OBJECTIVE COMPARISON OF SPEECH ENHANCEMENT ALGORITHMS WITH HEARING LOSS SIMULATION

Zhuohuang Zhang[⋆†] *Yi Shen*[⋆] *Donald S. Williamson*[†]

[⋆] Department of Speech and Hearing Sciences, Indiana University, USA
[†] Department of Computer Science, Indiana University, USA
zhuozhan@iu.edu { shen2, williads } @indiana.edu

ABSTRACT

Many speech enhancement algorithms have been proposed over the years and it has been shown that deep neural networks can lead to significant improvements. These algorithms, however, have not been validated for hearing-impaired listeners. Additionally, these algorithms are often evaluated under a limited range of signal-to-noise ratios (SNR). Here, we construct a diverse speech dataset with a broad range of SNRs and noises. Several enhancement algorithms are compared under both normal-hearing and simulated hearing-impaired conditions, where the perceptual evaluation of speech quality (PESQ) and hearing-aid speech quality index (HASQI) are used as objective metrics. The impact of the data's frequency scale (Mel versus linear) on performance is also evaluated. Results show that a long short-term memory (LSTM) network with data in the Mel-frequency domain yields the best performance for PESQ, and a Bidirectional LSTM network with data in the linear frequency scale performs the best in hearing-impaired settings. The Mel-frequency scale results in improved PESQ scores, but reduced HASQI scores.

Index Terms— speech enhancement, hearing loss, deep neural networks, long short-term memory

1. INTRODUCTION

Speech degradation in the presence of noise is a common problem for individuals, especially for people with hearing impairments [1]. A plethora of speech enhancement algorithms have been proposed for improving speech intelligibility and perceived speech quality in noisy environments, including those based on statistical-model [2], non-negative matrix factorization (NMF) [3–6], deep neural networks (DNNs), and recurrent neural networks (RNNs) [7–10]. Recent DNN- and RNN-based speech enhancement algorithms have shown superior performance over traditional approaches, but there has been a lack of parallel comparisons amongst these algorithms. A previous study by Hu and Loizou [11] provides an overview of the relative performance of different speech enhancement algorithms, but many of the algorithms are no longer considered, largely due to the rise of DNN-based approaches. In many recent studies, speech materials and background noises are limited, and only a narrow range of signal-to-noise ratios (SNRs) are used. 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. Hearing loss affects tens of millions of individuals in the United States [12], and it is highly prevalent among older adults [13]. Approximately one in three people in the United States between 65 and 74 years of age live with a hearing impairment. In contrast, 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) [14]. It is not clear whether the findings using these metrics hold for the hearing-impaired population. It is crucial to evaluate speech enhancement algorithms for hearing-impaired listeners in order to generalize laboratory results to real-world applications involving listeners of all ages. The current study includes evaluations using the hearing-aid speech quality index (HASQI) [15], to quantify how speech enhancement algorithms impact those with hearing loss. Its important to note that HASQI's computations resemble the processing performed by normal and impaired auditory systems. Compared to PESQ, HASQI is more adept at predicting the perceived speech quality ratings that are provided by hearing-impaired listeners that use hearing aids [16, 17].

In the current study, the performance of various speech enhancement algorithms are compared using both PESQ and HASQI. These algorithms include one NMF-based approach [4] that serves as a strong baseline model, two DNN-based approaches [7, 8] that perform well for normal-hearing listeners, and two RNN-based approaches [9, 10], which determine the impact that recurrent structures have on speech enhancement

This research was supported in part by a NSF Grant (IIS-1755844).

for the hearing impaired. Meanwhile, the impact of the frequency scale (Mel versus linear) on the input and output data is also investigated in our study, since studies have used different frequency scales without performing direct comparisons. Hence, its impact on normal-hearing and hearing-impaired listeners is not fully understood. The evaluations of speech quality using HASQI are separately conducted for different genders and age groups according to the typical trajectories of age-related hearing loss for female and male listeners.

The rest of this paper is organized as follows. Section 2 introduces the speech enhancement algorithms that are investigated. A detailed experimental setup is given in Section 3. Results are provided in Section 4. Finally, conclusions are given in Section 5.

2. SPEECH ENHANCEMENT ALGORITHMS

2.1. Active-set Newton algorithm

NMF is an efficient method for extracting target signals from mixtures and it is widely used for speech enhancement and other applications [18]. As an extension of NMF, the active-set Newton algorithm (ASNA) [4,5] is expressed as $\hat{x} = Bw$, where $\hat{x}$ is the target speech signal, B is the trained speech dictionary and w represents the activation weights. ASNA applies the Newton method to update the weights more efficiently than other NMF approaches, and it has been shown to outperform them in different environments. It will serve as the baseline model for this study. Parameters match those of the original study.

2.2. DNN-based ideal ratio mask estimation

A DNN-based method that estimates the ideal ratio mask (IRM) in the time-frequency (TF) domain is included in our study, since it shows performance advantages over other DNN-based training targets [7]. The IRM is defined as

$$M_{t,f}^{rm} = |s_{t,f}| \,/(|s_{t,f}| + |n_{t,f}|) \tag{1}$$

where $|s_{t,f}|$ represents the magnitude response of the clean speech signal and $|n_{t,f}|$ is the magnitude response of the noise signal at time t and frequency f.

As mentioned in [7], this DNN-IRM network has three hidden layers with 1024 units each. The rectified linear (ReLU) [19] activation function is applied to the hidden layers and a linear activation function is applied to the output layer. A set of complementary features [7] are used as the input to the network. The window size is set to 40 ms with a step size of 20 ms. We use adaptive gradient descent as the optimizer, with a mini-batch size of 512, and maximum epoch number of 80. The mean square error is used as the cost function. The output of the network is an estimated IRM in the linear frequency scale. To investigate the impact of frequency scales (linear or Mel), a Mel-frequency (100-bin) domain implementation is also investigated in our study. A signal can be converted between the linear and Mel-frequency scales using the following transformations

$$|s_{t,f}^{Mel}| = B|s_{t,f}|, \quad |s_{t,f}^{iMel}| = B^T|s_{t,f}^{Mel}| \tag{2}$$

where $|s_{t,f}^{Mel}|$ is the Mel-domain signal and $|s_{t,f}^{iMel}|$ is the linear scale signal after an inverse-Mel transformation. B represents a matrix of weights to combine short-time Fourier transform (STFT) bins into Mel bins, and B^T represents the transpose of B. Note that Mel-transformation is a lossy process, so some information from the original spectrogram will be lost during reconstruction.

2.3. DNN-based complex ideal ratio mask estimation

The authors in [8] propose a network that estimates the complex ideal ratio mask (cIRM) in the TF domain. This enables the DNN to predict the phase response in addition to the magnitude response. We include this method, since it is shown to outperform other training targets in objective and subjective evaluations, but the importance of phase for the hearing impaired is not well understood. The cIRM is defined as

$$M_{t,f}^{crm} = \frac{|s_{t,f}|}{|y_{t,f}|}\cos(\theta_{t,f}) + j\frac{|s_{t,f}|}{|y_{t,f}|}\sin(\theta_{t,f}) \tag{3}$$

where $|y_{t,f}|$ represents the magnitude response of the noisy speech, j indicates a imaginary number, and $\theta_{t,f} = \theta_{t,f}^{s} - \theta_{t,f}^{y}$, e.g., the phase difference between the speech and noisy speech. 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. Other parameters for this network are the same as the ones for the DNN-IRM. A Mel-frequency domain implementation is also included, which has not been previousely done for the cIRM.

2.4. LSTM-based ideal ratio mask estimation

Long short-term memory (LSTM) is a special type of RNN, which solves the problem of exploding and vanishing gradient of traditional RNNs [20]. The recurrent structure within LSTM networks makes it powerful in time series prediction, such as problems dealing with stock price prediction, speech recognition, and speech enhancement.

We implement the LSTM network architecture described in [9], since it shows impressive performance on speech enhancement tasks. The network has two LSTM layers with 256 nodes in each layer, followed by a third sigmoidal layer. It takes 100-bin log-Mel magnitude spectrograms as input and predicts an IRM for the clean speech signal in the Mel scale. During training, the window size is set to 25 ms with a hop size of 10 ms. Mask approximation (MA) [9] is used as the

cost function and it is defined as

$$E^{MA}\left(M_{pred}\right) = \sum_{t,f}\left(M_{true} - M_{pred}\right)^2 \quad (4)$$

where M_{pred} is the predicted mask and M_{true} is the IRM. Note that the previously mentioned DNN approaches also use a mask approximation cost function. The network is trained with time steps of 100, a mini-batch size of 25 sequences, and a maximum epoch number of 100. RMSprop is applied as the optimizer, since its been proven as a good choice for RNNs [21]. We further trained and tested this LSTM structure using inputs and outputs in the linear frequency scale for comparison.

2.5. Bidirectional LSTM-based Phase-sensitive Mask estimation

A bidirectional-LSTM (BLSTM) is a LSTM network that considers 'memory' in both directions (i.e., past series and future series). To investigate the influence of this memory difference, a BLSTM architecture developed by Erdogan et al. [10] is investigated in this study. The BLSTM estimates a Mel-domain phase-sensitive mask (PSM) that is defined as

$$M_{t,f}^{psm} = \frac{|s_{t,f}|}{|y_{t,f}|} cos(\theta_{t,f}). \quad (5)$$

The PSM is truncated between 0 and 1. A phase-sensitive spectrum approximation (PSA) is used as the cost function, since it leads to significant improvements over the mask-based cost function [10]. This is defined as

$$E^{PSA}\left(M_{pred}\right) = \sum_{t,f}\left(M_{true}|y_{t,f}| - M_{pred}|y_{t,f}|\right)^2 \quad (6)$$

where M_{true} is the ideal PSM and M_{pred} is the estimated one. The network has two BLSTM layers with 256 nodes in each layer. Other settings are identical to the LSTM method. A linear frequency domain implementation of the BLSTM is also included for comparison.

3. EXPERIMENTAL SETUP

3.1. Material

Utterances from three speech corpora are combined, in order to investigate the performance of the above-described algorithms on diverse speech materials. The speech data includes 1440 IEEE utterances [22] for both male and female speakers, 250 male-speech utterances from the Hearing in Noise Test (HINT) corpus [23] and 2342 male and female utterances from the TIMIT database [24]. This results in a total of 4032 clean speech utterances, where 2822 (70%) of them are used for the training set and 605 (15%) are used for both the testing and development sets. The clean utterances are further corrupted by four types of noises at different levels, including airplane, babble, dog barking, and train noises. Noises are extracted from the Azbio [25] and ESC-50 datasets [26]. The clean speech and noise are mixed at several SNRs ranging from -5 dB to 20 dB with a step of 5 dB. All speech and noise signals are downsampled to a 16 kHz sampling rate before mixing. In total, the training set contains 16932 mixtures for each noise type. The development and testing sets consist of 3630 mixtures.

Table 1: Hearing thresholds (dB HL) of male (M) and female (F) subjects across different age groups.

Age Group	Frequency (Hz)					
	250	500	1000	2000	4000	6000
50-59 M	12.3	12.6	16.4	30.4	55.1	57.5
50-59 F	11.6	10.9	10.4	13.2	21.1	27.4
60-69 M	14.8	14.8	17.7	29.9	58.3	64.5
60-69 F	15.1	14.9	14.7	19.5	29.8	40.0
70-79 M	18.3	19.1	24.7	40.4	66.1	72.1
70-79 F	20.7	21.3	23.1	30.1	41.5	51.4
80+ M	28.0	31.2	38.3	49.6	67.5	76.7
80+ F	29.9	30.9	31.7	42.4	54.3	64.1

3.2. Objective metrics

We use PESQ as the objective speech quality metric for simulating evaluations by normal-hearing listeners. It was originally designed for evaluating speech signals that are transmitted over telephone lines, (see the ITU-T P.862 [14]), but it has been shown to have strong correlations with subjective evaluations by individuals with normal hearing [27]. It predicts a mean opinion scores (MOS) that ranges from -0.5 (bad) to 4.5 (excellent). It accomplishes this by comparing a signal of interest to a reference clean speech signal.

HASQI, which is a newer metric, captures the noise effects, nonlinear distortions, linear filtering and spectral changes between two signals in order to resemble the processing that is performed by normal and impaired auditory systems. HASQI even achieves comparable performance to PESQ for normal-hearing based evaluations [16]. HASQI requires audiometric thresholds as a function of frequency to model the hearing loss of hearing-impaired individuals. Audiometric characteristics of various age groups for typical females (based on 936 listeners) and males (based on 756 listeners) [13] are implemented within HASQI under different testing conditions. The average audiometric thresholds used in these conditions are summarized in Table 1. Higher audiometric thresholds (in dB HL) indicate a greater degree of hearing loss. High-frequency hearing loss is typical among older adults. The severity of hearing loss grows with age and is greater for male listeners.

The clean reference signals are spectrally shaped to compensate for hearing loss before they are passed through the impaired auditory model in HASQI for comparison. A standard formula for hearing-aid fitting [28] is used to determine

Table 2: PESQ scores for speech enhancement algorithms at each SNR. **Bold** font indicates the highest score.

	Input SNR (dB)						
	-5	0	5	10	15	20	Avg.
Mixture	1.55	1.84	2.14	2.42	2.69	2.97	2.27
ASNA	1.68	1.97	2.25	2.53	2.80	3.06	2.38
D-IRM	1.67	1.97	2.28	2.58	2.86	3.13	2.42
D-cIRM	1.74	2.07	2.40	2.71	2.98	3.22	2.52
L-IRM	1.84	2.17	2.48	2.77	3.03	3.26	2.59
BL-PSM	1.87	2.22	**2.53**	**2.80**	3.05	3.27	2.63
D-MIRM	1.75	2.04	2.33	2.62	2.90	3.15	2.47
D-McIRM	1.84	2.15	2.46	2.75	3.02	3.25	2.58
L-MIRM	**1.91**	**2.23**	2.52	2.80	**3.06**	**3.28**	**2.63**
BL-MPSM	1.88	2.19	2.47	2.74	2.99	3.23	2.58

the amount of amplification in each frequency region. Therefore, the predicted speech quality [between 0 (poorest) and 1 (perfect)] simulates the rating given by hearing-impaired people without hearing aids.

4. RESULTS

4.1. Normal-hearing evaluation results

Table 2 lists the PESQ scores for the original mixtures without enhancement, as well as the enhanced signals that are outputted by the speech enhancement algorithms. We show the PESQ scores averaged across the four noises for brevity. D-IRM and D-cIRM represent the DNN-based IRM and cIRM approaches, respectively. L-IRM and BL-PSM represent LSTM and BLSTM structures with IRM and PSM training targets, respectively. D-MIRM, D-McIRM, L-MIRM, BL-MPSM indicate the structures in the Mel-frequency domain.

All these algorithms improve the quality of the noisy mixtures according to the PESQ scores. Deep learning algorithms also significantly outperform the NMF-based ASNA approach, which is consistent with results from [7–9]. At the lower SNRs (i.e., -5 and 0 dB), the LSTM Mel-scale method performs the best. In SNRs from 5 to 10 dB, the BLSTM linear-scale approach performs better than other deep learning methods. While at higher SNRs (i.e., 15 and 20 dB), the LSTM Mel-scale method performs the best. Averaging across the SNRs, the LSTM Mel-scale method slightly outperforms the other algorithms. Mel-frequency domain processing often leads to improved performance for both DNN- and RNN-based structures. RNN-based structures perform better than conventional DNN-based methods, but we attribute this to LSTM's advantages of dealing with time series data and solving the problem for vanishing gradients [20] in RNNs.

4.2. Hearing-impaired evaluation results

Tables 3 provides the HASQI scores for various hearing loss conditions, averaged over all noises and genders for brevity. In general, most algorithms show improvement in speech quality for hearing-impaired listeners. Moreover, the amount of improvement decreases with increasing age, which is expected since these are the more challenging cases. Among the speech enhancement algorithms, the BL-PSM linear-scale method performs the best across all age groups, but its results are almost identical to L-IRM. Within the DNN-based approaches, the IRM linear-scale approach and cIRM approach perform similarly (linear and Mel scale).

Table 3: HASQI scores (averaged across genders) for speech enhancement algorithms across noise conditions and SNRs.

	Age Groups				
	50-59	60-69	70-79	80+	Avg.
Mixture	0.336	0.336	0.305	0.270	0.312
ASNA	0.354	0.353	0.320	0.284	0.328
D-IRM	0.377	0.375	0.341	0.298	0.348
D-cIRM	0.370	0.367	0.331	0.284	0.338
L-IRM	0.415	0.407	0.362	0.306	0.372
BL-PSM	**0.421**	**0.413**	**0.368**	**0.310**	**0.378**
D-MIRM	0.252	0.250	0.228	0.209	0.235
D-McIRM	0.285	0.284	0.262	0.239	0.268
L-MIRM	0.312	0.308	0.281	0.251	0.288
BL-MPSM	0.299	0.296	0.274	0.250	0.280

Surprisingly, we notice that for the DNN- and RNN-based methods, the Mel-domain processing results in reduced HASQI scores as compared to the linear-frequency domain approaches. This is contrary to the PESQ results, where the Mel-domain processing usually improves speech quality. We infer that this may result from the deteriorated frequency resolution of the enhanced signals following the Mel-domain transformation, especially at higher frequencies. PESQ is unaffected by this, since it assesses the speech quality on a narrower frequency range (3.1 kHz) [14] than HASQI (12 kHz) does [15].

5. CONCLUSIONS AND FUTURE WORK

We investigate 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.

6. ACKNOWLEDGEMENTS

We thank James Kates for providing HASQI code and Indiana University Pervasive Technology Institute [29] for providing HPC (Karst) resources that have contributed to the research results reported within this paper.

7. REFERENCES

[1] H. Glyde, L. Hickson, S. Cameron, and H. Dillon, "Problems hearing in noise in older adults: a review of spatial processing disorder," *Trends in amplification*, vol. 15, pp. 116–126, 2011.

[2] Y. Ephraimm and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. ASSP*, vol. 33, pp. 443–445, 1985.

[3] C. Févotte, J. Le Roux, and J. R. Hershey, "Non-negative dynamical system with application to speech and audio," in *Proc. ICASSP*, 2013, pp. 3158–3162.

[4] T. Virtanen, J. F. Gemmeke, and B. Raj, "Active-set newton algorithm for overcomplete non-negative representations of audio," *IEEE Trans. ASSP*, vol. 21, pp. 2277–2289, 2013.

[5] T. Virtanen, B. Raj, and J. F. Gemmeke, "Active-set newton algorithm for non-negative sparse coding of audio," in *Proc. ICASSP*, 2014, pp. 3092–3096.

[6] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Trans. ASSP*, vol. 21, pp. 2140–2151, 2013.

[7] Y. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," *TASLP*, vol. 22, pp. 1849–1858, 2014.

[8] D. Williamson, Y. Wang, and D. L. Wang, "Complex ratio masking for monaural speech separation," *TASLP*, vol. 24, pp. 483–492, 2016.

[9] F. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *Proc. GlobalSIP*, 2014.

[10] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. ICASSP*, 2015, pp. 708–712.

[11] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech communication*, vol. 49, pp. 588–601, 2007.

[12] J. Shargorodsky, S. G. Curhan, G. C. Curhan, and R. Eavey, "Change in prevalence of hearing loss in us adolescents," *JAMA*, vol. 304, pp. 772–778, 2010.

[13] R. A. Schmiedt, "The physiology of cochlear presbycusis," in *The aging auditory system*, pp. 9–38. 2010.

[14] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, 2001, vol. 2, pp. 749–752.

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

[16] A. A. Kressner, D. V. Anderson, and C. J. Rozell, "Robustness of the hearing aid speech quality index (HASQI)," in *Proc. WASPAA*, 2011, pp. 209–212.

[17] A. A. Kressner, D. V. Anderson, and C. J. Rozell, "Evaluating the generalization of the hearing aid speech quality index (HASQI)," *IEEE Trans. ASSP*, vol. 21, pp. 407–415, 2013.

[18] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *NIPS*, 2001, pp. 556–562.

[19] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. AISTATS*, 2011, pp. 315–323.

[20] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, pp. 1735–1780, 1997.

[21] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural networks for machine learning*, vol. 4, pp. 26–31, 2012.

[22] E. H. Rothauser, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust*, vol. 17, pp. 225–246, 1969.

[23] M. Nilsson, S. D. Soli, and J. A. Sullivan, "Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise," *JASA*, vol. 95, pp. 1085–1099, 1994.

[24] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continuous speech corpus," *NASA STI/Recon technical report*, vol. 93, 1993.

[25] A. J. Spahr, M. F. Dorman, L. M. Litvak, S. Van Wie, R. H. Gifford, P. C. Loizou, L. M. Loiselle, T. Oakes, and S. Cook, "Development and validation of the azbio sentence lists," *Ear and hearing*, vol. 33, pp. 112, 2012.

[26] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *ACM Multimedia*, 2015, pp. 1015–1018.

[27] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. ASSP*, vol. 16, pp. 229–238, 2008.

[28] D. Byrne and H. Dillon, "The national acoustic laboratories' (NAL) new procedure for selecting the gain and frequency response of a hearing aid," *Ear and hearing*, vol. 7, pp. 257–265, 1986.

[29] C. A. Stewart, V. Welch, B. Plale, G. Fox, M. Pierce, and T. Sterling, "Indiana university pervasive technology institute," 2017.