A Data-Driven Non-intrusive Measure of Speech Quality and Intelligibility

Dushyant Sharma, Yu Wang, Patrick A. Naylor, and Mike Brookes

***Abstract*—Speech signals are often affected by additive noise and distortion which can degrade the perceived quality and intelligibility of the signal. We present a new measure, NISA, for estimating the quality and intelligibility of speech degraded by additive noise and distortions associated with telecommunications networks, based on a data driven framework of feature extraction and tree based regression. The new measure is non-intrusive, operating on the degraded signal alone without the need for a reference signal. This makes the measure applicable to practical speech processing applications operating in the single-ended mode. The new measure has been evaluated against the intrusive measures PESQ and STOI. The results indicate that the accuracy of the new non-intrusive method is around 90% of the accuracy of the intrusive measures, depending on the test scenario. The NISA measure therefore provides non-intrusive (single-ended) PESQ and STOI estimates with high accuracy.**

I. Introduction

SPEECH acquisition and communication systems deployed in real-world (non-ideal) conditions suffer from various degradations, reducing the perceived quality and intelligibility of the speech signal. In the field of law enforcement audio collection, severely degraded audio is commonplace, reducing the intelligence value of the audio by making it inadmissible as evidence in a court of law [1]. In less extreme conditions, a loss in quality is encountered which has adverse effects on transcription speed resulting in higher transcription costs [2]. In the telecommunications sector, a related issue is the delivery of a consistent Quality of Service (QoS) to the consumer which could be achieved by exploiting measurement of the speech signal quality in order to adjust system parameters in the service infrastructure. The widespread use of mobile communication devices means that such systems are increasingly being used in highly variable and often adverse operating conditions, resulting in a loss in perceived quality for which customer experience monitoring might be very important. We therefore identify the need for a non-intrusive objective metric of the speech signal quality for the following tasks: (a) aiding or guiding the collection and transcription of law enforcement audio, (b) to help optimize the performance of speech telecommunication systems in general and (c) to enable selection of data to provide appropriately constructed corpora for training of speech, speaker and language recognition systems.

Speech quality is a judgment of a perceived multidimensional construct that is internal to the listener and is typically considered as a mapping between the desired and observed features [3]. The quality of a speech signal can be obtained in a listening test with a number of human subjects (subject-based methods) or algorithmically (instrumental methods). A popular scale for measuring speech quality is the 5-point Absolute Category Rating (ACR) scale proposed in the International Telecommunications Union (ITU) P.800 standard [4]. Speech quality scores from a number of subjects are typically averaged to obtain the Mean Opinion Score (MOS). Subject-based tests are time consuming and expensive to administer for larger amounts of audio data [5]. Instrumental methods for speech quality assessment aim to overcome these issues by modeling the relationship between the desired and perceived characteristics of the signal algorithmically.

There are three main categories of objective speech quality assessment: (a) those which require a reference signal in addition to the signal under test are referred to as intrusive methods, (b) those that operate only on the signal under test are referred to as non-intrusive methods and (c) those that rely on the parameters of the processing system are commonly referred to as parametric methods [6]. An ITU industry standard for intrusive quality testing is the Perceptual Evaluation of Speech Quality (PESQ) [7] metric, which is an integration of two previous intrusive methods [8]: an extended version of PSQM [9] and the PAMS method [10], [11]. In PESQ, quality scores are determined on a scale from -0.5 to 4.5 and a mapping function is then used to map the PESQ score to MOS on the ACR scale [12]. A correlation coefficient of 0.935 between PESQ and MOS has been reported for a number of telecommunication relevant databases [13], [8]. More recently, an extension of PESQ has been standardized as POLQA [14].

In situations where the reference signal is not available, a non-intrusive technique may be applied. A number of non-intrusive techniques have been proposed over the past decade, see [13], [3] for a review. The current ITU-T industry standard algorithm for non-intrusive speech quality assessment is P.563 [15], which uses a number of features from the audio stream to estimate the quality score directly on the MOS scale. More recently a number of techniques that use machine learning methods such as Gaussian mixture models (GMMs) to model perceptual speech features have been proposed by Falk *et al.* [16], [17], [18]. Additionally, speech quality metrics based on a data-mining approach using CART have also been developed [19], [20], [21]. The Low Complexity Speech Quality Assessment (LCQA) algorithm [22] derives a number of features from the speech signal and has been shown to outperform the P.563 metric for a large set of degradations. Due to its low complexity, novel feature set and favorable

Mike Brookes, Patrick A. Naylor and Dushyant Sharma were with the Centre of Law Enforcement Audio Research (CLEAR) at Imperial College London, UK, during the course of this work. E-mail: p.naylor@imperial.ac.uk.

performance this method is selected for baseline comparison along with the P.563 method in this paper.

Speech intelligibility can be defined as a measure of the proportion of a speech signal's content correctly recognised by a listener. It is an important quantifier for applications such as telecommunications, where the quality of a channel may be evaluated in terms of its effect on speech intelligibility [23], [24], as a performance metric for hearing aids [25], for determining the impact of an acoustic space on speech [26] and for intelligence gathering in law enforcement applications [2]. Subjective speech intelligibility scores are obtained through listening experiments where subjects listen to speech samples and their performance in a particular linguistic task is measured. The linguistic task may be to recognize nonsense syllables [27], [28], isolated words [29], [30] or a number of key-words in a sentence [31], [32]. Objective intelligibility assessment methods operate without the need for human subjects. One of the earliest intrusive intelligibility metrics was proposed by French and Steinberg [33] as the Articulation Index (AI), which was further refined with improved methods for calculating the AI by Kryter in 1962 [34] and finally led to an ANSI standard in 1969 [35]. The AI was further developed into the Speech Intelligibility Index (SII) and led to an ANSI standard for intrusive intelligibility assessment in 1997 [36]. The SII evaluates the effects of degradations in a number of frequency bands, weighted by their importance to speech intelligibility and quantifies the proportion of the speech signal that is audible to the listener. The SII score is monotonically related to the intelligibility of speech and is given in the range 0 to 1 (where a score of 0.5 means that half of the speech cues are audible and usable to the listener) [37]. More recently, the Short-Time Objective Intelligibility (STOI) method for intrusive intelligibility assessment has been proposed which has been shown to have a high correlation (greater than 0.92) with subjective intelligibility scores for noisy and noise-suppressed speech [38], [39], [40]. A non-intrusive method for speech quality and intelligibility estimation of reverberant and de-reverberated speech was presented in [41] and a method for predicting the intelligibility of reverberant speech for cochlear implant listeners was presented in [42]. A recent method for non-intrusive estimation of the C50 room acoustic parameter was presented in [43] and shown to correlate with speech quality and speech recognition performance. The low complexity speech intelligibility algorithm (LCIA) is a data-driven non-intrusive measure that has been shown to have a high per-condition correlation with subjective intelligibility scores of noisy and noise suppressed speech [44]. However, this method is only validated with a single speaker using condition averaging and a limited number of degradation conditions. LCIA will be further evaluated in this paper.

A. Overview

We propose a data-driven method for non-intrusive speech quality and intelligibility assessment, referred to as NISA. Since a key requirement for developing such methods is large quantities of labeled data, an automatic method for labeling large databases based on intrusive techniques is also proposed. We then apply the non-intrusive NISA method to estimate intrusive measures, specifically PESQ for speech quality and STOI for speech intelligibility. In this paper we limit our discussion to the case of additive noise and the effects of speech codecs. Other degradations such as reverberation and effects of noise suppression are not considered here but the reader is referred to the appropriate literature elsewhere on non-intrusive reverberation estimation such as [43]. The remainder of this paper is organized as follows, in Section II we review the non-intrusive LCQA method [22] and describe LCIA, a development of LCQA for non-intrusive speech intelligiblity assessment. This is followed by a description of the new NISA method in Section III. The test and training databases and evaluation metrics are outlined in Section IV. The results for the PESQ and STOI estimation experiments are presented in Sections V and VI respectively, followed by the conclusions in Section VII.

II. Background to data-driven speech assessment methods

The LCQA method [22] is a data driven approach to non-intrusive speech quality assessment and has been shown to outperform P.563 [15] for a number of speech databases. The algorithm begins with a pre-processing stage that splits the input signal into 20 ms non-overlapping frames for further processing. The algorithm extracts 11 features per-frame that include the pitch period and spectral features derived from a 10^{th} order LPC analysis of the speech signal. The next step is a frame selection procedure which applies thresholds on three per-frame features and retains only those frames that satisfy this threshold. This has been described as a generalization of a Voice Activity Detector (VAD) and typically discards between 50% to 80% of the frames [22]. The 11 per-frame features are characterized by their mean, variance, skewness and kurtosis over the entire signal and the resulting 44 dimensional global feature vector is used to perform feature subset selection on labeled training data. The final quality estimate is obtained with a GMM mapping using the dimensionally reduced global features. The GMM has to be trained on appropriate data.

A. LCIA

The Low-Cost Intelligibility Assessment (LCIA) method [44] is a data-driven approach for low-complexity, non-intrusive speech intelligibly assessment. LCIA differs from the LCQA method [22] by employing an additional feature (iSNR), an external VAD, the use of a two-step feature selection and projection technique and training on databases labeled with STOI [38]. The LCIA method begins by deriving per frame features from the speech waveform, then applying a statistical model followed by a two-step dimensionality reduction and GMM mapping. The pitch period is not used as a feature in LCIA due to the poor correlation of this feature with subjective intelligibility scores [44]. The 11 per-frame features are outlined in Table I. The mean, variance, skewness and kurtosis of the per-frame features results in a 44 dimensional feature vector per utterance, which is further reduced by a correlation based feature selection and

principal component analysis (PCA) based feature projection. A joint GMM is trained on the projected features and the intelligibility score for each speech utterance in the training data.

III. Proposed Non-Intrusive Speech Assessment Methods

Our Non-Intrusive Speech Assessment (NISA) uses a data-driven, machine learning approach consisting of a combination of feature extraction followed by a tree-based regression model for speech quality and intelligiblity assessment. Whereas this overall approach is commonly adopted in the literature, as has been discussed in Section II, the novelty in the NISA method lies both in the way NISA extracts features more discriminative in the given task, and in the way it performs the classification of these features for which it employs a Classification and Regression Tree (CART). The first step is the application of a Hamming window to segment the input signal into non-overlapping 20 ms frames, denoted as $y(i)$, where i is the frame index. The second step is application of a VAD, based on the P.56 method [45], to select frames where speech is present. Level normalization is applied so that the feature extraction is independent of the speech level. This is then followed by short-term feature extraction and the mean, variance, skewness and kurtosis are used to characterize the entire signal and combined with long-term features, as described below, to create the final feature vector, $\boldsymbol{\varphi}$ (of dimension 116), for the current signal. The features, $\boldsymbol{\varphi}$, are used to infer a trained CART regression model, that has been previously trained on a feature matrix, $\boldsymbol{\Phi}$, with corresponding ground truth scores from a training database. The features used in the NISA method, and the use of CART, are further described in the following subsections and the overall structure is presented in Fig. 1.

A. Short-term features

The short-term feature extraction follows the time segmentation of the input speech signal into voice-active frames using a voice activity detector based on the P.56 method [45], [46]. The following short-term features are included as defined below: fundamental frequency, iSNR, Hilbert envelope based features, zero crossing rate, LTASS deviation and LPC based features. Additionally, the first order numerical derivatives of some of the features are employed as shown in Table I.

1) Fundamental frequency: Fundamental frequency is an important feature that is used widely in many speech processing algorithms, such as the P.563 [15]. The task of fundamental frequency estimation in low SNR scenarios is a challenging problem, where many pitch estimation algorithms fail [47]. The NISA method makes use of fundamental frequency estimates from the PEFAC algorithm [48] that has been shown to be robust to high levels of additive noise.

2) Importance weighted signal to noise ratio (iSNR): The following additive model for noise is assumed,

$$y(n) = s(n) + v(n)$$

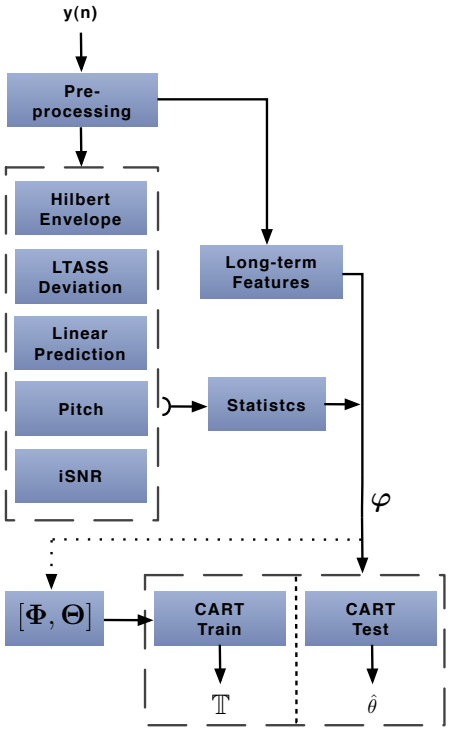


Figure 1. Block diagram of the Non Intrusive Speech Assessment (NISA) method.

where $y(n)$ is the noisy speech signal, $s(n)$ is the clean speech signal, $v(n)$ is the noise signal and $Y(i,k)$ is the Discrete Fourier Transform (DFT) of the noisy signal at time frame i and frequency bin k. The iSNR feature used in NISA is a non-intrusive SNR measure that performs the SNR calculation in short-time frames and also applies a frequency weighting function based on speech intelligibility measurement. The iSNR feature uses the 1/3 octave frequency band importance function from the SII standard ([44], [36]) that applies more weight to frequencies that have a higher importance to speech intelligibility. The iSNR for time frame i is defined as:

$$\text{iSNR}(i) = 10\times\sum_{k=1}^{N_k} I(k)\times\log_{10}\left(\frac{\max(0, P_y(i,k) - P_{\tilde{v}}(i,k))}{P_{\tilde{v}}(i,k)}\right) \quad (1)$$

where $I(k)$ is the SII weighting function, N_k is the number of frequency bands, $P_{\tilde{v}}(i,k) = Y(i,k)Y*(i,k)$ is the estimated noise power spectrum obtained by the minimum statistics algorithm [49], [46] and $P_y(i,k)$ is the power spectrum of the noisy speech signal.

3) Hilbert envelope: The Hilbert decomposition of a signal results in a slowly varying envelope, which has been shown to be an important factor in speech reception [50], and a rapidly varying fine structure component. The envelope for frame i is calculated as:

$$e(i) = \sqrt{y(i)^2 + \mathscr{H}(y(i))^2}, \quad (2)$$

where $\mathscr{H}\{.\}$ is the Hilbert transform. The variance, $\sigma_{e(i)}$ and dynamic range, $\Delta_{e(i)}$ of the envelope for each of the N_i frames are computed as:

$$\sigma_{e(i)} = \frac{1}{N_i}\sum_{i=1}^{N_i}(e(i) - \mu_{e(i)})^2 \quad (3)$$

$$\Delta_{e(i)} = |\max(e(i)) - \min(e(i))|. \quad (4)$$

4) LTASS deviation: The long term average speech magnitude spectrum (LTASS), defined in [51], has a characteristic shape that is often used as a model for the clean speech spectrum and has been used in a number of speech processing algorithms, such as blind channel identification [52]. The Power spectrum of Long term Deviation (PLD) feature for frame i and frequency bin k is defined as:

$$\text{PLD}(i,k) = \log(P_y(i,k)) - \log(P_{LTASS}(k)), \quad (5)$$

where $P_y(i,k)$ is the magnitude power spectrum of noisy signal and $P_{LTASS}(k)$ is the LTASS power spectrum. This deviation spectrum measures the effects on the magnitude spectrum due to the distortion. The per-frame LTASS deviation spectrum is used to derive the spectral flatness (SF), spectral centroid (SC) and spectral dynamics (SD) features as defined below:

$$SF(i) = \frac{\exp\left(\frac{1}{N_k}\sum_{k=1}^{N_k}\log(\text{PLD}(i,k))\right)}{\frac{1}{N_k}\sum_{k=1}^{N_k}\text{PLD}(i,k)}, \quad (6)$$

$$SC(i) = \frac{\sum_{k=1}^{N_k}\tau(k)\times\log(\text{PLD}(i,k))}{\sum_{k=1}^{N_k}\log(\text{PLD}(i,k))}, \quad (7)$$

$$SD(i) = \frac{1}{N_k}\sum_{k=1}^{N_k}(\log(\text{PLD}(i,k) - \log(\text{PLD}(i,k)))^2, \quad (8)$$

where τ is a vector consisting of the centre frequency for the k^{th}DFT bin and N_k is the number of FFT bins. The spectral flatness, dynamics and centroid of LTASS deviation spectrum and their rate of change are included as short-term features.

5) LPC features : A 10th order linear predictive coding (LPC) is performed on the speech signal using the auto-correlation method. The LPC residual variance and its rate of change over the signal are included as features. The spectral centroid, flatness and dynamics of the LPC spectral magnitude and their rate of change are computed as in [22], and and also used as features.

6) Zero crossing rate: The zero crossing rate has been successfully used as a feature for voiced-unvoiced speech and silence classification [53] and is also included here.

B. Long-term features

1) LTASS deviation: The long-term deviation of the magnitude spectrum of the signal (calculated over the entire utterance) is defined as follows

$$P_{LTLD}(k) = \frac{1}{N_i}\sum_{i=1}^{N_i}\text{PLD}(i,k) \quad (9)$$

where k is the frequency index and PLD is the power spectrum of long-term deviation (5). The resulting P_{LTLD} spectrum is then mapped into 16 bins each with a bandwidth of 500 Hz with 50 % overlap. The energy in each bin as a percentage of the total energy is then computed to form the long term features in NISA, as follows

$$\phi_j = \frac{\sum_{g\in w}P_{LTLD}(g)}{\sum_{k=1}^{K}P_{LTLD}(k)},$$

where ϕ_j is the j^{th} global feature and w is a 500 Hz window centered on the frame of interest and the numerator is the energy of the current frame and the numerator is the total energy in the residual spectrum. It is expected that this feature can identify the long-term frequency characteristics of different types of degradations.

C. Classification and Regression Tree

Let N_S and N_f be the number of utterances and features respectively, in any particular experiment. Then an $N_S \times N_f$ feature matrix $\mathbf{\Phi}$ forms the measurement space for the regression problem, for which the CART algorithm [54] is used to construct a regression tree. A vector of observed scores is also available during the training of the regression tree. In CART, the measurement space is recursively partitioned by binary splits resulting in a number of terminal nodes, with each terminal node containing a constant predicted response value [54]. In this use of CART, first, an over-sized and sub-optimal tree is grown using a minimum mean square error (MSE) criterion. This is followed by a pruning process where a 10-fold cross-validation of the training data is used to merge those tree branches that result in the smalllest reduction in MSE.

IV. Evaluation Methodology

Section IV-A outlines the databases used for training and testing the non-intrusive algorithms. As the performance of a machine learning algorithm is influenced by the quality and quantity of training data [55], the development of a large corpus of human labeled training data is an important issue. However, the use of human subjects for labeling data for a particular application such as speech quality is an expensive and time consuming task [13]. The solution we propose is to use an intrusive objective technique to label large amounts of training data automatically. However, it is necessary to ensure that the degradations that are being labeled are appropriate for the intrusive algorithm, which can be achieved by analyzing the performance of the intrusive techniques on databases for which human labeled scores are available. This is further outlined in Section IV-B.

A. Speech Quality and Intelligibility Databases

The training and validation database (referred to here as the TN database) is based on speech from the TIMIT database [56]. The speech material consists of 10 utterances[1]

[1]Here we define an utterance to mean a segment of speech for which the measure of interest is assumed approximately constant. The duration of an utterance should be suitably long as to permit estimation of the various features to be employed. In our work, we typically use utterance durations in the range 3 to 8 seconds. Long speech segments with varying quality can, without loss of generality, be segmented into shorter segments with less variability in the measure of interest.

		LCQA		LCIA		NISA	
Per-frame feature	Feature Id	x	$\Delta(x)$	x	$\Delta(x)$	x	$\Delta(x)$
LPC spectral flatness	ϕ_1	x	x	x	x	x	x
LPC spectral dynamics	ϕ_2	x		x		x	
LPC spectral centroid	ϕ_3	x	x	x	x	x	x
LPC excitation variance	ϕ_4	x	x	x	x	x	x
Speech variance	ϕ_5	x	x	x	x	x	x
Pitch (LPC)	ϕ_6	x	x				
Pitch (PEFAC)	ϕ_7					x	x
Zero crossing rate	ϕ_8					x	x
iSNR	ϕ_9			x	x	x	x
Hilbert envelope variance	ϕ_{10}					x	x
Hilbert envelope range	ϕ_{11}					x	x
Spectral flatness (PLD)	ϕ_{12}					x	x
Spectral dynamics (PLD)	ϕ_{13}					x	x
Spectral centroid (PLD)	ϕ_{14}					x	x

Table I
THE SHORT-TIME FEATURES USED IN THE DATA-DRIVEN ALGORITHMS PRESENTED IN THIS PAPER.

by each of the speakers, of which 8 distinct utterances are uniquely spoken by one speaker and 2 are common to all speakers. In the TN database, only the distinct utterances for all the speakers are used. An extensive additive noise database is then created by adding 15 noise types from the NATO noise database [57] at SNR's in the range -24 to 30 dB in 3 dB steps, with speech level calculated using the P.56 method [45] and referred to in this paper as the ADN-TN database (representing the additive noise partition of the TN database). Additionally, for the purpose of evaluating the effect on quality of degradations found in realistic telecommunication channels, the CTIMIT [58] and NTIMIT [59] databases are also included to represent telecommunications degradations. NTIMIT contains speech from the original TIMIT database transmitted through the telephone network and recorded at the listener end. Similarly, the CTIMIT is transmitted through the cellular network. The TN database is therefore composed of 285 additive noise conditions (due to 19 SNR's $\times$ 15 noises per speaker) and the union of the CTIMIT and NTIMIT databases.

The C-Qual database [60] includes examples of the types and levels of degradations commonly found in law enforcement applications, with speech material composed of the English language partition of the ITU-T P.23 database [61]. The non-intrusive experiments reported in this paper make use the additive noise partition of the C-Qual database, comprising of car, babble and hum noise representing 21 conditions for each of 4 speakers. The C-Qual database is used only as a generalization test database as it contains unseen speech and noise material compared to the training material. All databases were down-sampled to 8 kHz to represent narrow-band speech transmission.

B. Labeling

The labeling of our database for the purpose of speech quality testing is carried out with the intrusive PESQ algorithm [7]. In the case of degradations relevant to law enforcement, PESQ works well for the additive noise conditions [60], achieving a correlation coefficient of 0.93 with subjective quality scores. Since our database contains additive noise and CODEC distortions, the PESQ score is expected to be well correlated with speech quality and is therefore used as the ground truth for the non-intrusive quality experiments.

For an illustration of the performance of current intrusive speech intelligibility techniques we test the STOI method on a database where subjective intelligibility scores are available [62]. The STOI [38], [40] method achieved a Spearman correlation coefficient of 0.94 with subjective word intelligibility scores validating its use for the ADN-TN and additive noise partition of the C-Qual database. The speech intelligibility estimation experiments are restricted to the additive noise partitions of the database for which we aim to predict the output of the intrusive STOI method (which is well correlated with inteligibility) using our non-intrusive estimator.

C. Training

The original TIMIT database is partitioned into a training and test partition, which is maintained throughout the degradation processing described above. The TIMIT training partition contains 455 speakers and the remaining 168 speakers in the test partition. There is no overlap of speech material in the test and training set (different text and different speakers). A similar partitioning is maintained in the TN database, with the training partition consisting of 168 speakers randomly selected from the original 455 speakers in the TIMIT training partition. All data-driven algorithms are trained on the TN training partition and also tested on the entire test partition. Additionally, for the additive noise partition, each noise file is split into a training and test partition. The resulting TN database consists of more than 45 hours of speech material.

D. Metrics

In this section we define the metrics used for measuring the performance of algorithms for speech quality assessment.

1) Spearman Correlation Coefficient (SCC): The Spearman rank correlation coefficient (SCC) is a non-parametric measure that describes the monotonic relationship between two ranked variables [63] in the range -1 to +1.

2) Root Mean Square Error (RMSE): The root mean square error between the estimated and true scores is calculated as a measure of the estimation accuracy of each algorithm.

3) Bin Error : For speech quality assessment, the bin error evaluates the absolute mean residual error in the true and estimated PESQ scores in bins of size 0.25 PESQ (thereby dividing the PESQ scale into 20 bins), similar to the case of MOS prediction [7], [64].

4) Two Class Hit Rate (TCHR): This measure quantifies the hit rate (HR) achieved by splitting the ground truth scores into two classes, seperated by the mid point of the relevant scale. The motivation for this metric is to evaluate the algorithms in terms of an acceptability criteria. For speech quality assessment, the acceptance threshold is set to be the midpoint of PESQ=2.0. For speech intelligibility assessment, the acceptance threshold is set to the STOI score corresponding to 75% intelligibility, as provided by the mapping function proposed in [40] to be a STOI score of 0.62. A threshold of acceptance at 75% was found to be a practical choice in [65]. The TCHR metric is also evaluated at a number of other threshold values to assess how the performance changes with thresholds of 0.5 to 0.7 STOI, corresponding to intelligibly scores covering the range from 28% to 93%.

V. PESQ Estimation Results

This section presents the results for non-intrusive speech quality estimation using PESQ for the P.563, LCQA and NISA algorithms on the databases described in Section IV-A. The methods are trained on the training partition of the TN database and the ground truth is taken to be the intrusively calculated using the PESQ measure as described in Section IV.

A. Performance for TN database

Table II shows the results for non-intrusive speech quality estimation and taking PESQ as the ground truth, on the test partition of the TN database for the metrics previously defined in Section IV-B. The most accurate performance is achieved by the NISA algorithm on all the test metrics, with an SCC of 0.93 (the scatter of this data is presented in Fig. 3) and an RMSE of 0.32. Moreover, 91.2% of the estimation errors for the NISA algorithm are less than 0.5 PESQ score. Figure 2 shows the RMS performance for PESQ estimation as a function of SNR, where the results for each SNR are averaged over all noise types in the TN database. It can be seen that the NISA algorithm has a consistent performance across SNRs, unlike the P.563 and LCQA algorithms, which have a highly variable performance with SNR.

Table II
NON-INTRUSIVE PESQ ESTIMATION PERFORMANCE ON THE TEST PARTITION OF THE TN DATABASE.

			Bin Error				
Method	SCC	RMSE	0.25	0.5	0.75	1.0	TCHR
NISA	**0.93**	**0.32**	**63.8**	**91.2**	**97.5**	**98.9**	**93.2**
LCQA	0.83	0.72	23.6	46.1	66.3	82.9	77.3
P563	0.77	0.94	22.2	42.7	59.3	71.5	82.3

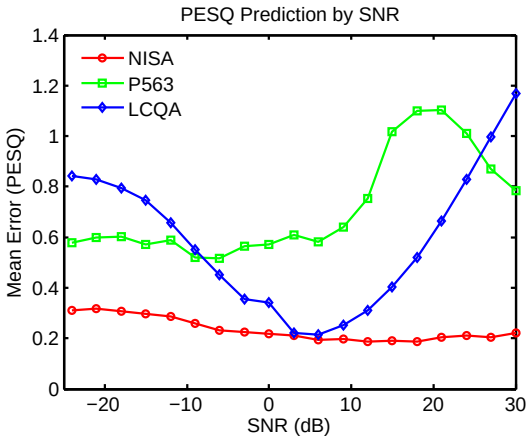


Figure 2. Per-SNR condition averaged results for PESQ prediction, based on the ADN-TN database.

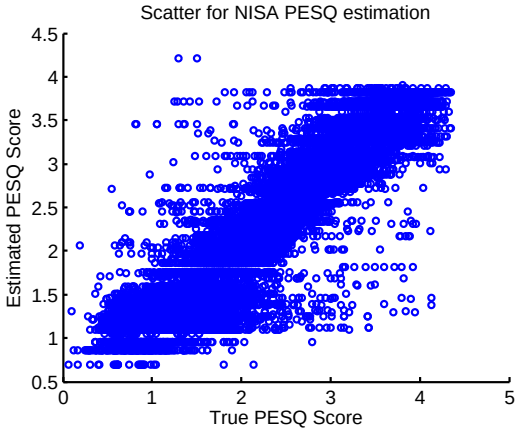


Figure 3. Scatter plot of results for PESQ prediction on the ADN-TN database.

B. Performance for C-Qual database

The generalization performance of the algorithms is ascertained by testing on the additive noise partition of the C-Qual database labeled with PESQ. No part of the C-Qual database was included in the training procedure. The results are shown in Table III. It can be seen that NISA outperforms the other methods with a marginally lower SCC and higher RMSE than for the TN database. This is attributed to the difference in the speech and noise material in C-Qual. The LCQA algorithm performs better than the standard P563 algorithm in the generalization test, with a significantly lower RMSE.

Table III
NON-INTRUSIVE PESQ ESTIMATION PERFORMANCE ON THE ADDITIVE NOISE PARTITION OF THE C-QUAL DATABASE, REPRESENTING A GENERALIZATION TEST.

			Bin Error				
Method	SCC	RMSE	0.25	0.5	0.75	1.0	TCHR
NISA	**0.91**	**0.33**	**60.7**	**88.1**	**96.4**	**98.8**	**90.9**
LCQA	0.87	0.55	41.7	60.7	77.4	90.5	88.6
P563	0.77	0.95	22.7	42.7	62.7	74.7	68.6

C. Feature analysis

The 10 most significant features for the NISA algorithm are listed in Table IV, which allows a comparison of the

importance of the different features for non-intrusive PESQ estimation in NISA. The feature significance is computed by summing the change in the Mean Squared Error (MSE) caused by splits in the final pruned CART model for each feature and dividing by the corresponding number of branch nodes.

The dynamics of the signal are considered important features in speech coding for example, where the dynamics of the envelope is more perceptually audible than spectral distortion [66] and similarly for speech enhancement [67]. In the case of NISA, the rate of change of spectral flatness and spectral envelope range along with the LTASS deviation based features are important for PESQ estimation. The NISA method selects 52 features for the binary tree regression model. The LCQA algorithm selects 12 features after feature selection. The number of features and linear combinations are obtained experimentally by re-substitution estimation on the training data.

Table IV
THE 10 MOST IMPORTANT FEATURES (RANK ORDER) FOR NON-INTRUSIVE PESQ ESTIMATION BASED ON THE TRAINING PARTITION OF THE TN DATABASE.

Rank	NISA
1	$\mu(\Delta\phi_1)$
2	$\sigma(\Delta\phi_{11})$
3	ϕ_{25}
4	$s(\phi_{14})$
5	$s(\Delta\phi_4)$
6	$\kappa(\phi_{10})$
7	$\kappa(\phi_1)$
8	$\sigma(\phi_{11})$
9	$s(\phi_3)$
10	ϕ_{16}

VI. STOI ESTIMATION RESULTS

This section presents the results for non-intrusive estimation of STOI scores using the databases outlined in Section IV-A for the LCIA, LCQA and NISA methods. These methods are trained on the training partition of the ADN-TN database and the ground truth is the intrusively computed value of STOI. The LCQA method is included for here for illustrative purposes as the LCIA method is a development of this method and we note that LCQA was not designed for a speech intelligibility estimation task.

A. Performance for TN database

Table V shows the performance of the 3 methods in the estimation of STOI on the ADN-TN database, with the NISA method outperforming the remaining methods, achieving a correlation (SCC) of 0.95 (the scatter plot of this data is presented in Fig. 6) and an RMSE of 0.08 STOI. The NISA method has a high accuracy, with 93.3% of errors less than 0.15 STOI and for an acceptance threshold of 0.62 STOI, the TCHR performance is nearly 95%. The LCIA method achieves a high correlation in this task (SCC = 0.91) but has a poor estimation accuracy, with an RMSE of 0.18. The performance of the methods for different acceptance thresholds is presented in Fig. 4, where NISA can be seen to have a consistent performance with a TCHR higher than 90% in the region of 0.5 to 0.7 STOI. A similar observation can be made for the STOI estimation performance (in terms of RMS error) according to SNR, as shown in Fig. 5, where the NISA algorithm is seen to have a more consistent performance across the range of SNR.

Table V
NON-INTRUSIVE STOI ESTIMATION PERFORMANCE FOR THE ADDITIVE NOISE PARTITION OF THE TN DATABASE.

			Bin Error				
Method	SCC	RMSE	0.05	0.10	0.15	0.20	TCHR
NISA	**0.95**	**0.08**	**64.0**	**85.4**	**93.3**	**97.0**	**94.7**
LCIA	0.91	0.18	26.6	45.4	61.3	72.9	80.9
LCQA	0.65	0.19	9.4	19.9	35.9	59.8	62.6

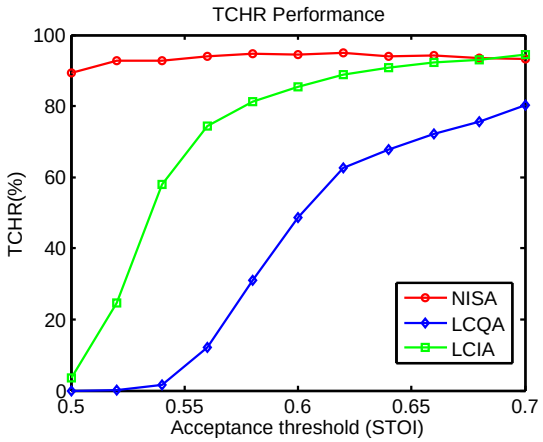


Figure 4. Results for the TCHR metric using thresholds on the STOI score in the range 0.5 to 0.7 for the ADN-TN database. This corresponds approximately to intelligibility in the range 28% to 93%.

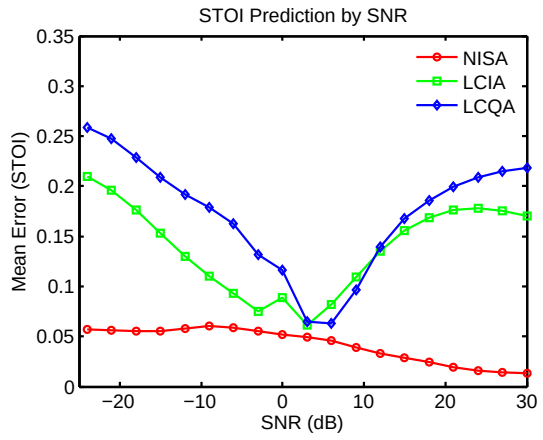


Figure 5. Per-SNR condition averaged results for STOI prediction, based on the ADN-TN database.

B. Performance for C-Qual database

The generalization performance of the methods is presented in Table VI, where the methods are trained on the ADN-TN database and tested on the additive noise partition of the C-Qual database. The performance for all methods in this task is

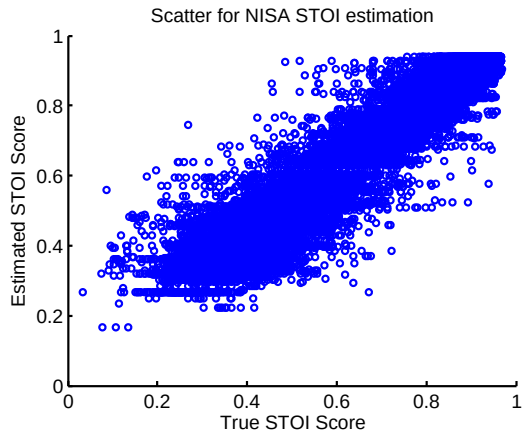


Figure 6. Scatter plot of results for STOI prediction on the ADN-TN database.

much lower, with the best performance provided by NISA in terms of the SCC (0.89) and RMSE metrics (0.17). The LCIA method is seen to outperform LCQA with an SCC of 0.88 and RMSE of 0.17.

Table VI
NON-INTRUSIVE STOI ESTIMATION PERFORMANCE FOR THE ADDITIVE NOISE PARTITION OF THE C-QUAL DATABASE.

			Bin Error				
Method	SCC	RMSE	0.05	0.10	0.15	0.20	TCHR
NISA	**0.89**	**0.17**	**10.7**	**33.3**	**54.8**	**75.0**	**93.0**
LCIA	0.88	0.17	20.2	35.7	42.8	69.0	100
LCQA	0.59	0.19	17.1	32.9	45.1	68.3	39.3

C. Feature analysis

The performance of the LCQA, LCIA and NISA methods on the test databases presented in the preceding sub-sections are based on the optimum parameters selected using a grid search on a subset of the training data (results for this tuning data are not presented here). For the LCQA method, 12 linear combinations of 16 features and a GMM with 9 mixtures was used for non-intrusive STOI estimation and the LCIA method performed best with 7 linear combinations of 8 features and a 7 mixture GMM[2]. The NISA algorithm constructs a CART regression tree with 90 features for this task, of which the 10 most significant features are presented in Table VII. This shows that the most important features for the task of STOI estimation of noisy speech using NISA are the mean of the rate of change of the envelope dynamic range, the kurtosis and skewness of the spectral flatness of the LPC spectrum and the kurtosis of the envelope variance. The iSNR and the LTASS based features are also used in the CART model along with the pitch and long term features.

VII. CONCLUSIONS

The non-intrusive assessment of the quality and intelligibility of speech signals in diverse applications including law enforcement and customer experience monitoring in telecommunications is a challenging problem due to the high levels of variation of degradation encountered. The data-driven non-intrusive NISA algorithm was proposed for this task, using a CART binary tree to model a number of novel features from the received speech signal. The intrusive PESQ algorithm was used to label a large database comprising of speech degraded by additive noise and transmission over real telephone channels and the additive noise partition of the C-Qual database was used as a generalization test database, for speech quality experiments. Similarly the intrusive STOI algorithm was used to label the additive noise partitions of the same databases for speech intelligibility estimation experiments. The use of instrusive methods for labeling the data allowed large quantity of speech to be labeled without the need for expensive and time consuming listening tests.

Two CART binary trees were trained, one for speech quality estimation and another for speech inettigiblity estimation. As part of the CART tree learning algorithm, the importance of each feature was obtained as the relative reduction in the Gini measure over the training data and it was found that six of the ten most important features used in modeling the quality and intelligiblity of speech were the same. These were the rate of chaneg of the LPC magnitude spectral flatness and centroid, the LPC excitation variance, Hilbert envelope dynamic range and variance and the deviation from LTASS in the 2 kHz to 2.5 kHz frequency band.

In the speech quality estimation task, the proposed NISA method achievied a Spearman correlation of 0.93 with PESQ and an RMSE of 0.32, outperforming the baseline LCQA algorithm and the industry standard P.563 algorithm. Similarly, for speech intelligibility assessment, NISA achieved a correlation of 0.95 with STOI with an RMS error of 0.08 on the additive noise database. Moreover, NISA was shown to be robust across a wide range of SNR values for both PESQ and STOI estimation tasks. This allows the NISA method to be used for monitoring the quality of service in a telecommunications service or for analysing past data in a non intrusive manner.

Table VII
THE 10 BEST RANKED FEATURES FOR NON-INTRUSIVE STOI ESTIMATION BASED ON THE ADDITIVE NOISE PARTITION OF THE TN DATABASE.

Rank	NISA
1	$\sigma(\Delta(\phi_{11}))$
2	$\kappa(\phi_1)$
3	$s(\phi_1)$
4	$\kappa(\phi_{10})$
5	$s(\phi_3)$
6	$\mu(\Delta(\phi_1))$
7	ϕ_{25}
8	$\sigma(\Delta(\phi_1))$
9	$s(\Delta(\phi_4))$
10	ϕ_{18}

REFERENCES

[1] R. C. Maher, "Audio forensic examination," *IEEE Signal Processing Magazine*, vol. 26, pp. 84–94, 2009.

[2]These were obtained by optimizing the parameters on a subset of the training data.

[2] P. A. Naylor, N. D. Gaubitch, D. Sharma, G. Hilkhuysen, M. Huckvale, and M. Brookes, "Intelligibility estimation in law enforcement speech processing," in *Proc ITG Conf on Speech Communication*, Bochum, Germany, Oct. 2010.

[3] S. Moller, W.-Y. Chan, N. Cote, T. H. Falk, A. Raake, and M. Waltermann, "Speech quality estimation: Models and trends," *IEEE Signal Processing Magazine*, vol. 28, pp. 18–28, 2011.

[4] *Methods for subjective determination of transmission quality*, Online, International Telecommunications Union (ITU-T) Recommendation P.800, Aug. 1996. [Online]. Available: http://www.itu.int/rec/T-REC-P.800/en

[5] T. H. Falk, H. Yuan, and W.-Y. Chan, "Single-Ended Quality Measurement of Noise Suppressed Speech Based on Kullback-Leibler Distances," *Journal of Multimedia*, vol. 2, no. 5, pp. 17 – 26, September 2007.

[6] T. H. Falk and W.-Y. Chan, "Hybrid Signal-and-Link-Parametric Speech Quality Measurement for VoIP Communications," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 8, pp. 1579 – 1589, November 2008.

[7] ITU-T, *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs*, International Telecommunications Union (ITU-T) Recommendation P.862, Feb. 2001.

[8] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, 2001, pp. 749–752.

[9] J. Beerends and J. Stemerdink, "A perceptual speech-quality measure based on a psychoacoustic sound representation," *Journal Audio Eng. Soc.*, vol. 42(3), pp. 115–123, 1994.

[10] A. Rix, R. Reynolds, and M. Hollier, "Perceptual measurement of end-to-end speech quality over audio and packet-based networks," in *Audio Engineering Society Convention 106*, May 1999.

[11] A. Rix and M. Hollier, "The perceptual analysis measurement for robust end-to-end speech quality assessment," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 3, 2000, pp. 1515–1518.

[12] ITU-T, *Mapping function for transforming P.862 raw result scores to MOS-LQ*, International Telecommunications Union (ITU-T) Recommendation P.862.1, 2003.

[13] A. W. Rix, J. G. Beerends, D.-S. Kim, P. Kroon, and O. Ghitza, "Objective assessment of speech and audio quality - technology and applications," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 1890–1901, 2006.

[14] ITU-T, *Perceptual objective listening quality assessments*, International Telecommunications Union (ITU-T) Recommendation P.863, 2011.

[15] *Single-ended method for objective speech quality assessment in narrow-band telphony applications*, International Telecommunications Union (ITU-T) Recommendation P.563, 2004.

[16] T. H. Falk and W.-Y. Chan, "Nonintrusive speech quality estimation using Gaussian mixture models," *IEEE Signal Process. Lett.*, vol. 13, no. 2, pp. 108–111, Feb. 2006.

[17] ——, "Single-ended speech quality measurement using machine learning methods," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 1935–1947, Nov. 2006.

[18] ——, "Enhanced non-intrusive speech quality measurement using degradation models," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, May 2006, pp. I–I.

[19] ——, "Feature mining for GMM-based speech quality measurement," in *Proc. Asilomar Conf. on Signals, Systems and Computers*, vol. 2, Nov. 2004, pp. 2290–2294.

[20] W. Zha and W.-Y. Chan, "Voice quality assessment using classification trees," in *Thirty-Seventh Asilomar Conference on Signals, Systems and Computers*, vol. 1, November 2003, pp. 537 – 541.

[21] ——, "Objective speech quality measurement using statistical data mining," *EURASIP Journal on Applied Signal Processing*, vol. 9, pp. 1410 – 1424, 2005.

[22] V. Grancharov, D. Y. Zhao, J. Lindblom, and W. B. Kleijn, "Low-complexity, nonintrusive speech quality assessment," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 1948–1956, Nov. 2006.

[23] W. Voiers, "Diagnostic acceptability measure for speech communication systems," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 1977, pp. 204–207.

[24] H. J. M. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *J. Acoust. Soc. Am.*, vol. 67, no. 1, pp. 318–326, Jan. 1980.

[25] R. J. M. van Hoesel and R. S. Tyler, "Speech perception, localization, and lateralization with bilateral cochlear implants," *J. Acoust. Soc. Am.*, vol. 113, no. 3, pp. 1617–1630, 2003.

[26] T. Houtgast and H. J. M. Steeneken, "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," *J. Acoust. Soc. Am.*, vol. 77, no. 3, pp. 1069–1077, 1985.

[27] H. Fletcher and J. Steinberg, "Articulation testing methods," *Bell Syst. Tech. J.*, vol. 8, pp. 806–854, 1929.

[28] G. Miller and P. Nicely, "An analysis of perceptual confusions among some English consonants," *J. Acoust. Soc. Am.*, vol. 27, no. 2, pp. 338–352, 1955.

[29] J. Egan, "Articulation testing methods," *Laryngoscope*, vol. 58(9), pp. 955–991, 1948.

[30] W. D. Voiers, "Evaluating processed speech using the diagnostic rhyme test," *Speech Technology*, vol. 1, no. 4, pp. 30–39, 1983.

[31] D. Kalikow, K. Stevens, and L. Elliott, "Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability," *J. Acoust. Soc. Am.*, vol. 61, no. 5, pp. 1337–1351, 1977.

[32] M. Nilsson, S. Soli, and J. Sullivan, "Development of hearing in noise test for the measurement of speech reception thresholds in quiet and in noise," *J. Acoust. Soc. Am.*, vol. 95, no. 2, pp. 1085–1099, 1994.

[33] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Am.*, vol. 19, no. 1, pp. 90–119, 1947.

[34] K. Kryter, "Methods for the calculation and use of the articulation index," *J. Acoust. Soc. Am.*, vol. 34, no. 11, pp. 1689–1697, 1962.

[35] ANSI, "Methods for the calculation of the articulation index," American National Standards Institute, New York, ANSI Standard ANSI S3.5–1969, 1969.

[36] ——, "Methods for the calculation of the speech intelligibility index," American National Standards Institute, ANSI Standard S3.5–1997 (R2007), 1997.

[37] B. W. Y. Hornsby, "The Speech Intelligibility Index: What is it and what's it good for?" *Hearing Journal*, vol. 57, no. 10, pp. 10–17, October 2004.

[38] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objecitve intelligibility measure for time-frequency weighted noisy speech," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2010, pp. 4214–4217.

[39] ——, "On predicting the difference in intelligibility before and after single-channel noise reduction," in *Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC)*, 2010.

[40] ——, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, September 2011.

[41] T. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 7, pp. 1766–1774, sept 2010.

[42] F. Chen, O. Hazrati, and P. C. Loizou, "Predicting the intelligibility of reverberant speech for cochlear implant listeners with a non-intrusive intelligibility measure," *Biomedical Signal Processing and Control*, vol. 8, no. 3, pp. 311 – 314, 2013.

[43] P. P. Parada, D. Sharma, J. Lainez, D. Barreda, T. van Waterschoot, and P. Naylor, "A single-channel non-intrusive c50 estimator correlated with speech recognition performance," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2016.

[44] D. Sharma, G. Hilkhuysen, N. D. Gaubitch, P. A. Naylor, M. Brookes, and M. Huckvale, "Data driven method for non-intrusive speech intelligibility estimation," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Denmark, Aug. 2010.

[45] ITU-T, *Objective Measurement of Active Speech Level*, International Telecommunications Union (ITU-T) Recommendation P.56, Mar. 1993.

[46] D. M. Brookes, "VOICEBOX: A speech processing toolbox for MATLAB," http://www.ee.imperial.ac.uk/hp/staff/dmb/voicebox/voicebox.html, 1997.

[47] D. Sharma and P. A. Naylor, "Evaluation of pitch estimation in noisy speech for application in non-intrusive speech quality assessment," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Glasglow, Aug. 2009.

[48] S. Gonzalez and M. Brookes, "A pitch estimation filter robust to high levels of noise (PEFAC)," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Barcelona, Aug. 2011.

[49] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, pp. 504–512, Jul. 2001.

[50] Z. M. Smith, B. Delgutte, and A. J. Oxenham, "Chimaeric sounds reveal dichotomies in auditory perception," *Letters to Nature*, vol. 416, pp. 87 – 90, 2002.

[51] ITU-T, *Artificial Voices*, International Telecommunications Union (ITU-T) Recommendation P.50, Sep. 1999.

[52] N. D. Gaubitch, M. Brookes, and P. A. Naylor, "Blind channel identification in speech using the long-term average speech spectrum," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Glasgow, Aug. 2009.

[53] B. Atal and L. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 3, pp. 201–212, Jun. 1976.

[54] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. CRC Press, 1984.

[55] A. K. Jain, R. P. Duin, and J. Mao, "Statistical Pattern Recognition: A Review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 1, pp. 4 – 37, January 2000.

[56] J. S. Garofolo, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database," National Institute of Standards and Technology (NIST), Gaithersburg, Maryland, Technical Report, Dec. 1988.

[57] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition II: NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 3, no. 3, pp. 247–251, Jul. 1993.

[58] K. L. Brown and E. B. George, "CTIMIT: a speech corpus for the cellular environment with applications to automatic speech recognition," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, 1995, pp. 105–108.

[59] C. R. Jankowski, Jr., , A. Kalyanswamy, S. Basson, and J. Spitz, "NTIMIT: a phonetically balanced, continuous speech, telephone bandwidth speech database," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 1990, pp. 109–112.

[60] D. Sharma, G. Hilkhuysen, N. D. Gaubitch, M. Brookes, and P. A. Naylor, "C-Qual - a validation of PESQ using degradations encountered in forensic and law enforcement audio," in *Proc. AES Conf. on Audio Forensics*, Hillerød, Denmark, Jun. 2010.

[61] *ITU-T coded-speech database*, International Telecommunications Union (ITU-T) Supplement P.Sup23, Feb. 1998.

[62] G. Hilkhuysen, N. Gaubitch, M. Brookes, and M. Huckvale, "Effects of noise suppression on intelligibility: dependency on signal-to-noise ratios," *J. Acoust. Soc. Am.*, vol. 131, no. 1, pp. 531–539, 2012.

[63] E. L. Lehmann and H. J. M. D'Abrera, *Nonparametrics: Statistical Methods Based on Ranks*. Englewood Cliffs, NJ: Prentice-Hall, 1998.

[64] ITU-T, *Application guide for objective quality measurement based on Recommendations P.862, P.862.1 and P.862.2*, International Telecommunications Union (ITU-T) Recommendation P.862.3, Nov. 2007.

[65] K. Worrall and R. Fellows, "Practical and affordable intelligibility testing for engineers and algorithm developers," in *Proc. AES Conf. on Audio Forensics*, Hillerod, Denmark, June 2010, pp. 194–201.

[66] H. Knagenhjelm and W. B. Kleijn, "Spectral dynamics is more important than spectral distortion," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, Michigan, USA, May 1995.

[67] T. Quatieri and R. Dunn, "Speech enhancement based on auditory spectral change," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)SS*, vol. 1, Florida, USA, May 2002.