Harmonic Constraint in Music Spectrogram Reconstruction

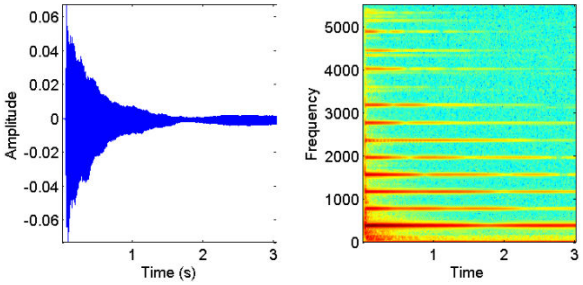
by Tongyu Lu https://lucainiaoge.github.io lutongyucn@foxmail.com, 15 Jan, 2021

It is assumed that readers have known the concepts about harmonic (overtone) structure of musical notes. And readers are recommended to know the non-negative matrix factorization (NMF) algorithm. But anyway, I tried to briefly elaborate it in this article.

Introduction: The Need to Design Harmonic Constraint

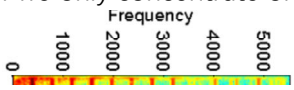
In automatic music transcription (AMT, or more precisely, multipitch estimation), sometimes we maintain a timbre dictionary W to store the overtone structure of musical notes.

For example, when we play the 70th piano key and record the sound for duration J, we may have a spectrogram of the audio recording, which is written as $w_{-,70}$. It may look like this:



(This figure is cited from B. Jacobson, Combined-channel instantaneous frequency analysis for audio source separation based on comodulation, Thesis (Ph. D.))

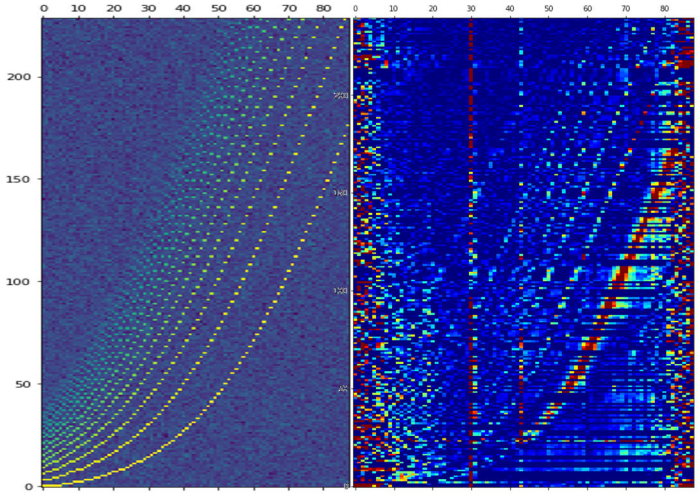
If we only concentrate on one moment, we extract a time bin like this:



We observe that the high-intensity frequency bins distribute uniformly in the frequency axis.

Now, we get to our problem: what if we want to learn to learn such a timbre dictionary from given audio? Apparently, this is a large topic. It requires us to **1. design representations of inputs/outputs/parameters/model-structure**, and **2. to design the optimization algorithm in order to converge into such a reasonable dictionary**.

Suppose we have already done step 1 and our parameter involves a timbre dictionary $W_{K\times J\times F}$, where $K = 88$ is the number of different piano keys (**components**) and J is the timbre diversity we want to consider for each piano key. And the problem is that, when we operate our optimization algorithm, the timbre dictionary may not converge into the desired overtone-structured codebook for piano keys, but may look like this:



(In the two figures above, the x-axis is K, the key dimension, while the y-axis is F, the frequency-bin dimension. The J dimension is squeezed for simplicity. **In the rest of this article, the J dimension is squeezed. Readers may interpret K dimension as considering both J and K.** The left one is the desired timbre dictionary, while the right one is the timbre dictionary converged. The two figures are generated by Tongyu Lu through experiments.)

We can see that the second one is poorly structured. For example, component #0, #30 and #87 are noisy. This is undesirable, because: when we apply this dictionary to reconstruct the audio spectrogram, it is not semantically meaningful. Therefore, we want to force our model parameter W to converge into a harmonically correct timbre dictionary. What can we do about it?

A Hard-Constraint Method for Harmonic Constraints

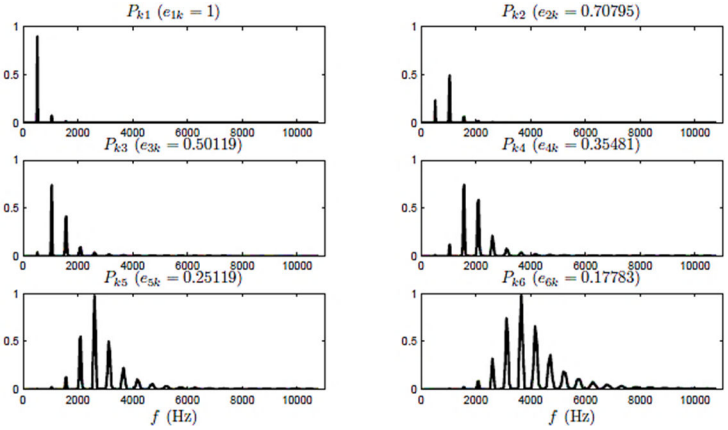
Emmanuel Vincent proposed a template-based harmonic constraint method in 2018, in paper **[1]***Harmonic and inharmonic Nonnegative Matrix Factorization for Polyphonic Pitch transcription*. That method evolved into a probabilistic version by Nancy Bertin, Roland Badeau and Emmanuel Vincent in paper **[2]***Enforcing Harmonicity and Smoothness in Bayesian Non-negative Matrix Factorization Applied to Polyphonic Music Transcription*, 2018. This section reviews that method.

They factorize each dictionary component $w_{-,k} = W[:,k]$ as the weighted sum of templates $p_{-,k,m} \in \mathbb{R}_+^F$, written as

$$w_{-,k} = \sum_m p_{-,k,m} e_{m,k}$$

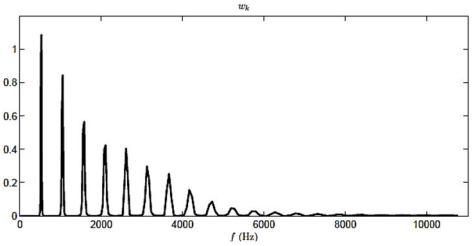
where $e_{m,k} \geq 0, m = 1, \dots, M$

Vectors $p_{-,k,m}$ are pre-defined overtone-templates, which look like:



(this figure is cited from paper [2]. According to [1,2], such templates are perceptually defined as narrow-band spectra)

And finally, the weighted summation of those templates becomes $w_{-,k}$. As you may see, $e_{m,k}$ are the weights for each component $p_{-,k,m}$. The summation may look like:



(this figure is cited from paper [2])

We can understand this factorization from the point of view of basis.

Define the harmonic space $\mathbb{H}_k \subset \mathbb{R}_+^F$ as such: choose M "basis" vectors $p_{-,k,m} \in \mathbb{R}_+^F$ and define $\mathbb{H}$ as the conic combination of $p_{-,k,m}, m = 1, \dots, M$, i.e., $\mathbb{H}_k = \left\{\sum_m p_{-,k,m} e_{m,k} : e_{m,k} \geq 0, m = 1, \dots, M\right\}$. Our task is then translated from directly estimating $w_{f,k}, f = 1, \dots, F;\ k = 1, \dots, K$ into estimating weight factors $e_{m,k}, m = 1, \dots, M;\ k = 1, \dots, K$. Easy to see, $\mathbb{H}_k$ is the domain where $w_{-,k}$ is defined.

In an NMF context (if NMF seems unfamiliar to you, feel free to check **[3]***A tutorial on Nonnegative Matrix Factorisation with applications to audiovisual content analysis*), the formula $V = WH$ becomes $V = (PE)H$. The product PE is not the matrix product in a traditional sense, but is defined as $W = PE = [w_{1,k}, \dots, w_{f,k}]$ where $w_{-,k} = \sum_m p_{-,k,m} e_{m,k}$

Actually, when we define the timbre dictionary W, we want to solve the optimization problem:

minimize $D(\hat{V}H|V)$ over W
subject to $W \geq 0,\ \hat{V} = WH$
where: V is the given spectrogram and H is the target music score

Enforcing the harmonic constraint above, this problem becomes

minimize $D(WH|V)$ over E
subject to $E \geq 0,\ \hat{V} = WH, W = PE$
where: V is the given spectrogram, H is the target music score and P is the pre-defined templates

Probabilistic Representation of the Hard Harmonic Constraints

The previous section showed that how hard-constraint proposed in [1] is represented. In [2], the hard-constraint in [1] was derived into a probabilistic version, which was inspired by a theorem proved in **[4]***Bayesian extensions to non-negative matrix factorisation for audio signal modelling*. Here is an introduction to such theorem:

- Paper [4] says that in a typical NMF context, when $D(\hat{V}|V)=\sum_{f,t} d_{KL}(\hat{v}_{f,t}|v_{f,t})$, then the distributions of $\hat{v}_{f,t}$ are Poisson. Specifically, the spectrogram components $\hat{v}_{f,t}$ could be written as summations of components, i.e., $\hat{v}_{f,t}=\sum_k|c_{f,k,t}|$. Such components $c_{f,k,t}$ are treated as a sample of random variable $|c_{f,k,t}|\sim\text{Poisson}(w_{f,k}h_{k,t})$, which leads to $\hat{v}_{f,t}\sim\text{Poisson}(\sum_k w_{f,k}h_{k,t})$.
- Paper **[5]***Nonnegative Matrix Factorization with the Itakura-Saito Divergence: With Application to Music Analysis subsequently* proved that, when Itakura-Saito difference is applied, i.e., $D(\hat{V}|V)=\sum_{f,t} d_{IS}(\hat{v}_{f,t}|v_{f,t})$, then the distributions of $\hat{v}_{f,t}$ are Gaussian. Specifically, $\hat{v}_{f,t}=\sum_k|c_{f,k,t}|$, and $c_{f,k,t}\in\mathbb{C}$, which satisfies $c_{f,k,t}\sim\mathcal{N}_c(0,w_{f,k}h_{k,t})$, where $\mathcal{N}_c(\mu,\Sigma)$ is the proper complex Gaussian distribution.

The probabilistic interpretation of the original NMF problem (as introduced above) is actually a detour: it converts the original matrix-product construction in NMF into a generative process under parameter W. Solving the maximum likelihood problem for $\hat{v}_{f,t}$ under parameter W certain pre-defined distributions (Poisson/Gaussian) is actually equivalent to solving the optimization problem for variable W with certain cost functions (KL/IS).

This gives a hint to solve the optimization problem: it is transformed into a maximum likelihood (ML) problem, which could be solved using EM algorithm. Paper [2] gave the update rule in the NMF context:

$$h_{kn}^{(\ell+1)}=h_{kn}^{(\ell)}\times\left(1+\frac{1}{FM}\sum_f\sum_m\frac{h_{kn}^{(\ell)}e_{mk}^{(\ell)}P_{km}(f)}{\hat{v}_{fn}}\left(\frac{v_{fn}}{\hat{v}_{fn}}-1\right)\right)\quad(14)$$

$$e_{mk}^{(\ell+1)}=e_{mk}^{(\ell)}\times\left(1+\frac{1}{FN}\sum_n\sum_f\frac{h_{kn}^{(\ell+1)}e_{mk}^{(\ell)}P_{km}(f)}{\hat{v}_{fn}}\left(\frac{v_{fn}}{\hat{v}_{fn}}-1\right)\right)\quad(15)$$

Enforcing Harmonic Constraints Using Regularization

Introduction: ML vs. MAP

Under the hard-constraint mentioned in [1], each dictionary component $w_{-,k}$ is restricted within $\mathbb{H}_k\subset\mathbb{R}_+^F$. This may lead to a problem: if the instruments do generate noise (e.g., the hammer of a piano key hitting the string, bringing noisy spectrum), the dictionary is not an expert to consider that. [2] did not solve this problem because its probabilistic modeling is not a permission for noise, but actually a detour to solve the deterministic optimization problem with EM algorithm in an ML framework.

What if we allow the model to consider a little bit noise? To do this, we must derive a soft-constraint on $w_{-,k}$ without confining it within $\mathbb{H}_k$. A reasonable way is to define prior distribution functions $p_k\colon\mathbb{R}_+^F\to[0,1]$ for every $w_{-,k}$. The distribution tells us the likelihood for every point in $\mathbb{R}_+^F$ that $w_{-,k}$ may situate. If the distribution tells us the likelihood is small, it does not say that it is impossible. The prior distribution $p(W)$ could be used to model the harmonic constraint, which is the idea of maximum a-posteriori (MAP).

To consider the prior, we just need to modify the original problem into:

minimize $D(WH|V)+R(W)$ over W
subject to $W\geq 0$
where: V is the given spectrogram and H is the target music score

The function $R(W)$ is a regularization term which represents the prior knowledge on W.

Why it could represent prior knowledge? This leads to the comparison between maximum likelihood (ML) and maximum a-posteriori (MAP).

- ML says: given parameter θ and data D, we want to maximize $p(D|\theta)$, or to minimize $-\log p(D|\theta)$, which leads to the loss function;
- MAP says, given parameter θ and data D, we want to maximize $p(\theta|D)$, or to minimize $-\log p(\theta|D)$, which leads to the loss function and the regularization term. Here is the trick:

$$-\log p(\theta|D)=-\log\frac{p(D|\theta)p(\theta)}{p(D)}=-\log p(D|\theta)-\log p(\theta)+\text{const}$$

The term $-\log p(\theta)$ is the regularization term.

In the context of timbre dictionary, the parameter is W. Therefore, $R(W)$ is proportional to $-\log p(W)$. Then comes the question: how to define $p(W)$?

It should tell us such prior information:

1. if $w_{-,k}$ has overtone on $f_{0,k}, f_{1,k}, \dots$, we expect that $w_{f_{0,k},k}$ is roughly the strongest among all frequency dimensions.
2. $f_{0,k}, f_{1,k}, \dots$ should define peaks in the frequency axis

The first prior knowledge tells us that: $\mathbb{E}[w_{f_{0,k},k}] > \mathbb{E}[w_{f,k}], \forall f \in [0, F], f \neq f_{0,k}$

The second prior knowledge tells us that: define a function $E_k(f) = \mathbb{E}[w_{f,k}]$, then $E_k(f)$ has peaks at $f = f_{0,k}, f_{1,k}, \dots$

Now, if we treat $E_k(f)$ in the manner of [1], we say that $E_k(f) = \sum_m p_{-,k,m} e_{m,k}$. And we observe that such interpretation considers both prior conditions above. But where is the difference of this MAP idea between [1]?

Here comes my model:

Adjustable Harmonic Regularization: to be Realistic or Idealistic

Assume that each dictionary component is subject to a log-normal prior distribution. Which says:

$$\log w_{f,k} \sim \mathcal{N}([E_k(f)]_{dB}, \sigma^2)$$

Define $W_{dB} = 10 \lg W$ (log by element), and $\widetilde{W} = [E_k(f)]$

Assume that $p(W_{dB}) = \prod_{k,f} \frac{1}{\sqrt{2\pi}\sigma} \exp - \frac{\left([w_{f,k}]_{dB} - [E_k(f)]_{dB}\right)^2}{2\sigma^2} = \frac{1}{\sqrt{2\pi}\sigma} \left(\exp - \left\| W_{dB} - \widetilde{W}_{dB} \right\|_F^2 / 2\sigma^2 \right)$. (Note that this assumption does not indicate that $w_{f,k}$ are independent, because they are linked by $E_k(f)$.)

Then, $-\log p(W_{dB}) = \left\| W_{dB} - \widetilde{W}_{dB} \right\|_F^2 / 2\sigma^2 + \text{const} \triangleq \lambda \left\| W_{dB} - \widetilde{W}_{dB} \right\|_F^2$

Therefore, the MAP version of NMF becomes

$$\begin{aligned} &\text{minimize } D(WH|V) + \lambda \left\| W_{dB} - \widetilde{W}_{dB} \right\|_F^2 \text{ over } W \\ &\text{subject to } W \geq 0 \\ &\text{where: } V \text{ is the given spectrogram and } H \text{ is the target music score} \end{aligned}$$

There is one thing that remains uncertain: should $\widetilde{W}_{dB}$ be an adjustable parameter or fixed?

My choice is to let it be adjustable within the pre-defined $\mathbb{H}_{dB}$.

Here is the idea: W_{dB} is realistic, while $\widetilde{W}_{dB}$ is idealistic.

- W_{dB} is realistic means that it is devoted to approximate real spectrogram, and thus, its goal is to do spectrogram reconstruction regardless of harmonic constraints (unless $\widetilde{W}_{dB}$ tells so).
- $\widetilde{W}_{dB}$ is idealistic means that it does not care whether reconstruction is beautifully achieved, but cares about its ideal that it should always obey harmonic constraints.
- The role of regularization term to make W_{dB} more idealistic while making $\widetilde{W}_{dB}$ more realistic (for example, $\widetilde{W}_{dB}$ should obey the basic overtone structure of the target instrument).

We have to exert our prior knowledge on $\widetilde{W}_{dB}$ such that it obeys harmonic constraint. Here is where paper [1] comes into use: we can assume that $\widetilde{w}_{f,k} = E_k(f) = \sum_m p_{-,k,m} e_{m,k}$.

Now, our problem becomes:

$$\begin{aligned} &\text{minimize } D(WH|V) + \lambda \left\| W_{dB} - \widetilde{W}_{dB} \right\|_F^2 \text{ over } W_{dB}, E \\ &\text{subject to } \widetilde{W} = PE,\ W = 10^{W_{dB}/10},\ \widetilde{W}_{dB} = 10 \lg \widetilde{W} \\ &\text{where: } V \text{ is the given spectrogram, } H \text{ is the target music score and } P \text{ is the pre-defined templates} \end{aligned}$$

By the Way: What is the dB Version of NMF?

In the statements above, the dB version of NMF problem was proposed.

In this problem, the parameter W_{dB} is not subject to negative constraint. This is good news for neural networks! When doing optimization, the neural network does not care if its parameters are negative or positive, which makes it impossible to leverage the NMF structure in neural networks. But equipped with the log trick, we do not need to worry about the feasibility of maintaining a timbre dictionary. What we need to do is to add an "Amp2dB" layer and a

"dB2Amp" layer, and both are differentiable.

Here comes the algorithm for spectrogram reconstruction and parameter update:

Input: spectrogram V, its piano-roll matrix H
Parameter: W_{dB}
Algorithm:
1. $W = \mathrm{dB2Amp}(W_{dB})$
2. $\hat{V} = WH$
3. compute reconstruction loss $L_r(\hat{V}|V)$
4. update W_{dB} using $\nabla_{W_{dB}} L_r(\hat{V}|V)$

Or we could do this totally in dB:

Input: spectrogram V_{dB}, its piano-roll matrix H
Parameter: W_{dB}
Algorithm:
1. $W = \mathrm{dB2Amp}(W_{dB})$
2. $\hat{V} = WH$
3. $\hat{V}_{dB} = \mathrm{Amp2dB}(\hat{V})$
4. compute reconstruction loss $L_{r\mathrm{dB}}(\hat{V}_{dB}|V_{dB})$
5. update W_{dB} using $\nabla_{W_{dB}} L_{r\mathrm{dB}}(\hat{V}_{dB}|\hat{V}_{dB})$

It is super simple! In the context of traditional NMF, we have to use MU algorithm or EM algorithm in order to maintain non-negativity. But now we could just use gradient descent!

I am now concerned with several questions: 1. is this log trick already considered? 2. is this log trick realistic? (for example, will the "dB2Amp" operation explode the gradient and subsequently make it impossible)

So far, I have not noticed relevant works on this log trick. But I still do not know if it is realistic, and I am going to do several experiments on that.

References

[1]. E. Vincent, N. Bertin and R. Badeau, "Harmonic and inharmonic Nonnegative Matrix Factorization for Polyphonic Pitch transcription," *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, Las Vegas, NV, 2008, pp. 109-112, doi: 10.1109/ICASSP.2008.4517558.

[2]. N. Bertin, R. Badeau and E. Vincent, "Enforcing Harmonicity and Smoothness in Bayesian Non-Negative Matrix Factorization Applied to Polyphonic Music Transcription," in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 538-549, March 2010, doi: 10.1109/TASL.2010.2041381.

[3]. A tutorial on Nonnegative Matrix Factorisation with applications to audiovisual content analysis, S. Essid & A. Ozerov, Telecom ParisTech / Technicolor, July 2014, URL: https://perso.telecom-paristech.fr/essid/teach/NMF_tutorial_ICME-2014.pdf

[4]. T. Virtanen, A. Taylan Cemgil and S. Godsill, "Bayesian extensions to non-negative matrix factorisation for audio signal modelling," *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, Las Vegas, NV, 2008, pp. 1825-1828, doi: 10.1109/ICASSP.2008.4517987.

[5]. C. Févotte, N. Bertin and J. Durrieu, "Nonnegative Matrix Factorization with the Itakura-Saito Divergence: With Application to Music Analysis," in *Neural Computation*, vol. 21, no. 3, pp. 793-830, March 2009, doi: 10.1162/neco.2008.04-08-771.