ROOM IDENTIFICATION USING FREQUENCY DEPENDENCE OF SPECTRAL DECAY STATISTICS

Alastair H. Moore, Patrick A. Naylor and Mike Brookes

Imperial College London
Electrical and Electronic Engineering, Exhibition Road, London

ABSTRACT

A method for room identification is proposed based on the reverberation properties of multichannel speech recordings. The approach exploits the dependence of spectral decay statistics on the reverberation time of a room. The average negative-side variance within 1/3-octave bands is proposed as the identifying feature and shown to be effective in a classification experiment. However, negative-side variance is also dependent on the direct-to-reverberant energy ratio. The resulting sensitivity to different spatial configurations of source and microphones within a room are mitigated using a novel reverberation enhancement algorithm. A classification experiment using speech convolved with measured impulse responses and contaminated with environmental noise demonstrates the effectiveness of the proposed method, achieving 79% correct identification in the most demanding condition compared to 40% using unenhanced signals.

***Index Terms*—** Room identification, Scene awareness, Reverberation, Classification, Microphone array

1. INTRODUCTION

Acoustic room identification is an emerging topic with many potential applications, including forensics [1], multimedia content labelling [2], robot navigation and location-aware voice interfaces. The aim is to be able to identify the specific room in which a speech recording was made given a limited number of previously seen rooms. A variety of features may be helpful including estimating the geometry [3, 4, 5, 6, 7], the background noise [2] and reverberation characteristics [1, 8].

Frequency-dependent reverberation time is largely independent of the source-microphone arrangement [9] and has shown promising discriminative power when calculated directly from measured acoustic impulse responses (AIRs) [1]. However, blindly estimating frequency-dependent reverberation time from reverberant speech is challenging and has, thus far, received little attention. In the recent Acoustic Characterization of Environments (ACE) Challenge [10, 11] there was only one submission which attempted it [12] (based on [13]) and it used a model to predict the low frequency values from high frequency values, rather than actually estimating them. Clearly such predictions do not add any additional information if the aim is to infer the identity of the room in which a recording was made.

Blind estimation of fullband reverberation time (RT) has received considerable attention [14, 15, 16, 17, 18, 19, 20, 21, 22, 23] due, predominantly, to its use in speech dereverberation [24, 25, 26]. The approach to RT estimation used in [21, 22, 23] exploits the observation that the distribution of spectral decay rates is dependent on the RT and the latter can be predicted by appropriate mapping of the negative-side variance (NSV) statistic. The NSV has also been used to predict the perceived level of reverberation [22], exploiting the fact that the NSV depends on the direct-to-reverberant ratio (DRR) as well as the RT. In [23], signals from 2 microphones were combined in such a way as to attenuate the coherent component of the signal, due to the direct path and early reflections. This was shown to reduce the dependence of the NSV on the source-microphone distance (or equivalently DRR) for synthesised AIRs under noise-free conditions.

The contributions of this work are to (i) propose room identification directly from frequency-dependent NSVs, (ii) propose a novel approach to estimating a blocking matrix which attenuates the direct path and early reflections, and (iii) demonstrate the benefit of the proposed enhancement to the proposed classification.

In Sec. 2 the concept of NSV is reviewed. Our proposed methods for room identification and reverberation enhancement are presented in Sec. 3 and Sec. 4, respectively, and evaluated in Sec. 5. Finally, conclusions are drawn in Sec. 6.

2. BACKGROUND TO NEGATIVE-SIDE VARIANCE

Expressed using a convolutive transfer function model [27, 28] in the short term Fourier transform (STFT) domain, the signal received at microphone m due to the source signal, $S(\ell, k)$, is

$$X_m(\ell, k) = \sum_{k'=0}^{K} \sum_{\ell'=-\infty}^{\infty} H_m(\ell', k, k') S(\ell - \ell', k) \qquad (1)$$

where $H_m(\ell', k, k')$ is the manifestation in the STFT domain of the transfer function from the source to microphone m, ℓ is the frame index and k is the frequency index. For simplicity, we assume that $H_m(\ell', k, k') = 0, \forall k \neq k'$.

The reverberant properties of a room are encapsulated in the AIR which Polack [29] modelled as a realisation of a zero-mean Gaussian random sequence multiplied by an exponential function. Extended to the convolutive transfer function model of (1), the energy decay can be modelled in terms of the power spectral density (PSD) of the direct path, $\beta_{m,d}(k)$, the PSD of the reverberation $\beta_{m,r}(k)$ and a frequency-dependent negative decay rate $\lambda_h(k)$ [27]

$$\mathbb{E}\left\{|H_m(\ell, k)|^2\right\} = \beta_{m,d}(k)\delta_{0\ell} + \beta_{m,r}(k)u(\ell - 1)e^{\lambda_h(k)\ell} \qquad (2)$$

where $\mathbb{E}\{\cdot\}$ is the expectation operator, $\delta_{0\ell}$ is the Kronecker delta function, which is 1 when $\ell = 0$ and 0 otherwise, and $u(\cdot)$ is the Heaviside step function.

This work was supported by the Engineering and Physical Sciences Research Council [grant number EP/M026698/1].

Considering the case when the DRR is low, i.e. $\beta_{m,d}(k) \ll \beta_{m,r}(k)$ the Heaviside function suggests that the reverberant signal $X_m(\ell,k)$ will quickly track the increase in energy at speech onsets but the corresponding decrease in energy at speech offsets will be determined by $\lambda_h(k)$ [21]. More formally, if $\lambda_s(\ell,k)$ is the negative decay rate of $S(\ell,k)$,

$$\lambda_x(\ell,k) \approx \max\{\lambda_h(k), \lambda_s(\ell,k)\} \tag{3}$$

is the negative decay rate of the reverberant speech. Fitting a straight line to the logarithm of an exponentially decaying sequence, the gradient is proportional to the negative decay rate. Positive gradients, due to speech onsets, track the increases in speech energy whilst negative gradients are limited by the room. Gradients which are more negative correspond to faster decays and smaller RTs. Since the distribution of gradients is skewed, and only the negative part is related to the reverberation, the negative part of the distribution is used to determine the NSV from which a good estimate of the RT can be obtained [21].

However, as observed in [23], when the DRR is high, i.e. $\beta_{m,d}(k) \gg \beta_{m,r}(k)$ the convolution of (1) is dominated by the Kronecker delta function and so the approximation in (3) is no longer valid. Large negative gradients can occur, regardless of the RT. It has been shown that attenuating the early reflection as well as the direct path through decorrelation leads to NSV values which are less dependent on the DRR and than those obtained from the microphone signals directly [23].

3. ROOM IDENTIFICATION USING NSV

Room identification is proposed in the following, based on frequency-dependent NSV features.

3.1. Feature extraction

Let $\mathbf{a}(\ell,k) = \left[A_1(\ell,k)\ A_2(\ell,k)\ \ldots\ A_N(\ell,k)\right]^T$, where $(\cdot)^T$ is the transpose operator, be an N channel reverberant speech signal. The negative decay rate, $\lambda_{a,n}(r,k)$ is the gradient of $\ln\{|A_n(\ell,k)|\}$ over frames $\ell = \{rR, rR+1, \ldots, rR+L_\lambda\}$, where L_λ and R determine the interval over which decays are estimated and the increment between successive estimates, respectively. The set of all negative-valued gradients in STFT frequency bin k across all channels is denoted $\boldsymbol{\lambda}(k)$. That is $\boldsymbol{\lambda}(k) = \{\lambda_{a,n}(r,k) : \lambda_{a,n}(r,k) < 0, \forall r, \forall n\}$. The NSV for STFT frequency bin k is $\zeta(k) = var\{\boldsymbol{\lambda}(k)\}$. Averaging across STFT frequency bins, the mean NSV within the j^{th} subband is denoted ψ_j and $\boldsymbol{\psi} = \left[\psi_1\ \psi_2\ \ldots \psi_J\right]^T$ is the proposed feature vector used for room identification.

3.2. Classifier

Each utterance in the dataset is represented by a J-element feature vector, $\boldsymbol{\psi}$ and is labelled according to one of I rooms. A subset of the utterances belonging to the i^{th} room are used to fit the parameters of a J-dimensional multi-variate Gaussian distribution with mean $\boldsymbol{\mu}_i = [\mu_{i,1}, ..., \mu_{i,J}]^T$ and diagonal covariance matrix $\boldsymbol{\Sigma}_i = \text{diag}\left(\sigma^2_{i,1}, ..., \sigma^2_{i,J}\right)$.

Assuming the NSVs in different frequency bands are uncorrelated, the probability that an unseen utterance with feature vector $\boldsymbol{\psi}' = \left[\psi'_1\ \psi'_2\ \ldots \psi'_J\right]^T$ was observed in a room with mean $\boldsymbol{\mu}_i$ and covariance $\boldsymbol{\Sigma}_i$ is given by

$$p(\boldsymbol{\psi}'|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \prod_{j=1}^{J} p(\psi'_j|\boldsymbol{\mu}_i, \sigma^2_{i,j}) \tag{4}$$

$$= \prod_{j=1}^{J} \frac{1}{\sigma_{i,j}\sqrt{2\pi}} e^{\left(-\frac{\psi'_j - \mu_{i,j}}{2\sigma^2_{i,j}}\right)}. \tag{5}$$

For the closed-set room identification task considered here, the room, i, which maximises $p(\boldsymbol{\psi}'|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ is chosen.

4. PROPOSED ENHANCEMENT

In order to reduce the effect of the direct path and early reflections on the NSV, we propose to design blocking filters based on an explicit estimate of the the relative transfer function of the direct path and early reflections.

The observed signal, $Y_m(\ell,k)$, at the m^{th} microphone in an array is composed of the reverberant speech, $X_m(\ell,k)$, as defined in (1), and noise, $V_m(\ell,k)$, which is uncorrelated with $X_m(\ell,k)$. Since the reverberation enhancement is performed independently at each frequency, the dependence on k is dropped for the remainder of this section. The signals received by an array of M microphones are

$$\mathbf{y}(\ell) = \mathbf{x}(\ell) + \mathbf{v}(\ell) \tag{6}$$

where $\mathbf{y}(\ell) = \left[Y_1(\ell)\ Y_2(\ell)\ \ldots\ Y_M(\ell)\right]^T$ and $\mathbf{x}(\ell)$ and $\mathbf{v}(\ell)$ are similarly defined.

From (1), let the early component of X_m be given by

$$X_m^{(e)}(\ell) = H_m(0)S(\ell) \tag{7}$$

where $H_m(0)$ represents the early transfer function which includes the direct path and earliest arriving reflections. The number of reflections depends on the geometry of the room, source-microphone configuration and STFT frame length. During speech onsets, before late reflections start to arrive, $\mathbf{y}(\ell)$ is assumed to consist of only the early component and noise

$$\mathbf{y}^{(e)}(\ell) = \mathbf{x}^{(e)}(\ell) + \mathbf{v}(\ell) \tag{8}$$

$$= \mathbf{h}^{(e)}S(\ell) + \mathbf{v}(\ell) \tag{9}$$

where $\mathbf{h}^{(e)} = \left[H_1(0)\ H_2(0)\ \ldots\ H_M(0)\right]^T$. Without loss of generality, taking $m = 1$ as the reference microphone, the relative early transfer function is

$$\mathbf{g} = \left[1\ \frac{H_2(0)}{H_1(0)}\ \ldots\ \frac{H_M(0)}{H_1(0)}\right]^T \tag{10}$$

and

$$\mathbf{x}^{(e)}(\ell) = \mathbf{g}X_1(\ell). \tag{11}$$

The following PSD matrices are defined $\boldsymbol{\Phi}_y = \mathbb{E}\left\{\mathbf{y}\mathbf{y}^H\right\}$, $\boldsymbol{\Phi}_{y^{(e)}} = \mathbb{E}\left\{\mathbf{y}^{(e)}(\mathbf{y}^{(e)})^H\right\}$, $\boldsymbol{\Phi}_{x^{(e)}} = \mathbb{E}\left\{\mathbf{x}^{(e)}(\mathbf{x}^{(e)})^H\right\}$ and $\boldsymbol{\Phi}_v = \mathbb{E}\left\{\mathbf{v}\mathbf{v}^H\right\}$, where $(\cdot)^H$ is the conjugate transpose operator.

Using (11), the PSD matrix of the noisy received signal during speech onsets is

$$\boldsymbol{\Phi}_{y^{(e)}} = \phi_{x_1{}^{(e)}}\mathbf{g}\mathbf{g}^H + \boldsymbol{\Phi}_v \tag{12}$$

Room	RT [s]
Lecture Room 1	0.64
Lecture Room 2	1.25
Meeting Room 1	0.44
Meeting Room 2	0.37
Office 2	0.48

Table 1. Fullband RT for each room in dataset, determined from the measured AIRs [11].

where ${\phi_{x_1}}^{(e)} = \mathbb{E}\left\{{X_1}^{(e)}(\ell){X_1}^{(e)}(\ell)^*\right\}$ and $(\cdot)^*$ denotes the complex conjugate operator. Since ${\mathbf{\Phi}_x}^{(e)}$ is rank 1 by definition, during time intervals when there is no late reverberation, the generalized eigenvalue (GEV) solution to the matrix pencil $({\mathbf{\Phi}_y}^{(e)}, \mathbf{\Phi}_v)$ has only one generalized eigenvalue greater than 1 [30, 31]. The corresponding eigenvector, $\mathbf{f}$, is a scaled and rotated version of the desired relative early transfer function, which can be recovered by the normalization

$$\mathbf{g} = \frac{\mathbf{\Phi}_v \mathbf{f}}{\mathbf{e}_1 \mathbf{\Phi}_v \mathbf{f}} \tag{13}$$

where $\mathbf{e}_1 = \begin{bmatrix}1 & 0 & \dots 0\end{bmatrix}$.

We propose to estimate ${\mathbf{\Phi}_y}^{(e)}$ as the ensemble mean of $\mathbf{y}(\ell)\mathbf{y}(\ell)^H$ over only those time intervals which are deemed to coincide with speech onsets where the level of residual reverberation is comparatively low. To identify suitable intervals, a time-varying estimate of $\mathbf{\Phi}_y(\ell)$ is obtained using recursive smoothing as

$$\hat{\mathbf{\Phi}}_y(\ell) = \alpha\hat{\mathbf{\Phi}}_y(\ell-1) + (1-\alpha)\mathbf{y}(\ell)\mathbf{y}(\ell)^H. \tag{14}$$

The covariance, i.e. off diagonal, terms of $\hat{\mathbf{\Phi}}_y(\ell)$ are a measure of the similarity between pairs of microphone signals that includes scaling due to the source power. Rising edges in the cross-PSD terms are therefore associated with speech onsets. To best satisfy the condition that the level of reverberation is low compared to the direct path and early reflections, a set of candidates frames are identified as the peaks in the cross-PSD. A candidate peak with level B is retained if it satisfies two criteria: i) Energy — B must be above the γ^{th} percentile across all the candidates; and ii) Onset — no other peaks that lie within the preceding τ s can exceed ωB. Having identified the locations of peaks associated with the end of speech onsets, the frames in the ρ s immediately preceding these peaks are included in the estimation of ${\mathbf{\Phi}_y}^{(e)}$.

The noise-only PSD matrix, $\mathbf{\Phi}_v$ is estimated as the ensemble mean of $\mathbf{y}(\ell)\mathbf{y}(\ell)^H$ over the first 100 frames of a recording.

Having obtained $\mathbf{g}$, $M-1$ enhanced reverberant signals are obtained [32]

$$Z_{m-1}(\ell) = Y_m(\ell) - G_m Y_1(\ell), \quad m \neq 1 \tag{15}$$

where G_m is the m^{th} element of $\mathbf{g}$.

5. EVALUATION

5.1. Test setup

Noisy reverberant recordings are taken from the evaluation dataset of the ACE challenge [11]. Longform recordings, denoted 's4', for the cruciform array ($M = 5$) in ambient noise with a signal-to-noise ratio of 18 dB are each segmented into 4 non-overlapping 8 s sections. The resulting dataset contains 5 rooms × 2 positions (i.e. source-microphone configurations) × 10 speakers (5 male, 5 female) × 4 utterances = 400 utterances. The fullband RT of the five rooms are listed in Table 1 and lie between 0.37 and 1.25 s.

	No processing	Baseline	Proposed
Train 1	80.5	**92.5**	90.5
Train 2	43.3	73.3	**80.3**
Train 3	40.0	70.0	**79.0**

Table 2. Percentage of correctly classified samples by training condition and reverberation enhancement method.

For each utterance, three alternative approaches to reverberation enhancement are considered. With 'No processing' applied, decay rates are estimated in each STFT band of all $N = M$ channels. The 'Baseline' approach applies the decorrelation method from [23] and the 'Proposed' approach, is as described in Sec. 4. The baseline and proposed approaches both result in $N = M - 1$ enhanced signals.

For each reverberation enhancement approach the N channels are used to obtain a subband NSV feature vector, as described in Sec. 3.1. Considering ⅓-octave bands centered between 200 Hz and 4 kHz gives a total of $J = 14$ features.

Room identification was performed using the classifier described in Sec. 3.2. Performance is expected to depend on how well matched the training and testing conditions are. Since the aim is to identify the specific room, of course it is necessary to include recordings from the test room in the training set. Using a cross-validation approach, every utterance in the dataset is classified by first retraining the classifier's models using an appropriate subset of the data, which, minimally, does not include the tested utterance. We define three training conditions with increasing disparity between the training and testing conditions in terms of those utterances which are excluded from the training set. 'Train 1': No utterances from the same speaker at the same position, 'Train 2': No utterances from any speaker at the same position, and 'Train 3': No utterances from any speaker at the same position and no utterances from the same speaker at any position in any room. In this way, 'Train 3' is the most demanding case since the tested speaker and the tested position are completely unseen.

5.2. Results and discussion

Classification results are presented in Table 2. In the easiest case, 'Train 1', where the training data includes utterances from the same position as the test data, excellent results are obtained as could be expected with both baseline and the proposed enhancement methods (92.5% and 90.5%) correct, respectively, while classification based on the unenhanced signals is substantially worse (80.5% correct). The nature of the misclassifications can be seen in Fig. 1(a-c). In all but one case, Lecture Room 2 is correctly identified. From the fullband RT, this is clearly the most distinct room. The most frequently confused rooms are Meeting Room 1 and Office 2 which have very similar fullband RTs of 0.44 s and 0.48 s, respectively. It is remarkable that such similar rooms can be separated at all.

Performance in 'Train 2' and 'Train 3' conditions is somewhat degraded for all enhancement methods, reflecting the more challenging classification task. However, the proposed method is least affected, with a correct identification rate 37–39 percentage points higher than the unprocessed case and 7–9 percentage points higher than baseline method. This suggests that the distributions of NSV

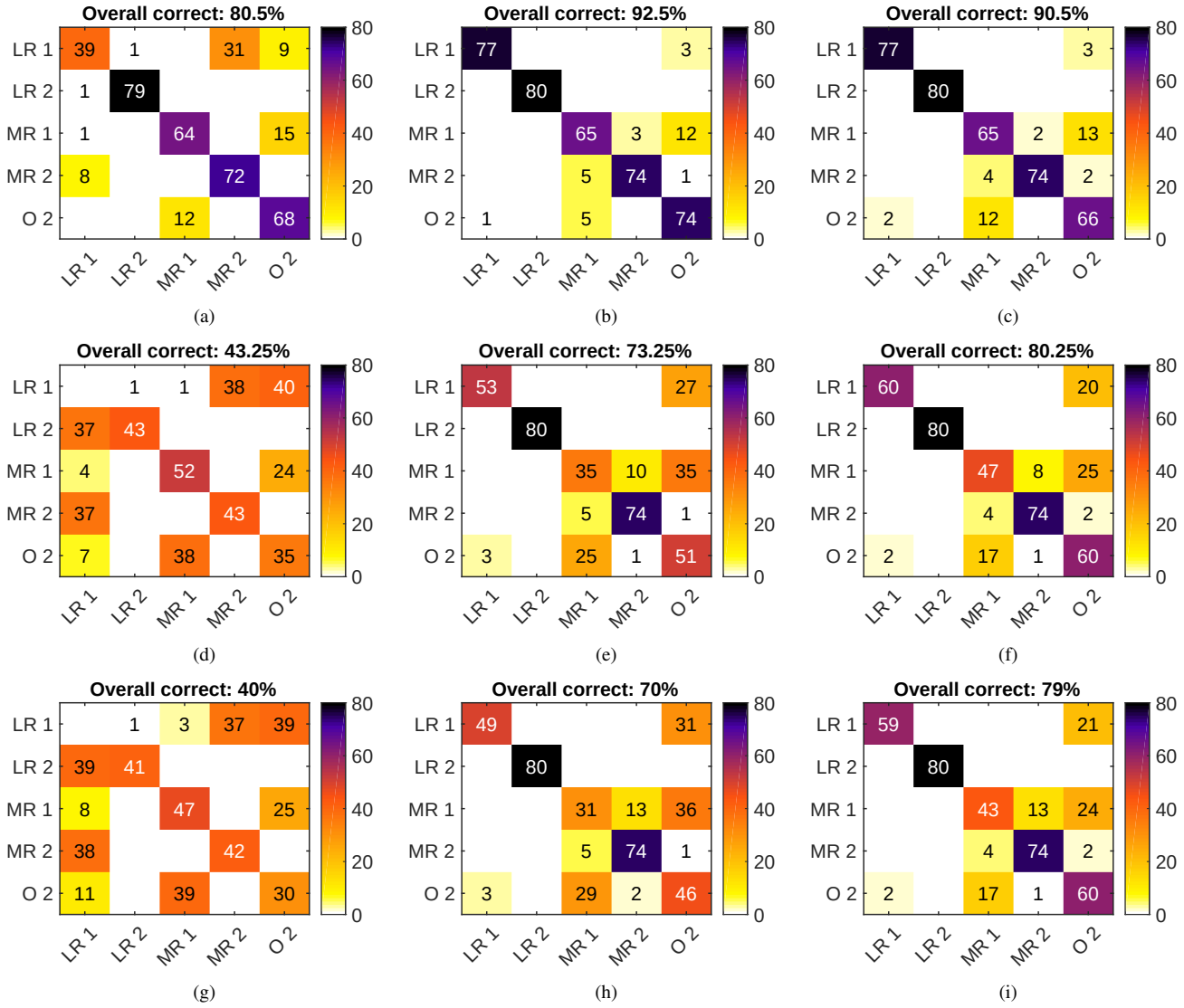

Fig. 1. Confusion matrix for classification using (a,b,c) 'Train 1', (d,e,f) 'Train 2' and (g,h,i) 'Train 3' training conditions with (a,d,g) no processing, (b,e,h) baseline enhancement [23] and (c,f,i) proposed enhancement of microphone signals.

features in each room are more separable when the proposed reverberation enhancement is used. Comparing the confusion matrices in Fig. 1(e) and (f) for 'Train 2' and Fig. 1(h) and (i) for 'Train 3', the specific rooms which are confused are the same but the proposed method is better able to discriminate between Lecture Room 1, Meeting Room 1 and Office 2, which are the rooms with mid-range fullband RTs.

6. CONCLUSIONS

A method for identifying the room in which a multichannel speech recording was made based on subband NSV features has been proposed. A novel method for reverberation enhancement based on estimating the relative early transfer function during speech onsets was also proposed. In the most demanding condition, it was found that the accuracy of room identification was better when using the proposed enhancement (79% correct) than with either a baseline enhancement method (70% correct) or no enhancement (40% correct).

7. REFERENCES

[1] A. H. Moore, M. Brookes, and P. A. Naylor, "Room-prints for forensic audio," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2013. [Online]. Available: http://www.commsp.ee.ic.ac.uk/~sap/uploads/publications/Moore2013a.pdf

[2] N. Peters, H. Lei, and G. Friedland, "Name that room: Room identification using acoustic features in a recording," in *Proceedings of the 20th ACM International Conference on Multimedia*, 2012, pp. 841–844.

[3] F. Antonacci, J. Filos, M. R. P. Thomas, E. A. P. Habets, A. Sarti, P. A. Naylor, and S. Tubaro, "Inference of room geometry from acoustic impulse responses," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 10, pp. 2683–2695, Dec. 2012.

[4] N. R. Shabtai, Y. Zigel, and B. Rafaely, "Feature selection for room volume identification from room impulse response," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2009, pp. 249–252.

[5] ——, "Room volume identification from reverberant speech," in *Proc. Intl. on Workshop Acoust. Echo and Noise Control (IWAENC)*, 2010.

[6] I. Dokmanic, Y. Lu, and M. Vetterli, "Can one hear the shape of a room: The 2-D polygonal case," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, May 2011, pp. 321–324.

[7] I. Dokmanić, R. Parhizkar, A. Walther, Y. M. Lu, and M. Vetterli, "Acoustic echoes reveal room shape," *Proc. National Academy of Sciences*, 2013.

[8] C. Papayiannis, C. Evers, and P. A. Naylor, "Discriminative feature domains for reverberant acoustic environments," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, Louisiana, USA, Mar. 2017, pp. 756–760.

[9] X. Pelorson, J.-P. Vian, and J.-D. Polack, "On the variability of room acoustical parameters: Reproducibility and statistical validity," *Applied Acoustics*, vol. 37, pp. 175–198, 1992.

[10] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "The ACE Challenge - corpus description and performance evaluation," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2015.

[11] ——, "Estimation of room acoustic parameters: The ACE challenge," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 10, pp. 1681–1693, Oct. 2016.

[12] H. W. Löllmann, A. Brendel, P. Vary, and W. Kellermann, "Single-channel maximum-likelihood T_{60} estimation exploiting subband information," in *Proc. ACE Challenge Workshop, a satellite of IEEE-WASPAA*, New Paltz, NY, USA, Oct. 2015.

[13] H. Löllmann and P. Vary, "Estimation of the frequency dependent reverberation time by means of warped filter-banks," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, May 2011, pp. 309–312.

[14] T. J. Cox, F. Li, and P. Darlington, "Extracting room reverberation time from speech using artificial neural networks," *J. Audio Eng. Soc. (AES)*, vol. 49, no. 4, pp. 219–230, 2001.

[15] R. Ratnam, D. L. Jones, B. C. Wheeler, W. D. O'Brien, C. R. Lansing, and A. S. Feng, "Blind estimation of reverberation time," *J. Acoust. Soc. Am.*, vol. 114, no. 5, pp. 2877–2892, Nov. 2003.

[16] M. Wu and D. Wang, "A pitch-based method for the estimation of short reverberation time," *Acta Acustica united with Acustica*, vol. 92, no. 2, pp. 337–339, Apr. 2006.

[17] H. W. Löllmann and P. Vary, "Estimation of the reverberation time in noisy environments," in *Proc. Intl. on Workshop Acoust. Echo and Noise Control (IWAENC)*, Sept. 2008, pp. 1–4.

[18] H. W. Löllmann, E. Yilmaz, M. Jeub, and P. Vary, "An improved algorithm for blind reverberation time estimation," in *Proc. Intl. on Workshop Acoust. Echo and Noise Control (IWAENC)*, Tel-Aviv, Israel, Aug. 2010, pp. 1–4.

[19] T. d. 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," *J. Acoust. Soc. Am.*, vol. 131, no. 4, pp. 2811–2816, Apr. 2012.

[20] P. Murgai, M. Rau, and J.-M. Jot, "Blind estimation of the reverberation fingerprint of unknown acoustic environments," in *Proc. Audio Eng. Soc. (AES) Convention*. Audio Engineering Society, Oct. 2017, preprint 9905. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=19302

[21] 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 *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, USA, Apr. 2008, pp. 329–332.

[22] J. Eaton, N. D. Gaubitch, and P. A. Naylor, "Noise-robust reverberation time estimation using spectral decay distributions with reduced computational cost," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013, pp. 161–165.

[23] B. Dumortier and E. Vincent, "Blind RT60 estimation robust across room sizes and source distances," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 5187–5191.

[24] K. Lebart, J. M. Boucher, and P. N. Denbigh, "A new method based on spectral subtraction for speech de-reverberation," *Acta Acoustica*, vol. 87, pp. 359–366, 2001.

[25] E. A. P. Habets, "Single- and multi-microphone speech dereverberation using spectral enhancement," Ph.D. dissertation, Technische Universiteit Eindhoven (TU/e), 2007.

[26] P. A. Naylor and N. D. Gaubitch, Eds., *Speech Dereverberation*. Springer-Verlag, 2010.

[27] E. A. P. Habets, S. Gannot, and I. Cohen, "Late reverberant spectral variance estimation based on a statistical mode," *IEEE Signal Process. Lett.*, vol. 16, no. 9, pp. 770–774, Sept. 2009.

[28] R. Talmon, I. Cohen, and S. Gannot, "Convolutive transfer function generalized sidelobe canceler," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 7, pp. 1420–1434, Sept. 2009.

[29] J.-D. Polack, "La transmission de l'énergie sonore dans les salles," Ph.D. dissertation, Dec. 1988.

[30] S. Markovich, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 6, pp. 1071–1086, Aug. 2009.

[31] M. Taseska and E. A. P. Habets, "Relative transfer function estimation exploiting instantaneous signals and the signal subspace," in *Proc. European Signal Processing Conf. (EUSIPCO)*, 2015.

[32] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 451–459, Sept. 2004.