Effectiveness of Ideal Ratio Mask for Non-intrusive Quality Assessment of Noise Suppressed Speech

Meet H. Soni and Hemant A. Patil
Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT),
Gandhinagar, India
Email: {meet_soni, hemant_patil}@daiict.ac.in

***Abstract*—The Ideal Ratio Mask (IRM) has proven to be very effective tool in many applications such as speech segregation, speech enhancement for hearing aid design and noise robust speech recognition tasks. The IRM provides information regarding the amount of signal power at each Time-Frequency (T-F) unit in a given signal-plus-noise mixture. In this paper, we propose to use the IRM for non-intrusive quality assessment of noise suppressed speech. Since the quality of noise suppressed speech is dependent on the residual noise present in speech, IRM can be extremely useful for its quality assessment. The quality assessment problem is posed as a regression problem and the mapping between statistics of acoustic features, namely, Mel Filterbank Energies (FBEs) plus IRM features and the subjective score of the corresponding utterances was found using single-layer Artificial Neural Network (ANN). The results of our experiments suggest that by using the mean of FBEs and IRM features as the input, the quality prediction accuracy was significantly increased.**

I. Introduction

Non-intrusive quality assessment of speech is the problem of evaluating the perceptual quality of speech in the absence of any reference signal. There are many practical scenarios such as wireless communication, and Voice over IP (VoIP) in which the clean speech is not available as a reference. The absence of a reference signal makes the quality assessment a challenging problem. An early attempt towards non-intrusive assessment of speech based on spectrogram analysis is presented in [1]. The study reported in [2] uses Gaussian Mixture Models (GMMs) to create artificial reference models to compare degraded speech signals whereas in [3], speech quality is predicted by Bayesian inference and minimum mean square estimation (MMSE) based on trained GMMs. In [4], a speech quality assessment algorithm based on a temporal envelope representation of speech is presented. Different features extracted from speech have been detected to be useful for speech quality assessment. Spectral dynamics, spectral flatness, spectral centroid, spectral variance, fundamental frequency or pitch (F_0) excitation variance and perceptual linear prediction (PLP) coefficients were used for quality prediction in [5], [6]. In [7] and [8], the quality assessment problem is posed as a regression problem and the mapping between acoustic features and the subjective score was found using Mel Frequency Cepstral Coefficients (MFCCs) and filterbank energies, respectively. To find the mapping, Support Vector Regression (SVR) was used. Bag-of-Words (BoW) inspired codebook approach was presented in [9]. Spectro-temporal features and several combinations of auditory features were used for the same task in [10] and [11], respectively. Currently, ITU-T P.563 is the standard metric for non-intrusive quality assessment [12].

In this paper, we propose to use information of a Time-Frequency (T-F) mask for non-intrusive quality assessment. An ideal T-F mask gives information about whether, or to what extent, each T-F unit is dominated by target speech, which is the clean speech in this work. A binary decision about target dominant regions is represented using an Ideal Binary Mask (IBM) [13]. On the other hand, the decision about the ratio of target dominant power to mixture power is represented using an Ideal Ratio Mask (IRM) [14], [15]. We propose to use IRM for quality assessment task since it leads to better speech quality without compromising speech intelligibility [15]. Since IRM contains the information regarding the relative presence of residual noise in enhanced or noise suppressed speech, it provides a useful clue about the quality of a given utterance. The features extracted from IRM are used in addition to the standard acoustic feature set, namely, Mel Filterbank Energies (FBEs) to train a regression model and thus, the problem is posed as a regression problem [7], [8], [16]. The statistics of FBEs and IRM features are used as an input to the regression model. The model is trained to predict the subjective score of the given input pattern. Once trained, the model can be used to predict the quality of unknown test utterances. In the first part of our experiments, we show the effectiveness of IRM for the stated problem using true IRM extracted using clean speech signals. After establishing the effectiveness of the true IRM, we predict it directly from noise suppressed speech signal and use predicted IRM for the quality assessment task. Different experiments were conducted to check the robustness of proposed approach by dividing the data into different training and testing sets.

II. Ideal Ratio Mask (IRM) for Quality Assessment

A. IRM features

The IRM is widely used in speech segregation, speech enhancement and noise robust ASR [13], [15], [17], [18]. The IRM is defined as follows:

$$IRM(t,f) = \left(\frac{S^2(t,f)}{S^2(t,f) + N^2(t,f)} \right)^{\beta}, \qquad (1)$$

ISBN 978-0-9928626-7-1 © EURASIP 2017

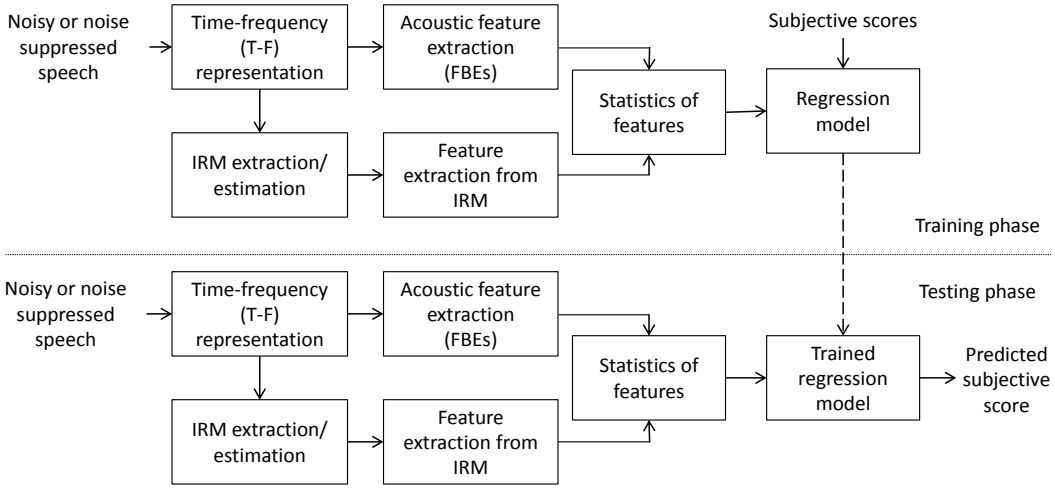

Fig. 1. Block diagram of proposed quality assessment system.

where $S^2(t,f)$ and $N^2(t,f)$ denote the speech and noise energy at a particular T-F point, respectively. β is a tunable parameter to scale the mask. We have used $\beta = 0.5$ [18]. With $\beta = 0.5$, equation 1 becomes similar to the square root Wiener filter, which is the optimal estimator of the power spectrum [15]. The IRM gives information about signal-to-noise ratio (SNR) at each T-F point which can be used to predict the quality of the utterance. IRM is applied on time-frequency (T-F) representation of noisy speech for different applications. In this paper, we have used IRM information for quality prediction along with spectral features extracted from noisy speech. Fast Fourier Transform (FFT) power spectrum and Gammatone power spectrum are used as the T-F representation of the speech signal and IRM is calculated from these T-F representations. The masks extracted using FFT power spectrum and Gammatone power spectrum are referred to as FFT-IRM and gammatone-IRM, respectively.

To use the IRM for quality prediction, low-dimensional features must be extracted from IRM which can provide effective representation. In this paper, recently proposed subband autoencoder (SBAE) [16], [19] is used for feature extraction from IRM. SBAE has been successfully used to learn the effective representation of speech spectrum for the non-intrusive speech quality assessment task [16] and for spoofed speech detection task [19]. In SBAE, the units of the first hidden layer are connected in a restricted manner with the units of the input layer. The restricted connectivity forces the units in the first hidden layer to learn the representation of one subband of the input spectrum. Hence, the first hidden layer is known as *subband layer*. More details about the architecture of SBAE and its advantages over AE can be found in [16], [19]. In presented work, SBAE is trained to learn *20-D* (dimensional) subband features from both FFT-IRM and gammatone-IRM.

B. Proposed quality assessment system

Fig. 1 shows the block diagram of the proposed non-intrusive quality assessment system. First of all, T-F representation is derived from noisy/enhanced speech signal. Acoustic features are then extracted from this T-F representation. As it can be observed from eq. (1) that IRM only consists local SNR information and not acoustic information of the underlying speech signal. Hence, some acoustic features must be used in order to capture the perceptual content of the speech signal under consideration. In this context, *40-D* Mel Filterbank Energies (FBEs) are used as acoustic features [8]. Initially, we extract the true IRM features to show their effectiveness for quality assessment task using a T-F representation of clean speech signal. The statistics of both FBEs and IRM features are then used as an input to the regression model. The subjective scores of training utterances are used while training the regression model. We have used the mean of FBEs, which is proven to perform better than using the variance of FBEs [8] and mean and variance of the IRM features at a time. However, in non-intrusive quality assessment, the clean speech signal is not available. Hence, true IRM cannot be calculated directly. In such case, IRM must be estimated or predicted using given noisy utterance only. To predict or estimate the IRM from given noisy or enhanced utterance, a Deep Neural Network (DNN) with three hidden layers is used [17]. More information about IRM prediction is given in Section III. Predicted IRM features are then used in further experiments.

III. Experimental Results

A. Experimental setup

All experiments were performed on NOIZEUS database [20]. The database has speech files which were corrupted by different kinds (and amount) of noise. It also had speech files enhanced using different noise suppression algorithms. The speech files were corrupted by four types of noise, namely,

ISBN 978-0-9928626-7-1 © EURASIP 2017

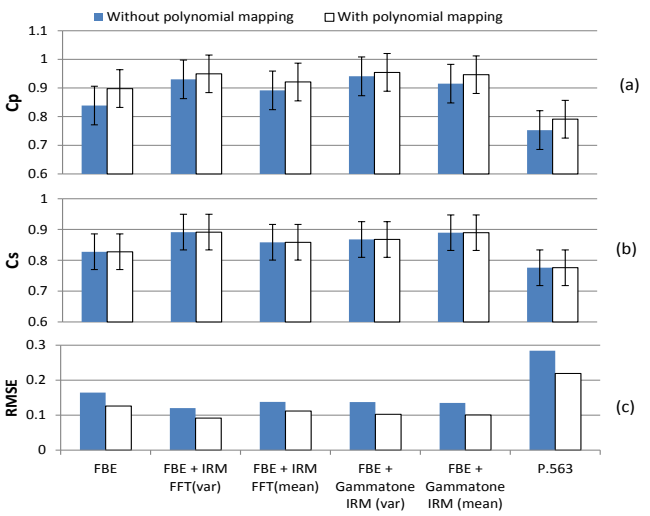

Fig. 2. (a) C_p, (b) C_s and (c) RMSE for *8*-fold CV experiment using true IRM features. C_p and C_s are shown with *95 %* confidence interval.

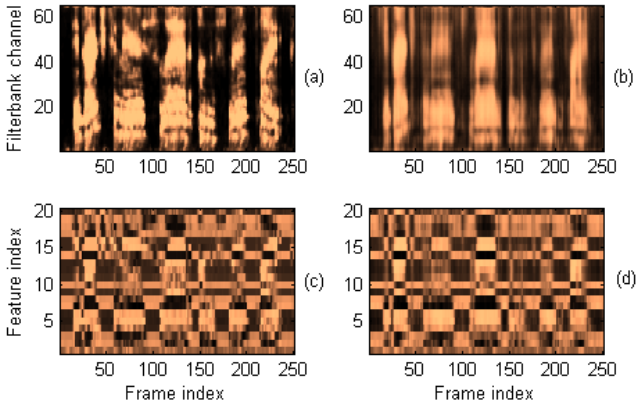

Fig. 3. (a) True IRM and (b) estimated IRM using trained DNN for an utterance. SBAE features extracted from (c) true IRM and (d) estimated IRM.

babble, car, street and train with two SNR levels, namely, *5 dB* and *10 dB*. The noise suppression algorithms fall under four different classes, namely, spectral subtraction, subspace, statistical model-based, and Wiener filter-based algorithms. A complete description of these algorithms can be found in [20]. Subjective evaluation of the speech files was done according to ITU-T Recommendation P.835 [21]. A single-layer artificial neural network (ANN) having *350* units was used as a regression model, as previously used in [16] and [22], to find the mapping between the acoustic feature vector and the corresponding subjective score. For reproducibility and consistency, all networks were initialized with the same random weights. Although a total of 1792 speech files was available in the database, for comparison between objective measure and subjective score, a usual way is to compare per-condition MOS with the per-condition average objective score [21]. The database included utterances enhanced by total 14 algorithms. Hence, 112 total conditions (= 14 algorithms $\times$ 2 SNR levels $\times$ 4 noise types) were available in database with per-condition MOS. Moreover, 240 total noisy utterances (30 utterances $\times$ 4 types of noises $\times$ 2 SNR) available in the database along with enhanced utterances. In order to test the robustness of proposed approach, data was divided into train and test dataset using different partitions. In the first test, 8-fold cross-validation (CV) was used. To evaluate the performance, 3 standard measures, namely, Root Mean Square Error (RMSE), Pearson's linear correlation coefficient (C_p) and Spearman's rank order coefficient (C_s) were used. We also used 3^{rd} order polynomial mapping suggested in [12] to eliminate offset between subjective and objective scores.

B. Quality prediction using true IRM features

To establish the effectiveness of IRM for quality prediction task, we used true IRM extracted using enhanced utterances and corresponding clean utterances in our experiments. We used both FFT-IRM and gammatone-IRM mask for initial experiments. To calculate the T-F representation, the speech signals were divided into frames using 25 ms window with 50 *%* overlap. *20-D* features were extracted using SBAE from true FFT-IRM and gammatone-IRM. For feature extraction from FFT-IRM, SBAE with Mel filterbank was used to incorporate perceptual scale in processing. On the other hand, while extracting features from gammatone-IRM, we used SBAE with linear filterbank since the perceptual scale is incorporated while using gammatone filterbank [23]. The architecture of SBAE which was used to extract features from FFT-IRM was *513-20-256-513* which implies *513* units in input layer, *20* units in subband layer, etc. The architecture of SBAE used to extract features from gammatone-IRM was *64-20-100-64*. The mean and variance of *20-D* IRM features were used as mask features along with *40-D* FBEs. The SBAE was trained using clean and noisy speech utterances only. Enhanced utterances were not used to train the SBAE. More details regarding SBAE training can be found in [16].

Fig. 2 shows results of *8*-fold CV by using true IRM features for quality prediction. As it can be seen from Fig. 2, adding mask information improves the overall performance of quality prediction system. Both mean and variance of true IRM features improve the performance of the system. Moreover, it is worth noting that performance after using IRM information from FFT mask and gammatone mask are almost similar. Hence, we consider using gammatone-IRM for further experiments due to its low dimensionality.

C. Prediction of IRM from noisy gammatone spectrum

To predict the IRM directly from a noisy speech signal, a Deep Neural Network (DNN) with *3* hidden layers was used. The input to the neural network was power *(1/15)* compressed gammatone spectrum of noisy speech [18] with *5*-frames context. *64*-channel gammatone spectrum was considered for experiments. Hence, the input dimension to the DNN was *64×5=320*. *3*-frames context of IRM was predicted simultaneously. Hence, output of the DNN was *64×3=192*-dimensional IRM. The number of units in each hidden layer

ISBN 978-0-9928626-7-1 © EURASIP 2017

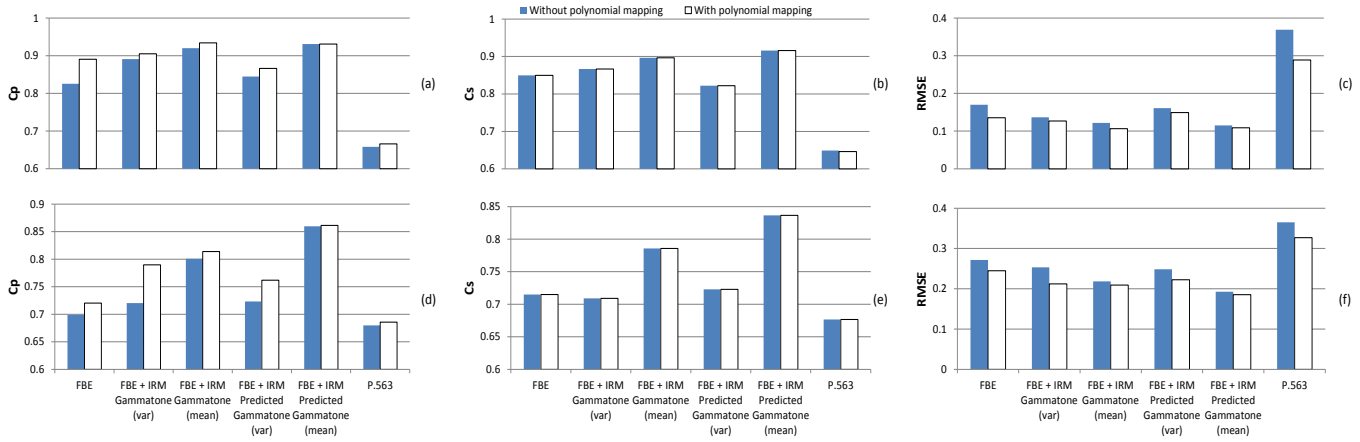

Fig. 5. (a) C_p, (b) C_s and (c) RMSE for *test 1*. (d) C_p, (e) C_s and (f) RMSE for *test 2*.

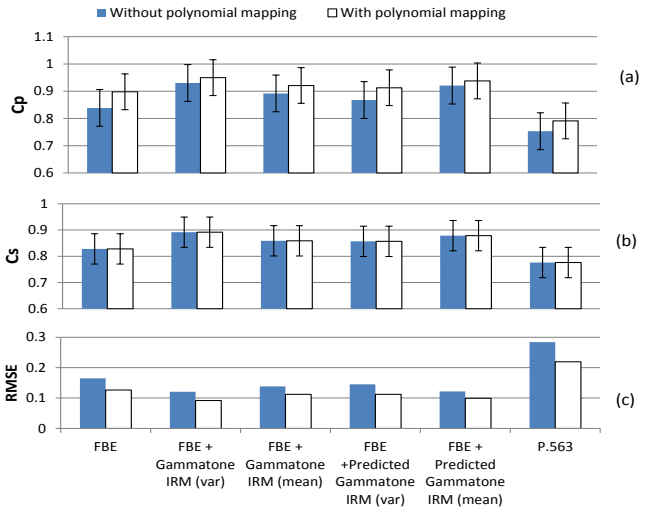

Fig. 4. (a) C_p, (b) C_s and (c) RMSE for *8*-fold CV experiment using predicted IRM features.

was *1024*. The units in hidden layer had rectified linear unit (ReLU) activation, while units in the output layer had sigmoid activation. Back-propagation algorithm with gradient descent optimization was used to train the DNN. The DNN was trained using IRMs of noisy speech utterances only. Enhanced utterances were *not* used while training the DNN. It implies that input of the DNN was a T-F representation of noisy utterance while output of the DNN was IRM calculated using the T-F representation of noisy as well as the clean speech signal. While testing, the DNN trained on noisy speech utterances was used to predict the IRM of enhanced speech utterances.

D. Quality assessment using predicted IRM

Fig. 3 shows original and predicted gammatone-IRM using the trained DNN. Fig. 3 also shows SBAE features extracted from these IRMs. In this Section, predicted IRM is used for non-intrusive quality assessment. The *8*-fold CV experiment was repeated using predicted IRM features. Fig. 4 shows results of *8*-fold CV using predicted IRM features. It is interesting to note that the mean of predicted IRM features gave better performance than the mean of true IRM features. On the other hand, the variance of predicted IRM features gave relatively worse performance than the variance of true IRM features. We believe that reason of this lies in the nature of predicted IRM. Fig. 3 shows that predicted IRM is a *smoother* version of true IRM. This smoothness can be observed in both time and frequency-axis. Hence, the variance of the predicted IRM will not be accurate to the true IRM which leads to performance degradation due to the variance of predicted IRM features. Further detailed analysis of these results is required.

In both the cases, the performance improvement by adding true or predicted IRM information is not statistically significant, due to overlap in 95% confidence interval corresponding to various features. However, the performance improvement is significant when compared to ITU-T P.563 standard, which is state-of-the-art non-intrusive speech quality assessment metric. Hence, all the improvements should be considered over ITU-T P.563 metric.

E. Robustness of IRM features

In *8*-fold CV experiment, the testing conditions were similar to the training conditions. To check the robustness of the proposed approach, two additional tests were performed by dividing utterances into training and testing data according to different mismatched conditions. In *test 1*, utterances having *5* dB SNR were used in training, while utterances having *10* dB SNR were used for testing. In *test 2*, utterances with street and train noise were used in training, while utterances having babble and car noise were used for testing. Fig. 5 shows results of *test 1* and *test 2*. As it can be observed from Fig. 5, mean of predicted IRM features gives a consistent improvement in overall performance for different training and testing conditions, too. Improvement of the performance is more significant in the results of *test 2*.

ISBN 978-0-9928626-7-1 © EURASIP 2017

IV. Summary and Conclusions

In this study, we proposed to use information captured by IRM features along with standard FBEs for non-intrusive quality assessment of noise suppressed speech. First, we showed that IRM information can be effective using true IRM and then IRM was predicted from the speech plus noise mixture using a DNN having *3*-hidden layers. We observed that mean of predicted IRM features gave slightly better performance than the mean of true IRM features. Moreover, we showed that proposed approach is also more robust for different training and testing conditions. IRM features gave the robust performance when noise types in training and testing datasets are different. In future, we plan to study the effectiveness of different IRM prediction techniques for the quality assessment task.

Acknowledgments

The authors would like to thank DeitY, Govt. of India for sponsoring two consortium projects, (1) TTS Phase-II (2) ASR Phase-II and authorities of DA-IICT. The authors are also thankful to NVIDIA for Titan X GPU as a research grant.

References

[1] O. Au and K. Lam, "A novel output-based objective speech quality measure for wireless communication," in *Fourth International Conference on Signal Processing Proceedings (ICSP)*, Beijing, China, 1998, pp. 666–669.

[2] T. H. Falk, Q. Xu, and W.-Y. Chan, "Non-intrusive GMM-based speech quality measurement." in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Philadelphia, Pennsylvania, USA, 2005, pp. 125–128.

[3] G. Chen and V. Parsa, "Bayesian model-based non-intrusive speech quality evaluation," in *IEEE Int. Conf. on Acoustics, Speech, and Sig. Process. (ICASSP)*, Philadelphia, Pennsylvania, USA, 2005, pp. 385–388.

[4] D.-S. Kim, "Anique: An auditory model for single-ended speech quality estimation," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 821–831, 2005.

[5] V. Grancharov, D. Y. Zhao, J. Lindblom, and W. B. Kleijn, "Low-complexity, nonintrusive speech quality assessment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1948–1956, 2006.

[6] T. H. Falk and W.-Y. Chan, "Single-ended speech quality measurement using machine learning methods," *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 14, no. 6, pp. 1935–1947, 2006.

[7] M. Narwaria, W. Lin, I. V. McLoughlin, S. Emmanuel, and C. L. Tien, "Non-intrusive speech quality assessment with support vector regression," in *Advances in Multimedia Modeling*, 2010, pp. 325–335.

[8] M. Narwaria, W. Lin, I. V. McLoughlin, S. Emmanuel, and L.-T. Chia, "Nonintrusive quality assessment of noise suppressed speech with mel-filtered energies and support vector regression," *IEEE Trans. on Audio, Speech, and Lang. Process.* , vol. 20, no. 4, pp. 1217–1232, 2012.

[9] Q. Li, W. Lin, Y. Fang, and D. Thalmann, "Bag-of-words representation for non-intrusive speech quality assessment," in *2015 IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP)*, Chengdu, China, 2015, pp. 616–619.

[10] Q. Li, Y. Fang, W. Lin, and D. Thalmann, "Non-intrusive quality assessment for enhanced speech signals based on spectro-temporal features," in *2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, 2014, pp. 1–6.

[11] R. K. Dubey and A. Kumar, "Non-intrusive speech quality assessment using several combinations of auditory features," *International Journal of Speech Technology (IJST)*, vol. 16, no. 1, pp. 89–101, 2013.

[12] T. Falk and W. Chan, "Single-ended method for objective speech quality assessment in narrowband telephony applications," *ITU-T*, 2004.

[13] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*. Springer, 2005, pp. 181–197.

[14] S. Srinivasan, N. Roman, and D. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Communication*, vol. 48, no. 11, pp. 1486–1501, 2006.

[15] 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.

[16] M. H. Soni and H. A. Patil, "Novel subband autoencoder features for non-intrusive quality assessment of noise suppressed speech," in *INTERSPEECH, San Francisco, USA*, 2016, pp. 3708–3712.

[17] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, pp. 7092–7096.

[18] J. Chen, Y. Wang, S. E. Yoho, D. Wang, and E. W. Healy, "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises," *The Journal of the Acoustical Society of America (JASA)*, vol. 139, no. 5, pp. 2604–2612, 2016.

[19] M. H. Soni, T. B. Patel, and H. A. Patil, "Novel subband autoencoder features for detection of spoofed speech," in *INTERSPEECH, San Francisco, USA*, 2016, pp. 1820–1824.

[20] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Communication*, vol. 49, no. 7, pp. 588–601, 2007.

[21] ITU-T, "ITU-T Rec 835, subjective test methodology for evaluating speech communication systems that include noise suppression algorithm," http://www.itu.int/rec/T-REC-P.835-200311-I, 2003, {Last Accessed: 30^{th} March, 2016}.

[22] M. H. Soni and H. A. Patil, "Novel deep autoencoder features for non-intrusive speech quality assessment," in *European Signal Processing Conference (EUSIPCO)*, 2016, pp. 2315–2319.

[23] R. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," *Paper presented at a meeting of the IOC Speech Group on Auditory Modelling*, RSRE, England, 1987.