MACHINE LEARNING BASED NON-INTRUSIVE QUALITY ESTIMATION WITH AN AUGMENTED FEATURE SET

Mona Hakami and W. Bastiaan Kleijn

School of Engineering and Computer Science, Victoria University of Wellington, New Zealand

ABSTRACT

We present a method that improves the objective quality estimation of a speech utterance. We show that including raw features that are presumably redundant reduces the effect of input noise and improves the performance of linear regressors. To exploit this effect we propose the novel idea to augment the feature set with redundant features. The proposed augmented feature set and the neural network that consists of an auto-encoder and a linear regressor leads to improved prediction accuracy of the single-ended quality assessment approach. Evaluating the system on the ITU-T Supplement 23 database illustrates that the proposed approach outperforms the current state-of-the-art.

Index Terms— Feature augmentation, machine learning, non-intrusive quality assessment.

1. INTRODUCTION

The success of any new speech transmission service in the telecommunication industry depends on the opinion of end-users about the perceived speech quality. Thus, a reliable estimation of speech quality is necessary to enable the developers of a new service to evaluate its quality or for the service providers to assess the quality of speech on a regular basis. *Subjective* assessment [1], in which human subjects score the quality of transmitted speech utterances, is the most reliable method for assessing voice quality. Subjective tests in general are expensive and time consuming. Therefore *objective* quality assessment algorithms that provide accurate automatic assessment of voice quality are desirable.

Objective algorithms are called *intrusive* [2, 3] if they require both reference and degraded signals to estimate the distortion introduced by the system under test. The algorithms are *non-intrusive* [4, 5] if they do not depend on a reference signal. Non-intrusive methods are important tools for monitoring speech quality of in-service systems, where the clean reference signals are not available. However, the design of a non-intrusive system is more complicated than intrusive models and its performance is generally lower than systems that use a reference signal.

In the conventional non-intrusive algorithms such as [4], the knowledge of specialists is used to design complex algorithms that model the interaction of the features and their contribution to the overall quality of the audio. In contrast, machine learning methods avoid designing an explicit model and apply the statistical learning from training data. This generally results in a performance improvement as the trained models are not based on poor assumptions or inadequate knowledge. Such systems are also desirable as they are flexible to be adapted to various applications and are not restricted to any particular service. For example, a method implemented for narrow-band data can be re-trained on wide-band database and used for a wide-band service.

In machine learning terminology, non-intrusive quality estimation can be described as a regressor or a multi-class classifier that maps the signal features to the quality score. In the absolute category rating (ACR) listening quality method [1], which is the most commonly used subjective test procedure in telecommunications [6], the subjects rate speech files using a five-level impairment scale. The average over all rating scores of a speech file represents its subjective listening quality mean opinion score (MOS). Our objective is to develop a regressor that predicts MOS values that are highly correlated with the MOS obtained from subjective tests.

Several non-intrusive methods that use machine learning algorithms for estimating the score of audio signals have been proposed [7, 8, 9, 10]. Our experimental results with state-of-the-art regressors [11, 12, 13] and reviewing the scores reported in the literature indicates the overall performance of speech quality assessment systems is to be improved by either 1) having more training data, or 2) enhanced features. Collecting more training data is expensive and time consuming. Hence the focus of this research is to form an enhanced feature set that results in a performance improvement.

In this work we present the novel idea to augment the feature set using raw features that are presumably redundant. This reduces the effect of input noise and hence improves the performance. Section 2 explains this novel idea in more detail. Section 3 explains how this novel idea is applied to quality assessment and discusses the implementation aspects. Section 4 represents the experimental results followed by the conclusion in section 5.

2. FEATURE SET AUGMENTATION

The objective we pursue in this work is to augment the feature set. That is we enlarge the number of features. This section analyses the scenario where the input features are noisy and shows including redundant features reduces the effect of input noise and that results in the better performance.

The term "curse of dimensionality" was first introduced by Bellman [14]. It states that if the dimensionality of the input feature set is very large in comparison to the number of observations, the convergence of predictors to the true value of a smooth function is very slow. The expression "blessing of dimensionality" [15] suggests an opposing viewpoint and declares although high-dimensional feature increases the cost of learning algorithm to overcome the curse effect, it constructs more informative features that lead to high performance. In particular [16] suggested that *including presumably redundant variables might result in a performance gain*. This statement became very well-known and a large number of papers (e.g., [17, 18]) refer to this phrase. However, a detailed analysis for the performance gain does not appear to exist. In this work we analyse how redundant features can improve the performance of machine learning models if they represent the same information, but contain independent noise. We initially focus on the linear regression meth-

ods. Our results suggest this performance gain can be generalised into nonlinear learning problems too.

In the following we study the behaviour of enlarged feature sets for two different models: 1) the ground truth model has low dimensional features and we enlarge our feature set by observing redundant features that decrease the effect of input noise, and 2) the ground truth model has high dimensional features and we enlarge our feature set by adding to it additional features that contain new information. The different behaviours of two models above are studied in section 2.1 and 2.2.

2.1. Model behaviour for redundant features

In this section we assume the observed features are redundant and contain the same information, but have independent noise. We first write the ground truth model for the linear quality estimation. Then we model the relation between the number of features and the performance of the linear quality estimator. We will show that the variance of error varies inversely as the number of redundant features in the augmented feature set.

Let $\mathbf{x}$ be the realization of an underlying random feature vector X. The MOS is computed as

$$\mathrm{MOS} = a^T \mathbf{x}. \tag{1}$$

We aim to develop an MOS estimator based on a set of observable features $\mathbf{y}$ that are redundant:

$$\hat{\mathrm{MOS}} = b^T \mathbf{y}. \tag{2}$$

We assume $X \sim \mathcal{N}(0, R_X)$, and that the random observations are of the form $Y = C(X+U)+W$, where C is a matrix, and U and W are random noise vectors called *intrinsic noise* and *observation noise* respectively.

The random variable V is the measurement error and defined as

$$V = (a^T - b^T C)X - b^T C U - b^T W. \tag{3}$$

Vector b must be estimated from the observations Y aiming to minimize the measurement error on the training data.

σ_V^2 is scalar and X, U, and W are independent. Hence the optimization criteria can be written as

$$\begin{aligned}\sigma_V^2 =& \mathrm{tr}[\mathrm{E}[X^T(a^T - b^T C)^T(a^T - b^T C)X \\ &+ U^T C^T b b^T C U + W^T b b^T W]].\end{aligned} \tag{4}$$

Exchanging the linear operators, the expectation and the trace, and using the cyclic property of the trace lead to

$$\sigma_V^2 = b^T(CR_X C^T + CR_U C^T + R_W)b - 2a^T R_X C^T b + a^T R_X a. \tag{5}$$

The optimal vector b that minimizes the variance of V is.

$$b^* = (CR_X C^T + CR_U C^T + R_W)^{-1} CR_X a. \tag{6}$$

Substituting (6) back into (5) gives

$$\begin{aligned}\sigma_V^2 =& a^T R_X C^T (CR_X C^T + CR_U C^T + R_W)^{-1} \\ &\times (CR_U C^T + R_W) C^{T\sharp} a.\end{aligned} \tag{7}$$

Let us assume $R_X = I_{d\times d}$ (so X is normalized to have a unit variance), $R_U = hI_{d\times d}$, and $R_W = gI_{t\times t}$, where d and t are dimensionality of X and Y respectively, and g and h are small. Then

$$\sigma_V^2 = a^T(C^T C + hC^T C + gI)^{-1}(hC^T C + gI)a. \tag{8}$$

For simplicity we initially assume each feature is repeated n times. Hence C is a tall matrix of n stacked identity matrices and $C^T C = nI$. We have

$$\sigma_V^2 = a^T(nI + hnI + gI)^{-1}(hnI + gI)a \tag{9}$$

$$= \frac{g + nh}{g + n(h+1)} a^T a. \tag{10}$$

The behaviour of equation (10) is shown in fig. 1.a, which indicates the variance of the error asymptotically goes down to $\frac{h}{h+1} a^T a$. This behaviour is clearer if we do not have intrinsic noise ($h = 0$):

$$\sigma_V^2 = \frac{g}{g+n} a^T a, \tag{11}$$

in which σ_V^2 goes to zero for large n.

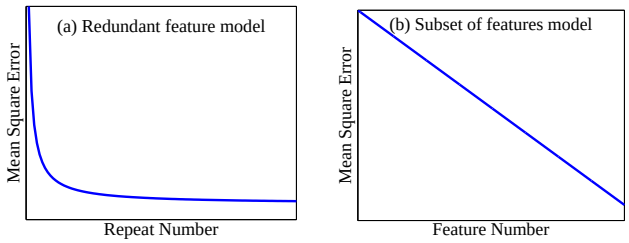

Fig. 1. Different model behaviours for enlarging feature set.

Now let us consider the more general case, where C is a tall $n \times d$ matrix, in which n and d are the dimensionality of the observed and the underlying features respectively. Let us assume $C \sim \mathcal{N}_N(0, \Sigma_c)$, where Σ_c is a diagonal matrix with diagonal elements equal to σ_c. We investigate the behaviour of σ_V^2 by estimating the expectation value of equation (8). To make this analytically tractable and show the principal model we assume $h = 0$. We analyse the main aspect of the model behaviour with re-writing equation (8) as

$$\sigma_V^2 = g a^T (C^T C + gI)^{-1} a, \tag{12}$$

$$E_C[\sigma_V^2] = g a^T E_C[(C^T C + gI)^{-1}] a. \tag{13}$$

The elements of C are i.i.d and have normal distribution. Thus $C^T C \sim \mathcal{W}_N(\Sigma_c, n)$ has Wishart distribution [19], and its mean is

$$E[(C^T C)_{ij}] = \begin{cases} n\sigma_c & i = j \\ 0 & i \neq j \end{cases}.$$

Since g is very small in compare with n, we estimate $E_C[(C^T C + gI)^{-1}]$ with $E_C[(C^T C)^{-1}]$, which follows

$$E_C[\sigma_c^2] \sim g a^T E[(C^T C)^{-1}] a. \tag{14}$$

C is a tall matrix with normal distribution. Thus $(C^T C)^{-1} \sim \mathcal{W}^{-1}(\Sigma_c^{-1}, n)$ has an Inverse Wishart distribution. With the assumption $\sigma_c = 1$ we have

$$E[(C^T C)^{-1}_{ij}] = \begin{cases} \frac{1}{n-d-1} & i = j \\ 0 & i \neq j \end{cases}.$$

Finally we estimate the mean of σ_c^2 as

$$E_c[\sigma_c^2] \sim \frac{g}{n-d-1} a^T a, \tag{15}$$

where n is the dimensionality of observed features. This suggests the variance of the error varies inversely as the the number of features , motivating the augmentation of the feature set with redundant features for higher performance.

2.2. Model behaviour for insufficient features

In this section we develop an MOS estimator based on the assumption that the observed feature set $\mathbf{y}$ is a subset of underlying feature set $\mathbf{x}$. We show that the relation between the performance of the system and the number of the features is linear and compare its behaviour with the model with redundant features explained in the previous section.

Let us assume $Y = SX + W$, where W is random observation noise. We define $S = [I_{n\times n} \quad 0_{(N-n)\times(N-n)}]$, in which n and N are the number of selected features and the number of full features respectively. Accordingly, the measurement error is

$$V = (a^T - b^T S)X - b^T W \tag{16}$$

and the optimization criteria is defined

$$\sigma_V^2 = a^T R_X a + b^T (S R_X S^T + R_W) b - 2a^T R_X S^T b. \tag{17}$$

The optimal vector b that minimizes σ_V^2 is

$$b^* = (S R_X S^T + R_W)^{-1} S R_X a. \tag{18}$$

Substituting (18) back into (17) provides

$$\sigma_V^2 = a^T R_X [I - S^T (S R_X S^T + R_W)^{-1} S R_X] a. \tag{19}$$

Let us assume $R_X = I$, so that X is normalized to have unit variance, and $R_W = g I_{n\times n}$. We obtain

$$\sigma_V^2 = a^T [I_{N\times N} - S^T (I_{n\times n} + g I_{n\times n})^{-1} S] a \tag{20}$$

$$= \sum_{i=1}^{N} \lambda_i a_i^2, \tag{21}$$

where $\lambda_i = 1$ if $i > n$ and $\lambda_i = \frac{g}{g+1}$ if $i \leq n$. With the assumption $a_i \sim N(E(a_i), \sigma_i^2)$ we have

$$E(\sigma_V^2) = \frac{N - \frac{n}{g+1}}{N} \sum_{i=1}^{N} E(a_i^2). \tag{22}$$

Equation (22) indicates the variance of error and the number of selected features have a linear relationship, whereas their relationship in the model with the redundant features is an inverse variation. The various behaviours of two models are shown in fig. (1).

3. AUGMENTED FEATURE SET FOR QUALITY ASSESSMENT

This section describes the proposed pre-processing approach to build an augmented feature set for quality assessment. The non-intrusive quality estimation P.563 and ANIQUE+ standards form a natural reference for our work and we use the features extracted from them to build our input vector. The feature sets from the both standards are expected to represent similar information about the quality of the speech. Hence the proposed input vector is likely to hold redundant features. It is expected that quality assessment system benefits from this redundancy as it results in reducing the impact of input noise based on discussion in section 2.

In ITU-T Recommendation P.563 [4], the incoming speech signal is analysed by several modules and a set of global parameters are determined. A restricted set of the parameters are used to determine the main distortion class of the speech signal: 1) Unnatural voice (male, female, robotization), 2) High additional noises (low static SNR, low segmental SNR), 3) Interruptions, mutes and time clipping. Each class distortion uses a linear combination of a set of parameters to generate the intermediate speech quality. P.563 parameters, $\Xi = \{\xi_i\}_{i=1}^{43}$, naturally form an informative global feature set for the quality assessment platforms.

In American national standard ANIQUE+ [5], the input signal is decomposed into successive time frames that are classified into *active speech* or *audible background noise* frames. The functional blocks, motivated by human auditory systems at peripheral and central levels, analyse the speech frames and obtains 69 features relevant to human speech quality perception. These perceptual features form a local feature vector Φ for frame quality degradation. We use the method suggested in [20] and compute the first four moments of the features to convert the per-frame features to per-utterance features:

$$\Psi = \{\mu_{\Phi_i}, \sigma_{\Phi_i}, s_{\Phi_i}, k_{\Phi_i}\}_{i=1}^{69}. \tag{23}$$

$\mu_{\Phi_i}, \sigma_{\Phi_i}, s_{\Phi_i}, ,$ and k_{Φ_i} are mean, variance, skewness, and kurtosis of the feature Φ_i that is computed over the speech active frames.

In ANIQUE+ we computed the per-utterance features from statistical attributes of the per-frame features. In contrast, the majority of the per-utterance parameters in P.563 are calculated based on the estimation of the distortion in the whole signal. The other per-utterance features in P.563, which are also based on the calculation of speech statistics, are extracted from *vocal tract* module, which unlike ANIQUE+ that focuses on the auditory system, models the speech production system. Although the feature sets generated from ANIQUE+ and P.563 have different natures, they both represent same information about the precieved quality of speech. Hence we build an augmented feature set $\Sigma = \{\Xi, \Psi\}$ with 319 features that are expected to contain independant observation noise.

Having a large number of features in Σ, naturally leads to the inclusion of the features that have poor behaviour. To facilitate the training, we standardize each feature in Σ to obtain pre-distorted feature x_i with uniform probability distribution and build our final augmented feature set $X = \{x_i\}_{i=1}^{319}$.

4. EXPERIMENTAL RESULTS

To evaluate the proposed system, we used seven data sets with absolute category ratings from ITU-T coded-speech data set, Supplement 23 [21]. The data sets contain 1328 speech files, where the MOS for each utterance is the average rating over 24 subjects. In the experiments with the features in section 4.1, we pooled all data sets together and did cross-validation using six-seventh of the data for the training and the remainder for the test. In the final experiment to evaluate the proposed system in section 4.2 we used a cross-validation procedure with leaving one data set out in each iteration. In each round of the cross-validation we computed Root Mean Squared Error (RMSE) and Pearson correlation-coefficient (PCC) for both per-file and per-condition. The scores are reported after applying a third-order monotonic polynomial, as is standard practice to reduce the effect of per-experiment variation [22].

Section 4.1 demonstrates how quality assessment system benefits from redundant features. Section 4.2 presents the experimental results from the proposed quality assessment system and compares its performance with the existing methods.

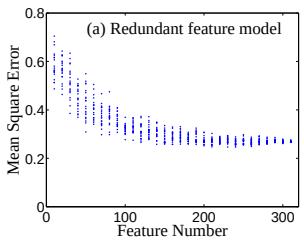

Fig. 2. Relation between the number of the features and the performance of linear quality predictor.

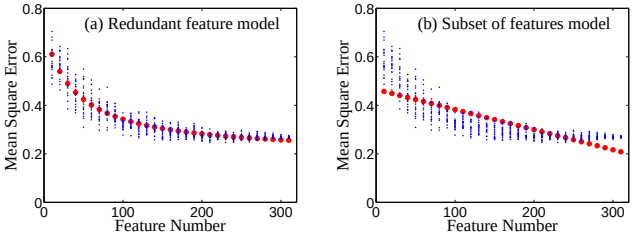

Fig. 3. The behaviour of non-linear model performance fits to models with redundant and subset of features.

4.1. Experiment with features

This section evaluates the relation between the number of the features and the performance of the quality predictor and shows that our experimental results fit the model with redundant features better. The experimental results show the augmented feature set X proposed in section 3, containing redundant features from P.563 and ANIQUE+, improves the performance of speech quality assessment.

We initially use a simple linear regressor and analyse the relation between its performance and the number of the features. We form the input vector, X_n, by randomly selecting $n = 10, 20, 30, ..., 310$ features from X. We repeat the experiment with 20 random X_n and compute the RMSE using cross-validation on SUPPL23. Fig. (2) presents the results, which suggests the behaviour of our system is strongly similar to the model for the redundant features from section 2.1. It is also observed that the variance of the performance of the system is small for larger values of n. This is to be expected as in our experiment the overall number of the features in X is fixed. Hence the different random X_ns are more likely to include same features for larger n and that results in a smaller variance in the performance of the quality predictor.

We then repeated our experiment with a non-linear regressor. We used a neural network that has one hidden layer with five nodes using sigmoid activation function, followed by a linear regressor. The experimental results with the proposed neural network show that the non-linear predictor has the same behaviour as linear one. The blue points in figure 3 represents the experimental results. The red points in figure 3.a and 3.b are the two candidate models from redundant and insufficient features fit to data respectively. We used the Akaike information criteria [23] to compare the fit of these two models to our experimental results. The model with redundant features fits to our data better than the model with the insufficient features and its evidence ratio is 2.5×10^{31} . This confirms the performance of the quality assessment is improved because the augmented feature set X contains redundant features from P.563 and ANIQUE+.

4.2. Experimental result for quality of speech

This section evaluates the performance of our proposed non-intrusive system with the augmented feature set. Experimental results with SUPPL23 database indicates our proposed system performs better than state-of-the-art in the field.

We configured our system to have one hidden layer that contains five nodes with sigmoid activation function, followed by a linear regressor. To evaluate the effect of our proposed augmented feature set, we performed three experiments with different feature sets shown in Table 1. The results suggest using redundant features from ANIQUE+ and P.563 increases the performance. The detailed scores from the augmented feature set is reported in Table 2.

Table 1. Model performance with different types of feature sets.

	RMSE		PCC	
Input Feature set	Per-File	Per-Cond	Per-File	Per-Cond
P.563	0.40	0.29	0.75	0.87
ANIQUE+	0.38	0.26	0.75	0.89
P.563 and ANIQUE+	0.30	0.21	0.82	0.92

Table 2. Model performance with the augmented feature set.

	RMSE		PCC	
Database	Per-File	Per-Cond	Per-File	Per-Cond
BNR-X1	0.18	0.08	0.86	0.95
BNR-X3	0.21	0. 11	0.84	0.95
CNET-X1	0.23	0.15	0.83	0.92
CNET-X3	0.47	0.38	0.79	0.89
CSELT-X3	0.60	0.51	0.80	0.88
NTT-X1	0.20	0.10	0.81	0.92
NTT-X3	0.27	0.18	0.84	0.92
Mean	0.31	0.21	0.82	0.92

Table 3 reviews the scores reported in the literature related to machine learning methods for assessing the quality of speech. The scores are based on seven-fold cross-validation on ITU-T SUPPL23. ANIQUE+ has the high score 0.98 as ITU-T SUPPL23 was included in the training data bases [24]. The next two high scores reported in [20] and [8] are expected as they used additional databases to train their system with. Although the score reported in [13] is 0.91, its author acknowledged an implementation error and the true score is 0.88. We are unable to do the comparison with other methods such as [25, 26, 9, 7] as the evaluations are performed with databases that are not publicly available, or with databases for which the subjective score is not available. From comparing our experimental results with the scores reported in Table 3 we conclude that the proposed non-intrusive quality assessment with the augmented feature set learns efficiently from a small training database and that it provides a performance that compares favorably to the state-of-the-art in the field.

Table 3. Review of the scores in the literature based on cross-validation on ITU-T SUPPL23. The high scores with asterisk are from systems that used additional databases for training.

Method	PCC
ANIQUE+ [5]	0.98*
Low Complexity, Non-Intrusive Speech Quality ... [20]	0.94*
Non-intrusive speech Quality Assessment Using ... [8]	0.92*
A Hierarchical Bayesian Approach to Modeling ... [13]	0.91
Probabilistic Non-Intrusive Quality Assessment ... [10]	0.91
A Bayesian Estimator for Non-intrusive Speech ... [27]	0.90
A Bayesian Approach to Non-Intrusive Quality ... [28]	0.89
Nonintrusive Speech Quality Evaluation Using ... [29]	0.88
ITU-T P.563 [4]	0.88
A Bayesian Hierarchical Mixture of Experts [30]	0.88

5. CONCLUSION

Our hypothesis was that a non-intrusive speech quality assessment performs better with an augmented feature set that contains features representing the same information but including independent noise. We studied the relation between the performance of the linear regressors and the number of the redundant features and showed that the variance of the error goes down by enlarging the feature set with redundant features. We defined experiments with linear and non-linear regressors to prove that. Based on our results, we can conclude that machine learning based non-intrusive systems benefit from redundant features by reducing the effect of input noise. Our experimental results with the ITU-T Supplement 23 database demonstrated the performance gain associated with the augmentation of the feature set and show that the proposed system outperforms the current state-of-the-art.

6. REFERENCES

[1] International Telecommunications Union (ITU-T), "Methods for subjective determination of transmission quality, Recommendation P.800," Online. http://www.itu.int/rec/T-REC-P.800-199608-I/en.

[2] J. G. Beerends and J. A. Stemerdink, "A perceptual speech-quality measure based on psychoacoustic sound representation," *J. Audio Eng. Soc.*, vol. 42, pp. 115–123, Mar. 1994.

[3] International Telecommunications Union (ITU-T), "P.862 : Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs,Recommendation P.862," Online. http://www.itu.int/rec/T-REC-P.862-200102-I/en.

[4] International Telecommunications Union (ITU-T), "Single-ended method for objective speech quality assessment in narrow-band telephony applications,Recommendation P.563," Online. https://www.itu.int/rec/T-REC-P.Imp563/en.

[5] D. Kim and A. Tarraf, "ANIQUE+: A new American national standard for non-intrusive estimation of narrowband speech quality," *Bell Labs Technical Journal*, vol. 12, pp. 221–236, May 2007.

[6] V. Grancharov and W. B. Kleijn, "Speech quality assessment," in *Springer Handbook of Speech Processing*, pp. 83–102, Nov. 2007.

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

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

[9] 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* (S. Boll, Q. Tian, L. Zhang, Z. Zhang, and Y.-P. P. Chen, eds.), pp. 325–335, Berlin, Heidelberg: Springer Berlin Heidelberg, 2010.

[10] P. N. Petkov and W. B. Kleijn, "Probabilistic non-intrusive quality assessment of speech for bounded-scale preference scores," in *Quality of Multimedia Experience (QoMEX), 2010 Second International Workshop on*, pp. 188–193, 2010.

[11] D. Shutin, T. Buchgraber, S. R. Kulkarni, and H. V. Poor, "Fast adaptive variational sparse Bayesian learning with automatic relevance determination," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pp. 2180–2183, May 2011.

[12] B. McWilliams, D. Balduzzi, and J. M. Buhmann, "Correlated random features for fast semi-supervised learning," in *Advances in Neural Information Processing Systems 26* (C. J. C. Burges, L. Bottou, Z. Ghahramani, and K. Q. Weinberger, eds.), pp. 440–448, 2013.

[13] I. Mossavat, P. N. Petkov, W. B. Kleijn, and O. Amft, "A hierarchical Bayesian approach to modeling heterogeneity in speech quality assessment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 136–146, 2012.

[14] R. Bellman, *Dynamic Programming*. Princeton, NJ, USA: Princeton University Press, 1 ed., 1957.

[15] D. L. Donoho, "High-dimensional data analysis: The curses and blessings of dimensionality," in *AMS Conference on Math Challenges of the 21st Century*, 2000.

[16] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[17] R. Ruiz, J. C. R. A, and J. S. A. ruiz B, "Incremental wrapper-based gene selection from microarray data for cancer classification," *Pattern Recognition*, pp. 2383–2392, 2006.

[18] V. Balasubramanian, S. S. Ho, and V. Vovk, *Conformal Prediction for Reliable Machine Learning: Theory, Adaptations and Applications*. Elsevier Science, 2014.

[19] J. Wishart, "The generalised product moment distribution in samples from a normal multivariate population," *Biometrika*, vol. 20A, no. 1/2, pp. 32–52, 1928.

[20] V. Grancharov, D. Zhao, J. Lindblom, and W. B. Kleijn, "Low-complexity, nonintrusive speech quality assessment," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, pp. 1948–1956, nov. 2006.

[21] International Telecommunications Union (ITU-T), "ITU-T coded-speech database." ITU-T Rec. P.Suppl. 23.

[22] A. W. Rix, "Comparison between subjective listening quality and P.862 PESQ score," *Proc. Meas. Speech Qual. Net. (MESAQIN)*, pp. 17–25, 2003.

[23] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Second International Symposium on Information Theory* (B. N. Petrov and F. Csaki, eds.), (Budapest), pp. 267–281, Akadémiai Kiado, 1973.

[24] D. Kim, *Personal Communication*.

[25] T. Falk and W. Chan, "Single-ended speech quality measurement using machine learning methods," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, pp. 1935–1947, Nov. 2006.

[26] 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.

[27] G. Chen and V. Parsa, "A Bayesian estimator for non-intrusive speech quality evaluation in psychoacoustic domain," in *2006 IEEE International Symposium on Signal Processing and Information Technology*, pp. 438–441, Aug 2006.

[28] P. N. Petkov, I. S. Mossavat, and W. B. Kleijn, "A Bayesian approach to non-intrusive quality assessment of speech.," in *INTERSPEECH*, pp. 2875–2878, ISCA, 2009.

[29] G. Chen and V. Parsa, "Nonintrusive speech quality evaluation using an adaptive neurofuzzy inference system," *IEEE Signal Processing Letters*, vol. 12, pp. 403–406, May 2005.

[30] S. I. Mossavat, O. Amft, B. de Vries, P. Petkov, and W. B. Kleijn, "A Bayesian hierarchical mixture of experts approach to estimate speech quality," in *QoMEX 2010: Second International Workshop on Quality of Multimedia Experience*, pp. 200–205, IEEE Signal Processing Society, IEEE Signal Processing Society, 2010.