SINGLE-CHANNEL BLIND ESTIMATION OF REVERBERATION PARAMETERS

Clement S. J. Doire[⋆], *Mike Brookes*[⋆], *Patrick A. Naylor*[⋆],
Dave Betts[†], *Christopher M. Hicks*[†], *Mohammad A. Dmour*[†], *Søren Holdt Jensen*[*]

[⋆]Electrical and Electronic Engineering, Imperial College London, UK.
[†]CEDAR Audio Ltd., Fulbourn, UK.
[*]Department of Electronic Systems, Aalborg University, Denmark.

ABSTRACT

The reverberation of an acoustic channel can be characterised by two frequency-dependent parameters: the reverberation time and the direct-to-reverberant energy ratio. This paper presents an algorithm for blindly determining these parameters from a single-channel speech signal. The algorithm uses an extended Kalman filter to estimate the parameters together with a hidden semi-Markov model to identify intervals of speech activity.

Index Terms— Single-channel, Reverberation time, Direct-to-Reverberant Ratio.

1. INTRODUCTION

Combating the damaging effects of reverberation has been a key research topic in recent years due to the increasing demand for effective methods of speech communication in real life scenarios [1]. While some progress has been made in both single- and multi-channel processing [2], the task of providing a blind single-channel dereverberation method suitable for real-time processing remains a challenge. Within this context, blind estimation of the room parameters that govern reverberation is an important prerequisite.

From the work of Schroeder [3] and Polack [4, 5], the acoustic reverberation in a room can be well characterised using two parameters: the reverberation time, T_{60}, and the Direct-to-Reverberant energy Ratio, DRR. T_{60} equals the time for the reverberant energy due to an acoustic impulse to decay by 60 dB while the DRR gives the ratio between the energy received via the direct path from source to microphone and the integrated reverberant energy. A number of researchers have proposed methods of estimating T_{60} and/or DRR from a recording of an unknown acoustic source, most commonly a speech signal. In [6] the sound decay is modelled as a stochastic process and an estimate of T_{60} is obtained using a maximum-likelihood estimator. Based on Polack's statistical model, the method originally presented in [7] and developed in [8] looks at the distribution of the decay rates and estimates T_{60} from the variance of its negative gradients. A single-channel estimator for both T_{60} and DRR was proposed in [9] by means of studying the short- and long-term temporal dynamics, obtained through the differential cepstral coefficients and the modulation spectrum respectively.

The goal of the present algorithm is to estimate blindly from a single-channel noisy reverberant speech signal frequency-dependent values of both T_{60} and DRR. We use an autoregressive reverberation model within an extended Kalman filter, the latter being conditioned by a semi-Markov model detecting speech activity.

The paper's structure is as follows: in the next section we detail the basic notation and concepts used throughout the paper as well as our reverberation model. In Sec. 3, a general description of the whole system as well as details of the EKF operation and speech activity detection are presented. In Sec. 4 experimental results are shown before concluding in Sec. 5.

2. PROBLEM STATEMENT

In the present work, a Short Time Fourier Transform (STFT) is applied to the noisy reverberant signal and the processing is done in the time-frequency domain. Let the speech, reverberation and noise power in frequency bin k of time frame l be given respectively by $S(l,k)$, $R(l,k)$, $N(l,k)$. Assuming uncorrelated powers are additive, we have for the total power

$$Y(l,k) = V(l,k)\,S(l,k) + R(l,k) + N(l,k) \qquad (1)$$

where V is a binary-distributed switch on S. We assume S, R and N follow Generalized Gamma distributions with shape parameters $\gamma = 1$ and κ_S, κ_R, κ_N respectively. Thus, using

$$p_{GG}(x;\gamma,\kappa,\theta) \triangleq \frac{\gamma}{\theta\,\Gamma(\kappa)} \left(\frac{x}{\theta}\right)^{\gamma\kappa-1} e^{-\left(\frac{x}{\theta}\right)^{\gamma}} \qquad (2)$$

we have $p(s(l)) = p_{GG}\left(s;1,\kappa_S,\frac{\mu_S}{\kappa_S}\right)$ and similarly for R and N. All frequency bins are processed independently, therefore the k index will be omitted in the remainder of the paper. Uppercase letters represent random variables, the corresponding lower case letters their realisations, and the mean

The research leading to these results has received funding from the EU 7th Framework Programme (FP7/2007-2013) under grant agreement ITN-GA-2012-316969.

of their distributions are denoted using the symbol μ associated with the corresponding subscript letter. Hatted symbols denote estimated quantities.
We model the reverberation in an enclosed space using the following autoregressive model:

$$R(l) = \sum_{\tau=1}^{+\infty} f\, V(l-\tau) S(l-\tau) \alpha^{\tau-1}. \quad (3)$$

Two parameters appear in this equation: the decay constant, α, and the energy drop, f. Both are frequency-bin dependent and are related to the more conventional T_{60} and DRR through the invertible equations DRR $= \frac{1-\alpha}{f}$ and $T_{60} = \frac{-6T}{\log_{10}(\alpha)}$ where T is the STFT frame increment.

3. PROPOSED METHOD

3.1. System Description

A block diagram of the proposed algorithm is shown in Fig. 1. For each frequency bin, the noisy reverberant speech power in frame l, $y(l)$ forms the input to a hidden semi-Markov model (HSMM) which estimates the speech presence indicator, $V(l)$. The signal $y(l)$ also forms the input to an Extended Kalman Filter (EKF) that estimates the parameter state vector, $\hat{\mathbf{x}}(l)$ which comprises estimates of μ_S, μ_R, μ_N, α and f. Computation of the HSMM observation probabilities is described below. Since they depend on the EKF state, the value of $V(l)$ cannot be reliably determined at time l. Accordingly, the HSMM keeps track of the n-best state sequences with the highest likelihoods and a separate EKF is used for each. In the following, $\mathcal{O}_l$ is one such state sequence up to time frame l in the frequency bin under consideration.

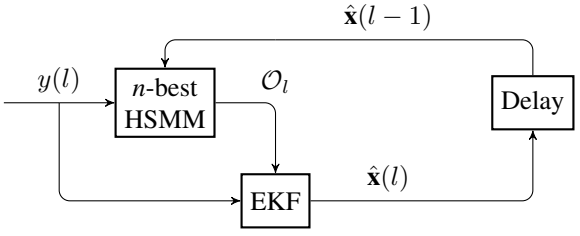

Fig. 1: Block diagram of the HSMM-conditioned EKF system.

3.2. Speech Activity Detection

In order to estimate $V(l)$, a 2-state HSMM is used. This speech activity detector conditions the EKF operation.
To get the probability of each path through this 2-state HSMM, we first need to derive the *a posteriori* Speech Presence Probability (SPP). Let $\mathcal{V}_1$ and $\mathcal{V}_0$ be the states corresponding to $V(l) = 1$ and 0 respectively. The posterior

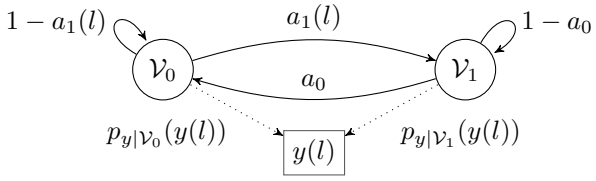

Fig. 2: Semi-continuous transition probability HSMM with 2 states. a_0, etc. represent the transition probabilities. $p_{y|\mathcal{V}_j}(y(l))$ is the likelihood of emitting observed energy $y(l)$ from state $\mathcal{V}_j$.

probability of having speech active at time frame l is

$$P(\mathcal{V}_1|y(l), \mathcal{O}_{l-1}) = \frac{P(\mathcal{V}_1|V(l-1))\, p_{y|\mathcal{V}_1}(y(l))}{P(\mathcal{V}_0|V(l-1))\, p_{y|\mathcal{V}_0}(y(l)) + P(\mathcal{V}_1|V(l-1))\, p_{y|\mathcal{V}_1}(y(l))} \quad (4)$$

with $p_{y|\mathcal{V}_j}(y(l))$ the likelihood of emitting observed energy $y(l)$ in state $\mathcal{V}_j$, $j \in \{0; 1\}$. Because of the Markov assumption, the prior probabilities $P(\mathcal{V}_j|V(l-1))$ only depend on the previous state and correspond to the a_j of Fig. 2.
In order to reduce false alarms, we use a dynamic scheme for the transition probabilities $a_1(l)$, in a similar fashion to the semi-Markov model described in [10].

$$a_1(l+1) = \begin{cases} \max\left(\underline{a}_1, a_1(l) e^{-\frac{1}{\tau}}\right) & \text{if } V(l) = 1 \\ \min\left(\overline{a}_1, a_1(l) e^{\frac{1}{\tau}}\right) & \text{if } V(l) = 0 \end{cases} \quad (5)$$

where $\underline{a}_1$ and $\overline{a}_1$ are lower and upper bounds on $a_1(l)$ respectively.
To calculate the likelihood functions, we model the power distribution of the noisy reverberant speech in each frequency bin using a Generalized Gamma distribution (2) with $\gamma = \frac{1}{2}$. It is shown in [15] that this is equivalent to the assumption that the magnitude coefficients follow a Gamma distribution [11, 12, 13]. The shape parameter κ was determined experimentally for speech active, κ_1, and inactive, κ_0 by fitting Gamma distributions to histograms of the magnitude coefficients. The results were then averaged between frequency bins. We used the TIMIT database [14] with SNRs in the range 10 to 20 dB and reverberation times in the range 0.3 to 2.2 s. This leads to

$$p_{y|\mathcal{V}_0}(y(l)) = p_{GG}\left(y(l); \frac{1}{2}, \kappa_0, \frac{(\mu_R + \mu_N)\,\Gamma(\kappa_0)}{\Gamma(\kappa_0 + 2)}\right) \quad (6)$$

$$p_{y|\mathcal{V}_1}(y(l)) = p_{GG}\left(y(l); \frac{1}{2}, \kappa_1, \frac{(\mu_R + \mu_N)\,(1+\xi)\,\Gamma(\kappa_1)}{\Gamma(\kappa_1 + 2)}\right) \quad (7)$$

with ξ the long-term *a priori* Signal-to-Interference Ratio (SIR). In [16], ξ is a long-term Signal-to-Noise Ratio (SNR), fixed so that the total probability of errors averaged over possible SNR values is minimised. In our application however, ξ cannot be fixed at a single value or easily derived analytically as it is dependent on both the SNR and the

Signal-to-Reverberant Ratio. Instead, ξ is allowed to take two different values, depending on the previous state of the path under consideration. If $V(l-1) = 1$, we assume that significant reverberation energy is present and therefore choose $\xi_1 = 0\,\text{dB}$. If $V(l-1) = 0$, we use the optimal value derived in [16] for the noise only case of $\xi_0 = 15\,\text{dB}$.
The mean noise power is assumed to be quasi-stationary so that $\mu_N(l) = \mu_N(l-1)$ is used in (6,7). To compute the mean reverberant power μ_R at time l, we use its estimate from the EKF operation at time $l-1$ and predict its evolution from(3) as

$$\mu_R(l) = \alpha\mu_R(l-1) + V(l-1)\, f\, \mu_S(l-1). \quad (8)$$

To find the probability of the path leading to $V(l) = \mathcal{V}_j$, $j \in \{0; 1\}$ through this HSMM, we compute

$$P\left(\mathcal{V}_j, \mathcal{O}_{l-1}|y(l)\right) = P\left(\mathcal{V}_j|y(l), \mathcal{O}_{l-1}\right) \times P(\mathcal{O}_{l-1}). \quad (9)$$

Because the computation of the observation likelihoods depend on the continuous EKF operation and we have a semi-continuous transition probability scheme, information about the past frames is needed. Therefore, we keep track of the n-best list of possible paths as well as the associated state vectors.

3.3. Extended Kalman Filtering

In each frequency bin, the EKF state vector encompasses the five quantities that we need to estimate: μ_S, μ_R, μ_N, α and f. The first three of these lie in the range $(0, +\infty)$ while the last two lie in the range $(0, 1)$. To avoid range constraints on the state vector elements, we define the state vector as

$$\mathbf{x}(l) = \log\left[\,\mu_S(l),\, \mu_R(l),\, \mu_N(l),\, \rho(\alpha(l)),\, \rho(f(l))\,\right]^T \quad (10)$$

where $\rho(x) \triangleq \frac{x}{1-x}$. Because of this parameterisation, both the prediction and update stages of a conventional Kalman Filter operation have non-linear terms. As they are analytically differentiable, we are able to use an EKF [17].
The prediction stage of the algorithm is described by the following set of equations.

$$\hat{\mathbf{x}}(l|l-1) = g\left(\hat{\mathbf{x}}(l-1),\, \hat{\mathrm{v}}(l-1)\right) \quad (11)$$

$$\mathbf{C}_{l|l-1} = \mathbf{G}_{l-1}\mathbf{\Sigma}_{l-1}\mathbf{G}_{l-1}^T + \mathbf{Q}_{l-1} \quad (12)$$

with $g(\mathbf{x}, v) = [\,x_1,\, \eta(\mathbf{x}, v),\, x_3,\, x_4,\, x_5\,]^T$ where x_i is the i^{th} element of $\mathbf{x}$. Eq. (8) translates into

$$\eta(\mathbf{x}, v) = \log\left[\frac{e^{x_2}}{1+e^{-x_4}} + v\,\frac{e^{x_1}}{1+e^{-x_5}}\right]. \quad (13)$$

$\mathbf{C}_{l|l-1}$ is the covariance matrix of the prediction stage, with $\mathbf{G}_{l-1} = \frac{\partial g}{\partial x}\big|_{\hat{\mathbf{x}}(l-1)}$ the Jacobian matrix of the prediction function, Σ_{l-1} the covariance matrix of the state at the previous time frame and $\mathbf{Q}_{l-1}$ the covariance matrix of the additive noise process of the prediction stage, which represents the natural variation of the state vector.
The update stage of the EKF algorithm is then defined by

$$e(l) = y(l) - h(\hat{\mathbf{x}}(l|l-1), \hat{\mathrm{v}}(l)) \quad (14)$$

$$\mathbf{U}_l = \mathbf{H}_l\mathbf{C}_{l|l-1}\mathbf{H}_l^T + \mathbf{M}_l \quad (15)$$

$$\mathbf{K}_l = \mathbf{C}_{l|l-1}\mathbf{H}_l^T\mathbf{U}_l^{-1} \quad (16)$$

$$\hat{\mathbf{x}}(l) = \hat{\mathbf{x}}(l|l-1) + \mathbf{K}_l e(l) \quad (17)$$

$$\mathbf{\Sigma}_l = \mathbf{C}_{l|l-1} - \mathbf{K}_l\mathbf{U}_l\mathbf{K}_l^T. \quad (18)$$

In Eq. (14), $e(l)$ is the prediction error, which is computed using

$$h(\mathbf{x}, v) = v\, e^{x_1} + e^{x_2} + e^{x_3}. \quad (19)$$

The covariance of the update stage $\mathbf{U}_l$, is calculated using $\mathbf{H}_l = \frac{\partial h}{\partial x}\big|_{\hat{\mathbf{x}}(l|l-1)}$ and $\mathbf{M}_l$, the variance of the predicted output knowing the probability distributions of the random variables involved. As S, R and N each follow a gamma distribution, we have $\mathbf{M}_l = \hat{\mathrm{v}}(l)\,\kappa_S^{-1}\hat{\mu}_S^2 + \kappa_R^{-1}\hat{\mu}_R^2 + \kappa_N^{-1}\hat{\mu}_N^2$. $\mathbf{K}_l$ is the Kalman gain, which is used to update the mean and covariance of the state vector according to equations (17,18).

4. EVALUATION

To evaluate the algorithm, clean speech signals were generated from the concatenation of different sentences pronounced by a male speaker from one of the CMU ARCTIC databases [18], resulting in speech files sampled at 16kHz of length varying between 7 and 12 seconds. These were then convolved with different room impulse responses (RIR) taken from the Aachen database [19].

In the first experimental setup, a highly reverberant cathedral RIR was used and non-stationary restaurant noise from ITU-T P.501 [20] was added to a reverberant speech file at an SNR of 15 dB. The STFT was computed using a Hann window of length 20ms and overlap factor 4. As all the frequency bins are processed independently, only the results corresponding to the bin centred on 700Hz are shown. By analysing the power spectrogram of the impulse response in this frequency bin, the ground truth for α and f were found to correspond to $T_{60} = 2.58\,\text{s}$ and $\text{DRR} = -1.73\,\text{dB}$. To initialise the method, the first 40ms of the degraded audio file were assumed to be only noise, and therefore $\mu_N(0)$ was set to the mean observed power during this time interval. $\mu_S(0)$ was initialised to 10 dB above the value of $\mu_N(0)$, and $\mu_R(0)$ was set to $-20\,\text{dB}$. Both the decay constant and the drop were initially set up to erroneous values corresponding to $T_{60}(0) = 3.4\,\text{s}$ and $\text{DRR}(0) = -7\,\text{dB}$.

The ground truth for speech activity was obtained by assuming that the time frames of the clean speech spectrogram that were in a 15 dB range of its maximum in the frequency bin under consideration corresponded to a speech active state. Applying our method resulted in 13.7% of frames being wrongly classified by the HSMM with 20.3% of these errors

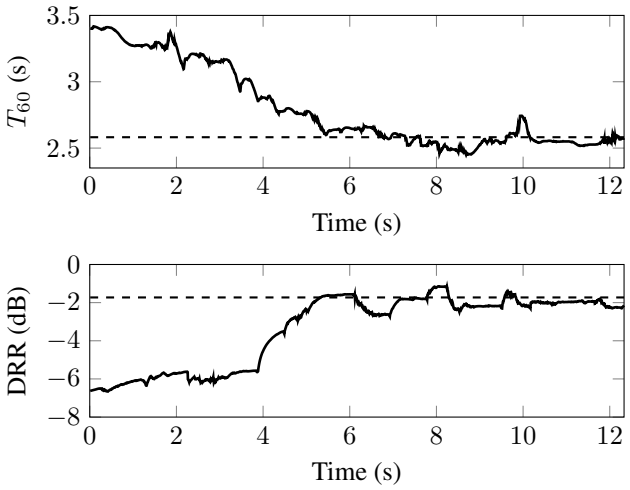

Fig. 3: Estimated T_{60} and DRR (solid lines) and true values (dashed lines) in the frequency bin under consideration.

being false alarms. The results of the estimation of T_{60} and DRR are shown Fig. 3. It can be seen that both converge to the ground truth within 5 seconds with small fluctuations around the ground truth thereafter. The dashed lines in Fig. 4 show the estimated mean noise and reverberant powers while the solid lines show the corresponding instantaneous ground truth.

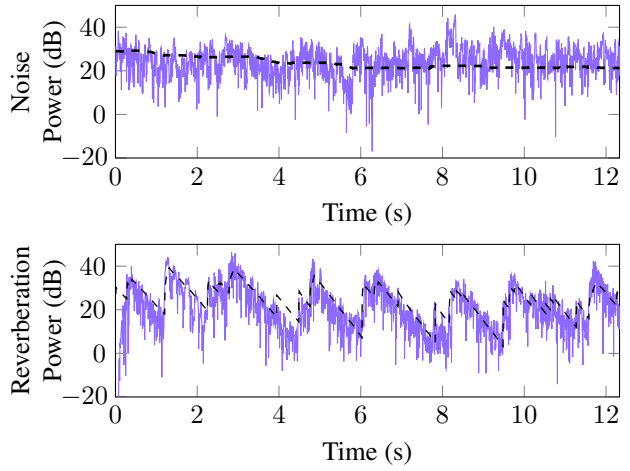

Fig. 4: Observed noise and reverberation powers (solid lines) and corresponding estimated mean powers (dashed lines) in this frequency bin.

To assess the consistency of the results obtained using the algorithm, the second experimental setup consisted of 50 speech files convolved with two different RIRs corresponding to a lecture room and a meeting room. White noise was added at 5, 10, 15 and 20 dB SNR. The measured ground truth for frequency bins centred at 0.7, 2 and 7 kHz were:

		700 Hz	2000 Hz	7000 Hz
Lecture Room	DRR	1.47 dB	−4.24 dB	3.27 dB
	T_{60}	0.95 s	0.93 s	0.63 s
Meeting Room	DRR	1.93 dB	3.67 dB	6.95 dB
	T_{60}	0.35 s	0.38 s	0.24 s

In all cases, DRR and T_{60} were initialised to 0 dB and 1.3 s respectively. The Root Mean Square Errors (RMSE) of the DRR and T_{60} estimates in these frequency bins are plotted in Fig. 5 against the SNR. In most cases, the errors increase at poor SNRs primarily because of more frequent errors in speech activity detection. At high SNRs, the DRR of the Meeting Room at 700 Hz is overestimated by about 3 dB even though the T_{60} estimate is very accurate. Further investigation of the energy decay curve in this frequency bin suggests the EKF sometimes converges to the Direct-to-Late Reverberation energy ratio when the DRR is low.

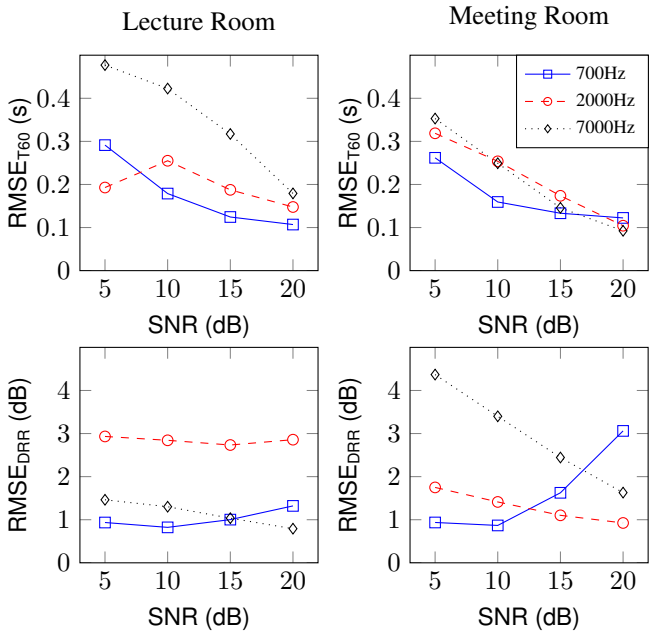

Fig. 5: Root Mean Squared Error of the estimated T_{60} and DRR values. Left panel: lecture RIR, right panel: meeting RIR.

5. CONCLUSION

Fast and accurate estimation of the reverberation parameters in each frequency bin is a key requirement of dereverberation algorithms. In this paper we proposed an online method estimating the speech activity, T_{60}, DRR, noise and reverberant mean powers by using an extended Kalman filter conditioned by a 2-state HSMM. Experimental results on individual frequency bins show that the estimates of the reverberation parameters T_{60} and DRR generally converge quickly and reliably to the true values.

6. REFERENCES

[1] Keisuke Kinoshita, Marc Delcroix, Takuya Yoshioka, Tomohiro Nakatani, Emanuel Habets, Reinhold Haeb-Umbach, Volker Leutnant, Armin Sehr, Walter Kellermann, Roland Maas, Sharon Gannot, and Bhiksha Raj, "The REVERB challenge," Website, 2014.

[2] Emanuel A. P. Habets, *Single- and Multi-Microphone Speech Dereverberation using Spectral Enhancement*, Ph.D. thesis, Technische Universiteit Eindhoven, 2007.

[3] Manfred R. Schroeder, "Statistical parameters of the frequency response curves of large rooms," *J. Audio Eng. Soc*, vol. 35, no. 5, pp. 299–306, 1987.

[4] Jean-Dominique Polack, *La transmission de l'énergie sonore dans les salles*, Ph.D. thesis, Université du Maine, Le Mans, France, 1988.

[5] Jean-Dominique Polack, "Playing billiards in the concert hall: The mathematical foundations of geometrical room acoustics," *Applied Acoustics*, vol. 38, no. 2, pp. 235–244, 1993.

[6] Heinrich W. Lollmann and Peter Vary, "Estimation of the reverberation time in noisy environments," *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2008.

[7] Jimi Y. C. Wen, Emanuel A. P. Habets, and Patrick A. Naylor, "Blind estimation of reverberation time based on the distribution of signal decay rates," *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 329–332, 2008.

[8] James Eaton, Nikolay D. Gaubitch, and Patrick A. Naylor, "Noise-robust reverberation time estimation using spectral decay distributions with reduced computational cost," *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 161–165, 2013.

[9] Tiago H. Falk and Wai-Yip Chan, "Temporal dynamics for blind measurement of room acoustical parameters," *IEEE Transactions on Instrumentation and Measurement*, vol. 59, no. 4, pp. 978–989, 2010.

[10] H. Othman and T. Aboulnasr, "A semi-continuous state-transition probability HMM-based voice activity detector," *EURASIP Journal on Audio, Speech and Music Processing*, p. 43218, 2007.

[11] Theodoros Petsatodis, Christos Boukis, Fotios Talantzis, Zheng-Hua Tan, and Ramjee Prasad, "Convex combination of multiple statistical models with application to VAD," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2314–2326, 2011.

[12] Timo Gerkmann and Rainer Martin, "Empirical distributions of DFT-domain speech coefficients based on estimated speech variances," *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2010.

[13] Thomas Lotter and Peter Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model," *EURASIP Journal on Applied Signal Processing*, vol. 7, pp. 1110–1126, 2005.

[14] W. Fisher, G. R. Doddington, and K. Goudie-Marshall, "The DARPA speech recognition research database: Specification and status," *Proceedings DARPA Speech Recognition Workshop*, 1986.

[15] A. Dadpay, E. S. Soofi, and R. Soyer, "Information measures for Generalized Gamma family," *Journal of Econometrics*, vol. 138, pp. 568–585, 2007.

[16] Timo Gerkmann and Richard C. Hendricks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, May 2012.

[17] M. S. Grewal and A. P. Andrews, *Kalman Filtering, Theory and Practice Using MATLAB*, Wiley, 2001.

[18] John Kominek and Alan W. Black, "CMU ARCTIC databases for speech synthesis," Tech. Rep. CMU-LTI-03-177, Language Technologies Institute, Carnegie Mellon University, 2003.

[19] Marco Jeub, Magnus Schäfer, and Peter Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," *Proceedings of International Conference on Digital Signal Processing*, 2009.

[20] ITU-T P.501, "Test signals for use in telephonometry," Recommendation, International Telecommunication Union, 2012.