International Journal of Speech Technology (2020) 23:585–594
https://doi.org/10.1007/s10772-020-09688-w

Low-complexity disordered speech quality estimation

Yousef S. Ettomi Ali[1] · Vijay Parsa[1,2] · Phillip Doyle[2] · Soulaimane Berkane[3]

Received: 11 June 2019 / Accepted: 11 February 2020 / Published online: 20 February 2020

Abstract
Tracheoesophageal (TE) speech is generated by patients who have undergone a total laryngectomy where the larynx (voice box) is removed and replaced by a tracheoesophageal puncture. This work presents a novel low complexity algorithm to estimate the degree of severity of disordered TE speech. The proposed algorithm has two output scores which are computed from 20 ms voiced frames of the speech signal. An 18th order Linear Prediction (LP) analysis is performed on each voiced frame of the speech signal. The first output score uses features derived from high order statistics (mean, variance, skewness and kurtosis) which are calculated from the LP coefficients, the cepstral coefficients and the LP residual signal. These high order statistics (HOS) along with the pitch value are averaged over all voiced frames yielding a total of 14 HOS quality features. The second output score is derived from features derived from the estimated vocal tract model parameters (cross-sectional tubes areas). Statistical vocal tract parameters (VTPs) across all voiced speech frames were used as speech quality features. Forward stepwise regression as well as K-fold cross validation are then used to select the best sets of features to be fed to the regression models. The results show high correlations with subjective scores for several regression techniques that can provide a correlation up to 0.91 when VTP-Gaussian model is used.

Keywords Tracheoesohageal speech · Speech quality · Linear prediction · Vocal tract parameters

1 Introduction

Voice and speech quality estimation is an important topic of research with many applications in telecommunication and biomedical engineering. Early algorithms that assesses voice and speech quality were developed in the telecommunication industry to evaluate the performance of telecommunication channels, the accuracy of speech coding algorithms and often the efficiency of speech enhancement methods (Union 1996; Rix et al. 2001; Malfait et al. 2006; Beerends et al. 2013). In the biomedical field, voice and speech quality estimation algorithms were developed to evaluate the severity of dysphonia (abnormality in the pereived quality of voice production) (Awan et al. 2010) and the associated voice quality of pathological speech (Parsa and Jamieson 2001; Ritchings et al. 2002; Gu et al. 2005). Besides, algorithms for speech quality evaluation have been developed to monitor Hearing Aid (HA) performance which is important for HA designers and audiologists (Kates and Arehart 2010). Our aim in the present study was to develop an algorithm for disordered speech quality estimation for applications in clinical speech language pathology.

TE speech is a voice restoration method used by those who had undergone total laryngectomy and utilize TE speech as a postlaryngectomy speech communication method (Maniglia et al. 1989). In a total laryngectomy, the entire larynx is removed (including the vocal folds, hyoid bone, epiglottis, thyroid and cricoid cartilage and a few tracheal cartilage rings) (Ward and van As-Brooks 2014). After laryngectomy is performed, TE puncture voice restoration is one voice and speech rehabilitation option. A TE puncture involves

✉ Yousef S. Ettomi Ali
yali23@uwo.ca

Vijay Parsa
parsa@nca.uwo.ca

Phillip Doyle
pdoyle@uwo.ca

Soulaimane Berkane
berkane@kth.se

[1] Department of Electrical and Computer Engineering, University of Western Ontario, London, ON, Canada

[2] School of Communications and Speech Disorders, University of Western Ontario, London, ON, Canada

[3] Department of Computer Sciences and Engineering, University of Quebec in Outaouais, Gatineau, QC, Canada

the creation of a small, controlled opening in the common tissue wall between the trachea and the esophagus.Following creation of the TE puncture, a small, one-way prosthesis is inserted. This allows for the speaker to direct pulmonary air through the prosthesis into the esophagus which can then be used to form TE speech.

The speech produced through the TE prosthesis has often a substantially poorer quality compared to normal speech since the sound source is abnormal and contains different anatomical asymmetries. TE speech is, generally, characterized by lowered fundamental frequency, normal or slightly greater than normal intensity, and because of access to the large volume of pulmonary air, generally normal temporal features (rate of speech) when compared to normal speakers (Robbins et al. 1984). However, the overall sound quality of TE voice and speech is best described as highly aperiodic, rough, and noisy. However, voice and speech quality is not invariant and, considerable variability across TE speakers does exist (Eadie and Doyle 2002, 2005). This necessitates assessment and monitoring of TE voice and speech quality.

Overall, there are two different speech quality estimation paradigms: subjective and objective. In the subjective evaluation of voice and speech quality, a group of listeners is asked to rate a voice/speech sample based on a given quality scale. For instance the mean opinion score (MOS) method has been widely used in telecommunication to evaluate speech quality and to validate standardized quality estimation algorithms (Union 1996). The GRBAS (Grade, Roughness, Breathiness, Asthenia, Strain) and Consensus Auditory Perceptual Evaluation-Voice (CAPE-V) scales are used in the speech pathology field where the clinician rates the perceived quality along different speech attributes such as the roughness, strain, breathiness and overall severity of the sample (Hirano 1981; Kempster et al. 2009).

Although subjective methods for speech quality estimation are considered to be the gold standard, they are often time and resource intensive. On the other hand, objective methods for speech quality estimation are fully automated and are usually developed to computationally predict the subjective scores by studying the correlation between the objective and subjective scores. In general, there are two schemes for objective speech quality estimation: algorithms that require a clean (reference) speech signal, termed intrusive methods, and algorithms which do not use any reference signal, termed non-intrusive methods, where the quality estimation is done solely based on the degraded speech signal.

Many intrusive (also called double-ended) algorithms for speech quality evaluation have been used in telecommunication industry (Rix et al. 2001) and in HA applications (Kates and Arehart 2010). However, these methods are not suitable for pathological voice and speech applications where a clean reference signal is not available. During the last few decades, several research studies have been conducted to assess the voice quality of patients with voice and speech disorders based on acoustical, aerodynamic and physiological measurements. Most of the computationally effective non-intrusive speech quality methods have been validated only on sustained vowels and usually fail to report good correlation when used on continuous speech samples (Parsa and Jamieson 2001). On the other hand, non-intrusive speech quality estimation methods which report good correlation with subjective scores of continuous speech samples are either computationally demanding (Ali et al. 2017) or developed for network assessment (Grancharov et al. 2006).

In this paper, our goal is to propose acoustical features which are easily extracted (computationally simple) from a given speech signal and which are shown to correlate well with subjective ratings of TE speech. First, the voiced frames of the acoustical speech signals are extracted using the simple autocorrelation method (Rabiner et al. 1976) and the corresponding pitch estimation per voiced frame is obtained. The voiced frames of the speech are evaluated using an 18th order Linear Prediction (LP) analysis based on the Levinson-Durbin algorithm. Speech quality features are extracted by computing the average over all frames of high order statistics (mean, standard deviation, skewness and kurtosis) of the LP coefficients, the cepstral coefficients and the LP residual signal. Furthermore, a vocal tract model has been extracted for each voiced frame by computing the parameters of an acoustical tube formed by interconnecting 18 uniform cross sectional tubes. The vocal tract parameters yield extra speech quality features. Finally, the extracted speech quality features have been used to train and test different support vector machine models on a dataset of 35 TE speech samples. The remainder of the paper is organized as follows. In Sect. 2, we describe the proposed voice/speech quality evaluation method by detailing all the different stages and processing blocks. The voice/speech databases used to evaluate our method, as well as the obtained results, are reported in Sects. 3 and 4 respectively. Concluding remarks and recommendations for future work are provided in Sect. 5.

2 Speech quality evaluation method

Our proposed approach for extracting speech quality features from disordered voice signals consists of three main stages. First, preprocessing is conducted to detect voiced and unvoiced speech frames. We use a temporal approach based on the autocorrelation method. Then, Linear Prediction (LP) analysis is performed to extract the LP coefficients, the cepstral coefficients and the residual signal from each frame marked as voiced by the first preprocessing stage.

The LP coefficients are used to derive a vocal tract model by calculating the reflection and the cross sectional areas of the acoustic tube model which provides the first group of

acoustic features. Besides, high-order statics are obtained from LP analysis coefficients and residual signal which constitute the second group of acoustical features. Each group of features is used in a regression-based mapping to provide quality scores for TE voice signals. The schematic of the proposed method for voice quality estimation is depicted in Fig. 1. The different stages listed above are detailed in the next subsections.

2.1 Pitch period estimation and voiced frames extraction

Pathological voice and speech signals are different in terms of their pitch period estimate. It is suggested that inclusion of pitch average estimates in computational models for voice quality may help improve the accuracy of these models. In non-intrusive speech quality measurement algorithms, such as the ITU standard P.563 and the Low-Complexity Speech Quality Assessment (LCQA) proposed in Grancharov et al. (2006), pitch is used as a feature for quality assessment. We use the autocorrelation method to provide an estimate of the pitch length for the frames marked as voiced. The speech signal is divided into 20 ms frames with 50% overlap using the Hann window. The autocorrelation function is then calculated and normalized for each 20 ms frame. The current nth speech frame is marked as voiced when the second maximum peak of the normalized autocorrelation exceeds 0.5. This extraction method is summarized in Fig. 2. The corresponding pitch length $T(n)$ is obtained by computing the time distance from the origin to the peak.

2.2 Linear prediction analysis

As the degree of severity of abnormal vocal quality becomes higher, the speech signal tends to have more and more aperiodic, irregular and noncoherent components. This has been observed for pathological voices in sustained vowels (Lee and Hahn 2009). The linear prediction (LP) analysis performed in Lee and Hahn (2009) has been used to derive high order statistics (skewness and kurtosis) from the LP residual signal from each frame of the sustained vowel signal. Since continuous pathological voices may contain voiced and/or unvoiced frames, we propose to perform the LP analysis only on voiced frames. In fact, voiced frames are quite quasi-periodic which suggests the value of using an Auto Regressive (AR) filter to model the production of each speech frame.

The Levinson–Durbin algorithm is used to derive an 18th-order all pole LP model for each 20 ms frame marked as voiced by the preprocessing done in Sect. 2.1. The model is characterized by a set of 18 LP coefficients $\{a_i(n)\}_{1\le i\le 18}$ where n denotes the frame number.

2.2.1 Cepstral coefficients

Cepstral coefficients are the coefficients of the inverse Fourier transform representation of the log magnitude of the spectrum of the signal. Once LP coefficients are obtained, it is possible to directly extract cepstral coefficients from them. Assume we want to extract $p < 18$ cepstral coefficients from the obtained 18 LP coefficients $\{a_i(n)\}_{1\le i\le 18}$ then we use the following formula:

$$c_i(n) = a_i(n) + \sum_{l=1}^{i-1} \frac{l}{i} c_l(n) a_{i-l}(n), \quad 2 \le i < p, \tag{1}$$

where $c_1(n) = r_{xx}(0)$ representing the maximum autocorrelation of the nth frame of the speech signal. In this work we extracted $p = 5$ cepstral coefficients per frame.

2.2.2 LP residual

LP residual may bring information on the abnormal behaviour of the voice and speech production system (vocal folds,

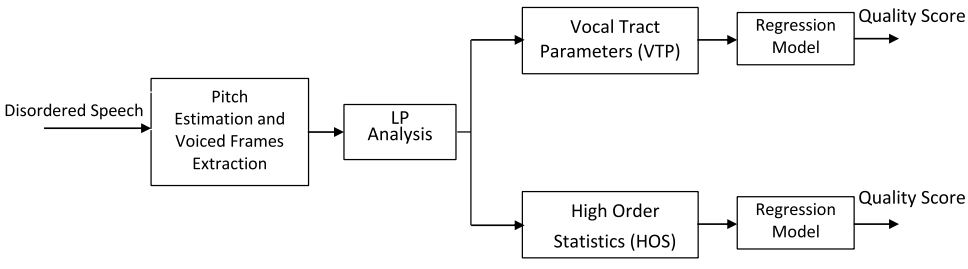

Fig. 1 The proposed speech quality algorithm

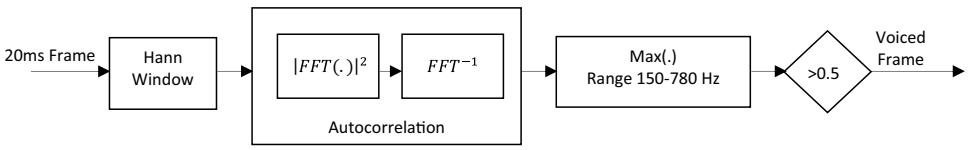

Fig. 2 Pitch period estimation and voiced frames extraction method using the autocorrelation method

vocal tract, turbulence noise...etc) which could be used for disordered voice and speech quality assessment (Lee and Hahn 2009). LP residual represents the error between the original signal and the synthesized (estimated) signal using the derived LP coefficients. The residual of the LP analysis for the nth voiced frame is obtained as

$$e_n(k) = x_n(k) - \sum_{i=1}^{18} a_i(n) x_n(k-i), \tag{2}$$

where $x_n(k)$ represents the value of the original signal at the kth sample of the nth frame. Once the LP analysis has been performed on each voiced frame of the speech signal, we derive different quality features as detailed in the following subsections.

2.3 Vocal tract modelling

This speech assessment block focuses on the voice and speech production system. The human voice production system is composed of an air source (lungs), a modulator (vocal folds) and a resonating system (vocal tract). Airflow created by the lungs excites the vocal cords to generate either a voiced sound or an unvoiced sound (also called voiceless sound). During voiced sounds, a low-frequency (quasi-periodic) sound is generated. The vocal tract acts as a filter that shapes the spectral content of the sound. Controlled contractions and relaxations of the vocal tract muscles change the shape of the vocal tract, and thus its resonant frequencies, to produce the different voiced sounds. During unvoiced sounds a turbulent, a periodic excitation is created by forcing air through a constriction in the vocal tract, for example, when the tongue is placed between the teeth.

In Gray et al. (2000), vocal tract models are used to design a non-intrusive speech quality assessment method that was later implemented in the ITU-T P.563 standard used in telecommunication (Malfait et al. 2006). The idea is to model the vocal tract as a set of acoustic tubes (with uniform cross-section area) arranged in a series configuration, see Fig. 3. Each segment of the tube has a different cross-sectional that changes over time. The idea is to use Linear Prediction (LP) to extract the reflection coefficients and the tube section areas for voiced speech frames. The number of tubes is equal to the order of the LP (number of LP coefficients). In Malfait et al. (2006), the vocal tract is modelled as eight concatenated acoustic tubes which is suitable for narrowband signals sampled at 8 kHz. In our work, we model the vocal tract using a series of 18 acoustic tubes (LP order equals 18) which is suitable for wideband signals associated with the disordered speech. This justifies our approach in using a vocal tract model to extract TE voice/speech quality features.

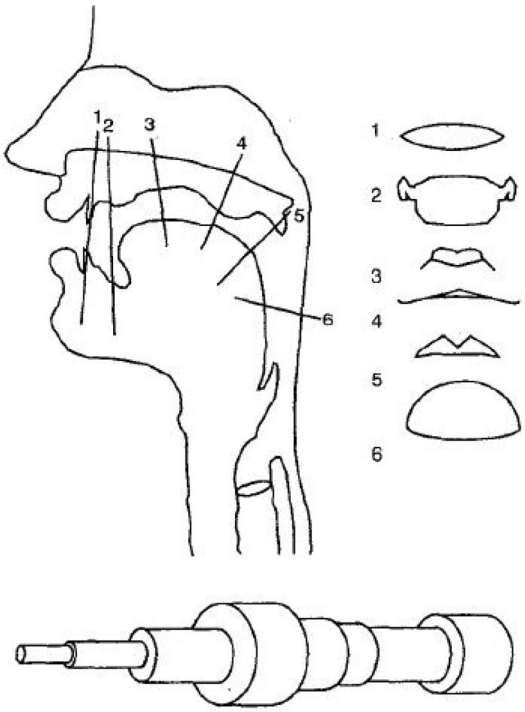

Fig. 3 Illustration of the vocal tract uniform-cross-sectional-area tube model (Gray et al. 2000). Top: true cross-section shapes of the vocal tract sketched at different locations. Bottom: a simplified uniform-cross-sectional-area tube model (with 8 tubes) of the vocal tract. In this work we consider a tube model with 18 acoustic tubes

For each voiced frame of the signal, the reflection coefficients are calculated from the LP coefficients using the following recursion:

$$r_i(n) = \alpha_{i,i}(n), \quad 1 \le i \le 18, \tag{3}$$

$$\alpha_{i-1,l}(n) = \frac{\alpha_{i,l}(n) - r_i(n)\alpha_{i,i-l}(n)}{1 - r_i(n)^2}, \quad 1 \le l < i, \tag{4}$$

such that $\alpha_{18,i} = a_i(n)$ corresponding to the ith coefficient for the LP model of the nth frame. Once the reflection coefficients $\{r_i(n)\}_{1 \le i \le 18}$ are extracted, the cross section areas can be computed using the recursion:

$$S_i(n) = \frac{1 + r_i(n)}{1 - r_i(n)} S_{i+1}(n), \quad i = 18, 17, \ldots, 1. \tag{5}$$

The cross section area S_{18} can be obtained by letting $S_{19} = 1$.

2.4 Features extracted

Based on the above LP analysis and vocal tract modelling, we derive two groups of features which will allow us to assess the quality of our TE speaker samples.

2.4.1 Higher-order statistics

High-order statistics (HOS) analysis has been used in classification of pathological voices (Alonso et al. 2001) and in robust voice activity detection (Nemer et al. 2001) with very promising results. It has the advantage of not requiring a periodic or quasiperiodic voice signal to permit a reliable analysis.

Given a real vector $x = \{x_k\}_{1\le k\le K}$ we define its HOS (mean, standard deviation, skewness and kurtosis) as follows:

$$\mu_x = \frac{1}{K}\sum_{k=1}^{K} x_k,$$

$$\sigma_x = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(x_k-\mu_x)^2},$$

$$\gamma_x = \frac{\frac{1}{K}\sum_{k=1}^{K}(x_k-\mu_x)^3}{\sigma_x^3},$$

$$\kappa_x = \frac{\frac{1}{K}\sum_{k=1}^{K}(x_k-\mu_x)^4}{\sigma_x^4}.$$

In this work, we derive 12 HOS for each frame of the speech signal by considering the 4 HOS (mean, variance, skewness and kurtosis) of the LP coefficients $\{a_i(n)\}_{1\le i\le 18}$, the cepstral coefficients $\{c_i(n)\}_{1\le i\le 5}$ and the LP residual signal $\{e_i(n)\}_{1\le i\le N}$ where N is the number of speech samples within one frame and n is the corresponding frame index. The 12 HOS statistics are averaged across all the voiced frames to yield the features $\mathtt{HOS}_1$,...,$\mathtt{HOS}_{12}$.

To this group of features, we add the $\mathtt{HOS}_{13}$ feature which is computed by taking the average of the different pitch lengths $T(n)$ for all the voiced speech frames. Also the number of voiced frames is taken as a quality feature and denoted $\mathtt{HOS}_{14}$.

To illustrate the dependencies of these high order statics on the voice/speech quality, we consider the mean value of the LP coefficients, denoted $\mu_a(n)$, for the nth frame. The transfer function of the all poles LP model, for a given frame is given by

$$H_n(z) = \frac{1}{1+\sum_{i=1}^{18} a_n(i)z^{-i}}. \tag{6}$$

Therefore, one has

$$\mu_a(n) = \frac{1}{18}\sum_{i=1}^{18} a_n(i) = \frac{1-H_n(1)}{18H_n(1)}. \tag{7}$$

This implies that the mean of the LP coefficients $\mu_a(n)$ will increase as the value of the DC-gain $H_n(1)$ decreases. For TE speech samples, it is observed that the voiced segments of the speech produced by by TE patients will tend to have a gain attenuation (lower values of $H_n(1)$) as the quality of the speech signal gets worse (see Fig. 4). Therefore, the average

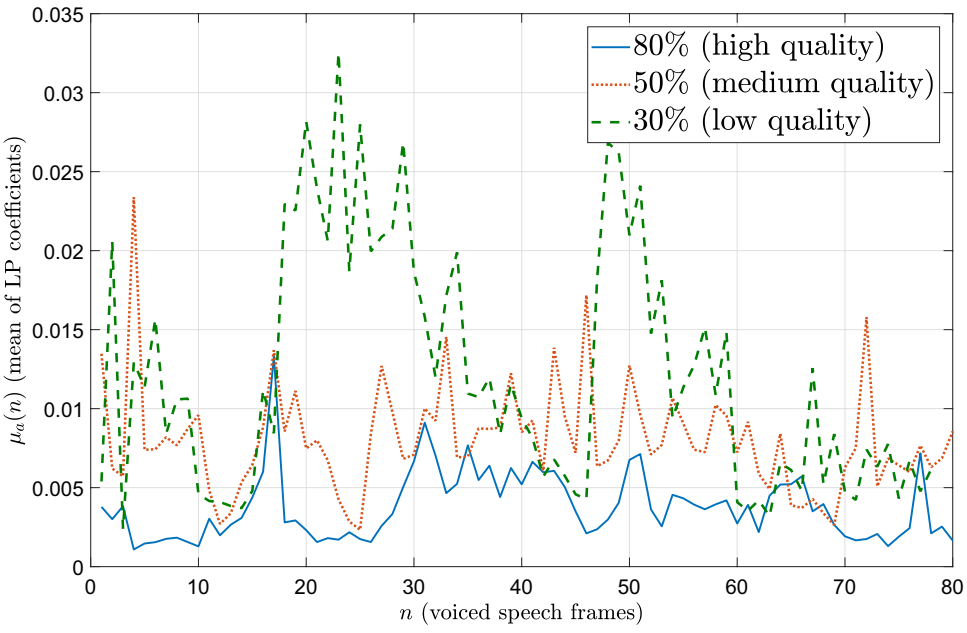

Fig. 4 Average value of LP coefficients for each voiced TE speech frame

of $\mu_a(n)$ across all frames is likely to be inversely proportional to the overall quality of the speech.

2.4.2 Vocal tract parameters

The second group of voice/speech quality features is based on the vocal tract modelling done in Sect. 2.3. To extract quality features from the instantaneous vocal tract model we use the idea that, due to the removal of the larynx, TE speech can be thought to have an "imperfect" speech production system. In this work we wanted to extract as many voice features as possible. We consider the maximum, minimum and average of each cross cross-sectional area which results in $18 \times 3 = 54$ different features. These features were assigned the labels $\texttt{VTP}_1, \ldots, \texttt{VTP}_{54}$ and are defined as follows:

$$\texttt{VTP}_i = \max_n(S_i(n)) \tag{8}$$

$$\texttt{VTP}_{i+18} = \min_n(S_i(n)) \tag{9}$$

$$\texttt{VTP}_{i+36} = \text{avg}_n(S_i(n)) \tag{10}$$

for $i \in \{1, \ldots, 18\}$. The extracted features are then feed to different models which are fitted and compared using advanced regression analysis performed on a TE disordered speech database as detailed in the next section.

3 Speech database

We used a database of 35 TE speech recordings. The speech samples were recorded from adult patients (males and females) with an age range of 45–65 years. All patients have undergone total laryngectomy and TE puncture at least one year prior to their participation. All recordings were gathered in a sound-treated environment using stereo recordings at 44.1 kHz sampling rate with 16-bit quantization. The sentence "*The rainbow is a division of white light into many beautiful colors*" was recorded from all the speakers and used for acoustic and perceptual measurements. The TE speech samples were played back to different groups of naive listeners who have no prior exposure to TE speech. The signals were played back in a random order and 38 listeners were instructed to rate the overall perceived quality on a scale from 1 (low quality) to 10 (high quality). The average of listener ratings was then used to determine the speech sample with the best perceptual rating and in the computation of correlation coefficients between objective and subjective ratings.

4 Results

The features extracted from the vocal tract modelling ($\texttt{VTP}_1, \ldots, \texttt{VTP}_{54}$) and from the high-order statistics ($\texttt{HOS}_1, \ldots, \texttt{HOS}_{14}$) are used to train different regression models. First, for each group of features, forward stepwise regression (FSR) (Stolzenberg 2004) is performed to prioritize the features within the group. Initially no predictors are included in the model. Then, at a first step, we check all the possible models with one predictor against the coefficient of determination R^2 (R squared)

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{11}$$

where the y_i's are the subjective scores (true observations), $\hat{y}_i$'s are the estimation scores and $\bar{y}$ is the mean value of the y_i's data. Then, the feature that gives a model with the highest R^2 is retained. The second step consists in checking all the models with two features by adding another feature to the previously selected feature. This procedure is repeated until we select all the available features. Note that the FSR algorithm stops also if the value of R^2 reaches 1, and in this case the remaining features are discarded. Finally, we obtain a natural ordering of the features by their importance. These results are provided in Table 1.

For example, if we want to use a model with 3 HOS features then the best set of 3 features (from the set of 14 features) is $\texttt{HOS}_5, \texttt{HOS}_9, \texttt{HOS}_{11}$. Similarly, a model with 3 VTP features would contain $\texttt{VTP}_5, \texttt{VTP}_{20}$ and $\texttt{VTP}_4$. Note that the FSR algorithm has stopped after selecting 34 features (out of 54 features) because the value of R^2 reached 1 and the addition of any other features will not bring further information.

Then, we use K-folds cross validation method (Picard and Cook 1984) to select the best set of features that guarantees the lowest prediction error (test error). This allows to avoid the problem of overfitting. For each number of selected features (obtained from the FSR), we use a 7-folds cross validation by training and testing support vector machines regression models (Cortes and Vapnik 1995) with two different kernel functions: linear and Gaussian.

Table 1 Forward stepwise regression results

R^2	Added HOS feature	R^2	Added VTP feature
0.1469	HOS_{11}	0.382	VTP_5
0.260	HOS_9	0.480	VTP_{20}
0.435	HOS_5	0.560	VTP_4
0.473	HOS_2	0.663	VTP_{38}
0.546	HOS_{13}	0.712	VTP_{18}
0.560	HOS_{10}	0.739	VTP_{24}
0.657	HOS_1	0.762	VTP_8
0.679	HOS_3	0.783	VTP_{22}
0.708	HOS_6	0.804	VTP_{21}
0.731	HOS_7	0.831	VTP_{35}
0.744	HOS_4	0.860	VTP_{34}
0.761	HOS_{12}	0.870	VTP_{54}
0.772	HOS_{14}	0.878	VTP_{28}
0.803	HOS_8	0.886	VTP_{36}
		0.893	VTP_{50}
		0.900	VTP_{51}
		0.926	VTP_{46}
		0.939	VTP_{14}
		0.947	VTP_{17}
		0.965	VTP_{33}
		0.973	VTP_7
		0.980	VTP_6
		0.983	VTP_{26}
		0.988	VTP_{37}
		0.990	VTP_{53}
		0.991	VTP_1
		0.993	VTP_{39}
		0.994	VTP_{23}
		0.998	VTP_9
		0.999	VTP_3
		0.999	VTP_{49}
		0.999	VTP_{52}
		0.999	VTP_{47}
		1	VTP_2

Figures 5 and 6 plot the out-of-sample mean square error (MSE) for each cross-validated model resulted from the selected features for the HSO predictors group and the VTP predictors group, respectively. From these figures we can determine the set of features from each group that minimizes the out-of-sample MSE. These sets of features are given in Table 2 for each group and each kernel function.

Table 2 Selected features for each model

Model	Selected features
HOS statistics	
Linear	HOS_1, HOS_2, HOS_3, HOS_4, HOS_5, HOS_6, HOS_7, HOS_9, HOS_{10}, HOS_{11}, HOS_{12}, HOS_{13}
Gaussian	HOS_5, HOS_9, HOS_{11}
VTP parameters	
Linear	VTP_4, VTP_5, VTP_8, VTP_{18}, VTP_{20}, VTP_{21}, VTP_{22}, VTP_{24} VTP_{34}, VTP_{35}, VTP_{38}, VTP_{54}
Gaussian	VTP_4, VTP_5, VTP_{20}, VTP_{38}

Once, the sets of features are selected, each set of features is used to train a model (linear or Gaussian). The data set consists of 35 recordings and is divided into two separate groups. The first group contains 25 recordings and serves as a training set to train the regression model, while the other ten recordings are used to test the prediction capabilities of this regression model. The performance of our proposed algorithms is evaluated using the Pearson's correlation coefficient (Pearson 1895) which measures the linear dependence between the objective measures, x, and the subjective voice quality ratings, y, as

$$\text{Correlation} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}},$$

where $\bar{x}$ is the mean of the objective measures x_i's, $\bar{y}$ is the mean of the subjective measures y_i's and $N = 35$ is the number of speech samples.

Table 3 shows the results obtained from the proposed objective metrics. Applying support vector regression (SVR) with linear kernel to the selected HOS features yields a correlation of 0.89 with the training dataset samples, while a correlation of 0.78 is obtained with the test dataset. Using the SVR technique with a Gaussian kernel to get an objective model for the selected HOS features has a slightly weaker performance in terms of prediction capabilities and overfitting avoidance. The correlation values are 0.78 and 0.63 for the training and the test datasts respectively. Applying SVR model with a linear kernel to the vocal tract VTP features led to a better performance in terms of overfitting avoidance and

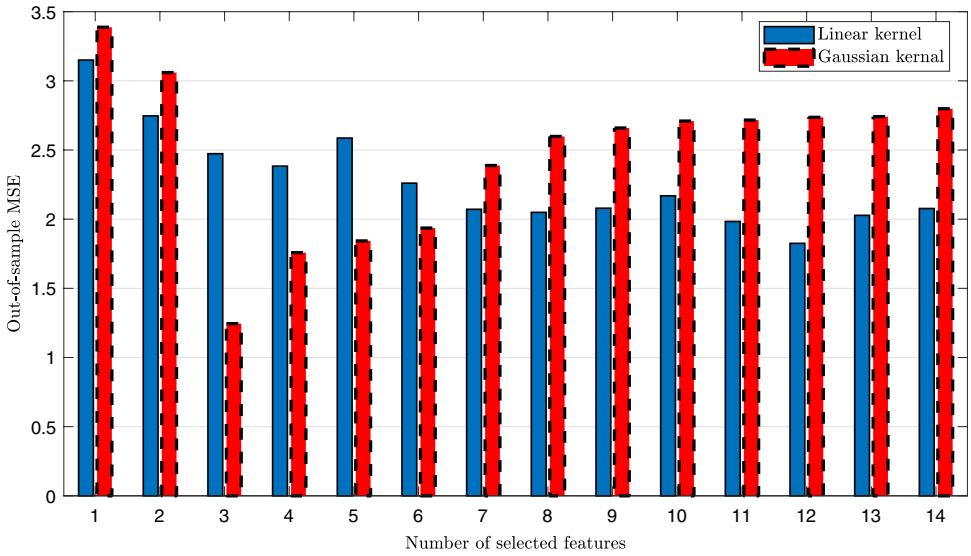

Fig. 5 Feature selection from the HOS statistics group

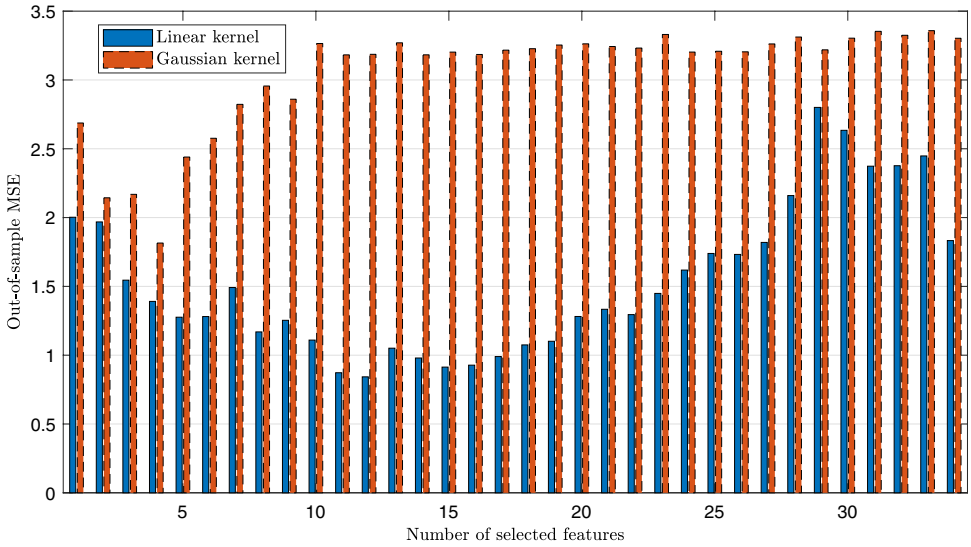

Fig. 6 Feature selection from VTP parameters group

Table 3 Correlation values of the proposed objective metrics

Metric	Correlation (training dataset)	Correlation (test dataset)
HOS-Linear	0.89	0.78
HOS-Gaussian	0.78	0.63
VTP-Linear	0.93	0.84
VTP-Gaussian	0.98	0.70

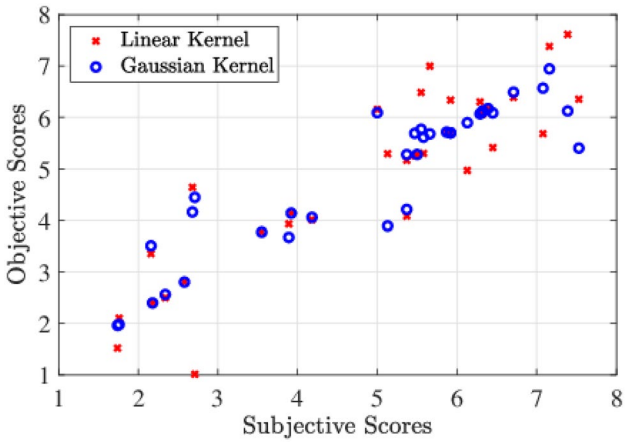

Fig. 7 Scatter plot of subjective scores against the objective scores derived from the VTP parameters-based model

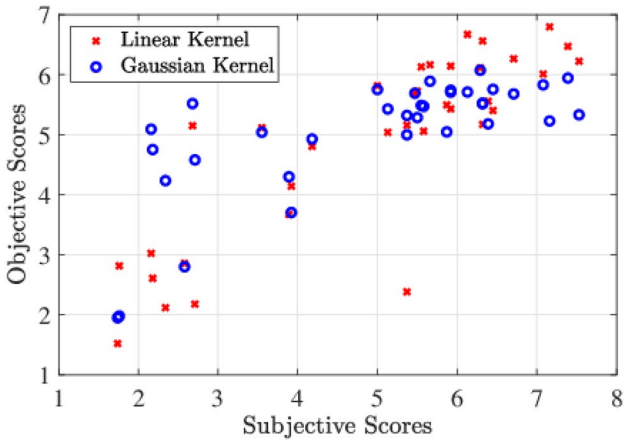

Fig. 8 Scatter plot of subjective scores against the objective scores derived from the HOS statistics-based model

bias minimization. The correlation values for the traininig and the test datasets were 0.93 and 0.84. Changing the kernel to Gaussian has increased the training correlation to 0.98 while decreasing the testing correlation to 0.70. Figures 7 and 8 shows the scatter plot of objective scores against subjective scores for the each of VTP- and HOS-based metrics. These results suggest that an SVR model with linear kernel would perform better than an SVR model with Gaussian kernel although the latter uses less number of features as shown in Table 2. Also, the VTP-based models have performed slightly better than the HOS-based features which shows that features extracted from the vocal tract modelling (speech production system) consist of good predictors for disordered speech quality estimation. The obtained correlation results for the proposed algorithms are much better than the correlation obtained from previously proposed features in the literature such as the Harmonics-to-Noise-Ratio (HNR), Cepstral Peak Prominence (CPP), the ITU-T recommendation P.563 amongst others, see Table 4.

5 Conclusion

This paper introduces a new nonintrusive algorithm, with low computational complexity, suitable for disordered speech quality estimation. Using an 18-order LP analysis applied to voiced frames of the acoustic speech signal, we derived up to 14 high-order statistical (HOS) based features and 54 vocal tract parameters (VTP) based features. We used a set of 35 TE speech samples to train different support vector regression models after performing features selection using forward stepwise regression and K-folds cross validation. The obtained models are shown to be able to predict the quality scores of the subjective scores with a correlation coefficient than ranges from 0.78 to 0.98 for the training dataset and from 0.63 to 0.84 for the test dataset. The obtained results of this paper suggest that the HOS and VTP features, which are extracted from a simple LP analysis of the acoustic speech signal, can be an efficient and effective alternative to the more complex existing nonintrosive algorithms for quality estimation of pathological voice samples.

Table 4 Comparison of the correlation values obtained using different quality estimation methods

Algorithm	Correlation
Literature	
Voice breaks	0.32
Harmonic-to-Noise-Ratio (HNR) (Awan and Frenkel 1994)	0.23
Cepstral Peak Prominence (CPP) (Maryn et al. 2009)	0.54
CPP (smoothed) (Maryn et al. 2009)	0.20
ITU-T P.563 (Malfait et al. 2006)	0.22
Our work	
HOS-Linear	0.85
HOS-Gaussian	0.73
VTP-Linear	0.90
VTP-Gaussian	0.91

Acknowledgements Funding from the Natural Sciences and Engineering Research Council of Canada is gratefully acknowledged.

References

Ali, Y., Parsa, V., Doyle, P., & Berkane, S. (2017). Disordered speech quality estimation using the matching pursuit algorithm. In *The 30th annual IEEE Canadian conference on electrical and computer engineering*.

Alonso, J. B., De Leon, J., Alonso, I., & Ferrer, M. A. (2001). Automatic detection of pathologies in the voice by HOS based parameters. *EURASIP Journal on Applied Signal Processing*, *4*, 275–284.

Awan, S. N., & Frenkel, M. L. (1994). Improvements in estimating the harmonics-to-noise ratio of the voice. *Journal of Voice*, *8*(3), 255–262.

Awan, S. N., Roy, N., Jetté, M. E., Meltzner, G. S., & Hillman, R. E. (2010). Quantifying dysphonia severity using a spectral/cepstral-based acoustic index: Comparisons with auditory-perceptual judgements from the cape-v. *Clinical Linguistics & Phonetics*, *24*(9), 742–758.

Beerends, J. G., Schmidmer, C., Berger, J., Obermann, M., Ullmann, R., Pomy, J., et al. (2013). Perceptual objective listening quality assessment (POLQA), the third generation ITU-T standard for end-to-end speech quality measurement part i—Temporal alignment. *Journal of the Audio Engineering Society*, *61*(6), 366–384.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Eadie, T. L., & Doyle, P. C. (2002). Direct magnitude estimation and interval scaling of naturalness and severity in tracheoesophageal (te) speakers. *Journal of Speech, Language, and Hearing Research*, *45*(6), 1088–1096.

Eadie, T. L., & Doyle, P. C. (2005). Scaling of voice pleasantness and acceptability in tracheoesophageal speakers. *Journal of Voice*, *19*(3), 373–383.

Grancharov, V., Zhao, D. Y., Lindblom, J., & Kleijn, W. B. (2006). Low-complexity, nonintrusive speech quality assessment. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*(6), 1948–1956.

Gray, P., Hollier, M., & Massara, R. (2000). Non-intrusive speech-quality assessment using vocal-tract models. *IEEE Proceedings on Vision, Image and Signal Processing*, *147*(6), 493–501.

Gu, L., Harris, J. G., Shrivastav, R., & Sapienza, C. (2005). Disordered speech assessment using automatic methods based on quantitative measures. *EURASIP Journal on Advances in Signal Processing*, *2005*(9), 768125.

Hirano, M. (1981). *Clinical examination of voice* (Vol. 5). New York: Springer.

Kates, J. M., & Arehart, K. H. (2010). The hearing-aid speech quality index (HASQI). *Journal of the Audio Engineering Society*, *58*(5), 363–381.

Kempster, G. B., Gerratt, B. R., Abbott, K. V., Barkmeier-Kraemer, J., & Hillman, R. E. (2009). Consensus auditory-perceptual evaluation of voice: Development of a standardized clinical protocol. *American Journal of Speech-Language Pathology*, *18*(2), 124–132.

Lee, J., & Hahn, M. (2009). Automatic assessment of pathological voice quality using higher-order statistics in the LPC residual domain. *EURASIP Journal on Advances in Signal Processing*,. https://doi.org/10.1155/2009/748207.

Malfait, L., Berger, J., & Kastner, M. (2006). P. 563–The ITU-T standard for single-ended speech quality assessment. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*(6), 1924–1934.

Maniglia, A. J., Lundy, D. S., Casiano, R. C., & Swim, S. C. (1989). Speech restoration and complications of primary versus secondary tracheoesophageal puncture following total laryngectomy. *The Laryngoscope*, *99*(5), 489–491.

Maryn, Y., Roy, N., De Bodt, M., Van Cauwenberge, P., & Corthals, P. (2009). Acoustic measurement of overall voice quality: A meta-analysis. *The Journal of the Acoustical Society of America*, *126*(5), 2619–2634.

Nemer, E., Goubran, R., & Mahmoud, S. (2001). Robust voice activity detection using higher-order statistics in the LPC residual domain. *IEEE Transactions on Speech and Audio Processing*, *9*(3), 217–231.

Parsa, V., & Jamieson, D. G. (2001). Acoustic discrimination of pathological voice: Sustained vowels versus continuous speech. *Journal of Speech, Language, and Hearing Research*, *44*(2), 327–339.

Pearson, K. (1895). Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, *58*, 240–242.

Picard, R. R., & Cook, R. D. (1984). Cross-validation of regression models. *Journal of the American Statistical Association*, *79*(387), 575–583.

Rabiner, L., Cheng, M., Rosenberg, A., & McGonegal, C. (1976). A comparative performance study of several pitch detection algorithms. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *24*(5), 399–418.

Ritchings, R., McGillion, M., & Moore, C. (2002). Pathological voice quality assessment using artificial neural networks. *Medical Engineering & Physics*, *24*(7), 561–564.

Rix, A. W., Beerends, J. G., Hollier, M. P., & Hekstra, A. P. (2001). Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs. In *IEEE international conference on acoustics, speech, and signal processing* (pp. 749–752).

Robbins, J., Fisher, H. B., Blom, E. C., & Singer, M. I. (1984). A comparative acoustic study of normal, esophageal, and tracheoesophageal speech production. *Journal of Speech and Hearing disorders*, *49*(2), 202–210.

Stolzenberg, R. M. (2004). Multiple regression analysis. *Handbook of Data Analysis*, *165*, 208.

Union, I. T. (1996). ITU-T recommendation P.800: Methods for subjective determination of transmission quality. *International Telecommunication Union*.

Ward, E. C., & van As-Brooks, C. J. (2014). *Head and neck cancer: Treatment, rehabilitation, and outcomes*. San Diego: Plural Publishing.