Model-based Blind Estimation of Reverberation Time: Application to Robust ASR in Reverberant Environments

Laurent Couvreur[†], *Christophe Ris*[†] *Christophe Couvreur*[‡]

[†]Faculté Polytechnique de Mons, Belgium | [‡]Lernout & Hauspie Speech Products, Belgium
{lcouv,ris}@tcts.fpms.ac.be | christophe.couvreur@lhs.be

Abstract

This paper presents a method for blind estimation of reverberation times in reverberant enclosures. The proposed algorithm is based on a statistical model of short-term log-energy sequences for echo-free speech. Given a speech utterance recorded in a reverberant room, it computes a Maximum Likelihood estimate of the room full-band reverberation time. The estimation method is shown to require little data and to perform satisfactorily. The method has been successfully applied to robust automatic speech recognition in reverberant environments by model selection. For this application, the reverberation time is first estimated from the reverberated speech utterance to be recognized. The estimation is then used to select the best acoustic model out of a library of models trained in various artificial reverberant conditions. Speech recognition experiments in simulated and real reverberant environments show the efficiency of our approach which outperforms standard channel normalization techniques.

1. Introduction

For the past decades, the number of applications which involve transmitting speech in reverberant rooms has continuously increased. Let's mention hands-free telephony, video-conference or hands-free automatic speech recognition (ASR) among others. Room reverberation is often considered as an undesirable effect which corrupts any speech signal propagating from a speaker to a microphone in a reverberant room. The reverberation time T_{60} [7] has been a central parameter for characterizing room reverberation. Hence, measurement of the reverberation time is a valuable tool for designing room acoustics and developing audio signal processing applications in order to reduce the impact of room reverberation. The reverberation time is classically measured [7] either from an energy decay curve, or from a room impulse response using Schroeder's method. The former directly measures the time interval in which the sound energy in a room reaches one millionth of its initial value (-60dB) after interrupting a sound source. The latter derives the reverberation time after reverse-time integration of an acoustic impulse response measured between a source and a microphone in the reverberant room.

In this paper, we propose an algorithm for blindly estimating the reverberation time of a room from speech signal recorded in that room. We first model the impact of room reverberation on echo-free short-term log-energy L_{eq} sequences by a parametric distortion model whose coefficients are related to the reverberation time. Then, given a reverberated L_{eq} sequence, Maximum Likelihood (ML) estimates of the distortion model parameters are obtained using a statistical model for echo-free L_{eq} sequences. The last model is trained beforehand on a clean speech database. Finally, the estimate of the reverberation time is derived from the estimates of the distortion model parameters. The algorithm requires little data, e.g. a few seconds of speech signal, and is useful for on-line estimation of the reverberation time.

We also present in this paper a technique for robust ASR in reverberant environments. The method is based on selecting the best acoustic model out of a library of models trained separately in different reverberant conditions. The best model is the model trained in the reverberant conditions most closely matching the reverberation time of the operating environment. Hence, the selection is directed by prior estimation of the reverberation time from the reverberated speech utterance to be recognized via our newly proposed method.

The paper is organized as follows. In the next section, we describe the algorithm for the blind estimation of the reverberation time. Results for recognition of connected digit sequences in reverberant environments by model selection are reported in section 3. Concluding remarks are given in section 4.

2. Estimation Algorithm

2.1. Room Reverberation Model

The most detailed model of room reverberation is the room impulse response between the speaker and the microphone. One can propose to identify blindly the room impulse response from recorded reverberated speech, and then compute the reverberation time from the estimated impulse response, e.g. using Schroeder's method [7]. Since the blind identification of a room impulse response is a badly conditioned task, we propose instead to use a simpler model of room reverberation in order to simplify the T_{60} estimation problem. We decide to model the impact of room reverberation on the short-term log-energy (L_{eq}) sequence X_m instead of on the clean speech signal x_n,

$$X_m \triangleq 10 \log_{10} \left(\frac{1}{N_w} \sum_{n=mN_r}^{mN_r+N_w-1} x_n^2 \right) \tag{1}$$

with $N_w \triangleq T_w \times F_s$ and $N_r \triangleq F_s/F_r$, where n, m, T_w, F_s and F_r denote the sample index, the frame index, the analysis frame length [s], the sampling frequency [Hz] and the frame rate [Hz], respectively. Figure 1 gives an example of a clean speech utterance x_n and its reverberated version y_n obtained by convolving x_n with a typical room impulse response h_n. The figure also shows the distortion on the corresponding L_{eq} sequences X_m and Y_m computed after proper normalization of the speech signals. Under the assumption that the sound field is diffuse and that the reverberation time is frequency independent, the decays of Y_m from peak to valley should be exactly linear, and thus exponential in the linear energy domain. That is, the impact of room reverberation can be modeled by a first order auto-regressive (AR) filter,

$$W_m = \alpha_0 Z_m + \alpha_1 Z_{m-1} \tag{2}$$

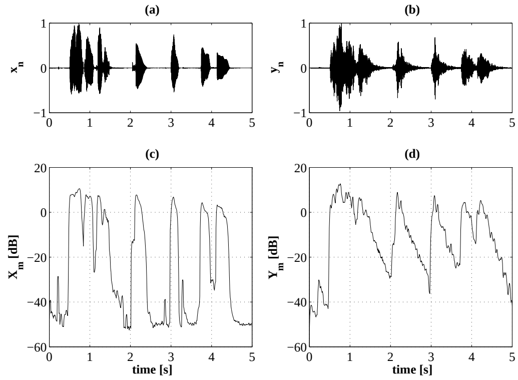

Figure 1: Waveforms for (a) a clean speech utterance x_n and (b) its reverberant version y_n, and the corresponding L_{eq} sequences (c) X_m and (d) Y_m for T_w=30ms and F_r=100Hz.

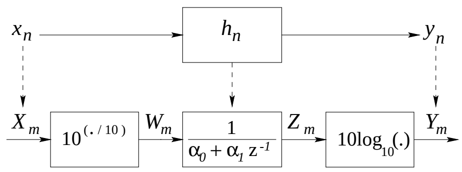

Figure 2: Room reverberation process (upper part) for temporal signal and equivalent diffuse model (lower part) for L_{eq} sequence.

where $W_m \triangleq 10^{X_m/10}$ and $Z_m \triangleq 10^{Y_m/10}$ denote the short-term linear energy sequences of the clean and reverberated speech signals, respectively. The assumed reverberation model is summarized in figure 2. In the sequel, we describe a method for ML estimation of the AR coefficients (α_0, α_1) from the observed L_{eq} sequence Y_m only. Our estimation algorithm requires a statistical model for the echo-free L_{eq} sequence X_m which is briefly presented in the next section. Once α_1 has been estimated, T_{60} can be derived via [7]

$$T_{60} = \log 10^6 / (-\log(-\alpha_1) \times F_r). \qquad (3)$$

2.2. Echo-free L_{eq} Sequence Model

The clean speech L_{eq} sequence X_m is typically non-stationary and characterized by two states, called the silence and speech states. Furthermore, successive values are undoubtedly not statistically independent: they are correlated (see figure 1.c). Hence, we choose to model X_m by a 2-states one-dimensional Linear Predictive Hidden Markov Model (LP-HMM) [6]. In this model, the L_{eq} sequence X_m is generated by processing the emission sequence E_m with an AR filter of order P,

$$\beta_0(s_m) X_m = E_m - \sum_{p=1}^{P} \beta_p(s_m) X_{m-p} \qquad (4)$$

whose coefficients $\beta_p(s_m)$, $p = 0, \ldots, P$, are function of the LP-HMM state sequence s_m. The emissions E_m are assumed to be conditionally independent given the state sequence s_m and have a Gaussian distribution with mean μ_i and variance σ_i for $s_m = i$, $i = 0, 1$. To complete our model, we define the transition probabilities $a_{ij} \triangleq P[s_m = j | s_{m-1} = i]$. All the parameters can be estimated by an Expectation-Maximization

Table 1: Parameters of a 2-states one-dimensional first-order LP-HMM for L_{eq} sequence of clean speech for T_w=30ms and F_r=100Hz.

$s_m = i$	a_{ii}	a_{ij}	μ_i	σ_i	$\beta_0(i)$	$\beta_1(i)$
0	0.95	0.05	-4.3	4.2	1.0	-0.92
1	0.03	0.97	1.1	3.2	1.0	-0.77

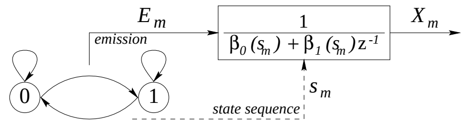

Figure 3: 2-states one-dimensional first-order LP-HMM for modeling L_{eq} sequence of clean speech (0: silence state, 1: speech state).

(EM) algorithm [6] from L_{eq} sequences extracted from a clean speech database. Note that the modeling is done over the log-energy sequences because these data have a better conditioned dynamics than the linear energy sequences. Figure 3 illustrates a 2-states one-dimensional LP-HMM with AR filters limited to first order ($P = 1$) which is used in this work. Table 1 gives the parameters of the model trained on a clean part of the AURORA speech database [2].

2.3. Maximum Likelihood T_{60} Estimation

In this section, we present the algorithm for blindly estimating the reverberation time T_{60}. The algorithm is inspired by ML stochastic matching [8]. Given a statistical model of the unobserved clean L_{eq} sequence X_m (section 2.2), the parameters (α_0, α_1) of the distortion model (section 2.1) are estimated so as to maximize the likelihood of the observed reverberated L_{eq} sequence Y_m. Maximization is performed by an iterative EM-like algorithm. Given an observed L_{eq} sequence Y_0^M of length $M+1$, or equivalently its linear counterpart Z_0^M, and current estimates $(\alpha_0^{(\ell)}, \alpha_1^{(\ell)})$ of the distortion model, we first compute (E-step) the auxiliary function,

$$Q(\alpha_0^{(\ell+1)}, \alpha_1^{(\ell+1)} | \alpha_0^{(\ell)}, \alpha_1^{(\ell)})$$
$$\triangleq E\left[\log p(Z_0^M, s_0^M | \alpha_0^{(\ell+1)}, \alpha_1^{(\ell+1)}) \;\middle|\; Z_0^M, \alpha_0^{(\ell)}, \alpha_1^{(\ell)}\right] \qquad (5)$$

where s_0^M denotes the hidden state sequence of the LP-HMM. We then find re-estimation formulae by maximizing (M-step) equation (5) with respect to $(\alpha_0^{(\ell+1)}, \alpha_1^{(\ell+1)})$. Since the distortion parameters do not depend on the state sequence, we can equivalently solve,

$$(\alpha_0^{(\ell+1)}, \alpha_1^{(\ell+1)}) = \arg\max E\left[\log p(Z_0^M | s_0^M)\right] \qquad (6)$$

where conditioning arguments have been dropped for ease of notation. In order to derive closed-form re-estimation formulae, we assume that the ℓ-th estimate $W_m^{(\ell)} = \alpha_0^{(\ell)} Z_m + \alpha_1^{(\ell)} Z_{m-1}$ is close to the next estimation $W_m^{(\ell+1)}$. Hence, one can write the following linear approximation,

$$10 \log_{10}(W_m^{(\ell+1)}) \simeq 10 \log_{10}(W_m^{(\ell)})$$
$$+ (W_m^{(\ell+1)} - W_m^{(\ell)}) / \gamma W_m^{(\ell)}, \quad \gamma = \log(10)/10 \qquad (7)$$

using Taylor series expansion limited to first order. Equation (7) can be rewritten as

$$X_m^{(\ell+1)} = \sum_{k=0}^{1} A_{k,m}^{(\ell+1)} Z_{m-k} + B_m^{(\ell+1)} \qquad (8)$$

with $A_{k,m}^{(\ell+1)} \triangleq \alpha_k^{(\ell+1)}/\gamma W_m^{(\ell)}$ and $B_m^{(\ell+1)} \triangleq X_m^{(\ell)} - 1/\gamma$. Using equation (4) for $P = 1$, we obtain the linear relation,

$$E_m^{(\ell+1)} = \sum_{p=0}^{1} \beta_p(s_m) \left[\sum_{k=0}^{1} A_{k,m-p}^{(\ell+1)} Z_{m-p-k} + B_{m-p}^{(\ell+1)} \right]. \tag{9}$$

Hence, we have now to solve,

$$(\alpha_0^{(\ell+1)}, \alpha_1^{(\ell+1)}) = \arg\max E \left[\log \left(p(E_0^M | s_0^M) |J| \right) \right] \tag{10}$$

with $|J|$ denoting the determinant of the Jacobian of the matrix form of the linear transformation (9) between the observation vector Z_0^M and the emission vector E_0^M. By proper normalization of the speech utterances during training of the LP-HMM and during estimation of T_{60}, we can constrain α_0 to be equal to 1. Under this constrain, one can easily show that $|J|$ is independent of (α_0, α_1). Furthermore, since the emissions are conditionally independent, equation (10) becomes

$$\alpha_1^{(\ell+1)} = \arg\max \sum_{m=0}^{M} E \left[\log p(E_m^{(\ell+1)} | s_m) \right]. \tag{11}$$

Equation (11) can be further developed by introducing the *a posteriori* state probabilities $\gamma_{m,i}^{(\ell)} \triangleq P[s_m = i | Z_0^M, \alpha_1^{(\ell)}]$, $i = 0, 1$. Using the assumption that E_m is Gaussian distributed, we find

$$\alpha_1^{(\ell+1)} = \arg\max \sum_{m=0}^{M} \sum_{i=0}^{1} \gamma_{m,i}^{(\ell)} (E_m^{(\ell+1)} - \mu_i)/\sigma_i^2 \tag{12}$$

where constant terms have been cancelled out. The re-estimation formula for α_1 is obtained by setting the first derivative of the right-hand term of (12) with respect to $\alpha_1^{(\ell+1)}$ to zero after using relation (9). We finally obtain,

$$\alpha^{(\ell+1)} = (-p^{(\ell)} - q^{(\ell)})/r^{(\ell)} \tag{13}$$

with $p^{(\ell)}$, $q^{(\ell)}$ and $r^{(\ell)}$ defined as,

$$c_{m,i}^{(\ell)} \triangleq \sum_{p=0}^{1} \beta_p(i) Z_{m-p-1}/\gamma W_{m-p}^{(\ell)} \tag{14}$$

$$r^{(\ell)} \triangleq \sum_{m=0}^{M} \sum_{i=0}^{1} \frac{\gamma_{m,i}^{(\ell)}}{\sigma_i^2} \left(c_{m,i}^{(\ell)} \right)^2 \tag{15}$$

$$q^{(\ell)} \triangleq \sum_{m=0}^{M} \sum_{i=0}^{1} \frac{\gamma_{m,i}^{(\ell)}}{\sigma_i^2} \left(\sum_{p=0}^{1} \frac{\beta_p(i) Z_{m-p}}{\gamma W_{m-p}^{(\ell)}} \right) c_{m,i}^{(\ell)} \tag{16}$$

$$p^{(\ell)} \triangleq \sum_{m=0}^{M} \sum_{i=0}^{1} \frac{\gamma_{m,i}^{(\ell)}}{\sigma_i^2} \left(\sum_{p=0}^{1} \beta_p(i) (X_{m-p}^{(\ell)} - \frac{1}{\gamma}) \; - \; \mu_i \right) c_{m,i}^{(\ell)}. \tag{17}$$

The resulting iterative estimation algorithm is outlined below:

1. Initialize the estimate of the distortion parameter $\alpha_1^{(0)}$ ($\alpha_0 = 1$) and set $\ell = 0$, and compute $Z_m = 10^{Y_m/10}$, $m = 0, \ldots, M$;
2. Apply the inverse distortion filter to obtain $W_m^{(\ell)} = \alpha_0^{(\ell)} Z_m + \alpha_1^{(\ell)} Z_{m-1}$, and compute $X_m^{(\ell)} = 10 \log_{10} W_m^{(\ell)}$, $m = 0, \ldots, M$;
3. Estimate the *a posteriori* state probabilities $\gamma_{m,i}^{(\ell)}$ via the Forward-Backward algorithm [6] given the LP-HMM parameters and $X_m^{(\ell)}$, $m = 0, \ldots, M$;
4. Obtain updated estimate of $\alpha_1^{(\ell+1)}$ by applying the re-estimation formula (13)–(17);
5. Set $\ell = \ell + 1$ and go to 2 unless convergence is reached;
6. Derive the reverberation time T_{60} from $\alpha_1^{(\ell)}$ via (3).

3. Application to Robust ASR

In [3], we showed that training acoustic models on artificially reverberated speech can provide robust systems for hands-free speech recognition in reverberant environments. We proposed in [3] to produce reverberated speech by processing clean speech with a filter whose finite-length impulse response h_n is designed to match the reverberation time of the target reverberant operating environment. Under the assumptions that the sound field is diffuse and that the reverberation time is frequency independent, the synthetic impulse response h_n can be obtained by shaping a Gaussian white random sequence with a decaying exponential whose damping constant is directly related to the reverberation time. Once a reverberated database has been generated by convolving a clean speech database with h_n, an acoustic model corresponding to the specified T_{60} can be trained for speech recognition. Repeating the process for different values of T_{60}, a library of acoustic models can be build for various reverberation times. Since the reverberation time is rarely known in advance, T_{60} has to be estimated during operation from the speech utterance to be recognized. Hence, the proposed T_{60} estimation algorithm can be used for selecting the best acoustic model.

3.1. Speech Material and Recognizer Description

The speech material used in this work comes from the clean part of the AURORA [2] database and consists of connected digit sequences. Recognition experiments are performed with a phoneme-based hybrid Multilayer Perceptron(MLP)/HMM recognizer. The phoneme *a posteriori* probabilities are estimated by a MLP fed with acoustic features computed from 30ms long/10ms overlapping frames of speech signal sampled at 8kHz. The acoustic features are obtained by one of three following front-ends: Mel-warped frequency cepstral coefficients (MFCC), MFCC with cepstral mean subtraction (CMS) [4] and logRASTA-PLP [5] coefficients. The last two front-ends are known to be robust to channel distortion. Speech decoding is done by Viterbi search, with neither pruning nor grammar constraints.

3.2. Experimental Results

First, we trained nine acoustic models: one on echo-free speech and eight on artificially reverberated training sets for T_{60} varying uniformly from 200ms to 1600ms. For each T_{60}, the corresponding training set was obtained by using the method depicted in the previous section. Meanwhile, test sets were generated by convolving the clean test set with room impulse responses computed by the Image Method [1]. The wall absorption coefficients of the reverberant room simulator were chosen to get specific reverberation times. Table 2 reports cross-testing results. We see that the lowest word error rate (WER), i.e. the sum of the substitution (SUB), deletion (DEL) and insertion (INS) error rates, is always achieved by the acoustic model most closely matching the testing conditions (main diagonal). Even if there is no acoustic model which matches exactly the test T_{60}, the performance of the selected model does not degrade much if the grid for T_{60} in the library of acoustic models is tight enough.

Table 2: Performances WER [%] of MFCC-based acoustic models trained on artificially reverberated speech for various reverberant testing conditions.

Test set	Training set								
	clean	$T_{60} = 200$ms	400ms	600ms	800ms	1000ms	1200ms	1400ms	1600ms
clean	**1.7**	2.9	7.6	11.9	15.9	19.8	20.6	22.7	23.8
$T_{60} = 200$ms	7.0	**3.6**	4.5	6.4	9.8	12.5	13.7	15.1	16.1
300ms	7.8	**3.9**	**4.4**	6.4	9.8	12.3	13.9	15.0	15.7
400ms	18.7	9.6	**5.2**	5.7	8.5	12.2	12.8	14.7	15.3
500ms	20.1	11.2	**5.9**	**5.9**	8.7	12.2	12.8	14.7	15.4
600ms	29.7	20.2	11.3	**9.2**	10.0	12.6	13.7	15.6	16.4
700ms	33.2	24.7	14.9	**11.2**	**11.3**	13.6	14.4	16.1	17.1
800ms	41.0	33.7	22.1	17.3	**14.0**	15.9	16.6	18.2	19.1
1000ms	43.4	35.8	24.0	20.4	16.0	**17.0**	17.1	18.7	19.7
1200ms	49.3	43.1	32.0	27.9	20.9	20.7	**20.4**	21.6	22.1
1400ms	51.1	48.5	36.8	33.5	26.0	24.9	23.2	**24.5**	24.4
1600ms	52.9	50.1	37.3	36.6	28.1	26.7	25.3	25.1	**25.1**

Table 3: Median absolute value error (MAVE) [ms] and relative MAVE [%] for blind estimation of T_{60} and corresponding confusion rate (CR) [%] for model selection.

Test Set	MAVE [ms]	relative MAVE [%]	CR [%]
A	51.4	8.2	24.0
B	80.9	9.7	26.8

As could have been expected, WER increases for the matching acoustic model as the reverberation becomes stronger.

Next, we tested our model selection approach by blind estimation of T_{60}. Test sets were generated by mixing groups of utterances reverberated at different levels. Each group was at least 5s long and obtained by convolving clean utterances with a room impulse response corresponding to a specific T_{60}. Prior to its recognition, every group was processed: the L_{eq} sequence was computed for $T_w = 30$ms and $F_r = 100$Hz, T_{60} was estimated and the most closely matching MLP of the library was activated. Two sets of experiments were performed: test sets were generated by convolution with room impulse responses either computed in one of the previous simulated rooms (test A), or measured in real reverberant enclosures (test B) using a correlation method based on an optimal time-stretched pulse [9]. Table 3 gives the median absolute value errors (MAVE) for T_{60} estimation. Reference reverberation times for computing MAVE are known by design for test A while they were estimated by Schroeder's method from the measured impulse responses for test B. Table 3 also reports confusion rates, i.e. wrong acoustic model selection rates. Table 4 shows that the proposed model selection method outperforms systems based on standard channel-robust acoustic features. Furthermore, it approaches the performance of the "Oracle" method for which T_{60} is assumed to be known in advance and the best model is always selected.

4. Conclusion and Future Work

We have proposed an algorithm for blind estimation of the reverberation time, and successfully applied it to robust speech recognition in reverberant environments by acoustic model selection. Even if satisfactorily results were obtained in terms of both estimation error and selection error, further improvements can be expected by relaxing the main hypothesis which supposes that the reverberation time is frequency independent. To do so, the method has to be extended to a multiband approach for which T_{60} is assumed constant inside frequency subbands only.

Table 4: Comparison between performances WER (SUB/DEL/INS) [%] of two standard normalization techniques, our model selection method and the "Oracle" method.

Method	Test Set	
	A	B
MFCC-CMS	35.7(10.2/15.8/9.7)	40.3(12.1/18.7/9.5)
logRASTA-PLP	35.2(12.6/13.9/8.7)	38.2(11.5/18.1/8.6)
Model Selection	14.0(4.8/5.4/3.8)	18.1(6.5/8.4/3.2)
"Oracle"	13.6(4.7/6.1/2.8)	17.5(6.2/8.4/2.9)

5. References

[1] J. B. Allen and D. A. Berkley, "Image Method for Efficiently Simulating Small-Room Acoustics", *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[2] AURORA database - http://www.elda.fr/aurora2.html.

[3] L. Couvreur, C. Couvreur and C. Ris, "A Corpus-Based Approach for Robust ASR in Reverberant Environments", *Proc. of ICSLP'2000*, vol. 1, pp. 397–400, Beiging, China, Oct. 2000.

[4] S. Furui, "Cepstral Analysis Technique for Automatic Speaker Verification", *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 29, no. 2, pp. 254–272, Apr. 1981.

[5] H. Hermansky and N. Morgan, "RASTA Processing of Speech", *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, Oct. 1994.

[6] P. Kenny, M. Lennig and P. Mermelstein, "A Linear Predictive HMM for Vector-Valued Observations with Applications to Speech Recognition", *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 38, no. 2, pp. 220–225, Feb. 1990.

[7] H. Kuttruff, *Room Acoustics*, Elsevier, 3rd ed., 1991.

[8] A. Sankar and C.-H. Lee, "A Maximum-Likelihood Approach to Stochastic Matching for Robust Speech Recognition", *IEEE Trans. on Speech and Audio Processing*, vol. 4, no. 3, pp. 190–202, May 1996.

[9] Y. Suzuki, F. Asano, H.-Y. Kim and T. Sone, "An Optimum Computer-Generated Pulse Signal Suitable for the Measurement of Very Long Impulse Responses", *J. Acoust. Soc. Am.*, vol. 97(2), pp. 1119–1123, Feb. 1995.