Improving Performance of Hearing Aids for Speech-in-Noise

1. Introduction

Multiple studies have revealed challenges in understanding the real cause and effects related to hearing aids. One of the main problems for hearing impaired is the reduction of speech intelligibility in noisy environments, which is mainly caused by the loss of temporal and spectral resolution in the auditory processing of the impaired ear [1]. This ICASSP SP Clarity challenge aims to improve the performance of hearing aids for a typical domestic noise scenario. The signals captured by the microphones on a pair of behind-the-ear hearing aids and those captured at the eardrum are provided with the motive of improving speech intelligibility without excessive loss of quality.

The given work proposes a speech enhancement system that works in two stages. In the first stage, 'Spectral gating' (SG) technique is used that separates out speech signal from the background noise by applying a threshold to the magnitude spectrum of the signal, which effectively "gates" or filters out the noise components [2]. In the second stage, Long Short-Term Memory (LSTM), a recurrent neural network (RNN) was concatenated after the spectral gating denoising network for further enhancement of the denoised signals [3].

2. Method

The overall architecture of the method is shown in Fig.1. For each ear of a hearing-impaired listener, a denoising module and a source separation module is optimized in order to enhance the noisy signals in two stages. The two stages of the model are described as: (i) First stage uses a denoising module that works by applying a frequency dependent gate to the audio signal. The method computes spectrogram of the given signal and estimate a noise threshold (or a gate) for different frequency band of the signal. (ii) In the second stage, signal outcome from first stage is fed as an input to our recurrent neural network (LSTM) that effectively separates out different source signal present in the mixed signal. All components are implemented with TensorFlow [4], and the back-propagation algorithm is used for the optimization.

2.1 First stage: Spectral Gating (SG)

In the primary stage, SG method is implemented using noise reduce library in python. In it, mixed signal (target + interferers) is first divided into small overlapping frames. Power spectrum corresponding to each frame is evaluated by computing FFT of each frame. Then a threshold function followed by a smoothing function is applied to the power spectrum of the mixed signal that separate out speech components from the noise components. That threshold is used to compute a mask, which gates noise below the frequency-varying threshold and thus suppress the noise component present in the signal [5]. The final speech spectrum is then transformed back into the time domain using the Inverse Fast Fourier Transform (IFFT) to obtain the final speech signal with reduced noise. We have used non-stationary noise reduction algorithm that continuously updates the estimated noise threshold over time [2]. The advantage of using this stage is that we are able to remove insignificant noise components from our mixed signal.

2.2 Second stage: Frequency Domain Audio Separation Network

We take an STFT of pre-processed (denoised) signal which is then fed to our Deep Learning model which is a recurrent neural network (LSTM). LSTM is a type of RNN architecture that is capable of learning long-term dependencies and addresses vanishing gradient issue by adding memory cells, gates, and pathways [6]. Our Model consists of three sequential LSTM layers followed by two parallel time distributed dense layers which predict our clean speech and interference respectively. We use their corresponding reference signals (anechoic and interferer) as ground truth for gradient descent. The separation network is trained to separate the processed mixture signal to individual anechoic target and interference signals.

3. Experiments

3.1 Data

The Training and development datasets consist of 6000 and 2500 simulated scenes respectively. There will be two evaluation datasets of 1500 scenes each as: (i) 1.5k simulated scenes generated in the same way as the training and development data (ii) 1.5k real ecologically valid dataset. All dataset scenes were sampled at 44.1kHz in .WAV format and saved in 32-bit floating point. For each scene, channel-1, which corresponds to front microphone location is used as input for both left and right ear. A clean anechoic signal is used as ground truth are provided

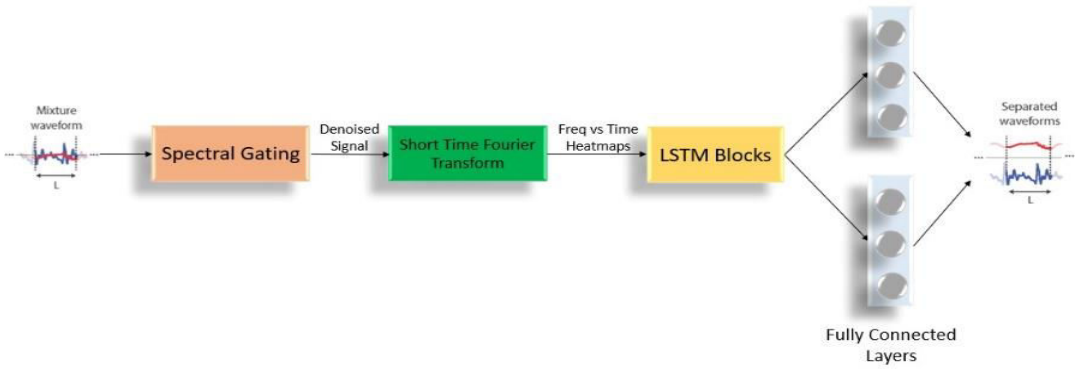

Fig.1: Overall architecture of the method

along with metadata consisting of audiograms of particular listeners and mapping of which listener will listen to which scene. All impulse responses that are used to model the propagation and modelling of sound and thus used for reproducing scenes are taken from OlHeadHRTF database [7]. For denoising module, channel-1(front microphone) is used as input, and one channel of the corresponding anechoic and interference signal is used as reference.

3.2 Setup

Recordings were carried out with two three channel BTE hearing aids and an additional internal microphone to record sound pressure near the location corresponding to the place of the human eardrum, resulting in a total of 8 recording channels. A NVIDIA GTX 1650 4GB GPU is used for training and inference for each ear of a listener. The signals are down sampled to 16 kHz in the optimization for faster training and inference. In the first stage, a window size of 2048 samples are considered with the hop length for number of audio samples between adjacent STFT columns being 512.

In the second optimization stage, RNN network used in this work have 3 LSTM layers with 600 units in each layer. Each windowed speech segment was processed with a 1024-point FFT with a hop length of 512. Sampling rate considered in this stage is 16kHz. LSTM model runs with hyperparameters as follows: Batch size: 10, learning rate: 10^{-5}, no. of epochs: 200, dropout rate: 0.2, activation function: ReLU and validation split of 0.1 out of 6k dataset.

4. Evaluation

The hearing-aid speech perception index (HASPI) and the hearing aid sound quality index (HASQI) which are two well-known objective evaluation metrics for speech intelligibility and speech quality, are used to evaluate the performance of the speech enhancement task for hearing aid application [8], [9]. The average of these two is computed and returned for each signal. The evaluation stage will first pass signals through a provided hearing aid amplification stage using a NAL-R [10] fitting amplification and a simple automatic gain compressor. The amplification is determined by the audiograms defined by the scene-listener pairs for the development set. The score for the baseline enhancement is 0.185 overall (0.239 HASPI; 0.132 HASQI). The enhanced stereo signals produced at output of enhancement stage are submitted for evaluation.

5. Results and Conclusion

In this work, we present a speech enhancement system, consisting of noise reduction in the former stage and source separation recurrent neural network in the later stage. Two separate evaluation datasets each consisting of 1.5k scenes are processed and submitted for evaluation.

References

[1] Dillon, H. “Hearing aids”, Boomerang Press, Sydney: New York, 2001

[2] Kumar, E. & Surya, K. & Varma, K. & Akash, A. & Kurapati, Nithish Reddy. (2023). Noise Reduction in Audio File Using Spectral Gatting and FFT by Python Modules.

10.3233/ATDE221305.

[3] Strake, Maximilian & Defraene, B. & Fluyt, Kristoff & Tirry, Wouter & Fingscheidt, Tim. (2020). Speech enhancement by LSTM-based noise suppression followed by CNN-based speech restoration. EURASIP Journal on Advances in Signal Processing. 2020. 10.1186/s13634-020-00707-1.

[4] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., & Zheng, X. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467.

[5] Tim Sainburg, "Noise reduction using Spectral Gating in Python", BSD License (MIT), January 24, 2020.

[6] M. Liu, Y. Wang, J. Wang, J. Wang and X. Xie, "Speech Enhancement Method Based On LSTM Neural Network for Speech Recognition," 2018 14th IEEE International Conference on Signal Processing (ICSP), Beijing, China, 2018, pp. 245-249, doi: 10.1109/ICSP.2018.8652331.

[7] F. Denk, S.M.A. Ernst, S.D. Ewert and B. Kollmeier (2018) Adapting hearing devices to the individual ear acoustics: Database and target response correction functions for various device styles. Trends in Hearing, vol 22, p. 1-19. DOI: 10.1177/2331216518779313.

[8] Kates, J.M. and Arehart, K.H., 2021. The hearing-aid speech perception index (HASPI) version 2. Speech Communication, 131, pp.35-46.

[9] Kates, J.M. and Arehart, K.H., 2014. "The hearing-aid speech quality index (HASQI) version 2". Journal of the Audio Engineering Society. 62 (3): 99–117.

[10] Byrne, Denis, and Harvey Dillon. "The National Acoustic Laboratories'(NAL) new procedure for selecting the gain and frequency response of a hearing aid." Ear and hearing 7.4 (1986): 257-265