Predicting the quality of processed speech by combining modulation-based features and model trees

Benjamin Cauchi[1,5], João F. Santos[2], Kai Siedenburg[4,5], Tiago H. Falk[2], Patrick A. Naylor[3], Simon Doclo[4,5], Stefan Goetze[1,5]

[1]Fraunhofer IDMT, Project Group Hearing, Speech and Audio Technology, Oldenburg, Germany
[2]INRS-EMT, University of Quebec, Montreal, QC, Canada
[3]Imperial College London, Dept. of Electrical Engineering, London, United Kingdom
[4]University of Oldenburg, Dept. of Medical Physics and Acoustics, Oldenburg, Germany
[5]Cluster of Excellence Hearing4all, Oldenburg, Germany
Email: benjamin.cauchi@idmt.fraunhofer.de

Abstract

Many signal processing methods have been proposed to improve the quality of speech recorded in the presence of noise and reverberation. The evaluation of these methods either requires the use of perceptual measures, i.e. listening tests, or instrumental measures. Perceptual measures are typically more reliable but are quite costly and time-consuming. On the other hand, instrumental measures may correlate poorly with the perceived speech quality. In this paper we propose to train an instrumental measure, combining modulation-based features and model trees, on the basis of perceptual scores obtained on a small corpus of speech data that has been processed by a combination of beamforming and spectral postfiltering. For evaluation purposes the resulting measure is then applied to a larger corpus. Results show that the use of model trees to train the predicting function of an instrumental measure increases its correlation with perceptual scores.

1 Introduction

In many speech communication applications, the speech of a distant user is recorded by a single or by multiple microphones. In such conditions, the recorded speech signal is typically corrupted by both noise and reverberation, which may severely degrade the perceived speech quality and speech intelligibility. To overcome these effects, many speech enhancement methods have been proposed [1–3]. Although many methods are able to substantially reduce the amount of noise and reverberation in the recorded signal, processing artefacts frequently result in a degraded speech quality [4], requiring the developed signal processing algorithms to be thoroughly evaluated.

Perceptual measures are generally considered the most reliable method for assessing the quality of processed speech. These measures require a group of human assessors to evaluate speech signals with respect to predefined attributes, such as overall quality, level of reverberation, coloration, etc. The evaluation is usually performed by grading each attribute on a scale which either consists of a few values, e.g., for the mean opinion score (MOS) [5], or continuous values, as in the multiple stimuli test with hidden reference and anchor (MUSHRA) [6]. However, such a procedure is quite costly and time-consuming and can therefore only be applied for a handful of stimuli. Consequently, many speech enhancement approaches are often only evaluated using instrumental measures, which can be applied to large corpora of processed signals. Instrumental measures can be broadly classified as either intrusive [7–11], i.e. requiring a (clean) reference signal, or non-intrusive [12–15], i.e. not requiring a reference signal [16]. Unfortunately, in the presence of processing artefacts, instrumental measures have been shown to correlate poorly with perceptual scores [4, 16, 17]. The non-intrusiveness property is particularly valuable for evaluation of real-life recordings, for which no reference is available, and hence this paper focuses on the development of non-intrusive measures to predict the perceived quality of processed speech.

Non-intrusive measures predict the value of an attribute by applying a predicting function to a feature vector extracted from the processed signal. The non-intrusive measures proposed in the literature differ in both the considered features and in the predicting function. In the case of ANIQUE+ [12], four features are used, which represent different types of distortions, namely, signal distortion, overall frame distortion, mute distortion and non-speech distortion. The predicted speech quality is obtained as a linear combination of these features. In the case of P.563 [13], several features are used, which represent parameters of the speech signal under test and several types of distortions, such as temporal clipping, "*robotization*" and noise. The predicted speech quality is obtained by applying a predicting function which combines decision rules as well as linear combination of features. In the case of both the speech-to-reverberation modulation ratio (SRMR) [14] and the normalized SRMR ($\text{SRMR}_{\text{norm}}$) [15], the used features comprise the modulation energy in a few modulation frequency bands, averaged over the whole length of the signal under test. The difference between SRMR and $\text{SRMR}_{\text{norm}}$ lies in the computation of these energies, where for $\text{SRMR}_{\text{norm}}$ the frequency range of the modulation filters was reduced and a per-frame energy treshold was implemented in order to reduce variability caused by pitch and speech content. However, both SRMR and $\text{SRMR}_{\text{norm}}$ use the same predicting function which is equal to the ratio between the energy in the lower and higher modulation frequencies.

Of all mentioned non-intrusive measures, $\text{SRMR}_{\text{norm}}$ has been shown to be the most reliable measure to predict the speech quality of speech processed by a combination of beamforming and spectral postfiltering [18]. The non-intrusive measure proposed in this paper uses the same features as $\text{SRMR}_{\text{norm}}$ but aims at an higher reliability by

The research leading to these results has received funding from the EU Seventh Framework Programme project DREAMS under grant agreement ITN-GA-2012-316969 and by the German Academic Exchange Service (DAAD).

using a different predicting function, trained on a small corpus of signals for which perceptual scores are available. The predicting function is a model tree [19], therefore combining classification rules and regression in a similar fashion as in the predicting functions used, in, e.g., [12] or [13]. Results on a large number of speech utterances show that the proposed measure yields higher correlation with perceptual scores than $\text{SRMR}_{\text{norm}}$, hence indicating that the proposed approach could be used to reliably evaluate algorithms by training the predicting function on a small corpus before applying the resulting measure to large corpora.

2 Proposed method

2.1 Considered features

The measure proposed in this paper considers the same features as $\text{SRMR}_{\text{norm}}$, i.e. the energy of the modulation spectra averaged over time and acoustic frequencies. The computation of these features is briefly summarized in this section. The signal under test $x[n]$, with n denoting the sample index, is filtered by a gammatone filterbank with J channels, resulting in J filtered signals $x_j[n]$, with j denoting the frequency index. The temporal envelope $e_j[n]$ is extracted from $x_j[n]$ as

$$e_j[n] = \sqrt{x_j^2[n] + \mathscr{H}\{x_j[n]\}^2}, \tag{1}$$

with $\mathscr{H}\{\cdot\}$ denoting the Hilbert transform. The temporal envelopes are divided into windowed overlapping frames of length L. The modulation spectral energy $E_j[k,\ell]$ is computed as the squared magnitude of the Fourier transform of the ℓ-th frame in the k-th modulation frequency bin. The modulation spectral energies $E_j[k,\ell]$ are only considered for frequency bins in the interval k_{min} to k_{max}, before being reduced to B bands by grouping the modulation frequency bins. After applying energy tresholding as proposed in [15, 20] and averaging over all frames, the signal $x[n]$ is represented by $J \times B$ coefficients $\varepsilon_j[b]$, with b denoting the modulation frequency index. Finally, the feature vector $\mathbf{x}$ of length B can be constructed as

$$\mathbf{x} = \left[\mathsf{x}[0]\ \mathsf{x}[1]\ \ldots\ \mathsf{x}[B-1]\right]^{\mathrm{T}}, \tag{2}$$

where the b-th element $\mathsf{x}[b]$ is computed as

$$\mathsf{x}[b] = \frac{\sum_{j=0}^{J-1} \varepsilon_j[b]}{\max\limits_{b}\left\{\sum_{j=0}^{J-1} \varepsilon_j[b]\right\}}. \tag{3}$$

In this paper, the different parameters of the feature extraction have been set as in [15], which is an extension of [14]. The gammatone filterbank uses $J = 23$ channels with center frequencies ranging from 125 Hz to 4 kHz. The temporal envelope $e_j[n]$ is divided in frames using a Hamming window of length L corresponding to 256 ms and using shifts of 32 ms. The indices k_{min} and k_{max} correspond to the range of modulation frequencies between 4 Hz and 40 Hz. These values have been shown to reduce the sensitivity of the extracted features to the pitch content [15] compared to the SRMR features proposed in [14]. The modulation frequency bins are grouped into $B = 8$ bands, resulting in a feature vector $\mathbf{x}$ with $B = 8$ elements.

2.2 Predicting function

The predicted value v of the perceived speech quality is obtained by applying a predicting function $f(\cdot)$ to the feature vector $\mathbf{x}$, i.e. $v = f(\mathbf{x})$. In the case of both SRMR [14] and $\text{SRMR}_{\text{norm}}$ [15], this predicting function is equal to the energy ratio between the lower and upper modulation frequencies, i.e.

$$v_{\text{SRMR}} = f_{\text{SRMR}}(\mathbf{x}) = \frac{\sum_{b=0}^{b_{\text{low}}-1} \mathsf{x}[b]}{\sum_{b=b_{\text{low}}}^{b_{\text{high}}-1} \mathsf{x}[b]}, \tag{4}$$

with b_{low} and and b_{high} denoting the lowest and highest considered modulation frequency index, respectively. As in [15], $b_{\text{low}} = 4$ and b_{high} is determined as the index for which 90% of the modulation energy is accounted for, as proposed in [14]. It should be noted that the denominator in (3) has no influence on the output of (4). However, it is still used here, as having features bounded between 0 and 1 is convenient to define decision rules.

Instead of using the energy ratio in (4) as the predicting function, we propose to train the predicting function based on a small corpus of (processed) speech data. It is hence assumed that a set of T signals are available for training, where for the t-th signal a perceptual score p_t of speech quality is available, e.g., from a previous perceptual measurement. Additionally, a set of M signals are available for testing, where for the m-th signal, the perceived speech quality $\tilde{p}_m$ would in practice be unknown. Using the training set, we aim to determine the predicting function $f(\cdot)$ which maximizes the Pearson correlation coefficient between predicted and perceived speech quality, i.e.

$$f(\cdot) = \underset{f(\cdot)}{\text{argmax}} \frac{\sum_{m=1}^{M}(\tilde{p}_m - \overline{p})(\tilde{v}_m - \overline{v})}{\sqrt{\sum_{m=1}^{M}(\tilde{p}_m - \overline{p})^2}\sqrt{\sum_{m=1}^{M}(\tilde{v}_m - \overline{v})^2}}, \tag{5}$$

with

$$\tilde{v}_m = f(\tilde{\mathbf{x}}_m), \tag{6}$$

with $\tilde{\mathbf{x}}_m$ denoting the feature vector extracted from the m-th signal of the testing set, cf. (2), and

$$\overline{p} = \frac{1}{M}\sum_{m=1}^{M} \tilde{p}_m, \qquad \overline{v} = \frac{1}{M}\sum_{m=1}^{M} \tilde{v}_m. \tag{7}$$

In this paper, we will use a model tree as the predicting function built using the so-called M5' algorithm [19], i.e., each leaf node of the decision tree consists of a linear model which predicts a value from the input features. The construction of the model tree requires a set of observations which are all associated with a feature vector and a target value. The set of observations here consists of the signals available in the training set, which are all associated with a feature vector $\mathbf{x}_t$ and a perceptual score p_t. The model tree is built in two stages. In the first stage, a conventional binary decision tree is built by recursively splitting the set of observations into subsets in which the variation of the perceptual scores is maximal. This results in a large binary decision tree which can be used to predict a value for each observation. In the second stage, each node, starting from the top of the constructed decision tree, is replaced by a linear model if the application of this model results in a lower prediction error than the binary classification. The resulting model tree is finally simplified using

ISBN 978-3-8007-4275-2

pruning and smoothing as described in [19], and contains D binary decision nodes.

Although small values of D may result in underfitting, $D << T$ is necessary to avoid overfitting, which would make the resulting predicting function unapplicable to signals not included in the training set. Additionally, it should be noted that the predicting function $f(\cdot)$ resulting from the model tree is actually not designed to maximize the correlation in (5) but rather to minimize the mean squared error obtained on the training set, i.e.,

$$f(\cdot) = \underset{f(\cdot)}{\operatorname{argmin}} \frac{1}{T} \sum_{t=1}^{T} \left(f(\mathbf{x}_t) - p_t \right)^2 . \tag{8}$$

The effectiveness of the resulting predicting function to maximize the correlation is discussed in Section 3.2.

3 Experiment

3.1 Corpus

For our evaluation, we used a subset of the evaluation set of the REVERB challenge [21], which consists of a large corpus of speech signals, corrupted by reverberation and noise, sampled at 16 kHz. As described in detail in [22], we obtained perceptual ratings for our subset by conducting a MUSHRA test. This corpus is divided into simulated and real data. The simulated data is composed of clean speech signals from the WSJCAM0 corpus [23], which have been convolved with recorded room impulse responses (RIRs) and to which measured noise at a signal-to-noise ratio (SNR) of 20 dB has been added. The real data is composed of utterances from the MC-WSJ-AV corpus [24] and contains speech recorded in a room in the presence of noise. For each room, two distances (denoted by "near" and "far") between the target speaker and the center of the circular array of 8 microphones have been considered. The combination of a room and a particular distance will be refered to as "condition" in the remainder of this paper. Four acoustic conditions have been tested, namely "S2, near", "S2, far", "R1, near" and "R1, far". The characteristics of these conditions are summarized in Table 1. The stimuli presented to the assessors consisted of unprocessed speech, an hidden reference, an anchor and the recorded signal processed using 3 processing schemes, namely single-channel spectral enhancement (SE_3), an MVDR beamformer (MVDR) and their combination (MVDR+SE_3). Detailed descriptions of the perceptual test and of the processing schemes can be found in [22].

A total of 21 self-reported normal-hearing listeners participated in the perceptual evaluation. The listening test was conducted in a soundproof booth and the assessors listened to diotic signals through headphones. Each assessor evaluated the "overall quality" of 3 utterances per condition and algorithm (i.e. 12 uterances per assessor) on a scale ranging from 0 to 100. For each assessor, the utterances to be evaluated were randomly picked from the REVERB challenge database, in order to reduce potential bias due to the choice of speech stimuli. The obtained MUSHRA scores are summarized in Fig. 1. It should be noted that the test was designed to evaluate the performance of processing schemes, i.e. only signals within the same acoustic condition were compared. Therefore, the evaluation described in the next section will assess the ability of the proposed method to improve the reliability of speech quality prediction when the predicting function is trained for each particular acoustic condition.

Set	Room	T_{60} [ms]	Distance [cm]	Label
Simulated	medium	500	50	S2, near
			200	S2, far
Real	large	700	100	R1, near
			250	R1, far

Table 1: Summary of the considered acoustic conditions.

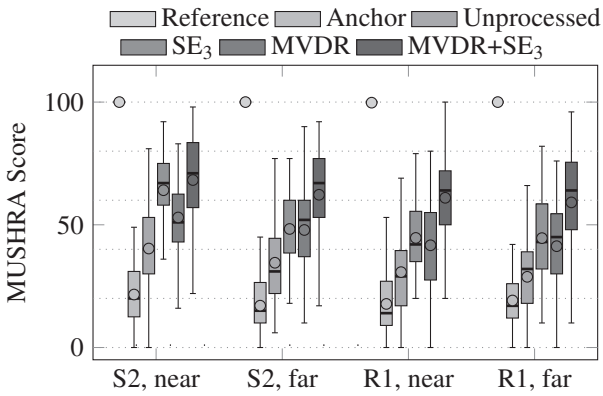

Figure 1: MUSHRA scores for the different acoustic conditions and processing schemes as well as for the reference, the unprocessed signal and the anchor. The means over all files and assessors are displayed by circles.

3.2 Evaluation procedure

As the method described in Section 2.2 requires a training set to train the predicting function, we used 3-fold cross-validation to assess its performance. The 21 assessors have been randomly divided into 3 equally sized disjoint subsets of 7 assessors. For each fold, the signals and corresponding perceptual scores contained in 2 of these subsets were considered as the test set, while the data corresponding to the remaining subset was used for training the predicting function. Using this folding, assessors, speech stimuli and noise segments always differ between training and testing and for each fold, the training and test sets contain $T = 126$ and $M = 252$ signals, respectively. For each recording in either training or test set, the true value of the perceived speech quality has been considered to be the mean of the perceptual scores attributed to signals in this set that have been processed by the same processing scheme. For each fold, results are obtained separately for each of the 4 considered acoustic conditions.

This evaluation procedure is used to compare the prediction of the perceived speech quality obtained by using the modulation-based features described in Section 2.1 as input to either the predicting function from (4), i.e. $\mathrm{SRMR}_{\mathrm{norm}}$, or the predicting function based on the model tree built using the training set and implemented using [25]. These predicting functions are evaluated using two figures of merits: 1) the Pearson correlation coefficient ρ between the true value of the perceived speech quality and the value estimated by the predicting function, 2) the standard deviation σ across different signals within the same processing schemes and acoustic conditions. For

ISBN 978-3-8007-4275-2

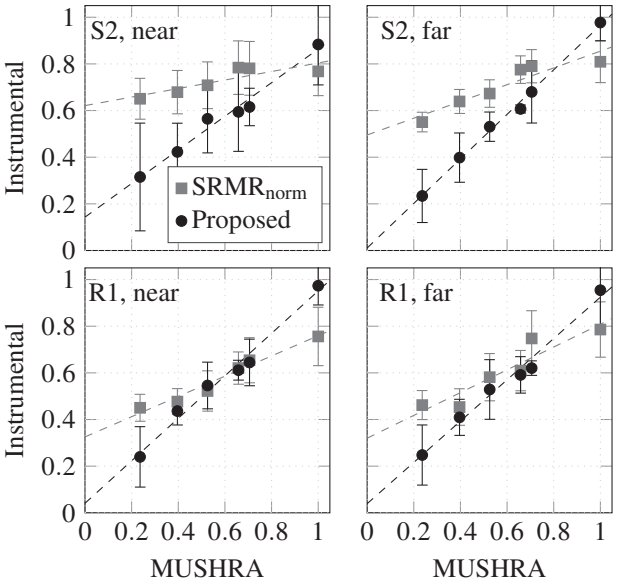

Figure 2: Scatterplot of the predicted value of the speech quality as a function of the true perceived speech quality, with values normalized between 0 and 1. The vertical bars represent the standard deviations of the predicted values for each algorithm in the considered corpus. The dashed lines repesent linear functions fitted to the mean of the predicted values.

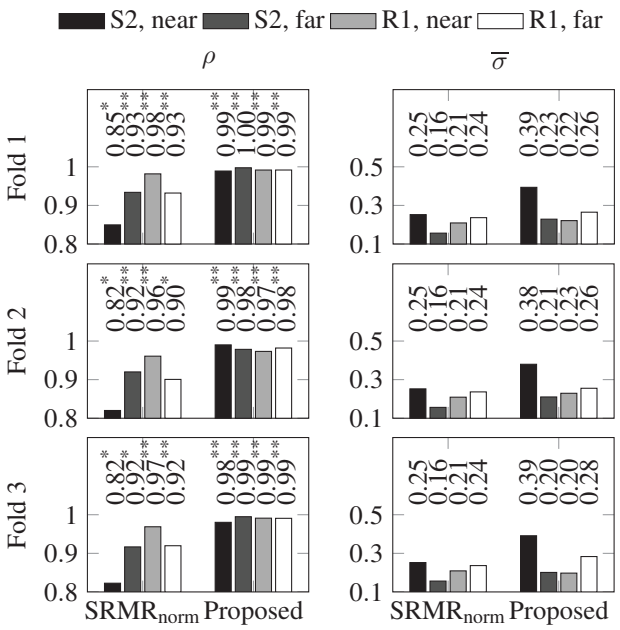

Figure 3: Pearson correlation coefficient ρ (left) and averaged standard deviation $\overline{\sigma}$(right) of SRMR$_{\text{norm}}$ and of the proposed approach. Each row represents one of the 3 folds considered in the evaluation setup. * and ** denote significance at the 5% and 1% level, respectively.

each condition, we present the average $\overline{\sigma}$, obtained as the root of the sum of the squared standard deviations for each processing scheme. Therefore, large values of ρ indicate the ability of the predicting function to predict the quality of the processed speech averaged over a large number of signals, while small values of $\overline{\sigma}$ indicate that the function is able to consistently predict between recordings.

3.3 Results

Fig. 2 depicts the scores predicted by SRMR$_{\text{norm}}$ and the proposed method as a function of the true perceived speech quality. Since the same conclusions can be drawn for all folds, only the results obtained for one fold are displayed. It can be observed that for all four acoustic conditions, both predicting functions result in predicted values increasing with the perceived speech quality. Additionally, in the case of the proposed method, the linear function fitted to the mean of the predicted values is closer to a proportional function than in the case of SRMR$_{\text{norm}}$. However, the spread of the predicted values is larger in the case of the proposed approach than in the case of SRMR$_{\text{norm}}$. These observations suggest that the proposed method might lead to higher correlations but larger standard deviations.

The observations from Fig. 2 seem to be confirmed by Fig. 3, in which the obtained values of the Pearson correlation coefficient ρ and the average standard deviation $\overline{\sigma}$ are depicted for the three folds and the four acoustic conditions. First, it appears that ρ and $\overline{\sigma}$ yield very similar values over the three folds, suggesting that the model tree used as a predicting function in the proposed method avoids overfitting and can be applied in the case of $M > T$, i.e., using a test set larger than the training set. This is partly the consequence of the number of decision rules being low, as, over all rules and conditions, the built model trees contain $D <= 28$ decisions while the number of signals in the training set is always $T = 126$. In addition, it appears that the Pearson correlation coefficient between SRMR$_{\text{norm}}$ and the perceived speech quality varies between acoustic conditions, namely from 0.85 to 0.98 in the first fold, suggesting that the reliability of SRMR$_{\text{norm}}$ depends on the environment in which the signals have been recorded. On the other hand, the Pearson correlation coefficient of the proposed method with the perceptual scores does not vary between acoustic conditions, which suggests that the model tree used as predicting function does generalize to signals recorded in the conditions under test, as expected. Notably, for all folds and acoustic conditions, the correlations are higher for the proposed approach. On the other hand, the standard deviation $\overline{\sigma}$ is larger for the proposed method, most notably for the condition "S2, near". These results suggest that the proposed method is the most effective in predicting the quality of the speech processed by particular algorithms over a large number of signals, but may not be the most effective when considering signals in isolation.

4 Conclusion

This paper proposed to use a model tree as a predicting function for the quality of speech processed by speech enhancement algorithms. We used modulation-based features, similarly as used in SRMR$_{\text{norm}}$ [15], which has been shown to be a promising non-intrusive measure to assess speech quality. Experimental results have shown that the proposed approach yields higher correlations than SRMR$_{\text{norm}}$, illustrating the ability of the proposed non-intrusive measure to predict ratings of speech quality. These results also showed that the proposed measure can be trained on a small training set and predict the speech quality of signals within a larger test set.

ISBN 978-3-8007-4275-2

References

[1] S. Doclo, W. Kellermann, S. Makino, and S. Nordholm, "Multichannel signal enhancement algorithms for assisted listening devices," *IEEE Signal Processing Magazine*, vol. 32, pp. 18–30, march 2015.

[2] P. A. Naylor and N. D. Gaubitch, eds., *Speech Dereverberation*. Springer, 2010.

[3] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Berlin, Germany: Springer-Verlag, 2008.

[4] K. Kinoshita, M. Delcroix, S. Gannot, E. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Mass, T. Nakatani, B. Raj, A. Sehr, and T. Yoshioka, "A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, pp. 1–19, Jan. 2015.

[5] ITU-T, "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithms," Standard P.835, International Telecommunications Union (ITU-T), Nov. 2003.

[6] ITU-T, "Method for the subjective assessment of intermediate quality levels of coding systems," Standard BS.1534–3, International Telecommunications Union (ITU-T), Nov. 2003.

[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," Feb. 2001.

[8] ITU-T, "Perceptual objective listening quality assessment: An advanced objective perceptual method for end-to-end listening speech quality evaluation of fixed, mobile, and IP-based networks and speech codecs covering narrowband, wideband, and super-wideband signals," Standard P.863, International Telecommunications Union (ITU-T), Jan. 2011.

[9] J. Ma, Y. Hu, and P. C. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *J. Acoust. Soc. Am.*, vol. 125, pp. 3387–3405, May 2009.

[10] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, pp. 2125–2136, Sept. 2011.

[11] R. Huber and B. Kollmeier, "PEMO-Q - a new method for objective audio quality assessment using a model of auditory perception," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, pp. 1902–1911, Nov. 2006.

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

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

[14] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, pp. 1766–1774, Sept. 2010.

[15] J. F. Santos, M. Senoussaoui, and T. H. Falk, "An improved non-intrusive intelligibility metric for noisy and reverberant speech," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, pp. 55–59, Sept. 2014.

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

[17] S. Goetze, A. Warzybok, I. Kodrasi, J. Jungmann, B. Cauchi, J. Rennies, E. Habets, A. Mertins, T. Gerkmann, S. Doclo, and B. Kollmeier, "A study on speech quality and speech intelligibility measures for quality assessment of single-channel dereverberation algorithms," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, pp. 233–237, Sept 2014.

[18] A. Avila, B. Cauchi, S. Goetze, S. Doclo, and T. Falk, "Performance comparison of intrusive and non-intrusive instrumental quality measures for enhanced speech," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, Sept 2016. submitted.

[19] E. Frank, Y. Wang, S. Ingliss, G. Holmes, and I. W, "Using model trees for classification," *Machine Learning*, vol. 32, no. 1, pp. 63–76, 1998.

[20] B. E. D. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *Speech communication*, vol. 25, no. 1, pp. 117–132, 1998.

[21] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, R. Haeb-Umbach, V. Leutnant, A. Sehr, W. Kellermann, R. Maas, S. Gannot, and B. Raj, "The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, (New Paltz, N.Y., U.S.A.), Oct. 2013.

[22] B. Cauchi, I. Kodrasi, R. Rehr, S. Gerlach, A. Jukić, T. Gerkmann, S. Doclo, and S. Goetze, "Combination of MVDR beamforming and single-channel spectral processing for enhancing noisy and reverberant speech," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, pp. 1–12, July 2015.

[23] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJCAMO: a British English speech corpus for large vocabulary continuous speech recognition," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, (Detroit, Michigan, U.S.A), pp. 81–84, May 1995.

[24] M. Lincoln, I. McCowan, J. Vepa, and H. Maganti, "The multichannel Wall Street Journal audio-visual corpus (MC-WSJ-AV): Specification and initial experiments," in *Proc. IEEE Workshop Autom. Speech Recognition and Understanding (ASRU)*, (Cancún, Mexico), pp. 357–362, Dec. 2005.

[25] G. Jekabsons, "M5PrimeLab: M5' regression tree, model tree, and tree ensemble toolbox for Matlab/Octave." `http://www.cs.rtu.lv/jekabsons/`, 2015.

ISBN 978-3-8007-4275-2