IMPROVING DEEP MODELS OF SPEECH QUALITY PREDICTION THROUGH VOICE ACTIVITY DETECTION AND ENTROPY-BASED MEASURES

Jasper Ooster and Bernd T. Meyer

Medizinische Physik and Cluster of Excellence Hearing4all
Carl von Ossietzky Universität Oldenburg, Germany

ABSTRACT

This paper explores Deep machine listening for Estimating Speech Quality (DESQ), which predicts the perceived speech quality based on phoneme posterior probabilities obtained from a deep neural network. The degradation of phonemes is quantified with the entropy-based Gini measure that is compared to the mean temporal distance (MTD) proposed earlier. Since long speech pauses might have a large effect on the speech quality, we investigate if a voice activity detection (VAD) has a beneficial or detrimental effect on the predictive power of our model. The evaluation is performed by correlating the model output and mean opinion scores (MOS) of normal-hearing listeners who rated signals degraded by typical VoIP artifacts. While the Gini-based measure and MTD result in very similar predictions (with a lower computational cost for the Gini-measure), the VAD increases performance from $r = 0.87$ to $r = 0.91$ which is higher than three competing baselines (ITU-P.563, ANIQUE+, and SRMRnorm).

Index Terms— subjective speech quality prediction, non-intrusive, deep neural network, voice activity detection, automatic speech recognition

1. INTRODUCTION

The perceived speech quality (SQ) of speech is an important measure for the analysis of telecommunication channels and speech enhancement algorithms. However, performing listening experiments to obtain subjective ratings of the perceived quality of a presented speech signal is time-consuming and expensive. Therefore, SQ models have been proposed for predicting the mean opinion score (MOS) of listeners from the acoustic stimulus [1]. Double-ended (or intrusive) models such as Perceptual Evaluation of Speech Quality (PESQ) [2] and Perceptual Objective Listening Quality Assessment (POLQA) [3] produce accurate predictions, but require both the degraded and the clean reference signal that are not available in most real-life scenarios.

To overcome this limitation, single-ended (or non-intrusive) models have been proposed that only require the degraded signal. Three single-ended algorithms that were shown to produce accurate predictions of subjective SQ are the ITU standard P.563 [4], Auditory Non-Intrusive Quality Estimation plus (ANIQUE+) [5], which is a standard of the American National Standard Institute, and the normalized speech-to-reverberation modulation energy ratio (SRMRnorm) [6]. The ITU standard P.563 [4] estimates separate quality features from signal characteristics such as the signal-to-noise ratio (SNR), linear prediction coefficients, and interruption indicators, and combines them into the SQ prediction. ANIQUE+ combines three intermediate measures of distortion, i.e., mute and non-speech distortion, as well as frame distortion. The latter is quantified by performing a spectral modulation analysis based on a perceptual model. The SRMRnorm as originally proposed by Falk and Chan [7] is computed by comparing the average energy in low modulation frequency bands to the average energy in high modulation frequency bands. A comprehensive overview of speech quality prediction algorithms is presented in [8].

In this paper, we explore a model that is based on Deep machine listening for Estimating Speech Quality (DESQ) as first introduced in [9]. It is structurally identical to the Listening Effort prediction from Acoustic Parameters (LEAP) model for predicting subjective listening effort as proposed in [10]. The model is motivated by the degradation of phoneme representations in suboptimal acoustic conditions obtained from an acoustic model of an ASR system, specifically, from a DNN that is trained to predict phoneme probabilities (or *posteriorgrams*, i.e., phoneme posterior probabilities over time). The degradation of posteriorgrams is quantified with a performance measure, i.e., the mean temporal distance (MTD) originally proposed to predict ASR error rates [11] and later on used for selecting the optimal stream in multi-stream ASR [12].The model output can then be compared to mean opinion scores (MOS) from listeners. When using an off-the-shelf training corpus (Aurora 4 [13]), the model outperformed ITU P.563 and was on average on par with ANIQUE+ [9]. Based on this approach, the dependence between predictive power and the number of training samples was explored for the TCD-VoIP database [14] in related work [15], which has shown that this approach profits from training sets with at least 80 hours of speech. The model is per se not suitable to predict degraded speech perception in the presence of a competing speaker (since it is based on speaker-independent ASR), but reached an average correlation of $r = 0.87$ for four other conditions. While this is a promising result, previous studies were limited to using one specific performance measure (MTD, as mentioned above) which is a crucial building block of the model and should bear potential for optimization. Second, the influence of speech pauses has not been taken into account, which might also be crucial for the SQ prediction: The absence of speech during pauses could cause severe problems for the DNN-based acoustic model exploited here. On the other hand, noise-only segments allow for a good estimate of the noise properties, which could be beneficial for estimating the effect of noise during speech segments. Finally, the two baselines chosen in previous work have been successful, but current state-of-the-art baselines based on machine learning have not been considered. In the current study, the resulting research questions are addressed for the first time by exploring the

This research was supported by the DFG (Cluster of Excellence 1077/1 Hearing4all; URL: http://hearing4all.eu and the CRC TRR 31, Transfer Project T01).

978-1-5386-4658-8/18/$31.00 ©2019 IEEE

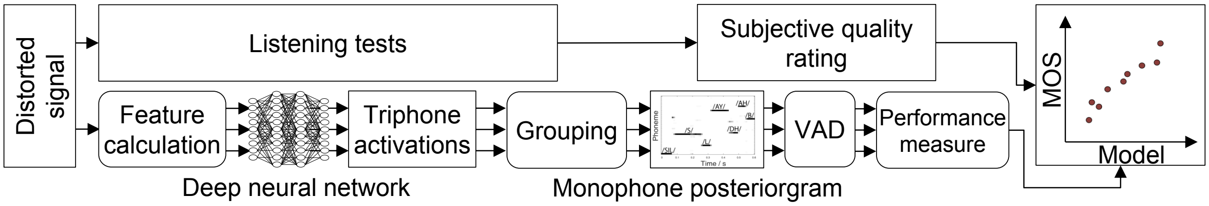

Fig. 1. Illustration of the DNN-based speech quality prediction system which has been adopted from [10].

entropy-based Gini impurity, which has been suggested as a measure of sparseness before [16], and which could help to differentiate between distinct, clear phoneme activations in clean speech and degraded phonemes. The effect of speech pauses is investigated by implementing a posteriorgram-based voice activity detection (VAD), and the model performance is compared to the three baselines mentioned above, including the state-of-the-art SRMRnorm model.

The remainder of this paper is structured as follows: The general concept of the ASR-based model is described in the next section along with the ASR architecture, the corresponding training data, and the SQ database. The results section presents model performance in five different types of distortion. Discussion and summary are presented in Sections 4 and 5, respectively.

2. METHODS

2.1. Speech quality prediction model

The model explored in this paper is illustrated in Figure 1. To obtain model predictions, we first train a standard ASR system that combines a feed-forward DNN (which serves as acoustic model) with a hidden Markov model (HMM). The training procedure is described in the next section in detail. We assume that phoneme posterior probabilities from the DNN degrade in the presence of factors that negatively affect speech quality; this degradation is quantified with two performance measures (as described below). The DNN output corresponds to context-dependent triphones, which are grouped to 42 monophones (including silence, spoken- and non-spoken-noise). This allows to visualize the output (Figure 2), is computationally cheaper, and produces similar results than using triphone activations [17]. In contrast to the original model [9], the monophone posteriorgram is also used for a VAD: Frames with the highest posterior probability for the silence class are interpreted as non-speech frames and discarded from further analyses. This was motivated from inconsistent predictions for speech utterances with longer speech pauses, which we hope to compensate by removing silence frames. Note that a forward-run of the model does not require a decoding step with the HMM or a word transcript, since it relies on the DNN output alone; the HMM is therefore not shown in Fig. 1.

The two performance measures investigated in this work are the MTD and the Gini measure: It assumes that signal degradations result in smeared phoneme activations of the DNN output, i.e., phoneme vectors should become more similar on average for noisy phoneme representations. On the other hand, clean vector representations should have very distinct class activations, and vectors representing different phonemes should be distant in vector space. The measure is defined as

$$M(\Delta t) = \frac{1}{T - \Delta t} \sum_{t=\Delta t}^{T} \mathcal{D}(\vec{p}(t - \Delta t), \vec{p}(t)) \qquad (1)$$

with phoneme vectors $\vec{p}$ and the symmetric Kullback-Leibler divergence $\mathcal{D}$. A scalar value is obtained by averaging values for Δt in the range of 350 ms to 800 ms as proposed in [9]. In the following, the resulting scalar value is referred to as MTD.

The Gini impurity $\mathcal{G}$ is used in the classification and regression tree (CART) algorithm [16]) and takes the sparseness of the temporal frames into account. It is calculated according to

$$\mathcal{G}(t) = 1 - \sum_{i}^{M} p_i(t)^2. \qquad (2)$$

where M is the number of phoneme classes, and p_i is the ith entry of the current phoneme probability vector $\vec{p}(t)$. To obtain positive correlation values, we calculate the Gini *purity* $\tilde{\mathcal{G}}$ given by

$$\tilde{\mathcal{G}}(t) = \sum_{i}^{M} p_i(t)^2. \qquad (3)$$

The Gini performance measure is the Gini purity averaged over each utterance.

2.2. ASR system

In this study we use the best-performing DNN from [15], as we investigated there different training procedures and the number of training samples. The DNN was trained on 40-dimensional log-Mel-spectral coefficients features with a splicing of ±5 frames using the *nnet1* recipe for the Aurora 4 database from the open source ASR software *Kaldi*. The DNN had six layers, each with 2048 neurons, a softmax output-layer and a sigmoid-nonlinearity. It was initialized with a layer-wise Restricted Boltzmann Machine (RBM) pre-training and fine-tuned with frame cross-entropy (CE) training. The training targets for the fine-tuning were alignments for ≈ 3.400 triphones created using a Gaussian Mixture Model - Hidden Markov Model (GMM-HMM) system. With this fine-tuned DNN, new phonetic alignments were created on which the pre-trained network was fine tuned again. This procedure is a standard approach for training ASR models in *Kaldi* and has not been optimized for SQ prediction.

The WSJ1 speech corpus with the full SI284 set containing 37,416 utterances and 81.27 h from 283 speakers is used as training set. Since previous studies have shown that training models from clean data does not provide sufficient robustness for SQ models [17], we created a multi-condition training set that resembles the Aurora 4 set by adding additive noise at random SNRs in the range of 10 dB to 20 dB to 75% of the utterance. Finally, all files were filtered according to the ITU-T recommendation P.341 [18]. We used the original Aurora 4 maskers as additive noise (airport, car, restaurant, subway, babble, exhibition, street, train). To ensure that the DNN does not overfit to the relatively short noise files from the Aurora 4 multi-condition training set, we also added additional noises from the Bits and Pieces sound effects library [19] that are similar to the original selection.

	ITU-P.563	ANIQUE+	SRMRnorm	MTD	MTD-VAD	Gini	Gini-VAD
clip	0.861	0.857	0.937	0.977	0.979	**0.992**	0.981
echo	0.524	0.746	0.758	0.867	0.890	0.814	**0.902**
chop	0.623	0.552	0.670	**0.816**	0.802	0.696	0.774
noise	0.831	0.837	0.860	0.807	**0.956**	0.896	0.954
compspkr	0.632	0.597	-0.189	-0.245	0.620	**0.636**	0.393
average	0.710	0.748	0.806	0.867	**0.906**	0.849	0.903

Table 1. Pearson correlation coefficients between MOS values from the TCD-VoIP speech quality corpus and the DNN-based model (right half of the table), and the baseline measures (columns 2-4). Averages are calculated without the competing speaker (compspkr) condition, since almost all correlation values in that condition had p values above 0.05. Only ITU-P.563 and the Gini measure without VAD showed significant correlations for compspkr with $p = 0.0498$ and $p = 0.0480$, respectively.

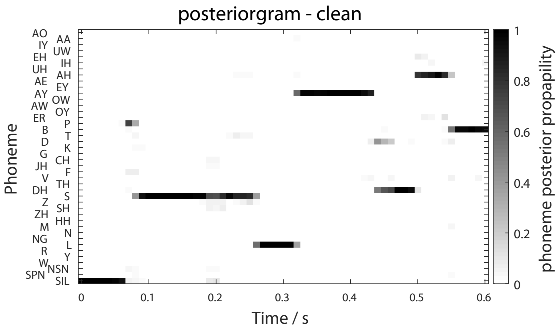

Fig. 2. Phoneme posteriorgram for a clean speech segment.

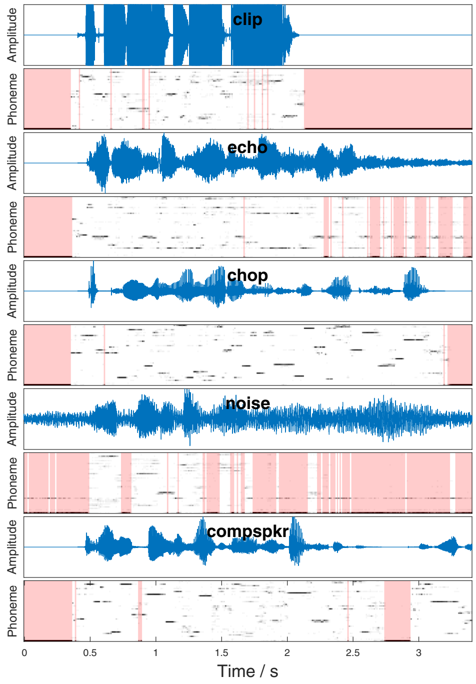

Fig. 3. Waveform and corresponding phoneme posteriorgrams for the five conditions of the TCD-VoIP database. The signal with the lowest MOS is selected in each condition. Highlighted areas denote frames with the highest posterior probability for the silence class (SIL), which are discarded for VAD-based processing. The phoneme order and the color mapping are identical to Figure 2 and omitted here for better readability.

2.3. Subjective listening data

The model was evaluated using the TCD-VoIP corpus [14]. This database contains subjective quality ratings in the presence of different degradations that can occur in VoIP applications. While the clean speech is also available from the database, only the corrupted signals were used in our experiments. In the following all conditions are briefly described:

Clipping effects: The time signal is multiplied with a factor between 1 and 55, causing some portion of the samples to be clipped (i.e. set to 1 or -1). *Echo effects*: One ore more copies of the signal are added to the original signal with a delay between 0 and 220 ms and a relative amplitude of the first delayed version related to the original between 0 and 0.5. *Chopped speech*: Samples with a length between 20 and 40 ms are either replaced with zeros, deleted entirely or overwritten with the previous portion of samples at a rate of 0 to 6 chops/s. *Background noise*: Car, street, office and babble noise are added to the signal at SNRs between 5 and 55 dB. The noise files are taken from [20]. *Competing speakers*: Two speakers (female/male, female/female, or male/male) talking in the background at SNRs between 10 and 50 dB. The target speech starts 500 ms before the competing speakers.

For each condition 20 parameter combinations are tested (except for clip and compspkr with ten parameter combinations). All subjective data is recorded accordingly ITU-T Rec. P.800 [1] with 13 male and 11 female normal-hearing subjects (except for the echo condition with 17 males and 7 females). Before the actual measurement was performed, subjects listened to speech files that represented the best and the worst speech quality contained in the test material.

3. RESULTS

Figure 2 shows the posteriorgram for a clean speech segment from the TCD-VoIP corpus with clear, distinct phoneme activations. Figure 3 shows posteriorgrams of distorted speech segments with the lowest perceived SQ per condition. In this figure, multiple phonemes are often activated in parallel as a result of the distortion. For instance, the posteriorgram for the noise condition (additive road noise at +5dB SNR corresponding to a MOS of 1.4) shows long activations for the silence state (SIL) and in parallel often at the *HH* phoneme (/h/ as in high). Ideally, the additive background noise should have activated the NSN (non-spoken-noise) model. However, the DNN is trained using the original labels for read speech recorded with a close-talk microphone in a quiet environment, i.e., the NSN labels are rarely used in the training data. Instead, for the database with added noise, the SIL class also covers noise-only segments, i.e., the result is a more general class that also covers noisy speech pauses due to the multi-condition training. To exclude these segments from further processing, the frames in which the SIL class assumes the highest value are discarded from further analysis for the two models MTD-VAD and Gini-VAD.

Table 1 shows the correlation values between the baseline measures and the MOS together with the correlations with the DNN-based MTD and with the Gini measure.

Among eight conditions (four degradations omitting competing speaker × two measures), the use of DNN-based VAD improves predictions in six conditions, which is also reflected in the improved average values by comparing the measures with and without VAD. In the competing speakers condition, it seems that the VAD does not distinguish between the target and the background speakers. Naturally, a lot of phonemes get activated from the compspkr noise so that frames only get discarded when both speakers remain silent.

The different DNN-based model implementations result in consistently high correlation values that are higher than the baseline scores in each single condition. In the presence of a competing speaker, both the baselines and the proposed DNN-based models fail to accurately predict the perceived SQ. Only ITU-P.563 and the Gini measure without VAD reach a significant correlation. This condition is per se problematic for models that are speech-specific but at the same time speaker-independent (as the DESQ model), since it requires the model to distinguish between target and the background speech while it was trained to ignore speaker-specific differences.

4. DISCUSSION

In the following, we discuss the effects of the VAD, differences between the two analyzed performance measures, as well as future work and application scenarios.

Effect of speech pauses on listeners' ratings: The MOS ratings from the subjects are based on the full audio files including segments without speech. These regions might also have influence the subjective ratings, since listeners are able to get a clear glimpse of noise-only segments and therefore could extrapolate their effect on speech quality. Hence, a removal of these segments with a VAD might remove information that played a role in the subjects rating. Nevertheless, the subjects are asked to rate the speech quality and not the general audio quality, i.e., there should be a bias towards a higher importance of active speech regions, and segments without speech should only play a minor role. This is consistent with the result that the highest improvements by using the VAD are observed in the noise condition, in which many noise-dominated segments are removed and therefore not accessible to the model.

Differences between the performance measures: Overall, the Gini measure produces results that are very similar to the previously used MTD. The Gini measure works on each frame separately and is therefore computationally less expensive than the MTD that requires multiple comparisons over a time span of 800 ms. On a typical workstation, the computational time with the Gini measure is approximately 70 times lower than with MTD. One conceptual flaw of the Gini measure is that it does not take the temporal context into account that we know to be relevant due to coarticulation effects [11] and the limited duration of phonemes. While the MTD penalizes implausible long phonemes (such as the HH activations in the noise condition in Figure 3), this is not the case for the Gini measure. These implausible long phonemes often occur in the presence of stationary background noise. Since the SIL model reflects these noises (as discussed in the results section) the posteriorgram-based VAD resolves this issue for the most part. This explains why MTD-VAD is the best system for this condition (cf. Table1). Although the temporal smearing of phoneme activations that motivated the MTD seems to be an important factor, the good performance with the Gini measure (especially in the echo condition) show that the sparseness of the different phonemes in each time frame might also be relevant. In future research, both measures should be combined to investigate if they carry complementary information for measuring the degradation of posteriorgrams.

Potential application scenario: The reason why we are interested in a small computational footprint of the model is a potential application in the context of hearing aids: Reference-free perception models could be used for online-monitoring of the speech quality in the acoustic scene and for selecting a speech enhancement algorithm (among several algorithms implemented on a hearing aid) that is optimal for this specific scene. In this context, the required time window for obtaining reliable estimates of speech quality needs to be explored. Even if this time window has a duration of several seconds (which is in the range of the duration of utterances used in this study), the approach could still be applicable for parameters that only change on longer time scales, e.g., for choosing the speech dereverberation algorithm that maximizes the perceived quality in a specific room.

Future work: In this study we used the Pearson correlation between the MOS and the MTD to evaluate the models. Future research should also consider the uncertainty of the subject ratings r_{sig} as well as the ϵ-RMSE to determine the differences between subjective ratings and models in terms in terms of MOS differences as proposed in [21], and these evaluation measures should be obtained for a wider range of test corpora. To obtain a mapping from model output values to the MOS, the training — material that was used for the subjective listening tests for adapting the listeners to the range of quality degradations (cf. end of Section 2.3) — could be used to adapt our model to the expected range as well. However, for application scenarios as outlined above, an *absolute* prediction of the MOS is not required: When the aim is to select the best speech enhancement algorithm for an acoustic scene, a monotonic relation between the model and the perceived speech quality is sufficient for selecting the best algorithm.

5. SUMMARY

The DESQ model based on deep machine learning for the prediction of speech quality was analyzed in this paper, which is a single-ended approach that uses degraded acoustic signals as input. As an extension to previous work, we investigated the effects of a DNN-based VAD on the predictive power of the model, used an alternative to quantify the degradation of phoneme posteriorgrams that was suggested as a measure of sparseness (the Gini measure), and compared the results with state-of-the-art baselines. Overall performance was found to be very similar on average for the Gini measure and the previously used MTD, with the MTD resulting in a slightly higher correlation while the Gini-based DESQ model requires less computational resources. The use of the VAD increased average scores and should be used in future models. Independently of the specific performance measure used, the DESQ models outperformed (in term of correlation) three baseline models (ITU-P.563, ANIQUE+ and SRMRnorm) in four of the five conditions of this particular test database. Since our approach is based on a speaker-independent acoustic model, it is not suited for predicting the speech quality in the presence of a competing speaker, but produces good results for clipped, reverberant, noisy, and chopped speech.

6. REFERENCES

[1] ITU-T, “Recommendation P.800: Methods for subjective determination of transmission quality,” *International Telecommunication Union, Geneva*, 1996.

[2] ITU-T, “Recommendation P.862,Perceptual Evaluation of Speech Quality (PESQ), An Objective Method for Endtoend Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs”,” *International Telecommunication Union, CH-Geneva*, 2011.

[3] ITU-T, “Recommendation P.863,Perceptual Objective Listening Quality Assessment (POLQA),” *International Telecommunication Union, CH-Geneva*, 2011.

[4] ITU-T, “Recommendation P.563: Single-ended method for objective speech quality assessment in narrow-band telephony applications,” *International Telecommunication Union, Geneva*, 2004.

[5] Doh-Suk Kim and Ahmed Tarraf, “ANIQUE+: A new American national standard for non-intrusive estimation of narrowband speech quality,” *Bell Labs Technical Journal*, vol. 12, no. 1, pp. 221–236, 2007.

[6] Joo F. Santos, Mohammed Senoussaoui, and Tiago H. Falk, “An Updated Objective Intelligibility Estimation Metric for Normal Hearing Listeners under Noise and Reverberation,” in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, September 2014.

[7] Tiago H. Falk, Chenxi Zheng, and Wai-Yip 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.

[8] Sebastian Möller, Wai-Yip Chan, Nicolas Cote, Tiago H Falk, Alexander Raake, and Marcel Waltermann, “Speech Quality Estimation: Models and Trends,” *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 18–28, nov 2011.

[9] Rainer Huber, Jasper Ooster, and Bernd T. Meyer, “Single-Ended Speech Quality Prediction Based on Automatic Speech Recognition,” *J. Audio Eng. Soc*, vol. 66, no. 10, pp. 759–769, 2018.

[10] Rainer Huber, Melanie Krüger, and Bernd T. Meyer, “Single-ended prediction of listening effort using deep neural networks,” *Hearing Research*, vol. 359, pp. 40–49, mar 2018.

[11] Hynek Hermansky, Ehsan Variani, and Vijayaditya Peddinti, “Mean temporal distance: Predicting ASR error from temporal properties of speech signal,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. may 2013, pp. 7423–7426, IEEE.

[12] Sri Harish Mallidi and Hynek Hermansky, “Novel neural network based fusion for multistream ASR,” in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. mar 2016, pp. 5680–5684, IEEE.

[13] Hans-Günter Hirsch and David Pearce, “The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions,” in *ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)*, 2000.

[14] Naomi Harte, Eoin Gillen, and Andrew Hines, “TCD-VoIP, a research database of degraded speech for assessing quality in VoIP applications,” *2015 7th International Workshop on Quality of Multimedia Experience, QoMEX 2015*, 2015.

[15] Jasper Ooster, Rainer Huber, and Bernd T. Meyer, “Prediction of Perceived Speech Quality Using Deep Machine Listening,” in *Proc. Interspeech 2018*, 2018, pp. 976–980.

[16] Leo Breiman, Jerome Friedman, RA Olshen, and Charles J Stone, “Classification and regression trees,” 1984.

[17] Bernd T. Meyer, Sri Harish Mallidi, Angel Mario Castro Martinez, Guillermo Paya-Vaya, Hendrik Kayser, and Hynek Hermansky, “Performance monitoring for automatic speech recognition in noisy multi-channel environments,” in *IEEE Workshop on Spoken Language Technology*, 2016, pp. 50–56.

[18] ITU-T, “Recommendation P.341 Transmission characteristics for wideband digital loudspeaking and hands-free telephony terminals,” p. 30, 2011.

[19] “Bits and Pieces Sound Effects Library,” www.bitsandpieces.co.uk, Accessed: 2018-03-01.

[20] European Telecommunications Standards Institute, “Speech quality performance in the presence of background noise - part 1: Background noise simulation technique and background noise database,” Tech. Rep. ETSI EG 202 396-1, 2008.

[21] Tiago h. Falk, Vijay Parsa, João F. Santos, Kathryn Arehart, Oldooz Hazrati, Rainer Huber, James M. Kates, and Susan Scollie, “Objective Quality Prediction for Users of and Intelligibility Assistive Listening Devices,” *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 114–124, 2015.