SPEECH DEREVERBERATION USING A LEARNED SPEECH MODEL

Dawen Liang∗
LabROSA, Dept. of Electrical Engineering
Columbia University
dliang@ee.columbia.edu

Matthew D. Hoffman, Gautham J. Mysore
Adobe Research
{mathoffm, gmysore}@adobe.com

ABSTRACT
We present a general single-channel speech dereverberation method based on an explicit generative model of reverberant and noisy speech. To regularize the model, we use a pre-learned speech model of clean and dry speech as a prior and perform posterior inference over the latent clean speech. The reverberation kernel and additive noise are estimated under the maximum-likelihood framework. Our model assumes no prior knowledge about specific speakers or rooms, and consequently our method can automatically adapt to various reverberant and noisy conditions. We evaluate the proposed model with both simulated data and real recordings from the REVERB Challenge in the task of speech enhancement and obtain results comparable to or better than the state-of-the-art.

Index Terms— dereverberation, Bayesian modeling, variational inference, non-negative matrix factorization

1. INTRODUCTION
Speech enhancement in the presence of reverberation and noise remains a challenging problem which draws attention from both academic and industrial communities. The REVERB (REVerberant Voice Enhancement and Recognition Benchmark) Challenge which was held last year is a successful attempt to provide a common dataset and evaluation metrics for speech dereverberation research.

In this paper, we present a novel single-channel speech dereverberation method that explicitly models the generative process of reverberant and noisy speech. The model treats the underlying clean speech as a set of latent variables. A generic speech model fit beforehand to a corpus of clean and dry speech is used as a prior on these variables, regularizing the model and making it possible to solve the otherwise underdetermined dereverberation problem. The model is capable of suppressing reverberation without any prior knowledge of or assumptions about the specific speakers or rooms. In fact, our experiments on both simulated and real data show that our approach can work on speech that is quite different than that used to train the speech model—specifically, we will show that a model of North American English speech can be very effective on British English speech.

1.1. Related work
Habets [1] provides a thorough review of various single- and multimicrophone speech dereverberation techniques. Using the categories developed in [1], explicit speech modeling techniques, which exploit the underlying structure of the anechoic speech signal, are particularly relevant to what we propose here.

Attias et al. [2] propose a unified probabilistic framework for denoising and dereverberation of speech signals. Their model shares a similar generative flavor with ours; they train a standard auto-regressive (AR) model on clean and dry speech signals as a speech model, and perform inference in the frequency domain. However, even though the model was designed for both tasks of denoising and dereverberation, the latter was not thoroughly evaluated.

Liang et al. [3] tackle the problem of speech decoloration (e.g. reverberation with short $T_{60}$) based on the product-of-filters (PoF) speech model. In this work, the PoF model is adopted as a strong speech model and a separate set of parameters is added to account for linear distortion effects; this work is the main inspiration behind the model presented in this paper. However, the model in [3] is fundamentally limited to short-time effects (i.e., distortions whose effects are smeared across only one DFT window), and consequently is not capable of suppressing longer reverberation.

Kumar et al. [5] propose a model for denoising and dereverberation in the context of ASR. Their method makes similar assumption to ours that the reverberant speech spectra are the convolution of clean speech spectra and room impulse response spectra. However, unlike what we will present in Section 2.2, they use non-negative matrix factorization (NMF) as a way to model this convolution operation and only put sparsity constraint on the clean speech spectra, instead of using a speech model learned beforehand.

2. PROPOSED APPROACH
The core of our approach is a probabilistic model of a process that generates reverberant and noisy speech. This model begins with a pre-learned model of clean and dry speech, and defines a random process by which clean speech is corrupted. Once we have defined the model, the problem of dereverberation and denoising can be recast as a statistical inference problem. We will first present the overall framework of our model. Then we will conscript a probabilistic non-negative matrix factorization (NMF) model [6] into the proposed framework for use as a model of clean speech, and derive the inference algorithm in detail. Finally, we will discuss possible extensions to other speech models.

2.1. General dereverberation framework
We adopt the notational conventions that upper case bold letters (e.g. $Y$, $X$, and $R$) denote matrices and lower case bold letters (e.g. $y$, $x$, $a$, and $r$) denote vectors. $f \in \{1, 2, \cdots, F\}$ is used to index frequency. $t \in \{1, 2, \cdots, T\}$ is used to index time. $k \in \{1, 2, \cdots, K\}$ is used to index latent components in the pre-learned speech model.
(e.g. NMF model), \( l \in \{0, \ldots, L - 1\} \) is used to index lags in time (we use 0-based indexing for lags).

Given magnitude spectra\(^2\) of reverberant speech \( Y \in \mathbb{R}^{K \times T} \), the general dereverberation model is formulated as follows:

\[
Y_{ft} \sim \mathcal{P} (\sum_j X_{j, f, t} R_{ft} + \lambda_f)
\]

\[
X_{ft} \sim \mathcal{S} (\theta)
\]

\( \mathcal{P}(\cdot) \) encodes the observational model and \( \mathcal{S}(\cdot) \) encodes the speech model. We choose \( \mathcal{P}(\cdot) \) to be a Poisson distribution, which corresponds to the widely used (e.g. in non-negative matrix factorization) pre-learned speech model \( X \) (e.g. stationary background noise). The latent random variables \( \lambda \) corresponds to the widely used (e.g. in non-negative matrix factorization) frequency-dependent additive noise (e.g. stationary background noise). The latent random variables \( X \) represent the spectra of clean and dry speech. The pre-learned speech model \( \mathcal{S}(\cdot) \) (parametrized by \( \theta \)) acts as a prior that encourages \( X \) to look like clean speech. The goal of the inference algorithm is to uncover \( X \), and incidentally to estimate \( \lambda \) from the observed reverberant spectra \( Y \).

Since we assume the reverberant effect comes from a patch of spectra \( R \) instead of a single spectrum (as in [3]), the model is capable of capture reverberation effects that span multiple analysis windows.

2.2. Speech model: probabilistic NMF

Non-negative matrix factorization (NMF)\(^3\) has been used in many speech-related applications, including denoising\([9\ 10]\) and bandwidth expansion\([11]\). Here we use a probabilistic version of NMF with exponential likelihoods, which corresponds to minimizing the Itakura-Saito divergence\([12]\). Concretely, the model is formulated as follows:

\[
Y_{ft} \sim \text{Poisson}(\sum_j X_{j, f, t} R_{ft} + \lambda_f)
\]

\[
X_{ft} \sim \text{Exponential}(c \sum_k W_{jk} H_{kt})
\]

\[
W_{jk} \sim \text{Gamma}(a, a), \quad H_{kt} \sim \text{Gamma}(b, b)
\]

Here \( a \) and \( b \) are model hyperparameters. \( c \) is a free scale parameter that we tune to maximize the likelihood of \( Y \). For the latent components \( W \in \mathbb{R}^{K \times K} \), we assume the posterior distribution \( q(W|X_{\text{clean}}) \) has been estimated from clean and dry speech using procedures similar to what is described in [13]. Therefore, we only need to compute the posterior over the clean and dry speech \( X \) as well as the weights \( H \in \mathbb{R}^{K \times T} \), which we denote as \( p(X, H|Y) \).

2.2.1. A variational EM algorithm

To estimate the reverberation kernel \( R \) and additive noise \( \lambda \), we maximize the likelihood \( p(Y|R, \lambda) \), by marginalizing out the latent random variables \( X \) and \( H \), which yields an instance of expectation-maximization algorithm.

E-step In the E-step, we compute the posterior \( p(X, H|Y) \) under the current value of model parameters. However, this is intractable to compute due to the non-conjugacy of the model. We approximate it via variational inference\([14]\) by choosing the following variational distribution:

\[
q(X, H) = \prod_f \left( \prod_j q(X_{j, f, t}) \prod_k q(H_{kt}) \right)
\]

\[
q(X_{ft}) = \text{Gamma}(X_{ft}; \nu^X_{ft}, \rho^X_{ft})
\]

\[
q(H_{kt}) = \text{GIG}(H_{kt}; \nu^H_{kt}, \rho^H_{kt}, \tau^H_{kt})
\]

GIG denotes the generalized inverse-Gaussian distribution, an exponential-family distribution with the following density:

\[
\text{GIG}(x; \nu, \rho, \tau) = \exp \left( \frac{(\nu - 1) \log x - \rho x - \tau x^2}{2\tau^2/\kappa^{(2)}} \right)
\]

for \( x \geq 0, \rho \geq 0, \) and \( \tau \geq 0 \). \( \kappa(\cdot) \) denotes a modified Bessel function of the second kind. We will see below that using the GIG distribution for \( q(H_{kt}) \) allows us to tune \( q(H) \) using closed-form updates.

We tune the variational parameters \( \nu^X, \rho^X \) and \( \nu^H, \rho^H, \tau^H \) so that the Kullback-Leibler divergence between the variational distribution \( q(X, H) \) and the true posterior \( p(X, H|Y) \) is minimized. This is equivalent to maximizing the following variational objective (Let \( S^f \in \mathbb{R}^{K \times L} \) be a patch \( X_{[t-L+1:t]} \)):

\[
\mathcal{L} \equiv \sum_{f, t} \left\{ Y_{ft} \left( \phi^X_{ft}(\log \lambda_f - \log \phi^X_{ft}) + \sum_j \phi^R_{ft} \text{E}_q[\log X_{j, f, t-1}] \right) \right.
\]

\[
+ \log R_{ft} - \log \phi^R_{ft}) \right) - \lambda_f - \sum_j \text{E}_q[\log X_{j, t-1}] R_{ft} \}
\]

\[
+ \sum_{f, t} \left( \rho^X_{ft} - \frac{(\rho^X_{ft})^2}{\text{E}_q[1/R_{ft}]} \right) \text{E}_q[X_{ft}] \log(c \omega_{ft}) \right. \]

\[
+ (1 - \nu^H) \text{E}_q[\log H_{kt}] + A^F(\nu^H, \rho^H) - \frac{1}{\tau^H} \sum_k \text{E}_q[W_{jk} H_{kt}] \}
\]

\[
+ \sum_{k, t} \left( b - \nu^H \right) \text{E}_q[\log H_{kt}] - (b - \rho^H) \text{E}_q[H_{kt}] - \tau^H \text{E}_q \left[ \frac{1}{\tau^H} \right] \}
\]

\[
+ A^G(\nu^H, \rho^H, \tau^H) \} + \text{const}
\]

where \( A^F(\cdot) \) and \( A^G(\cdot) \) denote the log-partition functions for gamma and GIG distribution, respectively\([15]\). Optimizing over \( \phi^X \) with Lagrangian multipliers, the bound for the first term in Equation [5] is tightest when

\[
\phi^X_{ft} = \frac{\lambda_f}{\lambda_f + \sum_j \exp[\text{E}_q[\log X_{j, t-1}] R_{ft}]} \}

\[
\phi^R_{ft} = \frac{\exp[\text{E}_q[\log X_{j, t-1}] R_{ft}]}{\lambda_f + \sum_j \exp[\text{E}_q[\log X_{j, t-1}] R_{ft}]} \}
\]

The explicit forms are not important here, because for exponential family distributions, taking derivative of log-partition function with respect to the natural parameter yields the expected sufficient statistic.
Similarly, we can optimize over $\phi_{ftk}^X$ and $\omega_{ft}$ and tighten the bound on the second term:

$$\phi_{ftk}^X \propto \left( E_q \left[ \frac{1}{W_{fk} H_{kt}} \right] \right)^{-1} ; \quad \omega_{ft} = \sum_k E_q[W_{fk} H_{kt}]$$  \hspace{1cm} (8)

Given the lower bound in Equation(8) we can maximize $L$ using coordinate ascent, iteratively optimizing each variational parameter while holding all other parameters fixed. To update $\{\nu^X, \rho^X\}$ by taking the derivative of $L$ with respect to $\nu^X$ and $\rho^X$, rather than considering them part of the clean speech model if it can.

Every time the value of variational parameters changes, the scale steps:

$$\nu^X_{ft} = 1 + \sum_l Y_{f,t+l} \phi^X_{f,t+l,l}; \quad \rho^X_{ft} = \frac{1}{c} \left( \sum_k E_q \left[ \frac{1}{W_{fk} H_{kt}} \right] \right)^{-1} + \sum_l R_{ft}.$$  \hspace{1cm} (9)

Similarly, the derivative of $L$ with respect to $\{\nu^H, \rho^H, \tau^H\}$ equals 0 and $L$ is maximized when

$$\nu^H_{kt} = b; \quad \rho^H_{kt} = b + \sum_l E_q[\nu^H_{kl}]; \quad \tau^H_{kt} = \sum_l E_q[X_{ftl}] \frac{(\phi^X_{ftl} - \phi^X_{f,t+l,l})}{\omega_{ft}}.$$  \hspace{1cm} (10)

Every time the value of variational parameters changes, the scale $c$ should be updated accordingly:

$$c = \frac{1}{1 - \tau^{X}_{ft}} \sum_f E_q[X_{ft}] \left( \sum_k E_q \left[ \frac{1}{W_{fk} H_{kt}} \right] \right)^{-1}$$  \hspace{1cm} (11)

Finally, the necessary expectations are ($\psi(\cdot)$ is the digamma function)

$$E_q[X_{ft}] = \frac{\nu^X_{ft}}{\rho^X_{ft}}; \quad E_q[\log X_{ft}] = \psi(\nu^X_{ft}) - \log \rho^X_{ft};$$

$$E_q[H_{kt}] = \frac{K_{\nu+1}(2\sqrt{\rho^H_{kt}})/\sqrt{\tau}}{K_{\nu}(2\sqrt{\rho^H_{kt}})/\sqrt{\tau}}; \quad E_q \left[ \frac{1}{H_{kt}} \right] = \frac{K_{\nu-1}(2\sqrt{\rho^H_{kt}})/\sqrt{\tau}}{K_{\nu}(2\sqrt{\rho^H_{kt}})/\sqrt{\tau}}.$$  \hspace{1cm} (12)

**M-step** In the M-step, given the approximated posterior estimated from the E-step, we can take the derivative of $L$ with respect to $\lambda$ and $\rho$ and obtain the following updates:

$$\lambda_f = \frac{1}{T} \sum_t \phi^X_{ft} Y_{ft}; \quad R_{ft} = \sum_k E_q[X_{ftl}].$$  \hspace{1cm} (13)

The overall variational EM algorithm alternates between two steps:

- In the E-step, the speech model attempts to explain the observed spectra as a mixture of clean speech, reverberation, and noise. In particular, it updates its beliefs about the latent clean speech via updating the variational distribution $q(X)$.
- In the M-step, the model updates its estimate of the reverberation kernel and additive noise given its current beliefs about the clean speech.

A good speech model should assign high probability to clean speech and lower probability to speech corrupted with reverberation and additive noise. The full model therefore has an incentive to explain reverberation and noises using the reverberation kernel and additive noise parameters, rather than considering them part of the clean speech. In other words, the model should try to “explain away” reverberation and noise and leave behind spectra that are likely under the speech model if it can.

By iteratively performing E- and M-steps, we are guaranteed to reach a stationary point of the objective $L$. To obtain the dereverberated spectra, we can simply take the expectation of $X$ under the variational distribution. To recover time-domain signals, we adopt the standard Wiener filter based on the estimated dereverberated spectra $E_q[X]$. However, in practice we notice that the Wiener filter aggressively takes energy from the complex spectra due to the crudeness of the estimated dereverberated spectra and produces artifacts. We apply a simple heuristic to smooth $E_q[X]$ by convolving it with an attenuated reverberation kernel $R$, where $R_{f,t} = \alpha R_{ft}$ and $R_{ft} = \alpha R_{ft}$ for $l \in \{1, \cdots, L-1\}$. $\alpha \in (0, 1)$ controls the attenuation level.

### 2.3. Extension with other speech models

As is evident from the general model in Equation(1) the speech model $S(\cdot)$ can take different forms. Here we will briefly describe an alternate speech model: Product-of-Filter ($PoF$) model $[4]$. The PoF model is motivated by the widely used homomorphic filtering approach to audio and speech signal processing $[13]$ and it attempts to decompose the log-spectra into a sparse and non-negative linear combination of “filters”, which are learned from data. Incorporating the PoF model into the framework defined in Equation(1) is straightforward:

$$Y_{ft} \sim \text{Poisson}(\sum_l X_{ft}R_{ft} + \lambda_f)$$

$$X_{ft} \sim \text{Gamma}(\gamma_f, \gamma_f \prod_k \exp(-U_{fk} H_{kt}))$$  \hspace{1cm} (14)

$$H_{kt} \sim \text{Gamma}(\alpha \delta, \alpha \delta)$$

where the filters $U \in \mathbb{R}^{F \times K}$, sparsity level $\alpha \in \mathbb{R}_+^K$, and frequency-dependent noise-level $\gamma \in \mathbb{R}^F_+$ are the PoF parameters learned from clean and dry speech. $H \in \mathbb{R}^{K \times T}$ denotes the weights of linear combination of filters. The inference can be carried out in a similar way as derived in Section $[2]$. Note that both NMF and PoF assume independence between frames, which is unrealistic. This assumption could be relaxed by imposing temporal structure to the speech model, e.g. with a non-negative hidden Markov model $[16]$ or a recurrent neural network $[17]$.

### 3. EXPERIMENTS

We evaluated the proposed model under the REVERB Challenge speech enhancement task. The challenge data comes from two sources: One is simulated reverberant and noisy speech, which is generated by convolving clean utterances from WSJAMO corpus $[18]$ with measured room impulse responses and then adding measured background noise signals; The other is real recording from MC-WSJ-AV corpus $[19]$, which is a re-recorded version of WSJAMO in a meeting room environment.

For simulated data, three rooms with increasing reverberation lengths ($T_{60}$’s of the three rooms are 0.25s, 0.5s, 0.7s, respectively) are used; for each room, two microphone positions (near and far) are adopted, which in total provides six different evaluation conditions.

For real data, three rooms with increasing reverberation lengths ($T_{60}$’s of the three rooms are 0.25s, 0.5s, 0.7s, respectively) are used; for each room, two microphone positions (near and far) are adopted, which in total provides six different evaluation conditions.

In the real evaluation, the meeting room has a measured $T_{60}$ of 0.7s. The detailed reverberation and recording specifications can be found on the REVERB Challenge website.

Speech enhancement methods are evaluated by several metrics, including cepstrum distance (CD) $[20]$, log-likelihood ratio (LLR) $[20]$, frequency-weighted segmental SNR (FWSegSNR) $[20]$, and speech-to-reverberation modulation energy ratio (SRMR) $[21]$. For real recordings, only the non-intrusive SRMR can be used.
Since our model processes each utterance separately without relying on any particular test condition, we compared our model with other utterance-based approaches from the REVERB Challenge: Cauchi et al. [22], González et al. [23], Wisdom et al. [24], and Xiao et al. [25]. We trained two exponential NMF speech models with $K = 50$ as the priors used in the dereverberation algorithm: one is from the clean training corpus of WSJCAM0 (British English) and the other is from the TIMIT corpus (American English). In our STFT, we used 1024-sample windows (zero-padded to 2048 samples) with 512-sample overlap. Inference was carried out as described in Section 2.2. We used model hyperparameters $\alpha = b = 0.1$, reverberation kernel length $L = 20$ (640 ms), and attenuation level $\alpha = 0.1$.

The speech enhancement results are summarized in Figure 1. The results are grouped by different test conditions. As we can see, on average our proposed model improves all metrics except LLR over the unprocessed speech by a large margin.

At first glance, our results do not stand out when the reverberant effect is relatively small (Room 1). However, as $T_{60}$ increases, we achieve results close to or better than the best reported in the REVERB Challenge, regardless of microphone position.

Note that our approach performs equally well when using a speech model trained on American English speech and tested on British English speech. That is, our performance is competitive with the state of the art even when we make no use at all of the provided WSJCAM0 training data. This robustness to training-set-test-set mismatch allows our method to be used in real-world applications where we have no prior knowledge about the specific people who are speaking or the room that is coloring their speech. Our ability to do without speaker/room-specific clean training data may also explain the superior performance of our model on the real recording: Xiao et al. [25] hypothesize that their deep-neural-network-based approach may overfit to the rooms in the training set, which is not a problem for our approach.

4. CONCLUSION

In this paper, we propose a general single-channel speech dereverberation model, which follows the generative process of the reverberant and noisy speech. A speech model, learned from clean and dry speech, is used as a prior to properly regularize the model. We adapt NMF as a particular speech model into the general algorithm and derive an efficient closed-form variational EM algorithm to perform posterior inference and to estimate reverberation and noise parameters. We evaluate our model on a speech enhancement task from the REVERB Challenge and obtain promising results on both simulated data and real recording.

There are a few natural directions in which our model can be extended. As pointed out in Section 2.2, a speech model with temporal structure could be adopted. Stochastic variational inference [26] might allow our model to perform real-time/online dereverberation.

5. REFERENCES


