Novel Deep Autoencoder Features for Non-intrusive Speech Quality Assessment

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

***Abstract*—To emulate the human perception in quality assessment, an objective metric or assessment method is required, which is a challenging task. Moreover, assessing the quality of speech without any reference or the ground truth is altogether more difficult. In this paper, we propose a new non-intrusive speech quality assessment metric for objective evaluation of speech quality. The originality of proposed scheme lies in using deep autoencoder to extract low-dimensional features from a spectrum of the speech signal and finds a mapping between features and subjective scores using an artificial neural network (ANN). We have shown that autoencoder features capture noise information in a better way than state-of-the-art Filterbank Energies (FBEs). Quantification of our experimental results suggests that proposed metric gives more accurate and correlated scores than an existing benchmark for objective, non-intrusive quality assessment metric ITU-T P.563 standard.**

I. Introduction

Speech quality assessment is very important in many applications including telephone networks, voice over Internet, multimedia applications, etc. The best way to assess the quality of speech is to take the opinion of the human listeners. To do so, listening tests are conducted which serves as a subjective quality assessment measure. However, some fundamental difficulties, including cost, time consumption and in some cases, the reliability of the test, make subjective tests unsuitable for many applications which require in-service, real-time or in-process quality assessment. Hence, to overcome these limitations, there is a requirement of a reliable objective measure to assess the speech quality. Objective speech quality assessment has attracted researchers over the past years [1]–[8].

The aim of objective quality evaluation is to find a replacement for human judgment of perceived speech quality. Objective evaluation techniques are less complex, less expensive in terms of resources and time and give more consistent results. Objective evaluation techniques are categorized in two ways, namely, intrusive and non-intrusive. Intrusive assessments are based on waveform comparison wherein reference speech signal is available for comparison. On the other hand, non intrusive quality assessment (also known as single-ended, no-reference or output-based quality assessment) is performed using single speech waveform, without any reference or the ground truth. Intrusive methods are more straightforward, less complex and more accurate than non-intrusive ones. However, in many practical scenarios such as wireless communication, voice over IP (VoIP) and other in-service applications requiring monitoring of speech quality, intrusive methods cannot be applied due to unavailability of reference speech signal. In such cases, it is necessary to have a reliable non-intrusive method for quality assessment. Excellent summary of the principles of existing quality estimation models, their advantages, limitations and future directions is given in [9].

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 model 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 perceptually motivated speech quality assessment algorithm based on temporal envelope representation of speech is presented. A low-complexity, non-intrusive speech quality assessment method based on commonly used speech coding parameters such as spectral dynamics is presented in [5]. Different features extracted from speech have been detected to be useful for speech quality assessment. Spectral dynamics, spectral flatness, spectral centroid, variance, pitch and excitation variance was used for quality prediction in [5]. The authors in [10] used perceptual linear prediction (PLP) coefficients for quality assessment. A method for speech quality assessment using temporal envelope representation of speech was proposed in [4]. A non-intrusive algorithm for quality assessment in VoIP is presented and studied in [11]. In [12], authors posed quality estimation as a regression problem and used average Mel Frequency Cepstral Coefficients (MFCCs) to find mapping to subjective scores using support vector regression (SVR). Similarly, [13] examined use of mean and variance of filterbank energies for mapping using SVR in a similar fashion.

Recently, deep learning methods are gaining popularity for feature extraction from raw speech data. Autoencoder is such network which uses Deep Neural Network (DNN) or Restricted Boltzmann Machine (RBM) to extract low-dimensional information from high-dimensional raw data [14]–[17]. Autoencoder has been widely used for automatic speech recognition (ASR) systems for noisy or reverberant conditions. In [18] and [19], authors used autoencoder as de-noising front-end for such ASR task. Autoencoder was used to find the mapping between noisy speech spectrum and clean speech spectrum in [20] for noise reduction in ASR

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

system. Autoencoder features were also used for speech enhancement application in [21]. For speech coding, autoencoder was used to encode speech spectrum in [22]. The popularity of autoencoder features in denoising of speech spectrum and features suggests that they are able to capture information about presence or absence of noise in the speech signal. Moreover, in [22] it is shown that speech spectrum can be reconstructed using features learned by an autoencoder. This ability of autoencoder features makes them suitable for quality assessment since they are able to capture underlying spectral information. In this paper, the problem of speech quality assessment is posed as a regression problem, same as previously done in [12] and [13]. However, we have used autoencoder features as the acoustic features and used an artificial neural network (ANN) as a regression model. ANN was chosen due to its universal approximation abilities and need of least tuning of parameters. We have shown that autoencoder features provide more variability in feature vectors for speech files having different types and amount of noise. Moreover, they are able to reconstruct speech spectrum more precisely than filterbank energies. These properties of autoencoder features suggest that they capture noise information in a better way than MFCC or filterbank energies.

II. Autoencoder

A. Basic Autoencoder

Autoencoder is an artificial neural network (ANN) with a bottleneck structure in hidden layers. A basic autoencoder structure is shown in figure 1. It is used to learn a low-dimensional representation of input data which is originally of high-dimension. Autoencoder consists of two blocks, namely, encoder and decoder. The encoding part will represent the high-dimensional data into low-dimension while the decoder will convert that low-dimensional representation into a high-dimensional output feature. Mathematically, encoding operation can be represented as follows:

$$y = f_\theta = s(\mathbf{W}x + \mathbf{b}), \tag{1}$$

where y is the low-dimensional feature vector representation and $\theta = \{\mathbf{W}, \mathbf{b}\}$. $\mathbf{W}$ and $\mathbf{b}$ represents weights and biases of network, respectively. s is a nonlinear activation function. At the decoding stage, the low-dimensional representation y is mapped back to high-dimensional representation z using the following formula:

$$z = g_{\theta'} = s(\mathbf{W}'y + \mathbf{b}'), \tag{2}$$

where $\theta' = \{\mathbf{W}', \mathbf{b}'\}$. Hence, the output of network can be seen as function of $\{\theta\ ,\ \theta'\}$, i.e., $z = g_{\theta'}(f_\theta(x))$. These parameters are optimized such that output of the network z is as close as possible to input x and maximizes $P(x|z)$. Optimization is done using minimization of mean square error (MSE) between target x and network output z. Autoencoder can be made deep by inserting more encoding and and decoding layers. More details regarding training of an deep autoencoder can be found in [17].

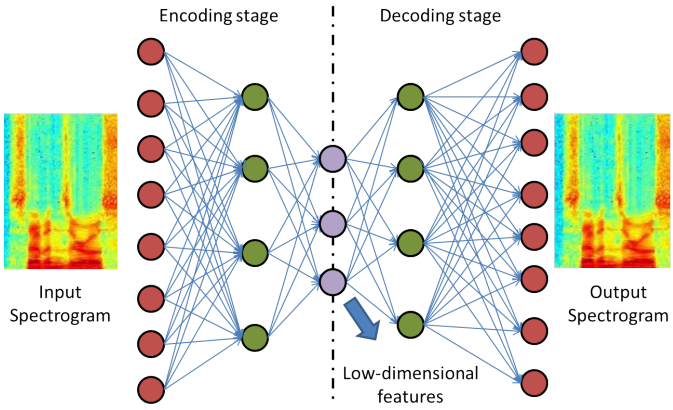

Fig. 1. General architecture of an autoencoder.

III. Analysis of Autoencoder Features Under the Influence of Noise

A. Analysis in terms of feature variability

Figure 2 shows autoencoder features and Mel filterbank energies for clean and noisy speech utterance along with the corresponding spectrum. It can be observed that both features get affected under the influence of noise present in speech signal. It can be seen from figure 2 (d) that the noise affects almost all the frequency components of the spectrum. While filterbank energies are able to capture the noise present in speech spectrum, each coefficient of filterbank energies captures information about noise present in one particular band of frequencies. On the other hand, it can be seen from figure 1 that each autoencoder feature captures information or variation in all the frequency regions. This happens due to the fact that every unit in subsequent layers of autoencoder is connected with all units of the previous layer. Hence, units in subsequent layers are forced to capture the information of all units in the previous layer. Thus, almost all autoencoder features get affected due to the noise present in all frequency components. This property makes them more suitable for quality assessment under the influence of noise.

Figure 3 shows mean features for clean and noisy speech with different amount of noise. The only effect of noise on filterbank energies is their average value increase with increasing amount of noise in speech. However, each autoencoder feature gets affected differently in the presence of different amount of noise. Hence, the autoencoder feature vector for the different amount of noise will be quite different in shape and value. Figure 4 shows average autoencoder features and filterbank energies having different Mean Opinion Scores (MOS). The features are extracted for the same utterance enhanced by different enhancement algorithms. Figure 4 (b) shows that if there is a large difference in MOS values then filterbank energies corresponding to that speech signals will have more difference in mean values. However, if the difference between MOS of speech files having different conditions is small, then the difference between their corresponding filterbank energies

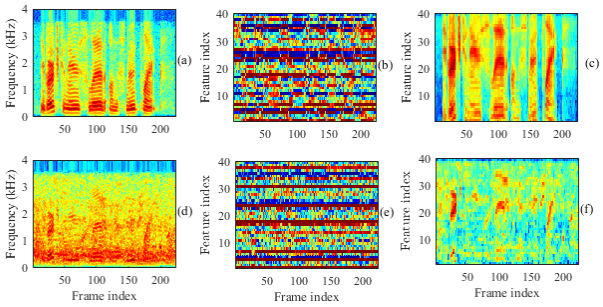

Fig. 2. *P*lots of (a) clean speech spectrogram, (b) its autoencoder features and (c) filterbank energies. similar plots for noisy speech having car noise of 5 dB SNR are shown in (d), (e) and (f).

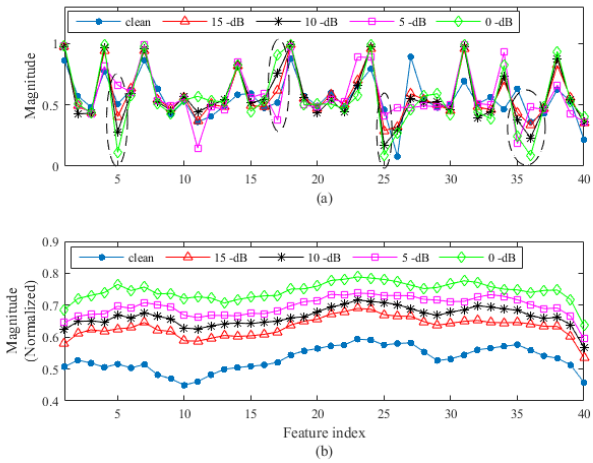

Fig. 3. *A*verage (a) autoencoder features and (b) filterbank energies of 20 speech files with different amount of babble noise. Dotted circles indicate features which vary more under the influence of additive noise.

is small and ambiguous. On the other hand, autoencoder features show different behavior than filterbank energies. Many of the autoencoder features show little difference for same utterance with different perceptual quality. However, finer details about the perceptual quality of the speech signal is reflected in some of the autoencoder features. These features are shown with dotted circles in figure 4 (a). The clue about the perceptual quality can be found in these features.

B. Analysis in terms of reconstruction ability

Figure 5 shows the spectrum of clean as well as noisy speech, and reconstructed spectrum using filterbank energies and autoencoder features. The spectrum using filterbank energies was reconstructed using method shown in [23]. By observing reconstructed spectrum, it is evident that autoencoder features are able to reconstruct both clean and noisy speech spectrum in a better way than filterbank energies. To support this observation, Log Spectral Distortion (LSD) between original spectrum and reconstructed spectrum is calculated for clean as well as noisy speech signals. The average LSD for *30* clean and noisy utterances was taken.

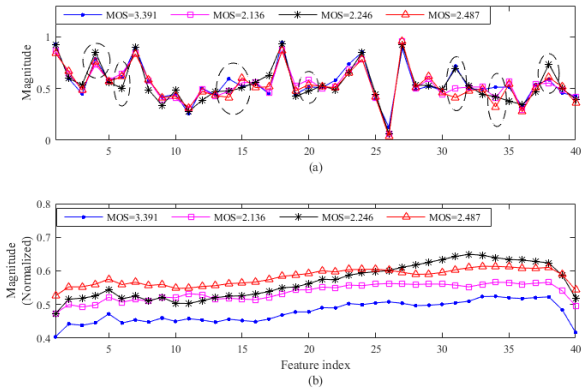

Fig. 4. *A*verage (a) autoencoder features and (b) filterbank energies of 20 speech files with different amount of babble noise. The dotted circles show autoencoder features that vary more for different perceptual quality.

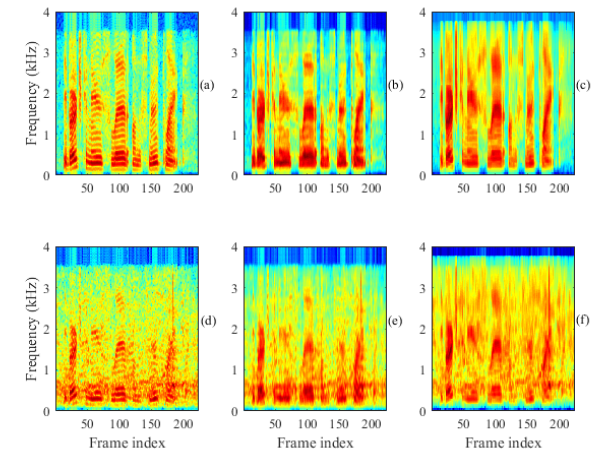

Fig. 5. (a) Speech spectrogram and reconstructed spectrogram using (b) autoencoder features and (c) filterbank energies for clean speech. Similar plots in (d), (e) and (f) for noisy speech having babble noise of 15 dB SNR. Dotted circles represent high frequency regions.

The LSD in case of clean speech utterances was *1.2318* and *3.3918* dB using autoencoder features and filterbank energies, respectively. In case of noisy speech with *15* dB SNR (additive babble noise) the LSD was *1.021* and *3.3614* dB for autoencoder features and filterbank energies, respectively. The effect of better reconstruction is more evident in high-frequency regions. Autoencoder features capture high-frequency variations more precisely than filterbank energies. This is due to the fact that at high-frequency, the bandwidth of Mel-filterbank is high. Hence, higher frequency filterbanks will contain average information of a wideband, which is not in the case of autoencoder features. As discussed in the previous Section, each autoencoder feature contains information about all frequency components. Hence, they are able to reconstruct speech spectrum in a much better way. This fact suggests that autoencoder features capture more information about speech spectrum than filterbank energies.

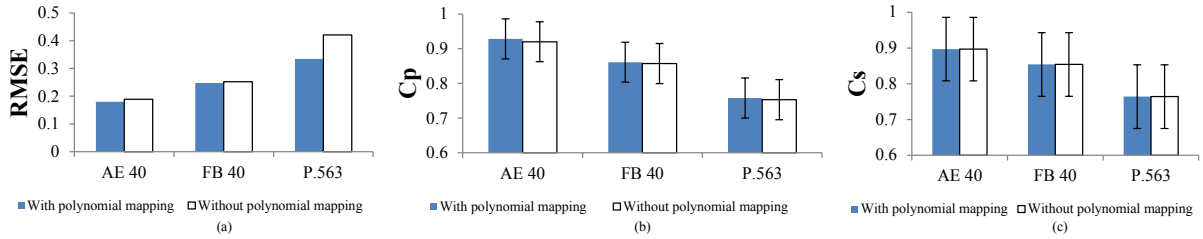

Fig. 6. (a) RMSE, (b) C_P and (c) C_s between predicted scores and actual subjective scores using proposed method, filterbank energies and ITU-T P.563 standard for test 1.

IV. Experiments and Results

A. Experimental setup

All experiments were performed on NOIZEUS database [24]. The database has speech files which were corrupted by different kind (and amount) of noise. It also had speech files which were enhanced by different noise suppression algorithms. The speech files were corrupted by four types of noise, namely, babble, car, street and train with two SNR levels *5* and *10dB*. The noise suppression algorithms fall under four different classes, namely, spectral subtraction, subspace, statistical model-based, and Wiener algorithms. A complete description of these algorithms can be found in [6], [24]. Subjective evaluation of the speech files was performed according to ITU-T Recommendation P.835 [6], [25]. Both autoencoder features and filterbank energies were extracted from this database. The performance comparison of the both features was done for *40-D* (dimensional) features. To extract autoencoder features from FFT spectrum, deep autoencoder was used. The architecture of deep autoencoder was *513-250-40-250-513*, meaning *513* units in the first layer, *250* units in the second layer and so on. All units had sigmoid as nonlinearity in all the layers. To demonstrate the ability of autoencoder features to capture general spectral information, autoencoder was trained only using *150* files which were not used for further experiments. Mel filterbank energies of same dimensions were extracted from speech files.

To find the mapping between features extracted from speech and their subjective score, artificial neural network (ANN) with single hidden layer was used. The total number of hidden units in ANN was *350* which was selected using validation data. The network was regularized using standard weight decay method to prevent over-fitting due to small database size. Although the 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 [25]. By including noisy speech files and their enhanced versions using *13* different algorithms, total *14* algorithms were available. Hence, total *112* conditions (= *14* algorithms $\times$ *2* SNR levels $\times$ *4* noise types) were available in database with per-condition MOS. In order to test the robustness of proposed approach against the data-dependency, data was divided into training and testing dataset using different partitions. In total, we evaluated the performance of proposed metric under *3* different test conditions. In the first test, *8*-fold cross-validation was used. Data was divided into *8* parts out of which *7* parts were used for training and *1* part for testing. Experiments were repeated till all *8* parts were used for testing. Results of test *1* are shown in figure 6. In second test, data was partitioned according to different types of noise added. Speech files for *3* noise types were kept for training and *1* noise type was used for testing. Experiments were repeated till all noise conditions were used for testing. In *test 3*, we divided the data according to noise suppression algorithm. Similar experiments were done as *test 2* in this case. Tabel I shows the results for *test 2* and *test 3*.

B. Results and discussions

To evaluate the performance, three common criteria was used: Pearson linear correlation coefficient C_p (for prediction accuracy), Spearman rank order correlation coefficient C_s (for prediction monotonicity) and Root Mean Squared Error (RMSE) between predicted objective score and subjective scores [6]. For an ideal match between the objective and subjective scores, C_p=C_s = 1 and RMSE= 0. Moreover, it is suggested in [7] that to eliminate the offset and non-linearity between objective scores and subjective score, it is advisable to use 3^{rd} order polynomial mapping between objective and subjective scores. All results are shown with and without polynomial mapping.

Figure 6 shows RMSE, C_p and C_s calculated for both autoencoder features and filterbank energies for *test 1*. We also compared our results with ITU P.563 standard [7], which is a standard objective measure for non-intrusive speech quality assessment. C_p and C_s are shown with *95 %* confidence intervals. Figure 6 clearly suggests that autoencoder features give more powerful mapping than filterbank energies with identical experimental conditions. Objective scores predicted using autoencoder features are more accurate as well as more correlated with actual subjective scores. These results are in coherence with our analysis of autoencoder features. Moreover, the overlap between *95 %* confidence intervals of C_p and C_s in case of P.563 and proposed method is very less. For C_p, it is zero. Hence, it can be said that proposed method gives objective scores which are nearer to actual subjective scores than state-of-the-art P.563 scores. Table I shows RMSE, C_p and C_s calculated for *test 2* and *3*. It can be said that

TABLE I
RESULTS FOR PROPOSED QUALITY MEASURE ALONG WITH SCORE USING FILTERBANK ENERGIES AND P.563 SCORE. DATA IN THIS CASE WAS PARTED ACCORDING TO TEST 2 AND 3. AVERAGE SCORES FOR DIFFERENT CONDITIONS ARE SHOWN

Test 2	Without mapping			With mapping			Test 3	Without mapping			With mapping		
Method	$RMSE$	C_p	C_s	$RMSE$	C_p	C_s	Method	$RMSE$	C_p	C_s	$RMSE$	C_p	C_s
AE 40	**0.200**	**0.884**	**0.867**	**0.194**	**0.886**	**0.867**	AE 40	**0.252**	**0.836**	**0.836**	**0.248**	**0.842**	**0.840**
FB 40	0.234	0.801	0.809	0.231	0.818	0.809	FB 40	0.281	0.768	0.761	0.278	0.772	0.761
P.563	0.374	0.721	0.732	0.330	0.736	0.732	P.563	0.374	0.726	0.740	0.350	0.738	0.740

for different test conditions, too, proposed metric gives more accurate and correlated objective scores. It is worth noting that scores predicted using both MFCC and autoencoder features perform better if data of all the test conditions are used for training. Performance using both the feature degrades if the test conditions are not involved in training. This effect is evident from results shown in Table I.

V. SUMMARY AND CONCLUSIONS

In this paper, we have proposed a new non-intrusive speech quality prediction system using autoencoder features and ANN. It uses autoencoder features as the acoustic features and ANN to find an optimal mapping between the features and subjective score. Hence, it poses speech quality prediction task as a *regression* problem. We have compared proposed approach with Mel filterbank energies as acoustic features as well as with ITU P.563 standard. Quantification of our experimental results suggests that proposed method has more prediction accuracy and more correlation with subjective scores than the existing systems. Future work includes using this method with various kinds of noise such as noise in the communication channel and testing the performance using proposed method.

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'98*, 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 International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Philadelphia, USA, 2005, pp. 125–128.

[3] G. Chen and V. Parsa, "Bayesian model based non-intrusive speech quality evaluation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Philadelphia, 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] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008.

[7] ITU-T P.563, "Single ended method for objective speech quality assessment in narrowband telephony applications," International Telecom Union (ITU), 2004.

[8] A. Hines, J. Skoglund, A. C. Kokaram, and N. Harte, "ViSQOL: an objective speech quality model," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, pp. 1–18, 2015.

[9] S. Möller, W.-Y. Chan, N. Côté, T. H. Falk, A. Raake, and M. Wältermann, "Speech quality estimation: Models and trends," *IEEE Signal Processing Magazine,*, vol. 28, no. 6, pp. 18–28, 2011.

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

[11] L. Ding, Z. Lin, A. Radwan, M. S. El-Hennawey, and R. A. Goubran, "Non-intrusive single-ended speech quality assessment in VoIP," *Speech Communication*, vol. 49, no. 6, pp. 477–489, 2007.

[12] 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.

[13] 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 Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1217–1232, 2012.

[14] J. Gehring, Y. Miao, F. Metze, and A. Waibel, "Extracting deep bottleneck features using stacked auto-encoders," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, pp. 3377–3381.

[15] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[16] T. N. Sainath, B. Kingsbury, and B. Ramabhadran, "Auto-encoder bottleneck features using deep belief networks," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, 2012, pp. 4153–4156.

[17] D. Yu and M. L. Seltzer, "Improved Bottleneck Features Using Pretrained Deep Neural Networks," in *INTERSPEECH*, Florence, Italy, 2011, pp. 237–240.

[18] T. Ishii, H. Komiyama, T. Shinozaki, Y. Horiuchi, and S. Kuroiwa, "Reverberant speech recognition based on denoising autoencoder." in *INTERSPEECH*, Lyon, France, 2013, pp. 3512–3516.

[19] X. Feng, Y. Zhang, and J. Glass, "Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014, pp. 1759–1763.

[20] A. L. Maas, Q. V. Le, T. M. O'Neil, O. Vinyals, P. Nguyen, and A. Y. Ng, "Recurrent Neural Networks for Noise Reduction in Robust ASR," in *INTERSPEECH*, Portland, USA, 2012, pp. 22–25.

[21] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder." in *INTERSPEECH*, Lyon, France, 2013, pp. 436–440.

[22] L. Deng, M. L. Seltzer, D. Yu, A. Acero, A.-R. Mohamed, and G. E. Hinton, "Binary coding of speech spectrograms using a deep auto-encoder." in *INTERSPEECH*, Makuhari, Japan, 2010, pp. 1692–1695.

[23] L. E. Boucheron and P. L. De Leon, "On the inversion of mel-frequency cepstral coefficients for speech enhancement applications," in *International Conference on Signals and Electronic Systems (ICSES)*, Krakov, Poland, 2008, pp. 485–488.

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

[25] ITU-T P.835, "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm," International Telecom Union (ITU), 2003.