Exploring Auditory-Inspired Acoustic Features for Room Acoustic Parameter Estimation from Monaural Speech

Feifei Xiong, *Student Member, IEEE,* Stefan Goetze, *Member, IEEE,* Birger Kollmeier, and Bernd T. Meyer

***Abstract*—Room acoustic parameters that characterize acoustic environments can help to improve signal enhancement algorithms such as for dereverberation, or automatic speech recognition by adapting models to the current parameter set. The reverberation time (RT) and the early-to-late reverberation ratio (ELR) are two key parameters. In this paper, we propose a blind ROom Parameter Estimator (ROPE) based on an artificial neural network that learns the mapping to discrete ranges of the RT and the ELR from single-microphone speech signals. Auditory-inspired acoustic features are used as neural network input, which are generated by a temporal modulation filter bank applied to the speech time-frequency representation. ROPE performance is analyzed in various reverberant environments in both clean and noisy conditions for both fullband and subband RT and ELR estimations. The importance of specific temporal modulation frequencies is analyzed by evaluating the contribution of individual filters to the ROPE performance. Experimental results show that ROPE is robust against different variations caused by room impulse responses (measured vs. simulated), mismatched noise levels and speech variability reflected through different corpora. Compared to state-of-the-art algorithms that were tested in the Acoustic Characterisation of Environments (ACE) challenge, the ROPE model is the only one that is among the best for all individual tasks (RT and ELR estimation from fullband and subband signals). Improved fullband estimations are even obtained by ROPE when integrating speech-related frequency subbands. Further, the model requires the least computational resources with a real time factor that is at least two times faster than competing algorithms. Results are achieved with an average observation window of 3 seconds, which is important for real-time applications.**

***Index Terms*—Reverberation time, early-to-late reverberation ratio, blind estimation, auditory-inspired acoustic features, machine learning.**

Manuscript received November 29, 2017; revised April 10, 2018; accepted 14 May, 2018. Date of publication; date of current version. This work was supported by the DFG Cluster of Excellence 1077/1 "Hearing4all". The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Federico Fontana. (Corresponding author: Feifei Xiong.)

F. Xiong and B. T. Meyer are with the Medical Physics Department and Cluster of Excellence "Hearing4all", University of Oldenburg, Oldenburg, Germany (e-mail: {feifei.xiong, bernd.meyer}@uni-oldenburg.de).

S. Goetze is with the Fraunhofer Institute for Digital Media Technology IDMT and Cluster of Excellence "Hearing4all", Oldenburg, Germany (e-mail: s.goetze@idmt.fraunhofer.de).

B. Kollmeier is with the Medical Physics Department and Cluster of Excellence "Hearing4all", University of Oldenburg, Oldenburg, and with the Fraunhofer Institute for Digital Media Technology IDMT, Oldenburg, Germany (e-mail: birger.kollmeier@uni-oldenburg.de).

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.2018.2843537.

I. Introduction

THE acoustic characteristics of a room have been shown to be important to predict the speech quality and intelligibility of speech signals, which are highly relevant in speech communication applications such as dereverberation in hands-free telecommunication devices, speech enhancement in hearing aids, or front-end processing in automatic speech recognition (ASR). The reverberation time (RT) and the early-to-late reverberation ratio (ELR) are two key parameters to represent such room acoustic characteristics [1], [2]. The RT is defined as the time interval for a 60 dB decay of the sound energy after the sound source is ceased, and the ELR refers to the energy ratio between the early reflections of signal transmission (including the direct path) and the late reverberation caused by multi-path propagation from the sound source to the receiver. Special cases of the ELR include the direct-to-reverberant ratio (DRR) (i.e., early components correspond to sounds from the direct path), and the clarity index [1] for which early and late reflections are separated at 50 ms (denoted by C_{50}) or at 80 ms (C_{80}). Naturally, the RT and the ELR are frequency-dependent parameters since the absorption reflection coefficients vary with frequency [1]. In addition to values obtained from fullband processing, the RT and the ELR are therefore often specified for subbands (e.g., for the octave band centered at 1 kHz [2]). Traditionally, the RT and the ELR are derived from the room impulse response (RIR) between the source and the receiver, which can be measured, e.g., using an excitation signal such as a swept-sine signal [3]. However, RIR recordings require time and other resources, and are not always practical in real-world scenarios. Consequently, it is of great interest to blindly (or non-intrusively) estimate the RT and the ELR directly from reverberant speech signals.

To estimate the RT, statistical models of the sound decay characteristics of reverberant speech have been explored in earlier research: Ratnam et al. [4] modeled the reverberation tail of the RIR using an exponentially damped Gaussian envelope, so that the RT can be obtained from the envelope that is fitted to the data using a maximum-likelihood (ML) criterion. This has been extended in [5] aiming at reducing the complexity using a pre-selection mechanism to detect the plausible exponential decays. Similarly, Vieira [6] focused on detecting the free decay regions of reverberant signals, where the exponential decay model and Schroeder's integral [7] were used to determine the RT. This method has been also applied

to subband processing mode in [8]. Wen et al. [9] developed an RT estimator using spectral decay distributions in which the negative-side variance was shown to strongly correlate with RT values, so that a linear mapping function can be generated based on samples with known RTs. A noise-robust version of [9] with reduced computational complexity was presented in [10]. A comparison of energies at high and low modulation frequencies, the so-called speech-to-reverberation modulation energy ratio (SRMR), was proposed in [11] to obtain linear estimates of room parameters.

Approaches based on machine learning were shown to be quite successful for estimating room parameters as well. For instance, artificial neural networks (ANNs) have been proposed to learn a mapping between the ground truth and target measure: Cox et al. [12] trained an ANN to blindly estimate the RT using the short-term root-mean square values of the speech signals as the ANN input, whereas [13] used the low-frequency envelope spectrum as ANN input feature.

In contrast to RT estimators, which usually utilize single-microphone audio, the majority of blind ELR estimators relies on multi-microphone data. This allows to exploit spatial information obtained by multi-microphone recordings for separating the early or the direct component from the reverberant speech, e.g., DRR estimation in [14]–[17]. However, multi-microphone configurations are not available in many scenarios, which motivated research on (usually supervised) single-channel ELR estimators. For instance, the ANN-based approach proposed in [12] for RT estimation has been modified in [13] to estimate C_{80} from single-microphone data. A classification combined with regression trees was used in [18] to estimate C_{50} with a complex combined set of acoustic features.

In this work, we explore a blind acoustic ROom Parameter Estimator (ROPE) which estimates the RT and the ELR from single-microphone speech recordings in both, fullband and subband processing. To this end, a multi-layer perceptron (MLP) is used as discriminative classifier to obtain different room acoustic parameters, motivated by the findings from subjectively perceptual experiments that human auditory system is able to distinguish various RTs and ELRs but with constrained just noticeable differences (JNDs) [2]. Input features to the MLP need to reflect changes for different RTs and ELRs in noisy environments, and at the same time, to generalize with respect to speech variability since speech encountered during test time is unknown to the classifier.

Motivated by the robustness of the human auditory system, previous research has shown auditory-inspired signal processing to be effective for speech separation [19] and ASR [20]. We explore auditory-inspired acoustic features for the ROPE system, which consist of a time-frequency (TF) representation of the speech signal and a temporal modulation filter bank: Gammatone filter bank based TF representation is used since it was shown to improve speech processing particularly in computational auditory scene analysis (CASA) [19], [21]. This approach allows a straightforward extension for our model from fullband analysis [22], [23] to subband processing due to the time-domain implementation of the Gammatone filters. Further, temporal processing in the human auditory system is crucial for speech perception in reverberant conditions (which result in temporal smearing of the speech signal) by analyzing acoustic temporal modulation cues [24]. Hence, we use a filter bank to extract features that capture different temporal modulation frequencies (TMFs) from TF representations, and analyze the relevance of TMFs as well as their robustness when estimating RTs and ELRs in noise-free and noisy conditions. Since ROPE requires labeled training data, we also explore its performance in mismatched training and testing conditions, which should provide insight into the robustness in the presence of different RIRs, noise levels, and speech signals. Our approach is validated using the ACE challenge single-microphone evaluation database recorded in real rooms [25] and compared to other state-of-the-art RT and ELR estimators in terms of the estimation accuracy and the computational complexity for real-time applications. Finally, the performance for each frequency subband is analyzed, as well as the benefit from combining subbands that cover frequencies important in speech perception.

The remainder of this paper is organized as follows: We first briefly introduce the two key room acoustic parameters in Section II, and then describe the calculation of the auditory-inspired acoustic features in Section III. Section IV illustrates the ROPE system structure, and the experimental setup for evaluation is presented in Section V. Results and discussion are presented in Section VI, which is structured by the experimental parameters that were systematically varied. Section VII concludes the paper.

II. Room Acoustic Parameters

Both the MLP training for ROPE as well as a subsequent model evaluation require the ground truth of the room parameters. The procedure for obtaining the ground truth from the RIR is described in the following, as well as the definitions for RT and ELR.

A. Reverberation Time (RT)

One classic intrusive measure of the RT is based on the energy decay curve computed by Schroeder's integral [7] from a measured RIR $h[k]$ with time index k. A linear fitting is then applied to a certain range of this curve to determine a decay by 60 dB [2]. Karjalainen et al. [26] found that nonlinear fitting algorithms produce more reliable results, especially against non-stationary noise floor in measured RIRs. In accordance with the ACE challenge [25], this nonlinear fitting is applied to the logarithmic magnitude of $h[k]$ to obtain the RT ground truth. For subband analysis, the RIR $h_f[k]$ in frequency band f is obtained from the fullband signal $h[k]$ (cf. Section III-A) and subsequently processed as the fullband counterpart.

B. Early-to-Late Reverberation Ratio (ELR)

For continuous signals, the ELR of an RIR $h(t)$ is defined in a decibel scale as in (1) with t_e denoting the boundary between early and late signal components. For a discrete RIR $h[k]$ of length L_h, the ELR is calculated according to (2) with

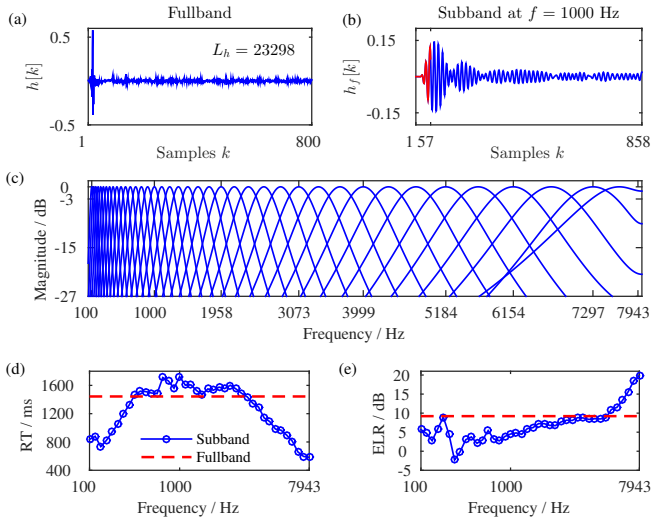

Fig. 1. (a) the early part ($t_e = 50$ ms at $f_s = 16$ kHz) of a fullband RIR $h[k]$ from the ACE challenge [25]; (b) the corresponding subband RIR $h_f[k]$ at $f = 1000$ Hz. The initial (red) part is removed due to the Gammatone filter delay; (c) frequency response of the Gammatone filter bank; (d) RT ground truth; (e) ELR ground truth.

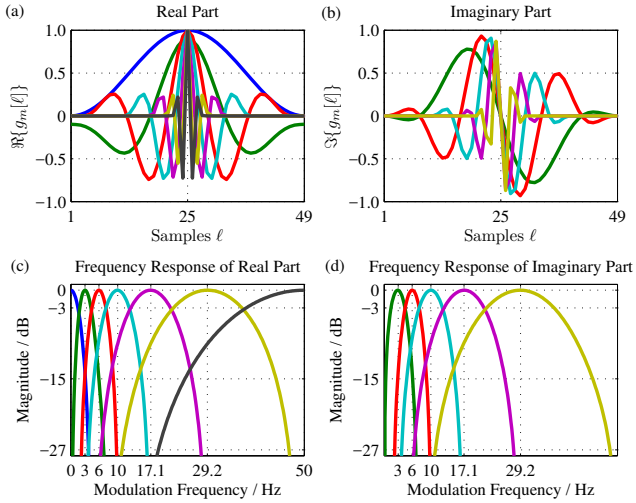

Fig. 2. Temporal modulation filter bank with the (a) real part and (b) imaginary part (short filters have been zero padded to the longest length for easy comparison), as well as (c) frequency responses of 7 real filters and (d) frequency responses of 5 imaginary filters.

k_e denoting the boundary sample obtained by rounding $f_s \cdot t_e$ to the nearest integer.

$$\mathrm{ELR} = 10\log_{10}\left(\frac{\int_{t=0}^{t=t_e} h^2(t)}{\int_{t=t_e}^{\infty} h^2(t)}\right), \tag{1}$$

$$\simeq 10\log_{10}\left(\frac{\sum_{k=1}^{k=k_e} h^2[k]}{\sum_{k=k_e+1}^{L_h} h^2[k]}\right). \tag{2}$$

To determine the ELR for our experiments, silence preceding the RIR is removed and the tail is cropped after a level reduction to -70 dB to limit artifacts. In the following, t_e is set to 50 ms (i.e., the ELR with this time constant is identical to C_{50}), which is motivated by the grouping of multi-path signal components of the human auditory system if the delay between paths does not exceed approximately 50 ms [27]. Note that the term ELR is a general description of the energy ratio room parameters, and the proposed ROPE model could be potentially applied to ELR measures with other time constants than 50 ms, which should be explored in future research. For the calculation of the subband ELR in (2) from the subband RIR $h_f[k]$, signal components introduced by the filter delay are removed (cf. the initial red part as illustrated in Fig. 1 (b)).

III. Auditory-Inspired Acoustic Features

For estimating acoustic room parameters, we exploit two different auditory-inspired features which are used as input to the neural network for parameter classification, as described in the following.

A. Time-Frequency Representation

Gammatone filter bank based TF representations [19], which are inspired by the human auditory system, are employed as the first feature type. For calculation, a Gammatone filter bank decomposes the speech signal into frequency bands, which models the auditory filters and exhibits a higher resolution for lower frequencies compared to the short-time Fourier transform. As shown in Fig 1 (c), we use 40 Gammatone filters with center frequencies ranging from 100 to 7943 Hz with a sampling frequency of 16 kHz. Short-time windowing is applied with a frame shift of 10 ms and an analysis block length of 25 ms, resulting in a two-dimensional time-frequency representation (cf. Fig. 3). A logarithmic compression function is then used for dynamic range reduction.

Besides these baseline TF representations (40-dimensional log-Gammatone filter bank features) without temporal context, we also investigate features with a context of several preceding and subsequent frames as MLP input. This is inspired by ASR experiments (e.g., in [28], [29]), where it was observed that neural networks profit from information about temporal dynamics through concatenated time frames, which is used here as well (with 5 preceding and 5 following frames labeled as ± 5, which results in 440-dimensional features).

B. Temporal Modulation Filter Bank

Motivated by the successful use of modulation filtering for feature extraction in ASR systems [29]–[31], we explore temporal modulation frequency (TMF) features that are extracted from the TF representations. The temporal modulation filter bank $g_m[\ell]$ with filter index m is given by a complex exponential carrier function $s_{\mathrm{carr}}[\ell, f_m]$ that is modulated by a zero-phase Hann-envelope function $w_{\mathrm{env}}[\ell, L_m]$,

$$g_m[\ell] = s_{\mathrm{carr}}[\ell, f_m] \cdot w_{\mathrm{env}}[\ell, L_m], \tag{3}$$

$$s_{\mathrm{carr}}[\ell, f_m] = \exp\left(-i2\pi f_m \cdot T \cdot (\ell - \ell_0)\right), \tag{4}$$

$$w_{\mathrm{env}}[\ell, L_m] = \begin{cases} \cos^2\left(\frac{\pi(\ell-\ell_0)}{L_m}\right) & \text{for } 1 < \ell < L_m \\ 0 & \text{for } \ell = \{1, L_m\} \end{cases} \tag{5}$$

In (3)-(5), f_m, L_m and T denote the center frequency, filter length, and the sampling period of modulation frequencies (in

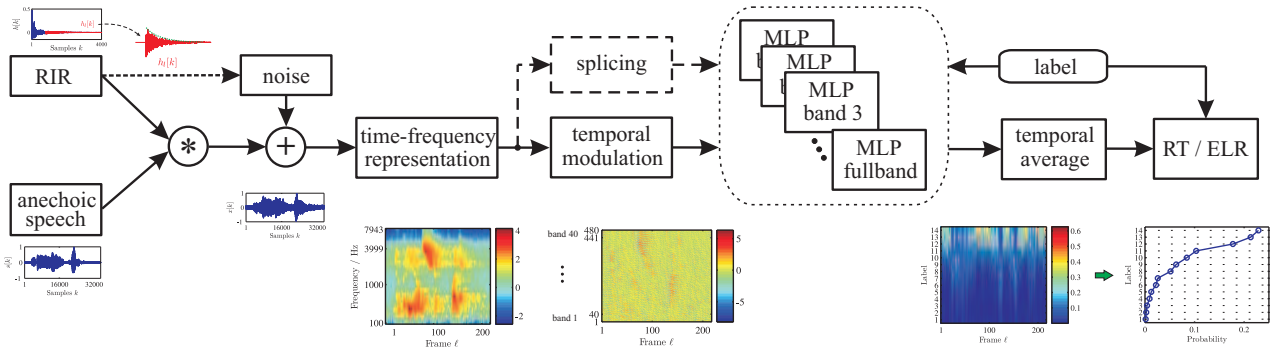

Fig. 3. System structure of ROPE to estimate the RT or the ELR in both fullband and subband analysis based on auditory-inspired features.

this case 100 Hz which follows from the 10 ms frame shift for TF representations). $\ell \in \{1, \ldots, L_m\}$ is the filter sample index and ℓ_0 represents the center sample computed as $\ell_0 = \lceil L_m/2 \rceil$.

The selection of center frequencies and the bandwidths of $g_m[\ell]$ are based on an auditory modulation filter bank proposed by Dau et al. [32]: A constant bandwidth is used for TMFs up to 10 Hz, whereas a logarithmic scaling with a constant Q-value of 2 is used above 10 Hz, resulting in 7 center frequencies: $\{0, 3, 6, 10, 17.1, 29.2, 50\}$ Hz. Both the real and the imaginary part of the band-pass modulation filters (cf. Fig. 2 (b) and (d)), are kept since the phase information is important to sustain the temporal alignment of the frequency components within the envelope window $w_{\text{env}}[\ell]$, particularly for the narrow-band filters $g_m[\ell]$ with a large temporal context. The convolution of the 40-dimensional TF representations with each filter (7 real filters and 5 imaginary filters) are concatenated, which results in 480-dimensional TMF features used as input to the MLP.

IV. ROOM PARAMETER ESTIMATOR (ROPE)

The proposed ROPE system is illustrated in Fig. 3. First, reverberant and noisy speech signals are synthesized using anechoic speech, available RIRs, and noise signals. Subsequently, the auditory-inspired features are extracted and used as input to the discriminative MLP. After temporal averaging of MLP outputs (which correspond to specific RT and ELR ranges), the output neuron with the highest activation/probability is chosen as estimated RT/ELR value (*winner-takes-all*).

A. Synthesized Reverberant and Noisy Speech

In order to simulate reverberant, diffuse noisy signals which are characteristic for the room that is considered (with the exception of test conditions that cover real noise data), we propose to use the late part $h_l[k]$ of the corresponding RIR. The early part is omitted since it can be assumed that it is correlated with reverberated speech signals, while the late part (usually after 50 ms, i.e., $k > \lceil f_s \cdot 50\text{ms} \rceil$) is assumed to be uncorrelated [1]. In other words, $n[k] * h_l[k]$ represents the diffuse noise recorded from the same room as the target speech signal $s[k]$. The synthesis model is therefore given by

$$x[k] = s[k] * h[k] + \beta \cdot n[k] * h_l[k]\,, \tag{6}$$

with the coefficient β that adjusts the signal-to-noise ratio (SNR) of the mixture according to ITU-T P.56 [33].

B. MLP Classifier

As shown in Fig. 3, classification of the RT or the ELR is performed using an MLP that maps the input features to binned classes for both output parameters. The MLP is implemented using the Kaldi ASR toolkit [34], and rectified linear units [35] are used as activation functions. The standard back-propagation via a stochastic gradient descent algorithm [36] is applied to train the MLP. The cost function is based on cross-entropy, and a softmax function is applied to the output layer to obtain posterior probabilities of RT and ELR classes. Preliminary experiments have shown that increasing the number of hidden layers does not improve performance for this specific task, therefore the results reported here were obtained using one hidden layer only. The dimensionality of the output layer corresponds to the number of RT and ELR output classes (cf. Section V-B1).

To smooth the classification result, the MLP output can be averaged over time. In this paper we consider different temporal integration window sizes that range from one to several hundreds of averaged frames. While single-frame decisions are expected to be noisy, very long observation windows introduce a delay (potentially of several seconds) that may not be acceptable for some applications. We therefore explore the trade-off between noisy classification results and integration time, and compare the result to utterance-based processing (in which all frames that belong to a longer utterance are averaged), which serves as an upper bound in this study.

V. EXPERIMENTAL SETUP

A. RT and ELR Resolution

The number of RT and ELR target classes defines the number of MLP output neurons. An adequate number of neurons is estimated on the basis of JNDs of both target values. As suggested in [2], [13], JNDs for RT and ELR are in the range of 100 ms and 1 dB, respectively. Center values are chosen based on these JNDs, and RTs within a ± 50 ms range around these values are grouped; similarly, ELR values within a ± 0.5 dB interval centered around the label are grouped. This introduces a quantization error, but at the same time

TABLE I
TEST SETS FOR ROPE TO ESTIMATE THE RT AND THE ELR IN VARIOUS ACOUSTIC ENVIRONMENTS IN BOTH FULLBAND AND SUBBAND ANALYSIS. NOISE TYPES CORRESPOND TO PINK, BABBLE, AND FAN NOISE.

Set	RIR (No.)	Noise	SNR/dB	Speech	No. of Utt.
A	Simu (12)	PN, BN	$\infty, 20, 10, 0$	TIMIT	1344×12
B	Real (12)	PN, BN	$\infty, 20, 10, 0$	TIMIT	1344×12
C	Simu+Real (12+12)	PN, BN	$30, 18, 12, -1$	TIMIT	1344×24
D	Simu+Real (12+12)	PN, BN	$30, 18, 12, -1$	ACE	50×8×24
E	Real (10)	AN, BN, FN	$18, 12, -1$	ACE	50×9×10

ensures sufficient training data for each class and seems to be acceptable from a perceptual point of view [1], [37].

B. Training and Test Data

1) Training Set: Anechoic speech from the TIMIT database [38] serves as basis for the ROPE training set (cf. Fig. 3). It contains recordings of phonetically-balanced prompted English speech in 3696 sentences uttered by 462 different speakers, recorded at 16 kHz sampling frequency. These are processed with RIRs characterized by uniformly distributed RTs (in an interval of $[200, 1500]$ ms) and ELRs (in an interval of $[-3, 30]$ dB) for fullband analysis. With the resolution described above, we obtain 14 and 34 output classes for RT and ELR, respectively. The same criterion of uniform class distribution is applied for subband analysis. For instance, the data selection for $f = 1$ kHz band results in the intervals of $[200, 1700]$ ms and $[-5, 25]$ dB. In summary, the number of RT labels is between 10 and 17, and the number of ELR classes ranges from 20 to 34. We choose 128 and 256 hidden neurons for the RT and the ELR estimation to account for the different number of output classes.

A set of RIRs for training is generated using the image method [39] [1], which provides 10 different RIRs for each RT and ELR class. Three different SNRs at $\{20, 10, 0\}$ dB are used for noisy environments and SNR $= \infty$ dB denotes the noise-free condition. Two types of noise signals are chosen, namely pink noise (PN) which exhibits similar noise energy in each frequency band of the Gammatone filter bank, and babble noise (BN) which is generated via a mixture of anechoic speech signals produced by 4 female and 4 male speakers from the WSJCAM0 corpus [40].

2) Test Sets: In order to evaluate the ROPE performance as well as to test its generalization to unseen acoustic environments, we create 5 different test sets that cover simulated and measured RIRs, various SNRs with different noise types, and speech recordings from two different sources (cf. Table I for details). Different speakers were chosen for training and test sets, resulting in speaker-independent models.

Test Sets A and B were created with the same procedure as the training set, but use different speech signals and RIRs. The speech signals of these sets are taken from the TIMIT evaluation set, which contains 1344 utterances collected from 168 speakers. Pink and babble noise at SNRs of $\{\infty, 20, 10, 0\}$ dB are added to the utterances as described earlier. Test Set A contains 12 simulated RIRs (Simu), while

[1]Implementation of the RIR generator according to *https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator*

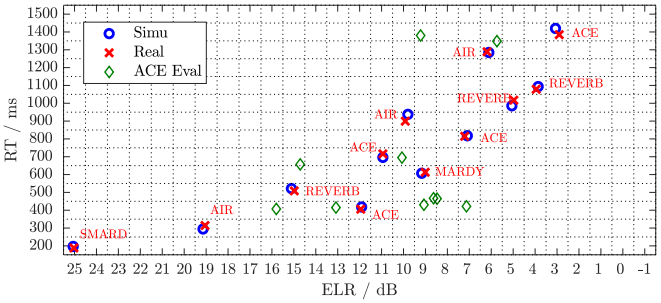

Fig. 4. Distribution of the RT and the ELR from the test sets in fullband analysis. 12 simulated (Simu) and 12 measured (Real, from [25], [41]–[44]) RIRs are used for Test Sets A, B, C and D. Test Set E contains 10 RIRs from the ACE challenge single-microphone evaluation test set [25].

Set B used 12 measured RIRs from several open-source databases as illustrated in Fig. 4. To test the effect of SNR mismatches, Test Set C combines the properties of Sets A and B with added noise at several SNRs ($\{30, 18, 12, -1\}$ dB) not used during training for both pink and babble noise. Test Set D is based on a different speech corpus (available through the ACE challenge [25]) that contains spontaneous and read speech, as well as long and short utterances. This is in contrast to the homogeneous structure of TIMIT, and is used here to study the effect of speech variability not encountered during training. The ACE database contains 50 utterances produced by 5 male and 5 female talkers in different dialects of international English with a mix of native and non-native English speakers.

Finally, the evaluation test set for single-channel processing from the ACE challenge is used as Set E, which contains 4500 utterances categorized by 3 noise types and 3 SNRs (-1, 12, and 18 dB). These noises are ambient (AN), babble (BN), and fan (FN) noise recorded in the same room as the corresponding RIRs. The RIRs were measured in 5 different rooms with 2 different microphone positions; the resulting values for RT and ELR are shown in Fig. 4. Test Set E is the one that differs the most from the training set, since it contains inhomogeneous speech material, different noise types obtained during measurements, and data that was reverberated using measured RIRs. Comparisons with other RT and ELR estimators are reported using this test set.

C. Evaluation Metrics

The first evaluation metric is estimation error $e_X = \hat{X} - X$, i.e., the difference between the estimated value and the ground truth with X denoting either the RT or the ELR. The root mean squared error (RMSE) is reported for each measure ($\mathrm{RMSE}_{\mathrm{RT}}$ and $\mathrm{RMSE}_{\mathrm{ELR}}$, respectively), as well as the underlying distribution of e_X using box plots. A system that always outputs the same value close to the median could produce a relatively low RMSE (although it does not actually perform a classification task). We therefore report a second measure to quantify the estimation accuracy, i.e., the Pearson correlation coefficient ρ between estimated and true parameters, as proposed in the ACE challenge. Higher ρ towards 1 exhibits more accurate estimations. Third, the real-time factor (RTF) is reported, which

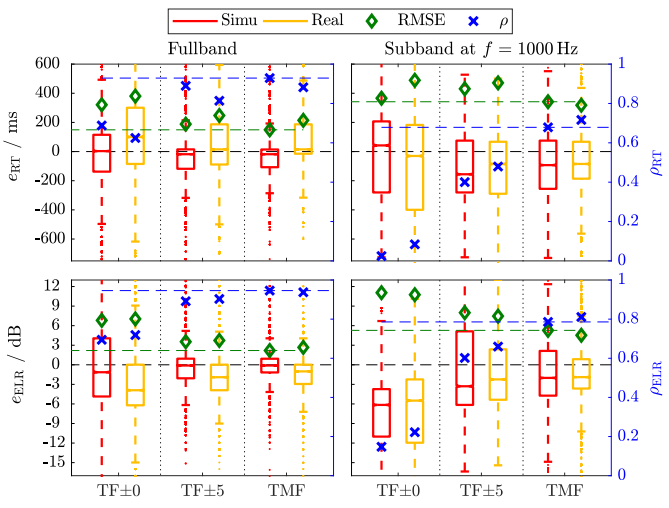

Fig. 5. ROPE performance for estimating the RT and the ELR in fullband (left panels) and subband at $f = 1$ kHz analysis (right panels) with Test Sets A and B for both feature types. Box plot of the estimation error, RMSE and ρ (right y-axis) are shown for *Simu* and *Real* RIRs, both with 1344×12 utterances, which cover pink and babble noise at multiple SNRs. Horizontal dashed lines correspond to RMSE and ρ results for TMF with *Simu*.

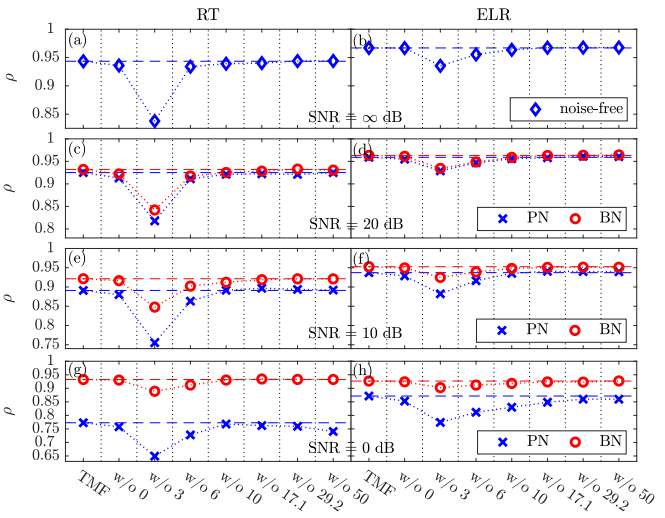

Fig. 6. ROPE performance ρ_{RT} and ρ_{ELR} for fullband data using different sets of TMFs in noise-free and noisy conditions in pink and babble noise. Horizontal dashed lines correspond to results for the full set of modulation frequencies. 'w/o' X specifies the modulation frequency that was omitted for the corresponding data points.

is the total computation time divided by the total duration of all processed speech data. The RTF is used to evaluate the potential of models for practical real-time applications that are constrained by computational complexity such as hearing aids or the front-end speech processors in mobile devices.

VI. Results and Discussion

A. Overall Results

Test Sets A and B are used to validate the ROPE model based on RIR sets in both *Simu* and *Real* conditions that widely cover RTs and ELRs. The TF representations with and without temporal context (denoted as TF±0 and TF±5, respectively), as well as temporal modulation frequency (TMF) features are tested as explained in Section III to investigate the importance of temporal modulations explicitly captured on feature level.

The results in Fig. 5 show that the ROPE model with TMF performs principally well for the RT and the ELR estimation, e.g., with median errors close to 0 (left y-axis) and correlation values between 0.93 and 0.95 (right y-axis) in fullband processing using Set A. This prediction performance is later compared to other approaches in Section VI-F3. Further, TMF features perform consistently better both in terms of RMSE and correlation in comparison to the TF representations with temporal context, while TF features without temporal context perform far worse with almost doubled RMSEs compared to TMF. This highlights the importance of temporal modulation cues to characterize the effect of reverberation, and the dedicated temporal modulation filter bank (cf. Section III-B) performs more effectively than a mere concatenation that supplies additional temporal context. The contribution of individual modulation frequencies will be analyzed in the next section. Third, the estimation from subband data (right panels in Fig. 5) is degraded in comparison to the fullband processing which is reflected by the median error (which is -100 ms or -2 dB on average), the standard deviation of estimated values, and a lower correlation. This is especially notable for TF features without temporal context from subband data with rather low correlation. With TMF features, a good performance is still achieved, which indicates that the redundant information contained in spoken languages can partially be preserved when explicitly extracting modulation features. The ROPE performance for all individual subbands is presented in Section VI-F4. Finally, Fig. 5 also shows the performance difference between *Simu* and *Real* RIRs to be quite small, indicating that this approach generalizes from simulated training data to measured RIRs encountered during testing.

B. Analysis of Modulation Filters

To pinpoint the feature characteristics of TMF features that contribute to the improved performance when compared to TF features, we analyze the contribution of individual modulation frequencies. To this end, each filter is evaluated based on a *leave-one-out* procedure (i.e., feature components that correspond to one of the seven TMFs are removed from the feature vector, and an MLP is trained with the resulting feature) that is performed in different noise types at different SNRs. Experiments are conducted using Sets A and B in both fullband and subband at $f = 1$ kHz processing. Fig. 6 (fullband) and Fig. 7 (subband) show the results for the omission of one specific modulation filter. The result in terms of RMSE follows the very same trend and is therefore not shown for better readability.

The correlation values show that modulation filters with center frequencies below 10 Hz are more important, with the 3 Hz filter being by far the most important. This could be due to the temporal smearing caused by reverberation, which especially affects the temporal on- and offsets of syllables, which often exhibit a peak modulation frequency around

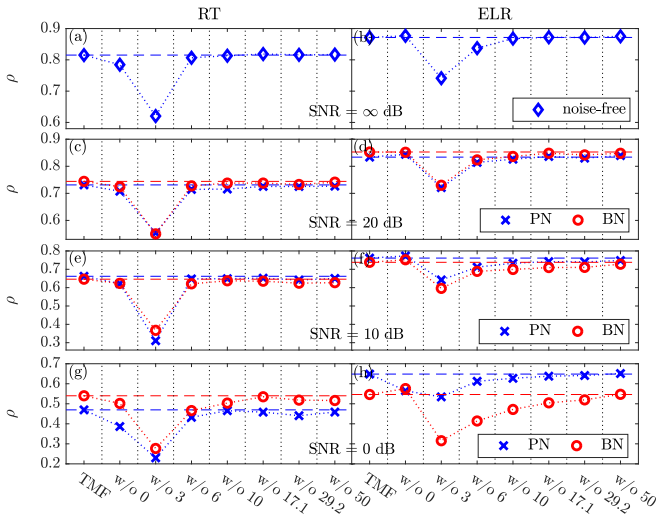

Fig. 7. ROPE performance ρ_{RT} and ρ_{ELR} for subband data ($f = 1$ kHz) using different sets of TMFs in noise-free and noisy conditions in pink and babble noise. Horizontal dashed lines correspond to results for the full set of modulation frequencies. 'w/o' X specifies the modulation frequency that was omitted for the corresponding data points.

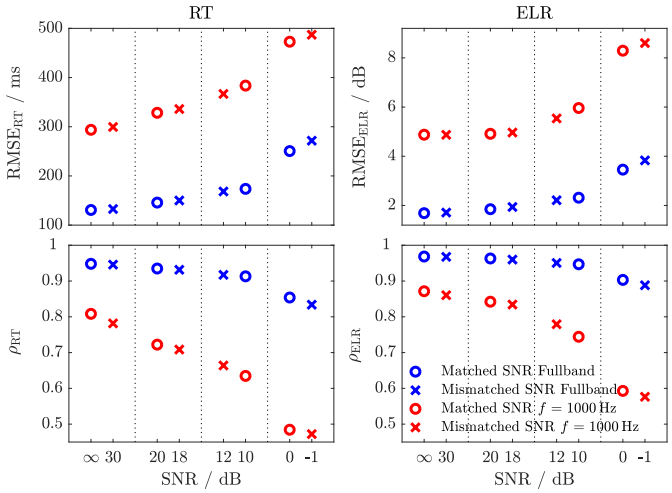

Fig. 8. ROPE performance in terms of RMSE and ρ averaged over *Simu* and *Real* RIRs as well as pink and babble noise against deviated SNRs when estimating the RT and the ELR in both fullband and subband at $f = 1$ kHz analysis. Test Sets A and B are used as the matched SNR scenario ($\{\infty, 20, 10, 0\}$ dB), while Test Set C is used as the mismatched SNR scenario ($\{30, 18, 12, -1\}$ dB).

$3 - 4$ Hz [24]. It seems that the 3 Hz component is more important for RT than for ELR estimation, since the decrease of correlation to the ground truth is stronger for RT.

Filters with modulation frequencies above 10 Hz do not contribute to the ROPE performance in many scenarios, especially in low-noise conditions. At the same time, none of the filters has a detrimental effect on the model quality, and in some conditions all high-frequency filters are required to reach the optimal model performance. For example, ρ decreases when any filter with a frequency of $\{17.1, 29.2, 50\}$ Hz is omitted for both RT and ELR estimation in fullband analysis at 0 dB SNR in pink noise (cf. Fig. 6 (g)-(h)). In the following experiments, we therefore continue to use the complete modulation filter set.

Fig. 6 also shows that babble noise has a very limited effect on the model performance, while both RT and ELR estimates are degraded in pink noise when using fullband data. We assume this is a specific property of the babble masker: Time-frequency patterns similar to the target should result in similar degradation and hence similar feature patterns. Further, its temporal modulation cues carry the same RT information as the reverberated target speech, while its temporal modulation cues should not affect the ELR since it is diffuse. On the other hand, the diffuse component of the noise could mask the temporal onset and offset of the syllables, resulting in the relative importance of the modulation filter at 3 Hz compared to other filters. For subband data, babble noise has a very strong effect on estimation performance (cf. Fig. 7). Since the SNR is calculated using fullband data (in accordance with [33]), strong local SNR fluctuations in each subband can be expected due to the non-uniform spectral distribution of (babble) speech. For the $f = 1$ kHz band, we assume a lower local SNR in comparison to pink noise is the reason for the degraded subband performance.

C. SNR Mismatch for Training and Testing

In realistic environments, arbitrary SNRs are encountered, while training sets are often limited to specific SNRs. This is also true for the open-source training sets used in our study, which usually cover SNRs in steps of 5 or 10 dB. While this allows for an SNR-dependent analysis, it also bears the risk of creating SNR-specific models. Hence, it is important to test the robustness of speech processing algorithms for mismatched SNRs as well. We therefore use the Test Set C with SNRs of $\{30, 18, 12, -1\}$ dB that differ from the training SNRs of $\{\infty, 20, 10, 0\}$ dB, which is compared to the performance for matched SNR testing (Sets A and B). The results are shown in Fig. 8: Generally, a lower SNR results in a consistent decrease of estimation accuracy, and this consistency is observed for matched as well as mismatched SNRs. This indicates that the ROPE approach generalizes well to the deviation from SNR mismatches. This is also reflected in the similar results obtained with neighboring unseen SNRs and seen SNRs, e.g., the average absolute differences of RMSE for the SNRs at 12 dB and 10 dB are only 11 ms and 0.27 dB.

D. Effect of Speech Variability

For previous results, training and testing have been carried out using data from disjunct speaker sets from the TIMIT database. Although this introduces speaker-dependent variability between training and test, a higher variability would be encountered in realistic test settings, since the TIMIT data is relatively homogeneous. Test Set D (cf. Table I for details) from the ACE database contains utterances with strongly varying duration recorded with different microphones from a different group of speakers, and therefore adds factors of additional speech variability. Mismatched SNRs ($\{30, 18, 12, -1\}$ dB) between training and test are used since this is a more realistic assumption than matched SNRs.

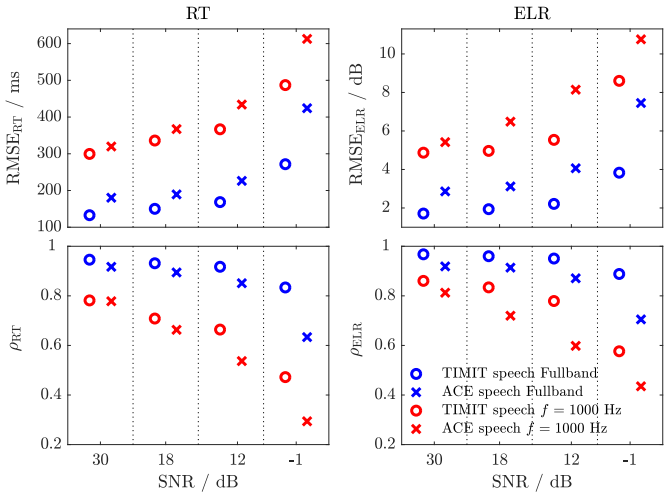

Fig. 9. ROPE performance in terms of RMSE and ρ averaged over *Simu* and *Real* RIRs as well as pink and babble noise against two different speech sources when estimating the RT and the ELR in both fullband and subband at $f = 1$ kHz analysis. Test Set C is used as the homogenous speech source scenario from TIMIT corpus, whereas Test Set D applies the speech recordings from the ACE challenge.

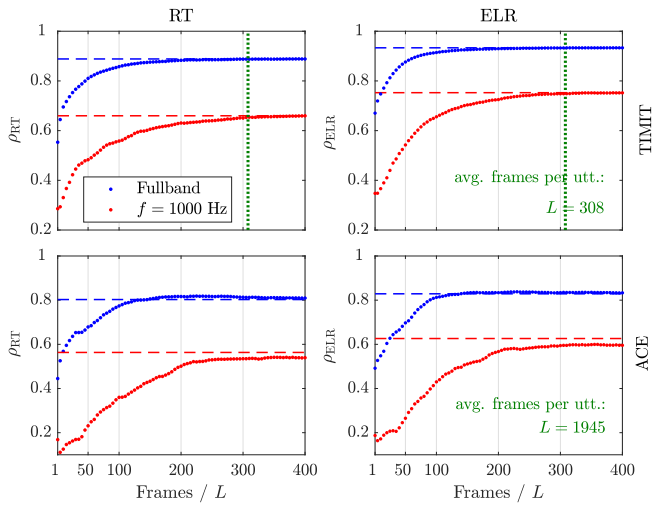

Fig. 10. ROPE performance for *window-based* processing for utterances from Set C (TIMIT, upper panels) and Set D (ACE, lower panels). RMSE results follow the same trend as the performance for ρ and are therefore not shown. The results for *utterance-based* processing are shown as horizontal dashed lines for comparison.

Fig. 9 shows that for high SNRs ($\{30, 18\}$ dB), $\mathrm{RMSE}_{\mathrm{RT}}$ and $\mathrm{RMSE}_{\mathrm{ELR}}$ increase by about 30 ms and 1 dB, and ρ_{RT} and ρ_{ELR} decrease by approximately 0.03 and 0.06, respectively. These minor degradations indicate that ROPE is robust against the added variability described above. On the other hand, as SNR decreases, a stronger performance degradation can be observed with the ACE speech data, i.e., there is an interaction between SNR and added variability, resulting in a further degradation of estimation accuracies at SNRs below 0 dB.

E. Temporal Integration Window

The previous results were obtained by averaging MLP output obtained from the complete utterance. This introduces a considerable delay, which is not compatible with applications that require low latency. In this section, we therefore analyze the temporal integration window required for accurate room parameter estimates. Results for smaller windows are obtained by calculating estimates from only the first frame, which is systematically extended to the window containing the first L frames of the utterance. We refer to the two processing modes as *utterance-based* and *window-based*. Scores for both processing modes are shown in Fig. 10, which includes the performance with different values for L shown on the x-axis. The results for fullband *window-based* processing saturate at $170 - 200$ frames (with frame rate of 10 ms), at which point the *utterance-based* scores are reached. This observation is consistent over two corpora tested, i.e., the TIMIT speech data (Set C) with average duration of utterances of 308 frames, as well as data from the ACE challenge (Set D) which exhibits a relatively long average duration of utterances (1945 frames).

For subband data centered at 1 kHz, a longer integration time is required to approach *utterance-based* performance, in most cases about 100 additional frames, which is presumably caused by the limited information in each subband compared to the fullband information. This result shows that integration over a complete utterance is not required by the ROPE system to produce accurate results, and a reasonable integration time is between 1.7 and 3 seconds (depending on the choice of fullband or subband processing).

F. Performance for the ACE Challenge Evaluation Database

In order to test ROPE in realistic recording environments, we use the single-microphone evaluation database from the ACE challenge [25] (Test Set E in Table I).

1) RT and ELR Estimation: For the RT estimation from fullband and 1 kHz-subband data (cf. upper panels of Fig. 11), ROPE performance increases with the SNR, and correlations above 0.8 and median errors close to 0 ms are obtained for ambient and babble noise at SNRs of $\{18, 12\}$ dB. Note that babble noise from the ACE challenge is not identical to the babble noise used for training: The ACE babble noise consists of recordings of $4 - 7$ continuously talking people positioned around the microphone [25], which adds a strong non-diffuse/spatial component to this noise type. In contrast, the babble noise for training is completely diffuse (cf. (6)) and contains speech from 8 different talkers (cf. Section V-B1). ROPE performance is degraded in the presence of fan noise (especially at -1 dB), which could arise due to its different noise characteristics. On the other hand, ambient noise (which is also not seen in training but is similar to pink noise) produces good results, which hints at the generalization capabilities of the ROPE approach if training and testing noise types share similarities.

Similar trends can be observed for the ELR estimation as illustrated in Fig. 11 (bottom panels) with the exception of babble noise, for which ELR is underestimated. We assume this is caused by the differences between babble noise for training and testing (see above) and the fact that spatial source positions influence the ELR, while the RT is mostly invariant

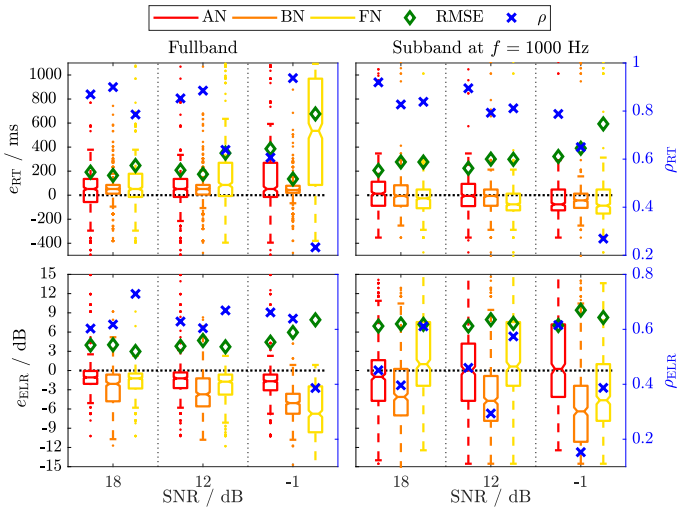

Fig. 11. ROPE performance for estimating the RT and the ELR in fullband and subband at $f = 1$ kHz analysis with Set E, i.e., the single-microphone evaluation database from the ACE challenge [25]. Box plot of the estimation error, RMSE and ρ (right y-axis) are illustrated in terms of ambient, babble and fan noise at SNRs of $\{18, 12, -1\}$ dB, each with 500 utterances.

to them: Since the ROPE algorithm is a speech-specific approach, ELR estimates are influenced by the babble noise from the ACE testing set that includes spatial components associated with masking speech. Since the masking speakers are usually farther away from the microphone than the original target speaker, the ELR would tend to be underestimated, especially at low SNRs.

2) Computational Complexity: The computational complexity of ROPE is analyzed since it is an important factor when estimation algorithms should be implemented on small-footprint devices. In this work, we evaluate the RTF using a GPU-based workstation, i.e., the specific factor does not correspond to small scale hardware; however, the relation to RTFs of other approaches (see below) does. The complexity of ROPE during test time is dominated by the feature extraction and MLP forward processing. In our set of experiments, the calculation of auditory-inspired TMF features is implemented in Matlab running on an *Intel x86_64 64bit CPU 2.0 GHz* platform, for which an average RTF of 0.142 is obtained for the 4500 test speech files (cf. Table I). The MLP output posterior probabilities are obtained via neural networks compiled with a *Tesla K20c NVIDIA* GPU with an average RTF of 0.033. Hence, the average RTF of ROPE for fullband data equals 0.175. For subband data, the average RTF for feature extraction from a single frequency channel decreases to 0.027, while the MLP forward run is just slightly changed with an RTF of 0.029. Therefore, the resulting average RTF for one frequency band is 0.056.

3) Performance Comparison: ROPE results are compared to other single-microphone state-of-the-art RT and ELR estimators that were implemented by their respective authors and tested on data provided by the ACE challenge database (cf. [25], [50] for detailed descriptions of these algorithms). Note that the ACE challenge focused on the DRR estimation using $t_e \approx 2.5$ ms in (1), which is different from $t_e = 50$ ms

TABLE II
PERFORMANCE COMPARISON WITH OTHER SINGLE-MICROPHONE STATE-OF-THE-ART RT AND ELR ESTIMATORS BASED ON THE ACE CHALLENGE EVALUATION DATABASE. † DENOTES THE MULTI-MICROPHONE CONFIGURATION.

Fullband Estimation

Estimator	RMSE/ms	ρ_{RT}	RTF	RMSE/dB	ρ_{ELR}	RTF
QAReverb [45]	255	0.778	0.400	4.86	0.058	0.391
NIRA [46]	389	0.302	0.899	3.85	0.558	0.899
SRMR [47]	380	0.220	0.457	5.82	-0.084	0.540
ROPE	285	0.716	0.175	4.81	0.556	0.175

Subband Estimation at $f = 1000$ Hz

Estimator	RMSE/ms	ρ_{RT}	RTF	RMSE/dB	ρ_{ELR}	RTF
ML-RTE [48]	358	0.699	0.939	-	-	-
ParVel† [49]	-	-	-	3.21	0.415	0.134
ROPE	338	0.751	0.056	7.63	0.421	0.056

used throughout this paper. Nevertheless, since DRR and ELR estimation are similar problems because both are based on energy ratios with different time constants, the performance should be comparable in terms of RMSE, the correlation coefficient ρ, and RTF. To the best of our knowledge, a blind algorithm for subband DRR or ELR estimation from single-channel data does not exist. As a baseline method, we therefore exploit the algorithm based on particle velocity (ParVal) [49] that was developed for multi-channel data (a spherical microphone array with 32 microphones).

As shown in Table II, the ROPE approach achieves competitive performance in terms of RMSE and correlation when compared to the best result for single-channel RT estimation, i.e., QAReverb [8], [45]. For fullband ELR estimation, ROPE provides nearly the same ρ_{ELR} compared to the best ACE challenge contribution for single-channel data (Non-Intrusive Room Acoustic (NIRA) estimator [18], [46]), despite a slightly higher $\mathrm{RMSE}_{\mathrm{ELR}}$. Further, slightly better performance is achieved by ROPE in subband analysis at $f = 1$ kHz, when compared to the subband RT estimator ML-RTE [48] (originally from [5] which is also the only algorithm submitted to subband RT estimation task in the ACE challenge). Compared to the the multi-microphone ParVal method [49] in subband analysis at $f = 1$ kHz, ROPE shows worse $\mathrm{RMSE}_{\mathrm{ELR}}$ but better ρ_{ELR}.

As summarized in Table II, most algorithms perform well for one specific task, but strongly degrade (or are even not applicable) for other tasks. For subband estimation from a single microphone in ACE challenge, only one algorithm was proposed for the RT estimation and no one for the ELR estimation. ROPE seems to the only algorithm that potentially provides reliable results for both the RT and the ELR estimation in fullband and subband processing. Furthermore, the relatively low RTF achieved by ROPE indicates its potential for practical applications. Since RTFs of other algorithms were provided by their respective authors based on different (but presumably similar) hardware processors, an exact comparison on identical hardware should be performed in the future for algorithms that are freely accessible.

4) Integration of Subband Data: Results from subband data reported so far were obtained with the 1 kHz frequency band. In this section, the performance for all 40 individual frequency bands is reported, and the benefit from integration of specific

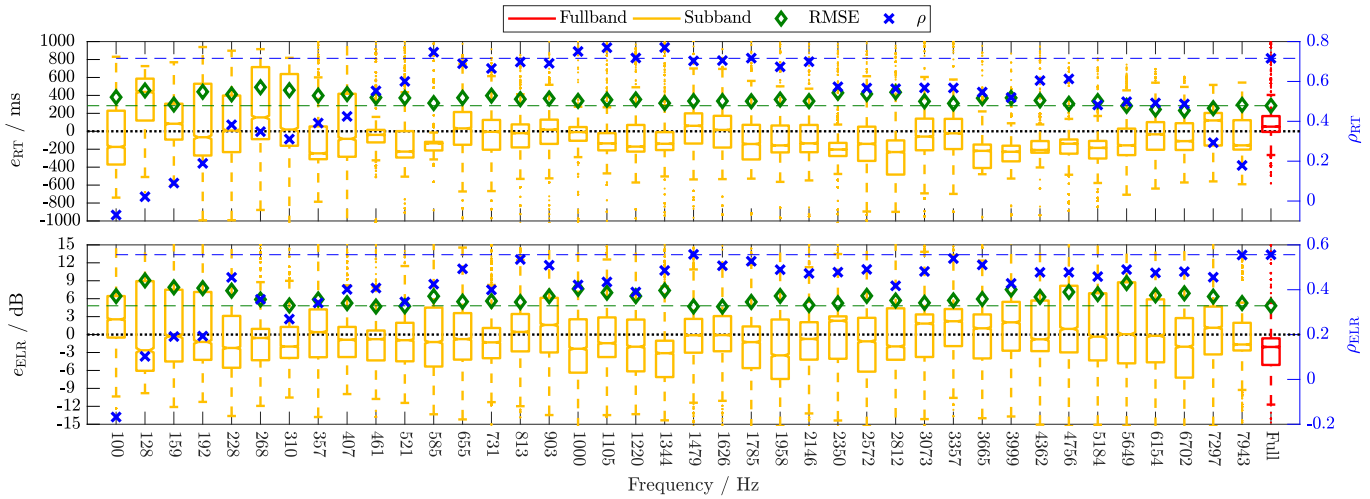

Fig. 12. ROPE performance of the RT and the ELR estimation with the ACE single-microphone evaluation database in both fullband and all 40 subband analysis. Box plot of the estimation error, RMSE and ρ (right y-axis) are illustrated in all noises (AN, BN, FN) for all SNRs ($\{18, 12, -1\}$ dB). Horizontal dashed lines correspond to fullband results of RMSE and ρ.

subbands is analyzed.

Results for individual bands are shown in Fig. 12 (with the fullband result on the right and as dashed horizontal line). For RT estimation, performance consistently degrades for center frequencies below 550 Hz, and appears to be stable from 550 to 2200 Hz. For higher frequencies, performance deteriorates again, but only slowly. For ELR estimation, the results have a higher variance across center frequency. However, correlations for data below 550 Hz are consistently low, while individual filters approach (and in some cases achieve) fullband performance levels for higher frequencies. This overall frequency dependence is roughly consistent with the design of the algorithm, which is based on speech-specific stimuli and therefore can be expected to perform best in the frequency range associated with high speech energy, i.e., 300 to 3400 Hz [51].

Individual frequency bands potentially carry complementary information, which could provide better results after integration compared to fullband processing. The integration of three different frequency ranges is therefore tested: Group A ranges from 400 to 1250 Hz and is motivated by the suggestion in ISO-3382 [2]. Second, a frequency range corresponding to high average energy in (narrow-band) speech is chosen (300 to 3400 Hz, Group B), which is in line with the speech-specific approach. Third, a frequency range based on adjacent subbands with high RT performance is chosen (585 to 2146 Hz, Group C). Table III compares fullband results to each of these selections. While Group A performs worse than the fullband approach for ELR data, Groups B and C improve estimation accuracy for all four measures that quantify model prediction quality. Results for Groups B and C are also very consistent, which indicates that the specific selection of frequency bands seems not to be crucial, as long as speech-relevant components are included. These integrated subband results outperform almost all baseline fullband models reported in the previous section (which however have not been optimized by integrating information from different frequency channels). Note that this benefit comes at the cost of increased computational cost (see values for RTF in Table III). For applications with limited resources, a selection and integration of individual high-performance subbands might be required.

TABLE III
FULLBAND RT AND ELR ESTIMATIONS OBTAINED BY AVERAGING SUBBAND ESTIMATIONS WITHIN DIFFERENT FREQUENCY RANGES. HORIZONTAL DASHED LINES CORRESPOND TO RMSE AND ρ RESULTS FOR FULLBAND PROCESSING.

Fullband estimation	RMSE		ρ		RTF
	RT/ms	ELR/dB	ρ_{RT}	ρ_{ELR}	
ROPE Fullband	285	4.81	0.716	0.556	0.175
Avg. $[400, 1250]$ Hz	216	5.20	0.830	0.430	0.616
Avg. $[300, 3400]$ Hz	206	3.94	0.853	0.601	1.344
Avg. $[585, 2146]$ Hz	207	4.15	0.861	0.583	0.784

VII. CONCLUSIONS

A novel blind acoustic room parameter estimator (ROPE) to estimate the RT and the ELR directly from single-channel speech in both fullband and subband processing has been presented and analyzed. The use of temporal modulation features as direct input to a multi-layer perceptron improved performance over simpler time-frequency features. The arrangement of modulations into a filter bank enabled a systematic analysis of the importance of modulation frequencies, from which features centered around 3 Hz emerged to be the most important. By using test sets with different characteristics, we showed that ROPE is robust against different RIRs, SNRs, and variability covered in different speech databases, despite the fact that an interaction of speech data variability and high-noise conditions below 0 dB can severely affect the prediction results. Further, ROPE was compared to other blind state-of-the-art RT and ELR estimators using test data from the Acoustic Characterisation of Environments (ACE) challenge. Comparable results with the best RT estimation of the competition were obtained, as well as comparable correlation coefficients of the ELR estimation in comparison to the best DRR estimators. ROPE

was applied for predicting RT as well as ELR in both subband and fullband single-microphone data, which is a unique feature of this approach. The computational cost quantified in terms of the real-time factor was found to be low in comparison to other approaches, and the temporal integration window required for stable results was in the range of a few seconds, which means that the algorithm should be of interest for real-time applications. In addition, fullband RTs and ELRs can be obtained alternatively by averaging the subband results within a certain frequency range, and the narrow-band speech frequency range from 300 Hz to 3400 Hz related to speech production was found to be a good candidate.

ACKNOWLEDGMENT

The authors would like to thank James Eaton for the data release of the ACE challenge.

REFERENCES

[1] H. Kuttruff, *Room Acoustics*, 4th ed. London: Spon Press., 2000.
[2] *Acoustics - Measurement of Room Acoustic Parameters*, the International Organization for Standardization (ISO) Std., Aug. 2009.
[3] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *Audio Engineering Society 108th Convention*, no. 5093, Paris, France, Feb. 2000, pp. 1–23.
[4] R. Ratnam, D. L. Jones, B. C. Wheeler, J. W. D. O'Brien, C. R. Lansing, and A. S. Feng, "Blind estimation of reverberation time," *Journal of the Acoustical Society of America*, vol. 114, no. 5, pp. 2877–2892, 2003.
[5] H. Löllmann, E. Yilmaz, M. Jeub, and P. Vary, "An improved algorithm for blind reverberation time estimation," in *International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Tel Aviv, Isreal, Aug. 2010.
[6] J. Vieira, "Estimation of reverberation time without test signals," in *Audio Engineering Society 118th Convention*, no. 6499, Barcelona, Spain, May 2005, pp. 1–7.
[7] M. R. Schroeder, "New method of measuring reverberation time," *Journal of the Acoustical Society of America*, vol. 37, no. 3, pp. 409–412, 1965.
[8] T. de M. Prego, A. A. de Lima, S. L. Netto, B. Lee, A. Said, R. W. Schafer, and T. Kalker, "A blind algorithm for reverberation-time estimation using subband decomposition of speech signals," *Journal of the Acoustical Society of America*, vol. 131, no. 4, pp. 2811–2816, 2012.
[9] J. Y. C. Wen, E. A. P. Habets, and P. A. Naylor, "Blind estimation of reverberation time based on the distribution of signal decay rates," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, USA, Apr. 2008, pp. 329–332.
[10] J. Eaton, N. D. Gaubitch, and P. A. Naylor, "Noise-robust reverberation time estimation using spectral decay distributions with reduced computational cost," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013, pp. 161–165.
[11] T. H. Falk and W.-Y. Chan, "Temporal dynamics for blind measurement of room acoustical parameters," *IEEE Transactions on Instrumentation and Measurement*, vol. 59, no. 4, pp. 978–989, Apr. 2010.
[12] T. J. Cox, F. Li, and P. Dalington, "Extracting room reverberation time from speech using artificial neural networks," *Journal of Audio Engineering Society*, vol. 94, no. 4, pp. 219–230, 2001.
[13] P. Kendrick, T. J. Cox, F. F. Li, Y. Zhang, and J. A. Chambers, "Monaural room acoustic parameters from music and speech," *Journal of the Acoustical Society of America*, vol. 124, no. 1, pp. 278–287, 2008.
[14] Y.-C. Lu and M. Cooke, "Binaural estimation of sound source distance via the direct-to-reverberant energy ratio for static and moving sources," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1793–1805, 2010.
[15] Y. Hioka, K. Niwa, S. Sakauchi, K. Furuya, and Y. Haneda, "Estimating direct-to-reverberant energy ratio using D/R spatial correlation matrix model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2374–2384, 2011.
[16] J. Eaton, A. H. Moore, P. A. Naylor, and J. Skoglund, "Direct-to-reverberant ratio estimation using a null-steered beamformer," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, Apr. 2015, pp. 46–50.
[17] M. Kuster, "Estimating the direct-to-reverberant energy ratio from the coherence between coincident pressure and particle velocity," *Journal of the Acoustical Society of America*, vol. 130, no. 6, pp. 3781–3787, 2011.
[18] P. P. Parada, D. Sharma, and P. A. Naylor, "Non-intrusive estimation of the level of reverberation in speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, May 2014, pp. 4718–4722.
[19] D. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. NY: Wiley/IEEE Press, 2006.
[20] R. M. Stern and N. Morgan, "Hearing is believing: Biologically inspired methods for robust automatic speech recognition," *IEEE Signal Processing Magazine*, vol. 29, pp. 34–43, 2012.
[21] Z. Chen and V. Hohmann, "Online monaural speech enhancement based on periodicity analysis and a priori SNR estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 11, pp. 1904–1916, 2015.
[22] F. Xiong, S. Goetze, and B. T. Meyer, "Blind estimation of reverberation time based on spectro-temporal modulation filtering," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013, pp. 443–447.
[23] ——, "Joint estimation of reverberation time and direct-to-reverberation ratio from speech using auditory-inspired features," in *ACE Challenge Workshop, satellite event of IEEE-WASPAA*, New Paltz, NY, USA, Oct. 2015.
[24] R. Drullman, J. M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *Journal of the Acoustical Society of America*, vol. 95, no. 2, pp. 1053–1064, 1994.
[25] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "Estimation of room acoustic parameters: The ACE challenge," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 10, pp. 1681–1693, 2016.
[26] M. Karjalainen, P. Antsalo, A. Mäkivirta, T. Peltonen, and V. Välimäki, "Estimation of modal decay parameters from noisy response measurements," *Journal of the Acoustical Society of America*, vol. 11, pp. 867–878, 2002.
[27] P. A. Naylor and N. D. Gaubitch, Eds., *Speech Dereverberation*. London: Springer, 2010.
[28] A. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 14–22, 2012.
[29] F. Xiong, B. T. Meyer, N. Moritz, R. Rehr, J. Anemüller, T. Gerkmann, S. Doclo, and S. Goetze, "Front-end technologies for robust ASR in reverberant environments - spectral enhancement-based dereverberation and auditory modulation filterbank features," *EURASIP Journal on Advances in Signal Processing*, vol. 2015:70, pp. 1–18, 2015.
[30] M. R. Schädler, B. T. Meyer, and B. Kollmeier, "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition," *Journal of the Acoustical Society of America*, vol. 131, no. 5, pp. 4134–4151, 2012.
[31] N. Moritz, J. Anemüller, and B. Kollmeier, "An auditory inspired amplitude modulation filter bank for robust feature extraction in automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 11, pp. 1926–1937, 2015.
[32] T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling auditory processing of amplitude modulation. I. detection and masking with narrow-band carriers," *Journal of the Acoustical Society of America*, vol. 102, no. 5, pp. 2892–2905, 1997.
[33] *Objective Measurement of Active Speech Level*, International Telecommunications Union (ITU-T) Recommendation P.56 Std., Mar. 1993.
[34] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, "The Kaldi speech recognition toolkit," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Big Island, HI, USA, Jul. 2011.
[35] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning*, Haifa, Israel, Jun. 2010, pp. 807–814.
[36] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," *Parallel distributed processing: Explorations in the microstructure of cognition*, vol. 1: Foundations. MIT Press, 1986.
[37] J. S. Bradley, R. Reich, and S. G. Norcross, "A just noticeable difference in C_{50} for speech," *Applied Acoustics*, vol. 58, pp. 99–108, Oct. 1999.
[38] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus LDC93S1," Linguistic Data Consortium (LDC), 1993.

[39] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.
[40] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJCAM0: A British English speech corpus for large vocabulary continuous speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Detroit, Michigan, USA, May 1995, pp. 81–84.
[41] J. Wen, N. D. Gaubitch, E. Habets, T. Myatt, and P. A. Naylor, "Evaluation of speech dereverberation algorithms using the (MARDY) database," in *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Paris, France, Sep. 2006.
[42] M. Jeub, M. Schäfer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Proceedings of International Conference on Digital Signal Processing*, Santorini, Greece, July 2009, pp. 1–4.
[43] J. K. Nielsen, J. R. Jensen, S. H. Jensen, and M. G. Christensen, "The single- and multichannel audio recordings database (SMARD)," in *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Antibes, France, Sep. 2014, pp. 40–44.
[44] K. Kinoshita, M. Delcroix, S. Gannot, E. A. P. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, A. Sehr, and T. Yoshioka, "A summary of the REVERB challenge: State-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 7, pp. 1–19, 2016.
[45] T. de M. Prego, A. A. de Lima, R. Zambrano-López, and S. L. Netto, "Blind estimators for reverberation time and direct-to-reverberant energy ratio using subband speech decomposition," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA*, New Paltz, NY, USA, Oct. 2015.
[46] P. P. Parada, D.Sharma, T. van Waterschoot, and P. A. Naylor, "Evaluating the non-intrusive room acoustics algorithm with the ACE challenge," in *ACE Challenge Workshop, satellite event of IEEE-WASPAA*, New Paltz, NY, USA, Oct. 2015.
[47] M. Senoussaoui, J. F. Santos, and T. H. Falk, "Srmr variants for improved blind room acoustics characterization," in *ACE Challenge Workshop, satellite event of IEEE-WASPAA*, New Paltz, NY, USA, Oct. 2015.
[48] H. Löllmann, A. Brendel, P. Vary, and W. Kellermann, "Single-channel maximum-likelihood T60 estimation exploiting subband information," in *ACE Challenge Workshop, satellite event of IEEE-WASPAA*, New Paltz, NY, USA, Oct. 2015.
[49] H. Chen, P. N. Samarasinghe, T. D. Abhayapala, and W. Zhang, "Estimation of the direct-to-reverberant energy ratio using a spherical microphone array," in *ACE Challenge Workshop, satellite event of IEEE-WASPAA*, New Paltz, NY, USA, Oct. 2015.
[50] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "ACE challenge results technical report," Imperial College London, Tech. Rep., 2016. [Online]. Available: https://arxiv.org/abs/1606.03365
[51] *Transmission Performance Characteristics of Pulse Code Modulation Channel*, International Telecommunications Union (ITU-T) Recommendation G.712 Std., Nov. 2001.

Feifei Xiong (S'12) received a B.Sc. degree in electronic and information engineering from Shandong University, China, in 2007, and an M.Sc. in communication and information technology from the University of Bremen, Germany, in 2009. From 2009 to 2013, he worked as a scientific engineer in the Project Group Hearing, Speech, and Audio Technology at the Fraunhofer Institute for Digital Media Technology IDMT in Oldenburg, Germany. Since 2013, he is pursuing his Ph.D. degree in the Medical Physics group at the University of Oldenburg, Germany, in collaboration with Project Group Hearing, Speech, and Audio Technology at the Fraunhofer Institute for Digital Media Technology IDMT in Oldenburg, Germany. Since 2017, he is a Research Associate with the Medical Physics group at the University of Oldenburg, Oldenburg, Germany. His research interests include room acoustics, speech dereverberation, distant and dysarthric speech recognition, and machine learning.

Stefan Goetze (M'09) is Dept. Head of Department "Hearing, Speech, and Audio Technology" and Head of "Automatic Speech Recognition" group at the Fraunhofer Institute for Digital Media Technology IDMT in Oldenburg, Germany. He received his Ph.D. in 2013 at the University of Bremen, Germany, where he was a Research Engineer from 2004 to 2008. His research interests are sound pick/up, processing and enhancement, such as noise reduction, acoustic echo cancellation and dereverberation, as well as assistive technologies, humanmachine-interaction, detection and classification of acoustic events and automatic speech recognition. He is a lecturer at the University of Bremen and project leader of national and international research projects in the field of acoustic signal enhancement and recognition technologies.

Birger Kollmeier received the Ph.D. degree in physics (supervisor: Prof. Dr. M.R. Schroeder) and the Ph.D. degree in medicine from the Universität Göttingen, Germany, in 1986 and 1989, respectively. Since 1993, he has been a Full Professor of physics at the Universität Oldenburg, Oldenburg, Germany. He is head of the Cluster of Excellence Hearing4All, director of the Department for medical physics and acoustics at the Universität Oldenburg, and scientific director of the Hörzentrum Oldenburg, HörTech gGmbH and Fraunhofer IDMT division for hearing, speech and audio technology. He supervised more than 55 Ph.D. theses and authored and coauthored more than 200 scientific papers in various areas of hearing research, speech processing, auditory neuroscience, and audiology. Prof. Kollmeier was awarded several scientific prizes, including the Alcatel-SEL research prize for technical communication, the International Award of the American Academy of Audiology and the German Presidents prize for Science and Innovation. He is Vice Chairman of the European Federation of Audiological Societies, past-president and board member of the German Audiological Society and advisory board member of the German Acoustical Society.

Bernd T. Meyer received the Ph.D. degree from the University of Oldenburg, Germany, in 2009. He was a Visiting Researcher in the speech group with the International Computer Science Institute, Berkeley, CA, USA, worked in the Center for Language and Speech Processing at the Johns Hopkins University, Baltimore, MD, USA, and is currently with the Medical Physics group at the University of Oldenburg, Oldenburg, Germany. His research interests include the relation of speech and hearing, with a special interest in models of human speech perception, automatic speech processing, and neurophysiological data.