An Instrumental Quality Measure for Artificially Bandwidth-Extended Speech Signals

Johannes Abel, Magdalena Kaniewska, Cyril Guillaumé, Wouter Tirry, and Tim Fingscheidt, *Senior Member, IEEE*

***Abstract*—Various studies have shown that the instrumental measures wideband PESQ and POLQA are not reliably predicting speech quality for artificial speech bandwidth extension (ABE) test conditions, as this has never been their scope. Based on data from a coordinated subjective listening test with 12 ABE variants developed by 6 different institutions, conducted in 4 languages, we propose in this work a novel instrumental quality measure that is specifically suited for narrowband-to-wideband ABE test conditions. In particular, our contributions are fourfold: First, we propose quality indicators particularly being able to detect ABE-related distortions. Second, we investigate the combination of perceptually and nonperceptually motivated distortion-related statistics. Third, we propose a support-vector-machine-based high-performance MOS predictor for ABE speech quality assessment, finally, we present the training process based on the subjective listening test data. A k-fold cross-validation test on 1) disjoint languages, 2) disjoint speakers, and 3) disjoint ABE solutions proves the superiority of our proposed measure in the ITU-T-recommended categories accuracy, consistency, and linearity compared to both, wideband PESQ and POLQA.**

***Index Terms*—Artificial speech bandwidth extension, objective speech quality assessment, perceptual model.**

I. Introduction

SPEECH quality degradation in telephone calls originates from numerous reasons. Environmental influences, such as additive background noise or the interference of speech signals into the transmitting path (acoustic echoes) can cause severe degradation in terms of speech quality. Furthermore, device characteristics of the employed handsets often result in a non-flat frequency response of the transmission system causing additional distortion to the received speech signal. Additionally, coding of speech is usually performed by lossy compression algorithms. Most speech enhancement approaches try to compensate for these quality degradation factors, however, operating at the sampling frequency determined by the transmission.

Manuscript received May 12, 2016; revised September 6, 2016 and November 17, 2016; accepted November 17, 2016. Date of publication December 1, 2016; date of current version January 10, 2017. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Richard Christian Hendriks.

J. Abel and T. Fingscheidt are with the Institute for Communications Technology, Technische Universität Braunschweig, D - 38106 Braunschweig, Germany (e-mail: j.abel@tu-bs.de; t.fingscheidt@tu-bs.de).

M. Kaniewska, C. Guillaumé, and W. Tirry are with NXP Semiconductors, 3001 Leuven, Belgium (e-mail: magdalena.kaniewska@nxp.com; cyril.guillaume@nxp.com; wouter.tirry@nxp.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2016.2635022

Still most of today's phone calls are established at a sampling rate of 8 kHz, i.e., a theoretical acoustical bandwidth of up to 4 kHz. While syllable intelligibility in this so-called narrowband (NB) speech is at around 90% [1], up to 98% can be reached by an acoustical bandwidth of 7 kHz at a sampling frequency of 16 kHz. This is referred to as wideband (WB) speech. Especially the intelligibility of fricative sounds, exhibiting most of their energy at frequencies above 4 kHz, is increased. Furthermore, speech quality, measured on the 5-point mean opion score (MOS) [2] scale (1: "bad", 5: "excellent"), rises when switching from NB to WB speech [3]. The synonym for telephony calls based on WB-capable speech codecs is HD Voice. Successfully establishing an HD Voice call, however, depends on a lot of factors: (1) near-end and far-end handsets need to be WB-capable, (2) the network cell infrastructure needs to support HD Voice, and (3) the provider's backbone network must be capable of WB (tandem speech coding or transcoder-free) operation throughout the entire connection. Further requirements must be met to establish international mobile HD Voice calls and to support inter-operator handovers. If any of these prerequisites is not met, the call can only be established in NB mode.

Bridging the transition from NB- to a purely WB-based telecommunication service, artificial speech bandwidth extension (ABE) algorithms aim at restoring the missing upper band (UB), i.e., the frequency components between 4 and 8 kHz, which have either never been captured by the microphone or have been omitted during NB transmission. ABE solutions are a downlink feature, i.e., they are employed in the receiving path of the transmission. ABE solutions often split the extension by means of the source-filter model for speech production into two sub-problems: calculation of an UB residual signal and an UB spectral envelope. While estimation of the UB residual typically utilizes simple modulation techniques of the NB residual, the UB spectral envelope estimation might be codebook-based [4], Gaussian mixture model-based [5], [6], hidden Markov model-based [7]–[10], or neural network-based [11]. Direct log spectra estimation was investigated using sum product networks, restricted Boltzman machines (RBMs), conditional RBMs, (deep) autoencoders, and sum-product networks [12], [13]. A deep neural network (DNN) is employed in [14] to directly estimate the UB log power spectrum, while in [15], [16] a cepstral representation of the UB is estimated using a DNN.

In [17] a higher intelligibility especially for the important /s/ sound [18] could be shown. Regarding speech quality, several subjective listening tests [3], [11], [19]–[21] showed an increased speech quality for at least some ABE solutions. In

these conducted listening tests the ABE input NB signals were all clean speech data subject to speech transcoding, and in some cases also further preprocessing steps were taken, applying device and/or transmission characteristics to the NB speech signal.

To evaluate subjective speech quality in general, the International Telecommunications Union (ITU-T) defined several test methodologies (P.800, [2]). Since subjective tests demand a huge amount of time and financial resources, instrumental quality measures predicting an average human vote have been developed to overcome these disadvantages. Obviously, instrumental measures aim at having only a small deviation between their estimated values (objective listening quality, dubbed $\mathrm{MOS_{LQO}}$) and the ground truth subjective votes (subjective listening quality, dubbed $\mathrm{MOS_{LQS}}$), obtained from an absolute category rating (ACR) listening test. Well-known reference-based measures are perceptual evaluation of speech quality (PESQ, P.862) [22] and its still heavily used WB extension WB-PESQ (P.862.2) [23]. Amongst other improvements, the successor perceptual objective listening quality assessment (POLQA, P.863) [24] is also capable of processing speech signals with even higher sampling frequencies. These algorithms have been proven to be very reliable in terms of predicting accurate $\mathrm{MOS_{LQO}}$ values w.r.t. a low root mean square error (RMSE) and high correlation coefficients, but they have not been developed for any use with ABE-processed speech signals. Besides reference-based measures, so-called non-intrusive measures exist, which do not require a reference signal for $\mathrm{MOS_{LQO}}$ estimation.

According to [23], WB-PESQ suffers from lower prediction accuracy when the signal under test has severe bandwidth limitations. POLQA, however, claims to be able to operate well enough on bandwidth-extended[1] speech signals [24]. In [20] two ABE solutions with different parameterization were evaluated in a subjective listening test and results were compared to $\mathrm{MOS_{LQO}}$ values predicted by WB-PESQ and POLQA. Pulakka *et al*. found that both instrumental measures performed acceptably, while in their study POLQA outperformed WB-PESQ w.r.t. the correlation coefficient. It was indicated, however, that both instrumental measures might not accurately predict the rank order of the ABE conditions under test. Furthermore, in [25] five ABE solutions were evaluated in a subjective listening test and compared to the results from WB-PESQ and POLQA. For WB-PESQ a correlation of 0.93 and for POLQA of 0.87 was found. In the subjective listening test, a statistically significant difference between NB and the ABE solutions could not be shown. In [19] six ABE conditions were part of a subjective listening test. The results show a correlation to the predicted subjective scores from WB-PESQ of 0.82 and from POLQA of only 0.75. In a statistical analysis one of the ABE conditions under test was found to be better than the underlying NB condition. Concluding on these prior works, the capability of WB-PESQ and POLQA to accurately predict $\mathrm{MOS_{LQO}}$ values for ABE-processed signals is questionable. In scientific publications, often spectral or cepstral distortion measures have been used [8], [14]–[16], [26]. MOS-predicting measures, however, have the big advantage of providing ratings that can be qualitatively understood and compared by non-experts. Moreover, correlation to subjective listening test results is typically reported for these measures. Obviously, a robust instrumental measure for ABE signals would be highly desirable for those being concerned with development and evaluation of ABE algorithms.

In this work, a reference-based measure for quality assessment of artificially bandwidth-extended speech signals is presented, focusing on the upper band extension. The main contributions are: (1) new quality indicators particularly being able to detect ABE-related distortions, (2) a combination of perceptually and non-perceptually influenced distortion-related statistics, (3) a high performance MOS predictor for linking the measured distortions to an adequate speech quality score, and (4) model parameter training based on a sophisticated subjective listening test [3], conducted in 4 different languages including ABE solutions from 6 different institutions and consortia, leading the field of ABE research.

The article is structured as follows: For ease of presentation, abbreviations for some distortion-related statistics are defined in Sec. II. The underlying listening test, serving as ground truth for development and evaluation of the instrumental measure, is briefly sketched in Sec. III. Afterwards, the algorithmic design of the instrumental measure is explained in Sec. IV, including the description of the newly developed quality indicators and the employed MOS predictor. In Sec. V evaluation metrics and three cross-validation experiments are outlined. The results of the proposed instrumental measure for the three experiments are reported and compared to the existing measures POLQA and WB-PESQ. Finally, conclusions are drawn in Sec. VI.

II. Notations of Distortion-Related Statistics

For better reading, some distortion-related statistics that are frequently used throughout this work are abbreviated as follows. First of all,

$$\mu(D(z);\mathcal{Z}) := \frac{1}{|\mathcal{Z}|}\sum_{z\in\mathcal{Z}} D(z),$$

is the mean of a distortion-related entity $D(z)$, calculated over all frames and/or frequency bands $z \in \mathcal{Z}$, with $|\mathcal{Z}|$ denoting the cardinality of set $\mathcal{Z}$. Secondly,

$$\sigma^2(D(z);\mathcal{Z}) := \frac{1}{|\mathcal{Z}|}\sum_{z\in\mathcal{Z}} \big(D(z)-\mu(D(z);\mathcal{Z})\big)^2,$$

denotes the variance of $D(z)$, calculated over all $z \in \mathcal{Z}$. Accordingly,

$$\sigma(D(z);\mathcal{Z}) = \sqrt{\sigma^2(D(z);\mathcal{Z})}$$

abbreviates the standard deviation of $D(z)$, calculated over all $z \in \mathcal{Z}$. Here $D(z)$ is only dependent on z and the statistics is computed over set $\mathcal{Z}$. If the entity D depends on two variables ($D(y,z)$), the statistics could be calculated over set $\mathcal{Y}$, set $\mathcal{Z}$, or over both sets $\{\mathcal{Y},\mathcal{Z}\}$. This means that

$$\mu(D(y,z);\mathcal{Z})$$

[1]not: *artificially* bandwidth-extended

TABLE I
OVERVIEW OF CONDITIONS IN THE SUBJECTIVE LISTENING TEST

c	Conditions
$1-6$	NB-MNRU: 6, 12, 18, 24, 30, ∞ dB
$7-12$	WB-MNRU: 5, 15, 25, 35, 45, ∞ dB
13	AMR-NB @ 12.2 kbps
$14-25$	$\mathrm{ABE}_{01}, \ldots, \mathrm{ABE}_{12}$
$26-28$	AMR-WB @ 8.85, 23.05, 23.85 kbps

provides a mean value for each $y \in \mathcal{Y}$, while

$$\mu(D(y,z);\mathcal{Y},\mathcal{Z}) := \mu(\mu(D(y,z);\mathcal{Y});\mathcal{Z})$$

results in a single mean value. The same holds for calculation of variance and standard deviation.

Finally, Pearson's correlation coefficient [22] ρ for $D(z)$ and $\hat{D}(z)$, calculated over all $z \in \mathcal{Z}$, is given as

$$\rho(D(z),\hat{D}(z);\mathcal{Z}) := \frac{\sum_{z\in\mathcal{Z}}(D(z)-\mu(D(z);\mathcal{Z}))\cdot\left(\hat{D}(z)-\mu(\hat{D}(z);\mathcal{Z})\right)}{\sqrt{\sum_{z\in\mathcal{Z}}(D(z)-\mu(D(z);\mathcal{Z}))^2}\cdot\sqrt{\sum_{z\in\mathcal{Z}}\left(D(z)-\mu(\hat{D}(z);\mathcal{Z})\right)^2}}.$$

III. UNDERLYING SUBJECTIVE LISTENING TEST

In order to develop an instrumental measure for artificially bandwidth-extended speech signals, a subjective listening test is needed to deliver ground truth $\mathrm{MOS}_{\mathrm{LQS}}$ values. Such a test is supposed to reflect some variety of ABE solutions to model the link between ABE-processed speech signals and subjective votes. In order to achieve this variety, Technische Universität Braunschweig and NXP Software cooperated with six different institutions or consortia that processed speech data using their ABE solutions [5], [7], [10], [11], [26], [27]. Some of them provided more than one solution or more than one parameterization of their solution, leading to a total of 12 different ABE variants ABE_{01}, ..., ABE_{12}. Please note that only ABE solutions with a maximum algorithmic latency of 30 ms were considered. In the final ACR test setup, presented in Table I, the ABE conditions are accompanied by so-called anchor conditions, namely NB- and WB-modulated noise reference units (MNRUs) at different speech-to-modulated noise power ratios as well as adaptive multirate NB (AMR-NB)- and AMR-WB-coded speech at different AMR bit rates. AMR-NB-coded speech served as input to all 12 ABE variants. In addition, language dependency of the perceived subjective quality shall be considered, thus conditions were tested in English, Chinese (Mandarin), German and Korean. A speech corpus, providing high-quality speech data for all languages under test was acquired from SpeechOcean [28], providing an identical recording environment thus ensuring comparability throughout the languages.

In each language, two female (F1, F2) and two male (M1, M2) speakers provided the sentences under test. Even though these speakers differ across languages, for ease of presentation, we will not use different notations of speakers in each language. Each speaker in the database contributed four utterances, consisting of two sentences with a short pause between them. One of those utterances was used during the familiarization phase of the subjective test and is therefore not considered in the remaining part of this paper.

Due to the huge amount of conditions under test, two separate subtests were conducted in *each* of the languages. Those subtests contained all of the anchor conditions, but only half of the ABE conditions. To enable a joint view onto the results, the subjective scores obtained from the subtests needed to be merged. Simply adding them would bias the statistical properties, thus a mapping was necessary before merging them. Since the anchor conditions are common ground of both subtests, they served as reference point for calculating linear regression coefficients. For each language, a global mean per anchor condition over both subtests was calculated. Then linear regression coefficients were found, mapping *only* the anchor conditions from each subtest towards this global mean. Subsequently, the subjective votes for *all* of the conditions including the ABE conditions were subject to the respective linear regression. Finally, the mapped subjective votes from the subtests were merged for each condition. More details of the underlying listening test are not necessary for the subject of this work; the interested reader may consult [3] for further details.

Each speech signal being used in the listening test was evaluated by more than one subject. In the context of instrumental measures for speech quality assessment, these multiple votes per speech signal can only be processed as one average value per signal. Therefore, in the following, $\mathrm{MOS}_{\mathrm{LQS}}$ will refer to the mean of individual votes given to the respective speech signal.

The speech signals presented in the listening test along with the $\mathrm{MOS}_{\mathrm{LQS}}$ will be taken up again in Sec. V-A, where the training of the MOS predictor is described.

IV. THE INSTRUMENTAL MEASURE FOR BANDWIDTH-EXTENDED SPEECH SIGNALS

In this section, the proposed approach to instrumental quality assessment of artificially bandwidth-extended speech signals is presented (see Fig. 1).

After an initial signal preprocessing step described in Sec. IV-A, frame-wise calculated quality indicators (QI) are obtained which constitute a comparison of the reference to the degraded signal particularly focusing on typical errors evoked from ABE algorithms. QIs are calculated on time domain representations (Sec. IV-B) or on perceptually-processed frequency domain representations (Sec. IV-C) of the reference and degraded speech signal. Features are derived from the presented QIs by calculating their mean and variance over time, however, some QIs directly result in a single value and do not need further time integration. Finally, the obtained features are concatenated to feature vector $\mathbf{x}$, which serves as input to the MOS predictor, explained in Sec. IV-D.

In a preceding *training phase*, a parameterization of the MOS predictor will be found, establishing a link between the observed feature vector $\mathbf{x}$ and the respective $\mathrm{MOS}_{\mathrm{LQS}}$ value from the subjective listening test. Once the parameterization is found, the MOS predictor will compute $\mathrm{MOS}_{\mathrm{LQO}}$ values based on feature vectors during the *operative phase* as shown in Fig. 1.

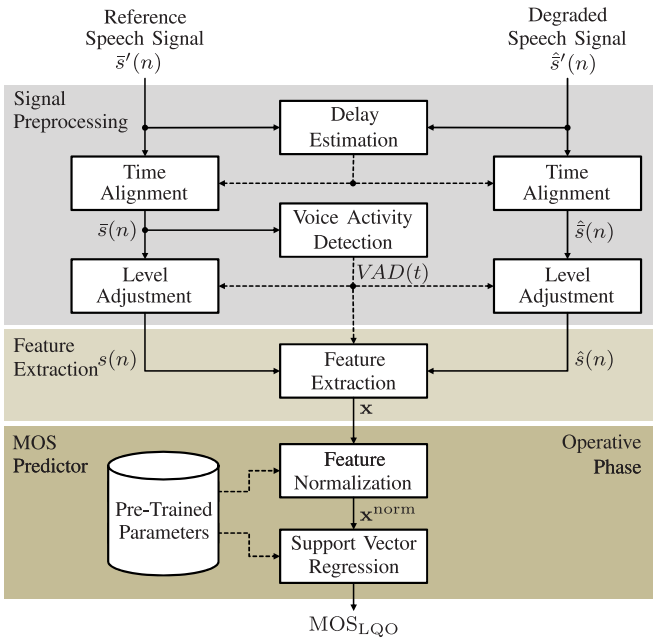

Fig. 1. **High-level processing overview** of the proposed instrumental speech quality measure for artificially bandwidth-extended speech signals.

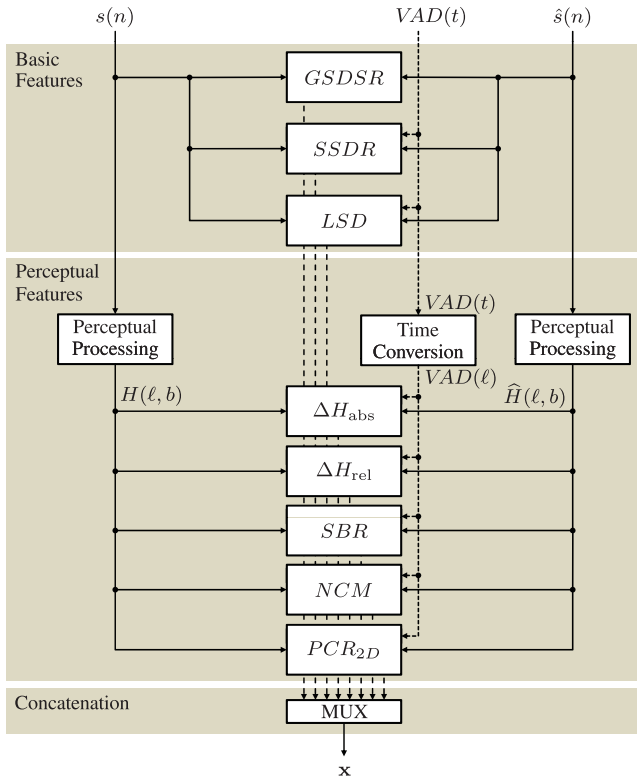

Fig. 2. Detailed overview of the processing for **feature extraction**. Time-aligned and level-adjusted signals $s(n)$ and $\hat{s}(n)$ as well as the frame-wise voice-activity information $VAD(t)$ are input to the feature extraction. First, basic features are calculated. Afterwards perceptual features are extracted, and finally all features are concatenated into feature vector $\mathbf{x}$.

A. Signal Preprocessing

The proposed instrumental quality measure inputs a quality-degraded speech signal $\hat{\bar{s}}'(n)$, i.e., the bandwidth-extended speech signal, along with a respective reference speech signal $\bar{s}'(n)$, with n being the sample index for signals sampled at a frequency of $f_s = 16$ kHz. Please note that the reference signal is uncoded and free of any distortion. Based on the input signals, the instrumental measure outputs an MOS_{LQO} value.

As depicted in Fig. 1, processing starts with a delay estimation and time alignment of the input speech signals. This is done by means of maximizing the correlation and potential removal of initial and final samples. After this both resulting signals $\bar{s}(n)$ and $\hat{\bar{s}}(n)$ share a common length $|\mathcal{N}|$ with sample index $n \in \mathcal{N} = \{0, 1, \ldots, |\mathcal{N}|-1\}$.

The reference speech signal $\bar{s}(n)$ is then subject to a simple voice activity detection (VAD), computing

$$VAD(t) = \begin{cases} 1, & \text{if } \dfrac{\mu(\bar{s}^2(n);\ \mathcal{N}(t))}{\mu(\bar{s}^2(n);\ \mathcal{N})} > \theta_{\text{VAD}} \\ 0, & \text{else,} \end{cases} \tag{1}$$

with $\theta_{\text{VAD}} = 0.0001$ and $\mathcal{N}(t) = \{n \mid (t-1)L \leq n \leq tL-1\}$ defining the sample indices of frame $t \in \mathcal{T} = \{1, ..., |\mathcal{T}| = \lfloor |\mathcal{N}|/L \rfloor\}$. Note that $\lfloor . \rfloor$ specifies the floor operator, $|\mathcal{T}|$ being the amount of frames included in the set $\mathcal{T}$. The frame length is set to $L = 256$ samples per frame, i.e., 16 ms at a sampling frequency of $f_s = 16$ kHz. Subsequently, we identify two sets of frame indices: $\mathcal{T}_1 = \{t \mid VAD(t) = 1\}$ and $\mathcal{T}_0 = \{t \mid VAD(t) = 0\}$.

The time-aligned speech signals $\bar{s}(n)$ and $\hat{\bar{s}}(n)$ are then subject to a modified active speech level normalization to match -26 dBov (c.f. P.56, [29]), only considering frequency components below 4 kHz. This modification eliminates the influence of the UB during level adjustment and takes into account that a synthetic UB might be under- or overestimated w.r.t. its energy by an ABE solution and thus normalization of this energy might interfere with the necessity to determine the degree of such degradation. To accomplish all this, the reference and degraded speech signals are subject to a high-order finite impulse response low-pass filter with cut-off frequency at 4 kHz and subsequent decimation by a factor of two. Based on these downsampled signals, the scaling factors α_s and $\alpha_{\hat{s}}$ are determined by applying ITU-T P.56 [30]. Level adjustment is then performed by $s(n) = \alpha_s \cdot \bar{s}(n)$ and $\hat{s}(n) = \alpha_{\hat{s}} \cdot \hat{\bar{s}}(n)$, respectively. The resulting speech signals provide the input for subsequent feature extraction.

B. Feature Extraction: Basic Features

Figure 2 depicts the detailed block diagram of the feature extraction (cf. Fig. 1). The time-aligned and level-adjusted speech signals $s(n)$ and $\hat{s}(n)$, as well as the voice activity information $VAD(t)$ serve as input to the feature extraction. The result is a single composite feature vector $\mathbf{x}$ per speech file, representing different aspects of the distortion contained in the degraded speech signal.

1) Global Signal-to-Degraded-Speech Ratio (GSDSR): The global signal-to-degraded-speech ratio (GSDSR), adapted from

[31], is defined as

$$GSDSR = 10\log_{10}\left(\frac{\sum_{n\in\mathcal{N}} s^2(n)}{\sum_{n\in\mathcal{N}} \hat{s}^2(n)}\right). \quad (2)$$

This value is directly used as one of the features in vector $\mathbf{x}$. Due to its 'global SNR' nature, it is able to detect high energy (short-term) disturbances in the degraded file.

2) Segmental Speech-to-Speech Distortion Ratio (SSDR): The QI segmental speech-to-speech distortion ratio (SSDR) is based on a frame-wise comparison of the two input speech signals according to [32]. A lookahead and a look-back are not employed, therefore the analysis window length is L as well, with a frame shift of $L_s = L$. It is calculated by first deriving

$$SSDR'(t) = 10\log_{10}\left(\frac{\sum_{n\in\mathcal{N}(t)} s(n)^2}{\sum_{n\in\mathcal{N}(t)} \left(\hat{s}(n) - s(n)\right)^2}\right),$$

and subsequent limitation steps resulting in $SSDR(t) = \max\{\min\{SSDR'(t), 30\text{ dB}\}, -10\text{ dB}\}$. Calculated over voice-active frame set $\mathcal{T}_1$, the mean and variance

$$\mu(SSDR(t);\ \mathcal{T}_1), \quad (3)$$

$$\sigma^2(SSDR(t);\ \mathcal{T}_1), \quad (4)$$

can detect energy and phase errors over all frequencies. On the other hand, deriving a mean and variance feature only over voice-inactive frame set $\mathcal{T}_0$,

$$\mu(SSDR(t);\ \mathcal{T}_0), \quad (5)$$

$$\sigma^2(SSDR(t);\ \mathcal{T}_0), \quad (6)$$

allows to identify ABE reconstruction artifacts particularly for background noise.

3) Logarithmic Spectral Distance (LSD): The QI logarithmic spectral distance (LSD) is employed according to [33]. In contrast to the SSDR QI, a lookahead and a look-back of $L_- = L_+ = 128$ samples is used, hence the analysis window length is $N_w = 512$. To tackle the spectral leakage effect, a Hamming window of length N_w is applied to the extended speech frames, yielding $s_w(n)$ and $\hat{s}_w(n)$, $n \in \mathcal{N}_w(t) = \{n \mid (t-1)L - L_- \le n \le tL - 1 + L_+\}$ for the reference and the degraded speech signal, respectively. The frames of windowed speech signals $s_w(n)$ and $\hat{s}_w(n)$ are zero-padded to a length of $K = 1024$ and subsequently the short-term spectra $S(t,k)$ and $\widehat{S}(t,k)$ of the reference and degraded input signal are derived by means of the K-point discrete Fourier transform (DFT). The LSD measure is then calculated as follows:

$$LSD(t) = \sqrt{\frac{1}{k_u - k_l + 1}\sum_{k=k_l}^{k_u}\left[10\log_{10}\left(\frac{|S(t,k)|^2}{|\widehat{S}(t,k)|^2}\right)\right]^2}.$$

The lower and upper frequency bin bounds k_l and k_u limit the measure to a certain frequency range. For the implementation of the instrumental measure the frequencies between 50 Hz $\le f \le$ 7 kHz are taken into account, therefore $k_l = \lfloor \frac{K}{16000\text{ Hz}} \cdot 50\text{ Hz} \rfloor = 3$ and $k_u = \lfloor \frac{K}{16000\text{ Hz}} \cdot 7000\text{ Hz} \rfloor = 448$. This yields two further single features, derived from speech frames only ($VAD(t)$ from (1) can still be used due to the same frame rate):

$$\mu(LSD(t);\ \mathcal{T}_1), \quad (7)$$

$$\sigma^2(LSD(t);\ \mathcal{T}_1). \quad (8)$$

These features are used to identify errors in the estimated UB spectrum of the ABE solution under test. Modifications of the NB part of the degraded signal, e.g., speech coding, are also reflected.

TABLE II
DEFINITION OF BARK BANDS b [34]

		b	$f_l(b)$ [Hz]	$f_c(b)$ [Hz]	$f_u(b)$ [Hz]	$f_\Delta(b)$ [Hz]
$\mathcal{B}$	-	1	0	50	100	100
		2	100	150	200	100
		3	200	250	300	100
		4	300	350	400	100
		5	400	450	510	110
		6	510	570	630	120
		7	630	700	770	140
		8	770	840	920	150
		9	920	1000	1080	160
	$\mathcal{B}_{\text{LB}}$	10	1080	1170	1270	190
		11	1270	1370	1480	210
		12	1480	1600	1720	240
		13	1720	1850	2000	280
		14	2000	2150	2320	320
		15	2320	2500	2700	380
		16	2700	2900	3150	450
		17	3150	3400	3700	550
	-	18	3700	4000	4400	700
	$\mathcal{B}_{\text{UB}}$	19	4400	4800	5300	900
		20	5300	5800	6400	1100
		21	6400	7000	7700	1300

C. Feature Extraction: Perceptual Features

Besides basic features, the input speech signals are also subject to a processing that adopts aspects of human speech perception. Sottek's hearing model [35], [36] processes the input signals via an outer and middle ear filtering and subsequent decomposition of the signals using a Bark band filter bank [34]. Besides merging the frequency components of the input signal into critical bands $b \in \mathcal{B} = \{1, \ldots, 21\}$, human perception is taken into account by reducing the excitation level at each band to match the frequency-dependent threshold in quiet. An overview of the lower and upper edge frequency $f_l(b)$ and $f_u(b)$, as well as the center frequency $f_c(b)$ and the bandwidth $f_\Delta(b)$ for different Bark bands is given in Table II.

In Sottek's model, each Bark band b represents a set of adjacent auditory hair cells located at the corresponding area on the human cochlea. By means of the Hilbert transform a temporal envelope is calculated for each band, representing the mean firing rates, i.e., the perceptual sensation at the respective auditory hair cells over time. The auditory hair cells have a frequency-dependent maximum firing rate, meaning that per Bark band only an upper-limited information rate can be processed by humans. Therefore, the Hilbert envelopes, representing the perceptual sensation at a certain Bark band are low-pass filtered with the respective cut-off frequency. The hearing model

representations $H(\ell,b)\in\mathbb{R}^+$ and $\widehat{H}(\ell,b)\in\mathbb{R}^+$ (see Fig. 2) are the result of the perceptual processing step for the reference and the degraded speech signals, respectively. Since a higher time resolution of a 3.3 ms frame rate is used for the perceptual processing, a further frame index $\ell\in\mathcal{L}$ is introduced. The sets $\mathcal{L}_1$ and $\mathcal{L}_0$ are obtained from $\mathcal{T}_1$ and $\mathcal{T}_0$, respectively, and contain the frame indices of voice active and inactive portions of the input signals. This is achieved by assigning that $VAD(t)$ value to $VAD(\ell)$, where the center of frame ℓ is closest to the center of frame t.

Please note that frequency and Bark band are non-linearly related. The higher the Bark band index b, the larger the frequency bandwidth $f_\Delta(b)$ covered by Bark band b. Thus, when calculating mean or variance according to Section II over set $\mathcal{B}$, a compensation factor

$$P(b)=\frac{f_\Delta(b)}{\sum_{\beta\in\mathcal{B}}f_\Delta(\beta)} \tag{9}$$

has to be applied.

Based on the hearing model representations, further frame-wise calculated QIs are calculated.

1) Absolute Bark Band-Based Distortion: The QI absolute Bark band-based distortion is calculated as

$$\Delta H_{\mathrm{abs}}(\ell,b)=10\log_{10}\left(\frac{|H(\ell,b)|^2}{|\widehat{H}(\ell,b)|^2}\right).$$

General distortion features derived from this QI are

$$\mu\big(P(b)\cdot|\Delta H_{\mathrm{abs}}(\ell,b)|;\ \mathcal{L}_1,\mathcal{B}\big), \tag{10}$$

$$\sigma^2\big(P(b)\cdot|\Delta H_{\mathrm{abs}}(\ell,b)|;\ \mathcal{L}_1,\mathcal{B}\big). \tag{11}$$

To specifically detect distortions related to added $(+)$ and omitted $(-)$ energy over all Bark bands $b\in\mathcal{B}$, meaning also increased or decreased loudness, respectively, the following frame sets are defined:

$$\mathcal{L}_+=\{\ell\mid\mu(\Delta H_{\mathrm{abs}}(\ell,b);\ \mathcal{B})>0\},$$
$$\mathcal{L}_-=\{\ell\mid\mu(\Delta H_{\mathrm{abs}}(\ell,b);\ \mathcal{B})\le 0\}.$$

Please note that also frequency bands below 4 kHz influence these frame sets. Means and variances are calculated for added and omitted components ($\cap$ stands for intersection of sets):

$$\mu(P(b)\cdot|\Delta H_{\mathrm{abs}}(\ell,b)|;\ \mathcal{L}_+\cap\mathcal{L}_1,\mathcal{B}), \tag{12}$$

$$\sigma^2(P(b)\cdot|\Delta H_{\mathrm{abs}}(\ell,b)|;\ \mathcal{L}_+\cap\mathcal{L}_1,\mathcal{B}), \tag{13}$$

$$\mu(P(b)\cdot|\Delta H_{\mathrm{abs}}(\ell,b)|;\ \mathcal{L}_-\cap\mathcal{L}_1,\mathcal{B}), \tag{14}$$

$$\sigma^2(P(b)\cdot|\Delta H_{\mathrm{abs}}(\ell,b)|;\ \mathcal{L}_-\cap\mathcal{L}_1,\mathcal{B}). \tag{15}$$

To further focus specifically on over- and underestimation errors of ABE solutions w.r.t. synthesized UB energy, meaning hissing or lisping artifacts, respectively, the following frame sets for added and omitted energy over all Bark bands defined in the UB are calculated:

$$\mathcal{L}_{\mathrm{UB}+}=\{\ell\mid\mu(\Delta H_{\mathrm{abs}}(\ell,b);\mathcal{B}_{\mathrm{UB}})>0\},$$
$$\mathcal{L}_{\mathrm{UB}-}=\{\ell\mid\mu(\Delta H_{\mathrm{abs}}(\ell,b);\mathcal{B}_{\mathrm{UB}})\le 0\},$$

with $\mathcal{B}_{\mathrm{UB}}=\{b\mid 4\text{ kHz}\le f_l(b)<f_u(b)\le 8\text{ kHz}\}=\{19,20,21\}$. From this, the following features are computed:

$$\mu(P(b)\cdot|\Delta H_{\mathrm{abs}}(\ell,b)|;\ \mathcal{L}_{\mathrm{UB}+}\cap\mathcal{L}_1,\mathcal{B}), \tag{16}$$

$$\sigma^2(P(b)\cdot|\Delta H_{\mathrm{abs}}(\ell,b)|;\ \mathcal{L}_{\mathrm{UB}+}\cap\mathcal{L}_1,\mathcal{B}), \tag{17}$$

$$\mu(P(b)\cdot|\Delta H_{\mathrm{abs}}(\ell,b)|;\ \mathcal{L}_{\mathrm{UB}-}\cap\mathcal{L}_1,\mathcal{B}), \tag{18}$$

$$\sigma^2(P(b)\cdot|\Delta H_{\mathrm{abs}}(\ell,b)|;\ \mathcal{L}_{\mathrm{UB}-}\cap\mathcal{L}_1,\mathcal{B}). \tag{19}$$

2) Relative Bark Band-Based Distortion: The QI relative Bark band-based distortion, inspired by [37], is calculated as follows:

$$\Delta H_{\mathrm{rel}}(\ell,b)=10\log_{10}\left(\frac{|H(\ell,b)|^2}{\big(|H(\ell,b)|-|\widehat{H}(\ell,b)|\big)^2}\right).$$

After limiting $\Delta H_{\mathrm{rel}}(\ell,b)$ to a maximum of 45 dB, two features focusing on the UB frequency components are calculated:

$$\mu(P(b)\cdot|\Delta H_{\mathrm{rel}}(\ell,b)|;\ \mathcal{L}_1,\mathcal{B}_{\mathrm{UB}}), \tag{20}$$

$$\sigma^2(P(b)\cdot|\Delta H_{\mathrm{rel}}(\ell,b)|;\ \mathcal{L}_1,\mathcal{B}_{\mathrm{UB}}). \tag{21}$$

These features exhibit a high sensitivity to smaller distortions.

3) Spectral Balance Ratio (SBR): The spectral balance ratio (SBR) links lower and upper frequency components of both input signals and therefore is an important QI for the spectral balance as restored by ABE algorithms. First, both inputs $H(\ell,b)$ and $\widehat{H}(\ell,b)$ need to be integrated over different frequency bands. The SBR is then defined as

$$SBR(\ell)=$$
$$10\log_{10}\left(\frac{\mu(P(b)\cdot|H(\ell,b)|^2;\mathcal{B}_{\mathrm{UB}})}{\mu(P(b)\cdot|H(\ell,b)|^2;\mathcal{B}_{\mathrm{LB}})}\Big/\frac{\mu(P(b)\cdot|\widehat{H}(\ell,b)|^2;\mathcal{B}_{\mathrm{UB}})}{\mu(P(b)\cdot|\widehat{H}(\ell,b)|^2;\mathcal{B}_{\mathrm{LB}})}\right),$$

with Bark band set $\mathcal{B}_{\mathrm{LB}}=\{b|1\text{ kHz}\le f_l(b)<f_u(b)\le 4\text{ kHz}\}=\{10,\ldots,17\}$, covering frequency components from 1.08 kHz up to 3.70 kHz, thus discarding low-frequent noise in the degraded signal. Note that the term $P(b)$ in (9) employs set $\mathcal{B}_{\mathrm{LB}}$ or $\mathcal{B}_{\mathrm{UB}}$ here, respectively. For calculation of features, the following frame sets are defined:

$$\mathcal{L}_{\mathrm{SBR}+}=\{\ell\mid SBR(\ell)>0\},$$
$$\mathcal{L}_{\mathrm{SBR}-}=\{\ell\mid SBR(\ell)\le 0\}.$$

Five features quantifying the amount of imbalance in both directions as well as their relative frequency are obtained:

$$\mu(SBR(\ell);\ \mathcal{L}_{\mathrm{SBR}+}), \tag{22}$$

$$\sigma^2(SBR(\ell);\ \mathcal{L}_{\mathrm{SBR}+}), \tag{23}$$

$$\mu(SBR(\ell);\ \mathcal{L}_{\mathrm{SBR}-}), \tag{24}$$

$$\sigma^2(SBR(\ell);\ \mathcal{L}_{\mathrm{SBR}-}), \tag{25}$$

$$\frac{|\mathcal{L}_{\mathrm{SBR}+}|}{|\mathcal{L}_{\mathrm{SBR}-}|}. \tag{26}$$

4) Modified Normalized Covariance Metric (NCM): The modified normalized covariance metric (NCM) (c.f. [38]) is a correlation-based comparison of filter bank outputs resulting in a single value. While in [38] the reference and degraded speech signal are subject to a gammatone filter bank and subsequent calculation of band-wise temporal envelopes via the Hilbert transform on the filter bank outputs, we use the perceptually influenced temporal envelopes contained in the hearing model representations $H(\ell, b)$ and $\hat{H}(\ell, b)$ for further processing. First, a Bark band-dependent normalized correlation ratio (NCR), based on the correlation over time, is calculated:

$$NCR'_{\rho}(b) = 10\log_{10}\left(\frac{\rho(|H(\ell,b)|, |\hat{H}(\ell,b)|; \mathcal{L}_1)^2}{\left(1-\rho(|H(\ell,b)|, |\hat{H}(\ell,b)|; \mathcal{L}_1)\right)^2}\right).$$

After limiting the $NCR'_{\rho}(b)$ to a range of $[-15,\ 15]$ dB and subsequently scaling it to a range of $[0, 1]$ following

$$NCR_{\rho}(b) = \frac{\min\left(\max\left(NCR'_{\rho}(b), -15\text{ dB}\right), 15\text{ dB}\right) + 15\text{ dB}}{30\text{ dB}},$$

a single feature is obtained:

$$\mu(NCR_{\rho}(b), \mathcal{B}). \tag{27}$$

5) Two-Dimensional Pearson's Correlation: The Pearson's correlation coefficient ρ_{2D} for two-dimensional input signals is used to compare the two hearing model-processed signals $H(\ell, b)$ and $\widehat{H}(\ell, b)$. This feature sets the focus on the temporal and spectral progress, while precise equality of frequency components over time is less important. It is calculated as follows:

$$\rho_{2D} = \frac{\sum_{\ell\in\mathcal{L}}\sum_{b\in\mathcal{B}}\left(|H(\ell,b)| - \overline{H}\right)\left(|\widehat{H}(\ell,b)| - \overline{\widehat{H}}\right)}{\sqrt{\sum_{\ell\in\mathcal{L}}\sum_{b\in\mathcal{B}}\left(|H(\ell,b)| - \overline{H}\right)^2}\sqrt{\sum_{\ell\in\mathcal{L}}\sum_{b\in\mathcal{B}}\left(|\widehat{H}(\ell,b)| - \overline{\widehat{H}}\right)^2}},$$

with $\overline{H} = \mu(|H(\ell,b)|; \mathcal{L}, \mathcal{B})$ and $\overline{\widehat{H}} = \mu(|\widehat{H}(\ell,b)|; \mathcal{L}, \mathcal{B})$. Finally, we compute a 2D Pearson's correlation ratio (PCR) as

$$PCR'_{2D} = 10\log_{10}\left(\frac{(\rho_{2D})^2}{(1-\rho_{2D})^2}\right),$$

which will be negative for correlation coefficients $\rho_{2D} < 0.5$, and positive high-valued for correlation coefficients close to one. Final limitation to the range $[-10, 55]$ dB leads to the feature

$$PCR_{2D} = \min\left(\max\left(PCR'_{2D}, -10\text{ dB}\right), 55\text{ dB}\right). \tag{28}$$

D. MOS Predictor

In the final operative phase, the MOS predictor maps the obtained feature vector $\mathbf{x}$ to an $\mathrm{MOS}_{\mathrm{LQO}}$ value, as shown in Fig. 1. To achieve this, a training has to be performed. In the training phase, a set of speech files presented in the subjective listening test (Sec. III) is subject to the signal preprocessing (Sec. IV-A) and feature extraction (Sec. IV-B and IV-C) steps, resulting in one feature vector $\mathbf{x}$ per file. Together with the respective $\mathrm{MOS}_{\mathrm{LQS}}$ values given by the participants of the listening test, they form an input/target pair $\{\mathbf{x}(i), \mathrm{MOS}_{\mathrm{LQS}}(i)\}$, indexed by $i \in \mathcal{I}_{\mathrm{Tr}}$, where $|\mathcal{I}_{\mathrm{Tr}}|$ is the number of different speech files in the underlying subjective listening test that is assigned to a training set.

Mathematically speaking, the MOS predictor can be considered as function

$$g : \mathbf{x} \mapsto \mathrm{MOS}_{\mathrm{LQO}},$$

which is found under the condition that

$$\mu\left(\left(g(\mathbf{x}(i)) - \mathrm{MOS}_{\mathrm{LQS}}(i)\right)^2; \mathcal{I}_{\mathrm{Tr}}\right) \overset{!}{\rightarrow} \min. \tag{29}$$

Note that a neural network is a possible choice for the MOS predictor, but due to the rather small amount of available training data, a robust set of weights and biases is hard to find. Therefore the MOS predictor function $g(\cdot)$ is defined as the well-known ϵ-support vector regression (ϵ-SVR) [39], exhibiting excellent generalization capabilities also on smaller input/target data sets.

The training phase is shown in the left part of Fig. 3, aiming at finding a function $g(\cdot)$ that fulfills (29). The ϵ-SVR is trained in a normalized input domain w.r.t. zero mean and unit variance. This is achieved by

$$\mathbf{x}^{\mathrm{norm}}(i) = \frac{\mathbf{x}(i) - \mu(\mathbf{x}(i); \mathcal{I}_{\mathrm{Tr}})}{\sigma(\mathbf{x}(i); \mathcal{I}_{\mathrm{Tr}})}$$

for all $i \in \mathcal{I}_{\mathrm{Tr}}$. Subsequently, the normalized data is used to derive the function $g(\cdot)$ trained via the `LibSVM` framework [40]. The resulting ϵ-SVR model is characterized by a Gaussian radial basis function and parameter ϵ, which defines the allowed error on the training data set indexed by $\mathcal{I}_{\mathrm{Tr}}$ during calculation of function $g(\cdot)$. In addition, parameter C which represents the cost for misclassification during model training, and parameter γ which is the inverse width parameter of the Gaussian radial basis function, are selected in the training phase.

In the operative phase, feature vector $\mathbf{x}$ is calculated from the two input signals, i.e., a reference and degraded speech signal according to Fig. 1 upper and center part. After normalizing the feature vector using the mean and the standard deviation values as calculated during the training phase according to

$$\mathbf{x}^{\mathrm{norm}} = \frac{\mathbf{x} - \mu(\mathbf{x}(i); \mathcal{I}_{\mathrm{Tr}})}{\sigma(\mathbf{x}(i); \mathcal{I}_{\mathrm{Tr}})},$$

the normalized vector is subject to the ϵ-SVR, according to

$$\mathrm{MOS}_{\mathrm{LQO}} = g(\mathbf{x}^{\mathrm{norm}}).$$

V. Evaluation Metrics, Setup, and Results

First, we describe the evaluation setup for benchmarking the proposed instrumental measure in Sec. V-A. In our work we have to cope with a scarce data issue, which is why we chose cross-validation experiments. Afterwards, evaluation metrics for the proposed measure are presented in Sec. V-B. The presentation of the performance results of the new instrumental speech quality measure and their discussion will constitute Sec. V-C.

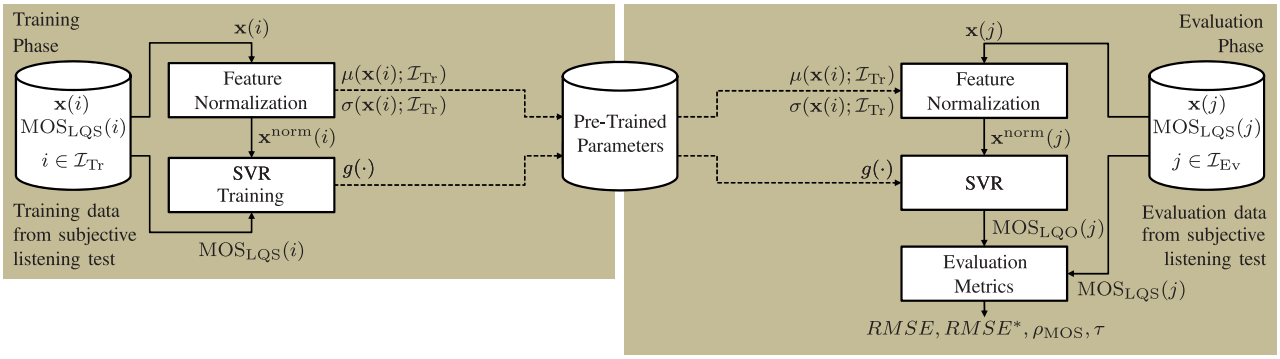

Fig. 3. **Evaluation setup** of the statistical model. *Left part*: Detailed block diagram of the **training phase of the MOS predictor**: Input/target pairs of $\mathbf{x}(i)$ and $\mathrm{MOS}_{\mathrm{LQS}}(i)$ indexed by training data set $\mathcal{I}_{\mathrm{Tr}}$ are each normalized and then subject to SVR training. Function $g(\cdot)$ and parameters for normalizing the input/target pairs are the result of the training phase. *Right part*: Detailed block diagram of the **evaluation phase of the MOS predictor**: Based on the preceding training phase using training data indexed by set $\mathcal{I}_{\mathrm{Tr}}$, $\mathrm{MOS}_{\mathrm{LQO}}$ values are calculated by means of the SVR for all feature vectors $\mathbf{x}(j)$ from the evaluation data set indexed by $j \in \mathcal{I}_{\mathrm{Ev}}$. Together with the ground truth $\mathrm{MOS}_{\mathrm{LQS}}(j)$ values, evaluation metrics are calculated.

TABLE III
DEFAULT CONDITIONS INCLUDED IN THE TRAINING AND EVALUATION DATA SET FOR THE CROSS-VALIDATION EXPERIMENTS A AND B

Sets	Condition(s)
$\mathcal{I}_{\mathrm{Tr}}$	WB-MNRU 35, 45, ∞ dB
	AMR-NB @ 12.2 kbps
	$\mathrm{ABE}_{01}, \ldots, \mathrm{ABE}_{12}$
	AMR-WB @ 8.85, 23.05, 23.85 kbps
$\mathcal{I}_{\mathrm{Ev}}$	AMR-NB @ 12.2 kbps
	$\mathrm{ABE}_{01}, \ldots, \mathrm{ABE}_{12}$
	AMR-WB @ 23.05 kbps

TABLE IV
EXPERIMENT A: DISJOINT LANGUAGES

Sets	A1	A2	A3	A4
$\mathcal{I}_{\mathrm{Tr}}$	*left out*	English	English	English
	Chinese	*left out*	Chinese	Chinese
	German	German	*left out*	German
	Korean	Korean	Korean	*left out*
$\mathcal{I}_{\mathrm{Ev}}$	English	Chinese	German	Korean

A. Evaluation Setup: Cross-Validation

This section describes the evaluation setup of the MOS predictor from subsection IV-D. Due to the relatively small amount of available data, three different k-fold cross-validation experiments are defined to rate the performance of the instrumental measure in three different disciplines: language independence, speaker independence, and the ability to generalize on ABE solutions, which were not part of the training data set. Fig. 3 visualizes the training and evaluation procedure for each of the experiments presented in this subsection. All available input/target pair indices $\mathcal{I}$ are divided into a set $\mathcal{I}_{\mathrm{Tr}}$ for SVR training and an evaluation set $\mathcal{I}_{\mathrm{Ev}}$ for measuring the predictive power of the MOS predictor. While evaluation metrics will be introduced in the next subsection, in this subsection the focus lies on defining a variety of data splits, constituting the conducted experiments.

In general, the training and evaluation scheme shown in Table III was used for the cross-validation experiments. To direct the instrumental measure towards the influence of acoustical bandwidth on perceptual quality, the speech data and subjective votes from the first 9 MNRU conditions were omitted from the SVR training. The AMR-NB condition is used as one of lower performance, while the remaining WB-MNRU conditions and AMR-WB conditions serve as upper performance cases. For evaluation of the measure, the AMR-NB condition, AMR-WB at 23.05 kbps, and all of the ABE conditions were used.

The underlying data for training and evaluation was analyzed in [3]. We found that statistical properties of the MOS ratings would change if the results from the different languages were merged to obtain a single global result. Therefore we conducted the result analysis for each language under test separately [3]. For the proposed measure, however we train the MOS predictor using multiple languages at once. This is due to the fact that we want to design a language-independent measure, which does not require a new training process before being applied to speech material in a new language.

With respect to the SVR, the parameter ϵ has direct influence on the number of support vectors and thus on complexity and generalization capabilities of the SVR model. For each subexperiment presented in the following, we select an $\epsilon \in [0.1, 1.0]$ in steps of 0.1 for which an SVR model results that has a maximum of 50% of input data pairs $|\mathcal{I}_{\mathrm{Ev}}|$ as support vectors. Furthermore, we bias the found ϵ by 0.2 to prevent overfitting to the rather small training data set. Parameter C is also found in an iterative process on the training data set via minimizing the RMSE error on the training data set. This parameter is found using equidistant sampling points on the logarithmic range $[1, 10000]$. Parameter $\gamma = \frac{1}{2\sigma^2}$ is set to $1/2$ using the *a priori* knowledge that the input data is of unit variance.

1) Experiment A - Disjoint Languages: In experiment A, the available data is split after the respective language of the utterances. This experiment aims at quantifying the ability of generalization w.r.t. unseen languages. Please note that this experimental design also implicitly means that there is no speaker overlap. Experiments follow the scheme presented in Table IV (Table III is still valid). A total of four subexperiments A1-A4 is conducted.

TABLE V
EXPERIMENT B: DISJOINT SPEAKERS

Sets	B1	B2	B3	B4
$\mathcal{I}_{\mathrm{Tr}}$	*left out*	F1	F1	F1
	F2	*left out*	F2	F2
	M1	M1	*left out*	M1
	M2	M2	M2	*left out*
$\mathcal{I}_{\mathrm{Ev}}$	F1	F2	M1	M2

TABLE VI
EXPERIMENT C: DISJOINT ABE SOLUTIONS

Sets	C1	C2
$\mathcal{I}_{\mathrm{Tr}}$	WB-MNRU $35, 45, \infty$ dB	
	AMR-NB @ 12.2 kbps	
	ABE_{01}	ABE_{02}
	ABE_{03}	ABE_{04}
	ABE_{05}	ABE_{06}
	...	...
	ABE_{11}	ABE_{12}
	AMR-WB @ $8.85, 23.05, 23.85$ kbps	
$\mathcal{I}_{\mathrm{Ev}}$	AMR-NB @ 12.2 kbps	
	ABE_{02}	ABE_{01}
	ABE_{04}	ABE_{03}
	ABE_{06}	ABE_{05}
	...	...
	ABE_{12}	ABE_{11}
	AMR-WB @ 23.05 kbps	

2) Experiment B - Disjoint Speakers: Experiment B is set up to evaluate speaker dependency in general. All combinations of three speakers in training and one speaker in evaluation define subexperiments B1-B4 as shown in Table V (Table III is still valid). A total of four subexperiments is conducted.

3) Experiment C - Disjoint ABE Solutions: To quantify the generalization ability of the instrumental measure on unseen ABE conditions, experiment C splits the available ABE conditions into two equally-sized sets, where always one is used for training and the other for testing (this is different to Table III). Using both ABE condition splits once in training and once in evaluation results in two subexperiments C1 and C2. This modified setup is shown in Table VI. When used in practice, the instrumental measure will rather have to cope with unseen speakers than languages. Therefore we will conduct subexperiments C1 and C2 each under the constraints defined in experiment B, leading to disjoint speakers as well. Consequently $2 \times 4 = 8$ subsubexperiments are conducted.

B. Evaluation Metrics

The metrics presented in the following evaluate an instrumental measure w.r.t. its predictive power, after the MOS predictor has been trained on training set $\mathcal{I}_{\mathrm{Tr}}$. For better distinction of training and evaluation phase, input/target pairs used for evaluation are indexed by variable $j \in \mathcal{I}_{\mathrm{Ev}}$. Based on the feature vectors $\mathbf{x}(j)$ the new measure calculates $\mathrm{MOS}_{\mathrm{LQO}}(j)$ values as shown in Fig. 3.

In the remaining part of this section, metrics taken from ITU-T P.1401 [41] for evaluating instrumental measures are presented. According to ITU-T P.1401, statistical evaluation of instrumental measures (of any kind) needs to be tested in three categories: Accuracy, consistency, and linearity against subjective data. For determining the accuracy of the predicted results we employ the RMSE of the absolute prediction error. Also being file-based, consistency is evaluated using the epsilon-insensitive RMSE, which qualitatively measures errors due to outliers. As check for linearity we employ Pearson's correlation coefficient as well as rank order, calculated over the predicted condition means.

1) Accuracy - RMSE: The file-based absolute prediction error is calculated as follows:

$$P_{\mathrm{error}}(j) = \mathrm{MOS}_{\mathrm{LQS}}(j) - \mathrm{MOS}_{\mathrm{LQO}}(j). \tag{30}$$

The respective RMSE over all files from the evaluation data set indexed by $j \in \mathcal{I}_{\mathrm{Ev}}$ is then given as [41]

$$RMSE = \sqrt{\frac{1}{|\mathcal{I}_{\mathrm{Ev}}| - 1} \sum\nolimits_{j \in \mathcal{I}_{\mathrm{Ev}}} \left(P_{\mathrm{error}}(j)\right)^2}. \tag{31}$$

A smaller RMSE value indicates better prediction performance and is employed for evaluation of accuracy of the proposed measure.

2) Consistency - Epsilon-Insensitive RMSE: To evaluate to which degree the predicted results are consistent, the epsilon-insensitive absolute prediction error P^*_{error} is calculated. Only those predicted $\mathrm{MOS}_{\mathrm{LQO}}$, which are scored *outside* of the 95% confidence interval are taken into account. The calculation follows

$$P^*_{\mathrm{error}}(j) = \max\left(0, |P_{\mathrm{error}}(j)| - CI^+_{95}(j)\right), \tag{32}$$

with $CI^+_{95}(j)$ being half of the confidence interval calculated over all votes from the test subjects on the respective condition of speech signal indexed by j. From this, the epsilon-insensitive $RMSE^*$ can be calculated [41]:

$$RMSE^* = \sqrt{\frac{1}{|\mathcal{I}_{\mathrm{Ev}}| - 1} \sum\nolimits_{j \in \mathcal{I}_{\mathrm{Ev}}} \left(P^*_{\mathrm{error}}(j)\right)^2}. \tag{33}$$

Consistency is indicated by a small $RMSE^*$ metric, which is caused by a small number of outliers or small-valued P^*_{error} of an instrumental measure. Implicitly, also a large $CI^+_{95}(j)$ for the respective file can cause a small $RMSE^*$ value, however, please keep in mind that a large $CI^+_{95}(j)$ means that the subjects' votes varied a lot.

3) Linearity - Correlation and Rank Order: As linearity check, we calculate Pearson's correlation coefficient ρ and Kendal's τ over condition-based MOS means. For the instrumentally-measured $\mathrm{MOS}_{\mathrm{LQO}}$ values, condition-based means are obtained as follows:

$$\mathrm{MOS}_{\mathrm{LQO}}(c) = \mu\left(\mathrm{MOS}_{\mathrm{LQO}}(j); \mathcal{I}^{(c)}_{\mathrm{Ev}}\right),$$

with $\mathcal{I}^{(c)}_{\mathrm{Ev}}$ indexing all files belonging to condition c (cf. Table I) in the evaluation data set. The condition-based means for the subjective votes $\mathrm{MOS}_{\mathrm{LQS}}(c)$ are derived accordingly.

The condition-based Pearson's correlation coefficient is then defined as

$$\rho_{\mathrm{MOS}} = \rho\left(\mathrm{MOS}_{\mathrm{LQS}}(c), \mathrm{MOS}_{\mathrm{LQO}}(c); \mathcal{C}_{\mathrm{Ev}}\right), \tag{34}$$

TABLE VII
PROPOSED QABE MEASURE BENCHMARK AND COMPARISON TO POLQA AND WB-PESQ EVALUATING THE THREE CATEGORIES ACCURACY, CONSISTENCY, AND LINEARITY IN THE THREE DISCIPLINES DISJOINT LANGUAGE, DISJOINT SPEAKERS, AND DISJOINT ABE SOLUTIONS

Exp. ID	SVR Parameters		**Accuracy** $RMSE$			**Consistency** $RMSE^*$			**Linearity** Correlation ρ_{MOS}			**Linearity** Rank Order τ		
	C	ϵ	Proposed	WB-PESQ	POLQA	Proposed	WB-PESQ	POLQA	Proposed	WB-PESQ	POLQA	Proposed	WB-PESQ	POLQA
Experiment A: Disjoint Languages (Sec. V-A1)														
A1	1000	0.5	**0.388**	0.608	0.523	**0.135**	0.325	0.257	**0.959**	0.888	0.688	**0.758**	0.670	0.670
A2	1000	0.5	0.617	0.831	**0.567**	0.315	0.497	**0.273**	0.875	**0.882**	0.619	**0.758**	0.692	0.648
A3	1000	0.5	0.589	**0.423**	0.740	0.315	**0.188**	0.465	**0.971**	0.849	0.723	**0.780**	0.516	0.670
A4	1000	0.5	**0.396**	0.511	0.654	**0.129**	0.271	0.353	**0.916**	0.772	0.614	**0.789**	0.478	0.633
Mean			**0.497**	0.593	0.621	**0.223**	0.320	0.337	**0.930**	0.848	0.661	**0.771**	0.589	0.656
Experiment B: Disjoint Speakers (Sec. V-A2)														
B1	1000	0.5	**0.443**	0.658	0.528	**0.134**	0.323	0.311	**0.979**	0.889	0.725	**0.868**	0.560	0.626
B2	1000	0.5	**0.482**	0.615	0.607	**0.182**	0.294	0.274	**0.926**	0.873	0.675	**0.604**	0.560	0.560
B3	1000	0.5	**0.536**	0.564	0.682	**0.187**	0.225	0.301	**0.955**	0.862	0.645	**0.758**	0.714	0.495
B4	1000	0.5	**0.443**	0.609	0.678	**0.146**	0.304	0.335	**0.959**	0.786	0.626	**0.846**	0.560	0.648
Mean			**0.476**	0.612	0.624	**0.162**	0.286	0.280	**0.955**	0.852	0.668	**0.775**	0.599	0.582
Experiment C: Disjoint ABE Solutions (Sec. V-A3)														
C1 ∩ B1	1000	0.5	**0.472**	0.671	0.723	**0.165**	0.336	0.358	**0.919**	0.774	0.638	**0.857**	0.643	0.786
C2 ∩ B1	1000	0.5	**0.431**	0.489	0.593	**0.121**	0.172	0.223	**0.969**	0.923	0.868	**0.714**	0.500	0.643
C1 ∩ B2	1000	0.5	**0.567**	0.609	0.650	**0.187**	0.241	0.249	**0.891**	0.882	0.750	**0.786**	0.643	0.714
C2 ∩ B2	1000	0.5	0.522	**0.462**	0.676	0.175	**0.133**	0.277	0.924	**0.950**	0.783	0.786	**0.857**	0.429
C1 ∩ B3	1000	0.5	**0.588**	0.632	0.609	0.270	0.292	**0.260**	0.895	**0.900**	0.745	**0.714**	0.643	**0.714**
C2 ∩ B3	1000	0.5	**0.421**	0.610	0.560	**0.112**	0.273	0.215	**0.936**	0.882	0.707	**0.571**	0.500	0.286
C1 ∩ B4	1000	0.5	**0.484**	0.688	0.524	**0.164**	0.344	0.197	**0.916**	0.913	0.813	**0.857**	0.643	0.714
C2 ∩ B4	1000	0.5	**0.467**	0.638	0.502	**0.122**	0.268	0.169	**0.962**	0.908	0.759	**0.857**	0.643	0.500
Mean			**0.494**	0.600	0.605	**0.164**	0.257	0.243	**0.927**	0.892	0.758	**0.768**	0.634	0.598

with $\mathcal{C}_{\text{Ev}}$ being the set of condition indices occurring in the evaluation set. Besides the correlation coefficient, Kendal's τ is employed quantifying correctness of the conditions' rank order. It is calculated as follows

$$\tau = \frac{N_c - N_d}{\frac{1}{2}|\mathcal{C}_{\text{Ev}}| \cdot (|\mathcal{C}_{\text{Ev}}| - 1)}, \tag{35}$$

with N_c and N_d being the number of concordant and discordant pairs, respectively, and $|\mathcal{C}_{\text{Ev}}|$ being the amount of conditions contained in the test data set. For calculating this metric, all consecutive score pairs $\{\text{MOS}_{\text{LQS}}(c), \text{MOS}_{\text{LQO}}(c)\}$ and $\{\text{MOS}_{\text{LQS}}(c{+}1),\ \text{MOS}_{\text{LQO}}(c{+}1)\}$ are evaluated. They count as concordant if

$$\text{MOS}_{\text{LQS}}(c) < \text{MOS}_{\text{LQS}}(c{+}1) \text{ and}$$
$$\text{MOS}_{\text{LQO}}(c) < \text{MOS}_{\text{LQO}}(c{+}1)$$
$$\text{or}$$
$$\text{MOS}_{\text{LQS}}(c) > \text{MOS}_{\text{LQS}}(c{+}1) \text{ and}$$
$$\text{MOS}_{\text{LQO}}(c) > \text{MOS}_{\text{LQO}}(c{+}1).$$

Else, they are counted to the number of discordant pairs. Pearson's correlation coefficient and Kendal's τ are used to investigate the performance of the proposed measure w.r.t. the P.1401-recommended category of linearity against the underlying subjective data.

C. Results and Discussion

The results of the SVR parameter tuning and of experiments A,B,C can be found in Table VII.

Parameter tuning according to Sec. V-A on the respective training data set led to $\epsilon = 0.3$ (without bias) and $C = 1000$ in each of the subexperiments. This is fortunate since it allows further analysis based on a single parameter set. The proposed instrumental quality measure using these found parameters is in the following referred to as QABE.

In experiment A (disjoint languages), all ABE conditions are both in the training and evaluation data set. Just considering the results of QABE, it can be confirmed that the elements of the feature vector contain all the information needed to learn and subsequently predict MOS_{LQO} for ABE-processed signals. In subexperiment A2, the instrumental measure was evaluated on the Chinese part of the available data, which is a rather challenging task, since the result analysis of the underlying subjective listening test showed a gap between NB and WB of only 0.31 MOS points for Chinese, while in the other languages higher gaps up to 1.39 MOS points were identified. Even in this rather challenging subexperiment, QABE's correlation of 0.875 is similar to WB-PESQ and much better than POLQA. Interestingly, WB-PESQ has to cope with its highest $RMSE$ of 0.831 and thus, considering the other metrics, shows an inconsistent picture concerning subexperiment A2. Looking at all subexperiments A1-A4 the $RMSE$ values for WB-PESQ vary quite a lot. Concluding on all disjoint language experiments, QABE outperforms POLQA and WB-PESQ in terms of rank order in each of the subexperiments. While on average POLQA shows a somewhat better rank order than WB-PESQ, POLQA is worse in all other categories, and particularly poor in correlation. Furthermore, results show a significant advantage of QABE vs. POLQA and WB-PESQ *on average in all* three quality categories.

Investigating the influence of speaker dependency, experiment B splits all available speakers in disjoint sets for training and evaluation of QABE. Concerning accuracy, consistency, and rank order, POLQA and WB-PESQ perform on average equally bad. Again, POLQA shows very poor correlation performance. Whatever speaker was left out of the training set and then used for evaluation, the instrumental measure shows superior generalization capabilities *in all* of the tested categories *for each subexperiment*, with an average correlation even above 0.95.

Experiment C simulates the probably most common use case of the instrumental measure: Being trained on several ABE solutions in a number of languages and then being employed to assess the speech quality of several ABE solutions (unseen in training). Although the training data set is reduced by taking only every other ABE solution, the absolute performance of QABE drops only slightly. However, considering the results of QABE, consistently high correlation values of on average 0.927 are a proof for good generalization capabilities. While WB-PESQ performs best in subexperiment C2 $\cap$ B2, results for QABE are comparable to WB-PESQ in experiment C1 $\cap$ B3 w.r.t. $RMSE^*$ and correlation. The rank order of 0.857 scored by WB-PESQ in this subexperiment deviates quite much from the average of 0.643. *On average*, QABE outperforms WB-PESQ and POLQA *in all categories* of this important experiment, again being the only measure providing an average correlation of more than 0.9.

Comparing the results of experiments A-C, a high and robust performance of QABE is proven by the small deviations of the averaged results of the three categories. In general, there is only one subexperiment where the instrumental measure did not score the highest rank order. Furthermore the results suggest that the instrumental measure takes profit from the additional information contained in the variety of features used as input for the MOS predictor. In combination with the powerful ϵ-SVR, the proposed measure was able to outperform both WB-PESQ and POLQA in almost all subexperiments, but most importantly, on average in all experiments. We should remind ourselves that WB-PESQ and POLQA have not been developed with focus on ABE solutions.

Looking back at our proposed features, it becomes clear that indeed they focus on ABE-typical distortions in the upper band. Nevertheless, all features depend on the complete wideband signal which is important for detection of ABE artifacts, but shall not mean that the here proposed ABE quality measure adequately also judges lower band artifacts. This is due to the fact that none of the ABE solutions used in training showed any severe lower band artifacts: For these cases our proposed measure cannot and shall not compete with, e.g., POLQA as applied on the downsampled lower band. In practice, therefore, we propose a 2-step ABE speech quality assessment approach: In a first step a pure lower band speech quality assessment using, e.g., POLQA in NB mode is performed which should yield a quality not worse than typical AMR conditions. Only if this test is passed, in a second step our new measure would be applied to further judge the ABE-processed speech.

Inquiries about availability of the QABE software can be directed to qabe@ifn.ing.tu-bs.de.

VI. Conclusion

In this work, the development and evaluation of an instrumental measure for artificially bandwidth-extended (ABE) speech signals was presented. The proposed measure outperforms both POLQA and wideband PESQ in all conducted k-fold cross-validation experiments, showing a smaller RMSE, a higher correlation of well above 0.90, and a higher rank order correlation. These experiments are also proof for good generalization capabilities of the proposed instrumental measure w.r.t. unseen languages, speakers and most-importantly unseen artificial bandwidth extension solutions. We should note that neither POLQA nor wideband PESQ were intended for ABE test conditions.

Acknowledgment

The authors would like to thank the anonymous reviewers for their helpful and very suitable suggestions to improve the quality of this paper.

References

[1] E. Terhardt, *Akustische Kommunikation*. New York, NY, USA: Springer, 1998.
[2] ITU, "ITU-T Recommendation P.800, Methods for Subjective Determination of Transmission Quality," International Telecommunication Union, Geneva, Switzerland, Aug. 1996.
[3] J. Abel *et al.*, "A subjective listenining test of six different artificial bandwidth extension approaches in English, Chinese, German, and Korean," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Shanghai, China, Mar. 2016, pp. 5915–5919.
[4] H. Carl and U. Heute, "Bandwidth enhancement of narrow-band speech signals," in *Proc. Eur. Signal Process. Conf.*, Edinburgh, U.K., Sep. 1994, pp. 1178–1181.
[5] A. H. Nour-Eldin, "Quantifying and exploiting speech memory for the improvement of narrowband speech bandwidth extension," Ph.D. dissertation, Dept. Elect. Comput. Eng., McGill University, Montreal, QC, Canada, 2013.
[6] A. H. Nour-Eldin and P. Kabal, "Memory-based approximation of the Gaussian mixture model framework for bandwidth extension of narrowband speech," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (Interspeech)*, Florence, Italy, Aug. 2011, pp. 1185–1188.
[7] P. Bauer and T. Fingscheidt, "A statistical framework for artificial bandwidth extension exploiting speech waveform and phonetic transcription," in *Proc. Eur. Signal Process. Conf.*, Glasgow, Scotland, Aug. 2009, pp. 1839–1843.
[8] C. Yagli, M. A. T. Turan, and E. Erzin, "Artificial bandwidth extension of spectral envelope along a Viterbi path," *Speech Commun.*, vol. 55, pp. 111–118, Jan. 2013.
[9] P. Jax and P. Vary, "On artificial bandwidth extension of telephone speech," *Signal Process.*, vol. 83, no. 8, pp. 1707–1719, Aug. 2003.
[10] T. Schlien, F. Heese, M. Schäfer, C. Antweiler, and P. Vary, "Audiosignalverarbeitung für Videokonferenzsysteme," in *Proc. Workshop Audiosignal- und Sprachverarbeitung; INFORMATIK 2013*, Koblenz, Germany, Sep. 2013, pp. 2987–3001.
[11] H. Pulakka and P. Alku, "Bandwidth extension of telephone speech using a neural network and a filter bank implementation for highband mel spectrum," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2170–2183, Sep. 2011.
[12] R. Peharz, G. Kapeller, P. Mowlaee, and F. Pernkopf, "Modeling speech with sum-product networks: Application to bandwidth extension," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Florence, Italy, May 2014, pp. 3699–3703.
[13] M. Zöhrer, R. Peharz, and F. Pernkopf, "On representation learning for artificial bandwidth extension," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (Interspeech)*, Dresden, Germany, Sep. 2015, pp. 791–795.
[14] K. Li and C.-H. Lee, "A deep neural network approach to speech bandwidth expansion," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Brisbane, Australia, Apr. 2015, pp. 4395–4399.

[15] Y. Wang, S. Zhao, W. Liu, M. Li, and J. Kuang, "Speech bandwidth expansion based on deep neural networks," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (Interspeech)*, Dresden, Germany, Sep. 2015, pp. 2593–2597.
[16] J. Abel, M. Strake, and T. Fingscheidt, "Artificial bandwidth extension using deep neural networks for spectral envelope estimation," in *Proc. Int. Workshop Acoust. Signal Enhancement*, Xi'an, China, Sep. 2016, pp. 1–5.
[17] P. Bauer, J. Jones, and T. Fingscheidt, "Impact of hearing impairment on fricative intelligibility for artificially bandwidth-extended telephone speech in noise," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Vancouver, BC, Canada, May 2013, pp. 7039–7043.
[18] T. Fingscheidt and P. Bauer, "A phonetic reference paradigm for instrumental speech quality assessment of artificial speech bandwidth extension," in *Proc. 4th Int. Workshop Perceptual Quality Syst.*, Vienna, Austria, Sep. 2013, pp. 36–39.
[19] P. Bauer, C. Guillaumé, W. Tirry, and T. Fingscheidt, "On speech quality assessment of artificial bandwidth extension," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Florence, Italy, May 2014, pp. 6082–6086.
[20] H. Pulakka, V. Myllylä, A. Rämö, and P. Alku, "Speech quality evaluation of artificial bandwidth extension: Comparing subjective judgments and instrumental predictions," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (Interspeech)*, Dresden, Germany, Sep. 2015, pp. 2583–2587.
[21] P. Bauer, J. Abel, and T. Fingscheidt, "HMM-based artificial bandwidth extension supported by neural networks," in *Proc. Int. Workshop Acoust. Signal Enhancement*, Juan les Pins, France, Sep. 2014, pp. 1–5.
[22] ITU, "ITU-T Recommendation P.862, Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs," International Telecommunication Union, Geneva, Switzerland, Feb. 2001.
[23] ITU, "ITU-T Recommendation P.862.2, Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs," International Telecommunication Union, Geneva, Switzerland, Nov. 2007.
[24] ITU, "ITU-T Recommendation P.863, Perceptual Objective Listening Quality Assessment," International Telecommunication Union, Geneva, Switzerland, Jan. 2011.
[25] S. Möller *et al.*, "Speech quality prediction for artificial bandwidth extension algorithms," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (Interspeech)*, Lyon, France, Aug. 2013, pp. 3439–3443.
[26] I. Katsir, D. Malah, and I. Cohen, "Evaluation of a speech bandwidth extension algorithm based on vocal tract shape estimation," in *Proc. Int. Workshop Acoust. Signal Enhancement*, Aachen, Germany, Sep. 2012, pp. 1–4.
[27] M. A. T. Turan and E. Erzin, "Synchronous overlap and add of spectra for enhancement of excitation in artificial bandwidth extension of speech," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (Interspeech)*, Dresden, Germany, Sep. 2015, pp. 2588–2592.
[28] "Speech Recognition Database," Speech Ocean., 2015. [Online]. Available: www.speechocean.com
[29] ITU, "ITU-T Recommendation P.56, Objective Measurement of Active Speech Level," International Telecommunication Union, Geneva, Switzerland, Dec. 2011.
[30] ITU, "ITU-T Recommendation G.191, Software Tool Library 2009 User's Manual," International Telecommunication Union, Geneva, Switzerland, Nov. 2009.
[31] M. Vondrasek and P. Pollak, "Methods for speech SNR estimation: Evaluation tool and analysis of VAD dependency," *Radioengineering*, vol. 14, no. 1, pp. 6–11, 2005.
[32] T. Fingscheidt, S. Suhadi, and S. Stan, "Environment-optimized speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 4, pp. 825–834, May 2008.
[33] I. Katsir, I. Cohen, and D. Malah, "Speech bandwidth extension based on speech phonetic content and speaker vocal tract shape estimation," in *Proc. Eur. Signal Process. Conf.*, Barcelona, Spain, Aug. 2011, pp. 461–465.
[34] E. Zwicker and H. Fastl, *Psychoacoustics - Facts and Models*, 2nd ed. New York, NY, USA: Springer, Jan. 1999.
[35] R. Sottek, "Modelle zur Signalverarbeitung im menschlichen Gehör," Ph.D. dissertation, Institut für Elektrische Nachrichtentechnik, RWTH Aachen, 1993.
[36] R. Sottek and K. Genuit, "Models of signal processing in human hearing," *Int. J. Electron. Commun.*, vol. 59, no. 3, pp. 157–165, Jun. 2005.
[37] P. Loizou and G. Kim, "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 1, pp. 47–56, Jan. 2011.
[38] J. F. Santos, S. Cosentino, O. Hazrati, P. C. Loizou, and T. H. Falk, "Objective speech intelligibility measurement for cochlear implant users in complex listening environments," *Speech Commun.*, vol. 55, no. 78, pp. 815–824, 2013.
[39] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag, 2009.
[40] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, 2011, Art. no. 27. [Online]. Available: http://www.csie.ntu.edu.tw/ cjlin/libsvm.
[41] ITU, "ITU-T Recommendation P.1401, Methods, Metrics and Procedures for Statistical Evaluation, Qualification and Comparison of Objective Quality Prediction Models," International Telecommunication Union, Geneva, Switzerland, Jul. 2012.

Johannes Abel received the M.Sc. degree in computer and communications systems engineering from Technische Universität Braunschweig, Braunschweig, Germany. During his studies, he worked as a Student Assistant in the field of speech enhancement and wrote his master's thesis at the Institute for Communications Technology on artificial bandwidth extension for automatic speech recognition. Since 2013, he has been working toward the Ph.D. degree in the field of artificial bandwidth extension for telephony applications. His research interests include speech enhancement, machine learning, and automatic speech recognition.

Magdalena Kaniewska received the M.Sc. degree and the Ph.D. degree in telecommunications from Gdansk University of Technology, Poland, in 2006 and 2012, respectively. She then continued her work in the field by joining Orange Labs in Lannion, France, for an 18-month post-doc. During that period, she was part of a team that co-developed the codec for Enhanced Voice Services, standardized by 3GPP in September 2014. In April 2014, she started working as a Researcher and Algorithm Designer with NXP Semiconductors, Leuven, Belgium, where she helps to develop speech enhancement algorithms for mobile communications.

Cyril Guillaumé received the Dipl.-Ing. degree in electronics and computer science from Ecole Superieure d'Electronique de l'Ouest, Angers, France, in 2004, with a major in signal processing and telecommunications. After his studies he worked on speech coding at Orange. He joined NXP in 2007, where he is working on speech enhancement for mobile devices.

Wouter Tirry received the M.Sc. degree in physics and the Ph.D. degree in solar physics from the University of Leuven, Belgium, in 1994 and 1998, respectively. As a Post-Doc, he further pursued his research at the National Centre for Atmospheric Research, Boulder, CO, USA. Since 1999, he has been building up expertise in the domain of speech enhancement for mobile devices at Philips and NXP as Research Engineer and System Architect. He is currently the Senior Principal at the Product Line Mobile Audio Solutions, NXP, leading the speech technology development activities.

Tim Fingscheidt (S'93–M'98–SM'04) received the Dipl.-Ing. degree in electrical engineering in 1993 and the Ph.D. degree in 1998 from RWTH Aachen University, Germany. He further pursued his work on joint speech and channel coding as a consultant in the Speech Processing Software and Technology Research Department at AT&T Labs, Florham Park, NJ, USA. In 1999, he entered the Signal Processing Department of Siemens AG (COM Mobile Devices) in Munich, Germany, and contributed to speech codec standardization in ETSI, 3GPP, and ITU-T. In 2005, he joined Siemens Corporate Technology in Munich, Germany, leading the speech technology development activities in recognition, synthesis, and speaker verification. Since 2006, he has been a Full Professor in the Institute for Communications Technology, Technische Universität Braunschweig, Germany. His research interests include speech and audio signal processing, enhancement, transmission, recognition, and instrumental quality measures. Dr. Fingscheidt received several awards, among them the prize of the Vodafone Mobile Communications Foundation in 1999, and the 2002 prize of the Information Technology branch of the Association of German Electrical Engineers (VDE ITG), where he is leading the Speech Acoustics Committee ITG FA4.3 since 2015. From 2008 to 2010, he served as Associate Editor for IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, and since 2011, as Member of the IEEE Speech and Language Processing Technical Committee.