A Generative Product-of-Filters Model of Audio

Dawen Liang*
Columbia University
New York, NY 10027
dliang@ee.columbia.edu

Matthew D. Hoffman
Adobe Research
San Francisco, CA 94103
mathoffm@adobe.com

Gautham J. Mysore
Adobe Research
San Francisco, CA 94103
gmysore@adobe.com

Abstract

We propose the product-of-filters (PoF) model, a generative model that decomposes audio spectra as sparse linear combinations of "filters" in the log-spectral domain. PoF makes similar assumptions to those used in the classic homomorphic filtering approach to signal processing, but replaces decompositions built of basic signal processing operations with a learned decomposition based on statistical inference. This paper formulates the PoF model and derives a mean-field method for posterior inference and a variational EM algorithm to estimate the model's free parameters. We demonstrate PoF's potential for audio processing on a bandwidth expansion task, and show that PoF can serve as an effective unsupervised feature extractor for a speaker identification task.

1 Introduction

Some of the most successful approaches to audio signal processing of the last fifty years have been based on decomposing complicated systems into an excitation signal and some number of simpler linear systems. One of the simplest (and most widely used) examples is linear predictive coding (LPC), which uses a simple autoregressive model to decompose an audio signal into an excitation signal and linear filter [17]. More broadly, homomorphic filtering methods such as cepstral analysis [16] try to decompose complicated linear systems into a set of simpler linear systems that can then be analyzed, interpreted, and manipulated independently.

One reason that this broad approach has been successful is because it is consistent with the way many real-world objects actually generate sound. An important example is the human voice: human vocal sounds are generated by running vibrations generated by the vocal folds through the rest of the vocal tract (tongue, lips, jaw, etc.), which approximately linearly filters the vibrations that come from the larynx or lungs.

Traditional approaches typically rely on hand-designed decompositions built of basic operations such as Fourier transforms, discrete cosine transforms, and least-squares solvers. In this paper we take a more data-driven approach, and derive a generative product-of-filters (PoF) model that *learns* a statistical-inference-based decomposition that is tuned to be appropriate to the data being analyzed. Like traditional homomorphic filtering approaches, PoF decomposes audio spectra as linear combinations of filters in the log-spectral domain. Unlike previous approaches, these filters are learned from data rather than selected from convenient families such as orthogonal cosines, and the PoF model learns a sparsity-inducing prior that prefers decompositions that use relatively few filters to explain each observed spectrum. The result when applied to speech data is that PoF discovers some filters that model excitation signals and some that model the various filtering operations that the vocal tract can perform. Given a set of these learned filters, PoF can infer how much each filter contributed to a given audio magnitude spectrum, resulting in a sparse, compact, interpretable representation.

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

The rest of the paper proceeds as follows. First, we formally introduce the product-of-filters (PoF) model, and give more rigorous intuitions about the assumptions that it makes. Next, we review some previous work and show how it relates to the PoF model. Then, we derive a mean-field variational inference algorithm that allows us to do approximate posterior inference in PoF, as well as a variational EM algorithm that fits the model's free parameters to data. Finally, we demonstrate PoF's potential for audio processing on a bandwidth expansion task, where it achieves better results than a recently proposed technique based on non-negative matrix factorization (NMF). We also evaluate PoF as an unsupervised feature extractor, and find that the representation learned by the model yields higher accuracy on a speaker identification task than the widely used mel-frequency cepstral coefficient (MFCC) representation.

2 Product-of-Filters Model

We are interested in modeling audio spectrograms, which are collections of Fourier magnitude spectra $\mathbf{W}$ taken from some set of audio signals, where $\mathbf{W}$ is an $F \times T$ non-negative matrix; the cell W_{ft} gives the magnitude of the audio signal at frequency bin f and time window (often called a frame) t. Each column of $\mathbf{W}$ is the magnitude of the fast Fourier transform (FFT) of a short window of an audio signal, within which the spectral characteristics of the signal are assumed to be roughly stationary.

The motivation for our model comes from the widely used homomorphic filtering approach to audio and speech signal processing [16], where a short window of audio $w[n]$ is modeled as a convolution between an excitation signal $e[n]$ (which might come from a speaker's vocal folds) and the impulse response $h[n]$ of a series of linear filters (such as might be implemented by a speaker's vocal tract):

$$w[n] = (e * h)[n] \tag{1}$$

In the spectral domain after taking the FFT, this is equivalent to:

$$|\mathcal{W}[k]| = |\mathcal{E}[k]| \circ |\mathcal{H}[k]| = \exp\{\log|\mathcal{E}[k]| + \log|\mathcal{H}[k]|\} \tag{2}$$

where $\circ$ denotes element-wise multiplication and $|\cdot|$ denotes the magnitude of a complex value produced by the FFT. Thus, the convolution between these two signals corresponds to a simple addition of their log-spectra. Another attractive feature is the symmetry between the excitation signal $e[n]$ and the impulse response $h[n]$ of the vocal-tract filter—convolution commutes, so mathematically (if not physiologically) the vocal tract could just as well be exciting the "filter" implemented by vocal folds.

We will likewise model the observed magnitude spectra as a product of filters. We assume each observed log-spectrum is approximately obtained by linearly combining elements from a pool of L log-filters[1] $\mathbf{U} \equiv [\boldsymbol{u}_1|\boldsymbol{u}_2|\cdots|\boldsymbol{u}_L] \in \mathbb{R}^{F\times L}$:

$$\log W_{ft} \approx \textstyle\sum_l U_{fl} a_{lt}, \tag{3}$$

where a_{lt} denotes the activation of filter $\boldsymbol{u}_l$ in frame t. We will impose some sparsity on the activations, to allow us to encode the intuition that not all filters are always active. This assumption expands on the expressive power of the simple excitation-filter model of equation 1; we could recover that model by partitioning the filters into "excitations" and "vocal tracts", requiring that exactly one "excitation filter" be active in each frame, and combining the weighted effects of all "vocal tract filters" into a single filter.

We have two main reasons for relaxing the classic excitation-filter model to include more than two filters, one computational and one statistical. The statistical rationale is that the parameters that define the human voice (pitch, tongue position, etc.) are inherently continuous, and so a very large dictionary of excitations and filters might be necessary to explain observed inter- and intra-speaker variability with the classic model. The computational rationale is that clustering models (which might try to determine which excitation is active) can be more fraught with local optima than factorial models such as ours (which tries to determine how active each filter is), and there is some precedent for relaxing clustering models into factorial models [3].

[1] We will use the term "filter" when referring to $\mathbf{U}$ for the rest of the paper.

Formally, we define the product-of-filters model:

$$\begin{aligned} a_{lt} &\sim \text{Gamma}(\alpha_l, \alpha_l) \\ W_{ft} &\sim \text{Gamma}\Big(\gamma_f, \gamma_f / \exp(\textstyle\sum_l U_{fl} a_{lt})\Big) \end{aligned} \tag{4}$$

where γ_f is the frequency-dependent noise level. We restrict the activations $\boldsymbol{a}_t$ (but not the filters $\boldsymbol{u}_l$) to be non-negative; if we allowed negative a_{lt}, then the the filters would be inverted, reducing the interpretability of the model.

Under this model

$$\begin{aligned} \mathbb{E}[a_{lt}] &= 1 \\ \mathbb{E}[W_{ft}] &= \exp(\textstyle\sum_l U_{fl} a_{lt}). \end{aligned} \tag{5}$$

For $l \in \{1, 2, \cdots, L\}$, α_l controls the sparseness of the activations associated with filter $\boldsymbol{u}_l$; smaller values of α_l indicate that filter $\boldsymbol{u}_l$ is used more rarely. From a generative point of view, one can view the model as first drawing activations a_{lt} from a sparse prior, then applying multiplicative gamma noise with expected value 1 to the expected value $\exp(\sum_l U_{fl} a_{lt})$. A graphical model representation of the PoF model is shown in Figure 1.

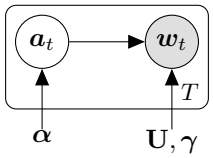


Figure 1: Graphical model representation of the PoF model.

In this paper we focus on speech applications, but the homomorphic filtering approach has also been successfully applied to model other kinds of sounds such as musical instruments. For example, [13] treat the effect of the random excitation, string, and body as a chain of linear systems, which can therefore be modeled as a product of filters.

3 Related Work

The PoF model can be interpreted as a matrix factorization model, where we are trying to decompose the log-spectrogram. A closely related model is non-negative matrix factorization (NMF) [14] and its variations, e.g., NMF with sparseness constrains [10], convex NMF [4], and fully Bayesian NMF [2]. In NMF, a $F \times T$ non-negative matrix $\mathbf{W}$ is approximately decomposed into the product of two non-negative matrices: a $F \times K$ matrix $\mathbf{V}$ (often called the dictionary) and a $K \times T$ matrix $\mathbf{H}$ (often called the activations). NMF is widely used to analyze audio spectrograms [6, 19], largely due to its additivity property and the parts-based representation that it induces. It also often provides a semantically meaningful interpretation. For example, given the NMF decomposition of a piano sonata, the components in the dictionary are likely to correspond to notes with different pitches, and the activations will indicate when and how strongly each note is played. NMF's ability to isolate energy coming from different sources in mixed recordings has made it a popular tool for addressing source separation problems.

Although both models decompose audio spectra using a linear combination of dictionary elements, NMF and PoF make fundamentally different modeling assumptions. NMF models each frame of a spectrogram as an additive combination of dictionary elements, which approximately corresponds to modeling the corresponding time-domain signal as a summation of parts. On the other hand, PoF models each frame of the spectrogram as a product of filters (sum of log-filters), which corresponds to modeling the corresponding time-domain signal as a convolution of filters. NMF is well suited to decomposing *polyphonic* sounds into mixtures of independent sources, whereas PoF is well suited to decomposing *monophonic* sounds into simpler systems.

In the compressive sensing literature, there has been a great deal of work on matrix factorization and dictionary learning by solving an optimization problem with sparseness constraints—adding ℓ^1 norm penalty as a convex relaxation of ℓ^0 norm penalty [5]. Online algorithms have also been

proposed to handle large data sets [15]. In principle, we could have formulated the PoF model similarly, using an ℓ^1 penalty and convex optimization in place of a gamma prior and Bayesian inference. However, in such a formulation it is unclear how we might fit the L hyperparameters $\boldsymbol{\alpha}$ controlling the amount of sparsity in the activations associated with each filter. We found that PoF best explains speech training data when each filter $\boldsymbol{u}_l$ has its own sparsity hyperparameter α_l, and performing cross-validation to select so many hyperparameters would be impractical.

4 Inference and Parameter Estimation

From Figure 1, we can see that there are two computational problems that will arise when using the PoF model. First, given fixed $\mathbf{U}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\gamma}$ and input spectrum $\boldsymbol{w}_t$, we must (approximately) compute the posterior distribution $p(\boldsymbol{a}_t|\boldsymbol{w}_t, \mathbf{U}, \boldsymbol{\alpha}, \boldsymbol{\gamma})$. This will enable us to fit the PoF model to unseen data and obtain a different representation in the latent filter space. Second, given a collection of training spectra $\mathbf{W} = \{\boldsymbol{w}_t\}^{1:T}$, we want to find the maximum likelihood estimates of the free parameters $\mathbf{U}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\gamma}$. In this section, we will tackle these two problems respectively. The detailed derivations can be found in the appendix.

4.1 Mean-Field Posterior Inference

The posterior $p(\boldsymbol{a}_t|\boldsymbol{w}_t, \mathbf{U}, \boldsymbol{\alpha}, \boldsymbol{\gamma})$ is intractable to compute due to the nonconjugacy of the model. Therefore, we employ mean-field variational inference [12].

Variational inference is a deterministic alternative to Markov Chain Monte Carlo (MCMC) methods. The basic idea behind variational inference is to choose a tractable family of variational distributions $q(\boldsymbol{a}_t)$ to approximate the intractable posterior $p(\boldsymbol{a}_t|\boldsymbol{w}_t, \mathbf{U}, \boldsymbol{\alpha}, \boldsymbol{\gamma})$, so that the Kullback-Leibler (KL) divergence between the variational distribution and the true posterior $\mathrm{KL}(q_a \| p_{a|\mathbf{W}})$ is minimized. In particular, we are using the mean-field family which is completely factorized, i.e., $q(\boldsymbol{a}_t) = \prod_l q(a_{lt})$. For each a_{lt}, we choose a variational distribution from the same family as a_{lt}'s prior distribution: $q(a_{lt}) = \mathrm{Gamma}(a_{lt}; \nu^a_{lt}, \rho^a_{lt})$. $\boldsymbol{\nu}^a_t$ and $\boldsymbol{\rho}^a_t$ are free parameters that we will tune to minimize the KL divergence between q and the posterior.

We can lower bound the marginal likelihood of the input spectrum $\boldsymbol{w}_t$:

$$\begin{aligned} &\log p(\boldsymbol{w}_t|\mathbf{U}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \\ &\geq \mathbb{E}_q[\log p(\boldsymbol{w}_t, \boldsymbol{a}_t|\mathbf{U}, \boldsymbol{\alpha}, \boldsymbol{\gamma})] - \mathbb{E}_q[\log q(\boldsymbol{a}_t)] \triangleq \mathcal{L}(\boldsymbol{\nu}^a_t, \boldsymbol{\rho}^a_t) \end{aligned} \tag{6}$$

To compute the variational lower bound $\mathcal{L}(\boldsymbol{\nu}^a_t, \boldsymbol{\rho}^a_t)$, the necessary expectations $\mathbb{E}_q[a_{lt}] = \nu^a_{lt}/\rho^a_{lt}$ and $\mathbb{E}_q[\log a_{lt}] = \psi(\nu^a_{lt}) - \log \rho^a_{lt}$, where $\psi(\cdot)$ is the digamma function, are both easy to compute. For $\mathbb{E}_q[\exp(-U_{fl}a_{lt})]$, we seek for the moment-generating function of a gamma-distributed random variable and obtain the expectation as:

$$\mathbb{E}_q[\exp(-U_{fl}a_{lt})] = \left(1 + \frac{U_{fl}}{\rho^a_{lt}}\right)^{-\nu^a_{lt}} \tag{7}$$

for $U_{fl} > -\rho^a_{lt}$, and $+\infty$ otherwise[2].

The nonconjugacy and the exponents in the likelihood model preclude optimizing the lower bound by traditional closed-form coordinate ascent updates. Instead, we compute the gradient of $\mathcal{L}(\boldsymbol{\nu}^a_t, \boldsymbol{\rho}^a_t)$ with respect to variational parameters $\boldsymbol{\nu}^a_t$ and $\boldsymbol{\rho}^a_t$ and use Limited-memory BFGS (L-BFGS) to optimize the variational lower bound, which guarantees to find a local optimum and optimal variational parameters $\{\hat{\boldsymbol{\nu}}^a_t, \hat{\boldsymbol{\rho}}^a_t\}$.

Note that in the posterior inference, the optimization problem is independent for different frame t. Therefore, given input spectra $\{\boldsymbol{w}_t\}^{1:T}$, we can break down the whole problem into T independent sub-problems which can be solved in parallel.

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

4.2 Parameter Estimation

Given a collection of training audio spectra $\{\boldsymbol{w}_t\}^{1:T}$, we carry out parameter estimation for the PoF model by finding the maximum likelihood estimates of the free parameters $\mathbf{U}, \boldsymbol{\alpha}$, and $\boldsymbol{\gamma}$, approximately marginalizing out $\boldsymbol{a}_t$.

We formally define the parameter estimation problem as

$$\begin{aligned}\hat{\mathbf{U}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}} &= \underset{\mathbf{U},\boldsymbol{\alpha},\boldsymbol{\gamma}}{\arg\max} \sum_t \log p(\boldsymbol{w}_t|\mathbf{U}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \\ &= \underset{\mathbf{U},\boldsymbol{\alpha},\boldsymbol{\gamma}}{\arg\max} \sum_t \log \int_{\boldsymbol{a}_t} p(\boldsymbol{w}_t, \boldsymbol{a}_t|\mathbf{U}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \mathrm{d}\boldsymbol{a}_t\end{aligned} \tag{8}$$

This problem can be solved by variational Expectation-Maximization (EM) which maximizes a lower bound on marginal likelihood in equation 6 with respect to the variational parameters, and then for the fixed values of variational parameters, maximizes the lower bound with respect to the model's free parameters $\mathbf{U}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\gamma}$.

E-step For each $\boldsymbol{w}_t$ where $t = 1, 2, \cdots, T$, perform posterior inference by optimizing the values of the variational parameters $\{\hat{\boldsymbol{\nu}}_t^a, \hat{\boldsymbol{\rho}}_t^a\}$. This is done as described in Section 4.1.

M-step Maximize the variational lower bound in equation 6, which is equivalent to maximizing the following objective:

$$\mathcal{Q}(\mathbf{U}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) = \textstyle\sum_t \mathbb{E}_q[\log p(\boldsymbol{w}_t, \boldsymbol{a}_t|\mathbf{U}, \boldsymbol{\alpha}, \boldsymbol{\gamma})] \tag{9}$$

This is accomplished by finding the maximum likelihood estimates using the expected sufficient statistics for each $\boldsymbol{a}_t$ that were computed in the E-step. There are no simple closed-form updates for the M-step. Therefore, we compute the gradient of $\mathcal{Q}(\mathbf{U}, \boldsymbol{\alpha}, \boldsymbol{\gamma})$ with respect to $\mathbf{U}, \boldsymbol{\alpha}, \boldsymbol{\gamma}$, respectively, and use L-BFGS to optimize the bound in equation 9.

The most time-consuming part for M-step is updating $\mathbf{U}$, which is a $F \times L$ matrix. However, note that the optimization problem is independent for different frequency bins $f \in \{1, 2, \cdots, F\}$. Therefore, we can update $\mathbf{U}$ by optimizing each row independently, and in parallel if desired.

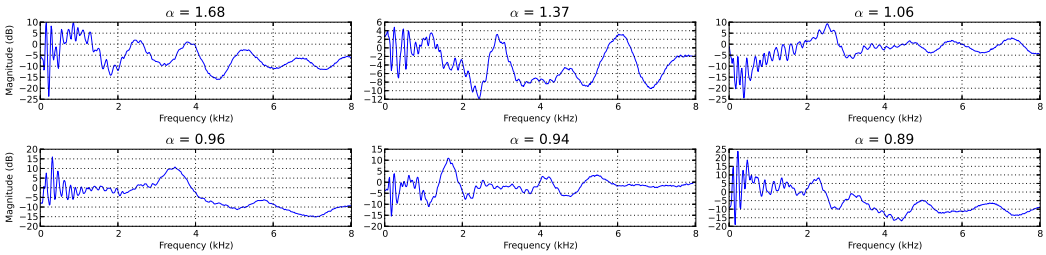
5 Evaluation

We conducted experiments to assess the PoF model on two different tasks. We evaluate PoF's ability to infer missing data in the bandwidth expansion task. We also explore the potential of the PoF model as an unsupervised feature extractor for the speaker identification task.

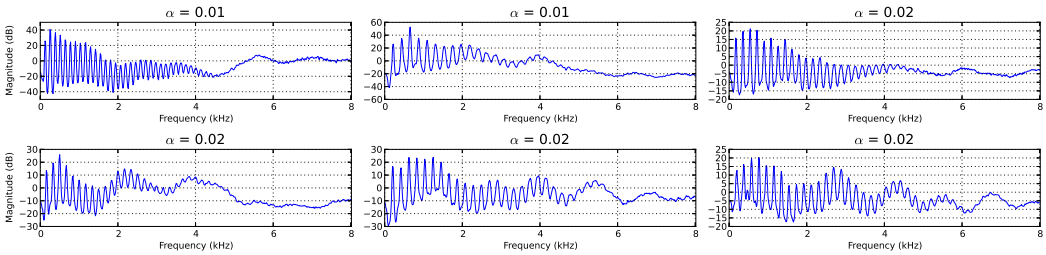
Both tasks require pre-trained parameters $\mathbf{U}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\gamma}$, which we learned from the TIMIT Speech Corpus [7]. It contains speech sampled at 16000 Hz from 630 speakers of eight major dialects of American English, each reading ten phonetically rich sentences. The parameters were learned from 20 randomly selected speakers (10 males and 10 females). We performed a 1024-point FFT with Hann window and 50% overlap, thus the number of frequency bins was $F = 513$. We performed the experiments on magnitude spectrograms except where specified otherwise.

We tried different model orders $L \in \{10, 20, \cdots, 80\}$ and evaluated the lower bound on the marginal likelihood $\log p(\boldsymbol{w}_t|\mathbf{U}, \boldsymbol{\alpha}, \boldsymbol{\gamma})$ in equation 6. In general, larger L will give us a larger variational lower bound and will be slower to train. In our experiments, we set $L = 50$ as a compromise between performance and computational efficiency. We initialized all the variational parameters $\boldsymbol{\nu}_t^a$ and $\boldsymbol{\rho}_t^a$ with random draws from a gamma distribution with shape parameter 100 and inverse-scale parameter 100. This yields a diffuse and smooth initial variational posterior, which helped avoid bad local optima. We ran variational EM until the variational lower bound increased by less than 0.01%.

Figure 2 demonstrates some representative filters learned from the PoF model with $L = 50$. The six filters $\boldsymbol{u}_l$ associated with the largest values of α_l are shown in Figure 2a, and the six filters associated with the smallest values of α_l are shown in Figure 2b. Small values of α_l indicate a prior preference to use the associated filters less frequently, since the $\mathrm{Gamma}(\alpha_l, \alpha_l)$ prior places more mass near 0 when α_l is smaller. The filters in Figure 2b, which are used relatively rarely, tend to have the



(a) The top 6 filters $\boldsymbol{u}_l$ with the largest α_l values (shown above each plot).



(b) The top 6 filters $\boldsymbol{u}_l$ with the smallest α_l values (shown above each plot).

Figure 2: The representative filters learned from the PoF model with $L = 50$.

strong harmonic structure displayed by the log-spectra of periodic signals, while the filters in Figure 2a tend to vary more smoothly, suggesting that they are being used to model the filtering induced by the vocal tract. The periodic "excitation" filters tend to be used more rarely, which is consistent with the fact that normally there is not more than one excitation signal contributing to a speaker's voice. (Very few people can speak or sing more than one pitch simultaneously.) 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 be combined to filter the excitation signal generated by a speaker's vocal folds.

Despite the non-convexity inherent to all dictionary-learning problems, which causes the details of the filters vary from run to run, training with multiple random restarts and different speakers produced little impact on the filters that the PoF model learned; in all cases with different model order L, we found the same "excitation/filter" structure, similar to what is shown in Figure 2.

5.1 Bandwidth Expansion

We demonstrate PoF's potential in audio processing applications on a bandwidth expansion task, where the goal is to infer the contents of a full-bandwidth signal given only the contents of a band-limited version of that signal. Bandwidth expansion has applications to restoration of low-quality audio such as might be recorded from a telephone or cheap microphone.

Given the parameters $\mathbf{U}, \boldsymbol{\alpha}$, and $\boldsymbol{\gamma}$ fit to full-bandwidth training data, we can treat the bandwidth expansion problem as a missing data problem. Given spectra from a band-limited recording $\mathbf{W}^{\text{bl}} = \{\boldsymbol{w}_t^{\text{bl}}\}^{1:T}$, the model implies a posterior distribution $p(\boldsymbol{a}_t|\boldsymbol{w}_t^{\text{bl}})$ over the activations $\boldsymbol{a}_t$ associated with the band-limited signal, for $t = \{1, 2, \cdots, T\}$. We can approximate this posterior using the variational inference algorithm from Section 4.1 by only using the band-limited part of $\mathbf{U}$ and $\boldsymbol{\gamma}$. Then we can reconstruct the full-bandwidth spectra by combining the inferred $\{\boldsymbol{a}_t\}^{1:T}$ with the full-bandwidth $\mathbf{U}$. Following the model formulation in equation 4, we might either estimate the full-bandwidth spectra using

$$\mathbb{E}_q[W_{ft}^{\text{fb}}] = \textstyle\prod_l \mathbb{E}_q[\exp(U_{fl} a_{lt})] \tag{10}$$

or

$$\mathbb{E}_q[W_{ft}^{\text{fb}}] = \exp\{\textstyle\sum_l U_{fl} \cdot \mathbb{E}_q[a_{lt}]\}. \tag{11}$$

We use equation 11, both because it is more stable and because human auditory perception is logarithmic; if we are summarizing the posterior distribution with a point estimate, the expectation on the log-spectral domain is more perceptually natural.

As a comparison, NMF is widely used for bandwidth expansion [1, 18, 20]. The full-bandwidth training spectra $\mathbf{W}^{\text{train}}$, which are also used to learn the parameters $\mathbf{U}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\gamma}$ for the PoF model, are decomposed by NMF as $\mathbf{W}^{\text{train}} \approx \mathbf{VH}$, where $\mathbf{V}$ is the dictionary and $\mathbf{H}$ is the activation. Then given the band-limited spectra $\mathbf{W}^{\text{bl}}$, we can use the band-limited part of $\mathbf{V}$ to infer the activation $\mathbf{H}^{\text{bl}}$. Finally, we can reconstruct the full-bandwidth spectra by computing $\mathbf{VH}^{\text{bl}}$.

Based on how the loss function is defined, there can be different types of NMF models: KL-NMF [14] which is based on Kullback-Leibler divergence, and IS-NMF [6] which is based on Itakura-Saito divergence, are among the most commonly used NMF decomposition models in audio signal processing. We compare the PoF model with both KL-NMF and IS-NMF with different model orders K = 25, 50, and 100. We used the standard multiplicative updates for NMF and stopped the iterations when the decrease in the cost function was less than 0.01%. For IS-NMF, we used power spectra instead of magnitude spectra, since the power spectrum representation is more consistent with the statistical assumptions that underlie the Itakura-Saito divergence.

We randomly selected 10 speakers (5 male and 5 female) from TIMIT that do not overlap with the speakers used to fit the model parameters $\mathbf{U}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\gamma}$, and took 3 sentences from each speaker as test data. We cut off all the contents below 400 Hz and above 3400 Hz to obtain band-limited recordings of approximately telephone-quality speech.

In previous NMF-based bandwidth expansion work [1, 18, 20], all experiments are done in a speaker-dependent setting, which means the model is trained from the target speaker. What we are doing here, on the other hand, is speaker-independent: we use no prior knowledge about the specific speaker whose speech is being restored[3]. To our knowledge, little if any work has been done on speaker-independent bandwidth expansion based on NMF decompositions.

To evaluate the quality of the reconstructed recordings, we used the composite objective measure [11] and short-time objective intelligibility [21] metrics. These metrics measure different aspects of the "distance" between the reconstructed speech and the original speech. The composite objective measure (will be abbreviated as OVRL, as it reflects the overall sound quality) was originally proposed as a quality measure for speech enhancement. It aggregates different basic objective measures and has been shown to correlate with humans' perceptions of audio quality. OVRL is based on the predicted perceptual auditory rating and is in the range of 1 to 5 (1: bad; 2: poor; 3: fair; 4: good; 5: excellent). The short-time objective intelligibility measure (STOI) is a function of the clean speech and reconstructed speech, which correlates with the intelligibility of the reconstructed speech, that is, it predicts the ability of listeners to understand what words are being spoken rather than perceived sound quality. STOI is computed as the average correlation coefficient from 15 one-third octave bands across frames, thus theoretically should be in the range of -1 to 1, where larger values indicate higher expected intelligibility. However, in practice, even when we filled out the missing contents with random noise, the STOI is 0.306 $\pm$ 0.016, which can be interpreted as a practical lower bound on the test data.

The average OVRL and STOI with two standard errors[4] across 30 sentences for different methods, along with these from the band-limited input speech as baseline, are reported in Table 1. We can see that NMF improves STOI by a small amount, and PoF improves it slightly more, but the improvement in both cases is fairly small. This may be because the band-limited input speech already has a relatively high STOI (telephone-quality speech is fairly intelligible). On the other hand, PoF produces better predicted perceived sound quality as measured by OVRL than KL-NMF and IS-NMF by a large margin regardless of the model order K, improving the sound quality from poor-to-fair (2.71 OVRL) to fair-to-good (3.25 OVRL).

[3]When we conducted speaker-dependent experiments, both PoF and NMF produced nearly ceiling-level results. Thus we only report results on the harder and more practically relevant speaker-independent problem.

[4]For both OVRL and STOI, we used the MATLAB implementation from the original authors.

Table 1: Average OVRL (composite objective measure) and STOI (short-time objective intelligibility) score with two standard errors (in parenthesis) for the bandwidth expansion task from different methods. OVRL is in the range of 1 to 5 [1: bad; 2: poor; 3: fair; 4: good; 5: excellent]. STOI is the average correlation coefficient, thus theoretically should be in the range of -1 to 1, where larger values indicate higher expected intelligibility.

		OVRL	STOI
Band-limited input		1.72 (0.16)	0.762 (0.012)
KL-NMF	K=25	2.60 (0.12)	0.786 (0.013)
	K=50	2.71 (0.14)	0.790 (0.013)
	K=100	2.41 (0.10)	0.759 (0.012)
IS-NMF	K=25	2.43 (0.15)	0.779 (0.013)
	K=50	2.62 (0.12)	0.774 (0.014)
	K=100	2.15 (0.10)	0.751 (0.012)
PoF		**3.25 (0.13)**	**0.804 (0.010)**

5.2 Feature Learning and Speaker Identification

We explore PoF's potential as an unsupervised feature extractor. One way to interpret the PoF model is that it attempts to represent the data in a latent filter space. Therefore, given spectra $\{\boldsymbol{w}_t\}^{1:T}$, we can use the coordinates in the latent filter space $\{\boldsymbol{a}_t\}^{1:T}$ as features. Since we believe the inter- and intra-speaker variability is well-captured by the PoF model, we use speaker identification to evaluate the effectiveness of these features.

We compare our learned representation with mel-frequency cepstral coefficients (MFCCs), which are widely used in various speech and audio processing tasks including speaker identification. MFCCs are computed by taking the discrete cosine transform (DCT) on mel-scale log-spectra and using only the low-order coefficients. PoF can be understood in similar terms; we are also trying to explain the variability in log-spectra in terms of a linear combination of dictionary elements. However, instead of using the fixed, orthogonal DCT basis, PoF learns a filter space that is tuned to the statistics of the input. Therefore, it seems reasonable to hope that the coefficients $\boldsymbol{a}_t$ from the PoF model, which will be abbreviated as PoFC, might compare favorably with MFCCs as a feature representation.

We evaluated speaker identification under the following scenario: identify different speakers from recordings where each speaker may start and finish talking at random time, but at any given time there is only one speaker speaking (like during a meeting). This is very similar to speaker diarization [8], but here we assume we know *a priori* the number of speakers and certain amount of training data is available for each speaker. Our goal in this experiment was to demonstrate that the PoFC representation captures information that is missing or difficult to extract from the MFCC representation, rather than trying to build a state-of-the-art speaker identification system.

We randomly selected 10 speakers (5 males and 5 females) from TIMIT outside the training data we used to learn the free parameters $\mathbf{U}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\gamma}$. We used the first 13 DCT coefficients which is a standard choice for computing MFCC. We obtained the PoFC by doing posterior inference as described in Section 4.1 and used $\mathbb{E}_q[\boldsymbol{a}_t]$ as a point estimate summary. For both MFCC and PoFC, we computed the first-order and second-order differences and concatenated them with the original feature.

We treat speaker identification problem as a classification problem and make predictions for each frame. We trained a multi-class (one-vs-the-rest) linear SVM using eight sentences from each speaker and tested with the remaining two sentences, which gave us about 7800 frames of training data and 1700 frames of test data. The test data was randomly permuted so that the order in which sentences appear is random, which simulates the aforementioned scenario.

The frame level accuracy is reported in the first row of Table 2. We can see PoFC increases the accuracy by a large margin (from 49.1% to 60.5%). To make use of temporal information, we used a simple median filter smoother with length 25, which boosts the performance for both representations equally; these results are reported in the second row of Table 2.

Table 2: 10-speaker identification accuracy using PoFC, MFCC, and combined. The first row shows the raw frame-level test accuracy. The second row shows the result after applying a simple median filter with length 25 on the frame-level prediction.

	MFCC	PoFC	MFCC + PoFC
Frame-level	49.1%	60.5%	65.5%
Smoothing	74.2%	85.0%	89.5%

Although MFCCs and PoFCs capture similar information, concatenating both sets of features yields better accuracy than that obtained by either feature set alone. The results achieved by combining the features are summarized in the last column of Table 2, which indicates that MFCCs and PoFCs capture complementary information. These results, which use a relatively simple frame-level classifier, suggest that PoFC could produce even better accuracy when used in a more sophisticated model (e.g. a deep neural network).

6 Discussion and Future Work

In this paper, we proposed the product-of-filters (PoF) model, a generative model which makes similar assumptions to those used in the classic homomorphic filtering approach to signal processing. We derived variational inference and parameter estimation algorithms, and demonstrated experimental improvements on a bandwidth expansion task and showed that PoF can serve as an effective unsupervised feature extractor for speaker identification.

In this paper, we derived PoF as a standalone model. However, it can also be used as a building block and integrated into a larger model, e.g., as a prior for the dictionary in a probabilistic NMF model.

Although the optimization in the variational EM algorithm can be parallelized, currently we cannot fit PoF to large-scale speech data on a single machine. Leveraging recent developments in stochastic variational inference [9], it would be possible to learn the free parameters from a much larger, more diverse speech corpus, or even from streams of data.

References

[1] D. Bansal, B. Raj, and P. Smaragdis. Bandwidth expansion of narrowband speech using non-negative matrix factorization. In *INTERSPEECH*, pages 1505–1508, 2005.

[2] A. T. Cemgil. Bayesian inference for nonnegative matrix factorisation models. *Computational Intelligence and Neuroscience*, 2009.

[3] C. Ding and X. He. K-means clustering via principal component analysis. In *Proceedings of the International Conference on Machine learning*, page 29. ACM, 2004.

[4] C. Ding, T. Li, and M. I. Jordan. Convex and semi-nonnegative matrix factorizations. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(1):45–55, 2010.

[5] D. L. Donoho and M. Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization. *Proceedings of the National Academy of Sciences*, 100(5):2197–2202, 2003.

[6] C. Févotte, N. Bertin, and J.-L. Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence: with application to music analysis. *Neural Computation*, 21(3):793–830, Mar. 2009.

[7] W. M. Fisher, G. R. Doddington, and K. M. Goudie-Marshall. The DARPA speech recognition research database: specifications and status. In *Proc. DARPA Workshop on speech recognition*, pages 93–99, 1986.

[8] E.B. Fox, E.B. Sudderth, M.I. Jordan, and A.S. Willsky. A Sticky HDP-HMM with Application to Speaker Diarization. *Annals of Applied Statistics*, 5(2A):1020–1056, 2011.

[9] M. Hoffman, D. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14:1303–1347, 2013.

[10] P. O. Hoyer. Non-negative matrix factorization with sparseness constraints. *The Journal of Machine Learning Research*, 5:1457–1469, 2004.

[11] Y. Hu and P. C. Loizou. Evaluation of objective quality measures for speech enhancement. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(1):229–238, 2008.

[12] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

[13] M. Karjalainen, V. Välimäki, and Z. Jánosy. Towards high-quality sound synthesis of the guitar and string instruments. In *Proceedings of the International Computer Music Conference*, pages 56–56, 1993.

[14] D.D. Lee and H.S. Seung. Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems*, 13:556–562, 2001.

[15] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *The Journal of Machine Learning Research*, 11:19–60, 2010.

[16] A. Oppenheim and R. Schafer. Homomorphic analysis of speech. *Audio and Electroacoustics, IEEE Transactions on*, 16(2):221–226, 1968.

[17] D. O'Shaughnessy. Linear predictive coding. *Potentials, IEEE*, 7(1):29–32, 1988.

[18] B. Raj, R. Singh, M. Shashanka, and P. Smaragdis. Bandwidth expansionwith a Pólya urn model. In *Acoustics, Speech and Signal Processing, IEEE International Conference on*, volume 4, pages IV–597. IEEE, 2007.

[19] P. Smaragdis and J. C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *Applications of Signal Processing to Audio and Acoustics, IEEE Workshop on.*, pages 177–180. IEEE, 2003.

[20] D. L. Sun and R. Mazumder. Non-negative matrix completion for bandwidth extension: A convex optimization approach. In *Machine Learning for Signal Processing (MLSP), IEEE International Workshop on*, pages 1–6. IEEE, 2013.

[21] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(7):2125–2136, 2011.

A Variational EM for Product-of-Filters Model

A.1 E-step (Posterior Inference)

Following Section 4.1, the variational lower bound for the E-step (equation 6):

$$\begin{aligned}
&\log p(\boldsymbol{w}_t|\mathbf{U},\boldsymbol{\alpha},\boldsymbol{\gamma}) \\
=&\log \int_{\boldsymbol{a}_t} q(\boldsymbol{a}_t)\frac{p(\boldsymbol{w}_t,\boldsymbol{a}_t|\mathbf{U},\boldsymbol{\alpha},\boldsymbol{\gamma})}{q(\boldsymbol{a}_t)}\mathrm{d}\boldsymbol{a}_t \\
\geq&\int_{\boldsymbol{a}_t} q(\boldsymbol{a}_t)\log\frac{p(\boldsymbol{w}_t,\boldsymbol{a}_t|\mathbf{U},\boldsymbol{\alpha},\boldsymbol{\gamma})}{q(\boldsymbol{a}_t)}\mathrm{d}\boldsymbol{a}_t \\
=&\mathbb{E}_q[\log p(\boldsymbol{w}_t,\boldsymbol{a}_t|\mathbf{U},\boldsymbol{\alpha},\boldsymbol{\gamma})]-\mathbb{E}_q[\log q(\boldsymbol{a}_t)] \\
\equiv&\mathcal{L}(\boldsymbol{\nu}_t^a,\boldsymbol{\rho}_t^a)
\end{aligned} \tag{12}$$

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

$$\begin{aligned}
&-\mathbb{E}_q[\log q(\boldsymbol{a}_t)] \\
=&\sum_l\Big(\nu_{lt}^a-\log\rho_{lt}^a+\log\Gamma(\nu_{lt}^a)+(1-\nu_{lt}^a)\psi(\nu_{lt}^a)\Big)
\end{aligned}$$

For the first term, we can keep the parts which depend on $\{\boldsymbol{\nu}_t^a,\boldsymbol{\rho}_t^a\}$:

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

Take the derivative of $\mathcal{L}(\boldsymbol{\nu}_t^a,\boldsymbol{\rho}_t^a)$ with respect to ν_{lt}^a and ρ_{lt}^a:

$$\frac{\partial\mathcal{L}}{\partial\nu_{lt}^a}=\sum_f\gamma_f\Big\{W_{ft}\log(1+\frac{U_{fl}}{\rho_{lt}^a})\prod_{j=1}^L\mathbb{E}_q[\exp(-U_{fj}a_{jt})]-\frac{U_{fl}}{\rho_{lt}^a}\Big\}+(\alpha_l-\nu_{lt}^a)\psi_1(\nu_{lt}^a)+1-\frac{\alpha_l}{\rho_{lt}^a}$$

$$\frac{\partial\mathcal{L}}{\partial\rho_{lt}^a}=\frac{\nu_{lt}}{(\rho_{lt}^a)^2}\sum_f\gamma_f\Big\{U_{fl}-W_{ft}(1+\frac{U_{fl}}{\rho_{lt}^a})^{-1}U_{fl}\prod_{j=1}^L\mathbb{E}_q[\exp(-U_{fj}a_{jt})]\Big\}+\alpha_l(\frac{\nu_{lt}}{(\rho_{lt}^a)^2}-\frac{1}{\rho_{lt}^a})$$

where $\psi_1(\cdot)$ is the trigamma function.

A.2 M-step

The objective function for M-step is:

$$\begin{aligned}
\mathcal{Q}(\mathbf{U},\boldsymbol{\alpha},\boldsymbol{\gamma})=&\sum_t\mathbb{E}_q[\log p(\boldsymbol{w}_t,\boldsymbol{a}_t|\mathbf{U},\boldsymbol{\alpha},\boldsymbol{\gamma})] \\
=&\sum_t\mathbb{E}_q[\log p(\boldsymbol{w}_t|\boldsymbol{a}_t,\mathbf{U},\boldsymbol{\gamma})]+\mathbb{E}_q[\log p(\boldsymbol{a}_t|\boldsymbol{\alpha})]
\end{aligned} \tag{13}$$

where

$$\begin{aligned}
&\mathbb{E}_q[\log p(\boldsymbol{w}_t|\boldsymbol{a}_t,\mathbf{U},\boldsymbol{\gamma})] \\
=&\sum_f\Big(\gamma_f\log\gamma_f-\gamma_f\sum_l U_{fl}\mathbb{E}_q[a_{lt}]-\log\Gamma(\gamma_f)+(\gamma_f-1)\log W_{ft}-W_{ft}\gamma_f\prod_l\mathbb{E}_q[\exp(-U_{fl}a_{lt})]\Big) \\
&\mathbb{E}_q[\log p(\boldsymbol{a}_t|\boldsymbol{\alpha})] \\
=&\sum_l\Big(\alpha_l\log\alpha_l-\log\Gamma(\alpha_l)+(\alpha_l-1)\mathbb{E}_q[\log a_{lt}]-\alpha_l\mathbb{E}_q[a_{lt}]\Big)
\end{aligned}$$

Take the derivative with respect to $\mathbf{U}, \boldsymbol{\alpha}, \boldsymbol{\gamma}$, we obtain the following gradients:

$$\frac{\partial \mathcal{Q}}{\partial U_{fl}} = \sum_t \Big(- \mathbb{E}_q[a_{lt}] + W_{ft}\mathbb{E}_q[a_{lt}](1 + \frac{U_{fl}}{\rho_{lt}^a})^{-(\nu_{lt}^a+1)} \prod_{j \neq l} \mathbb{E}_q[\exp(-U_{fj}a_{jt})]\Big) \tag{14}$$

$$\frac{\partial \mathcal{Q}}{\partial \alpha_l} = \sum_t \Big(\log \alpha_l + 1 - \psi(\alpha_l) + \mathbb{E}_q[\log a_{lt}] - \mathbb{E}_q[a_{lt}]\Big) \tag{15}$$

$$\frac{\partial \mathcal{Q}}{\partial \gamma_f} = \sum_t \Big(\log \gamma_f - \sum_l U_{fl}\mathbb{E}_q[a_{lt}] + 1 - \psi(\gamma_f) + \log W_{ft} - W_{ft} \prod_l \mathbb{E}_q[\exp(-U_{fl}a_{lt})]\Big) \tag{16}$$

Note that the optimization problem for $\mathbf{U}$ is independent for different frequency bin $f \in \{1, 2, \cdots, F\}$, as reflected by the gradient.