

COVER PAGE

Document downloaded by @DAEL

Fri Nov 27 15:00:26 2026
For personal use

When automatic English translation is provided, only the original document is authentic.
The EAA cannot be held responsible of any translation error

Bibliographical reference

Towards Room-Volume Classification from Reverberant Speech using Room-Volume Feature Extraction and Room-Acoustics Parameters, Noam R. Shabtai, Yaniv Zigel and Boaz Rafaely, Acta Acustica **vol. 99** (Number 4), 2013, pp. 658-669

DOI
https://doi.org/10.3813/AAA.918644

ACTA ACUSTICA UNITED WITH ACUSTICA
Vol. 99 (2013) 658 – 669

Technical and Applied Papers
DOI 10.3813/AAA.918644

Towards Room-Volume Classification from Reverberant Speech using Room-Volume Feature Extraction and Room-Acoustics Parameters

Noam R. Shabtai[1)], Yaniv Zigel[2)], Boaz Rafaely[1)]
[1)] Department of Electrical and Computer Engineering, Ben-Gurion University of the Negev, P.O.B. 653, Beer-Sheva, 84105 Israel. shabtai.noam@gmail.com
[2)] Department of Biomedical Engineering, Ben-Gurion University of the Negev

Summary
Classification of the room volume from reverberant speech signals can be useful in acoustic scene analysis applications such as forensic audio, rescue and security. In the past, such work explicitly required knowledge of the room impulse response (RIR) to either estimate or classify the room volume. In this work, different approaches to extract room-volume features from speech rather than from the RIR are investigated. These include using abrupt stops in speech, dereverberation, extraction of room-volume features directly from speech and using speech recognition features, such as mel frequency cepstral coefficients (MFCC). The room volume is classified using a pattern recognition-based system in which the room-volume features are employed using a feature selection algorithm. Three experimental studies using (1) speech convolved with simulated RIRs, (2) speech convolved with measured RIRs and (3) recorded speech in reverberant environments are presented. It has been shown for all experimental studies that the approach of using abrupt stops in speech outperforms all other approaches. It has also been shown that the classification of room volume is affected mainly by room-volume features that are calculated from the low frequency bands of the room transfer function, in addition to reverberation time or early decay time.

PACS no. 43.60.-c

1. Introduction

Room-volume classification from reverberant speech signals can be useful in acoustic scene analysis applications to help characterize the type of enclosure in which the speech signal was received, e.g., car, office, concert hall, etc. Potential applications may include forensic audio, rescue and security. In forensic audio applications, the claimed location of a conversation can be verified. For rescue purposes, information on the rescue environment that is extracted from a received speech signal may be useful.

However, research work on room-volume classification from reverberant speech is limited in the current literature. As opposed to estimation of other room parameters from speech signals [1], or from music signals [2], methods for room-volume classification seem to exist only in feasibility studies in which the *room impulse response* (RIR) is known. Kuster [3] compared four methods for estimating room volume from the RIR. These were the geometrical acoustic method, the eigenmode method, the diffuse-field method and correlation with *reverberation time* (RT). Among the four methods, only the diffuse-field method and the correlation with RT were reported to deliver consistent results. However, the diffuse-field method required exact information on the source-to-receiver distance and the correlation with RT is sensitive to differences in absorption.

Room-volume feature extraction from the RIR has been developed in research studies that were performed by the authors [4, 5, 6]. Room-volume features are parameters that are extracted from observations of audio signals which were recorded in a room and are related to the room volume. These room-volume features have been employed for room-volume classification using a statistical pattern recognition approach, which is potentially robust both to differences in source-receiver distance outside the radius of reverberation and to differences in absorption. However, this method still requires the known RIRs in order to classify the room volume.

In the current work, room-volume features are extracted directly or indirectly from speech signals with no previous knowledge of the RIRs. This is performed using three different approaches. The first is by extracting room-volume features from abrupt stops in speech. In this approach it is assumed that the decay pattern of a reverberant speech signal following a single short word may resemble an impulse response signal. The second is by performing system iden-

Received 4 July 2012,
accepted 5 May 2013.

© S. Hirzel Verlag · EAA

tification using dereverberation and then extracting room-volume features from the resulting RIR. The third is by extracting room-volume features directly from speech without pre-processing. In addition, the approach of extracting mel-frequency cepstral coefficient (MFCC) features from reverberant speech and using these features in order to classify room volumes is also investigated.

This work extends an early study conducted by the authors [7] showing an experimental study using convolution of anechoic speech with measured RIRs. This work extends the discussion in [7] on the room-volume features and their relation to the volume. In addition, an experimental study is provided with higher room-volume resolution using RIR simulations, which facilitate the estimation of room volume using room classification. Furthermore, another important experimental study is presented in which recorded speech in real reverberant enclosures is used.

Section 2 presents room-acoustics parameters that are relevant to the room volume and which may be calculated from the RIR. Section 3 presents the methods in which room-volume features are extracted from a speech signal. Section 4 presents the room-volume classification system. Section 5 presents an experimental study using speech signals which are convolved with simulated RIRs with high resolution on the room-volume. Section 6 presents an experiment where the speech signal is convolved with measured RIRs. Due to limited resolution on room volume, section 6 only shows how this method can be used to classify rooms with similar characteristics, rather than to strictly classify room volumes. section 7 presents an experimental study using speech segments that were recorded in reverberant environments, using a sub-set of the rooms that were used in section 6. Section 8 concludes the paper.

2. Room-acoustics parameters

In this section, room-acoustics parameters that are related to the room volume are presented. Most of these parameters are more likely to be calculated using the RIR than from actual speech. Nevertheless, these parameters may be calculated from reverberant speech segments, in accordance with the methods presented in section 3. These parameters have been presented in [4] and are outlined here for completeness.

2.1. Density of reflections

The density of reflections arriving from the room boundaries is closely related to the room volume. The larger the room, the sparser the reflections, and vice versa. The temporal density of reflections in a room of volume V can be approximated by [8]

$$\frac{\mathrm{d}N\,(t)}{\mathrm{d}t} = 4\pi\frac{c^3t^2}{V} \qquad (1)$$

where $N\,(t)$ is the number of reflections arriving up to time t and c is the speed of sound. Figures 1a and 1b show the first 60 ms of a measured RIR in a small office (4 m^3) and in a concert hall ($\sim$ 18200 m^3), respectively. The difference between the RIRs is significant.

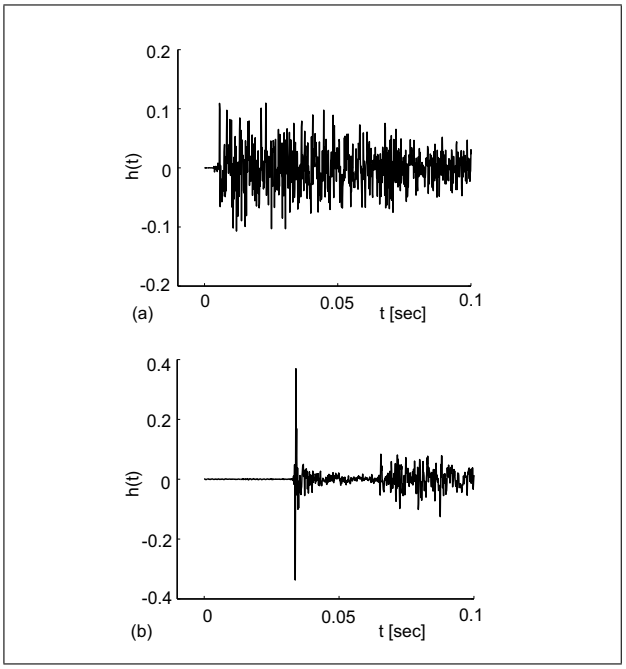


Figure 1. Measured RIR (a) in a small office (42 m^3), and (b) in a concert hall ($\sim$ 18200 m^3).

In order to calculate the temporal density of reflections, it is necessary to count the peaks in an early part of the RIR and to record the number of peaks divided by the part length; this is the temporal density at the time in the center of that RIR part. If the same part is used for all the RIRs from which the room volume is to be identified, it may only be necessary to record the number of peaks in that part. However, it may be difficult to trace these peaks and, moreover, to count them. This is probably a result of acoustic scattering and noise that are not taken into account in equation (1). For this reason, enveloping may be applied on the RIRs prior to calculating the number of reflections.

2.2. Kurtosis in the time domain

Kurtosis of a time domain signal describes the peakedness and tail behavior of its *probability density function* (PDF) [9]. Higher kurtosis means that more of the variance is due to infrequent extreme deviations, as opposed to lower kurtosis, which means that more of the variance is due to frequent modestly sized deviations. Hence, kurtosis of the RIR in the time domain can be related to the volume because large rooms typically have sparse early reflections, while small rooms have more dense reflections over time. For example, the measured RIRs in Figures 1a and 1b show kurtosis of 69 and 615, respectively, within the first 60 ms.

2.3. Reverberation time

The *reverberation time* (RT) is the time it takes the energy in a room to decay by 60 dB from a steady-state excitation. The RT may be extracted from the RIR using the ISO 3382 standard [10]. In statistical room acoustics where a diffuse

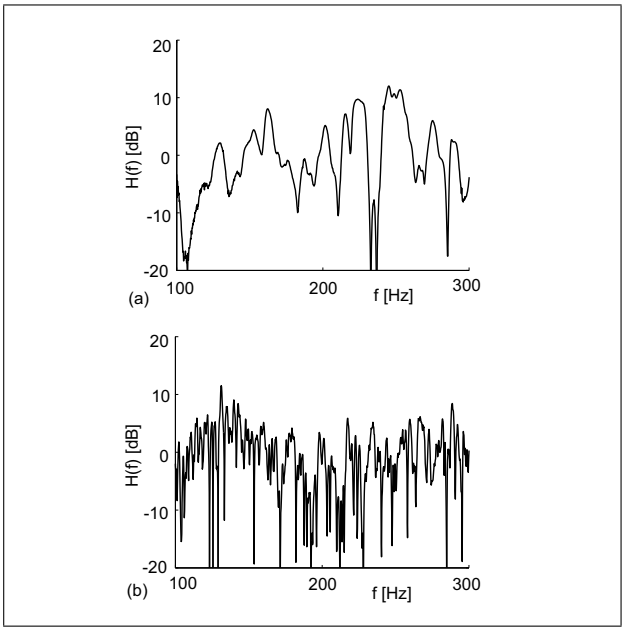


Figure 2. Measured RTF at [100, 300] Hz (a) in an office (42 m^3), with f_{sc}=240 Hz and (b) in a concert hall ($\sim$ 18200 m^3), with f_{sc}=20 Hz.

field is assumed to exist, the RT is related to the volume V according to the Sabine formula [11]

$$\mathrm{RT} = 0.161\frac{V}{A} \tag{2}$$

where A is the absorption area of the room. However, since the absorption in the room is usually unknown, RT by itself cannot discriminate between rooms of different volumes.

2.4. Room transfer function and eigenmodes

Room response from a source to a receiver can be described in the frequency domain by the *room transfer function* (RTF). In rectangular rooms, the RTF is known to be a combination of eigenmodes. An eigenmode may be noticed as a peak in the RTF magnitude. The density of eigenmodes around frequency f is related to the room volume V by [12]

$$\frac{\mathrm{d}N_f(f)}{\mathrm{d}f} = 4\pi V\frac{f^2}{c^3} \tag{3}$$

where $N_f(f)$ is the number of eigenmodes in the frequency band $[0, f]$ of the RTF magnitude. Figures 2a and 2b show the [100, 300] Hz band in RTFs of a small office (42 m^3) and a concert hall ($\sim$ 18000 m^3), respectively, calculated from measured RIRs. In Figure 2a, a low density of eigenmodes may be detected as distinct peaks in the RTF for the small office. On the other hand, Figure 2b shows a high density of eigenmodes in the RTF for the concert hall. Hence, the number of peaks in different RTF magnitude bands may be used as room-volume features.

However, it is worth mentioning that room-volume estimation based only on the density of eigenmodes might be sensitive to room architecture since equation (3) applies only to rectangular rooms. Furthermore, a simple approach comprising a selection of frequency bands where the lowest modes are observed, followed by counting eigenmodes in these bands might not prove consistent since these may be associated with the longest dimension of the room in case it is considerably longer than the other two dimensions.

2.5. Schroeder frequency

At frequencies where the density of the eigenmodes is above three eigenmodes for a 3 dB bandwidth, the sound field may be considered as sufficiently satisfying the assumptions of diffuse-field theory. The threshold frequency is referred to as the *Schroeder frequency*, f_{sc}. The Schroeder frequency is related to the volume V and RT by [13]

$$f_{sc} \approx 2000\sqrt{\frac{\mathrm{RT}}{V}} \quad . \tag{4}$$

The Schroeder frequency as an indicator of eigenmode density may therefore be affected by room volume.

2.6. Kurtosis and standard deviation in the frequency domain

In Kuster's work [3], a calculation of f_{sc} from the RTF was performed by using the RTF magnitude statistics parameters of *standard deviation* (STD) and kurtosis. It was demonstrated that the STD and the kurtosis of the RTF magnitude in different frequency bands may be used as room-volume features that together with RT may discriminate between rooms with different volumes. Along with the STD and the kurtosis of the RTF magnitude, the number of mean-value crossings can also indicate on the modal density. If the eigenmodes are sparse, a small number of mean-value crossings is expected in the RTF magnitude, and vice versa.

3. Extraction of room-volume features from reverberant speech

Using the room-acoustics parameters that were presented in section 2, room-volume features may be extracted from the RIRs. In this work it is required to extract room-volume features from reverberant speech signals. For this reason, different approaches may be considered, as illustrated in Figure 3. The approaches that are considered in this work are:

1. extracting room-volume features from abrupt stops in speech,
2. performing system identification with dereverberation and then extracting room-volume features,
3. extracting room-volume features directly from speech without pre-processing, or
4. to ignore the room-volume features and instead to extract features that are usually extracted in speech processing applications, e.g., MFCC.

Also, the room-volume features are extracted from the known RIRs as a reference for performance evaluation.

The purpose of the "room-volume feature extraction" blocks in Figure 3 is to extract room-volume features from a given signal at the input, which is considered to be the RIR. The features that are extracted from the RIR are presented in Table I. A set of scalar room parameters which are selected from the different sets in Table I, is extracted to form a feature vector, $\mathbf{x}$. The extraction of room-volume features from the RIR has been previously presented [4] and is outlined here in the following paragraphs for completeness.

Energetic features include the RT, *early decay time* (EDT), *definition* (D_{50}), *clarity index* (C_{80}) and *centre time* (T_s). These are all calculated using the ISO 3382 standard [10]. An energetic feature variant to the D_{50} of ISO 3382 that is used in this work is the *frame-definition* D_{20} [14], which describes the ratio between the energy in the first 20ms of the RIR and the entire RIR signal, instead of the usual 50ms used in D_{50}.

The temporal reflections and the temporal kurtosis are calculated from an early part of the RIR, where reflections are sparser and of greater amplitude than in later parts. The number of reflections is counted by tracing peaks in the RIR envelope. Two approaches for envelope calculation are used, namely the *short-time root-mean-square* (STRMS) [15] to form N_S, and using the Hilbert transform of the RIR [16] to form N_H. The kurtosis of the RIR samples is calculated from the RIR, from its STRMS envelope and from its Hilbert envelope, to form K_h, K_S, and K_H, respectively.

The kurtosis and the STD of the RTF magnitude, $K_{|\mathrm{RTF}|}[f_1, f_2]$ and $\sigma_{|\mathrm{RTF}|}[f_1, f_2]$, are calculated from different frequency bands $[f_1, f_2]$. The number of modes $N_{|\mathrm{RTF}|}[f_1, f_2]$ is counted by tracing peaks in the RTF magnitude in those bands. The number of mean-value crossings, $MC_{|\mathrm{RTF}|}[f_1, f_2]$, is calculated by subtracting the mean value of the RTF magnitude in the band $[f_1, f_2]$ and counting the number of zero crossings.

Various frequency bands $[f_1, f_2]$ are used in the form of

1. $\left[\frac{kf_0}{2}, kf_0 \times 2\right]$ where $f_0 = 100Hz$ and $k = 1, \ldots, 3$.
2. $\left[\frac{kf_0}{1.5}, kf_0 \times 1.5\right]$ where $f_0 = 100Hz$ and $k = 1, \ldots, 5$.
3. $\left[\frac{kf_0}{1.2}, kf_0 \times 1.2\right]$ where $f_0 = 100Hz$ and $k = 1, ..., 10$.

such that bands are overlapping in frequency and have different widths. This will ensure that frequencies which are lower than the typical cut-off frequency of telephony or hand-set media will not be used.

3.1. Extraction of room-volume features from abrupt stops in speech

The first approach for room-volume feature extraction from speech is from abrupt stops in speech. Ideally, words which end with a plosive consonant should be selected, but it may be a difficult task to indicate the location of these based on a signal form of a reverberant signal. Hence, abrupt stops are considered in this work to be short words. The abrupt stop detector in Figure 3 receives a reverberant

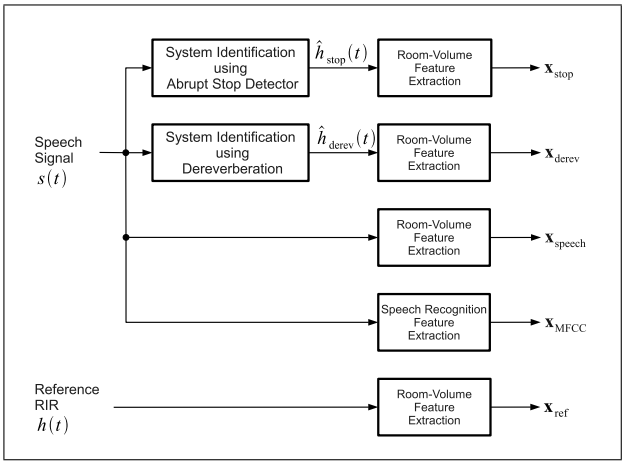


Figure 3. Different approaches for feature extraction from reverberant speech.

speech signal, $s(t)$, as an input and outputs a detected isolated word, $\hat{h}_{\mathrm{stop}}(t)$, which is typically shorter than 1s, but is associated with a 1s silence before and 2s silence after the isolated words. Hence the total length of the signal is typically 3-4 seconds and windowing is not required. Although automatic stop detectors are available for extracting room parameters from running speech [2], experimental studies in this work will only use manual detection of these.

In this approach it is assumed that the decay pattern following that short single word may resemble a RIR signal decay characteristic. This is not entirely true considering the details of the information contained in the signal, as it is not possible to refer to the speech unit as having an abrupt end-point. Furthermore, the number of stops and the silence durations are expected to strongly impact the performance of the feature vectors that are extracted using this approach. However, this method has been used before in the context of RT estimation from reverberant speech [17].

This approach may require more pre-processing than the following approaches, which may have an effect on real-time applications. However, in data mining, where real time may not be the issue of interest, manual selection of parts in speech can be performed prior to the execution of the core algorithm. Furthermore, a *voice activity detector* (VAD) in combination with peak detectors may replace this manual selection, however this is outside the scope of this paper. Assuming that an abrupt stop detector performs automatically or manually, a feature vector, $\mathbf{x}_{\mathrm{stop}}$, is extracted from $\hat{h}_{\mathrm{stop}}(t)$ according to Table I.

3.2. Extraction of room-volume features from identified system using dereverberation

The second approach for room-volume feature extraction from speech signals is to perform system identification using dereverberation, and then extracting room-volume features from the resulting estimated RIR. The dereverbera-

Table I. Extraction of room-volume features from the RIR.

Features in the set	Calculated from	Symbols
RIR energetic features.	Integration of the RIR.	RT, EDT, D_{20}, D_{50}, C_{80}, T_s
Number of peaks (reflections) in an early part of the RIR.	STRMS RIR Envelope. Hilbert-transform RIR Envelope.	N_S N_H
Kurtosis of RIR samples in an early part of the RIR.	RIR STRMS RIR Envelope. Hilbert transform RIR Envelope.	K_h K_S K_H
Kurtosis of RTF magnitude.	RTF magnitude in the frequency band $[f_1, f_2]$.	$K_{\|\mathrm{RTF}\|}[f_1, f_2]$
STD of RTF magnitude.		$\sigma_{\|\mathrm{RTF}\|}[f_1, f_2]$
Number of peaks (modes).		$N_{\|\mathrm{RTF}\|}[f_1, f_2]$
Number of mean-value crossings.		$MC_{\|\mathrm{RTF}\|}[f_1, f_2]$

tion block in Figure 3 receives a reverberant speech signal, $s(t)$, and identifies the RIR as $\hat{h}_{\mathrm{derev}}(t)$ by applying

$$\hat{h}_{\mathrm{derev}}(t) = \mathcal{F}^{-1}\left\{\frac{\mathcal{F}\{s(t)\}}{\mathcal{F}\{\widetilde{s}(t)\}}\right\} \tag{5}$$

where $\mathcal{F}\{\cdot\}$ represents the Fourier transform operator and $\widetilde{s}(t)$ is the reverberation-suppressed speech signal. Then, a feature vector, $\mathbf{x}_{\mathrm{derev}}$, is extracted from $\hat{h}_{\mathrm{derev}}(t)$ using the features indicated in Table I.

In this work, dereverberation is performed using a single-microphone spectral enhancement algorithm [18], also referred to as reverberation-suppression. This algorithm was designed for suppressing the effect of late room reflections which makes a speech signal sound echoic. However, reconstructing the impulse response of the room using this kind of blind dereverberation algorithm is considered a more difficult task, which may be the reason why this approach may prove inconsistent for room-volume classification. Figure 4 shows an example of a measured room impulse response and its reconstruction after filtering with a speech signal and applying dereverberation. The frequency responses are also displayed. It is clearly seen that the temporal structure of the impulse response is not maintained.

The input signals to the dereverberation blocks are complete sentences with connected speech. Short breaks between sentences are indicating the beginning and the end of each sentence. These breaks may also be detected using a VAD algorithm. The result would be reverberant speech signals containing a single sentence, each following a (reverberant) break, which typically lasts 3 to 4 seconds.

3.3. Extraction of room-volume features directly from speech

The third approach for room-volume feature extraction from speech is to extract these features directly from reverberant speech with no prior attempt of estimating the RIR. In consequence, minimal pre-processing may be required. The motivation for this approach stems mainly from its simplicity. Furthermore, humans classify auditory scenes based on perceived speech signals that contain complete sentences of connected speech, which are typically available in a normal conversation. Therefore, this approach may be used in a future comparison with human performance for classifying the room-volume from perceived speech signals. A justification for its use may be the potential ability of the feature selection algorithm (to be discussed in section 4) to focus on frequency bands in reverberant speech where more information is available on the room than in the speech itself. These frequency bands may include low frequency eigenmodes, for example.

Here, instead of using isolated words, complete sentences with connected speech are used. In a similar manner to the dereverberation approach, short breaks between sentences are used to indicate the start and end of a sentence, resulting in 3-4 seconds speech input signals including the breaks. The successive break after the sentence maintains a decay pattern. The resulting signal, $s(t)$, which includes the signal and the break, is not required to have any further windowing, since the non-continuity at the beginning of the signal is actually desired and since the signal is already smooth at the end due to the decay. The signal $s(t)$ is processed directly by the room-volume feature extraction block in Figure 3 as if it was the RIR itself, using the features from Table I. This procedure results in a single feature vector, $\mathbf{x}_{\mathrm{speech}}$, for each speech sentence.

3.4. Extraction of MFCC features from speech

Figure 3 shows that both $\mathbf{x}_{\mathrm{speech}}$ and $\mathbf{x}_{\mathrm{MFCC}}$ feature vectors are extracted directly from speech. The difference, however, is that $\mathbf{x}_{\mathrm{speech}}$ contains room acoustic features, while $\mathbf{x}_{\mathrm{MFCC}}$ contains features that are usually extracted in speech processing applications such as speech or speaker recognition. The *mel frequency cepstral coefficients* (MFCC) [19] are usually extracted in such applications, and will be denoted by $\mathbf{x}_{\mathrm{MFCC}}$. It should be mentioned that longer frames are expected to be used in room-volume classification than the usual 20–30 ms frames that are often used in speech processing. The reason is that room-volume classification requires features that are neither related to the contents of the speech nor to the physiological characteristics of the speaker, but to the characteristics of the room from which the reverberant speech signal was re-

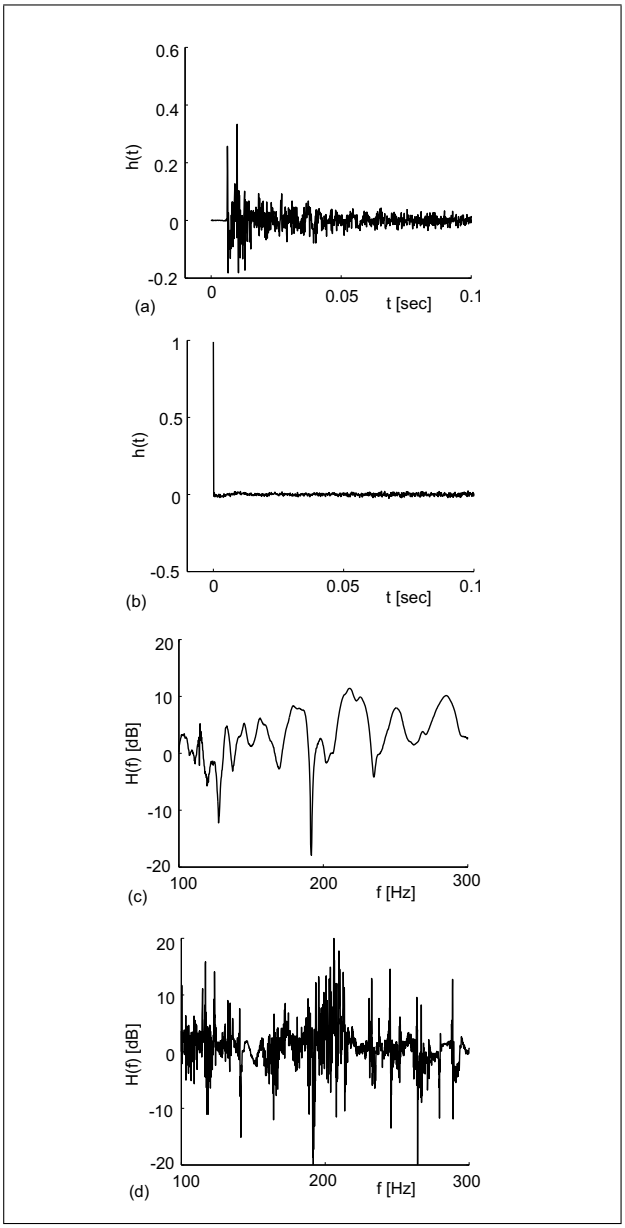


Figure 4. (a) Measured RIR in a teaching lab with a volume of 340 m^3 vs. (b) the reconstruced RIR using dereverberation of a speech signal that was convolved with the same RIR. (c) The corresponding measured RTF and (d) the reconstructed RTF.

ceived. Hence, it is not necessary in this work to maintain a stationary speech signal within the analysis frames. Furthermore, the length of the analysis windows in the MFCC feature extraction for room-volume features is expected to be on the scale of seconds, since that is the scale in which the decay characteristics of a RIR are usually measured.

The MFCC feature extraction approach uses 12 MFCC coefficients c_1-c_{13}, 12 ΔMFCC coefficients Δc_1 - Δc_{13}, and 12 ΔΔMFCC coefficients $\Delta\Delta c_1$ - $\Delta\Delta c_{13}$ such that first MFCC coefficient c_0 and its derivatives were omitted. Time domain windowing is performed using 50% overlapping 1s Hamming windows. The reason for the relatively long windows is the slow-varying room characteristics with respect to speech signal characteristics.

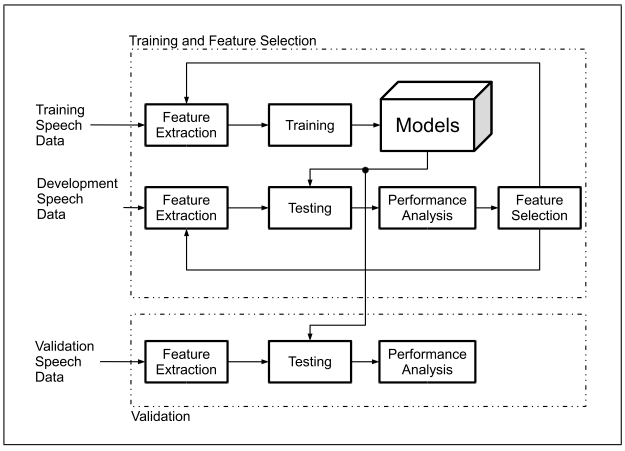


Figure 5. The room-volume classification process.

4. Room-volume classification

A statistical pattern recognition approach for classification of room-volume from speech signals is proposed using the room-volume features that are extracted according to section 3. Figure 5 shows a block diagram of the proposed room-volume classification process, which consists of two main steps, namely, training and feature selection, and validation. The purpose of the training and feature selection step is to derive a selected set of coordinates in the feature vector and to model their distribution in the resulting feature space. In the validation step, new feature vectors that were not used in training and feature selection are classified using the trained models in order to evaluate the classification performance of the system.

In addition to the feature extraction blocks that operate according to section 3, this block diagram includes several more blocks denoted by training, testing, performance analysis and feature selection. In the training blocks, room-volume classes are trained using the *Gaussian mixture model* (GMM) approach [20]. Each GMM is trained from a training set of speech data that is known to have been recorded in rooms with a typical volume. Hence, each GMM is trained to identify a single room-volume class of index i, with associated volume V_i. The GMMs are used later in the testing blocks, where feature vectors from different rooms are tested for the likelihood that they have been extracted from a similar room. In addition, a background model (BGM) [20] is trained for each room-volume class using speech data that was recorded in rooms with different volumes. The BGM of each room-volume class represents the hypothesis that the feature vector does not belong to the room-volume class. The GMMs and BGMs are trained using the *expectation maximization* (EM) [21] method.

In the testing blocks, a log-likelihood-ratio score function, $\Lambda_i(\mathbf{x})$, is applied for evaluating the likelihood of every feature vector, $\mathbf{x}$, to be from any room-volume class i, in the form of

$$\Lambda_i(\mathbf{x}) = \log p(\mathbf{x} \mid \mathrm{GMM}_i) - \log p(\mathbf{x} \mid \mathrm{BGM}_i) \quad (6)$$

Table II. Selected features using simulated RIRs. All frequencies are given in Hz.

	$\mathbf{x}_{\text{ref}}$	$\mathbf{x}_{\text{stop}}$	$\mathbf{x}_{\text{speech}}$	$\mathbf{x}_{\text{derev}}$	$\mathbf{x}_{\text{MFCC}}$
1	$MC_{\lvert\text{RTF}\rvert}[100, 400]$	RT	RT	$MC_{\lvert\text{RTF}\rvert}[225, 506]$	$\Delta\Delta c_3$
2	$K_{\lvert\text{RTF}\rvert}[100, 400]$	K_H	$MC_{\lvert\text{RTF}\rvert}[200, 800]$	K_h	$\Delta\Delta c_2$
3	EDT	K_h	$\sigma_{\lvert\text{RTF}\rvert}[144, 207]$	$\sigma_{\lvert\text{RTF}\rvert}[200, 800]$	Δc_{11}
4	$\sigma_{\lvert\text{RTF}\rvert}[67, 150]$	$MC_{\lvert\text{RTF}\rvert}[200, 800]$	$MC_{\lvert\text{RTF}\rvert}[100, 400]$	K_H	Δc_6
5	K_h	$\sigma_{\lvert\text{RTF}\rvert}[83, 120]$	K_h	$MC_{\lvert\text{RTF}\rvert}[430, 619]$	$\Delta\Delta c_{11}$
E_V	**34.03%**	**70.16%**	**83.52%**	**89.98%**	**118.62%**

where $p(\mathbf{x} \mid \text{GMM}_i)$ and $p(\mathbf{x} \mid \text{BGM}_i)$ are the likelihood that a feature vector $\mathbf{x}$ belongs or does not belong to a room-volume class i, respectively. Each feature vector is assigned to a value of a room volume using the maximum-score criterion, such that the estimated volume is

$$\widehat{V}(\mathbf{x}) = V_{\arg\max_i \Lambda_i(\mathbf{x})} \tag{7}$$

The performance analysis that is used yields the room-volume classification error given by

$$E_V = 100 \times E\left[\frac{\left|\widehat{V} - V\right|}{\frac{1}{2}\left(\widehat{V} + V\right)}\right] \tag{8}$$

which is based on the mean of the difference between the classified volume, $\widehat{V}$, and the true volume, V. This difference is normalized by the average between the classified and true volumes and the error is given as a percentage of the volume difference from that average. In this work, the GMMs and BGMs of the room-volume classes use a diagonal covariance matrix for every Gaussian component [20]. The GMMs contain two or three Gaussian components depending if the RIRs were simulated or measured, respectively. The BGMs always contain three Gaussian components. These models were selected after some preliminary experiments.

Feature selection is used to overcome potential overfitting that may be a result of the high dimensionality of the feature vectors. For this purpose, the *sequential floating forward selection* (SFFS) [22, 23, 24] feature selection algorithm is used, with E_V as an optimization criterion. Feature selection is performed on a development speech database, which was not used during training. In this work, five features were selected and it was shown that increasing the number of parameters does not improve the performance in terms of minimum E_V.

5. Experimental study using simulated RIRs

In this section, an experimental study is presented, in which an anechoic speech signal is convolved with simulated RIRs. The simulation of RIRs allows high resolution in room volumes such that room-volume estimation can be performed using classification to different room-volume classes. Furthermore, rooms with different geometries and absorption characteristics may be used for each room-volume class.

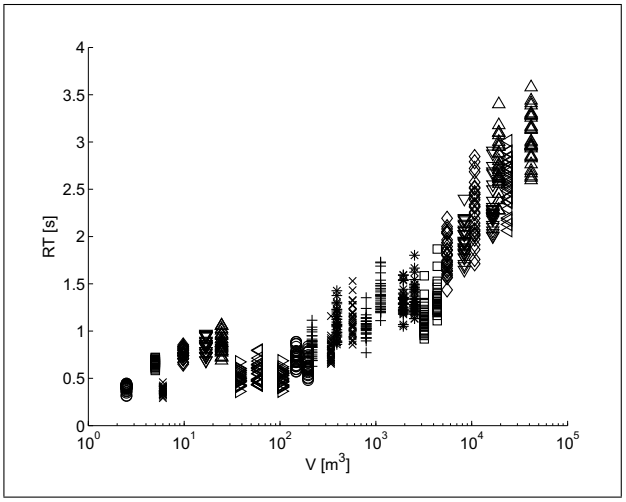


Figure 6. Scatter plot of simulated RIRs in the training database, showing room-volume vs. Schroeder RTs.

The image method of Allen and Berkley [25] was used to simulate RIRs from a large database of room-volume classes. Room dimensions were selected to mimic typical environments ranging from a phone booth and a car to lecture rooms and concert halls, under the shoe-box shape limitation. Three different room geometries were selected for each room-volume class such that the volume of the room is fixed, and such that there are no identical rooms in the different databases for training, development (for feature selection) and testing. For each geometry, 10 wall reflection coefficients were randomly chosen around a mean value that satisfies a predefined Sabine RT equal to either 0.5, 1, or 2s, depending on the volume. This formed a total of 30 RIRs for each room-volume class on each database. The Distance between the source and the receiver was made well above the radius of reverberation on each simulated RIR. Figure 6 shows the simulated RIRs in the training database on a scatter plot of their volumes with respect to their Schroeder RT, measured using the ISO 3382 standard [10].

Once the simulated RIRs were available, anechoic speech segments were convolved with them. Anechoic speech segments were taken from recordings performed during the PhD work of Papanagiotou [26]. These record-

Table III. Confusion matrix for room-volume classification using short words convolved with simulated RIRs.

Class Vol. [m^3]	True Vol. [m^3] 3	5	6	10	17	25	37	60	104	147
3	**53.3**	**21.7**	**23.3**	3.3	0	0	10.0	6.7	5.0	0
5	**11.7**	**33.3**	**11.7**	**18.3**	21.7	15.0	15.0	23.3	16.7	5.0
6	**8.3**	**1.7**	**33.3**	**0**	**10.0**	0	1.7	5.0	5.0	0
10	1.7	**11.7**	**1.7**	**28.3**	**21.7**	**30.0**	5.0	6.7	11.7	6.7
17	1.7	3.3	**5.0**	**23.3**	**16.7**	**20.0**	**0**	1.7	1.7	1.7
25	1.7	8.3	1.7	**8.3**	**11.7**	**3.3**	**5.0**	**6.7**	0	1.7
37	11.7	5.0	3.3	5.0	**1.7**	**1.7**	**20.0**	**15.0**	**16.7**	8.3
60	3.3	6.7	1.7	3.3	1.7	**5.0**	**13.3**	**5.0**	**16.7**	**5.0**
104	1.7	0	8.3	1.7	1.7	0	**0**	**6.7**	**10.0**	**6.7**
147	0	0	1.7	0	0	1.7	1.7	**0**	**1.7**	**0**
196	0	0	1.7	0	0	1.7	5.0	6.7	**11.7**	**11.7**
216	0	0	0	0	3.3	1.7	3.3	0	1.7	**6.7**
340	3.3	5.0	6.7	1.7	5.0	8.3	18.3	11.7	1.7	21.7
392	1.7	1.7	0	3.3	5.0	10.0	0	0	0	8.3
572	0	0	0	3.3	0	1.7	1.7	3.3	0	3.3
793	0	1.7	0	0	0	0	0	1.7	0	10.0
1120	0	0	0	0	0	0	0	0	0	0
1935	0	0	0	0	0	0	0	0	0	0
2529	0	0	0	0	0	0	0	0	0	0
3200	0	0	0	0	0	0	0	0	0	3.3
4370	0	0	0	0	0	0	0	0	0	0
5530	0	0	0	0	0	0	0	0	0	0
8370	0	0	0	0	0	0	0	0	0	0
10734	0	0	0	0	0	0	0	0	0	0
16612	0	0	0	0	0	0	0	0	0	0
19152	0	0	0	0	0	0	0	0	0	0
24080	0	0	0	0	0	0	0	0	0	0
41606	0	0	0	0	0	0	0	0	0	0

ings include a database of words according to the *diagnostic rhyme test* (DRT) [27] words list and a database of sentences according to the *diagnostic acceptability measure* (DAM) [28] sentences list. The DRT database was used to generate $\mathbf{x}_{\text{stop}}$, from which 32 short words were narrated by a male and a female talker, forming a total of 64 short words sections. The DAM database was used to generate $\mathbf{x}_{\text{derev}}$, $\mathbf{x}_{\text{speech}}$ and $\mathbf{x}_{\text{MFCC}}$, from which 32 sentences were narrated by a male and a female talker, forming a total of 64 sections of connected speech sentences. This formed a total of $30 \times 64 = 1920$ convolved speech sections for each room-volume class on each database. Different speakers were used for training, development and testing, always containing a single male talker and a single female talker.

Table II shows the selected features using the SFFS algorithm for each feature extraction method. The first column indicates the feature vectors, which were extracted directly from the RIRs as a reference. The last row indicates the room-volume classification error, E_V, defined in equation (8). All the frequencies in the rectangular brackets of room-volume features are given in Hz. It may be seen from Table II that the feature extraction approach with the best performance is using abrupt stops in speech, the second best is using whole sentences with connected speech, the third is the dereverberation approach and the last is using MFCCs.

The reverberation time was selected in $\mathbf{x}_{\text{stop}}$ and in $\mathbf{x}_{\text{speech}}$. The reason might be the ability of RT to give a block separation between small, medium and large rooms. However, RT was not selected in $\mathbf{x}_{\text{derev}}$, where the reason might be that the direct sound in the reconstructed RIR using reverberation is much stronger than the reconstructed reflections, as can clearly be seen from Figure 4. The Number of reflections was not selected for any of the approaches, while the kurtosis of the time domain signal was selected in all approaches.

Most of the selected features are calculated from low frequency bands of the RTF, showing a combination of STD ($\sigma_{|\text{RTF}|}[f_1, f_2]$) and mean-value crossings ($MC_{|\text{RTF}|}[f_1, f_2]$), in relatively low frequency bands. The kurtosis in a low frequency band ($K_{|\text{RTF}|}[f_1, f_2]$) appears only in the reference selected room-volume features. Although the number of modes ($N_{|\text{RTF}|}[f_1, f_2]$) was not selected at all, these results imply that the difference in the modal density of the RTFs between small and large rooms is likely to be noticeable by room-volume features that are extracted from low frequencies. The reason might be that at the low frequencies the shape of the RTF is more well-defined than at high frequencies, mainly as a result of sparser eigenmodes. The sparse pattern of the eigenmodes at low frequencies also allows the calculation of the Schroeder frequency. As a result, the kurtosis and STD

Table III. (cont.)

Class Vol. [m^3]	True Vol. [m^3] 196	216	340	392	572	793	1120	1935	2529	3200
3	0	1.7	1.7	0	0	0	0	0	0	0
5	10.0	10.0	8.3	1.7	8.3	0	0	0	0	0
6	1.7	0	0	0	0	0	0	0	0	0
10	1.7	3.3	5.0	0	3.3	0	0	0	0	0
17	1.7	3.3	1.7	1.7	1.7	1.7	0	0	0	0
25	3.3	1.7	0	16.7	0	1.7	0	0	0	1.7
37	6.7	1.7	11.7	0	0	0	0	0	0	1.7
60	13.3	1.7	5.0	0	1.7	0	0	0	0	0
104	**3.3**	0	3.3	0	0	0	0	0	0	0
147	**3.3**	**3.3**	1.7	1.7	0	0	0	0	0	0
196	**25.0**	**3.3**	**11.7**	0	1.7	0	0	0	0	0
216	**0**	**5.0**	**0**	**1.7**	1.7	1.7	0	1.7	0	1.7
340	**26.7**	**28.3**	**28.3**	**18.3**	**15.0**	6.7	3.3	6.7	3.3	11.7
392	0	**5.0**	**5.0**	**6.7**	**3.3**	**10.0**	3.3	1.7	0	0
572	1.7	8.3	**3.3**	**23.3**	**18.3**	**16.7**	**10.0**	5.0	1.7	10.0
793	1.7	16.7	8.3	**13.3**	**13.3**	**15.0**	**6.7**	**5.0**	15.0	8.3
1120	0	0	1.7	6.7	**20.0**	**11.7**	**25.0**	**13.3**	**20.0**	16.7
1935	0	1.7	0	0	1.7	**8.3**	**11.7**	**5.0**	**11.7**	**1.7**
2529	0	0	0	0	0	3.3	**3.3**	**6.7**	**5.0**	**0**
3200	0	5.0	3.3	5.0	1.7	18.3	11.7	**18.3**	**6.7**	**28.3**
4370	0	0	0	1.7	6.7	5.0	10.0	13.3	**13.3**	**10.0**
5530	0	0	0	1.7	1.7	0	11.7	21.7	20.0	**6.7**
8370	0	0	0	0	0	0	1.7	0	0	0
10734	0	0	0	0	0	0	0	0	0	1.7
16612	0	0	0	0	0	0	1.7	1.7	3.3	0
19152	0	0	0	0	0	0	0	0	0	0
24080	0	0	0	0	0	0	0	0	0	0
41606	0	0	0	0	0	0	0	0	0	0

may be more affected by the room-volume at low frequencies than at high frequencies, due to their relation to the Schroeder frequency. Furthermore, the more traceable shape of the RTF at the low frequencies may enable more reliable counting of mean-value crossings.

More detailed classification results are presented using the confusion matrix in Table III. The element in row i and column j in the confusion matrix indicates the percentage of feature vectors from a validation database of room-volume V_j, which were estimated to be extracted from a room with volume V_i. Hence higher values near the diagonal of the matrix indicate better classification results.

6. Experimental study using measured RIRs

In this section, room-volume classification is applied to speech signals which were convolved with measured RIRs. Tables IV–VI present the rooms in which the RIRs were recorded to generate training, development and validation databases, respectively. The RIRs in the offices, classrooms, teaching labs and auditoriums, were measured using a Brüel & Kjær 4295 Omni-Source loudspeaker and a Brüel & Kjær 4942 $^1/_2$-inch diffuse-field microphone, which were located differently for each RIR recording. The RIRs in the concert halls were taken from the *Concert Hall Research Group* (CHRG) database [29]. For every room in each database 30 measured RIRs were used. Anechoic speech segments were convolved with these RIRs to yield reverberant speech databases in a similar manner to that described in section 5, such that a total of 1920 convolved speech sections were available for each room-volume class on each database.

It may be seen from Tables IV–VI that the volumes of the rooms in the training database are of the same order of magnitude as in the development and validation databases. The reason for not using identical room volumes is that only physically existing venues are used. Hence, this experimental study shows how this method can be used to classify rooms with similar characteristics. However, the objective in this experiment remains to minimize volume classification error.

Table VII shows the selected features for each feature extraction method, with the classification error for each approach. It may be seen that the sorting order of the approaches according to their performance in classification is identical to the order observed when the simulated RIRs were used to convolve the anechoic speech with. However, the errors were lower in the case where the measured RIRs were used. This might seem surprising, but it should be mentioned that less room-volume classes were tested (6 compared to 17 with simulated RIRs) due to the low

Table III. (cont.)

Class Vol. [m^3]	True Vol. [m^3] 4370	5530	8370	10734	16612	19152	24080	41606
3	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0
25	0	0	0	0	0	0	0	0
37	0	0	0	0	0	0	0	0
60	0	0	0	0	0	0	0	0
104	0	0	0	0	0	0	0	0
147	0	0	0	0	0	0	0	0
196	0	0	0	0	0	0	0	0
216	0	0	0	0	0	0	0	0
340	1.7	0	0	0	0	0	0	0
392	0	0	0	0	0	0	0	0
572	1.7	1.7	0	0	0	0	0	0
793	5.0	0	0	0	0	0	0	0
1120	28.3	6.7	1.7	1.7	0	0	0	0
1935	6.7	0	0	0	1.7	0	0	0
2529	**0**	0	1.7	0	0	0	0	0
3200	**13.3**	**0**	0	0	0	0	0	0
4370	**20.0**	**1.7**	**0**	0	1.7	0	0	0
5530	**23.3**	**63.3**	**38.3**	**15.0**	30.0	0	5.0	0
8370	**0**	**16.7**	**31.7**	**13.3**	**15.0**	6.7	8.3	1.7
10734	0	**8.3**	**18.3**	**23.3**	**15.0**	**1.7**	6.7	0
16612	0	1.7	**6.7**	**6.7**	**13.3**	**6.7**	**1.7**	0
19152	0	0	0	**16.7**	**11.7**	**23.3**	**15.0**	**16.7**
24080	0	0	1.7	21.7	**11.7**	**36.7**	**40.0**	**28.3**
41606	0	0	0	1.7	0	**25.0**	**23.3**	**53.3**

Table IV. Rooms from which measured RIRs were used for training.

Class	Room	Average RT [s]	V [m^3]
T1	Building 33 Office 427 PhD Room	0.61	42
T2	Building 33 Classroom 102	0.60	147
T3	Building 33 Teaching lab 204	0.62	340
T4	Building 26 Auditorium 4	1.47	793
T5	Academy Hall (Philadelphia, PA)	1.38	$\sim$ 10700
T6	Kleinhans Music Hall (Buffalo, NY)	1.89	$\sim$ 18200

Table V. Rooms from which measured RIRs were used for development, and validation.

Class	Room	Average RT [s]	V [m^3]
D1	Building 33 Office 126	0.77	37
D2	Building 34 Classroom 103	0.65	120
D3	Building 34 Classroom 002	0.89	291
D4	Building 26 Auditorium 5	1.22	$\sim$ 1100
D5	Mechanics Hall (Worchester, MA)	2.28	$\sim$ 8300
D6	Boston Symphony Hall (Boston, MA)	2.67	$\sim$ 16600

resolution of the volume. As in section 5, in this experimental study the RT was selected in $\mathbf{x}_{\text{stop}}$ and in $\mathbf{x}_{\text{speech}}$, while the number of reflections was not selected at all. A combination of STD and mean-value crossings at the low frequency bands of the RTF is also selected in this experiment. Moreover, the kurtosis is not only selected in the reference case. Detailed classification results using the abrupt stop approach are presented using the confusion matrix in Table VIII.

Table VI. Rooms from which measured RIRs were used for validation.

Class	Room	Average RT [s]	V [m^3]
V1	Building 33 Office 404	0.45	40
V2	Building 34 Classroom 003	0.59	120
V3	Building 34 Classroom 202	1.09	301
V4	Building 26 Auditorium 6	1.28	~ 1100
V5	Troy Music Hall (Troy, NY)	2.73	~ 11300
V6	Joseph Mayerhoff Symphony Hall (Baltimore, MD)	2.37	~ 18700

Table VII. Selected features using measured RIRs. All frequencies are given in Hz.

	$\mathbf{x}_{\text{ref}}$	$\mathbf{x}_{\text{stop}}$	$\mathbf{x}_{\text{speech}}$	$\mathbf{x}_{\text{derev}}$	$\mathbf{x}_{\text{MFCC}}$
1	$MC_{\lvert\text{RTF}\rvert}[100, 400]$	RT	$\sigma_{\lvert\text{RTF}\rvert}[358, 516]$	$\sigma_{\lvert\text{RTF}\rvert}[249, 358]$	Δc_1
2	$\sigma_{\lvert\text{RTF}\rvert}[358, 516]$	$MC_{\lvert\text{RTF}\rvert}[207, 299]$	RT	$MC_{\lvert\text{RTF}\rvert}[200, 800]$	Δc_8
3	$\sigma_{\lvert\text{RTF}\rvert}[100, 400]$	K_h	$\sigma_{\lvert\text{RTF}\rvert}[100, 400]$	$\sigma_{\lvert\text{RTF}\rvert}[430, 619]$	c_7
4	$K_{\lvert\text{RTF}\rvert}[200, 800]$	$K_{\lvert\text{RTF}\rvert}[249, 358]$	$MC_{\lvert\text{RTF}\rvert}[225, 506]$	K_H	c_6
5	$\sigma_{\lvert\text{RTF}\rvert}[249, 358]$	$\sigma_{\lvert\text{RTF}\rvert}[299, 430]$	$K_{\lvert\text{RTF}\rvert}[173, 249]$	$MC_{\lvert\text{RTF}\rvert}[120, 173]$	c_3
E_V	**44.64%**	**61.09%**	**66.01%**	**67.99%**	**75.28%**

Table VIII. Confusion matrix for room-volume classification using short words convolved with measured RIRs. Ti and Vj indicate trained class Ti and true volume class Vj, respectively.

	V1	V2	V3	V4	V5	V6
T1	5.0	3.3	6.7	3.3	0	0
T2	63.3	70.0	73.3	26.7	0	0
T3	30.0	16.7	6.7	0	0	0
T4	1.7	10.0	13.3	70.0	0	0
T5	0	0	0	0	0	0
T6	0	0	0	0	100	100

7. Experimental study using recorded speech

In this section, the performance of the classification approach is further investigated using speech segments that were recorded in reverberant environments. The first four rooms in Tables IV–VI are the rooms where the reverberant speech was recorded to generate training, development and validation databases, respectively. Speech was recorded using a Brüel & Kjær 4942 $\frac{1}{2}$-inch diffuse-field microphone. In each database and for each room-volume class, 32 words from the DRT database were recorded from 5 different locations in the room (where the first author was the speaker), resulting in $32 \times 5 = 160$ $\mathbf{x}_{\text{stop}}$ feature vectors in each room-volume class on each database.

Table IX shows the selected features using the SFFS algorithm for the abrupt stop approach showing a classification error of $E_V = 71\%$. It is interesting to see that RT was not selected. Detailed classification results are presented using the confusion matrix in Table X.

8. Conclusion

In this work, the volume of a room was classified from reverberant speech signals. Different approaches of room-volume feature extraction from speech signals were presented. Experimental studies where anechoic speech was convolved with simulated and measured RIRs show that the best performance is achieved when features are extracted from speech signals that contain abrupt stops after single isolated words, which resemble a RIR decay. The second best performance was obtained by extracting room-volume features from single sentences that contain connected speech and a break, without trying to identify the RIR. For every feature extraction approach a feature selection procedure has shown that a combination of RT and a low frequency band of the RTF are essential for reasonable classification. The feature extraction approach that uses single words was tested using recorded speech and reasonable results were shown, although RT was not used.

Table IX. Selected features using recorded speech. All frequencies are given in Hz.

	$\mathbf{x}_{\text{stop}}$
1	$K_{\lvert\text{RTF}\rvert}[50, 200]$
2	$\sigma_{\lvert\text{RTF}\rvert}[299, 430]$
3	$\sigma_{\lvert\text{RTF}\rvert}[338, 759]$
4	K_h
5	$K_{\lvert\text{RTF}\rvert}[144, 207]$
E_V	71.00%

Table X. Confusion matrix for room-volume classification using recorded short words. Ti and Vj indicate trained class Ti and true volume class Vj, respectively.

	V1	V2	V3	V4
T1	38.1	20.3	6.2	6.9
T2	16.2	69.2	85.0	9.4
T3	18.8	5.3	2.5	10.6
T4	26.9	5.3	6.2	73.1

These results imply that features originally designed for room-volume classification from the RIR can be used for room-volume classification from reverberant speech.

It is worth mentioning that the volume of the room was strictly classified only in the case study where anechoic speech was convolved with simulated RIRs, while experiments using convolution with measured RIRs and using recorded speech may have classified rooms with similar characteristics. Hence, future work is proposed for recorded speech which includes a full and comprehensive experiment with various rooms and speakers. This will require a standard library of speech, which is recorded from plenty of speakers in plenty of rooms with various volumes, at least in the room-volume resolution that was used in the simulated RIRs case study.

Further future work may include room-volume identification using microphone array processing, as the approaches presented in this paper used a single microphone input. This may result in better system identification using dereverberation, e.g. when beamforming is performed.

Acknowledgment

We would like to thank Dr. Emanuël A.P. Habets for the use of his spectral compensation-based dereverberation method [18] and for supplying the code to execute it.

References

[1] P. Kendrick, T. J. Cox, Y. Zhang, J. A. Chambers, F. F. Li: Room acoustic parameter extraction from music signals. Proc. 2006 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 2006), Toulouse, France, 2006, 801–804.

[2] P. Kendrick, F. F. Li, T. J. Cox, Y. Zhang, J. A. Chambers: Blind estimation of reverberation parameters for non-diffuse rooms. Acta Acust. united Acust. **93** (Mai 2007) 760–770.

[3] M. Kuster: Reliability of estimating the room volume from a single room impulse response. J. Acoust. Soc. Am. **124** (Aug. 2008) 982–993.

[4] N. R. Shabtai, Y. Zigel, B. Rafaely: Room volume identification from room impulse response using statistical pattern recognition and feature selection. J. Acoust. Soc. Am. **128** (Sep. 2010) 1155–1162.

[5] N. R. Shabtai, Y. Zigel, B. Rafaely: Feature selection for room volume identification from room impulse response. Proc. 2001 IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA 2001), New Paltz, New York, USA, 2009, 249–252.

[6] N. R. Shabtai, Y. Zigel, B. Rafaely: Estimating the room volume from room impulse response via hypothesis verification approach. Proc. 2009 IEEE Workshop Stat. Signal Process. (SSP 2009), Cardiff, Wales, 2009, 717–720.

[7] N. R. Shabtai, Y. Zigel, B. Rafaely: Room volume classification from reverberant speech. Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC 2010), Tel-Aviv, Israel, 2010.

[8] H. Kuttruff: Room acoustics. Fourth ed. Spon Press, New York, 2000, Ch. 4, 89–114.

[9] L. H. An, S. E. Ahmed: Improving the performance of kurtosis estimator. Comp. Stat. Data Analysis **52** (Jan. 2008) 2669–2681.

[10] ISO 3382:1997: Acoustics – measurement of the reverberation time of rooms with reference to other acoustical parameters. 1997.

[11] L. E. Kinsler, A. R. Frey, A. B. Coppens, J. V. Sanders: Fundamentals of acoustics. Fourth ed. John Wiley, New York, 2000, Ch. 12, 333–358.

[12] D. Y. Maa: Distribution of eigenmodes in a rectangular chamber at low frequency range. J. Acoust. Soc. Am. **10** (1938) 235–238.

[13] M. R. Schroeder: Normal frequency and excitation statistics in rooms: Model experiments with electric waves. J. Audio Eng. Soc. **35** (may 1987) 307–316.

[14] I. Peer, B. Rafaely, Y. Zigel: Reverberation matching for speaker recognition. Proc. 2008 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 2008), Las Vegas, Nevada, USA, 2008, 4829–4832.

[15] T. J. Cox, F. Li, P. Darlington: Extracting room reverberation time from speech using artificial neural networks. J. Audio Eng. Soc. **49** (Apr. 2001) 219–230.

[16] F. F. Li, T. J. Cox: Speech transmission index from running speech: A neural network approach. J. Acoust. Soc. Am. **113** (Apr. 2003) 1999–2008.

[17] R. Ratnam, D. L. Jones, B. C. Wheeler, W. D. O. Jr., C. R. Lansing, A. S. Feng: Blind estimation of reverberation time. J. Acoust. Soc. Am. **114** (Nov. 2003) 2877–2892.

[18] E. Habets: Single- and multi-microphone speech dereverberation using spectral enhancement. Dissertation. Technische Universiteit Eindhoven, Eindhoven, The Netherlands, 2007.

[19] S. B. Davis, P. Mermelstein: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust., Speech, Signal Process. **ASSP-28** (Aug. 1980) 357–366.

[20] D. A. Reynolds, T. F. Quatieri, R. B. Dunn: Speaker verification using adapted Gaussian mixture models. Digital Signal Processing **10** (Jan. 2000) 19–41.

[21] A. P. Dempster, N. M. Laird, D. B. Rubin: Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc. B: Methodilogical **39** (1977) 1–38.

[22] P. Pudil, F. Ferri, J. Novovicova, J. Kittler: Floating search methods for feature selection with nonmonotonic criterion functions. Proc. 12th IAPR Int. Conf. Pat. Rec. (ICPR 1994), Jerusalem, Israel, 1994, 279–283.

[23] Y. Zigel: Feature selection for speaker recognition. Dissertation. Ben-Gurion University of the Negev, Beer Sheva, Israel, March 2004.

[24] Y. Zigel, A. Cohen: On feature selection for speaker verification. Proc. IEEE Odyssey 2004: The Speaker and Language Recognition Workshop, Toledo, Spain, 2004, 329–336.

[25] J. B. Allen, D. A. Berkley: Image method for efficiently simulating small room acoustics. J. Acoust. Soc. Am. **65** (Apr. 1979) 943–950.

[26] K. Papanagiotou: Enhancement of body-conducted speech from an ear-microphone. Dissertation. University of Southampton, Southampton, England, Oct. 2003.

[27] W. Voiers: Diagnostic evaluation of speech intelligibility. – In: Speech Intelligibility and Speaker Recognition. M. E. Hawley, 1977, Chap. 10, 374–387.

[28] W. Voiers: Diagnostic acceptability measure for speech communication systems. Proc. 1997 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 1997), 1977, 204–207.

[29] Concert hall research group CD v.3. Concert Hall Research Group, 327F Boston Post Road. Sudbury, MA 01776, 2004. Attention: Timothy J. Foulkes, email: chrg@cavtocci.com, phone: 978.443.7871, fax: 978.443.7873.