

SPEECH DECOLORATION BASED ON THE PRODUCT-OF-FILTERS MODEL

Dawen Liang*, Daniel P. W. Ellis

LabROSA, Dept. of Electrical Engineering
Columbia University
{dliang, dpwe}@ee.columbia.edu

Matthew D. Hoffman, Gautham J. Mysore

Adobe Research
{mathoffm, gmysore}@adobe.com

ABSTRACT

We present a single-channel speech decoloration method based on a recently proposed generative product-of-filters (PoF) model. We take a spectral approach and attempt to *learn* the magnitude response of the actual coloration filter, given only the degraded speech signal. Experiments on synthetic data demonstrate that the proposed method effectively captures both coarse and fine structure of the coloration filter. On real recordings, we find that simply subtracting the learned coloration filter from the log-spectra yields promising decoloration results.

Index Terms— audio, decoloration, Bayesian modeling, variational inference.

1. INTRODUCTION

Linear distortion effects, such as those caused by recording in reverberant rooms or using non-transparent equipment, are a major cause of speech degradation in practice. For example, although the human auditory system can easily cope with moderately reverberant speech, it causes significant performance diminution for automatic speech recognition (ASR) systems [1].

Various techniques have been proposed in the literature for single-channel speech dereverberation. Reverberation is commonly modeled as the effect of a linear filter, making it susceptible to homomorphic filtering approaches (e.g. [2]). [3] proposes a spectral subtraction based method, which uses a non-stationary reverberation power spectrum estimator. Approaches which estimate the inverse filters to cancel the effect of reverberation have been proposed as well. For example, [4] leveraged harmonicity assumptions (which are particularly applicable for speech) to design a dereverberation filter.

[5] observes that the distortion caused by room reverberation is due to two factors: coloration and long-term reverberation. In this paper we present a new approach to speech decoloration¹. The technique employs the recently

proposed Product-of-Filters (PoF) model [6], a generative model of short-time magnitude spectra.

2. PROPOSED MODEL

2.1. Product-of-filters (PoF) model

We first briefly review the product-of-filters (PoF) model. The motivation for the PoF model comes from the widely used homomorphic filtering approach to speech signal processing [7], where a short window of speech $w[n]$ is modeled as a convolution between an excitation signal $e[n]$ and the impulse response $h[n]$ of a series of linear filters, which becomes a simple addition of their log-spectra in the log-spectral domain.

PoF generalizes the concept of the excitation-filter model: it models a matrix of T magnitude spectra $\mathbf{W} \in \mathbb{R}_+^{F \times T}$ as a product of many filters. PoF assumes that each observed log-spectrum is approximately obtained by linearly combining elements from a pool of L log-scale filters² $\mathbf{U} \equiv [\mathbf{u}_1 | \mathbf{u}_2 | \dots | \mathbf{u}_L] \in \mathbb{R}^{F \times L}$:

$$\log W_{ft} \approx \sum_l U_{fl} a_{lt}, \quad (1)$$

where a_{lt} denotes the activation of filter \mathbf{u}_l in frame t . Sparsity is imposed on the activations to encode the intuition that not all filters are always active.

Formally, PoF is defined:

$$\begin{aligned} a_{lt} &\sim \text{Gamma}(\alpha_l, \alpha_l) \\ W_{ft} &\sim \text{Gamma}(\gamma_f, \gamma_f / \exp(\sum_l U_{fl} a_{lt})) \end{aligned} \quad (2)$$

where γ_f is the frequency-dependent noise level. For $l \in \{1, 2, \dots, L\}$, α_l controls the sparseness of the activations associated with filter \mathbf{u}_l ; smaller values of α_l indicate a prior preference to use \mathbf{u}_l less frequently, since the $\text{Gamma}(\alpha_l, \alpha_l)$ prior places more mass near 0 when α_l is smaller. From a generative point of view, one can view PoF as first drawing activations a_{lt} from a sparse prior, then applying multiplicative gamma noise with expected value 1 to the expected value $\exp(\sum_l U_{fl} a_{lt})$.

*This work was performed in part while Dawen Liang was an intern at Adobe Research, and was supported in part by NSF project IIS-1117015.

¹The term “coloration” can be ambiguous. In this paper, we use it mainly to refer to short-time effects, e.g. reverberation with a short T_{60} .

²When there is no ambiguity, we will simply use “filter” to refer to these log-scale filters for the rest of the paper.

Variational inference [8] is adopted to infer the activation a_{tl} , and the free parameters \mathbf{U} , α , and γ are chosen to approximately maximize the marginal likelihood $p(\mathbf{W}^{\text{train}}|\mathbf{U}, \alpha, \gamma)$ of a set of training spectra. PoF replaces hand-designed decompositions built of basic signal processing operations with a learned decomposition based on statistical inference.

2.2. Decoloration with PoF

Training the PoF model on clean, dry speech will result in a model that assigns high probability to typical speech; that is, the trained model will be better able to explain dry speech than distorted speech. We can leverage this preference for clean speech to infer the characteristics of linear colorations that have been applied to dry speech signals.

Coloration is usually modeled as an effect of a linear filter (e.g., a room impulse response (RIR)) on the signal. The effect of a linear filter factors out as an addition in the log-spectral domain, so we can account for any global linear coloration in the PoF model by adding an extra coloration filter and keeping it on (i.e., setting its activation to 1) for the entire recording. If we hold the pretrained PoF parameters \mathbf{U} , α , and γ constant and tune the new coloration filter to a recording of linearly distorted speech, it is reasonable to suppose that the model will use the new filter to account for this linear distortion, allowing the pretrained PoF model to focus on the phonetic and speaker-level variation in the recording.

Formally, we define the coloration filter $\mathbf{h} \in \mathbb{R}^F$ and modify (2) as follows:

$$\begin{aligned} a_{lt} &\sim \text{Gamma}(\alpha_l, \alpha_l) \\ W_{ft} &\sim \text{Gamma}\left(\gamma_f, \gamma_f / \exp(h_f + \sum_l U_{fl} a_{lt})\right). \end{aligned} \quad (3)$$

Under the model,

$$\begin{aligned} \mathbb{E}[a_{lt}] &= 1 \\ \mathbb{E}[W_{ft}] &= \exp(h_f + \sum_l U_{fl} a_{lt}). \end{aligned} \quad (4)$$

A graphical model representation is shown in Figure 1. One potential problem with this formulation is that this approach will be limited by the length of the analysis window when we perform short-time Fourier transform (STFT). This can be addressed with a convolutive model and will be developed as future work.

We will learn the coloration filter \mathbf{h} given previously unseen degraded speech data using a variational Expectation-Maximization (EM) algorithm, which consists of an ‘‘E-step’’ and an ‘‘M-step.’’

2.2.1. E-step

In the E-step, the goal is to approximate the posterior $p(\mathbf{a}_t|\mathbf{w}_t)$, which is intractable to compute directly, with a variational distribution of the form $q(\mathbf{a}_t) = \prod_l q(a_{lt})$ where $q(a_{lt}) = \text{Gamma}(a_{lt}; \nu_{tl}^a, \rho_{tl}^a)$. We will tune ν_t^a and ρ_t^a to minimize

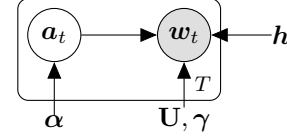


Fig. 1. Graphical model representation of the PoF-based decoloration model. Shaded nodes represent observed variables. Unshaded nodes represent hidden variables and parameters. In this model, \mathbf{U} , α , and γ are trained from clean, dry speech and assumed to be fixed.

the Kullback-Leibler (KL) divergence between the variational distribution q and the posterior p .

Minimizing the KL-divergence is equivalent to maximizing the following variational lower bound:

$$\begin{aligned} &\log p(\mathbf{w}_t|\mathbf{U}, \alpha, \gamma, \mathbf{h}) \\ &\geq \mathbb{E}_q[\log p(\mathbf{w}_t, \mathbf{a}_t|\mathbf{U}, \alpha, \gamma, \mathbf{h})] - \mathbb{E}_q[\log q(\mathbf{a}_t)] \\ &\equiv \mathcal{L}(\nu_t^a, \rho_t^a). \end{aligned} \quad (5)$$

For the first term,

$$\begin{aligned} &\mathbb{E}_q[\log p(\mathbf{w}_t, \mathbf{a}_t|\mathbf{U}, \alpha, \gamma, \mathbf{h})] \\ &= \mathbb{E}_q[\log p(\mathbf{w}_t|\mathbf{a}_t, \mathbf{U}, \gamma, \mathbf{h})] + \mathbb{E}_q[\log p(\mathbf{a}_t|\alpha)] \\ &= \text{const} - \sum_f \gamma_f \left(h_f + \sum_l U_{fl} \mathbb{E}_q[a_{lt}] \right. \\ &\quad \left. + W_{ft} e^{-h_f} \prod_l \mathbb{E}_q[\exp(-U_{fl} a_{lt})] \right) \\ &\quad + \sum_l \left((\alpha_l - 1) \mathbb{E}_q[\log a_{lt}] - \alpha_l \mathbb{E}_q[a_{lt}] \right) \end{aligned}$$

where $\Gamma(\cdot)$ is the gamma function. The necessary expectations $\mathbb{E}_q[a_{lt}] = \nu_{lt}^a / \rho_{lt}^a$ and $\mathbb{E}_q[\log a_{lt}] = \psi(\nu_{lt}^a) - \log \rho_{lt}^a$, where $\psi(\cdot)$ is the digamma function, are both easy to compute. An expression for $\mathbb{E}_q[\exp(-U_{fl} a_{lt})]$ follows from the moment-generating function of a gamma-distributed random variable:

$$\mathbb{E}_q[\exp(-U_{fl} a_{lt})] = \left(1 + \frac{U_{fl}}{\rho_{lt}^a} \right)^{-\nu_{lt}^a} \quad (6)$$

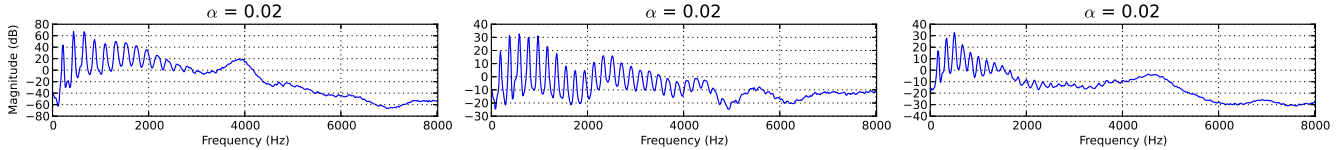
for $U_{fl} > -\rho_{lt}^a$, and $+\infty$ otherwise³.

The second term is the entropy of a gamma-distributed random variable:

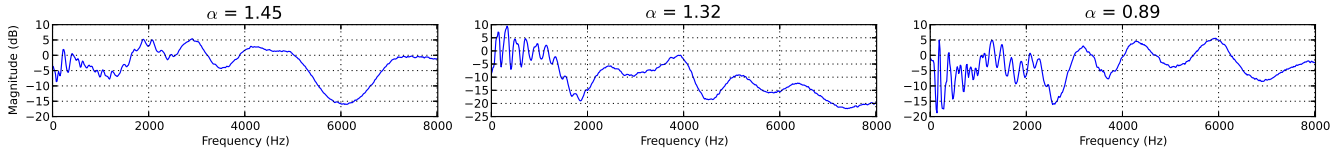
$$\begin{aligned} &-\mathbb{E}_q[\log q(\mathbf{a}_t)] \\ &= \sum_l \left(\nu_{lt}^a - \log \rho_{lt}^a + \log \Gamma(\nu_{lt}^a) + (1 - \nu_{lt}^a) \psi(\nu_{lt}^a) \right). \end{aligned} \quad (7)$$

Closed-form updates for the variational parameters ν_t^a and ρ_t^a are not available. We optimize the variational lower bound via gradient-based numerical optimization (specifically, limited-memory BFGS). Note that $\mathcal{L}(\nu_t^a, \rho_t^a)$ can be decomposed into T independent terms, and the E-step can therefore be done in parallel.

³Technically the expectation for $U_{fl} \leq -\rho_{lt}^a$ is undefined. Here we treat it as $+\infty$ so that when $U_{fl} \leq -\rho_{lt}^a$ the variational lower bound goes to $-\infty$ and the optimization can be carried out seamlessly.



(a) The top 3 filters u_l with the smallest α_l values (shown above each plot).



(b) The top 3 filters u_l with the largest α_l values (shown above each plot).

Fig. 2. The representative filters learned from the PoF model with $L = 30$.

2.2.2. M-step

In the M-step, we find an approximate maximum-likelihood estimate of the coloration filter h using the expected sufficient statistics for a_t obtained from the E-step. This is accomplished by maximizing the variational objective (5) with respect to h . Taking the derivative and setting it to 0, we obtain the closed-form update:

$$h_f^{\text{new}} = \log \left(\frac{1}{T} \sum_t W_{ft} \cdot \prod_l \mathbb{E}_q[\exp(-U_{fl} a_{lt})] \right) \quad (8)$$

Each E-step and M-step increases the objective \mathcal{L} , so iterating between them is guaranteed to find a stationary point. In practice, we iterate until the variational lower bound increases by less than 0.01%, which in our experiments typically takes less than 10 iterations. Once the coloration filter h is learned, decoloration can be done by subtracting h from each log-spectrum w_t .

3. EXPERIMENTS

To evaluate the effectiveness of the proposed method on decoloration, we conducted experiments on both synthetic data and real recordings.

The proposed model requires pretrained PoF parameters \mathbf{U} , α , and γ , which we learned from 20 randomly selected speakers (10 males and 10 females) in the TIMIT Speech Corpus. We performed a 1024-point FFT (64 ms) with a Hann window and 25% overlap. We performed the experiments on magnitude spectrograms, and set the number of filters used in the PoF model to $L = 30$.

To illustrate what the learned filters from PoF look like, the three filters u_l associated with the smallest and largest values of α_l are shown in Figure 2. The filters in Figure 2(a), which are used relatively rarely and therefore have smaller values of α_l , tend to have the strong harmonic structure displayed by the log-spectra of periodic signals, which is con-

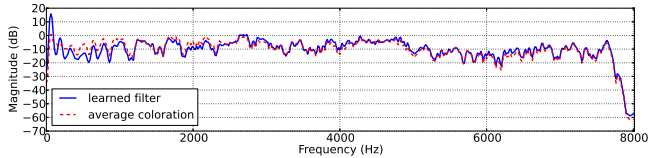
sistent with the fact that normally there is not more than one excitation signal contributing to a speaker’s voice. The filters in Figure 2(b), which are used relatively frequently and therefore have larger values of α_l , tend to vary more smoothly, suggesting that they are being used to model the filtering induced by the vocal tract. This indicates the model has more freedom to use several of the coarser “vocal tract” filters per spectrum, which is consistent with the fact that several aspects of the vocal tract may combine to filter the excitation signal generated by a speaker’s vocal folds.

It is worth noticing that speaker-specific coloration effects may bleed into the coloration filter h learned by the proposed model. We can partially compensate for this by learning an average speaker coloration filter by fitting the model (3) to the clean speech data used to learn the PoF model parameters \mathbf{U} , α , and γ . We subtracted this average speaker coloration filter from the learned filter h in all experiments.

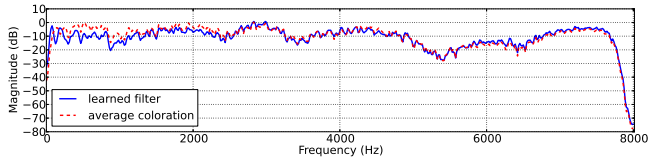
3.1. Synthetic data

We use short-time reverberation as a particular example of coloration. We selected three different room impulse responses (RIR) with various T_{60} from the Aachen impulse response (AIR) database [9]: studio booth ($T_{60} = 80$ ms), meeting room ($T_{60} = 210$ ms), and office ($T_{60} = 370$ ms). We convolved these RIRs with sentences from 6 randomly selected speakers that do not overlap with the speakers used to fit the model parameters \mathbf{U} , α and γ .

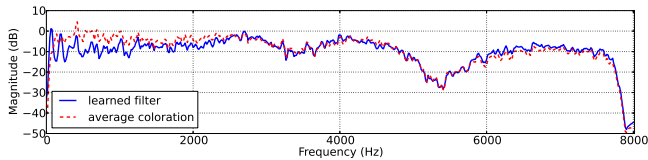
Since the lengths of the RIRs are longer than the analysis window used for the STFT, we cannot directly compare the learned filters with the magnitude responses of the RIRs. However, since we are dealing with short-time reverberation, it is reasonable to assume the difference of the log-spectra between the reverberant speech and the original clean speech is roughly consistent across time frames. By taking the average of the difference, we can obtain an average coloration which closely approximates the effects of the RIRs. One way



(a) Studio booth ($T_{60} = 80$ ms).



(b) Meeting room ($T_{60} = 210$ ms).



(b) Office ($T_{60} = 370$ ms).

Fig. 3. The comparison between learned coloration filter and the average coloration (average difference between the log-spectra of reverberant speech and that of clean speech) under three different room impulse responses with increasing T_{60} . We can see that the proposed model effectively recovers the coloration without the access to the clean speech.

to interpret this average coloration is as the log-magnitude response of the filter that, if subtracted from the observed log-spectra, would minimize the mean squared Euclidean distance between the colored spectra and the clean spectra. Figure 3 shows the comparison between our learned coloration filter and the average coloration under the three room impulse responses. We can see that the proposed model effectively recovers the structure of the average coloration.

3.2. Real recordings

We also evaluated the proposed method on real recordings. To test our method’s ability to correct for coloration from sources other than reverb, we used the *Voice Memos* application from an iPhone 5s to record the same TIMIT sentences used in Section 3.1 played from a Macbook Pro (the distance between the iPhone and the laptop speaker was roughly 30 cm) in a small room (10 feet by 10 feet). To decolor the recordings, we applied a zero-phase filter with log-magnitude response $-h$, effectively dividing out the impact of the coloration estimated by the PoF model. We compared with two alternative estimation methods for h : the average coloration (AC) “oracle” filter from the previous section (which cannot be used in practice), and a simple baseline obtained by computing each recording’s average magnitude spectrogram and dividing by

Table 1. Average scores across sentences on the speech enhancement metrics: cepstrum distance (CD), log likelihood ratio (LLR), frequency-weighted segmental SNR (FWSegSNR) and speech-to-reverberation modulation energy ratio (SRMR). Bold numbers indicate the best scores; the significance is assessed with a paired Wilcoxon signed-rank test at $\alpha = 0.05$ level.

	CD	LLR	FWSegSNR	SRMR
Input	5.69	1.64	5.87	4.87
Baseline	4.27	0.50	6.73	4.39
Proposed	3.61	0.50	9.60	6.19
AC	3.69	0.43	7.94	5.46

the average magnitude spectrum from the same speech data used to fit the PoF model parameters (this baseline uses the same data as our proposed method).

We evaluated the proposed method under the context of speech enhancement. We used the metrics from the Reverb Challenge⁴, which include cepstrum distance (CD), log likelihood ratio (LLR), and frequency-weighted segmental SNR (FWSegSNR) from [10], and speech-to-reverberation modulation energy ratio (SRMR) from [11]. The average scores across sentences are reported in Table 1. For each metric, scores statistically indistinguishable from the best score are indicated in bold; significance is assessed with a paired Wilcoxon signed-rank test at $\alpha = 0.05$ level.

From the results we can see that the proposed model outperforms the baseline by a large margin except on LLR, where all three methods perform equally well. Note that under some metrics, the proposed method even outperforms AC, which has access to the original clean speech. This may be due to the PoF’s ability to infer data at frequencies that are missing from the poorly recorded audio, which was demonstrated in a bandwidth expansion task from [6].

4. CONCLUSION AND FUTURE WORK

We proposed a single-channel speech decoloration method based on a generative product-of-filters (PoF) model. By adding an extra filter, we extend the original PoF model to learn global coloration effects while retaining PoF’s ability to capture speech characteristics. Experimental results on both synthetic data and real recordings demonstrate that the proposed method accurately estimates coloration filters.

A limitation of our approach is that it can only recover short-time coloration effects; moving to a convolutive model would allow us to handle longer reverberation times. Another goal for the future is to speed up the E-step, possibly by relaxing the model so that closed-form updates can be employed rather than gradient-based optimization.

⁴<http://reverb2014.dereverberation.com>

5. REFERENCES

- [1] Brian E. D. Kingsbury, Nelson Morgan, and Steven Greenberg, “Robust speech recognition using the modulation spectrogram,” *Speech communication*, vol. 25, no. 1, pp. 117–132, 1998.
- [2] James L. Caldwell, “Implementation of short-time homomorphic dereverberation,” M.S. thesis, Massachusetts Institute of Technology, 1971.
- [3] Katia Lebart, Jean-Marc Boucher, and P. N. Denbigh, “A new method based on spectral subtraction for speech dereverberation,” *Acta Acustica united with Acustica*, vol. 87, no. 3, pp. 359–366, 2001.
- [4] Tomohiro Nakatani, Keisuke Kinoshita, and Masato Miyoshi, “Harmonicity-based blind dereverberation for single-channel speech signals,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 1, pp. 80–95, 2007.
- [5] Mingyang Wu and DeLiang Wang, “A two-stage algorithm for enhancement of reverberant speech,” in *Proc. International Conference on Acoustics, Speech, and Signal Processing*. Citeseer, 2005, pp. 1085–108.
- [6] Dawen Liang, Matthew D. Hoffman, and Gautham J. Mysore, “A generative product-of-filters model of audio,” *arXiv:1312.5857*, 2013.
- [7] Alan V. Oppenheim and Ronald W. Schafer, “Homomorphic analysis of speech,” *Audio and Electroacoustics, IEEE Transactions on*, vol. 16, no. 2, pp. 221–226, 1968.
- [8] 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.
- [9] Marco Jeub, Magnus Schafer, and Peter Vary, “A binaural room impulse response database for the evaluation of dereverberation algorithms,” in *Digital Signal Processing, 2009 16th International Conference on*. IEEE, 2009, pp. 1–5.
- [10] Yi Hu and Philippos 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.
- [11] 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.