MAXIMUM-LIKELIHOOD-BASED CEPSTRAL INVERSE FILTERING FOR BLIND SPEECH DEREVERBERATION

Kshitiz Kumar[1] *and Richard M. Stern*[1,2]

Department of Electrical and Computer Engineering[1]
Language Technologies Institute[2]
Carnegie Mellon University, Pittsburgh, PA 15213
Email: {kshitizk, rms}@cs.cmu.edu

ABSTRACT

Current state-of-the-art speech recognition systems work quite well in controlled environments but their performance degrades severely in realistic acoustical conditions in reverberant environments. In this paper we build on the recent developments that represent reverberation in the cepstral feature domain as a filtering operation and we formulate a maximum likelihood objective to obtain an inverse reverberation filter. We show analytically that the optimal inverse filter can be approximately obtained under certain assumptions about the corresponding clean speech signal. We demonstrate that our approach reduces the relative gap in word error rate by 30 percent in large as well as small reverberation times.

Index Terms— Speech recognition, reverberation, blind deconvolution, maximum likelihood

1. INTRODUCTION

Current state-of-the-art automatic speech recognition (ASR) systems perform quite well in controlled environments when the speech signals are reasonably clean, but real-life environments are far less controlled. ASR accuracy deteriorates significantly in the presence of noise, interference, and reverberation. In this paper we study the problem of reverberation and seek to provide greater robustness to its effects.

ASR has a long history. While a number of algorithms have been successfully developed for robustness to additive noise (*e.g.* [1, 2]), reverberation remains a challenging problem [3]. Reverberation is a phenomenon in which delayed and attenuated versions of a signal are added to itself. It is typically modeled as a linear filtering of signals in the time domain. Compensation for reverberation becomes especially difficult because the room characteristics and hence the linear filter modeling reverberation change as people and other objects move around in rooms, thus requiring the compensating algorithms to be blind to the actual reverberation filter. Cepstral mean normalization (CMN) is a ubiquitous algorithm for spectral normalization that if implemented correctly can compensate for some effects of reverberation. CMN can be applied in the log-spectral or cepstral domains where the effects of reverberation are represented as an additive shift which can be removed in the long term by CMN. This approach is effective only if the reverberation time (RT) is small with respect to the duration of the feature-analysis window. In most practical settings the RT can be quite long (300-500 msec) and the implicit assumptions in CMN modeling do not hold.

Recently there has been a growing body of research in extending CMN-based modeling. Some approaches such as *long-term log-spectral subtraction (LTLSS)* work on long-duration windows (*e.g.* [4]) while others model reverberation itself as a linear filter (*e.g.* [5, 6, 7, 8]) in the log-spectral or cepstral domain. We build our research on those recent reverberation models in cepstral domain and formulate a problem to construct an inverse filter in cepstral domain to inverse the effects of room reverberation filter. We describe the motivation for a maximum likelihood criterion for estimating the inverse filter parameters, showing analytically that the inverse filter parameters can be optimally and uniquely recovered with some simple assumptions about the original clean signal. Later we extend our approach to dereverberation for ASR.

Various approaches using multiple microphones have also been developed for reverberation. For example, a multiple-microphone-based score fusion procedure has been described in [9]. Many of the multi-microphone approaches require training data from different environments which is infeasible in practice. The solution for reverberation compensation is also usually local and no guarantees are made about performance across different conditions especially when the reverberant environment changes.

The rest of the paper is organized as follows. We first provide a motivation for our approach in Sec. 2. In Sec. 3 we extend the approach to the problem of ASR. Sections 4 and 5, respectively, present our experiments and results and discussions of the findings. Sec. 6 summarizes this study.

2. MOTIVATION FOR THE MAXIMUM LIKELIHOOD CRITERION FOR ESTIMATING INVERSE FILTERS

In this work we seek to improve the robustness of ASR systems with respect to reverberation. Reverberation is conventionally modeled as a finite impulse response (FIR) linear time-invariant (LTI) system. Using the FIR representation of reverberation, a typical approach for reverberation compensation is to design a system which acts as an inverse for the reverberating LTI system. Nevertheless, the design of such an inverse system is difficult because the time domain reverberation filter is generally both unknown and potentially non-invertible. In this work, we propose the estimation of an inverse system to compensate for reverberation using a maximum likelihood (ML) criterion [10]. ML only requires knowledge of the probability density function (pdf) from which the signals are drawn, which can be obtained from a small amount of training data. ML transforms the reverberated signals into a space from which clean signals originated and thus dereverberates the signal.

We first demonstrate analytically the merit of the ML criterion through two simple illustrations. We show that under certain assumptions about the original signal it is possible to approximately estimate the optimal inverse LTI parameters from the recordings of reverberant input signal. We demonstrate our approach for both all-zero and all-pole inverse systems.

This work was supported by NSF Grant IIS-0916918.

978-1-4244-4296-6/10/$25.00

2.1. Inverse FIR Filter

In this illustration we formulate a simple reverberation problem and demonstrate that we can invert the effects of reverberation with an estimated all-zero filter. We assume that the reverberated signal x can be represented in terms of a convolution of the original unobserved signal s and an FIR filter H, which models the reverberation. We further assume that the the original signal s is white and Gaussian, with zero mean and an autocorrelation that is the Kronecker delta function. We assume that the filter H has only two taps (and hence only a single delay tap). The assumption that the reverberation filter is only of length 2 appears very restrictive at first but we can easily overcome this restriction by applying the approach in each of multiple narrow sub-bands. We assume that a large number of narrow sub-band versions of the H filter can approximate the actual H filter. Note that these assumptions are only for the present illustrations; we will allow the number of filter taps to be unconstrained in the actual ASR problem. We formalize our assumptions as follows:

$$\begin{aligned} s[n] &\sim N(0,1), \quad \text{original signal} \\ H(z) &= 1 + h\,z^{-1}, \quad \text{reverberation filter} \\ x[n] &= s[n] * h[n] = s[n] + h\,s[n-1], \ \text{reverberated signal} \end{aligned} \tag{1}$$

Next we formulate our problem in terms of designing a filter that operates on the reverberated signal x using a log-likelihood criterion with respect to the pdf of s to design the inverse filter parameters. In Eq. (2) below, P denotes the putative inverse FIR filter and y is the estimated dereverberated signal.

$$\begin{aligned} P(z) &= 1 + p\,z^{-1} \\ y[n] &= x[n] * p[n] = x[n] + p\,x[n-1] \end{aligned} \tag{2}$$

The filter parameter p is estimated by maximizing L, the likelihood of y with respect to the pdf of s.

$$L = \log\Pi_n \frac{1}{\sqrt{2\pi}} \exp(\frac{-y^2[n]}{2}) \tag{3}$$

The above can be simplified to minimizing

$$\mathcal{L} = E[y^2[n]] \tag{4}$$

where we replaced summation by expectation and ignored positive constants under the operation. The optimal filter parameter is obtained by differentiating $\mathcal{L}$ with respect to the unknown p.

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial p} &= 2E[(x[n] + p\,x[n-1])x[n-1]] \\ &= 2(R_{xx}[1] + p\,R_{xx}[0]) \end{aligned}$$

Setting the above to 0, we obtain: $p = -R_{xx}[1]/R_{xx}[0]$. Noting that x is the convolution of s and h, the relationship between the autocorrelation sequence of x and s becomes

$$R_{xx}[n] = R_{ss}[n] * R_{hh}[n] \tag{5}$$

It can be shown that

$$R_{xx}[n] = [h,\ 1+h^2,\ h], \quad n \in [-1, 0, 1] \tag{6}$$

from which we obtain $p = -h/(1+h^2)$. Next, assuming that $h \ll 1$ and making the first-order approximation of neglecting the squared term for h, we obtain

$$p \approx -h \tag{7}$$

The assumption of $h \ll 1$ holds because we work on narrow sub-bands of H. So far we had showed that under the assumptions in Eq. (1), we can devise a log-likelihood criterion to invert the H filter by $P = [1 \ \ -h]$, which indeed is expected to be the inverse of H under first-order approximations. Finally we note that Eq. (1) did not include an explicit gain term for the reverberation filter H. While the maximum likelihood procedure described cannot be used to estimate the gain term, gain inversion can be achieved via variance normalization.

2.2. Inverse IIR Filter

In the illustration in Sec. 2.1 we showed that we can estimate an inverse reverberation filter in terms of a FIR filter. In this illustration, we start with the same assumptions as in Eq. (1) but we model the inverse filter as an all-pole IIR filter, showing that we can approximately estimate the optimal inverse filter parameters. Specifically, we assume that the inverse filter P and the dereverberated signal y are of the form:

$$\begin{aligned} P(z) &= \frac{1}{1 + p\,z^{-1}} \\ y[n] &= x[n] * p[n] = x[n] - p\,y[n-1] \end{aligned} \tag{8}$$

Following the same principles as in Sec. 2.1 we obtain

$$\frac{\partial \mathcal{L}}{\partial p} = 2E[(x[n] - p\,y[n-1])y[n-1]]$$

As before, it can be shown that $p = R_{xy}[1]/R_{yy}[0]$ and

$$R_{xy}[1] = R_{xx}[1] = h, \quad R_{yy}[0] = \frac{1 + h^2 - 2p\,h}{1 - p^2}$$

from which we obtain $p = h(1-p^2)/1+h^2-2p\,h$. As in Sec. 2.1, if we can assume that $h \ll 1$ and $p \ll 1$,

$$p \approx h \tag{9}$$

Hence, the estimated compensation filter will be $P(z) = 1/(1 + hz^{-1})$, which is indeed the inverse of the filter H.

Sections 2.1 and 2.2 illustrated the maximum likelihood formulation for estimating a filter (FIR or IIR) that inverts the effects of reverberation. The illustrations showed analytically that our approach is well founded and can approximately guarantee the optimal performance under certain assumptions. These assumptions, of course, do not hold for speech signals, we relax some of those assumptions in Sec. 3 which follows, and we extend the approach to ASR. While analytical verification of our approach for realistic reverberant environments is not tractable, we validate our approach through the experiments and results in Sec. 5.

3. MAXIMUM-LIKELIHOOD-BASED INVERSE FILTERING (MAX LIFE)

In Sec. 2 we formulated the problem of reverberation compensation in terms of obtaining an appropriate inverse filter, proposing the use of a maximum likelihood criterion for obtaining that inverse filter. We demonstrated that the approach can approximately estimate the optimal inverse filter parameters. In the present section we extend our approach for reverberation compensation for speech data, referring to the extended approach as *Maximum Likelihood based Inverse Filtering* (Max-LIFE). There has been a great deal of recent research in modeling reverberation in the log-spectral or cepstral domain, including the characterization of reverberation as linear filtering in the cepstral domain (*e.g.* [5, 6, 7, 8]). These approximations extend the earlier representations of reverberation as a simple additive shift in the log-spectral or cepstral domains. Continuing along these lines we

seek to design an inverse reverberation filter that does not make any *a priori* assumptions about the nature of the actual room reverberation filter. We formulate a maximum likelihood objective function which requires the pdfs of the features of clean speech, which can be easily obtained from training data. The likelihood objective is expected to guide the features to the space from which the clean features originate, thereby dereverberating the features. In Sec. 2.1 the pdf was assumed to be a single Gaussian density in Eq. (1). Because a single Gaussian density is insufficient for real speech applications, we extend the pdf to be a Gaussian mixture model (GMM). The GMM is trained from a pool of clean speech features. Similarly, the number of filter taps to model reverberation was assumed to be 2 in the discussion of Sec. 2, but for practical ASR the number of filter taps modeling reverberation will need to be unconstrained.

We begin by initially assuming that the speech features are unidimensional, and we subsequently extend the model to multi-dimensional features by individually applying the unidimensional approach to different features. While we illustrate these developments only for the inverse IIR filter parameters, the approach can easily be adapted for the FIR filters. Assuming that the all-pole IIR filter de-reverberating P is M taps long, the reverberation-compensated features become:

$$y[n] = x[n] - \sum_{m=1}^{M-1} p[m]y[n-m] \tag{10}$$

where n indicates the feature frame index, with 10 ms between frames. The parameters that describe P are obtained by maximizing the log-likelihood with respect to the GMMs for speech. Specifically, $\mathbf{P} = \arg\max_P L$, where the log-likelihood L for the compensated features $\mathbf{y}$ is:

$$L = \frac{1}{N_y}\sum_{j=1}^{N_y} log\left(\sum_i \frac{w_i}{\sqrt{2\pi\sigma_i^2}}\exp\left(-\frac{(y[j]-\mu_i)^2}{2\sigma_i^2}\right)\right) \tag{11}$$

The GMM parameters are represented by the set $\{w_i, \mu_i, \sigma_i\}$ with N_w being the number of Gaussian densities and N_y being the length of the unidimensional feature y or equivalently the number of feature frames. For ease of writing and understanding the equations we define:

$$\gamma_i^j = \frac{w_i}{\sqrt{2\pi\sigma_i^2}}\exp\left(-\frac{(y[j]-\mu_i)^2}{2\sigma_i^2}\right),\ \gamma^j = \sum_{i=1}^{N_w}\gamma_i^j,\ L = \frac{1}{N_y}\sum_{j=1}^{N_y}\gamma^j$$

Using the above definitions we maximize Eq. (11) by gradient ascent via its partial derivative with respect to the parameters in P. It can be shown that:

$$\frac{\partial L}{\partial p[m]} = \frac{1}{N_y}\sum_{j=1}^{N_y}\sum_{i=1}^{N_w}\frac{\gamma_i^j}{\gamma^j}\ \frac{(y[j]-\mu_i)\ y[j-m]}{\sigma_i^2} \tag{12}$$

Next, we iteratively obtain the parameters for $p[m]$:

$$\hat{p}[m] = p[m] + \nu\frac{\partial L}{\partial p[m]} \tag{13}$$

where ν is a small-valued learning-rate parameter. The filter update term in Eq. (12) provide a deep understanding into the evolution of P. Summing over the j terms for a fixed i in Eq. (12) results in the update for $p[m]$ becoming proportional to the m^{th} auto-correlation sequence of y. Summing over the i terms for a fixed j in Eq. (12) results in $p[m]$ becoming proportional to the summed and weighted likelihoods of γ_j^i. Thus the overall filter updates are proportional to the "likelihood-weighted" auto-correlation sequences of y.

Note that Eq. (12) requires knowledge of $y[j]$ which in turn depends on P in Eq. (10), so $y[j]$ will also be updated after each iteration of P. We refer to the IIR filter as estimated above as *LIFE-IIR*. Update equations for the corresponding inverse FIR filter can similarly be obtained, and these filters will be referred as *LIFE-FIR*.

3.1. The Top-1 Approximation for Filter Updates

The filter update described in Eq. (13) may be simplified through suitable approximations. A common approximation in GMMs is to replace the overall GMM likelihood score in Eq. (11) by the top-scoring Gaussian density among the set of Gaussian mixtures. This approximation, referred to as the Top-1 approximation, results in:

$$\begin{aligned}\gamma^j &= \sum_{i=1}^{N_w}\gamma_i^j \approx \gamma_{i_*}^j, \quad i_* = \arg\max_i \gamma_i^j \\ \frac{\partial L}{\partial p[m]} &= \frac{1}{N_y}\sum_{j=1}^{N_y}\frac{(y[j]-\mu_{i_*})\ y[j-m]}{\sigma_{i_*}^2}\end{aligned} \tag{14}$$

Note that i_* is a function of j in Eq. (14). This approximation is more valid for sparsely-distributed features in terms of the Gaussian densities where only the top-scoring density can adequately describe the overall feature score. A Top-N approximation could be similarly derived by approximating Eq. (11) with the top N Gaussians.

4. EXPERIMENTS

We applied our dereverberation experiments to the DARPA Resource Management (RM) Database with 1600 training utterances and 600 test utterances. The RM database was collected in clean conditions and served as the clean database. A reverberated database was obtained by convolving the clean RM database with simulated room impulse responses obtained using the RIR package based on the image method[1] for different room reverberation times (RT). We used nominal room dimensions of $5 \times 4 \times 3m$ for simulations, with a single microphone located at the center of the room, with 1 m between the source and microphone. Conventional 13-dimensional Mel frequency cepstral coefficients (MFCC features) were derived from the speech signal and compensation was applied to these features. The window length was 25.6 msec with a frame period of 10 msec. A GMM with 32 densities was trained on the clean training features to model the speech features, and these densities were required to implement Eq. (11). 20 taps were used for both the FIR and IIR inverse filters, and the learning-rate parameter ν being 0.01. The estimated inverse filter tap weights were found to converge very rapidly in actual ASR experiments, typically within 5-10 iterations. The overall approach was computationally very efficient.

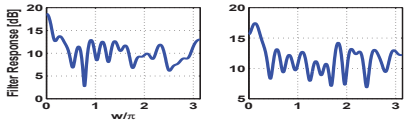
We used CMN to remove any constant additive shift in the cepstral features as well as Cepstral Post Filtering [5] to partially decorrelate the features. We noted in Sec. 2 that our approach works best if the original clean signal is completely uncorrelated. This is, however, an unrealistic assumption for features extracted from real-speech, and introducing some decorrelation through CPF helps LIFE-filters, as shown with the results presented below. We used the [2] for ASR training and decoding. All ASR experiments were conducted using 39-dimensional feature vectors obtained by appending delta and double-delta features to the parent 13-dimensional feature. The ASR training states consisted of 8 Gaussian Mixtures. The ASR language model was a bigram word model.

[1] http://2pi.us/rir.html

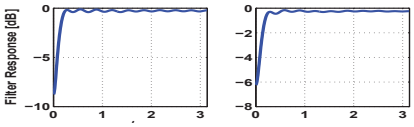
[2] The SPHINX open source speech recognition engines, available online at http://cmusphinx.sourceforge.net/html/cmusphinx.php

5. RESULTS

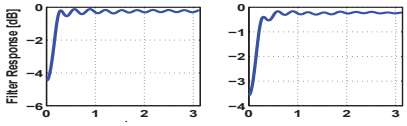
We present our results in this section. We first consider typical room reverberation filter responses in the cepstral domain along with the corresponding inverse filters as estimated in Sec.3. Figure 1(a) plots the room reverberation filter response for the 4^{th} and 5^{th} cepstral features. We note that the reverberation filter has a lowpass characteristic. This is expected, as reverberation essentially smears sound along time. Figures 1(b) and (c), respectively, plot the estimated FIR and IIR dereverberation filters. Ideally we would like these estimated frequency responses to be the exact inverse of those in Fig. 1(a). Although, our estimates do not achieve this ideal, we note that the inverse of a lowpass reverberation filter should have a highpass characteristic, which indeed is the case for the estimated FIR and IIR filters in Figs. 1(b)(c). Specifically, we note that the reverberation filter introduces an attenuation of 5 to 7 dB at higher frequencies which is close to the gain introduced at higher frequencies by the inverse filters. Figs. 1(a)(b)(c) thus validate our inverse filtering approach, at least in the broad sense.



(a) Reverb. filters in the cepstral domain.



(b) Estimated FIR dereverberation filter.



(c) Estimated IIR dereverberation filter.

Fig. 1. Filter responses in the cepstral domain, illustrated for the 4^{th} cepstral coefficient in the left panel and the 5^{th} cepstral coefficient in the right panel.

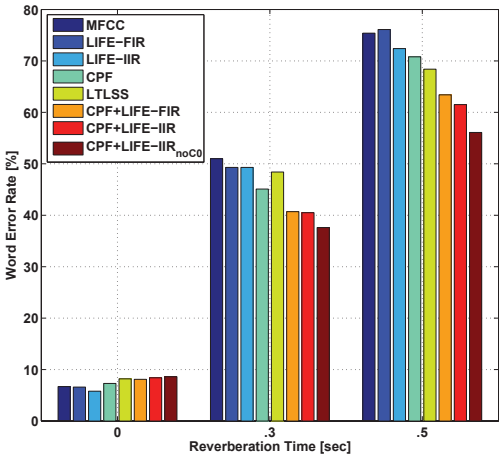


Fig. 2. WER Comparisons for LIFE family of dereverberation filters

Figure 2 summarizes ASR word error rates (WER) results for various RTs. All features include CMN pre-processing. The Top-1 approximation was used in updating the estimates of the P filters. Extending Top-1 to a Top-N approximation provided virtually no improvement in WER. We observe that LIFE-FIR and LIFE-IIR provide some improvement over baseline MFCC processing. In contrast, preprocessing the MFCC features using CPF and then designing LIFE-FIR or LIFE-IIR filters provides a very significant improvement. The strong boost in performance due to CPF preprocessing can be understood from the discussion in Sec. 2, which showed that LIFE filters work best for uncorrelated input sequences. CPF attempts to decorrelate the cepstral sequence and thus assists the LIFE filters. LIFE filters also provided a 15-20% relative improvement over LTLSS [4]. LTLSS was applied on a window length of duration 1 s. We also note that the 0^{th} cepstral feature (C0) is the most affected by reverberation. Our experiments on C0 also revealed that it is not helpful for ASR, at least in our experiments. Removing C0 provided additional gains for the LIFE filters, producing a 25% relative reduction compared to baseline results. Finally, we note that the use of CPF+LIFE processing reduces the difference between the WER obtained for de-reverberated speech and clean speech by about 30% compared to baseline MFCC processing.

6. CONCLUSIONS

In this study we considered the problem of dereverberation for ASR. We motivated and developed a maximum-likelihood-based inverse filtering technique for dereverberation. We showed analytically that the approach approximately estimated the optimal inverse filter under certain assumptions on the signal and its pdf. Our approach is blind to the actual nature of room characteristics and does not require any operator-assisted information. We developed inverse filters for both all-zero and all-pole filters. We validated our approach in reverberant environments, obtaining up to a 25% relative decrease in WER compared to the baseline, closing the WER gap by 30% in reverberant environments.

7. REFERENCES

[1] P. J. Moreno, B. Raj, and R. M. Stern, "A vector taylor series approach for environment-independent speech recognition," *Proc. ICASSP*, 1996.

[2] R. M. Stern, B. Raj, and P. J. Moreno, "Compensation for environmental degradation in automatic speech recognition," *Proc. of the ESCA Workshop on RSRUCC*, 1997.

[3] E. Habets, *Single- and multi-microphone speech dereverberation using spectral enhancement*, Ph.D. thesis, TU Eindhoven, 2007.

[4] D. Gelbart and N. Morgan, "Double the trouble: handling noise and reverberation in far-field automatic speech recognition," *Proc. ICSLP*, pp. 2185–2188, 2002.

[5] K. Kumar and R. M. Sterm, "Environment-invariant compensation for reverberation using linear post-filtering for minimum distortion," *Proc. IEEE ICASSP*, 2008.

[6] A. Sehr and W. Kellermann, "A new concept for feature-domain dereverberation for robust distant-talking asr," *Proc. IEEE ICASSP*, pp. IV–369–IV–372, 2007.

[7] H. Kameoka, T. Nakatani, and T. Yoshioka, "Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms," *Proc. IEEE ICASSP*, pp. 45–48, 2009.

[8] A. Krueger and R. Haeb-Umbach, "Model based feature enhancement for automatic speech recognition in reverberant environments," *Proc. InterSpeech*, pp. 1231–1234, 2009.

[9] Q. Jin, T. Schultz, and A. Waibel, "Far-field speaker recognition," *IEEE Tran. on ASLP*, vol. 15, No.7, pp. 2023–2032, 2007.

[10] Jerry M. Mendel, *Maximum-Likelihood Deconvolution: A Journey into Model-Based Signal Processing*, Sprinter-Verlag, 1990.