OBJECTIVE COMPARISON OF SPEECH ENHANCEMENT ALGORITHMS WITH HEARING LOSS SIMULATION

Zhuohuang Zhang[1,2]; Yi Shen[1]; Donald S. Williamson[2]
[1]Department of Speech and Hearing Sciences, Indiana University; [2]Department of Computer Science, Indiana University
zhuozhan@iu.edu; {shen2, williads}@indiana.edu

Introduction

One goal of the current study is to compare many traditional and newly-emerged speech enhancement algorithms, using a large database that contains diverse mixtures of speech and background noise under a broad range of SNRs. A second goal of this study is to evaluate the performance of these algorithms for people with hearing impairments. Most previous studies evaluated speech-enhancement outcomes using metrics developed for healthy young adults, such as the widely-adopted perceptual evaluation of speech quality (PESQ). It is not clear whether the findings using these metrics hold for the hearing-impaired population. The current study includes evaluations using the hearing-aid speech quality index (HASQI) (Kates and Arehart, 2014).

Speech Enhancement Algorithms

Active Set Newton Algorithm (ASNA) (Virtanen et al., 2013):
It applies the Newton method to update the weights more efficiently than other NMF approaches. Parameters match those of the original study.

DNN-based ideal ratio mask estimation (D-IRM) (Wang et al., 2014):
This DNN-IRM network has three hidden layers with 1024 units each. The rectified linear (ReLU) activation function is applied to the hidden layers and a linear activation function is applied to the output layer. The mean square error is used as the cost function.

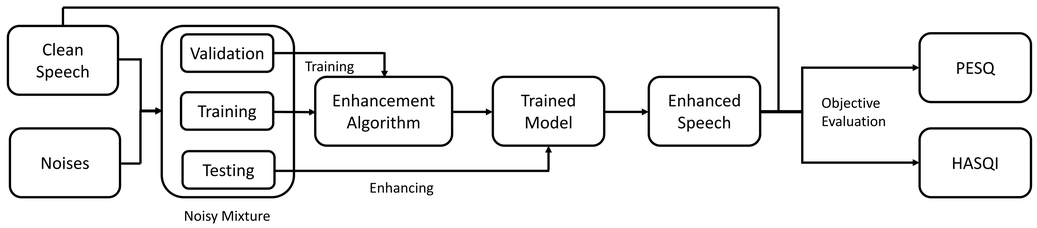
DNN-based complex ideal ratio mask estimation (D-cIRM) (Williamson et al., 2016):
The cIRM is predicted with a network that has three hidden layers with 1024 units each. All hidden layers use ReLU activation functions. The output layer uses a linear activation function.

LSTM-based ideal ratio mask estimation (L-IRM) (Weninger et al., 2014):
The network has two LSTM layers with 256 nodes in each layer, followed by a third sigmoidal layer. Mask approximation (MA) is used as the cost function.

BLSTM-based phase-sensitive mask estimation (BL-PSM) (Erdogan et al., 2015):
The network has two BLSTM layers with 256 nodes in each layer, followed by a third sigmoidal layer. Phase-sensitive spectrum approximation (PSA) is used as the cost function.

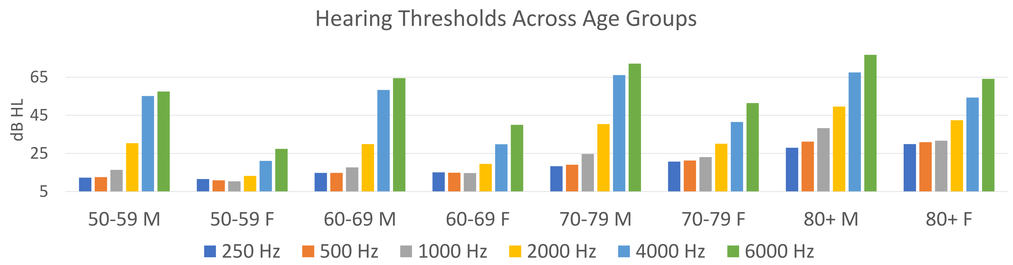
A Mel-frequency domain implementation is applied for all DNN and RNN-based methods.

Experiment Design and Results

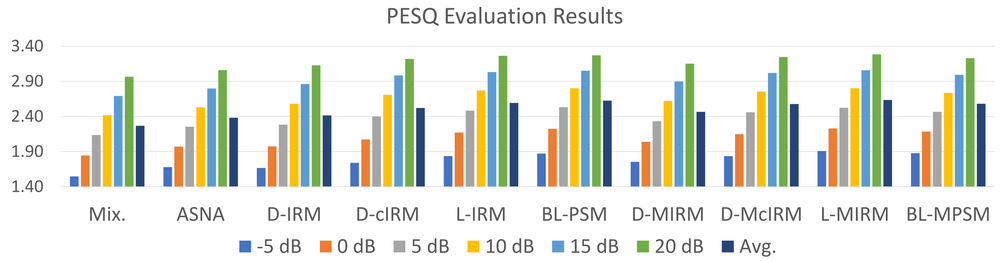


Experimental Design

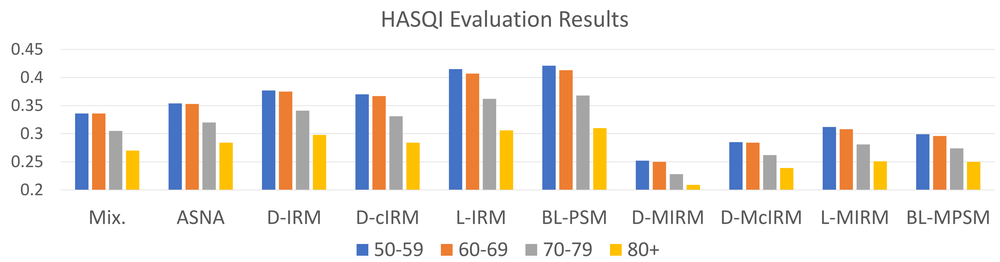
Hearing thresholds (dB HL) of male (M) and female (F) subjects across age groups (Schmiedt 2010):



PESQ (-0.5 to 4.5) evaluation results are averaged across noise types for brevity:



HASQI (0 to 1) evaluation results are averaged across noise types, genders and SNRs for brevity:



Materials

The speech data includes 1440 IEEE utterances, 250 utterances from the Hearing in Noise Test (HINT) corpus and 2342 utterances from the TIMIT database. 70% of them are used for the training set and 15% are used for both the testing and development sets. The clean utterances are further corrupted by four types of noises at different levels [-5 dB to 20 dB], including airplane, babble, dog barking, and train noises.

Conclusions

- We investigated the performance of several speech enhancement algorithms on a diverse speech dataset, with a particular interest in simulated hearing loss environments.
- The RNN-based methods result in significantly higher PESQ and HASQI scores for normal-hearing listeners.
- For hearing-impaired listeners, the BLSTM method achieves the best performance in all age groups for both genders.
- We also found that for both DNN- and RNN-based methods, Mel-frequency domain processing can often lead to improved PESQ scores, but reduced HASQI scores.
- Future studies that include subjective evaluations are warranted to confirm the performance of these algorithms for normal and hearing-impaired listeners.

Acknowledgements

This research was supported in part by a NSF Grant (IIS-1755844) and NVIDIA GPU Grant Program. We thank James Kates for providing HASQI code and Indiana University Pervasive Technology Institute for providing HPC (Karst) resources that have contributed to the research results reported within this paper.

INDIANA UNIVERSITY

NSF