9th ISCA Speech Synthesis Workshop
13-15 Sep 2016, Sunnyvale, USA

Non-intrusive Quality Assessment of Synthesized Speech using Spectral Features and Support Vector Regression

Meet H. Soni, Hemant A. Patil

Dhirubhai Ambani Institute of Information and Communication Technology, India

{meet_soni, hemant_patil}@daiict.ac.in

Abstract

In this paper, we propose a new quality assessment method for synthesized speech. Unlike previous approaches which uses Hidden Markov Model (HMM) trained on natural utterances as a reference model to predict the quality of synthesized speech, proposed approach uses knowledge about synthesized speech while training the model. The previous approach has been successfully applied in the quality assessment of synthesized speech for the German language. However, it gave poor results for English language databases such as Blizzard Challenge 2008 and 2009 databases. The problem of quality assessment of synthesized speech is posed as a *regression* problem. The mapping between statistical properties of spectral features extracted from the speech signal and corresponding speech quality score (MOS) was found using Support Vector Regression (SVR). All the experiments were done on Blizzard Challenge Databases of the year 2008, 2009, 2010 and 2012. The results of experiments show that by including knowledge about synthesized speech while training, the performance of quality assessment system can be improved. Moreover, the accuracy of quality assessment system heavily depends on the kind of synthesis system used for signal generation. On Blizzard 2008 and 2009 database, proposed approach gives correlation of *0.28* and *0.49*, respectively, for about *17 %* data used in training. Previous approach gives correlation of *0.3* and *0.09*, respectively, using spectral features. For Blizzard 2012 database, proposed approach gives correlation of *0.8* by using *12 %* of available data in training.

Index Terms Quality assessment, autoencoder, subband.

1. Introduction

Text-to-Speech (TTS) systems have seen a drastic improvement over the decades which enables their use in everyday applications, which includes spoken dialogue applications. TTS systems are useful in navigation, smart home assistance, human-computer interface, traffic information systems, email and SMS reading systems, etc. Many of these applications require generalized TTS system trained on a large database to handle large vocabulary and to provide flexibility. With many speech synthesis systems to choose from available state-of-the-art techniques, it is crucial to choose appropriate system for the provider according to the application. This decision requires comparison of perceived quality of generated speech from many systems. Moreover, it is also important to measure the quality of synthesized speech while developing or optimizing a new or existing algorithm for speech synthesis. Thus, quality assessment of synthesized speech is necessary for many applications.

So far, the evaluation of TTS systems is heavily relied on subjective tests, in which human participants listen to the speech utterances and give their judgment about a particular aspect of speech quality. Different types of tests are recommended for evaluation of different aspects of quality of speech generated by TTS systems [1–3]. The aspects of speech that has to be judged while evaluating the quality of synthesized speech include naturalness, intelligibility, pronunciation, intonation, prosody, voice pleasantness, rhythm, etc. Although the subjective tests are questioned about their reliability, they are most frequently used for judging the overall quality of synthesized speech [4], [5]. It is also recommended by International Telecommunication Union (ITU) for evaluating telephone-based services [3]. However, all of these methods require the human listener to judge the quality of speech which makes these methods time consuming and expensive. Hence, there is a need of objective measure which can replace human listener and gives accurate results about the underlying quality of a speech utterance.

In order to come up with the objective measure, several methods have been proposed to assess the quality of synthesized speech in an instrumental way. One way of objective measurement is to compare synthesized utterance with the natural utterance having same linguistic content. Method described in [6] uses perceptually weighted distance between natural and synthetic utterance to estimate the degradation in the speech signal. This approach is similar to the methods described in [7] and [8], which uses pattern comparison for predicting the quality of transmitted speech. In [9], authors proposed to use perceptual features which are extracted from synthesized speech utterance to compare with the perceptual features of natural utterance. The distance between both of these features may be used as the quality measure. In [10], the authors used a combination of magnitude and phase-based features for objective evaluation of TTS system. More details about the intrusive quality assessment of TTS systems is given in [11]. The first non-intrusive objective quality assessment system was proposed in [12], which uses Hidden Markov Model (HMM) trained on natural utterances as the reference and it is used to predict the quality of synthetic utterances. To train the HMM, perceptual features such as Mel Frequency Cepstral Coefficients (MFCCs) and fundamental frequency (F_0) is used. It is assumed that the temporal variations in natural speech will be quite different than that of synthesized speech. Hence, HMM trained on only natural utterances will give lower likelihood score if synthesized utterance is given to it. This likelihood is taken as the quality measure. This approach gave promising results on German database with the correlation between *0.54* to *0.81* for different quality dimensions collected in the auditory test. These results were per-stimulus basis. This approach was compared with ITU-T P.563 [13], [14], which is standard non-intrusive metric for quality assessment in narrowband telephony applications. Re-

http://dx.doi.org/10.21437/SSW.2016-21

sults reported in [15] suggest that P.563 is not as effective metric as HMM-based approach for predicting the quality of synthesized speech. Studies reported in [16] and [17] leads to conclusions that P.563 is not the most appropriate model for quality assessment of synthesized speech. Moreover, the performance is considerably different in the case of male and female speech. Comparison of the performance of HMM-based quality assessment system using different features was done in [18]. In this approach, authors used spectral parameters such as MFCC and F_0, parameters given in ITU-T P.563 [14] to model degradation in speech and some general speech parameters given in [19] for quality assessment. Linear combination of log-likelihood generated by HMM trained on natural utterances, P.563 parameters and general parameters was used to predict the quality of synthesized utterance. This approach worked very well on German database [18] in predicting quality for both per-stimulus and per-synthesizer cases. For male utterances, the maximum correlation of *0.9* was got for per-stimulus bases using the combination of features. However, the same approach failed on popular English databases [20] given by Blizzard Challenge organizers, in predicting the quality of utterances on per-stimulus bases. The maximum correlation reported on Blizzard Challenge 2008 [21] database is *0.3* for per-stimulus-based quality assessment. While in case for Blizzard Challenge 2009 [22] database, the maximum correlation reported in [20] is *0.49* for per-stimulus cases. However, the results for per-synthesizer based quality assessment were better in comparison to per-stimulus based ones.

This non-uniformity suggests that current methods for quality evaluation of synthesized speech are not generalized and there is a need of better approach towards the problem. In this paper, we propose a new quality assessment system for synthesized speech which uses information of synthesized utterances while training the system. Hence, the model will have prior knowledge about synthesized utterances and may lead to better prediction of quality. The proposed approach is significantly different than the one proposed in [18] since the later one uses information of only natural utterances while training the model which acts as the reference model for synthesized speech. In proposed approach, the spectral features from Short-Time Fourier Transform (STFT) power spectra was extracted and statistical properties of these features such as mean and variance were used to find a mapping between feature vectors and Mean Opinion Score (MOS) of the corresponding utterance. A similar approach has been used successfully for quality prediction of noise suppressed speech in [23], [24]. The mapping between features and MOS was established using Support Vector Regression (SVR). Two spectral features, namely, Mel filterbank energies and Subband Autoencoder (SBAE) features were considered to find the mapping. SBAE features were proposed for the quality assessment task and they have shown more effectiveness in predicting quality of noise suppressed speech than mel-filterbank energies [25]. Extensive experiments have been conducted on four Blizzard Challenge databases from year 2008, 2009, 2010 and 2012 [21], [22], [26], [27].

The rest of the paper is organized as follows. Section 2 gives information about the spectral features used in proposed approach. Analysis of the spectral features for different synthesized utterances and details about proposed approach is given in Section 3. Section 4 includes information about experimental setup, details of experiments and results of those experiments. Section 5 gives the summary and concluding remarks along with future research directions.

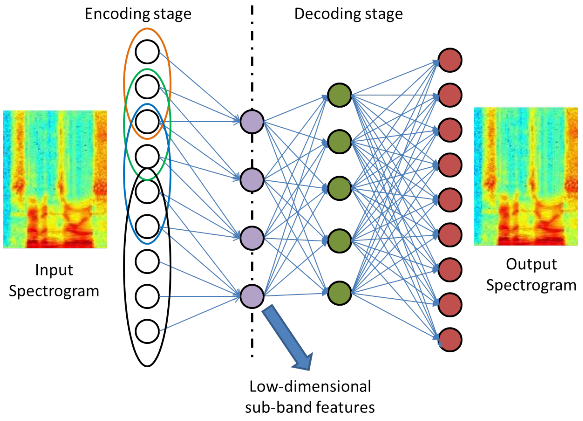

Figure 1: *Architecture of subband autoencoder.*

2. Spectral Features

2.1. Filterbank energies

To extract the spectral features from speech utterances, first of all, preprocessing steps were conducted. All utterances were normalized to have zero mean and unit variance. Simple Voice Activity Detection (VAD) algorithm was used to remove silence intervals longer than *75* ms, same as [20]. *40*-dimensional (i.e., *40*-D) filterbank energies were extracted from Short-Time Fourier Transform (STFT) power spectra. The framing of speech signal was done using the Hamming window of *25* ms duration with *50* % overlap.

2.2. Subband autoencoder features

Fig. 1 shows the architecture of subband autoencoder (SBAE). The main difference between the architecture of SBAE and architecture of an autoencoder is the connectivity of neurons or units immediately after the input layer [28]. In autoencoder, each unit in the layer immediately after input layer is connected with all the units of the previous layer. While in the case of SBAE, the connectivity is restricted. In SBAE, each unit of the first hidden layer is connected with a particular frequency band of the input spectrum. Hence, each unit in the first layer will encode the information about that particular frequency band only, with which it is connected. The decoding structure is same as a general autoencoder with full connectivity. The band structure of *restricted* connectivity for neurons is same as Mel filterbank, implying one neuron in the first layer is connected with the frequencies of one Mel filterbank. This architecture is nearer to HAS and provides more meaningful information than autoencoder in the case of the speech signal. Mathematically, operation of the subband layer can be represented as follows:

$$a_j = f(\sum_i W_{ij}^1 \times x_j), \qquad (1)$$

where a_j is j^{th} subband feature, x_j is short-time power corresponding to j^{th} filterbank frequencies and W_{ij}^1 are weights corresponding to j^{th} subband feature. f represents nonlinear activation function of the neuron. The functionality of preceding layers of subband autoencoder is same as of a simple autoencoder [28]. Proposed architecture can be trained by backpropagation similarly as an autoencoder. a_j learned by SBAE can be used as low-dimensional features for speech technology task. These features are different from filterbank energies in following ways: First, difference in the method of extracting

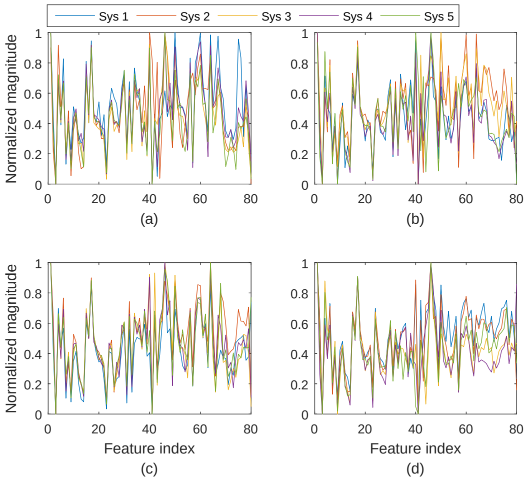

Figure 3: (a) Mean and variance of SBAE features for an utterance from different systems of Blizzard 2008 data for same text spoken. Similarly, (b), (c) and (d) are mean and variance of SBAE features from Blizzard 2009, 2010 and 2012 database, respectively.

features. In particular, MFCCs or filterbank energies are *hand-crafted* features while SBAE features are learned by machine learning approach. Second filterbank energies are extracted in a linear way, while SBAE features are extracted in a nonlinear manner. The latter property may provide some more useful information about speech spectrum variations for different conditions.

3. Analysis of Spectral Features and Proposed Approach

Figure 2 shows spectrogram, SBAE features and filterbank energies for three different utterances synthesized by different systems [system B, C and E] of Blizzard Challenge 2009 database for the same text. As it can be observed, for same text the behavior of features is different for speech synthesized by different systems and algorithms. The characteristics of synthesized speech such as naturalness, intelligibility and other properties can be captured by these features. Hence, they can be used in some manner to assess the quality of the synthesized speech. The question is, how to harvest the information from these features which can lead to the perceptual quality of the corresponding utterance. In this work, statistical properties of the features such as mean and variance are used for quality assessment.

Figure 3 shows mean and variance of SBAE features for utterances from different databases. For one Blizzard database, the text of utterances was same. However, due to the generation of speech using different synthesis systems, the statistical properties of spectral features of each utterance was different. In figure 3, first *40* features are mean and last *40* features indicates variance of the SBAE features. As it can be seen, the mean and variance for utterances generated by different speech synthesis systems were different for same text spoken. The MOS of these utterances were different, indicating their perceptual qualities were different than each other. Hence, it can be assumed that the perceptual quality of an utterance can be related with the statistical properties of the utterance in some manner. We assume that this relation between mean and variance of utterance and perceptual quality is very complex and a nonlinear model is required to find this relation. We propose to use SVR with Gaussian or Radial Basis Function (RBF) kernel to find the mapping between the features and MOS of an utterance. Although other models are available, SVR is used due to its higher generalization capability. Figure 4 shows mean and variance of filterbank energies for same data utterances shown in figure 3. Statistical properties of both features show variation for speech utterances synthesized by different systems for the same text. Hence, both features can be used for quality assessment of synthesized speech. Comparison of the performance and effectiveness of both features has been done in the following Sections. Both figure 3 and figure 4 suggests that variance of the spectral features shows more variation and gives more information about synthesis method. However, our experiments suggested that by including the mean of the features improved the overall performance. Overall working of proposed approach is given in figure 5. First of all, STFT is calculated from given synthesized utterance and spectral features are extracted from STFT power spectrum. During the training phase, the model (i.e., SVR in this case) is trained using labeled training data. Input to the model is mean and variance of spectral features and target is MOS of the input utterance. Once the model is trained, it can be used to predict perceptual quality scores of other utterances.

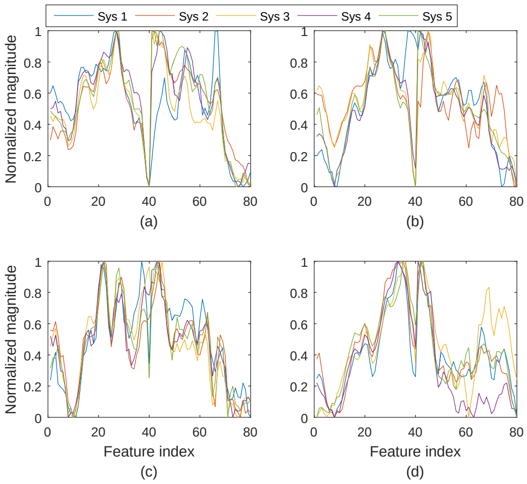

Figure 4: (a) Mean and variance of filterbank features for an utterance from different systems of Blizzard 2008 data for same text spoken. Similarly, (b), (c) and (d) are mean and variance of filterbank features from Blizzard 2009, 2010 and 2012 database, respectively.

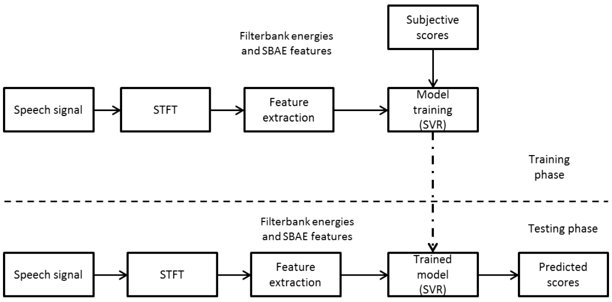

Figure 5: Block diagram of proposed quality assessment system.

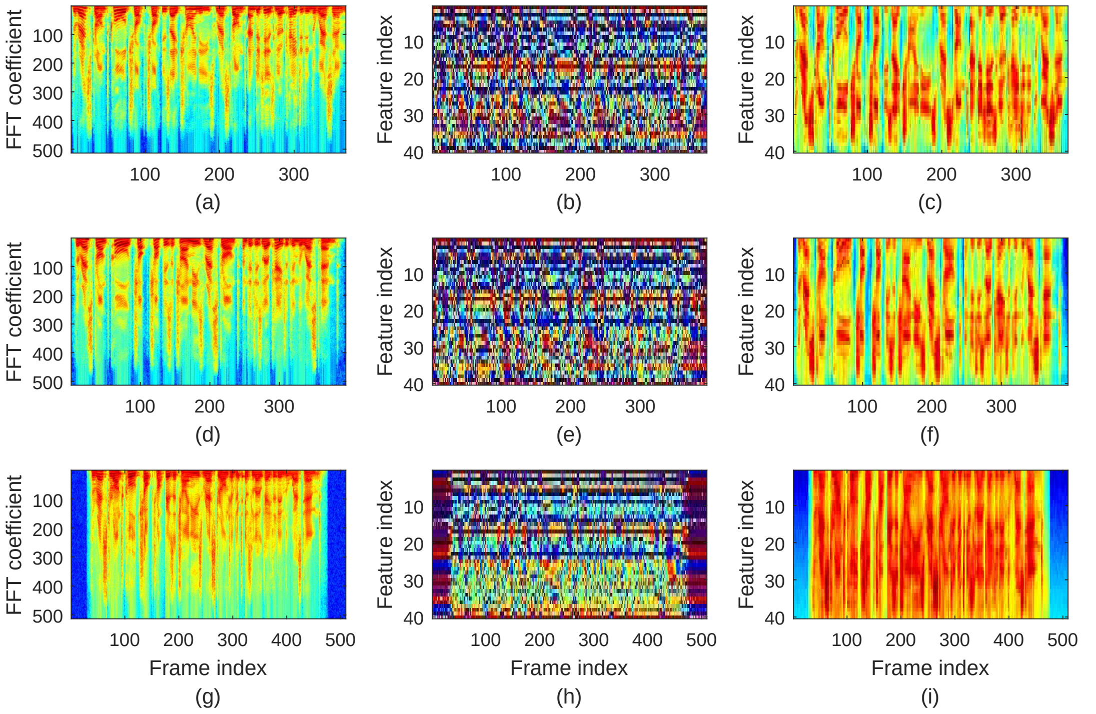

Figure 2: (a) Spectrum, (b) SBAE features and (c) filterbank energies of an utterance from system B of Blizzard 2009 database. Similarly, (d), (e) and (f) are spectrum, SBAE features and filterbank energies, respectively, for an utterance from system C. (g), (h) and (i) are spectrum, SBAE features and filterbank energies for an utterance from system E. The text was the same for all the utterances.

4. Experimental Results

4.1. Experimental Setup

All the experiments were done on Blizzard Challenge databases. Databases of challenge 2008, 2009, 2010 and 2012 were chosen for experimentation. Data of Blizzard Challenge 2011 was not used due to the reason that it consisted all female utterances while other databases consisted of male utterances. For gender uniformity in data and to check the possibility of using the different database for the training of the system than the testing database, Blizzard Challenge 2011 data was not included in experiments. All the databases had natural as well as synthetic speech utterances generated by different systems. The database also included MOS for selected utterances which are used for experimentation. Natural utterances were omitted from experiments since the focus of the work is to predict the quality of synthesized speech. To measure the quality, spectral features were extracted from pre-processed speech utterances. The pre-processing was carried out in the same way as [12]. *40*-D filterbank energies and SBAE features were extracted from speech utterances. The mean and variance of the feature for each utterance were calculated and *80*-D feature vector was generated. The mapping between *80*-D feature vector and MOS was found using SVR. The parameters of SVR were calibrated using a small validation dataset. SBAE was trained on Blizzard Challenge 2012 database and used for all other databases to extract features. The configuration of SBAE was *513-40-200-513*, meaning *513* units in the first layer, *40* units in the second layer which is subband layer and so on. Various experiments were conducted to check the performance of proposed approach.

4.2. Cross-validation to find suitable features

The first experiment was conducted to find suitable representation or features for the quality assessment task. To find the most suitable features for solving the problem, cross-validation was performed on all the Blizzard databases individually using both filterbank energies and SBAE features. To perform cross-validation, the data available for each database was divided into batches of *10* utterances. A number of total batches (B) were dependent upon the number of test utterances available for that particular database. To perform cross-validation, *1* batch out of B batches were used for testing purpose, while (B-1) batches were used for training the SVR. *80*-D input vector consisting mean and variance of spectral features were given to the SVR along with the MOS value of that utterance. Trained SVR was used for predicting MOS of test utterances. The experiment was repeated till all B batches were used for testing purpose. To evaluate the performance of proposed system, 3 metrics, namely, Pearson's correlation coefficient C_p, Spearman's rank-order correlation C_r and Root Mean Square Error (RMSE) between predicted scores and target MOS were used. Table 1 shows the result of cross-validation using filterbank energies and SBAE features. It can be observed that for Blizzard 2008, 2009 and 2012 databases, the SBAE performed significantly better than filterbank features. On Blizzard 2010 database, filterbank features gave more correlation than SBAE features, while RMSE was less using SBAE features. However, the difference between C_p and C_s using both features was *not* significant. Hence, it can be said that using SBAE features, more powerful mapping can be found between spectral features and MOS. Due to better performance, further experiments were conducted only using SBAE features. Figure 7 shows scatter plots for all Blizzard databases for cross-validation experiments. It can be observed that for Blizzard 2012, the scores predicted were very close to the ideal line, while in the case of other databases, the results were poorer than Blizzard 2012 database.

Table 1: C_p, C_s and RMSE between predicted scores and MOS for cross-validation experiment. Results are shown using SBAE features and filterbank energies (FBE). BC indicates Blizzard Challenge database used for experiments.

Database	Using SBAE			Using FBE		
BC	C_p	C_s	RMSE	C_p	C_s	RMSE
2008	**0.69**	**0.62**	**0.43**	0.63	0.56	0.47
2009	**0.79**	**0.68**	**0.34**	0.71	0.60	0.40
2010	**0.75**	0.67	**0.44**	**0.75**	**0.69**	0.45
2012	**0.89**	**0.87**	**0.23**	0.82	0.78	0.25

4.3. Experiments using Different Amount of Training data

In cross-validation, the amount of training data taken was very large. It is not practical to use such large data for training since large scale subjective tests has to be conducted for getting MOS of each training utterance. In another set of experiments, different amount of training data was used to train SVR. the amount of training data was varied from very less (*5 %*) to very high (*90 %*) and performance of the system was checked. In the first set of experiments in this Section, the training data was taken from the same database on which testing had to be performed. For example, if testing data consisted utterances from Blizzard 2012 database, then SVR was trained using utterances of Blizzard 2012 database only. The utterances used in training and testing were different. Training utterances included synthesized utterances from all the speech synthesis systems that were available in the particular database. We also tried to use utterances of different speech synthesis systems in training and testing, however, results were not encouraging. We concluded that for this method to work, utterances of all the systems must be included in the training of SVR. The first row of figure 6 shows results of these experiments. In another set of experiments, utterances from other database were also included while training the model. If test dataset includes *90 %* utterances of Blizzard 2012 database, then *10 %* utterances of Blizzard 2012 database and all the utterances of other database were used for testing. Ultimately, the training utterances from the test database were same as the earlier experiment. This experiment was performed to verify the possibility of using available labeled database for quality assessment problem. The second row of figure 6 shows results of these experiments.

4.4. Results and Discussion

As it can be observed from figure 6 that proposed approach works very well on Blizzard 2012 database. After using only about *12 %* of available utterances in training, the correlation (C_p) of *0.75* per stimulus was achieved in first experiment. Furthermore, after using utterances of other databases in training, the C_p of *0.8* was achieved. Values of C_s and RMSE were *0.8* and *0.54*, respectively, for the case. As training data was increased, the prediction accuracy was increased in the case of Blizzard 2012 database. The same behavior was observed for other metrics. C_s was increased and RMSE was decreased significantly after adding more utterances in the training set. Figure 8 shows predicted scores for Blizzard 2012 database for *12 %* data used in training. It can be observed that scores are correlated with actual MOS even after using a small amount of data in training. Figure 8 shows results of different combination of training and testing data. The amount of training data was same for the different combinations. Similar behavior was observed for Blizzard 2010 database. Performance in the case of Blizzard 2010 was improved after adding utterances of other databases in training. However, performance on 2010 database was not as good as 2012 database. For about *10 %* of data used for training, the C_p, C_s and RMSE were *0.58*, *0.55* and *0.68*, respectively for test utterances. The RMSE in the case of 2010 database was the highest amongst all the databases.

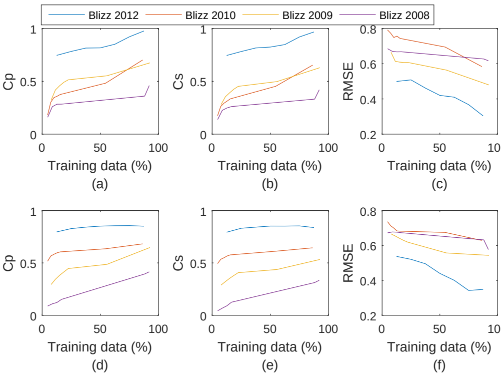

Figure 6: *(a), (b) and (c) shows C_p, C_s and RMSE, respectively after using different amount of training data. In this experiment, training and testing data was from same database. (d), (e) and (f) shows C_p, C_s and RMSE, respectively, after using different amount of training data. In this experiment, the training data was taken from all Blizzard databases.*

Results for Blizzard 2009 and 2008 database were worse than 2010 and 2012 database. In case of Blizzard 2009 database, the value of C_p, C_s and RMSE were *0.42*, *0.35* and *0.614*, respectively. However, the previous best results that were achieved on Blizzard 2009 only using spectral features were *0.09*, *0.08* and *1.34* for C_p, C_s and RMSE, respectively. Hence, proposed approach gives significant improvement in performance only using spectral features. However, after adding some more information such as ITU-T P.563 parameters and other general speech parameters described in [20], the maximum values of C_p, C_s and RMSE were *0.49*, *0.48* and *1.09*, respectively. Comparable results (C_p=*0.49*, C_s=*0.43* and RMSE=*0.6*) were observed using only spectral features in proposed approach if about *20 %* data was used in training. Results on Blizzard 2009 database were comparable to the results of [20] for about *88 %* data used in training. Similarly, results for Blizzard 2008 database (C_p=*0.28*, C_s=*0.26* and RMSE=*0.67*) were comparable with previous best (C_p=*0.3*, C_s=*0.3* and RMSE=*0.8*) using spectral features only by using around *16 %* data in training.

Figure 8 shows some limitations of proposed approach. The first limitation is that proposed approach did not give good scores for utterances with higher MOS value. The region marked by dotted circle suggests that utterances with higher MOS were given low scores by proposed method. The difference in MOS and predicted scores was significant in some cases. The similar observation can be done in the case of utterances with low quality or low MOS. Regions marked by solid curve suggests that poor quality utterances were given high scores. Moreover, the

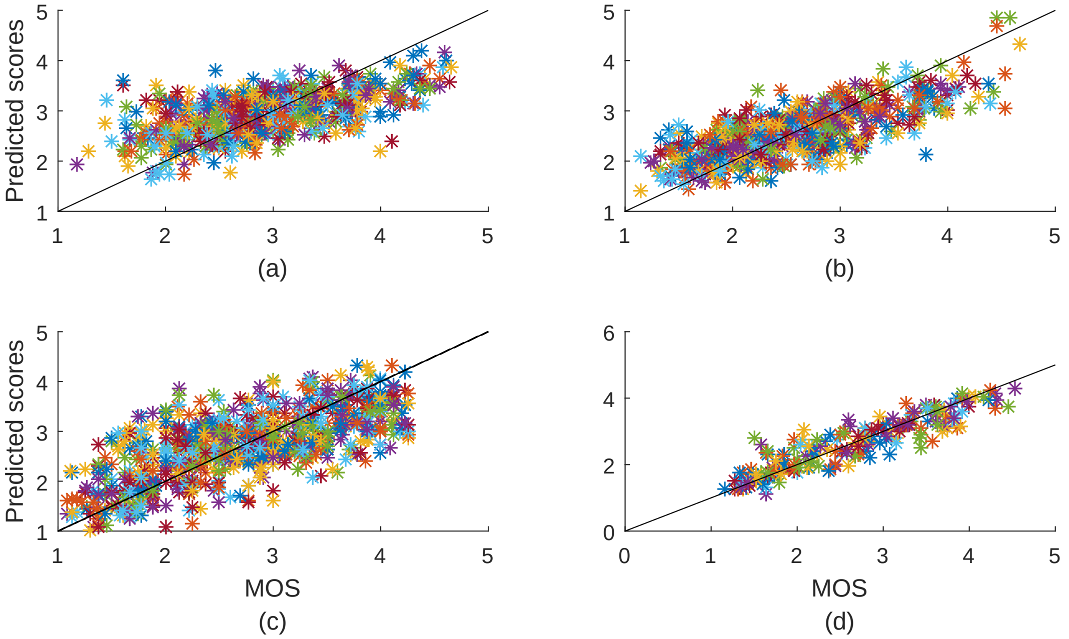

Figure 7: Scatter plot of predicted scores on Blizzard Challenge (a) 2008, (b) 2009, (c) 2010 and (d) 2012 databases for cross-validation experiments. Black line shows the ideal mapping between MOS and predicted scores.

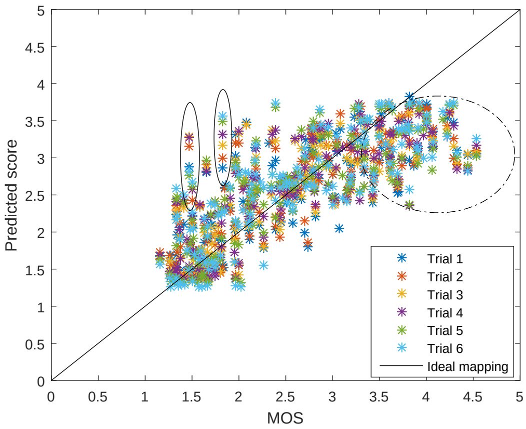

Figure 8: Scatter plot of predicted scores of Blizzard 2012 database for *12 %* data used in training. Total 6 trials were conducted for different combination of training and testing data. The dotted circle indicates the dependency of predicted scores on training data for utterances having higher MOS. The solid circles show same dependency for utterances having low MOS.

same utterance gave different score for different training data. The difference among the scores for same utterances was significant in some cases. It suggests that proposed approach is quite dependent on proper training data. The type of training data required for better results requires further investigation. Proposed approach gave nice results for utterances having MOS between 2.5-3.5. Overall results suggest that as the development in speech synthesis technology year-by-year, the synthesized utterances developed nice statistical properties which can be used in the objective quality evaluation. Moreover, with advancement in the performance of HTS technology from 2008 to 2012, the number of systems which used HTS and hybrid approach for synthesis increased more than USS systems. This may be the possible reason for the better performance of proposed approach in Blizzard 2012 database than other databases. However, to verify this further studies are required. Results of cross-validation experiments can be used to verify that whether the synthesized utterances shows uniformity in statistical properties for different synthesis systems or not.

5. Summary and Conclusions

In this paper, we proposed a new algorithm for non-intrusive objective quality assessment of synthesized speech. In proposed approach, statistics (mean and variance) of spectral features of an utterance, such as Mel filterbank energies and SBAE features was used to find the mapping between features and MOS of the utterance. Hence, quality assessment problem was posed as a regression problem. To find the mapping between features and MOS, Support Vector Regression (SVR) was used. Various experiments were conducted on Blizzard Challenge 2008, 2009, 2010 and 2012 databases. With the different amount of training data from the testing database, the model was trained and objective scores were found using trained model. Results of the experiments suggested that proposed approach worked quite well on Blizzard 2012 database. The Pearson's correlation coefficient between scores generated by using about *12 %* of total data in training was *0.8* for per-stimulus results. However, the proposed approach did not work that well on other databases, it worked better in the case of Blizzard 2009 database than baseline method using spectral features only. On Blizzard 2008 database, the results were comparable to baseline system using spectral features only. However, this approach requires careful studies and experimentation for further applications.

For future work, other features can be incorporated along with spectral features. Moreover, SVR only takes global properties of the utterance into consideration. A better model which can incorporate time variations in the utterance can be used for this application. Since time variations are important for the perception of speech, it would be better to use such models for stated task.

6. Acknowledgments

The authors would like to thank Dept. of Electronics and Information Technology (DeitY), Govt. of India, for sponsored project, Development of Text-to-Speech (TTS) System in Indian Languages (Phase-II) and the authorities of DA-IICT, Gandhinagar, India.

7. References

[1] R. Van Bezooijen and V. Van Heuven, "Assessment of speech output systems," *Handbook of standards and resources for spoken language systems*, pp. 481–563, 1997.

[2] C. Delogu, S. Conte, and C. Sementina, "Cognitive factors in the evaluation of synthetic speech," *Speech Communication*, vol. 24, no. 2, pp. 153–168, 1998.

[3] ITU-T Rec. P.85, "A method for subjective performance assessment of the quality of speech voice output devices," Geneva, 1994.

[4] D. Sityaev, K. Knill, and T. Burrows, "Comparison of the ITU-T P. 85 standard to other methods for the evaluation of text-to-speech systems," in *in Proc. of* 9^{th} *International Conference on Spoken Language Processing (ICSLP)*, Pittsburgh, PA, 2006.

[5] M. Viswanathan and M. Viswanathan, "Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (mos) scale," *Computer Speech and Language*, vol. 19, no. 1, pp. 55–83, 2005.

[6] M. Cernak and M. Rusko, "An evaluation of synthetic speech using the pesq measure," in *Proc. European Congress on Acoustics*, 2005, pp. 2725–2728.

[7] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective Measures of Speech Quality*. Prentice Hall, 1988.

[8] ITU-T Rec. P.862, "Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-to-End Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs," Geneva, 2001.

[9] A. Mariniak, "A global framework for the assessment of synthetic speech without subjects," in 3^{rd} *European Conference on Speech Communication and Technology (EUROSPEECH)*, Berlin, 1993, pp. 1683–1356.

[10] H. B. Sailor and H. A. Patil, "Fusion of magnitude and phase-based features for objective evaluation of tts voice," in *Chinese Spoken Language Processing (ISCSLP), 2014 9th International Symposium on*. Singapore: IEEE, 2014, pp. 521–525.

[11] S. Hardik, "Objective Evaluation of Quality of Text-To-Speech synthesis systems," Ph.D. dissertation, Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar, India, 2013.

[12] T. H. Falk and S. Möller, "Towards signal-based instrumental quality diagnosis for text-to-speech systems," *Signal Processing Letters, IEEE*, vol. 15, pp. 781–784, 2008.

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

[14] ITU-T Rec. P.563, "Single-ended Method for Objective Speech Quality Assessment in Narrow-band Telephony Applications," Geneva, 2004.

[15] T. H. Falk, S. Möller, V. Karaiskos, and S. King, "Improving instrumental quality prediction performance for the blizzard challenge," in *Proc. Blizzard Challenge Text-to-Speech Workshop*, vol. 5, Brisbane, 2008.

[16] S. Möller, D.-S. Kim, and L. Malfait, "Estimating the quality of synthesized and natural speech transmitted through telephone networks using single-ended prediction models," *Acta Acustica united with Acustica*, vol. 94, no. 1, pp. 21–31, 2008.

[17] S. Möller and T. Falk, "Single-ended quality estimation of synthesized speech: Analysis of the rec," in *Internal Signal Processing. ITU-T SG12 Meeting*, 2008, p. 563.

[18] S. Möller, F. Hinterleitner, T. H. Falk, and T. Polzehl, "Comparison of approaches for instrumentally predicting the quality of text-to-speech systems." in *in Proc. of INTERSPEECH*, 2010, pp. 1325–1328.

[19] W. Minker, G. G. Lee, S. Nakamura, and J. Mariani, *Spoken Dialogue Systems Technology and Design*. Boston: Springer Science and Business Media, 2010.

[20] F. Hinterleitner, S. Möller, T. H. Falk, and T. Polzehl, "Comparison of approaches for instrumentally predicting the quality of text-to-speech systems: Data from Blizzard Challenges 2008 and 2009," in *Blizzard Challenge Workshop 2010*, 2010, pp. 48–60.

[21] S. King, R. A. Clark, C. Mayo, and V. Karaiskos, "The Blizzard Challenge 2008," 2008.

[22] A. W. Black, S. King, and K. Tokuda, "The Blizzard Challenge 2009," 2009.

[23] 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.

[24] 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.

[25] M. Soni and H. Patil, "Novel subband autoencoder features for non-intrusive quality assessment of noise suppressed speech." in *accepted for publication in INTERSPEECH*, San Francisco, USA, 2016.

[26] S. King and V. Karaiskos, "The Blizzard challenge 2010," 2010.

[27] ——, "The Blizzard Challenge 2012," 2012.

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