Towards real-time speech enhancement on embedded devices for full-band audio," in *17th International Workshop on Acoustic Signal Enhancement (IWAENC 2022)*, 2022.

[9] H. Schröter, T. Rosenkranz, A. N. Escalante-B., and A. Maier, "DeepFilterNet: Perceptually motivated real-time speech enhancement," in *INTERSPEECH*, 2023.

[10] L. Liu, H. Guan, J. Ma, W. Dai, G. Wang, and S. Ding, "A mask free neural network for monaural speech enhancement," 2023.

[11] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.

[12] Y. Wang, J. Han, T. Zhang, and D. Qing, "Speech enhancement from fused features based on deep neural network and gated recurrent unit network," *EURASIP Journal on Advances in Signal Processing*, vol. 2021, no. 1, p. 104, 2021.

[13] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[14] C. K. A. Reddy, E. Beyrami, J. Pool, R. Cutler, S. Srinivasan, and J. Gehrke, "A scalable noisy speech dataset and online subjective test framework," 2019.

[15] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, vol. 2, 2001, pp. 749–752 vol.2.

[16] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 4214–4217.