OPTIMAL SPEECH ESTIMATOR CONSIDERING ROOM RESPONSE AS WELL AS ADDITIVE NOISE:
DIFFERENT APPROACHES IN LOW AND HIGH FREQUENCY RANGE

Lae-Hoon Kim and Mark Hasegawa-Johnson

Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
Urbana, IL, USA

ABSTRACT

This paper proposes minimum mean squared error (MMSE) speech signal estimation in a reverberant space using different optimal estimators in the low and high frequency ranges. At low frequencies, an MMSE spectral amplitude estimator divided by the spectral amplitude of a representative impulse response produces optimal performance. In the high frequency range, the MMSE estimator is computed based on its sufficient statistic: the maximum likelihood (ML) estimate. Inference is factored using a two-step algorithm: the maximum likelihood value of the source spectrum is first estimated using expectation-maximization (EM) under the assumption of the hidden room response with complex Gaussian pdf, then the MMSE source spectral estimate is computed.

Index Terms— Signal enhancement, channel inversion, room response estimation, statistical room response modeling.

1. INTRODUCTION

One of the main problems we have to solve, in order to achieve speech recognition in acoustic spaces such as classrooms and moving cars, is noise. Noise in acoustic spaces is mostly composed of two components; "additive background noise", and "convolutive noise" which is caused by the room impulse response (RIR). Additive noise may be removed by optimal filtering [1, 2], because it can be easily modelled as an independent Gaussian noise. Early echoes (short-lag convolutive noise) may be similarly removed by optimal filtering of the log short-time Fourier transform (STFT) or cepstrum, but long-lag convolutive noise (reverberation) is seldom addressed.

From the classical room acoustic theory for a closed space, first we would like to introduce how we can model the room response from one point to the other in a closed space, for example, like when a driver talks to the hands free microphone for conversation with his family in a car. This modelling will be done in the low frequency range and in the high frequency range separately. The crossover frequency, "Schroeder's frequency" [3], specifies the frequency above which the acoustic modes overlap too much, in frequency, to be discretely modeled:

$$f_s = c\sqrt{\frac{6}{A}} \approx 2000\sqrt{\frac{T}{V}} \quad (1)$$

where c is the speed of the sound, A is the absorbtion area of the room (m^2), Sabine's reverberation time T=0.163V/A(sec), and V (m^3) is the volume of the space. In actual computation, we use the measured reverberation time. From this approach, a probability density function (pdf) of the room response above the crossover frequency shall be introduced, which can be thought of as a statistical model of the room response. Based on this model, we derive a novel EM algorithm to complete the MMSE-optimal estimates of the speech spectrum produced in a space with convolutive and additive noise.

2. BACKGROUND: MODELLING OF THE ROOM RESPONSE

Two different models of the room response should be introduced, because the statistical characteristics of the room response in the high frequency range are different from the statistical characteristics of the room response in the low frequency range. In the high frequency range, the room response can be assumed to have a Gaussian pdf due to mode overlapping; by the central limit theorem, the sum of modal responses with independent uniformly distributed phase approaches a Gaussian pdf [3]. Note that the Gaussian-like histogram differs little depending on the measurement point; a histogram computed at a different

measurement point was nearly identical. In the low frequency range, modes are discrete and easily measurable, therefore we model the low frequency range using a deterministic frequency response. To get the crossover frequency, we use Schroeder's frequency formula (1). For example, a typical automobile space with volume of 5 m^3 and reverberation time of 0.2s will have a crossover frequency of around 400 Hz.

3. OPTIMAL ESTIMATORS: DETERMINISTIC ROOM RESPONSE

This section considers the problem of MMSE spectral amplitude estimation (MMSE-SA, [1]) and MMSE log amplitude estimation (MMSE-LSA, [2]) assuming a deterministic, measured room response. Assume a signal Y measured based on source X, room response H, and noise signal D:

$$Y = HX + D \tag{2}$$

Below the Schroeder frequency, we assume that H is deterministic, and that Hs within a measurement region of radius $|\vec{r}|$ are similar enough [4, 5], so that over the region we assume to be able to replace Hs as a representative H_{rep}:

$$Y(\vec{r}) = a|H|_{rep}e^{j\alpha} + N, \quad |\vec{r}| \ll \text{wavelength} \tag{3}$$

where $a = |X|$ is a Rayleigh random variable with parameter λ_X, and uniformly distributed α is the sum of phase of X and H:

$$p(a,\alpha) = \frac{a}{\pi\lambda_X} exp(-\frac{a^2}{\lambda_X}) \tag{4}$$

The measurement Y is complex Gaussian with mean $a|H_{rep}|e^{j\alpha}$:

$$p\left(Y\,|a,\alpha,|H|_{rep}\right) = \frac{1}{\pi\lambda_D} exp(-\frac{1}{\lambda_D}\left|Y - a|H_{rep}|e^{j\alpha}\right|^2) \tag{5}$$

Therefore the MMSE estimator of $|X|$ is

$$\begin{aligned} & |\hat{X}|_{MMSE-SA,det} \\ = \;& E[|X||Y,|H|_{\text{rep}}] \\ = \;& \frac{\frac{1}{|H|_{\text{rep}}}\frac{\int_0^\infty a^2|H|_{\text{rep}}^2 exp(-(\frac{a^2|H|_{\text{rep}}^2}{\lambda_X|H|_{\text{rep}}^2}+\frac{a^2|H|_{\text{rep}}^2}{\lambda_D}))}{\int_0^\infty a|H|_{\text{rep}} exp(-(\frac{a^2|H|_{\text{rep}}^2}{\lambda_X|H|_{\text{rep}}^2}+\frac{a^2|H|_{\text{rep}}^2}{\lambda_D}))}}{\cdots \quad \frac{\frac{1}{2\pi}\int_0^{2\pi} exp(2a|H|_{\text{rep}}\frac{Y}{\lambda_D}cos\beta)d\beta da_(|H|_{\text{rep}})}{\frac{1}{2\pi}\int_0^{2\pi} exp(2a|H|_{\text{rep}}\frac{Y}{\lambda_D}cos\beta)d\beta da_(|H|_{\text{rep}})}} \\ = \;& \frac{1}{|H|_{rep}}|H\hat{X}|_{EM84} \end{aligned} \tag{6}$$

where $|\hat{X}|_{EM84}$ is the Ephraim-Malah MMSE-SA estimator [1]. Note that the phase response α, which is the sum of the channel and source phase, can be marginalized, therefore we can have the intuitive simple way of getting MMSE estimation as in (6). Similarly MMSE-LSA estimator can be obtained as

$$\begin{aligned} \ln|\hat{X}|_{MMSE-LSA,det} &= [\ln|X|\,|Y,|H|_{rep}] \\ &= \ln|H\hat{X}|_{EM85} - \ln|H|_{rep} \end{aligned} \tag{7}$$

where $|\hat{X}_{EM85}|$ is the Ephraim-Malah MMSE-LSA estimator [2].

4. OPTIMAL ESTIMATORS: PROBABILISTIC ROOM RESPONSE

Above the Schroeder frequency, we can assume that H is a random variable with Gaussian pdf. In this case, optimal estimators should be based on the pdf of H instead of a deterministic representative H_{rep}. Because many modes overlap at each frequency, by the central limit theorem, we can assume that H has a complex Gaussian pdf, and therefore $|H|$ is a Rayleigh random variable. The MMSE estimator is

$$\begin{aligned} & |\hat{X}|_{MMSE-SA,prob} \\ = \;& E\left[|X|\,|Y\right] = E\left[E\left[|X|\,|Y,H\right]\right] \\ = \;& \int_0^\infty |\hat{X}|_{MMSE-SA,det}p(|H|)d|H| \end{aligned} \tag{8}$$

Because we have been unable to analytically integrate the equation above, a novel ML estimator followed by the conventional optimal filter is proposed.

4.1. ML Estimator as a Sufficient Statistic

Balan and Rosca [6] demonstrated that the ML estimate of X given H, $X_{ML|H}$, is a sufficient statistic for optimal estimation of any function $f(X)$:

$$E\left[f(X)\,|Y,H\right] = E\left[f(X)\,\middle|X_{ML|H}\right] \tag{9}$$

where

$$X_{ML|H} = (H^*H)^{-1}H^*Y \tag{10}$$

In our formulation, H is a random variable, therefore $X_{ML|H}$ is also a random variable. We propose a two-step procedure in which X, and H are jointly estimated according to

$$(X_{ML}, H_{ML}) = \arg\max p(Y\,|X,H,\lambda_D) \tag{11}$$

and then MMSE estimates of functions $f(X)$ are computed. Note that by including $\lambda_{D,ML}$ in (11) the noise variance λ_D is also able to be jointly estimated, but in this paper we simply assume that we can obtain λ_D without this ML estimation as in the previous methods [1, 2].

4.2. EM algorithm

The expectation-maximization algorithm (EM) is an iterative method to get the ML estimates of parameters [7]. Hidden parameters may be defined to include the expectation and variance of H. Define (Y,H) to have pdf

$$p(Y, H|X, \lambda_D) = p(Y|H, X, \lambda_D)p(H). \quad (12)$$

Assume that the noise D has a zero mean Gaussian pdf of variance λ_D, and that H is a zero mean Gaussian with variance λ_H. The objective function has the form below.

$$Q(X, X^{(i-1)}) = E\left[\ln p(Y, H|X, \lambda_D) \,\middle|\, Y, X^{(i-1)}\right] \quad (13)$$

From Eq. (12),

$$Q(X, X^{(i-1)}) = E\left[\ln p(Y|H, X, \lambda_D) \,\middle|\, Y, X^{(i-1)}\right] \quad (14)$$

because $p(H)$ has nothing to do with X. The EM algorithm has two steps: "Expectation" and "Maximization."

<EXPECTATION STEP>

$$\begin{aligned} & E[H^*|Y, X^{(i-1)}] \\ = \; & E[H^*] + \frac{\text{Cov}(H^*, Y^*|X^{(i-1)})}{\text{Cov}(Y^*|X^{(i-1)})}(Y^* - E[Y^*]) \\ = \; & \frac{\lambda_H X^{(i-1)} Y^*}{|X^{(i-1)}|^2\lambda_H + \lambda_D} \end{aligned} \quad (15)$$

where this is obtained from $Y = HX^{i-1} + D^{i-1}$.

$$\text{Cov}[H^*|Y^*, X^{(i-1)}] = \frac{\lambda_H \lambda_D}{|X^{(i-1)}|^2\lambda_H + \lambda_D} \quad (16)$$

and

$$\begin{aligned} E[H^*H|Y, X^{(i-1)}] &= \text{Cov}[H^*|Y^*, X^{(i-1)}] \\ &+ |E[H^*|Y^*, X^{(i-1)}]|^2 \end{aligned} \quad (17)$$

<MAXIMIZATION STEP>

$$X^i = \frac{E[H^*|Y^*, X^{(i-1)}]}{E[H^*H|Y^*, X^{(i-1)}]}Y \quad (18)$$

where this is obtained from $\frac{\partial Q}{\partial X} = 0$, and note that X is complex. Update formulas for $\lambda_{D,ML}$ may be similarly derived.

Multiple measurements around a target position may be modeled as independent and identically distributed. In this case, the Q function is:

$$Q(X, X^{(i-1)}) = \prod_{j=1}^{N} E\left[\ln p(Y_j, H_j|X, \lambda_D) \,\middle|\, Y_j, X^{(i-1)}\right], \quad (19)$$

and the maximization step for X^i is (20).

$$X^i = \frac{\sum_{j=1}^{N} E[H_j^*|Y_j^*, X^{(i-1)}]Y_j}{\sum_{j=1}^{N} E[H_j^* H_j|Y_j^*, X^{(i-1)}]} \quad (20)$$

5. EXPERIMENTAL EVALUATION

In this experimental evaluation, we focus on the newly derived EM algorithm for the sufficient statistic X_{ML}, because theoretically derived optimal estimator for deterministic room response has already been evaluated in the previous researches [4, 5]. However, note that the theoretical consideration given in this paper is missing there.

5.1. Experimental setup

To verify the proposed EM algorithm, 50 room impulse responses (RIRs) are simulated with randomly chosen position of the source and 50 receivers using conventionally used room simulation method "image-method" [8, 9]. A shoebox type of acoustic space has been used, the dimension of which is 6.25 m $\times$ 3.75 m $\times$ 2.5 m and volume of which is 58.59 m^3. The speed of sound, the average absorption coefficient and the reverberation time are set to 343 m/s, 0.45, and 0.25, respectively.

Source was one copy of the speech waveform "three," extracted from TIDigits [10]. Fig. 1(a) shows the source spectrum (an utterance of the word "three") only contaminated by additive noise, where signal to noise ratio was about 12 dB. 50 measurements have been simulated not only by convoluting the source signal with the 50 RIRs but also by adding noise. Fig 1(b) shows one of the measured responses.

5.2. Results

At each measurement location, the sufficient statistic X_{ML} can be estimated using EM. In this experiment, we use multiple measurements to obtain X_{ML} more accurately. Fig. 1(c) shows the source spectrum $|\hat{X}|_{ML}$ estimated using the proposed EM-based algorithm. Above about 200Hz, the estimated source spectrum above the noise floor is almost same as the source spectrum above the noise floor. In other words, we could successfully eliminate the effect of the room responses in the maximum likelihood sense, and this result is the sufficient statistic to the classical MMSE speech enhancement algorithms.

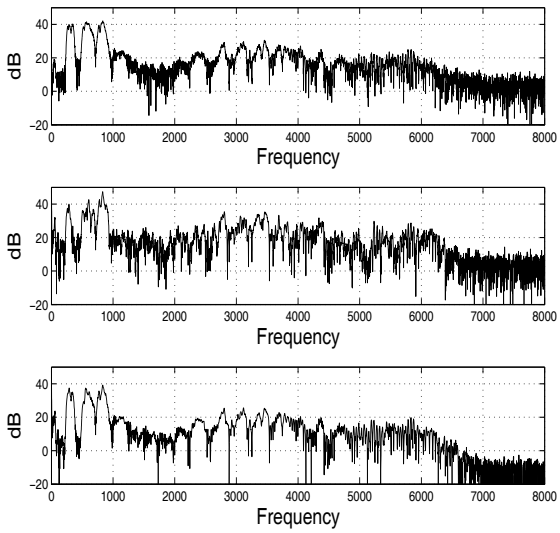

Fig. 1. EM-ML estimate result (a) Source spectrum (an utterance of the word "three") only contaminated by additive noise: SNR is about 12 dB, (b) One of the measurements: Source spectrum contaminated by RIR as well as additive noise, (c) EM ML estimated spectrum: sufficient statistic X_{ML}

Below 200Hz, the estimation starts to break down, apparently because the room response is not a Gaussian random variable at lower frequencies. Although the range between 100 - 200Hz is above the Schroeder frequency, it seems that the modes of the room response do not overlap thickly enough, in this frequency range, for the response to approach a Gaussian distribution.

6. CONCLUSION

In this paper MMSE optimal estimators of a signal which has been contaminated by convolutive noise as well as additive noise are defined. In the low frequency range (below the Schroeder frequency), the room response is estimated using a single representative measured response. In the high frequency range, the MMSE spectral estimator is expressed as the conditional expectation of $|X|$ given knowledge of the maximum likelihood estimate, $X_{ML|H}$. Rather then integrating over the pdf of random variable $X_{ML|H}$, we approximate $p(X_{ML|H})$ using a point distribution centered at the joint maximum likelihood estimates X_{ML} and H_{ML}; these joint maximum likelihood estimates are computed using EM.

7. REFERENCES

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time amplitude estimator," *IEEE Trans. on Acous., Speech, and Signal process.*, vol. ASSP-32, pp. 1109–1121, December 1984.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. on Acous., Speech, and Signal process.*, vol. ASSP-33, pp. 443–445, April 1985.

[3] H. Kuttruff, *Room Acoustics Fourth edition*, SPON PRESS, New York, NY, 2000.

[4] L.-H. Kim, J.-S. Lim, C. Choi, and K.-M. Sung, "Equalization of low frequency response in automobile," *IEEE Trans. on Consumer Electronics*, vol. 49, pp. 243–252, February 2003.

[5] S. Bharitkar and C. Kyriakakis, "Cascaded fir filters for multiple listener low frequency room acoustic equalization," in *Proc. Intern. Conf. Acoust., Speech, and Signal Proc. (ICASSP)*, 2006.

[6] R. Balan and J. Rosca, "Microphone array speech enhancement by bayesian estimation of spectral amplitude and phase," in *Proc. Sensor Array and Multichannel Signal Process. Workshop*, 2002.

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Ann. of the Royal Stat. Soc.*, pp. 1–38, December 1977.

[8] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, pp. 943–950, 1979.

[9] Douglas R. Campbell, Kalle J. Palomaki, and Guy J. Brown, "Roomsim, a matlab simulation of shoebox room acoustics for use in teaching and research," in *http://media.paisley.ac.uk/ campbell/Roomsim/*, last visited on January 3, 2008.

[10] R. G. Leonard, "A database for speaker-independent digit recognition," in *Proc. Intern. Conf. Acoust., Speech, and Signal Proc. (ICASSP)*, 1984.