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[1] 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 multi-microphone 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 autoregressive (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) [4] 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. $\mathbf{Y}$, $\mathbf{X}$, and $\mathbf{R}$) denote matrices and lower case bold letters (e.g. $\boldsymbol{y}$, $\boldsymbol{x}$, $\boldsymbol{\lambda}$, and $\boldsymbol{r}$) 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

*This work was performed while Dawen Liang was an intern at Adobe Research.

[1] http://reverb2014.dereverberation.com/

(e.g. NMF model). $l \in \{0, \cdots, L-1\}$ is used to index lags in time (we use 0-based indexing for lags).

Given magnitude spectra[2] of reverberant speech $\mathbf{Y} \in \mathbb{R}_+^{F \times T}$, the general dereverberation model is formulated as follows:

$$\begin{aligned} Y_{ft} &\sim \mathcal{P}(\textstyle\sum_l X_{f,t-l} R_{fl} + \lambda_f) \\ X_{ft} &\sim \mathcal{S}(\theta) \end{aligned} \tag{1}$$

$\mathcal{P}(\cdot)$ 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 methods [7]) generalized Kullback-Leibler divergence loss function.

The model parameter $\mathbf{R} \in \mathbb{R}_+^{F \times L}$ defines a reverberation kernel and $\boldsymbol{\lambda} \in \mathbb{R}_+^F$ defines the frequency-dependent additive noise (e.g. stationary background noise). The latent random variables $\mathbf{X} \in \mathbb{R}_+^{F \times T}$ represent the spectra of clean and dry speech. The pre-learned speech model $\mathcal{S}(\cdot)$ (parametrized by θ) acts as a prior that encourages $\mathbf{X}$ to look like clean speech. The goal of the inference algorithm is to uncover $\mathbf{X}$, and incidentally to estimate $\mathbf{R}$ and $\boldsymbol{\lambda}$ from the observed reverberant spectra $\mathbf{Y}$.

Since we assume the reverberant effect comes from a patch of spectra $\mathbf{R}$ 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) [8] 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:

$$\begin{aligned} Y_{ft} &\sim \text{Poisson}(\textstyle\sum_l X_{f,t-l} R_{fl} + \lambda_f) \\ X_{ft} &\sim \text{Exponential}(c \textstyle\sum_k W_{fk} H_{kt}) \\ W_{fk} &\sim \text{Gamma}(a, a),\ H_{kt} \sim \text{Gamma}(b, b) \end{aligned} \tag{2}$$

Here a and b are model hyperparameters. c is a free scale parameter that we tune to maximize the likelihood of $\mathbf{Y}$. For the latent components $\mathbf{W} \in \mathbb{R}_+^{F \times K}$, we assume the posterior distribution $q(\mathbf{W}|\mathbf{X}_{\text{clean}})$ 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 $\mathbf{X}$ as well as the weights $\mathbf{H} \in \mathbb{R}_+^{K \times T}$, which we denote as $p(\mathbf{X}, \mathbf{H}|\mathbf{Y})$.

2.2.1. A variational EM algorithm

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

E-step In the E-step, we compute the posterior $p(\mathbf{X}, \mathbf{H}|\mathbf{Y})$ 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:

$$\begin{aligned} q(\mathbf{X}, \mathbf{H}) &= \textstyle\prod_t \Big(\prod_f q(X_{ft}) \Big) \prod_k q(H_{kt}) \\ q(X_{ft}) &= \text{Gamma}(X_{ft}; \nu_{ft}^X, \rho_{ft}^X) \\ q(H_{kt}) &= \text{GIG}(H_{kt}; \nu_{kt}^H, \rho_{kt}^H, \tau_{kt}^H) \end{aligned} \tag{3}$$

[2]In this paper, we work in the magnitude spectral domain. For simplicity, we will use "spectra" for the rest of the paper.

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

$$\text{GIG}(x; \nu, \rho, \tau) = \frac{\exp\{(\nu-1)\log x - \rho x - \tau/x\} \rho^{\nu/2}}{2\tau^{\nu/2} \mathcal{K}_\nu(2\sqrt{\rho\tau})} \tag{4}$$

for $x \geq 0$, $\rho \geq 0$, and $\tau \geq 0$. $\mathcal{K}_\nu(\cdot)$ 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 $q(\mathbf{H})$ using closed-form updates.

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

$$\begin{aligned} &\textstyle\sum_t \Big(\mathbb{E}_q[\log p(\boldsymbol{y}_t, \mathbf{S}^t, \boldsymbol{h}_t | \boldsymbol{\lambda}, \mathbf{R})] - \mathbb{E}_q[\log q(\boldsymbol{x}_t, \boldsymbol{h}_t)] \Big) \\ &= \textstyle\sum_{f,t} \mathbb{E}_q[\log p(Y_{ft} | \boldsymbol{s}_f^t, \lambda_f, \boldsymbol{r}_f)] + \sum_{f,t} \mathbb{E}_q \left[\log \frac{p(X_{ft}|\boldsymbol{w}_f, \boldsymbol{h}_t)}{q(X_{ft})} \right] \\ &\quad + \textstyle\sum_{k,t} \mathbb{E}_q \left[\log \frac{p(H_{kt}|b)}{q(H_{kt})} \right] \end{aligned} \tag{5}$$

We cannot compute the expectations in the first and second terms analytically. However, we can compute lower bounds on both of them. For the first term, we can apply Jensen's inequality and introduce auxiliary variables $\phi_{ft}^\lambda \geq 0$ and $\phi_{ftl}^R \geq 0$ where $\phi_{ft}^\lambda + \sum_l \phi_{ftl}^R = 1$; For the second term, we can introduce auxiliary variables $\phi_{ftk}^X \geq 0$ where $\sum_k \phi_{ftk}^X = 1$ and $\omega_{ft} > 0$ to bound it as in [13]. The lower bound of the variational objective in Equation 5:

$$\begin{aligned} \mathcal{L} \stackrel{\triangle}{=} &\textstyle\sum_{f,t} \Big\{ Y_{ft} \Big(\phi_{ft}^\lambda (\log \lambda_f - \log \phi_{ft}^\lambda) + \sum_l \phi_{ftl}^R (\mathbb{E}_q[\log X_{f,t-l}] \\ &\quad + \log R_{fl} - \log \phi_{ftl}^R) \Big) - \lambda_f - \textstyle\sum_l \mathbb{E}_q[X_{f,t-l}] R_{fl} \Big\} \\ &+ \textstyle\sum_{f,t} \Big\{ \Big(\rho_{ft}^X - \sum_k \frac{(\phi_{ftk}^X)^2}{c} \mathbb{E}_q \Big[\frac{1}{W_{fk} H_{kt}} \Big] \Big) \mathbb{E}_q[X_{ft}] - \log(c\, \omega_{ft}) \\ &+ (1 - \nu_{ft}^X) \mathbb{E}_q[\log X_{ft}] + A^\Gamma(\nu_{ft}^X, \rho_{ft}^X) - \tfrac{1}{\omega_{ft}} \textstyle\sum_k \mathbb{E}_q[W_{fk} H_{kt}] \Big\} \\ &+ \textstyle\sum_{k,t} \Big\{ (b - \nu_{kt}^H) \mathbb{E}_q[\log H_{kt}] - (b - \rho_{kt}^H) \mathbb{E}_q[H_{kt}] - \tau_{kt}^H \mathbb{E}_q \Big[\frac{1}{H_{kt}} \Big] \\ &\quad + A^{\text{GIG}}(\nu_{kt}^H, \rho_{kt}^H, \tau_{kt}^H) \Big\} + \text{const} \end{aligned} \tag{6}$$

where $A^\Gamma(\cdot)$ and $A^{\text{GIG}}(\cdot)$ denote the log-partition functions for gamma and GIG distribution, respectively[3]. Optimizing over $\boldsymbol{\phi}$'s with Lagrangian multipliers, the bound for the first term in Equation 5 is tightest when

$$\begin{aligned} \phi_{ft}^\lambda &= \frac{\lambda_f}{\lambda_f + \sum_j \exp\{\mathbb{E}_q[\log X_{f,t-j}]\} R_{fj}}; \\ \phi_{ftl}^R &= \frac{\exp\{\mathbb{E}_q[\log X_{f,t-l}]\} R_{fl}}{\lambda_f + \sum_j \exp\{\mathbb{E}_q[\log X_{f,t-j}]\} R_{fj}}. \end{aligned} \tag{7}$$

[3]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 ϕ^X_{ftk} and ω_{ft} and tighten the bound on the second term:

$$\phi^X_{ftk} \propto \left(\mathbb{E}_q\left[\frac{1}{W_{fk}H_{kt}}\right]\right)^{-1}; \quad \omega_{ft} = \sum_k \mathbb{E}_q[W_{fk}H_{kt}] \tag{8}$$

Given the lower bound in Equation 6, we can maximize $\mathcal{L}$ using coordinate ascent, iteratively optimizing each variational parameter while holding all other parameters fixed. To update $\{\boldsymbol{\nu}^X_t, \boldsymbol{\rho}^X_t\}$ by taking the derivative of $\mathcal{L}$ and setting it to 0, we obtain:

$$\begin{aligned} \nu^X_{ft} &= 1 + \textstyle\sum_l Y_{f,t+l}\phi^R_{f,t+l,l}; \\ \rho^X_{ft} &= \tfrac{1}{c} \cdot \left(\textstyle\sum_k \mathbb{E}_q\left[\frac{1}{W_{fk}H_{kt}}\right]^{-1}\right)^{-1} + \textstyle\sum_l R_{fl}. \end{aligned} \tag{9}$$

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

$$\begin{aligned} &\nu^H_{kt} = b; \quad \rho^H_{kt} = b + \textstyle\sum_f \frac{\mathbb{E}_q[W_{fk}]}{\omega_{ft}}; \\ &\tau^H_{kt} = \textstyle\sum_f \frac{\mathbb{E}_q[X_{ft}]}{c}(\phi^X_{ftk})^2 \mathbb{E}_q\left[\frac{1}{W_{ft}}\right]. \end{aligned} \tag{10}$$

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

$$c = \tfrac{1}{FT}\textstyle\sum_{f,t} \mathbb{E}_q[X_{ft}]\left(\sum_k \mathbb{E}_q\left[\frac{1}{W_{fk}H_{kt}}\right]^{-1}\right)^{-1} \tag{11}$$

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

$$\begin{aligned} &\mathbb{E}_q[X_{ft}] = \frac{\nu^X_{ft}}{\rho^X_{ft}}; \quad \mathbb{E}_q[\log X_{ft}] = \psi(\nu^X_{ft}) - \log \rho^X_{ft}; \\ &\mathbb{E}_q[H_{kt}] = \frac{\mathcal{K}_{\nu+1}(2\sqrt{\rho\tau})\sqrt{\tau}}{\mathcal{K}_{\nu}(2\sqrt{\rho\tau})\sqrt{\rho}}; \; \mathbb{E}_q\left[\frac{1}{H_{kt}}\right] = \frac{\mathcal{K}_{\nu-1}(2\sqrt{\rho\tau})\sqrt{\rho}}{\mathcal{K}_{\nu}(2\sqrt{\rho\tau})\sqrt{\tau}}. \end{aligned} \tag{12}$$

M-step In the M-step, given the approximated posterior estimated from the E-step, we can take the derivative of $\mathcal{L}$ with respect to $\boldsymbol{\lambda}$ and $\mathbf{R}$ and obtain the following updates:

$$\lambda_f = \frac{1}{T}\sum_t \phi^\lambda_{ft} Y_{ft}; \quad R_{fl} = \frac{\sum_t \phi^R_{ftl} Y_{ft}}{\sum_t \mathbb{E}_q[X_{ft}]} \tag{13}$$

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 $q(\mathbf{X})$.
- 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 $\mathcal{L}$. To obtain the dereverbed spectra, we can simply take the expectation of $\mathbf{X}$ under the variational distribution. To recover time-domain signals, we adopt the standard Wiener filter based on the estimated dereverbed spectra $\mathbb{E}_q[\mathbf{X}]$. However, in practice we notice that the Wiener filter aggressively takes energy from the complex spectra due to the crudeness of the estimated dereverbed spectra and produces artifacts. We apply a simple heuristic to smooth $\mathbb{E}_q[\mathbf{X}]$ by convolving it with an attenuated reverberation kernel $\mathbf{R}^*$, where $R^*_{f,0} = R_{f,0}$ and $R^*_{fl} = \alpha R_{fl}$ 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 $\mathcal{S}(\cdot)$ can take different forms. Here we will briefly describe an alternate speech model: Product-of-Filters (PoF) model [4].

The PoF model is motivated by the widely used homomorphic filtering approach to audio and speech signal processing [15] 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:

$$\begin{aligned} Y_{ft} &\sim \text{Poisson}(\textstyle\sum_l X_{f,t-l}R_{fl} + \lambda_f) \\ X_{ft} &\sim \text{Gamma}(\gamma_f, \gamma_f \textstyle\prod_k \exp\{-U_{fk}H_{kt}\}) \\ H_{kt} &\sim \text{Gamma}(\alpha_k, \alpha_k) \end{aligned} \tag{14}$$

where the filters $\mathbf{U} \in \mathbb{R}^{F\times K}$, sparsity level $\boldsymbol{\alpha} \in \mathbb{R}^K_+$, and frequency-dependent noise-level $\boldsymbol{\gamma} \in \mathbb{R}^F_+$ are the PoF parameters learned from clean and dry speech. $\mathbf{H} \in \mathbb{R}^{K\times T}_+$ denotes the weights of linear combination of filters. The inference can be carried out in a similar way as derived in Section 2.2.

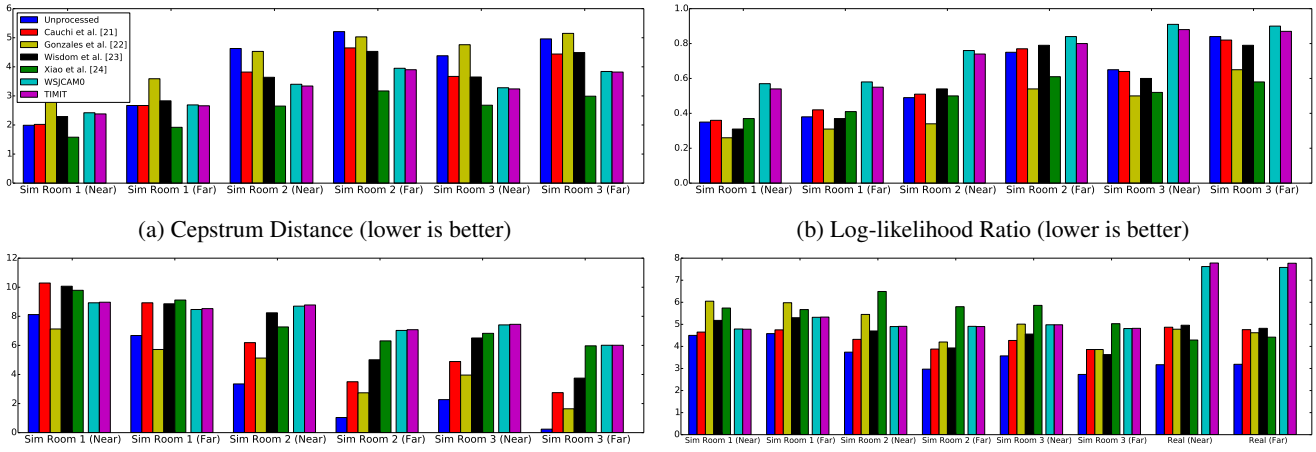
Note that both NMF and PoF assume independence between frames, which is unrealistic. This assumption could be relaxed by imposing temporal structure to the speech model, e.g. with a non-negative hidden Markov model [16] or a recurrent neural network [17].

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 WSJCAM0 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 WSJCAM0 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. In the real recording, 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.



(a) Cepstrum Distance (lower is better)

(b) Log-likelihood Ratio (lower is better)

(c) Frequency-weighted Segmental SNR (higher is better)

(d) SRMR (higher is better)

Fig. 1. The speech enhancement results from the REVERB Challenge evaluation data under different test conditions (Sim: simulated data. Real: real recording). For each condition, seven results are reported: unprocessed reverberant and noisy speech as baseline, four results from REVERB Challenge submissions, obtained from `http://reverb2014.dereverberation.com/result_se.html`, and our proposed approach using speech models learned from WSJCAM0 and TIMIT, respectively.

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 Challange: 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 $a = b = 0.1$, 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 2.3, 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

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

[2] Hagai Attias, John C. Platt, Alex Acero, and Li Deng, "Speech denoising and dereverberation using probabilistic models," *Advances in neural information processing systems*, pp. 758–764, 2001.

[3] Dawen Liang, Daniel P. W. Ellis, Matthew D. Hoffman, and Gautham J. Mysore, "Speech decoloration based on the product-of-filters model," in *Acoustics, Speech and Signal Processing, IEEE International Conference on*, 2014, pp. 2400–2404.

[4] Dawen Liang, Matthew D. Hoffman, and Gautham J. Mysore, "A generative product-of-filters model of audio," in *International Conference on Learning Representations*, 2014.

[5] Kshitiz Kumar, Rita Singh, Bhiksha Raj, and Richard Stern, "Gammatone sub-band magnitude-domain dereverberation for

ASR," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 4604–4607.

[6] Ali Taylan Cemgil, "Bayesian inference for nonnegative matrix factorisation models," *Computational Intelligence and Neuroscience*, vol. 2009, 2009.

[7] Paris Smaragdis and Judith C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on.* IEEE, 2003, pp. 177–180.

[8] Daniel D. Lee and H. Sebastian Seung, "Algorithms for non-negative matrix factorization," *Advances in Neural Information Processing Systems*, vol. 13, pp. 556–562, 2001.

[9] Zhiyao Duan, Gautham J. Mysore, and Paris Smaragdis, "Speech enhancement by online non-negative spectrogram decomposition in non-stationary noise environments.," in *INTERSPEECH*, 2012.

[10] Dennis L. Sun and Gautham J. Mysore, "Universal speech models for speaker independent single channel source separation," in *Acoustics, Speech and Signal Processing, IEEE International Conference on*, 2013, pp. 141–145.

[11] Dhananjay Bansal, Bhiksha Raj, and Paris Smaragdis, "Bandwidth expansion of narrowband speech using non negative matrix factorization," in *9th European Conference on Speech Communication (Eurospeech)*, 2005.

[12] Cédric Févotte, Nancy Bertin, and Jean-Louis Durrieu, "Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis," *Neural computation*, vol. 21, no. 3, pp. 793–830, 2009.

[13] Matthew D. Hoffman, David M. Blei, and Perry R. Cook, "Bayesian nonparametric matrix factorization for recorded music," in *Proceedings of the 27th Annual International Conference on Machine Learning*, 2010, pp. 439–446.

[14] Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul, "An introduction to variational methods for graphical models," *Machine learning*, vol. 37, no. 2, pp. 183–233, 1999.

[15] Alan Oppenheim and Ronald Schafer, "Homomorphic analysis of speech," *Audio and Electroacoustics, IEEE Transactions on*, vol. 16, no. 2, pp. 221–226, 1968.

[16] Gautham J. Mysore, Paris Smaragdis, and Bhiksha Raj, "Non-negative hidden markov modeling of audio with application to source separation," in *Latent Variable Analysis and Signal Separation*, pp. 140–148. Springer, 2010.

[17] Nicolas Boulanger-Lewandowski, Gautham J. Mysore, and Matthew D. Hoffman, "Exploiting long-term temporal dependencies in NMF using recurrent neural networks with application to source separation," in *Acoustics, Speech and Signal Processing, IEEE International Conference on*, 2014, pp. 6969–6973.

[18] Tony Robinson, Jeroen Fransen, David Pye, Jonathan Foote, and Steve Renals, "Wsjcam0: A british english speech corpus for large vocabulary continuous speech recognition," in *In Proc. ICASSP 95*. 1995, pp. 81–84, IEEE.

[19] Mike Lincoln, Iain McCowan, Jithendra Vepa, and Hari Krishna Maganti, "The multi-channel Wall Street Journal audio visual corpus (mc-wsj-av): Specification and initial experiments," in *Automatic Speech Recognition and Understanding, 2005 IEEE Workshop on*. IEEE, 2005, pp. 357–362.

[20] Yi Hu and Philipos C. Loizou, "Evaluation of objective quality measures for speech enhancement," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 1, pp. 229–238, 2008.

[21] Tiago H. Falk, Chenxi Zheng, and Wai-Yip Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 7, pp. 1766–1774, 2010.

[22] Benjamin Cauchi, Ina Kodrasi, Robert Rehr, and Stephan Gerlach, "Joint dereverberation and noise reduction using beamforming and a single-channel speech enhancement scheme," in *the Proceedings of REVERB Challenge*, 2014.

[23] Dayana Ribas González, Serguey Crespo Arias, and Jośe Ramon Calvo de Lara, "Single channel speech enhancement based on zero phase transformation in reverberated environments," in *the Proceedings of REVERB Challenge*, 2014.

[24] Scott Wisdom, Thomas Powers, Les Atlas, and James Pitton, "Enhancement of reverberant and noisy speech by extending its coherence," in *the Proceedings of REVERB Challenge*, 2014.

[25] Xiong Xiao, Shengkui Zhao, Duc Hoang Ha Nguyen, Xionghu Zhong, Douglas L Jones, Eng Siong Chng, and Haizhou Li, "The NTU-ADSC system for Reveration Challenge 2014," in *the Proceedings of REVERB Challenge*, 2014.

[26] Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley, "Stochastic variational inference," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 1303–1347, 2013.