Environment Classification via Blind Roomprints Estimation

Malte Baum*†, Luca Cuccovillo*, Artem Yaroshchuk*, and Patrick Aichroth*

*Fraunhofer Institute for Digital Media Technology, 98693 Ilmenau, Germany
†Ilmenau University of Technology, 98693 Ilmenau, Germany

***Abstract*—In this paper we present a novel approach for environment classification for speech recordings, which does not require the selection of decaying reverberation tails. It is based on a multi-band RT60 analysis of blind channel estimates and achieves an accuracy of up to 93.6% on test recordings derived from the ACE corpus.**

***Index Terms*—environment classification, audio forensics, multi-band RT60 estimation**

I. Introduction

Much of modern communication takes place via social networks, in real-time, and with the help of smartphones and other mobile devices [1], resulting in an ever-increasing amount of user-generated content, and potential misuse: For instance, recordings can be manipulated in order to spread rumors and fake news, to support phishing attacks, or to create forged evidences to influence a trial. One discipline that can support content verification and detect such manipulations is audio forensics, which aims at analyzing footprints left by the recordings process, using automatic and reproducible analysis algorithms.

Environment classification is a specific audio forensics technique, which aims at determining where a recording took place – e.g., outdoor, in a small room, in a large church, by exploiting the acoustic characteristics of the recording environment [2], [3]. Such information can be very useful for verification, e.g. in court cases, as it supports the comparison of *claims* against *facts* regarding environment and recording characteristics. However, this task is challenging, and proposed approaches have struggled to reach an acceptable level of accuracy for handling real-life speech material [4], [5].

Furthermore, while some research focused on detecting inconsistencies of acoustic parameters related to decaying reverberation tails of a recording, thereby identifying spliced speech segments within a pre-existing recording [6], [7], no attention has yet been devoted to environment classification in presence of speech signals: Existing literature requires the presence of impulsive signals created e.g. by hand claps to work, which however rarely occur in relevant audio recordings [2]–[5].

In this paper, we propose to address the problem by applying multi-band reverberation time (RT_{60}) analysis of blind channel estimates: Multi-band RT_{60} analysis has been proposed as essential component for deriving roomprint features for environment classification in [2], [3]. Within the original proposed approach, such features were however retrieved under laboratory conditions, via analysis of the noiseless room impulse response in the time domain. Our proposal, in contrast, uses blind estimate of the existing transmission channel [8], which is used as source for estimating the desired roomprints. As a consequence, it is applicable to speech recordings without the need for automatic or semi-automatic selection of decaying reverberation tails, which is an error-prone step potentially leading to increased error rates.

The paper is structured as follows: In Section II, we present the background terminology and concepts required to retrieve roomprints for environment classification, which we then use in the algorithm proposed in Section III to perform environment classification from speech recordings, rather than in laboratory conditions. In Section IV, we then evaluate the classification algorithm using recordings derived from the ACE corpus [9], then closing in Section V with a summary and possible future research steps.

II. Background

A. Room Impulse Response and Reverberation Time

Let us denote with $s(t)$ an input speech signal reverberating through an environment characterized by a Room Impulse Response (RIR) denoted by $h(t)$. The corresponding signal $x(t)$ recorded by the receiving microphone can be modeled by means of a convolution $*$ in the time domain:

$$x(t) = h(t) * s(t) + n_{\text{env}}(t), \tag{1}$$

where $n_{\text{env}}(t)$ denotes the environmental noise and is therefore assumed equal to zero for noiseless recordings.

As depicted in Figure 1, the RIR $h(t)$ not only defines the time-delay T_0 between the sound emission and its acquisition, but most importantly the amount and characteristics of *early reflections*, conveying most of the information on the geometry and materials of the surroundings, and of *late reverberations*, conveying information on the size of the surroundings and of the absorbing power of the materials within it.

An important parameter related to the RIR is the reverberation time RT_{60}, i.e., the amount of time required for the space-averaged sound energy density in an enclosure to decrease by 60 dB after the source emission has stopped [11]:

This paper was supported by the BMBF SpeechTrust+ project (grant no 13N16267) and by the EU H2020 AI4Media project (grant no 951911).

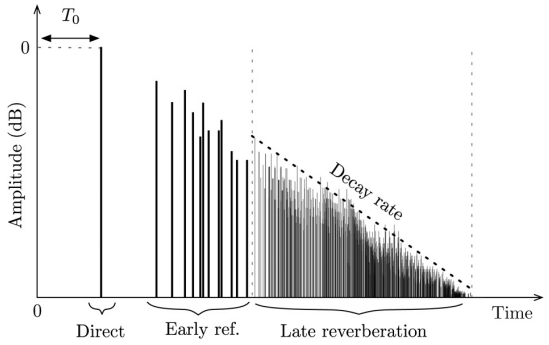

Fig. 1: Schematic example of a generic room impulse response in which $s(t)$ is a unitary pulse [10]

$$RT_{60} = \left(t \quad : \quad 10 \log_{10} \left(\left| \frac{x(0)}{x(t)} \right| \right) = 60 \right). \tag{2}$$

In practical cases, due to the presence of noise, RT_{60} is derived by linear interpolation:

$$RT_{60} = \alpha \cdot RT_{60/\alpha}, \tag{3}$$

where the parameter α can be adapted according to the noise conditions. Estimation is therefore often performed according to RT_{40} ($\alpha = 3/2$), RT_{30} ($\alpha = 2$) or RT_{20} ($\alpha = 3$).

B. Roomprints for Environment Classification

Roomprints for environment classification, introduced by Moore *et al.* in [3] and then refined in [2], propose to characterise a room by means of multi-band RT_{60} estimation.

The authors argued that the RT_{60} is largely related to the absorption properties of the materials present in the acoustic environment, which is frequency-dependent. Therefore, they proposed not to use a full-band RT_{60} estimation to characterize an acoustic environment, but rather to do so according to fractional octave filterbanks.

Fractional octave filterbanks can be calculated according to the ANSI standard [12] with the base-two system. Let us denote with B a positive integer to designate the fraction of an octave band (1/B).

The exact midband frequency of each bandpass filter of the filterbank is given by

$$f_m^{(b)} = \begin{cases} 2^{(b-30)/B} f_r & \text{if } B \text{ is odd,} \\ 2^{(2b-59)/(2B)} f_r & \text{if } B \text{ is even.} \end{cases} \tag{4}$$

f_r is the reference frequency of 1 kHz and b is any integer positive, negative or zero indicating the band number.

The frequencies of the lower and upper edges of the passband of the bandpass filter, called bandedge frequencies, can be expressed as

$$f_l^{(b)} = 2^{-1/(2B)} f_m^{(b)} \tag{5}$$

for the lower bandedge frequency and

$$f_u^{(b)} = 2^{1/(2B)} f_m^{(b)} \tag{6}$$

for the upper bandedge frequency.

The exact midband frequency is the geometric mean of the lower and upper bandedge frequencies:

$$f_m^{(b)} = \sqrt{f_l^{(b)} \cdot f_u^{(b)}}. \tag{7}$$

In order to estimate a roomprint, Moore *et al.* proposed to first apply the b-th filterbank to the RIR $h(t)$, and thus obtaining the fractional octave response $h^{(b)}(t)$ of the b-th sub-band:

$$h^{(b)}(t) = \text{bandpass-filter} \left(h(t), f_l^{(b)}, f_u^{(b)} \right). \tag{8}$$

Then, they proposed to compute the reverberation time independently for each band, obtaining one measurement $RT_{60}^{(b)}$ per each band.

The roomprint feature ψ of a single room can therefore be derived by aggregating the entire set of RT_{60} values:

$$\psi = \left[RT_{60}^{(1)}, RT_{60}^{(2)}, \ldots, RT_{60}^{(b)}, \ldots, RT_{60}^{(B)} \right], \tag{9}$$

where each $RT_{60}^{(b)}$ can be obtained by reverse integrating the energy of the band-passed filtered response $h^{(b)}(t)$, according to Schroeder's integration method [13]. A high-level schema of the process is depicted in Figure 2. Further details, including multi-path considerations and alternative definitions for the vector can instead be found in [2], [3].

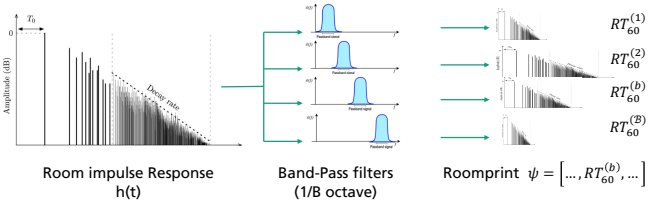

Fig. 2: High level schema of roomprint extraction

Roomprints have been tested extensively by the original authors, reaching a remarkable accuracy of 96.1% for a test set of 22 rooms, using $\psi' = \log(\psi)$ as feature vector and a Gaussian Mixture Model (GMM) for classification.

The main limitation of this feature and respective tests, however, lies in the fact that the test signals were obtained in noiseless laboratory environments by recording reverberations from sudden, impulsive bursts. The case of speech input signals was not addressed, due to the difficulties inherent its bandwidth limitation, and to the known issues related to the high variability of RT_{60} values estimated from decaying reverberation tails selected within speech recordings [5]. In the next section, we will therefore focus on the blind estimation of the room impulse response $h(t)$ from speech recordings, and retrieve a signal $\hat{h}(t)$ which can be used in conjunction with roomprints for effective environment classification.

III. Proposed Approach

A. Overview

Our approach for environment classification consists of four main steps, summarised here and described in the following sections:

1) A *blind channel magnitude estimation* step, which estimates the magnitude $|\widehat{H}(f)|$ of the RIR associated to the recording $x(t)$
2) A *minimum-phase digital filter estimation* step, which determines a digital filter $\widehat{H}(z)$ with minimum-phase response and magnitude as close as possible to $|\widehat{H}(f)|$
3) A *blind roomprint estimation* step, which computes the roomprint $\hat{\phi}$ associated to output in the time domain of the digital filter $\widehat{H}(z)$
4) A *closed-set classification* step, which according to the the estimated roomprint $\hat{\phi}$ returns a *label* of the room in which the recording $x(t)$ took place.

B. Blind Channel Magnitude Estimation

Under the assumption of time-invariant RIR, each frame $x_l(t)$ of the input recording $x(t)$ can be modeled by re-applying eq. (1) to the specific analysis window:

$$x_l(t) = h(t) * s_l(t) + n_{\text{env},l}(t) \tag{10}$$

An equivalent formulation in the log-power domain for the noiseless case of $n_{\text{env},l}(t)$ is:

$$X_l(f) = H(f) + S_l(f). \tag{11}$$

If we can estimate the dry input speech $S_l(f)$, i.e., if we can compute a term $\widehat{S}_l(f)$ which is accurate enough, $h(t)$ can be estimated blindly by applying:

$$\widehat{H}(f) = \frac{1}{L}\sum_{l=1}^{L}\left(X_l(f) - \widehat{S}_l(f)\right), \tag{12}$$

with L denoting the amount of frames and thus eq. (12) denoting the average difference between the input recording frames and the estimated ideal speech.

To further improve the stability of eq. (12), both terms $X_l(f)$ and $\widehat{S}_l(f)$ may be normalized to have zero mean, by applying

$$\underline{\mathrm{X}}_l(f) = X_l(f) - \frac{1}{N_{\text{stft}}}\sum_{f=1}^{N_{\text{stft}}} X_l(f), \tag{13a}$$

$$\underline{\widehat{\mathrm{S}}}_l(f) = \widehat{S}_l(f) - \frac{1}{N_{\text{stft}}}\sum_{f=1}^{N_{\text{stft}}} \widehat{S}_l(f), \tag{13b}$$

and obtaining the final eq. (14) for the mean-normalized blind estimation of the room impulse response:

$$\underline{\widehat{\mathrm{H}}}(f) = \frac{1}{L}\sum_{l=1}^{L}\left(\underline{\mathrm{X}}_l(f) - \underline{\widehat{\mathrm{S}}}_l(f)\right). \tag{14}$$

Intuitively, the better the estimated $\widehat{S}_l(f)$ is, the more accurate the estimated channel log-magnitude $\underline{\widehat{\mathrm{H}}}(f)$.

The basis for estimating $\widehat{S}_l(f)$ have been set by Gaubitch *et al.* in [8], both for the noiseless case we outlined so far, as well as for the case with additive noise. In the following, we are going to describe the estimation of the ideal dry speech in noiseless conditions and thus avoid a more cumbersome notation. If the noise is not negligible, it is possible to either follow the noise-aware estimation in [8], or – according to the outcome of recent experiments on the very same channel estimation procedure applied for estimating microphone frequency responses [14], [15] – to first apply denoising in the spectral domain and then to follow the noiseless estimation procedure we are going to describe herein.

The first step of the estimation procedure consists of processing a large speech corpus to extract a high amount of RASTA filtered Mel Frequency Cepstral Coefficients (MFCCs) [16]. In the following, MFCCs of the l-th frame of an input audio signal x will be denoted by the symbol c_{X_l}.

Given L_X training MFCC vectors c_{X_l}, used to fit a GMM with M mixtures, a key element of the estimation procedure is the relative mixture probability $p_i\left(c_{X_l}\right)$, i.e., the probability that the feature vector c_{X_l} belongs to the i-th mixture:

$$p_i\left(c_{X_l}\right) = \frac{\pi_i \cdot \mathcal{N}\left(c_{X_l} \mid \mu_i, \Sigma_i\right)}{\sum_{m=1}^{M} \pi_m \cdot \mathcal{N}\left(c_{X_l} \mid \mu_m, \Sigma_m\right)}. \tag{15}$$

In eq. (15), $\mathcal{N}\left(c_{X_l} \mid \mu_i, \Sigma_i\right)$ denotes the posterior probability of the vector c_{X_l} against the i-th mixture, having a normal distribution with mean μ_i, covariance Σ_i, and prior π_i.

With the help of these definitions, a model of the average log spectrum of the ideal speech can be obtained as follows:

1) Build a first normalized power spectrum matrix $\underline{\mathrm{X}}$, by collecting row-wise all mean-normalized log powers $\underline{\mathrm{X}}_l(f)$ of the GMM training set:
$$\underline{\mathrm{X}} \in \mathbb{R}^{L_X \times N_{\text{stft}}} = \left\{\underline{\mathrm{X}}_l(f)\right\}, \tag{16a}$$
2) Build a relative probability matrix $\underline{\mathrm{P}}_X$, by collecting row-wise all relative mixture probabilities $p_i\left(c_{X_l}\right)$ of the GMM training set:
$$\underline{\mathrm{P}}_X \in \mathbb{R}^{L_X \times M} = \left\{p_i\left(c_{X_l}\right)\right\} \tag{16b}$$
3) Compute the average speech spectrum matrix $\underline{\mathrm{S}}_X$:
$$\underline{\mathrm{S}}_X \in \mathbb{R}^{M \times N_{\text{stft}}} = \underline{\mathrm{P}}_X^t \cdot \underline{\mathrm{X}}, \tag{16c}$$
with t denoting the transposition.

The matrix $\underline{\mathrm{S}}_X$ is at the core of the ideal speech estimation procedure in [8]: Given an arbitrary input speech signal s having L_S frames and a relative probability matrix $\underline{\mathrm{P}}_S$, it is straightforward to compute:

$$\underline{\widehat{\mathrm{S}}} \in \mathbb{R}^{L_S \times N_{\text{stft}}} = \underline{\mathrm{P}}_S \cdot \underline{\mathrm{S}}_X, \tag{17}$$

i.e., a matrix whose rows can be applied directly in eq. (14) to obtain the desired estimate of the log-magnitude of the RIR in the frequency domain.

C. Minimum-phase Digital Filter Estimation

The estimated channel log-magnitude $\underline{\widehat{\mathrm{H}}}(f) \approx \log|H(f)|$ is not sufficient to recover an estimate $\hat{h}(t)$ of the RIR in the time domain by using the inverse Fourier transform: This operation requires an estimate of the phase component $\angle H(f)$, which is not retrieved by the algorithm in [8].

We thus propose to design a causal digital filter $\widehat{H}(z)$ with a response as close as possible to $\underline{\widehat{\mathrm{H}}}(f)$, and retrieve the RIR $\hat{h}(t)$ by filtering a unit impulse with $\widehat{H}(z)$.

Let us denote with $H(e^{j\omega}) \in \mathbb{C}$, $-\pi < \omega \leq +\pi$ the channel response we want to obtain, and assume our desired digital filter to be defined by

$$\widehat{H}(z) = \frac{\widehat{B}(z)}{\widehat{A}(z)}, \tag{18}$$

where

$$\begin{aligned} \widehat{B}(z) &= \hat{b}_0 + \hat{b}_1 z^{-1} + \ldots + b_{n_b} z^{-n_b}, \\ \widehat{A}(z) &= 1 + \hat{a}_1 z^{-1} + \ldots + a_{n_a} z^{-n_a}. \end{aligned} \tag{19}$$

The coefficients of the digital filter $\widehat{H}(z)$ can be obtained by minimizing the l^2-norm of the error

$$J(\hat{\theta}) = \left\| H(e^{j\omega}) - \widehat{H}(e^{j\omega}) \right\| \tag{20}$$

with respect to the filter coefficients

$$\hat{\theta} = [\hat{b}_0, \hat{b}_1, \ldots, \hat{b}_{n_b}, \hat{a}_1, \hat{a}_2, \ldots, \hat{a}_{n_a}] \tag{21}$$

by applying, e.g., Prony's method for filter design [17], [18].

The highest stability and accuracy of filter design methods is achieved in the case of minimum phase filters, for which poles and zeros fall inside the unitary circle. The phase of these filters can be determined analytically by recalling that the logarithm of their magnitude response is related to the phase response by means of the Hilbert transform $\mathcal{H}(\cdot)$ [19]:

$$\mathcal{H}\left(\log|H(f)|\right) = \log|H(f)| - j\angle H(f). \tag{22}$$

The minimum-phase digital filter $\widehat{H}(z)$ should therefore be designed to minimize its l^2-norm with respect to the transfer function $\widehat{H}(f)$ defined by:

$$\widehat{H}(f) = \exp\left\{\underline{\widehat{\mathrm{H}}}(f)\right\} \exp\left\{-j\,\mathrm{Im}\left(\mathcal{H}\left(\underline{\widehat{\mathrm{H}}}(f)\right)\right)\right\}, \tag{23}$$

with $\mathrm{Im}(\cdot)$ denoting the imaginary part of a complex number and $\exp\{\cdot\}$ the exponential operator.

D. Blind Roomprint Estimation

A roomprint describing the RIR associated to the digital filter $\widehat{H}(z)$ can be retrieved according to the procedure outlined in Section II.

The first step consists of recovering $\hat{h}(t)$, i.e., the estimated RIR in the time domain:

$$\hat{h}(t) = \text{digital-filter}\left(\delta(t), \widehat{H}(z)\right), \tag{24}$$

where $\delta(t)$ is the unit impulse:

$$\delta(t) = \begin{cases} 1 & \text{if } t = 0, \\ 0 & \text{if } t \neq 0. \end{cases} \tag{25}$$

$\hat{h}(t)$ is then filtered with fractional octave bandpass filters, to obtain band-limited signals

$$\hat{h}^{(b)}(t) = \text{bandpass-filter}(\hat{h}(t), f_l^{(b)}, f_u^{(b)}), \tag{26}$$

with $f_l^{(b)}$ and $f_u^{(b)}$ being the lower and upper bandedge frequencies of the b-th filter.

Finally, the estimated roomprint $\hat{\psi}$ can be computed by collecting several RT_{60} values estimated independently for each band using the Schroeder's integration method:

$$\hat{\psi} = \left[\widehat{RT}_{60}^{(1)}, \widehat{RT}_{60}^{(2)}, \ldots, \widehat{RT}_{60}^{(b)}, \ldots, \widehat{RT}_{60}^{(\mathcal{B})}\right]. \tag{27}$$

E. Closed-Set Classification

The last step of our proposed approach for environment classification consists of a feature vector computation, and of the actual training of the classifier. To ease the reproducibility, we use the roomprint estimate $\hat{\psi}$ in eq. (27) as feature vector for the classification.

The algorithm selected for the classification is a classic Support Vector Machine (SVM) with Radial Basis Function (RBF) kernel, with hyperparameters c and γ determined by grid search on the set $\mathcal{G}_c \times \mathcal{G}_\gamma$, where $\mathcal{G}_c = \{10^{-4}, \ldots, 10^3\}$ and $\mathcal{G}_\gamma = \{10^{-4}, \ldots, 10^3\}$.

IV. Evaluation

A. Datasets Involved

The evaluation of our proposed approach for environment classification involved several datasets in conjunction.

The SVM used for environment classification was evaluated on a dataset created by combining the ACE corpus [9] and the LibriSpeech corpus [20]. The ACE corpus consists of RIRs of seven different rooms, the general information of which is reported in Table I. For each room the impulse response was recorded in two positions, first in near-field conditions, and then in far-field conditions. The speech recordings drawn from the LibriSpeech corpus, which present no reverberation nor background noise, were first resampled to 16 kHz and then convolved with the seven pairs of RIRs from the ACE corpus.

TABLE I: Dimensions and RT_{60} values of rooms present in the ACE corpus for acoustic parameter estimation [9]

Label l	Name	Size (m)			Volume (m^3)	RT_{60} (s)
		L	W	H		
1	Office 1	4.8	3.3	3.0	47	0.34
2	Office 2	5.1	3.2	2.9	48	0.39
3	Meeting Room 1	6.6	4.7	3.0	92	0.44
4	Meeting Room 2	10.3	9.2	2.6	250	0.37
5	Lecture Room 1	6.9	9.7	3.0	200	0.64
6	Lecture Room 2	13.4	9.2	2.9	360	1.25
7	Building Lobby	5.1	4.5	3.2	72	0.65

The generated dataset consists of 200 noiseless reverberant recordings per RIR to be used for evaluation, with a total of 1400 recordings in near-field conditions as well as 1400 recordings in far-field conditions. The two sets were then split *once*, retaining 80% of the content for training and validation, and 20% for testing. Each set can be uniquely identified as follows:

$X_{\text{near}}^{\text{train}}$: Training examples in *near-field* conditions
$X_{\text{near}}^{\text{test}}$: Test examples in *near-field* conditions
$X_{\text{far}}^{\text{train}}$: Training examples in *far-field* conditions
$X_{\text{far}}^{\text{test}}$: Test examples in *far-field* conditions

To avoid any bias in the evaluation, we ensured an equal proportion of samples for each room in the training and testing set. Furthermore, no speech sample was convolved with two or more rooms, nor it appears in both the test and training set.

The last dataset involved in the evaluation is the VCTK corpus [21], which we used to train the GMM involved in the channel estimation procedure. In particular, we used the whole content of VCTK to train a 1024 mixture GMM, using 12 MFCCs per frame, where each frame was processed with an Hanning window having 128 ms of length and 50% overlap, which in [8] were found being optimal[1]. The speakers present in the VCTK corpus are absent in the LibriSpeech dataset related to the SVM training and testing. Therefore, this choice for the GMM should help in avoiding any evaluation bias related to the outcome of the environment classification. The sampling frequency was forced being equal to 16 kHz, to ensure compatibility between the channel estimation and the classification stage.

B. Bandwidth of the 1/B-octave filters

The first experiments to determine the quality of the environment classification aimed at determining the optimal bandwidth of the 1/B-octave fractional filters.

We tested the system training the SVM on roomprints computed using 1/3-octave, 1/4-octave and 1/8-octave fractional filters, obtaining the performance reported in Table II. All cases refer to RT_{60} values computed by linear interpolating RT_{20} values on the near-field dataset $X^{\text{test}}_{\text{near}}$.

The configuration with 1/4-octave bandwidth, corresponding to 26 bands ranging approximately from 100 Hz to 8 kHz, is the one performing the best, with an average accuracy of 89.6%. The outcome confirms the finding in [2], despite the higher sampling frequency used in our experiments.

TABLE II: Impact of filter bandwidths

Bandwidth	Performance (%)		
	Precision	Recall	Accuracy
1/3-octave	86.5	86.1	86.1
1/4-octave	90.2	89.6	89.6
1/8-octave	87.9	87.5	87.5

C. Interpolation parameter for RT60 estimation

A second experiment addressed the interpolation parameter α involved in the estimation of the RT_{60} values, once again in relation to the near-field test set $X^{\text{test}}_{\text{near}}$.

In particular, we set α in order to base the estimation of the reverberation time on the value of RT_{20} ($\alpha = 3$), RT_{30} ($\alpha = 2$), RT_{40} ($\alpha = 1.5$), RT_{50} ($\alpha = 1.2$) and RT_{60} itself ($\alpha = 1$), obtaining the performance reported in Table III. According to the previous experiments, the filter bandwidth was fixed and set equal to 1/4-octaves.

The configuration based on RT_{40} ($\alpha = 1.5$) is the one performing best. Intuitively, lower values of α at first improve the performance of the classification, since the uncertainty is decreasing accordingly. Very low values of α, however, cause numerical instability for bands that decrease too rapidly, since the Schroeder's integration method is disrupted by the floor noise level. This experiments confirms the tendency, and allows us to determine $\alpha = 1.5$ as sweet spot for the evaluation.

TABLE III: Impact of RT_{60} interpolation

$RT_{60/\alpha}$	α	Performance (%)		
		Precision	Recall	Accuracy
RT_{20}	3.0	90.2	89.6	89.6
RT_{30}	2.0	90.9	90.7	90.7
RT_{40}	1.5	93.8	93.6	93.6
RT_{50}	1.2	91.6	91.1	91.1
RT_{60}	1.0	90.7	90.4	90.4

D. Near-field and Far-field Comparison

In the third and last experiment, we used the parameters selected for being the most effective – namely a 1/4-octave filter bandwidth and the RT_{60} interpolation parameter $\alpha = 1.5$ – to evaluate the performance of the system in relation to near-field and far-field recording conditions.

The experiments are summarised in Table IV, in which the last rows refer to the case of performing the evaluation by merging the near-field and far-field datasets together. The outcome of the evaluation, which we are going to comment in the following, met our expectations in terms of the predicted system behavior.

In near-field conditions for both training and test data, the system performs the best with a satisfying accuracy of 93.6%, whereas in far-field conditions the accuracy drops to 89.6%. The drop is probably due to the difficulty of the GMM in estimating the ideal speech, due to the low ratio between the energy of the direct speech for the frame, and the one of the reverberant speech from previous ones.

When training and testing conditions do not match, the accuracy drops significantly by more than 20%. Once again, the drop could have been predicted beforehand: The distribution of training and test data are very different – to the point that is debatable whether the room should be identified or not, considering that the RIRs might differ significantly – and thus the SVM has severe problems in the classification task.

Lastly, the performance of the SVM in mixed conditions is better than for mismatching training and testing sets, but still

TABLE IV: Impact of Near-field and Far-field conditions

Training Dataset	Test Dataset	Performance (%)		
		Precision	Recall	Accuracy
$X^{\text{train}}_{\text{near}}$	$X^{\text{test}}_{\text{near}}$	93.8	93.6	93.6
$X^{\text{train}}_{\text{far}}$	$X^{\text{test}}_{\text{far}}$	90.2	89.6	89.6
$X^{\text{train}}_{\text{near}}$	$X^{\text{test}}_{\text{far}}$	71.1	70.0	70.0
$X^{\text{train}}_{\text{far}}$	$X^{\text{test}}_{\text{near}}$	62.7	63.9	63.9
$X^{\text{train}}_{\text{mixed}}$	$X^{\text{test}}_{\text{mixed}}$	84.8	84.6	84.6

[1]Their optimality can also be confirmed by comparing different GMM configurations by means of the Bayesian information criterion (BIC).

inferior to the initial one with training and test set properly selected. Indeed, we can imagine that the SVM is fitting two (in principle disjoint) distributions per room, leading to a high error rate.

To summarize, this last experiment proves that the proposed algorithm for environment classification should be applied only in presence of prior information about the relative distance between the speakers and the recording device, in order to select the training content appropriately.

E. Relation with state-of-the art results

Existing state-of-the-art methods, as stated upfront in the introduction, can be applied in the presence of impulsive signals created e.g. by hand claps, which however rarely occur in relevant audio recordings [2]–[5].

A comparison might have been possible by applying an algorithm for semi- or fully- automatic selection of reverberation tails, and by processing these tails as if they were created by an impulsive sound or white noise rather than by speech.

We decided not to perform such a comparison, however, to avoid incurring in any bias due to, e.g., faulty tail selection or mismatching statistical properties of the input audio signal. As absolute reference, the state-of-the-art roomprints proposal in [2] achieved a classification accuracy of 97.1% on a self-curated dataset with 22 rooms, given noiseless impulse responses as input.

V. Conclusions and Outlook

To our knowledge, the algorithm proposed in this work represents the first attempt to address environment classification for the specific case of speech recordings: Previous methods addressed recordings created under laboratory conditions, exploiting decaying reverberation tails produced by impulsive signals, resulting in difficulties to obtain stable estimates of the reverberation parameters. In contrast, our approach is based on multi-band RT_{60} analysis of blind channel estimates, which does not require selection of the decaying reverberation tails, and is suitable for retrieving roomprint features whenever speech is dominant in a recording.

Under near-field and noiseless conditions, the algorithm is achieving an accuracy of about 93.6% for recordings derived from the ACE corpus, dropping to about 89.6% in the case of far-field conditions. The accuracy of the system drops significantly in mismatching conditions, thereby providing implicit evidence of the sensibility of the algorithm to changes with respect to the room impulse responses of speech source and recording device.

For the future, we plan to determine whether noisy conditions are best addressed using the original derivation for channel estimation in [8], or rather by applying AI-based denoising as in [14], [15]. Furthermore, since the channel estimation procedure has been extensively used for performing microphone classification, as e.g. in [14], [15] and references thereof, we also plan to investigate the influence of the microphone on environment classification, and to determine whether the two contributions to the channel can be decoupled effectively or not.

References

[1] A. Mitchell, M. Jurkowitz, J. B. Oliphant, and E. Shearer. "Survey results: About one-in-five US adults say they get their political news primarily through social media." Pew Research Center. (2019), [Online]. Available: https://s.fhg.de/nn4.

[2] A. H. Moore, M. Brookes, and P. A. Naylor, "Room identification using roomprints," in *AES International Conference on Audio Forensics*, London, United Kingdom, 2014.

[3] ——, "Roomprints for forensic audio applications," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2013, pp. 1–4.

[4] R. K. Patole, P. P. Rege, and P. Suryawanshi, "Acoustic environment identification using blind de-reverberation," in *IEEE International Conference on Computing, Analytics and Security Trends (CAST)*, Pune, India, 2016, pp. 495–500.

[5] H. Malik and H. Farid, "Audio forensics from acoustic reverberation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Dallas, TX, USA, 2010, pp. 1710–1713.

[6] D. Capoferri, C. Borrelli, P. Bestagini, F. Antonacci, A. Sarti, and S. Tubaro, "Speech audio splicing detection and localization exploiting reverberation cues," in *IEEE International Workshop on Information Forensics and Security (WIFS)*, New York, NY, USA, 2020, pp. 1–6.

[7] R. Patole, G. Kore, and P. Rege, "Reverberation based tampering detection in audio recordings," in *AES International Conference on Audio Forensics*, Paper no 1-4, Arlington, VA, USA, 2017.

[8] N. D. Gaubitch, M. Brookes, and P. A. Naylor, "Blind channel magnitude response estimation in speech using spectrum classification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2162–2171, 2013.

[9] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "Estimation of room acoustic parameters: The ACE challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 24, no. 10, pp. 1681–1693, 2016.

[10] V. Valimaki, J. D. Parker, L. Savioja, J. O. Smith, and J. S. Abel, "Fifty years of artificial reverberation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 5, pp. 1421–1448, 2012.

[11] ISO Standard 3382-1:2009. "Acoustics – measurement of room acoustic parameters – part 1: Performance spaces." (2009), [Online]. Available: https://www.iso.org/standard/40979.html.

[12] ANSI/ASA S1.11-2004 (R2009). "Octave-band and fractional-octave-band analog and digital filters." (2004), [Online]. Available: https://webstore.ansi.org/standards/asa/ansiasas1112004r2009.

[13] M. R. Schroeder, "New method of measuring reverberation time," in *Journal of the Acoustical Society of America*, vol. 37, 1968, pp. 409–412.

[14] L. Cuccovillo, A. Giganti, P. Bestagini, P. Aichroth, and S. Tubaro, "Spectral denoising for microphone classification," in *ACM International Workshop on Multimedia AI against Disinformation (MAD)*, in press, Newark, NJ, USA, 2022.

[15] A. Giganti, L. Cuccovillo, P. Bestagini, P. Aichroth, and S. Tubaro, "Speaker-independent microphone identification in noisy conditions," in *European Signal Processing Conference (EUSIPCO)*, in press, Belgrade, Serbia, 2022.

[16] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

[17] J. O. Smith, *Introduction to Digital Filters with Audio Applications*. W3K Publishing, 2007, ISBN: 978-0-9745607-1-7.

[18] C. S. Burrus and T. W. Parks, "Time domain design of recursive digital filters," *IEEE Transactions on Audio and Electroacoustics (TAE)*, vol. 18, no. 2, pp. 137–141, 1970.

[19] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*. Upper Saddle River, NJ, USA: Prentice-Hall, 2014, pp. 788–789, ISBN: 0-13-754920-2.

[20] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, QLD, Australia, 2015, pp. 5206–5210.

[21] J. Yamagishi, C. Veaux, and K. MacDonald, *VCTK corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)*, 2019.