Article:
Xiong, F., Goetze, S., Kollmeier, B. et al. (1 more author) (2019) Joint estimation of reverberation time and early-to-late reverberation ratio from single-channel speech signals. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27 (2). pp. 255-267. ISSN 2329-9290

https://doi.org/10.1109/taslp.2018.2877894

Joint Estimation of Reverberation Time and Early-to-Late Reverberation Ratio from Single-Channel Speech Signals

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

***Abstract*—The reverberation time (RT) and the early-to-late reverberation ratio (ELR) are two key parameters commonly used to characterize acoustic room environments. In contrast to conventional blind estimation methods that process the two parameters separately, we propose a model for joint estimation to predict the RT and the ELR simultaneously from single-channel speech signals from either fullband or subband frequency data, which is referred to as joint ROom Parameter Estimator (jROPE). An artificial neural network is employed to learn the mapping from acoustic observations to the RT and the ELR classes. Auditory-inspired acoustic features obtained by temporal modulation filtering of the speech time-frequency representations are used as input to the neural network. Based on an in-depth analysis of the dependency between the RT and the ELR, a two-dimensional (RT, ELR) distribution with constrained boundaries is derived, which is then exploited to evaluate four different configurations for jROPE. Experimental results show that — in comparison to the single-task ROPE system which individually estimates the RT or the ELR — jROPE provides improved results for both tasks in various reverberant and (diffuse) noisy environments. Among four proposed joint types, the one incorporating multi-task learning with shared input and hidden layers yields the best estimation accuracies on average. When encountering extreme reverberant conditions with RTs and ELRs lying beyond the derived (RT, ELR) distribution, the type considering RT and ELR as a joint parameter performs robust in particular. From state-of-the-art algorithms that were tested in the Acoustic Characterization of Environments (ACE) challenge, jROPE achieves comparable results among the best for all individual tasks (RT and ELR estimation from fullband and subband signals).**

***Index Terms*—Reverberation time, early-to-late reverberation ratio, joint estimation, temporal modulation features, multi-task learning.**

I. Introduction

FOR speech communication applications such as teleconferencing, automatic speech recognition (ASR) and speech enhancement in hearing aids, the room acoustics have a great impact on speech quality and speech intelligibility [1]–[5]. As the reverberation time (RT) and the early-to-late reverberation ratio (ELR) are two key parameters to characterize a room [6], [7], and usually it is not sufficient to rely only on either of them to fully represent the considered room reverberation effect due to various room sizes and source-to-receiver distances (cf. the REVERB challenge [5]), an accurate estimation of both measures is desirable for improving system monitoring [8]–[10] and system performance [11], [12] in reverberant environments. Traditionally, the RT and the ELR are derived from a corresponding room impulse response (RIR) between the source and the receiver, which however, needs to be intrusively measured. Such intrusive RIR measurement requires time and other resources, and is not always practical in real-world scenarios. It is, therefore, of great interest to blindly (or non-intrusively) estimate the RT and the ELR directly from reverberant speech signals.

Numerical approaches for blind RT and ELR estimation have been proposed in recent decades (cf. [13] for a detailed categorization): Most RT estimators rely on single-microphone recordings (e.g., [14]–[17]) to determine parameters related to the reverberation tail that defines the RT, i.e., the time interval derived from a 60 dB sound energy decay. ELR estimators on the other hand commonly exploit multi-microphone data for capturing cues that separate early and late components (e.g., [18]–[21]). ELR estimation includes two special cases, i.e., the direct-to-reverberant ratio (DRR) (which assumes that early components correspond to the direct path), and the clarity index [6] for which early and late reflections are separated at 50 ms or at 80 ms, which is denoted as C_{50} or C_{80}, respectively. So far, the two estimation tasks usually have been treated independently. However, the RT and the ELR share properties that motivated our research on a model that integrates both measures jointly: First, the knowledge of the late reverberation tail is important for both the RT and the ELR estimation. Second, many common features have been found to be correlated with the RT and the ELR, so that methods based on these features could be beneficial for joint estimation. For example, the low-frequency envelope spectrum has been used in [22] as input to a neural network to obtain different room acoustic parameters including the RT and the ELR. A comparison of energies at high and low modulation frequencies, the so-called speech-to-reverberation modulation energy ratio (SRMR), was found to correlate with the RT and the DRR [23]. A complex combined set of acoustic features proposed in [24] was employed as input of a classification

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).

Manuscript received April 23, 2018; revised August 26, 2018 and October 04, 2018; accepted October 19, 2018. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Federico Fontana. *(Corresponding author: Feifei Xiong.)*

and regression tree to estimate the clarity index C_{50}, and this feature set has been exploited to estimate the RT as well [25].

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) [7], our earlier work proposed a ROom Parameter Estimator (ROPE) [26] to formalize the blind estimation as a classification task, and also showed that auditory-inspired modulation features are well-suited for discriminating different RTs or ELRs in combination with a multi-layer perceptron (MLP). However, the ROPE approached modeled each measure separately, not taking into account potential complementary effects between RT and ELR as motivated above, which could be beneficial for improving the estimation of room parameters. To overcome this limitation of the previous model and to test beneficial effects of complementary modeling, we explore ROPE for the *joint* estimation of the RT and the ELR from single-microphone reverberant speech from fullband or subband frequency data in this paper, which is referred to as jROPE. Building upon the previous model, an MLP is used in jROPE as discriminative classifier to learn the mapping from acoustic features to discrete classes of room acoustic parameters. The auditory-inspired acoustic features used as MLP input are extracted by filtering time-frequency representations of speech [27] using a temporal modulation filter bank [28], [29]. In contrast to our previous single-task algorithms [26], the multi-task jROPE algorithm has the potential to model the dependency of RTs and ELRs, which might result in an improved prediction performance. Further, to the best of our knowledge, this is the first study for jointly estimating room parameters in the field of blind acoustic room parameter estimation.

The integration of information related to room acoustics can be performed at different stages of the model, and thus, the optimal integration strategy is an important research challenge in this context. In this paper, we explore four different system architectures for joint estimation. The first two (referred to as jROPE-I and jROPE-II) are motivated by an analysis of RT and ELR distributions for typical reverberant conditions, which shows the classes of the RT and the ELR to be mutually dependent. The first approach (jROPE-I) exploits this dependency by imposing constraints on the ELR classification based on the outcome of the RT classification. The second approach (jROPE-II) implicitly models the relation by using joint labels (RT, ELR) instead of separate RT and ELR classes, i.e., the joint class distribution is taken into account based on the training data with (RT, ELR) pairs. A prototype of jROPE-II has been introduced in our earlier work [30] but only for a limited training set without systematical analysis of RT-ELR dependency. jROPE-III and jROPE-IV implement multi-task learning (MTL), which provides solutions for solving multiple learning tasks at the same time [31]. MTL is applied to fuse two separated MLPs (one for RT and the other for ELR) into one. Different combination schemes in MTL are explored: For jROPE-III, the hidden MLP layers between the ELR and RT classifier are shared, while for jROPE-IV both the hidden and the input layers are shared.

In this remainder of this paper, we first briefly introduce the original ROPE model [26] (Section II). The relation of RTs and ELRs corresponding to a wide range of typical everyday acoustic scenarios is carried out (Section III) as a prerequisite for the joint estimation algorithm. The resulting (RT, ELR) distribution motivated different jROPE architectures that are introduced subsequently. In order to test the robustness of jROPE against different reverberant environments in noisy conditions for both fullband and subband frequency processing, we generate several training and testing sets for evaluation as described in Section IV. Further, the single-microphone evaluation database from the ACE challenge recorded in realistic room environments [13] is used to validate jROPE, and results are compared to the performance of the ROPE model for separate estimation, as well as to other state-of-the-art RT and ELR estimators as summarized in [32]. Section V reports and discusses the experimental results, before we conclude the paper in Section VI.

II. ROOM PARAMETER ESTIMATOR (ROPE)

ROPE is a data-driven approach to perform blind room parameter estimation by mapping auditory-inspired features to RT and ELR classes. As illustrated in Fig. 1, ROPE is comprised of four main processing steps that are briefly described in the following. For a detailed description, please refer to [26].

A. Reverberant and Noisy Speech Synthesis

Training data for the ROPE model is simulated, while it is later evaluated for both simulated and realistic scenarios. The following signal model

$$x[k] = s[k] * h[k] + \beta \cdot n[k] \tag{1}$$

is commonly applied for speech enhancement in reverberant and noisy environments, where $x[k]$, $s[k]$ and $n[k]$ represent the received microphone signal, anechoic speech and additive noises, respectively. $*$ denotes the convolution operation and β is the coefficient to adjust the signal-to-noise ratio (SNR) of the mixture according to ITU-T P.56 [33]. In effect, for a reliable modeling, the additive noise $n[k]$ associated with the reverberant speech should be recorded in the same room with the same microphone position, since $n[k]$ also contains the room reverberation but with different impulse responses from $h[k]$.

However, it is not practical to collect all the corresponding room noises (e.g., ambient or background noises) which match with the available measured RIR database. In order to simulate reverberant speech with diffuse noises that are characteristic for the same room, we exploit the late part $h_l[k]$ of the corresponding RIR $h[k]$, based on the assumption that the late part (diffuse information) is assumed to be uncorrelated to the early part (spatial information) of the RIR [6]. To separate early from late components, a threshold of 50 ms is typically chosen, which corresponds to time indices $k > \lceil f_s \cdot 50\text{ms} \rceil$ with the sampling frequency f_s. Consequentially, the synthesis model is given by

$$x[k] = s[k] * h[k] + \beta \cdot n[k] * h_l[k]\,, \tag{2}$$

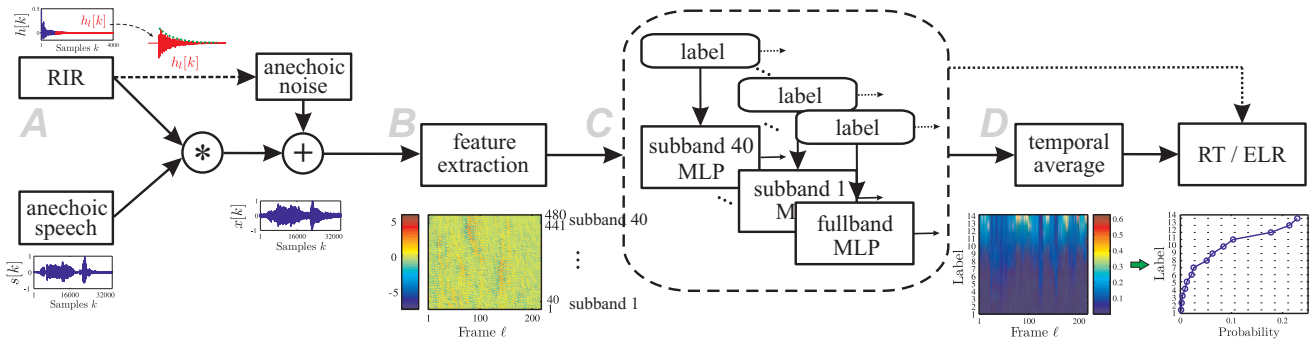

Fig. 1. System structure of ROPE to estimate the RT or the ELR. Fullband processing uses all frequency bands to generate the auditory-inspired features as MLP input, whereas subband processing at specific frequency band takes the corresponding frequency channel information for feature extraction. Steps A-D are briefly explained in Section II-A to II-D, respectively.

where $n[k] * h_l[k]$ represents the diffuse noise recorded from the same room as the target speech signal. In this notation, $n[k]$ represents an anechoic noise signal. Note that (2) does not cover the scenarios with localized noises, which are beyond the scope of the proposed model.

B. Auditory-Inspired Feature Extraction

Compared to the conventional temporal context splicing, auditory-inspired temporal modulation features have been shown to be more effective to extract temporal cues which are strongly related to RTs and ELRs [26]. These auditory-inspired features are calculated with a temporal modulation filter bank [28], [29] that uses time-frequency representations of speech signals as input [27]. First, a Gammatone filter bank decomposes the speech signal into frequency bands, and short-time windowing (length of 25 ms) is applied with a frame shift of 10 ms. A sequent logarithmic compression function is applied to the resulting two-dimensional time-frequency representations for dynamic range reduction. A temporal modulation filter bank with center frequencies at $\{0, 3, 6, 10, 17.1, 29.2, 50\}$ Hz (cf. [26]) is used to extract the temporal cues from speech time-frequency representations. It contains 7 real-valued filters and 5 imaginary filters, which are convolved with each frequency band, resulting in one 12-dimensional input vector (subband) per 10 ms time frame for the neural network, and are subsequently stacked if fullband processing (e.g., 40 bands with a sampling frequency of 16 kHz) is considered.

C. Neural Network Classifier

The RT or the ELR classification is performed using a multi-layer perception (MLP) that maps the input features to binned classes for each output parameter. The MLP is implemented using the Kaldi ASR toolkit [34] and uses building blocks typical for state-of-the-art ASR systems: Rectified linear units [35] are used as activation functions, and the standard back-propagation is applied to train the MLP via a stochastic gradient descent algorithm [36]. The cost function is based on cross-entropy, and a softmax function is applied to the output layer to obtain posterior probabilities of RT or ELR classes. One hidden layer is used and the dimensionality of the output layer corresponds to the number of RT or ELR output classes.

During training, labels for RT and ELR classes are required. To obtain these, we apply a nonlinear fitting [37] to the logarithmic magnitude of RIR $h[k]$, which is identical to the procedure used in the ACE challenge [13]. The ELR ground truth is calculated from $h[k]$ on a decibel scale with a division boundary of 50 ms that separates early and late components, which is motivated by the grouping effect of multi-path signal components in the human auditory system for delays below this time constant [2]. This is different from the DRR definition in the ACE challenge which used a shorter time constant as the division boundary, e.g., 2.5 ms as claimed in [13]. However, the true direct sound cannot be precisely determined when dealing with the measured RIRs, because the direct component depends on the source-to-microphone distance, source directivity, as well as the room structure and reflection factors [2] that are usually not provided together with the available measured RIR database. For subband analysis, the RIR $h_f[k]$ in frequency band f is decomposed by the Gammatone filter bank from the fullband representation $h[k]$ and subsequently processed analogously to the fullband counterpart.

D. Decision Strategy

Since the MLP generates one estimate per frame, and single-frame decisions are expected to be noisy, we smooth the classification result with a temporal averaging over all frames of the test utterance (*utterance-based* processing). The output neuron with the highest activation (or probability) corresponds to the RT or the ELR estimate (*winner-takes-all*). Although the results reported in this paper are based on utterance-wise processing, results from [26] show that results with *window-based* processing with a limited integration time should exhibit the same accuracy, e.g., 1.7 and 3 seconds are sufficient for fullband and subband estimation, respectively.

III. Joint ROom Parameter Estimator (jROPE)

In this section, we first analyze the relation of RTs and ELRs in detail. Based on this relation, four different system designs for a joint room parameter estimation derived from the original ROPE system are proposed and explored.

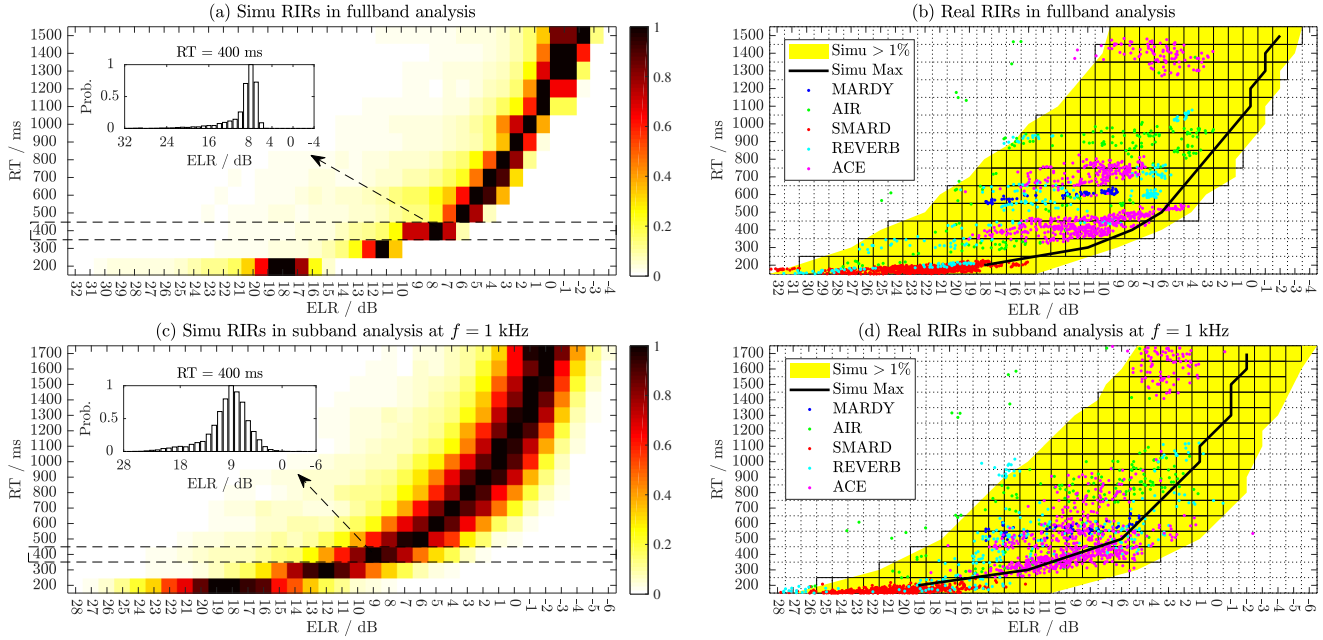

Fig. 2. Distribution of the RT and the ELR (fullband and subband analysis centered at $f = 1$ kHz). (a) and (c) show the fullband and the subband (RT, ELR) distributions based on 90000 simulated (Simu) RIRs (cf. Section IV-A) with row-wise normalization in each RT group, respectively. Panels (b) and (d) show distribution probabilities derived from (a) and (c) over 1% (yellow shading). The solid curves denote the positions of histogram peaks, and single data points refer to RIR parameters for measured data (Real) from the available open source databases including MARDY [38], AIR [39], SMARD [40], the REVERB challenge [5], and the ACE challenge [13].

A. Relation of Room Parameters

Fig. 2 (Panels (a) and (c)) shows the joint two-dimensional (RT, ELR) distribution in a numerical manner obtained from a large set of simulated RIRs (containing 90000 RIRs) which reflects typical acoustic conditions from everyday room scenarios with volumes from 30 to 1000 m^3 and fullband RTs ranging from 150 ms to 1550 ms. These RIRs are generated using the image method [41] [1], and uniformly distributed speaker-to-microphone distances are selected in each room in order to obtain a wide range of ELRs.

To determine the RT-ELR relation, we first group the data using a resolution of 100 ms (RT) and 1 dB (ELR), which is motivated by the just noticeable differences (JNDs) for the measures [7], [22], and is also later used for the classification task. Further, this resolution ensures a sufficient number of data points per tile of the grid in the RT-ELR space for probability analysis. On the other hand, due to the RIR simulation manner that is room-oriented, i.e., each room volume represents one RT with potentially uniformly distributed ELRs, it might not guarantee a uniformly distributed RTs within one specific ELR group. Therefore, we normalize the distribution for each RT group (row wise), by dividing all RIR samples of tiles by the maximum one (Max) in this RT group to represent the relative distribution probabilities, as illustrated by the embedded plots with RT value of 400 ms in Fig. 2 (a) and (c) as an example. The joint (RT, ELR) distribution is then formed by stacking all the RT groups, and the shape of this distribution illustrates the mutual dependency of both parameters, that is, ELRs vary within a limited range at a specific RT, and vice verse.

To quantify the area of this distribution (which is later used to constrain classification labels), RT-ELR pairs with a relative distribution probability below 1% (outliers as extreme reverberant conditions which rarely occur, cf. Section IV-B) are not considered in the following. The center points of the peripheral tiles are linearly connected to form the boundaries, and the resulting distribution is highlighted in Fig. 2 (b) and (d) for fullband and subband data, respectively. The width of the ELR distribution is relatively stable, while its center moves to lower ELRs with increasing RT. In effect, this is in line with the physical properties of RT-ELR relation, i.e., for small RTs, typically high ELRs are observed (cf. the solid curves with maximum probabilities). Outliers could arise from extremely small or large source-to-microphone distances. For instance, the source-to-microphone distances were chosen to be mouth-to-microphone distances within 15 cm for RIR recordings in the AIR database [39], [42] with scenarios of phone conversations (outlier green dots in Fig. 2 (b) and (d)), which is in contrast to most other databases considered in this study.

We refer to the highlighted region in Fig. 2 (b) and (d) as *core area* of (RT, ELR) pairs. The core area covers the (RT, ELR) pairs of measured RIRs from several open source databases [5], [13], [38]–[40] and therefore reflects typical scenarios, since these databases were also designed to simulate realistic settings.

[1]The RIR generator is implemented as detailed at *https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator*

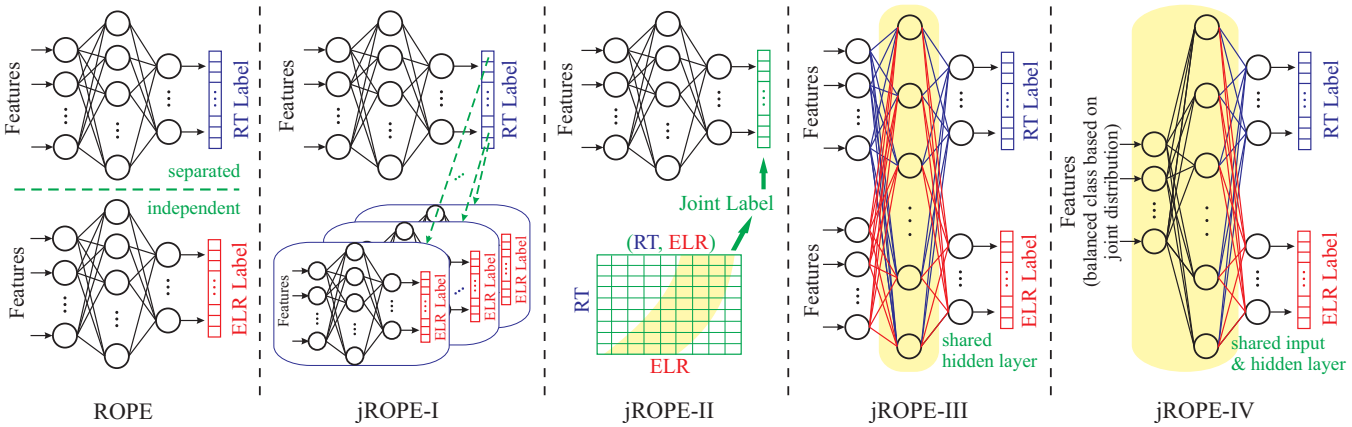

Fig. 3. Four different jROPE configurations in terms of the MLP topology and labeling (dashed box in Fig. 1) compared to the baseline ROPE system.

B. Design of Joint Estimators

As outlined in the introduction, a joint parameter estimation can be achieved by changing the MLP topology at different levels or by modifying the targets of the classifier. The original ROPE system (that serves as basis for jROPE) and four novel joint estimation approaches are illustrated in Fig. 3.

jROPE-I is based on a two-step procedure, i.e., the RT is estimated first, and ELR estimation is constrained by this first result. This differs from the ROPE system for ELR estimation, which takes a broad range of ELRs into account that range from -3 to 30 dB in fullband processing as shown in the ELR span of the core area in Fig. 2 (b). Instead, for jROPE-I, the number of classes is reduced (e.g., an RT of 400 ms results in an ELR class range 7 to 24 dB), which substantially reduces the ELR range, and thereby potentially reduces the standard deviation of ELR estimation errors. Through this approach, the training complexity increases (at a factor of the amount of RT labels) since each RT group requires its own ELR-MLP, the complexity during test does not increase (cf. Section V-E).

The second approach (jROPE-II) combines the RT and the ELR into a target parameter pair that is estimated in one step, i.e., the output neurons of the MLP directly map the input to a specific pair that can be visualized by a matrix spanned by RT and ELR classes (see inlay in Fig. 3, jROPE-II). In contrast to jROPE-I, which makes explicit use of a-priori knowledge of class distributions, the dependence of classes is learned implicitly in this case by providing the training labels that cover the core area as described above. This limits the amount of classes, which potentially improves the discrimination performance of the MLP.

Due to the mutual dependence of RT and ELR values, and because same signal processing chain resulted in good results when using the ROPE model, we explore multi-task learning (MTL) [31] for joint estimation (jROPE-III and IV). In MTL, several learning tasks are solved simultaneously using one classifier, e.g., a neural network that predicts two or more output measures. Potentially, the learning efficiency and prediction accuracy with MTL models can be improved over separate models by leveraging the domain-specific information contained in the training signals of related tasks.

The merging of neural net weights for implementing MTL for estimating room parameters can be performed on several levels; in this paper, two specific architectures are analyzed. For jROPE-III, we merge the hidden layers, while preserving the two (task-specific) output layers, as illustrated by the highlighted area of Model III in Fig. 3. Since RT and ELR can be estimated with the same features, we also explore the effect of shared input and hidden layer weights, as shown for Model IV in Fig. 3. Compared to jROPE-III, jROPE-IV further reduces the number of network parameters, which could improve the training efficiency. However, since the same input layer is used, but the RT and ELR label distribution in the core area is not symmetric as shown in Fig. 2 (b) and (d), an additional balancing of training data is required for jROPE-IV not only in terms of inter-labeling between the RT and the ELR tasks, but also in terms of intra-labeling in each task. For the single-task ROPE model, training data is balanced in terms of the output intra-labels to avoid overfitting for one specific class/label, which is generally achieved by generating the same amount of data for each class. Because of the independent input layer in jROPE-III, the two training sets with the same amount of data from ROPE-RT and ROPE-ELR can be directly merged to achieve the balance of the inter-labeling between the RT and the ELR task. The shared input layer in jROPE-IV ensures the data balance for the inter-labeling between these two tasks, which however causes an inherent imbalance for intra-labeling in either RT or ELR task because of the non-symmetric RT and ELR label distribution. In order to minimize the effect caused by such imbalance on the jROPE-IV training, we therefore adjust the training data from ROPE-RT with balanced data amount across RT labels (zero flatness) but imbalanced for ELR, until achieving nearly the same flatness across RT labels and ELR labels, though both flatness measures are not zero anymore.

The MTL approach with two classes as in our case typically uses a weighted-sum rule [31] for MLP training given by

$$\mathcal{L} = w\,\mathcal{L}_{\mathrm{RT}} + (1-w)\,\mathcal{L}_{\mathrm{ELR}}\,, \tag{3}$$

with the cross-entropy losses for the RT and the ELR estimation $\mathcal{L}_{\mathrm{RT}}$ and $\mathcal{L}_{\mathrm{ELR}}$, respectively. w denotes the weighting factor between both measures. Extreme measures of 1 and 0

TABLE I
PARAMETERS FOR MLP TRAINING OF ROPE AND JROPE SYSTEMS.

System	Task	Hidden Neurons	Label Amount	
			Fullband	$f = 1$ kHz
ROPE	RT	128	14	16
	ELR	256	34	31
jROPE-I	RT $\rightarrow$	128	14	16
	ELR	128	13−18	10−16
jROPE-II	(RT, ELR)	512	217	219
jROPE-III	(RT, ELR)	256	(14, 34)	(16, 31)
jROPE-IV	(RT, ELR)	256	(14, 34)	(16, 31)

result in a system that is functionally identical to single-task ROPE estimation for RT and ELR estimation, respectively.

IV. EXPERIMENTAL SETUP

A. Training Sets

Anechoic speech signals from the TIMIT corpus [43] are used as basis to generate the training set for the proposed data-driven approach (cf. Fig. 1 and Equation (2)). TIMIT contains recordings of phonetically-balanced prompted English speech in 3696 sentences (3.14 hours) from 462 speakers. Simulated RIRs (cf. Section III-A) are used, from which 10 different RIR samples for each RT and ELR class are selected to generate the training set. The sampling rate f_s is chosen as 16 kHz, and the Gammatone filter bank is accordingly characterized with 40 frequency bands with center frequencies starting from 100 Hz to 7943 Hz. For fullband frequency data, the range of RTs is $[200, 1500]$ ms with a 100 ms resolution, and the range of of ELRs is $[-3, 30]$ dB with a 1 dB resolution. For subband processing e.g., at $f = 1$ kHz, the ranges are $[200, 1700]$ ms and $[-5, 25]$ dB. The range of the ELR label for each specific RT is determined according to the highlighted core area (solid boxes in Fig. 2 (b) and (d)). In [26] performance of all the frequency channels have been investigated, and in this paper we focus on subband processing at typical $f = 1$ kHz for the sake of simplicity and better readability.

Three different SNRs ($\{20, 10, 0\}$ dB) and noise-free utterances (referred to as SNR $= \infty$ dB) are chosen for the training data. Two types of noise signals are chosen, namely pink noise (PN) which exhibits similar noise energy in each Gammatone frequency band, and babble noise (BN) which is generated using a mixture of anechoic speech signals produced by 4 female and 4 male speakers from the WSJCAM0 corpus [44]. The temporal modulation feature vectors (cf. Section II-B) have a dimensionality of 480 for fullband data; vectors for subband data are 12-dimensional.

Parameters for MLP training of different architectures in Fig. 3 are summarized in Table I. The amount of the MLP hidden nodes is determined based on pilot experiments to avoid overfitting of the MLP. For a fair comparison to ROPE, the number of computational operations during MLP training for both tasks of classifying RTs and ELRs is kept the same for the jROPE systems. The weighting factor w in (3) is initially set to 0.5 for training with MTL which assumes the RT and the ELR estimation to be equally important. A detailed analysis of the influence of the weighting factor is presented in Section V-D.

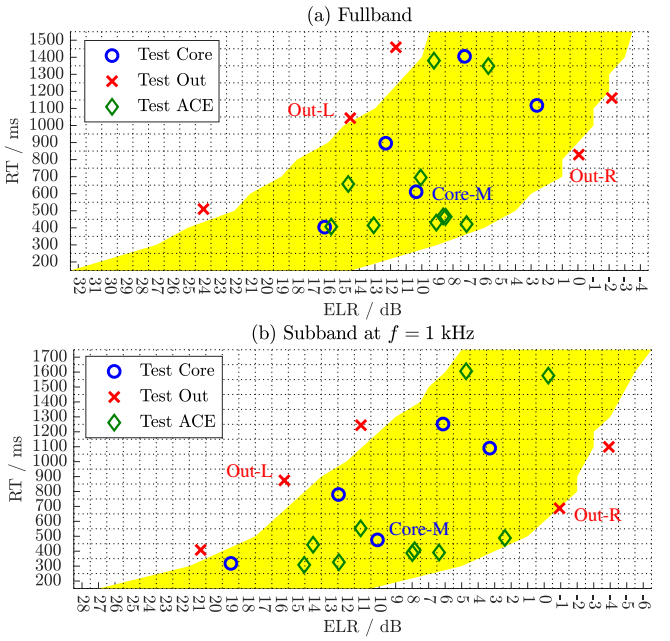

Fig. 4. Three test sets for system evaluation in various reverberant environments. *Test Core* contains 5 simulated RIRs with (RT, ELR) pairs inside the core area (cf. Fig. 2), while parameters of 5 RIRs lie outside (*Test Out*). *Test ACE* refers to the single-microphone evaluation test set from the ACE challenge with 10 measured RIRs [13].

B. Test Sets

In order to evaluate jROPE and to test its generalization to various acoustic environments, we generate 3 different test sets that include different (RT, ELR) distributions, noise types, SNRs, and sources of speech signals. Specifically, the first two data sets are created with the same procedure as the training data, but using different RIRs, speech signals, and SNRs. We select 10 simulated RIRs, 5 of which are within the (RT, ELR) core area, while the parameters of the remaining 5 are outside the core area for both fullband and subband data. The resulting sets are referred to as *Test Core* and *Test Out*, respectively. *Test Core* represents the common reverberant conditions, e.g., RIR data point 'Core-M'. *Test Out* represents more extreme reverberant scenarios that are not covered by the core distribution, i.e., data points with a very high (e.g., 'Out-L') or low ELR (e.g., 'Out-R') given a specific RT. The corresponding data points for both test sets are shown in Fig. 4. The speech signals for these two test sets are obtained from TIMIT evaluation test set, which contains 1344 different utterances (1.15 hours) from 168 speakers. Pink and babble noises are added according to (2) with the SNRs $\{30, 18, 12, -1\}$ dB that are different from training SNRs.

Further, the evaluation test set for single-channel processing from the ACE challenge [13] (*Test ACE*) is used, which is different from the training set with respect to RIR types (simulated versus measured), noise types (synthesized versus measured), and speech materials (by using different speech corpora). The RIRs were measured in 5 different rooms with 2 different microphone positions; the corresponding (RT, ELR) distributions are shown in Fig. 4. Three different SNR conditions are considered, which are referred to as low (−1 dB),

medium (12 dB), and high (18 dB). The anechoic speech signals contain 50 utterances (0.27 hours) produced by 5 male and 5 female speakers in different dialects of international English with a mix of native and non-native English speakers. The reverberant and noisy speech signal for evaluation was then synthesized following Equation (1). In summary, *Test ACE* contains 4500 utterances categorized by 10 RIRs, 3 noise types and 3 SNRs.

C. Evaluation Metrics

1) Estimation Errors: The estimation error is defined as the difference between the estimated value and the ground truth:

$$e_X = \hat{X} - X\,, \tag{4}$$

with X denoting either the RT or the ELR. When analyzing N measurement samples, the root mean squared error (RMSE) of e_X is reported:

$$\mathrm{RMSE}_X = \sqrt{\frac{1}{N}\sum_{n=1}^{N} e_{X,n}^2}\,. \tag{5}$$

Additionally, box plots illustrate the underlying distribution of e_X, where the central mark denotes the median error, the edges are the 25th and 75th percentiles, the whiskers show extreme values, and outliers are plotted individually.

2) Pearson correlation coefficient ρ: 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 an additional measure to quantify the estimation accuracy (as proposed in the ACE challenge), i.e., the Pearson correlation coefficient ρ for estimated and true parameters. It is defined by

$$\rho_X = \frac{E\{\hat{X}\cdot X\} - E\{\hat{X}\}\cdot E\{X\}}{\sqrt{\left(E\{\hat{X}^2\} - E\{\hat{X}\}^2\right)\cdot\left(E\{X^2\} - E\{X\}^2\right)}}\,, \tag{6}$$

where $E\{\cdot\}$ is the expectation value from N measurement samples.

3) Computational Complexity: The number of operations commonly expressed using big-O notation $\mathcal{O}(f(N))$ as a function of the input size N [45], is analyzed for quantifying computational complexity, which is used to indicate whether the proposed algorithm has the potential for practical real-time applications that are constrained in computational complexity such as hearing aids or the front-end speech processors in mobile devices.

V. Results and Discussion

A. General Comparison between ROPE and jROPE

First, the Set *Test Core* is used to compare the original single-task ROPE algorithm and four types of jROPE. As illustrated in Fig. 5, in fullband analysis, estimation errors e_{RT} and e_{ELR} are within ± 150 ms and ± 1.5 dB with median values close to 0 (left y-axis) for all the proposed algorithms, respectively, indicating that ROPE and jROPE systems provide accurate RT and ELR estimations for this test set. This is

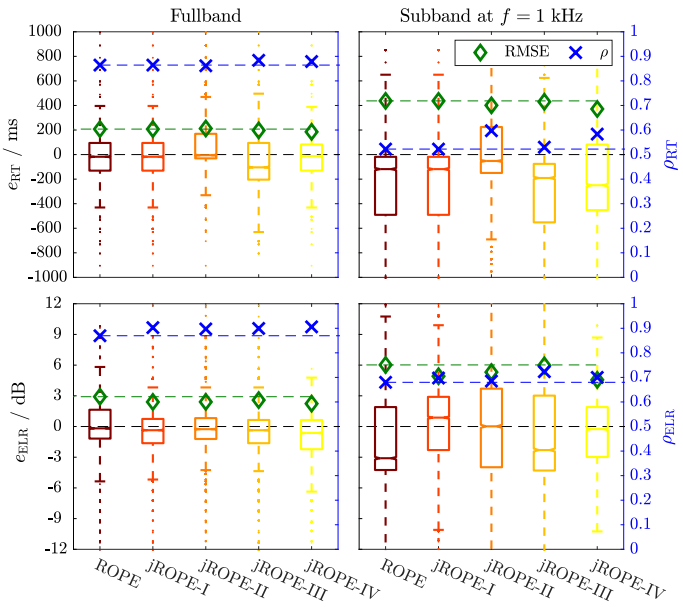

Fig. 5. ROPE and jROPE performance for estimating the RT and the ELR from fullband and subband data at $f = 1$ kHz analysis with Set *Test Core*. The estimation error, RMSE, and ρ were obtained from 1344×5 test utterances (with pink and babble noise at SNRs of $\{30, 18, 12, -1\}$ dB). For RMSE (left y-axis), lower is better, while for the correlation (right y-axis) higher is better. Dashed lines indicate the single-task ROPE performance.

in line with the obtained high correlation coefficients, since both ρ_{RT} and ρ_{ELR} are larger than 0.85 (right y-axis). In comparison, the performance of subband estimations degrades, probably due to the limited speech information within one specific frequency band: Average $\mathrm{RMSE}_{\mathrm{RT}}$ and $\mathrm{RMSE}_{\mathrm{ELR}}$ across all algorithms increase nearly twice, i.e., from 200 ms to 400 ms, and from 3 dB to 6 dB, respectively.

Compared to the baseline ROPE system that estimates RT and ELR separately, jROPE generally provides better results, indicating that joint estimation can further improve individual tasks. More specifically, jROPE-I further reduces the standard deviations of e_{ELR} for both fullband and subband ELR estimations, as shown by the boxplots of e_{ELR} in lower panels from Fig. 5. The underestimation of ELR values (with negative median e_{ELR}, particularly for subband data) obtained with ROPE are greatly mitigated when using jROPE-I. This might be due to the constrained range of the MLP labeling for ELR in jROPE-I, rather than the whole spanning of the ELR distribution used in ROPE (cf. Section III-B).

Results for jROPE-II shows that consistent improvements for RT and ELR estimations in both fullband and subband analysis are obtained compared the original ROPE system, i.e., the approach of considering the two key parameters as a combined two-domain parameter (RT, ELR) pair seems to be beneficial. This also indicates that the discriminative cues between different (RT, ELR) pairs can be captured well by the auditory-inspired temporal modulation features (cf. Section II-B) although the dimension of the MLP output label increases by a factor of ten (on average) in comparison to ROPE (cf. Table I).

The Improvement obtained with jROPE-III (for which the

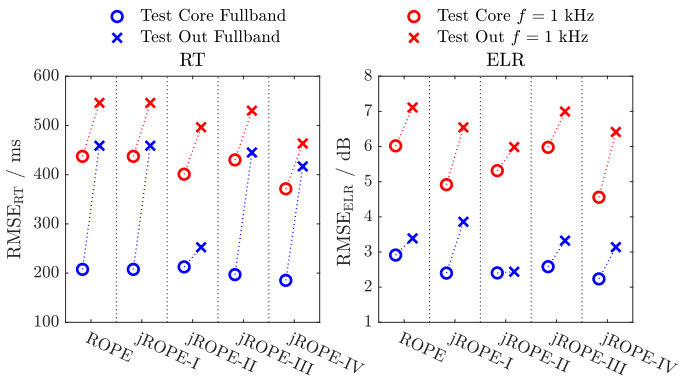

Fig. 6. Performance comparison between using *Test Core* and *Test Out* for ROPE and jROPE systems in terms of $\mathrm{RMSE_{RT}}$ and $\mathrm{RMSE_{ELR}}$ (each set with 1344×5 test utterances containing pink and babble noises at SNRs of $\{30, 18, 12, -1\}$ dB).

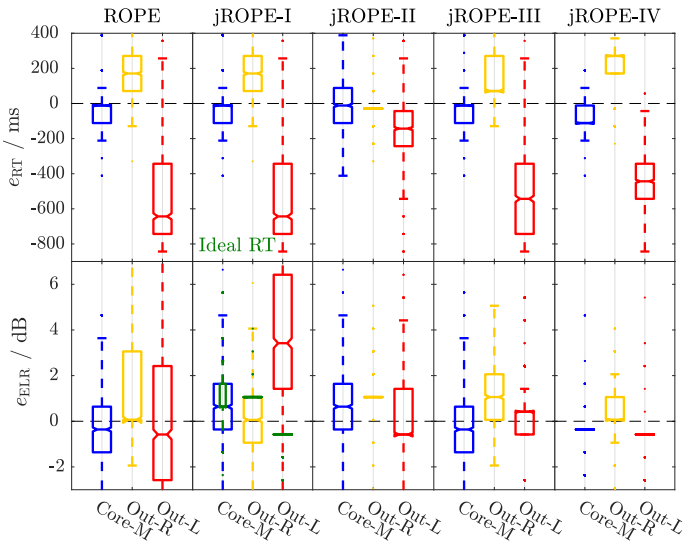

Fig. 7. ROPE and jROPE performance for estimating the RT and the ELR in fullband analysis with three specific test reverberant conditions (each with 1344 test utterances containing pink and babble noise at SNRs of $\{30, 18, 12, -1\}$ dB), where 'Core-M' belongs to Set *Test Core*, and 'Out-R' and 'Out-L' correspond to Set *Test Out*, as illustrated in Fig. 4. In the panel of jROPE-I, the error e_{ELR} obtained with ideal RT estimation (*oracle*) is plotted as well.

only difference compared to ROPE is the joint training with a shared hidden layer) shows that multi-task learning can further improve the discrimination by leveraging the mutual relations of the RT and the ELR via a shared hidden layer. This slight improvement, on the other hand, indicates that shared hidden layer is not sufficient to fully exploit the RT-ELR relation due to their inherent relation starting from the data level. jROPE-IV, which additionally shares the input layer, shows further improvement for both estimation tasks, indicating that more shared information for complementary measures seems to be beneficial for MTL in general to further improve task performance.

B. Impact of (RT, ELR) Mismatch

The Set *Test Out* measures the generalization of ROPE and jROPE systems when encountering extreme reverberant conditions with (RT, ELR) pairs outside the core area not seen during training. Fig. 6 compares the performance for such set to results for Set *Test Core* in terms of the RMSE and shows that the (RT, ELR) mismatch results in a performance degradation for all proposed algorithms. On average, $\mathrm{RMSE_{RT}}$ and $\mathrm{RMSE_{ELR}}$ increase approximately by 100 ms or 1 dB, respectively. The degradation is most noticeable for fullband RT estimation, with an RMSE increase over 200 ms. The results obtained with jROPE-II are an exception to this, with relatively stable results for *Test Out*. For instance, $\mathrm{RMSE_{RT}}$ and $\mathrm{RMSE_{ELR}}$ increase only by 40 ms and 0.03 dB in fullband analysis, respectively.

On the other hand, other jROPE types show similar trends of the performance degradation in comparison to ROPE, but still perform better than ROPE in general with Set *Test Out*. It is also interesting to notice that jROPE-I performance for the ELR estimation decreases more severely from *Test Core* to *Test Out* in comparison to other algorithms. Particularly in fullband analysis, the $\mathrm{RMSE_{ELR}}$ is even higher than the RMSE obtained with ROPE. In order to investigate how (RT, ELR) mismatches affect the proposed systems, an in-depth analysis of the estimation errors is carried out in the following section.

C. Estimation Error Analysis

To pinpoint the estimation errors in detail for different systems, we select three specific reverberant conditions for analysis, namely 'Core-M' from *Test Core*, 'Out-L' and 'Out-R' from *Test Out* (cf. Fig. 4). Fig. 7 shows the estimation performance in terms of e_{RT} and e_{ELR} in fullband processing, and the results with subband data are not shown since they follow the same performance trend.

In agreement with results from Section V-A, all algorithms perform well in condition 'Core-M' sampled from the core distribution, and jROPE-IV performs particularly good for the ELR estimation. For condition 'Out-R', a slight overestimation with median errors around 200 ms for RT estimation is observed, with the exception of jROPE-II that produces mostly correct RT estimates (median errors near 0 ms) with few outliers. With ROPE, the error of ELR estimation exhibits a large standard deviation, as well as the potential overestimation. This standard deviation is smaller for the joint estimation algorithms, with e_{ELR} in the range of ± 1 dB. jROPE-II provides a particularly small standard deviation but with a median error slightly larger than 1 dB. Small ELR overestimation errors also occur for jROPE-III, whereas jROPE-IV performs better and yields a median error of almost 0.

In contrast to this, ROPE performance severely degrades for condition 'Out-L' with strong underestimates of RT (median error: -600 ms, ground truth: 1043 ms). Although the median error of the ELR estimation is still close to 0 dB, the standard deviation is almost twice as large compared to 'Out-R'. 'Out-L' corresponds to a condition with a large room volume but with a rather short speaker-to-microphone distance, which is reflected by a long reverberation tail but also a high energy of the early reflections. For ROPE, such reverberant cases are

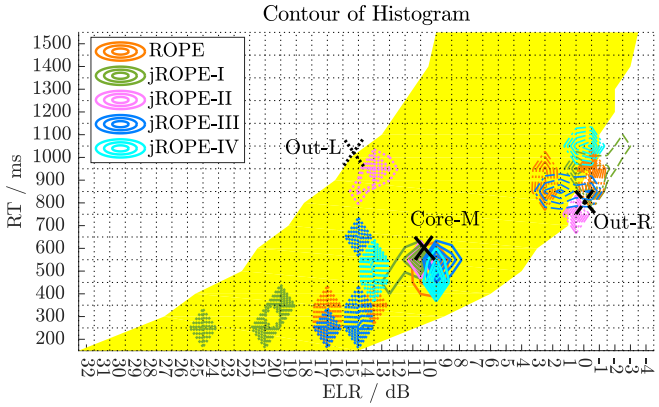

Fig. 8. Contour of the histograms of the estimated RTs and ELRs (cf. Fig. 7) from ROPE and jROPE systems with the three chosen reverberant conditions (also marked in Fig. 4), i.e. 'Core-M' (solid lines), 'Out-R' (dash-dotted lines) and 'Out-L' (dotted lines).

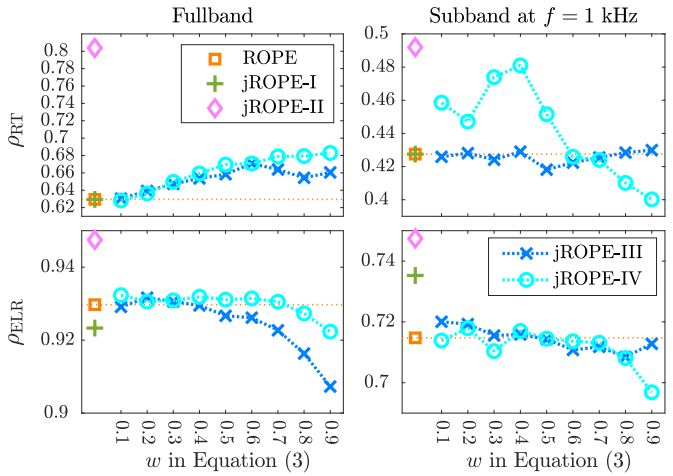

Fig. 9. Overall performance comparison in terms of correlation values ρ for ROPE and jROPE systems for both fullband and subband analysis. Horizontal dashed lines correspond to ROPE baseline results. The effect of the RT vs. ELR weighting factor w in Equation (3) during MTL training on the performance of jROPE-III and jROPE-IV is illustrated as well.

mapped to known class distributions with the lower RTs. The underestimated RT values are used in jROPE-I (first step) and result in a severe ELR overestimation (second step) due to the constraint ELR output values. The impact of RT errors on jROPE-I on the estimation of ELRs is quantified by using oracle knowledge, i.e., the RT estimate is replaced with the RT ground truth. Results with RT oracle knowledge are shown in Fig. 7 for jROPE-I (green box plot): Estimated ELRs are mapped to the nearest ELR class seen during training for current RT values, e.g., 14 dB at 1000 ms for 'Out-L'. As for condition 'Out-R', jROPE-II is robust to the mismatch introduced by 'Out-L' with a small performance degradation due to a higher standard deviation. While the RT estimation with jROPE-III and jROPE-IV is slightly improved compared to ROPE, the ELR estimates are much more accurate, particularly in the case of jROPE-IV.

To better illustrate whereto the estimated RTs and ELRs distribute in the two-dimensional (RT, ELR) area, histograms of the estimated results with the chosen three conditions are diagrammed, as shown in Fig. 8. Generally, when room parameters are within the core area (e.g., 'Core-M'), or on the right side (e.g., 'Out-R'), all models can provide RT and ELR estimates within ± 200 ms and ± 2 dB around the ground truth, respectively, albeit different properties: Compared to ROPE, jROPE-I provides ELR estimates with a smaller standard deviation but slightly more outliers. jROPE-II estimates towards the nearest (RT, ELR) class with a slight RT underestimation. jROPE-III performs slightly better than jROPE-IV for RT estimation, whereas jROPE-IV outperforms jROPE-III for ELR estimation. If parameters locates on the left side of the core distribution such as 'Out-L', jROPE-II performs consistently robust, while errors become notable by other models particularly for RT estimation, though jROPE-IV yields the best ELR estimates.

D. Overall Performance

Fig. 9 shows the average performance of the proposed algorithms in terms of correlation values ρ averaged over both sets *Test Core* and *Test Out*. The figure also presents the effect of the weighting factor w for trading off the importance of RT and ELR classification (cf. Equation (3) during MTL training) both for jROPE-III and jROPE-IV. The weighting factor is increased from 0.1 to 0.9 with 0.1 step size. Note that jROPE-III and jROPE-IV become equivalent to ROPE for single-task RT estimation when $w = 1$ or ELR estimation for $w = 0$. In general, as w increases, ρ_{RT} increases while ρ_{ELR} decreases, as shown in Fig. 9 particularly in fullband analysis (left panel). On the other hand, performance is not extremely sensitive to w, and it seems that $w = 0.4$ is a good choice for achieving good classification performance for both tasks in both fullband and subband analysis.

In comparison to ROPE, ELR estimation performance of jROPE-I is degraded in some cases which is attributed to a potentially inaccurate RT estimation from the first step. The best average results (considering RT and ELR classification, as well as fullband and subband data) are obtained with jROPE-II, which is very robust against mismatches between training and test conditions. On average, jROPE-IV outperforms jROPE-III, and both of them perform better than ROPE. Further, for subband RT estimation, jROPE-IV provides results comparable to jROPE-II. Overall, jROPE-II with joint label and jROPE-IV with MTL sharing both input and hidden layer seem to make good use of the RT-ELR relation for joint estimation in both fullband and subband analysis.

E. Complexity Analysis in Test Stage

Since the proposed models are blind estimators of room parameters, they could potentially be used in hearing devices, e.g., for speech enhancement or dereverberation algorithms that require RT and/or ELR as input. The computational complexity is an important factor for such mobile application scenarios, which is quantified in this work by analyzing the number of operations expressed by big-O notation. The complexity of ROPE/jROPE during test is mainly due to the

calculations of feature extraction (cf. Section II-B) and MLP forward processing (cf. Section II-C).

In the feature extraction stage (only calculated once for two tasks), the time-frequency representations were calculated using the Gammatone filter bank to filter the time-domain speech signal which was implemented by overlap-add method using fast Fourier transform (FFT) [46], resulting in $\mathcal{O}(N \cdot \log(N))$ per each frequency channel with N input samples. The convolution of the temporal modulation filter bank and the time-frequency representations was implemented via multiplication in the FFT domain, resulting in $\mathcal{O}(L \cdot \log(L))$ per each modulation filter with L input frames. The complexity of a forward-run for 1-hidden-layer MLP with I input neurons, H hidden neurons and O output neurons can be estimated as $\mathcal{O}(L \cdot (I + O) \cdot H)$ with L input frames. Among the different system architectures explored in this study, jROPE-II exhibits the highest computational complexity, while the lowest complexity is achieved with jROPE-IV. When using 1 second speech signals as an example, the numbers of multiplications for ROPE, jROPE-II and jROPE-IV for joint RT and ELR estimations in fullband processing are approximately 2.76, 4.36 and 2.17 ($\times 10^7$), respectively. It is also worthwhile noting that current GPU usage for numerical computing can efficiently handle dense matrix-matrix multiplications in neural networks.

F. Performance for the ACE Challenge Evaluation Database

In order to test ROPE and jROPE in realistic recording environments, we use the single-microphone evaluation database from the ACE challenge [13] (Set *Test ACE* in Fig. 4).

1) RT and ELR Estimation: For the RT estimation from fullband and 1 kHz-subband data (cf. upper panels of Fig. 10), performance increases with the SNR, and correlations above 0.75 and median errors close to 0 ms are obtained by all algorithms 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 [13] and therefore covers properties of spatially-localized maskers. In contrast, the babble noise for training is completely diffuse (cf. Equation (2)) and contains speech from 8 different talkers (cf. Section IV-A). Performance is degraded in the presence of fan noise (especially at -1 dB), which could arise from its totally different noise characteristics: It was generated by one or two fans near the corners of the recording rooms [13] and performs as localized noise rather than diffuse noise (cf. Equation (2)). On the other hand, jROPE produces good results in ambient noise that is also not seen during training, which hints at the generalization capabilities of the proposed data-driven approaches if training and testing noise types share similarities (pink noise usually serves to minic ambient noise). Among all tested algorithms, it seems that jROPE-IV performs the best particularly for the fullband RT estimation, while jROPE-II works slightly worse than others. Although jROPE-II is robust against the (RT, ELR) mismatch, it performs slightly worse than others when all test reverberant conditions

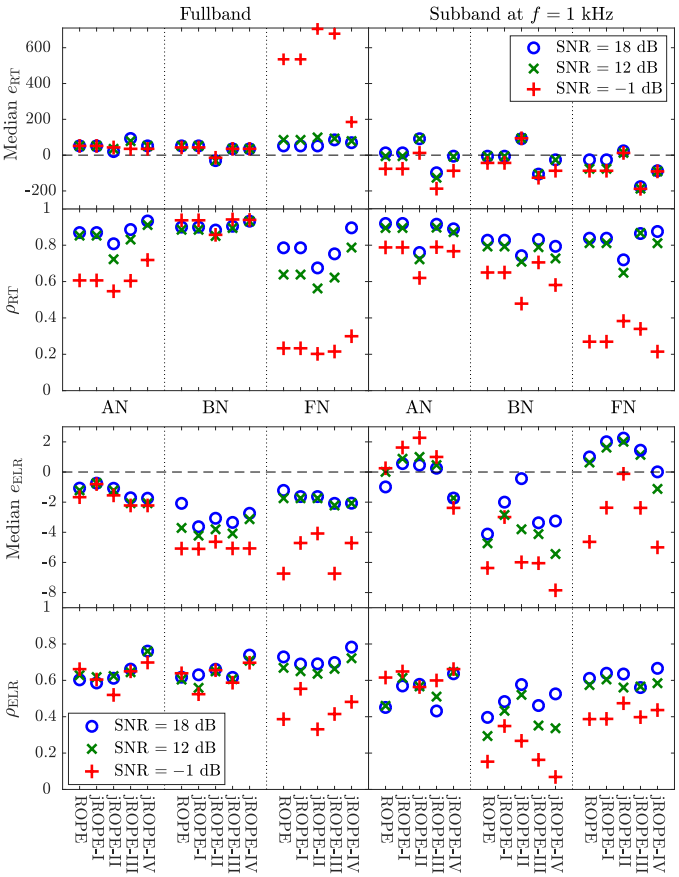

Fig. 10. ROPE and jROPE performance for estimating the RT and the ELR in fullband and subband at $f = 1$ kHz analysis with *Test ACE*, i.e., the single-microphone evaluation database from the ACE challenge [13]. Median errors of e_{RT} and e_{ELR}, as well as the correlation values ρ_{RT} and ρ_{ELR} (RMSE values follow the same trend) are illustrated in terms of ambient (AN), babble (BN) and fan noise (FN) at SNRs of $\{18, 12, -1\}$ dB, each with 500 test utterances (2.7 hours).

(in *Test ACE*) are inside the core distribution (cf. Fig. 4), which is line with the findings in Section V-A.

Similar trends can be observed for the ELR estimation as illustrated in Fig. 10 (lower panels) with the exception of babble noise, for which ELR is underestimated (median errors below 0 dB). We assume this is caused by the spatial component of the babble test data, as described above. This is supported by the fact that spatial source positions influence the ELR, while the RT is mostly invariant to them: Since the proposed models are tailored to speech processing, ELR estimates are influenced by the babble noise from the ACE testing set that includes spatial components associated with masking speech. Masking speakers are usually farther away from the microphone than the original target speaker, which would result in an underestimation of ELR, especially at low SNRs.

2) Performance Comparison: Results from ROPE and jROPE systems are compared to other single-microphone state-of-the-art RT and ELR estimators that were tested on data provided through the ACE challenge (cf. [13], [32] for detailed descriptions of these algorithms). Since these algorithms were implemented and specifically tailored to the ACE challenge by their respective authors and currently are not freely accessible,

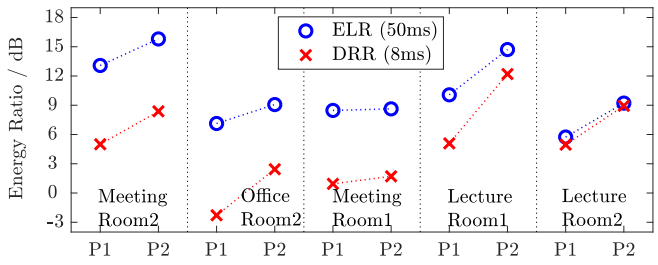

Fig. 11. Difference between fullband ELRs and DRRs in Set *Test ACE*, including 10 RIR samples recorded with two positions (P1 and P2) in 5 rooms. These rooms are sorted with increasing fullband RTs [13].

a comparison w.r.t. the computational complexity cannot be presented in this paper, although the ACE challenge primarily used real-time factor as an indication of time complexity (which cannot be fairly compared across different hardware).

Note that the ACE challenge focused on the DRR estimation with the assumption that the division boundary between the direct path and the reverberant part equals 8 ms [2], which is different from the 50 ms used throughout this paper (cf. Section II-C). The value differences between ELRs and DRRs in Set *Test ACE* are not consistent per measured RIR, as shown in Fig. 11, indicating that the performance between ELR and DRR estimators in terms of RMSE cannot be directly comparable. On the other hand, due to the similar trend of ELR/DRR varying in an almost same fixed range, a meaningful comparison can be obtained through the correlation coefficients ρ which reflects the relative relation between ground truth and the estimate, showing how good the energy ratio estimator works. Since a blind algorithm for subband DRR or ELR estimation from single-channel data was not proposed before, we exploit an algorithm based on particle velocity (ParVal) [47] as a baseline, although it was developed for multi-channel data (a spherical microphone array with 32 microphones).

As shown in Table II, jROPE-IV achieves competitive performance in terms of $\mathrm{RMSE}_{\mathrm{RT}}$ and ρ_{RT} when compared to the best result for single-channel RT estimation, i.e., QAReverb [17], [48]. Other jROPE types also provide comparable results, but perform slightly worse than ROPE. For fullband ELR estimation, jROPE-II and jROPE-IV produce a slightly better correlation ρ_{ELR} compared to the best ACE challenge contribution for single-channel data (Non-Intrusive Room Acoustic (NIRA) estimator [24], [25]). Improved performance is also achieved when comparing jROPE-II with a preceding model version jROPE-ACE [30] (which was introduced during the ACE challenge) which is attributed to more reliable data labels for ELR with 50 ms in comparison to a DRR with a fixed time constant of 8 ms which is not accurate in all cases since it does not represent the direct component for all RIRs under consideration. When compared to the subband maximum-likelihood RT estimator (ML-RTE) [16], [49], slightly better performance is achieved by all proposed algorithms with the exception of subband analysis at $f = 1$ kHz with jROPE-II.

[2]Note that, in [13] (after the competition), this value was claimed to be 2.5 ms with the same evaluation data. However, for the final competition results, the DRR ground truth [30] specified by the ACE challenge was 8 ms.

TABLE II
PERFORMANCE COMPARISON WITH OTHER SINGLE-MICROPHONE STATE-OF-THE-ART RT AND ELR ESTIMATORS BASED ON THE ACE CHALLENGE EVALUATION DATABASE. RESULTS WITH GRAY TEXT ARE NOT DIRECTLY COMPARABLE. † DENOTES THE MULTI-MICROPHONE CONFIGURATION.

Fullband Estimation

Estimator	RMSE/ms	ρ_{RT}	RMSE/dB	ρ_{ELR}
QAReverb [48]	255	**0.778**	4.86	0.058
NIRA [25]	389	0.302	3.85	0.558
SRMR [50]	380	0.220	5.82	–0.084
jROPE-ACE [30]	322	0.480	3.99	0.405
ROPE [26]	285	0.716	4.81	0.556
jROPE-I	285	0.716	5.01	0.524
jROPE-II	327	0.685	4.70	0.562
jROPE-III	316	0.696	4.88	0.556
jROPE-IV	288	0.758	4.09	**0.621**

Subband Estimation at $f = 1$ kHz

Estimator	RMSE/ms	ρ_{RT}	RMSE/dB	ρ_{ELR}
ML-RTE [49]	358	0.699	-	-
ParVel† [47]	-	-	3.21	0.415
ROPE [26]	338	0.751	7.63	0.421
jROPE-I	338	0.751	5.54	0.495
jROPE-II	389	0.693	5.61	**0.512**
jROPE-III	351	**0.776**	7.53	0.430
jROPE-IV	377	0.705	5.81	0.440

Compared to the the multi-microphone ParVal method [47] in subband analysis, all proposed algorithms show better ρ_{ELR}, where jROPE-II shows the best. Also, jROPE models can further reduce $\mathrm{RMSE}_{\mathrm{ELR}}$ in comparison to ROPE.

As summarized in Table II, most algorithms perform well for one specific task, but strongly degrade for other tasks (or are not applicable at all). For the single-channel subband task in the ACE challenge, only one algorithm was proposed for RT estimation and none for ELR estimation. In contrast, our proposed algorithms provide reliable results for both the RT and the ELR estimation in fullband and subband processing. Compared to ROPE, jROPE further improves the estimation accuracies and jROPE-IV emerges as the best system on average.

TABLE III
STATISTICAL DIFFERENCES USING PAIRED-SAMPLE T-TEST IN TERMS OF PERFORMANCES FROM ALGORITHMS IN TABLE II. ROPE MODEL IS CHOSEN AS THE REFERENCE FOR PAIR COMPARISON, AND THE SIGNIFICANCE LEVEL IS CHOSEN AS 0.01. MODELS THAT SHARE SIMILARITY TO ROPE ARE LISTED AND THE p-VALUES OF OTHER ALGORITHMS ARE BELOW 10^{-8}.

Models	p-value
jROPE-II fullband RT	0.613
jROPE-III fullband RT	0.610
jROPE-I fullband ELR	0.012
Others	$< 10^{-8}$

Furthermore, statistical tests of the employed algorithms show that almost all the algorithms perform significantly different than ROPE. The results for jROPE-II and jROPE-III for RT estimation in fullband analysis are not statistically different, which presumably arises from the very similar training data (which is different from the class-balanced data for jROPE-IV).

VI. Conclusions

This paper presented a system for blind estimation of two key room parameters from single-channel speech data in fullband and subband processing, i.e., the reverberation time (RT) and the early-to-late reverberation ratio (ELR). These two parameters were estimated jointly, and we refer to the resulting ROom Parameter Estimator (ROPE) as jROPE model. We first defined a core (RT, ELR) distribution that represents parameters of typical reverberant conditions. Four jROPE architectures were proposed that differ with respect to the integration stage of RT and ELR, and were compared to a related modeling approach that performs a separate estimation of room parameters (ROPE model). Results show that an improved estimation is achieved with joint estimation either by putting explicit constraints on the estimation value or by relying on multi-task learning to implicitly exploit the mutual relation of RT and ELR by shared network parameters. For unseen room parameters covered by the core distribution, the best prediction performance is obtained with multi-task learning that shares both input and hidden layers of a multi-layer perceptron (jROPE-IV). In the presence of extreme (RT, ELR) pairs that are not covered by the core distribution, the approach of explicit pair classification (jROPE-II) performs better than other joint approaches. The jROPE models were benchmarked against state-of-the-art models proposed for the Acoustic Characterization of Environments (ACE) challenge, and provided further improvements compared to the single-task ROPE system, which has already achieved comparable results with the best estimators of the competition for each individual tasks.

Acknowledgment

This research was funded by the Cluster of Excellence 1077/1 "Hearing4all". The authors would like to thank James Eaton for the data release of the ACE challenge.

References

[1] M. Wölfel and J. McDonough, *Distant Speech Recognition*. John Wiley & Sons Ltd, 2009.

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

[3] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 114–126, Nov. 2012.

[4] O. Hazrati, J. Lee, and P. C. Loizou, "Blind binary masking for reverberation suppression in cochlear implants," *Journal of the Acoustical Society of America*, vol. 133, no. 3, pp. 1607–1614, 2013.

[5] 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.

[6] H. Kuttruff, *Room Acoustics*, 4th ed. London: Spon Press., 2000.

[7] *Acoustics - Measurement of Room Acoustic Parameters*, the International Organization for Standardization (ISO) Std., Aug. 2009.

[8] T. Nishiura, Y. Hirano, Y. Denda, and M. Nakayama, "Investigations into early and late reflections on distant-talking speech recognition toward suitable reverberation criteria," in *Proceedings of Interspeech*, Antwerp, Belgium, Aug. 2007, pp. 1082–1085.

[9] A. Sehr, E. A. P. Habets, R. Maas, and W. Kellermann, "Towards a better understanding of the effect of reverberation on speech recognition performance," in *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Tel Aviv, Israel, Aug. 2010.

[10] A. Brutti and M. Matassoni, "On the relationship between early-to-late ratio of room impulse responses and asr performance in reverberant environments," *Speech Communication*, vol. 76, pp. 170–185, 2016.

[11] E. A. P. Habets, S. Gannot, and I. Cohen, "Late reverberant spectral variance estimation based on a statistical model," *IEEE Signal Processing Letters*, vol. 16, no. 9, pp. 770–773, Sep. 2009.

[12] F. Xiong, S. Goetze, and B. T. Meyer, "Estimating room acoustic parameters for speech recognizer adaptation and combination in reverberant environments," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, May 2014, pp. 5559–5563.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] F. Xiong, S. Goetze, B. Kollmeier, and B. T. Meyer, "Exploring auditory-inspired acoustic features for room acoustic parameter estimation from monaural speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, pp. 1809–1820, 2018.

[27] D. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. NY: Wiley/IEEE Press, 2006.

[28] 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.

[29] 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.
[30] F. Xiong, S. Goetze, and B. T. Meyer, "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.
[31] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, pp. 41–75, Jul. 1997.
[32] 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
[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] 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.
[38] 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.
[39] 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, Jul. 2009, pp. 1–4.
[40] 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.
[41] 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.
[42] M. Jeub, M. Schäfer, H. Krüger, C. Nelke, C. Beaugeant, and P. Vary, "Do we need dereverberation for hand-held telephony?" in *Proceedings of International Congress on Acoustics*, Sydney, Australia, Aug. 2010, pp. 1–7.
[43] 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.
[44] 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.
[45] M. Sipser, *Introduction to the Theory of Computation*, 3rd ed. Cengage Learning, 2012.
[46] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding and Error Concealment*. John Wiley & Sons, Ltd, 2006.
[47] 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.
[48] 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.
[49] 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.
[50] 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.

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 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 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 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.