PROCEEDINGS of the
23rd International Congress on Acoustics
9 to 13 September 2019 in Aachen, Germany

Impact of amplification on speech enhancement algorithms using an objective evaluation metric

Zhuohuang Zhang[(1,2)], Donald S. Williamson[(2)], Yi Shen[(1)]

[(1)]Department of Speech and Hearing Sciences, Indiana University, USA
[(2)]Department of Computer Science, Indiana University, USA
zhuozhan@iu.edu williads@indiana.edu shen2@indiana.edu

Abstract

Hearing loss is prevalent among elderly adults, which leads to speech-understanding difficulties in noisy environments. Speech enhancement algorithms are thus proposed to alleviate this speech-in-noise problem. However, most of these algorithms have not been evaluated for hearing-impaired people either with or without the use of hearing aids. In this study, we evaluated the performance of several speech enhancement algorithms [i.e., non-negative matrix factorization based, deep-neural-network based and long short-term memory (LSTM) based algorithms] for hearing-impaired listeners using an objective speech quality metric, namely the Hearing-Aid Speech Quality Index (HASQI). The HASQI is based on a physiologically inspired model of auditory processing, which also allows the simulation of hearing impairment. The evaluation was repeated separately for the typical hearing characteristics of different genders in various age groups. For the aided condition, linear amplification was implemented using the NAL-R prescription formula. The benefits from the speech enhancement algorithms decrease with increasing degrees of hearing loss. With amplification, the benefit diminishes for the listener group with the most severe hearing impairment. Among the various algorithms, the LSTM-based structures exhibit superior performance with and without amplification.

Keywords: Speech enhancement, Objective metric, Hearing loss

1 INTRODUCTION

In the United States, about one third of older adults between 65 and 74 years of age live with hearing impairment. Age-related hearing loss makes speech communication challenging, especially in noisy environments. The traditional approach to address the speech understanding difficulties experienced by older adults is the use of hearing aids. Hearing aids amplify the acoustic signals received at the ears and send the amplified signals into the ear canal. Typically, greater amounts of amplification would be prescribed to the frequency regions with greater degrees of hearing loss to ensure conversational speech is audible to the hearing-aid users. This strategy is capable to significantly improve speech understanding in quiet, however, when background noises are present both the speech and noise would be amplified, leading to limited benefits from amplification.

The advancements in speech enhancement techniques have created new opportunities to facilitate speech understanding among older adults. The goal of many speech enhancement algorithms is to remove background noise from noise-corrupted speech signals. It is hypothesized that applying amplification to the enhanced speech signals would improve speech understanding for listeners with age-related hearing loss.

The current study investigates the perceived speech quality following speech enhancement and amplification from hearing-impaired listeners using an objective speech-quality metric, namely the Hearing-Aid Speech Quality Index (HASQI) [1]. The HASQI metric is different from many other objective speech-quality metrics, such as the Perceptual Evaluation of Speech Quality (PESQ) index [2], in that it includes a physiologically inspired model of auditory processing. The inclusion of the model also allows HASQI to simulate auditory perception in impaired ears and predict the perceived speech quality by hearing-impaired listeners.

In a recent study, Zhang et al. [3] applied HASQI to compare the improvement in speech quality by a range of speech enhancement algorithms, including algorithms based on non-negative matrix factorization (NMF)

[4, 5, 6], deep neural networks (DNNs) [7, 8], and recurrent neural networks (RNNs) [9, 10]. The average hearing profiles for listeners in various age groups were used to configure HASQI to generate predictions on the perceived speech quality from these listener groups. Results showed that for both normal-hearing and hearing-impaired listeners the long short-term memory (LSTM) based network achieved the greatest improvement in speech quality over other enhancement techniques. Moreover, as the degree of hearing loss increased, the HASQI score decreased, indicating poorer speech quality. In this previous study, no compensation for the loss of audibility was applied for the hearing-impaired groups, therefore, it is possible that the decreased HASQI scores merely reflected the reduced available speech bandwidth due to hearing loss.

In the current study, the evaluations conducted by Zhang et al. [3] are repeated with and without amplification applied following speech enhancement. It will be shown that both speech enhancement and amplification improve speech quality; however, the benefits from them may not be additive depending on the degree of hearing loss and the specific speech enhancement algorithm.

The rest of this paper is organized as follows. Section 2 describes the speech enhancement algorithms that were investigated in the current study. The experimental setup is described in Section 3. Results are provided in Section 4. Finally, conclusions are drawn in Section 5.

2 SPEECH ENHANCEMENT ALGORITHMS

2.1 Active-set Newton algorithm (ASNA)

Non-negative matrix factorization is an efficient method for extracting target signals from mixtures of signal and noise, and it is widely used for speech enhancement and applications such as document clustering [11, 12]. The active-set Newton algorithm (ASNA) [5, 6] is an extension of NMF and applies the Newton method to update the weights. It can be 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 is more efficient compared to other NMF-based approaches in terms of processing time, and it has been shown to outperform many of them under various conditions. We use the same parameters as reported in the original study.

2.2 DNN-based ideal ratio mask estimation (D-IRM)

A DNN-based method that estimates the ideal ratio mask (IRM) in the time-frequency (TF) domain is included in the current study. Using the ideal ratio mask as the training target in a DNN-based algorithm has been shown to achieve better results than other training targets [7]. The IRM can be described as:

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

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

The implementation of this DNN-IRM network consists of three hidden layers with 1024 units each [7]. The rectified linear (ReLU) [13] activation function is used for all hidden layers, and linear activation function is used for the output layer. The input to the network is a set of complementary features [7]. When constructing the spectrogram, the window size is chosen as 40 ms with a step size of 20 ms. Adaptive gradient descent is used as the optimizer during training, with a mini-batch size of 512, and a maximum epoch number of 80. The mean squared error is used as the loss function. The network predicts an estimated IRM which is combined with the phase of the noisy mixture to reconstruct the enhanced speech.

2.3 DNN-based complex ideal ratio mask estimation (D-cIRM)

As mentioned above, the IRM estimated by the DNN-IRM network does not contain any phase information for the reconstruction of the enhanced speech, despite of the importance of phase for audio quality. An alternative DNN-based algorithm [8] utilizes the phase information by predicting a complex ideal ratio mask (cIRM) in

the TF domain. Instead of solely relying on magnitude information, this network is able to predict the phase as well. The cIRM is defined as:

$$M_{t,f}^{crm} = \frac{|s_{t,f}|}{|y_{t,f}|}\cos(\theta_{t,f}) + j\frac{|s_{t,f}|}{|y_{t,f}|}\sin(\theta_{t,f}), \tag{2}$$

where $|y_{t,f}|$ represents the magnitude response of the noisy speech, j is the imaginary unit, and $\theta_{t,f} = \theta_{t,f}^{s} - \theta_{t,f}^{y}$ is the phase difference between the speech and noisy speech. This network consists of three hidden layers with 1024 units each. The ReLU activation function is used as activation function for all hidden layers while the output layer uses a linear activation function. Other settings for this network are identical to the ones for the DNN-IRM approach.

2.4 LSTM-based ideal ratio mask estimation (L-IRM)

As a variant of RNN, LSTM network solves the problem of exploding and vanishing gradient that traditional RNNs face [14]. With the help of the recurrent structure, the LSTM networks are suitable for dealing with data that contain temporal information, such as those in speech translation and speech enhancement.

The structure of the implemented LSTM network is described in [9]. This network consists of two LSTM layers with 256 nodes in each layer, then connects to another dense layer with sigmodal activation function. It takes the log magnitude spectrogram as the input and outputs an estimated IRM for the enhanced speech. When generating the spectrogram, the window size is chosen as 25 ms with a hop size of 10 ms. The log magnitude spectrogram is then “chopped" into sequences of 100 timesteps. We use mask approximation (MA) [9] as the loss function which is defined as:

$$E^{MA}(M_{pred}) = \sum_{t,f}(M_{true} - M_{pred})^2, \tag{3}$$

where M_{pred} is the predicted IRM and M_{true} is the ground truth. During training, a mini-batch size of 25 sequences with a maximum epoch number of 100 is used. We also apply RMSprop as the optimizer since its been shown as a good choice when training RNNs [15].

2.5 Bidirectional LSTM-based phase-sensitive mask estimation (BL-PSM)

As an extension of LSTM, a bidirectional-LSTM (BLSTM) takes “memory" in both directions into account (i.e., past and future series). This BLSTM network structure is described in [10]. It has two BLSTM layers with 256 units each, which is followed by a third fully connected layer. It also takes the log magnitude spectrogram as the input, but predicts a phase-sensitive mask (PSM). The PSM is defined as:

$$M_{t,f}^{PSM} = \frac{|s_{t,f}|}{|y_{t,f}|}cos(\theta_{t,f}). \tag{4}$$

We truncate the PSM between 0 and 1 and use a phase-sensitive spectrum approximation (PSA) [10] as the loss function during training. The PSA is defined as:

$$E^{PSA}(M_{pred}) = \sum_{t,f}(M_{true}|y_{t,f}| - M_{pred}|y_{t,f}|)^2, \tag{5}$$

where M_{true} stands for the ideal PSM and M_{pred} is the estimated one. Other settings are the same as those described for the LSTM-based method.

3 EXPERIMENTAL SETUP

3.1 Speech material

Sentences from three speech corpora were used, including 250 male-speech utterances from the Hearing in Noise Test (HINT) corpus [16], 1440 IEEE utterances [17] containing both male and female talkers and 2342

Table 1. Hearing thresholds (dB HL) of male and female subjects from various age groups.

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

male and female utterances from the TIMIT dataset [18]. In total, there are 4032 clean speech utterances. We split them into three sets: (1) 2822 (70%) utterances in the training set, (2) 605 (15%) utterances for the development set and (3) 605 (15%) utterances for the test set. By mixing up the clean utterances with four different types of noise (i.e., airplane, babble, dog, train) at 6 signal-to-noise ratios (SNRs), we obtain 16932 mixtures for each type of noise. Among the four noise types, the ten-talker babble noise are extracted from the AzBio database [19] and the other noises are obtained from the ESC-50 dataset [20]. The SNRs range from -5 dB to 20 dB with a step size of 5 dB. All signals are resampled to 16 kHz before mixing.

3.2 Listener age groups and linear amplification

The typical audiometric thresholds for male and female listeners at various ages are provided in Table 1. The data listed in the table are based on 936 females and 756 males, reported in [21]. A higher hearing threshold (in dB HL) represents a greater degree of hearing loss. For each listener group, the perceived quality of enhanced speech is separately predicted using HASQI. This results in a HASQI score, ranging between 0 (poorest) to 1 (best). The computational procedure to calculate the HASQI score considers the audiometric thresholds for the listener group and configures the auditory model within HASQI accordingly to simulate the perceptual consequences according to the severity of the hearing loss. In the model, hearing loss not only reduces sensitivity to sound but also reduces the compressive nonlinearity associated with the healthy peripheral auditory system and degrades the resolving power to spectral details [1].

The HASQI metric is based on an intrusive algorithm, which means that it requires both the clean speech signal as the reference signal and the noisy speech mixture as the test signal at its input [22]. In the current study, the clean reference signal is amplified according to a standard hearing-aid prescription formula (i.e. NAL-R) [23]. This formula generates a fixed gain, independent of input level (i.e. linear amplification), for each of the frequency regions. Therefore, the reference signal used for computing the HASQI score represents speech heard in quiet through typical linear amplification.

For each combination of speech enhancement algorithm and listener group, speech quality is calculated using HASQI either before (“Mixture") or after (“Enhanced") speech enhancement is applied, and with (“Eq. On") or without (“Eq. Off") amplification. The test signal that is used to calculate the HASQI score is different depending on the condition. Specifically, when no amplification is applied, the noisy speech mixture is used in the “Mixture Eq. Off" condition, and the enhanced speech is used in the “Enhanced Eq. Off" condition. When amplification is applied, the test signal is the the noisy speech mixture amplified according to the NAL-R formula in the “Mixture Eq. On" condition, and it is the enhanced speech amplified according to the NAL-R formula in the “Enhanced Eq. On" condition.

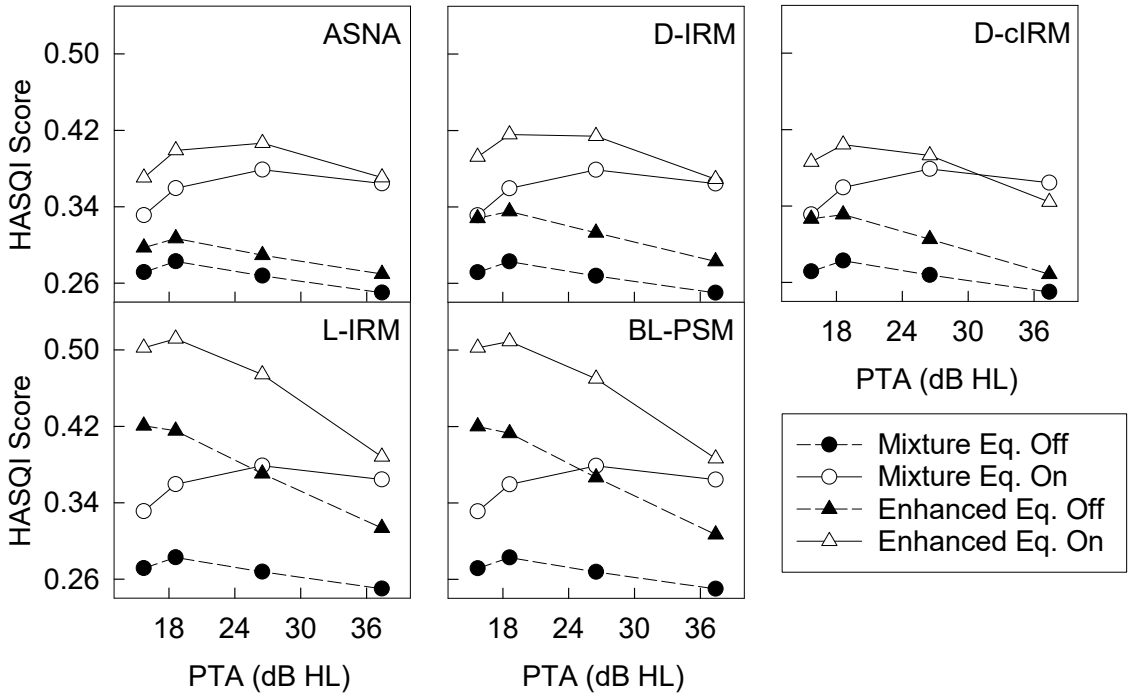

Figure 1. HASQI results with NAL-R amplification on and off

4 RESULTS

The calculated HASQI scores are shown in Figure 1 for the five speech enhancement algorithms evaluated in the current study (in separate panels). In each panel, the HASQI scores are plotted for the four test conditions as separate curves (with different symbols). Each curve in a panel shows the HASQI score as a function of pure-tone-average (PTA) threshold, which is the average of the thresholds at 500, 1000, and 2000 Hz. The PTA is commonly used to summarize the severity of hearing loss because 500, 1000, and 2000 Hz are frequencies that are important for speech understanding. The four data points along each curve indicate the four listener groups, the group with a higher age range corresponds to a higher PTA threshold.

As observed from Figure 1, amplification always leads to improvement in speech quality. Without speech enhancement, the benefit from amplification increases slightly as the degree of hearing loss increases (comparing filled and unfilled circles). This effect of hearing loss is less evident when speech enhancement is applied. Across all five speech enhancement algorithms, speech enhancement improved the predicted speech quality by HASQI (comparing triangles to circles). However, under the conditions with amplification, the benefit from speech enhancement diminishes as the PTA threshold increases. Comparing among the algorithms, the deep learning approaches exhibit significant improvements over the traditional NMF-based method (ASNA). The LSTM- and BLSTM-based structure perform the best among the algorithms both with and without amplification.

To investigate whether the benefits from speech enhancement and amplification are additive, we quantify the benefit from speech enhancement by the difference between the HASQI scores for the “Mixture Eq. Off" and “Enhanced Eq. Off" conditions (Δ_{enhance}) and the benefit from amplification by the difference between the HASQI scores for the “Mixture Eq. Off" and “Mixture Eq. On" conditions (Δ_{Eq}). Then, the expected HASQI

score for the “Enhanced Eq. On" condition assuming additivity of the benefits ($\widehat{\text{HASQI}}_{\text{Enhanced Eq. On}}$) would be:

$$\widehat{\text{HASQI}}_{\text{Enhanced Eq. On}} = \text{HASQI}_{\text{Mixture Eq. Off}} + \Delta_{\text{enhance}} + \Delta_{\text{Eq}}, \tag{6}$$

where $\text{HASQI}_{\text{Mixture Eq. Off}}$ is the HASQI score for the “Mixture Eq. Off" condition. The actual HASQI score for the “Enhanced Eq. On" condition ($\text{HASQI}_{\text{Enhanced Eq. On}}$) is close to $\widehat{\text{HASQI}}_{\text{Enhanced Eq. On}}$ for listener group with minimal hearing loss. On the other hand, when it is applied to listeners with higher degrees of hearing loss, the benefits from speech enhancement and amplification becomes subadditive:

$$\text{HASQI}_{\text{Enhanced Eq. On}} < \widehat{\text{HASQI}}_{\text{Enhanced Eq. On}}. \tag{7}$$

From our simulations, we infer that as the degree of hearing loss increases, there will be fewer benefits from wearing a digital hearing aid device with built-in speech enhancement. Although speech enhancement algorithm can bring many benefits in noisy environments when the amplification is not applied, the benefit decreases dramatically when amplification is applied.

5 CONCLUSIONS

The influence of amplification on speech quality is investigated using an objective metric in this study. Simulated results show that the benefits from speech enhancement are getting smaller as the hearing loss becomes more severe. The LSTM- and BLSTM-based methods achieve the best speech quality for listeners with various degrees of hearing loss. This holds true in both conditions when amplification is turned on and off. Future studies that include listening tests using hearing-impaired listeners will help to confirm the findings in this study.

ACKNOWLEDGEMENTS

This research is supported in part by the NSF Grant (IIS-1755844) and the Nvidia GPU Grant program. We also like to show our gratitude to James Kates for providing the software package for HASQI.

References

[1] J. M. Kates and K. H. Arehart, “The hearing-aid speech quality index (hasqi) version 2,” *Journal of the Audio Engineering Society*, vol. 62, no. 3, pp. 99–117, 2014.

[2] 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 *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.

[3] Z. Zhang, Y. Shen, and D. S. Williamson, “Objective comparison of speech enhancement algorithms with hearing loss simulation,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6845–6849.

[4] C. Févotte, J. Le Roux, and J. R. Hershey, “Non-negative dynamical system with application to speech and audio,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 3158–3162.

[5] T. Virtanen, J. F. Gemmeke, and B. Raj, “Active-set newton algorithm for overcomplete non-negative representations of audio,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 11, pp. 2277–2289, 2013.

[6] T. Virtanen, B. Raj, J. F. Gemmeke *et al.*, "Active-set newton algorithm for non-negative sparse coding of audio," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 3092–3096.

[7] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[8] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 3, 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 *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 2014, pp. 577–581.

[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 *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 708–712.

[11] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556–562.

[12] S. Sra and I. S. Dhillon, "Generalized nonnegative matrix approximations with bregman divergences," in *Advances in neural information processing systems*, 2006, pp. 283–290.

[13] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 2011, pp. 315–323.

[14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[15] 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, no. 2, pp. 26–31, 2012.

[16] 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," *The Journal of the Acoustical Society of America*, vol. 95, no. 2, pp. 1085–1099, 1994.

[17] E. Rothauser, "Ieee recommended practice for speech quality measurements," *IEEE Trans. on Audio and Electroacoustics*, vol. 17, pp. 225–246, 1969.

[18] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, 1993.

[19] 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, no. 1, p. 112, 2012.

[20] K. J. Piczak, "Esc: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 1015–1018.

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

[22] T. H. Falk, V. Parsa, J. F. Santos, K. Arehart, O. Hazrati, R. Huber, J. M. Kates, and S. Scollie, "Objective quality and intelligibility prediction for users of assistive listening devices: Advantages and limitations of existing tools," *IEEE signal processing magazine*, vol. 32, no. 2, pp. 114–124, 2015.

[23] 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, no. 4, pp. 257–265, 1986.