Audio Imputation using the Non-negative Hidden Markov Model

Jinyu Han[1]*, Gautham J. Mysore[2] and Bryan Pardo[1]

[1] EECS Department, Northwestern University.
[2] Advanced Technology Labs, Adobe Systems Inc.

Abstract. Missing data in corrupted audio recordings poses a challenging problem for audio signal processing. In this paper we present an approach that allows us to estimate missing values in the time-frequency domain of audio signals. The proposed approach, based on the Non-negative Hidden Markov Model, enables more temporally coherent estimation for the missing data by taking into account both the spectral and temporal information of the audio signal. This approach is able to reconstruct highly corrupted audio signals with large parts of the spectrogram missing. We demonstrate this approach on real-world polyphonic music signals. The initial experimental results show that our approach has advantages over a previous missing data imputation method.

1 Introduction

The problem of missing data in an audio spectrogram occurs in many scenarios. For example, the problem is common in signal transmission, where the signal quality is degraded by linear or non-linear filtering operations. In other cases, audio compression and editing techniques often introduce spectral holes to the audio. Missing values also occur frequently in the output of audio source separation algorithms, due to time-frequency component masking [2]. Audio imputation is the task of filling in missing values of the audio signal to improve the perceived quality of the resulting signal. An effective approach for audio imputation could benefit many important applications, such as bandwidth extension, sound restoration, audio declipping, and audio source separation.

Audio imputation from highly corrupted recordings can be a challenging problem. The popular existing generic imputation algorithm [1] is usually ill-suited for use with audio signals and results in audible distortions. Other algorithms such as those in [6] are suitable for imputation of speech, or in the case of musical audio [3] or [7]. However, these algorithms treat individual time frames of the spectrogram as independent of adjacent time frames, disregarding the important temporal dynamics of sound, which makes them less effective for complex audio scenes or severely corrupted audio.

In this paper, we propose an audio imputation algorithm, based on the Non-negative Hidden Markov Model (N-HMM) [4], which takes the temporal dynamics of audio into consideration. The N-HMM jointly learns several small spectral dictionaries as well as a Markov chain that describes the structure of transitions

* This work was supported in part by National Science Foundation award 0643752.

between these dictionaries. We extend the N-HMM for missing values imputation by formulating the imputation problem in an Expectation–Maximization (EM) framework. We show promising performance of the proposed algorithm by comparing it to an existing imputation algorithm on real-world polyphonic music audio.

2 Proposed Method

In this section, we describe the proposed audio imputation method. We first give an overview of the modeling strategy. We then briefly describe the probabilistic model that we employ, followed by the actual imputation methodology.

2.1 Overview

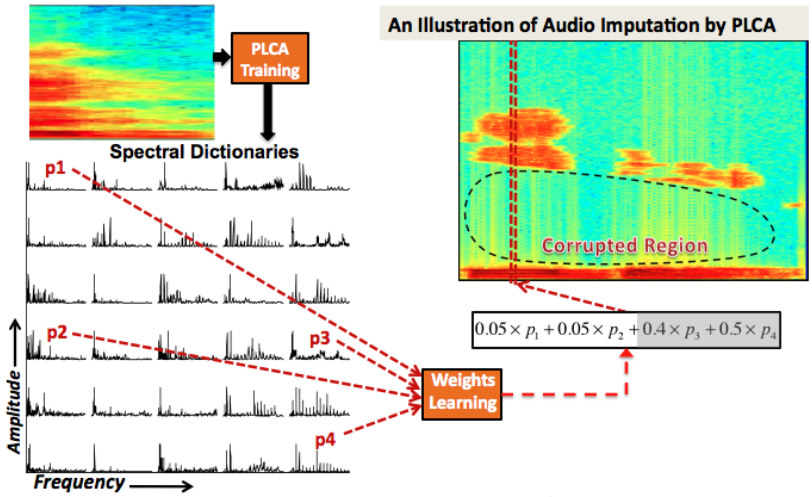

Fig. 1. General Procedure of Supervised Audio Imputation

The general procedure of supervised audio imputation methods [7] is as follows. First make a dictionary of spectral vectors from the training data using a non-negative spectrogram factorization technique, such as Non-negative Matrix Factorization (NMF) or Probabilistic Latent Component Analysis (PLCA). Each frame of the spectrogram is then modeled as a linear combination of the spectral vectors from the dictionary. Given the spectrogram of a corrupted audio, we estimate the weights for each spectral vector as well as the expected values for the missing entries of the spectrogram using an EM algorithm.

Fig.1 shows an example of Audio Imputation using PLCA. In this example, a dictionary of spectral vectors is learned from an intact audio spectrogram. Given corrupted audio that is similar to the training audio, the original audio spectrogram can be estimated by a linear combination of the spectral vectors from the dictionary.

Previous audio imputation methods [3][7] are based on NMF or PLCA to learn a single dictionary of spectral vectors to represent the entire signal. These approaches treat individual time frame independently, ignoring the temporal dynamics of audio signal. Furthermore, it is not always the case that the training

data has exactly the same characteristics as the corrupted audio. For example, the corrupted audio may contain a piano solo playing an intro of a song but the training audio from the same song may contain the piano source and the singing voice. In this case, a single dictionary learned from a mixture of the piano and singing voice may be less effective in reconstructing the piano sound from the corrupted audio. This may introduce interference to the reconstructed piano sound from the dictionary elements that are used to explain the singing voice.

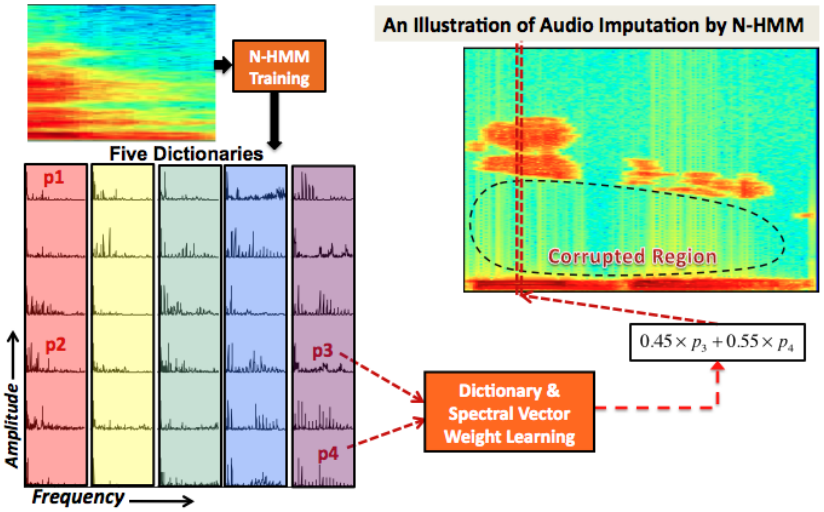

Fig. 2. Supervised Audio Imputation using a N-HMM

As shown in Fig.2, our proposed approach uses a N-HMM to learn several small dictionaries from the training audio. Dictionaries are associated with states in a model that incorporates the dynamic temporal structure of the given audio signal. Several small dictionaries are learned from the training data to explain different aspects of the audio signal. During the imputation process, only spectral vectors from one dictionary are used to reconstruct a certain frame of the corrupted spectrogram. In this way, it is less likely to introduce interference from other sources of the training data.

2.2 Probabilistic Model

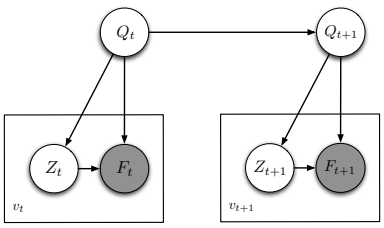

Fig. 3. Graphical Model of the N-HMM. {Q, Z, F} is a set of random variables and {q, z, f} are the realization of the random variables. v_t represents the number draws at time t.

Fig.3 shows the graphical model of the N-HMM, an extension of Hidden Markov Model (HMM) by imposing non-negativity on the observation model. The observation alphabet for each state q in the N-HMM is a dictionary of

spectral vectors. Each vector z can be thought of a magnitude spectrum. To maintain consistency with prior work [4] we treat it as a probability distribution. For frequency f and time t, we notate the contribution to the magnitude of the spectrogram from a spectral vector z in dictionary q as $P(f_t|z_t, q_t)$. Here, f is one of a set of K frequencies of analysis of a spectrogram. At time t, the observation model is obtained by a linear combination of all spectral vectors z from the current dictionary q:

$$P(f_t|q_t) = \sum_{z_t} P(z_t|q_t)P(f_t|z_t, q_t) \tag{1}$$

where $P(z_t|q_t)$ is the spectral vector mixture weight, given q_t. The transitions between states are modeled with a Markov chain, given by $P(q_{t+1}|q_t)$.

In our model, we assume the spectrum V_t at time t is generated by repeated draws from a distribution $P_t(f)$ given by

$$P_t(f) = \sum_{q_t} P(f_t|q_t)\gamma_t(q_t) \tag{2}$$

where $\gamma_t(q_t)$ is the distribution over the states, conditioned on all the observations over all time frames. We can compute $\gamma_t(q_t)$ using the forward-backward algorithm as in the traditional HMM. Please refer to [4] for the full formulation. Here, the resulting value $P_t(f)$ can be thought as an estimation of the relative magnitude of the spectrum at frequency f and time t.

A comparison between a N-HMM and a PLCA is illustrated in Fig.4. Compared to most other non-negative spectrogram decomposition techniques, the N-HMM has taken into account the temporal dynamics of the audio signal. Instead of using one large dictionary to explain everything in the audio, the N-HMM learns several small dictionaries, each of which will explain a particular part of the spectrogram. All the parameters of the N-HMM can be learned using the EM algorithm detailed in [4].

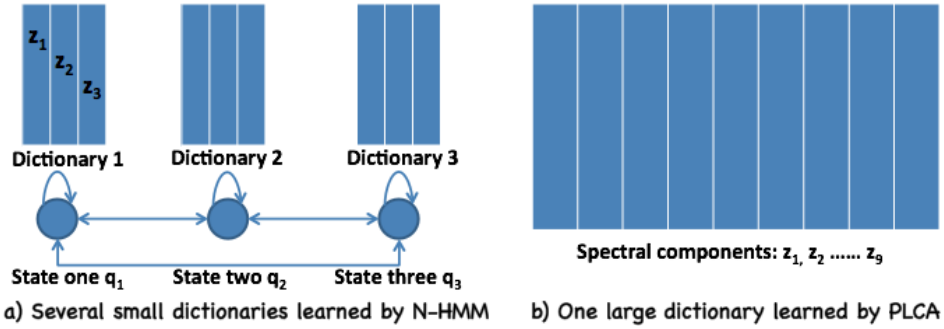

Fig. 4. A comparison between PLCA and N-HMM

3 Estimation of incomplete data

When the spectrogram is incomplete, a great deal of the entries in the spectrogram could be missing. In this paper, we assume the locations of the corrupted

bins are known. Identifying the corrupted region is beyond the scope of this paper. Our objective is to estimate missing values in the magnitude spectrogram of audio signals.

In the rest of the paper we use the following notation: we will denote the observed regions of any spectrogram V as V^o and the missing regions as $V^m = V \setminus V^o$. Within any magnitude spectrum V_t at time t, we will represent the set of observed entries of V_t as V_t^o and the missing entries as V_t^m. $\mathscr{F}_t^o$ will refer to the set of frequencies for which the values of V_t are known, i.e. the set of frequencies in V_t^o. $\mathscr{F}_t^m$ will similarly refer to the set of frequencies for which the values of V_t are missing, i.e. the set of frequencies in V_t^m. $V_t^o(f)$ and $V_t^m(f)$ will refer to the magnitude at frequency f of V_t^o and V_t^m respectively.

To estimate the magnitude of each value in V_t^m we need to scale the value $P_t(f)$ from Eq.2. We do not know the total amplitude at time t because some values are missing. Therefore, we must estimate a scaling factor. We sum the values of the uncorrupted frequencies in the original audio to get $n_t^o = \sum_{f \in \mathscr{F}_t^o} V_t^o(f)$. We sum the values of $P_t(f)$ for $f \in \mathscr{F}_t^o$ to get $p_t^o = \sum_{f \in \mathscr{F}_t^o} P_t(f)$. The expected amplitude at time t is obtained by dividing n_t^o by p_t^o. This gives us a scaling factor. The expected value of any missing term $V_t^m(f)$ can be estimated by:

$$E[V_t^m(f)] = \frac{n_t^o}{p_t^o} P_t(f) \tag{3}$$

The audio imputation process is as follows:

1. Learn the parameters of a N-HMM from the training audio spectrogram, using the EM algorithm.
2. Initialize the missing entries of the corrupted spectrogram to random values.
3. Perform the N-HMM learning on the corrupted spectrogram from step 2. During the learning process,
   - Fix most of the parameters such as $P(f|z,q)$ and $P(q_{t+1}|q_t)$ to the learned parameters from step 1.
   - Learn the remaining parameters in the N-HMM model using the EM algorithm. Specifically, learn the weights distributions $P(z_t|q_t)$. Then estimate the posterior state distribution $\gamma_t(q_t)$ using the forward-backward algorithm and update $P_t(f)$ using Eq.2.
   - At each iteration, update every missing entry in the spectrogram with its expected value using Eq.3.
4. Reconstruct the corrupted audio spectrogram by:

$$\bar{V}_t(f) = \begin{cases} V_t(f) & \text{if } f \in \mathscr{F}_t^o \\ E[V_t^m(f)] & \text{if } f \in \mathscr{F}_t^m \end{cases} \tag{4}$$

5. Convert the estimated spectrogram to the time domain.

This paper does not address the problem of missing phase recovery. Instead we use the recovered magnitude spectrogram with the original phase to re-synthesize the time domain signal. We found this to be more perceptually pleasing than a standard phase recovery method [5].

4 Experiments

We test the proposed N-HMM audio imputation algorithm on real-world polyphonic musical data. We performed the experiment on 12 real-world pop music songs. The proposed method is compared to a recent audio imputation method using PLCA [7].

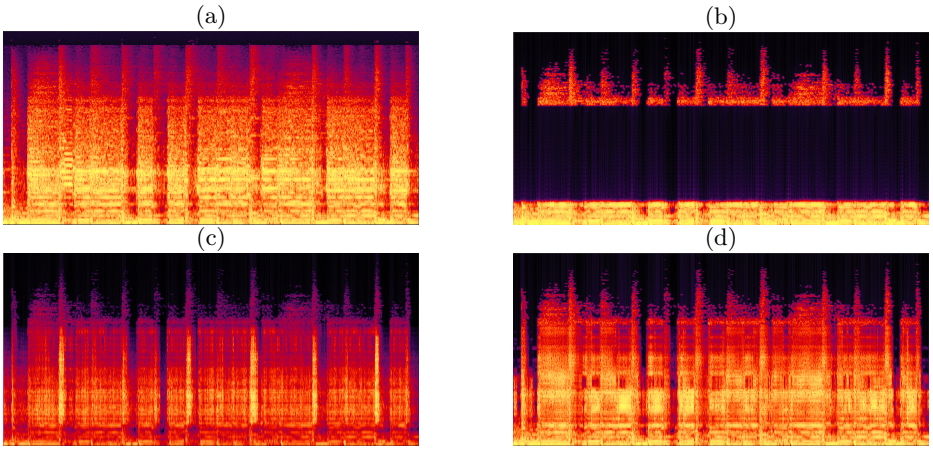

Fig. 5. 5.5-second audio clip from "Born to be wild" by "Steppenwolf". The x-axis represents time and the y-axis represents frequency. a)Original audio; b) Corrupted audio input (1.05 dB SNR); c) Imputation result by PLCA (1.41 dB SNR); d) Imputation result by proposed algorithm (4.89 dB SNR).

For a particular audio clip, both the testing data and training data are taken from the same song. The testing data is about 6-second long, taken from the beginning of a song. The corrupted audio is obtained from the testing data by removing all the frequencies between 800 Hz and 12k Hz in the spectrogram. Another clip (not containing the testing audio) of about 11-second long is taken from the same song as the training data. The details of each audio clip are listed in Table.1. We learn the N-HMM parameters for each song from the training data. Specifically, we learned 10 dictionaries of 8 spectral vectors each, as well as the transition matrix from the training data. We then update the N-HMM for the corrupted audio during the imputation process. When using PLCA, we learn one dictionary of 40 spectral vectors. The values for the parameters are determined by the authors empirically. Signal-to-Noise-Ratio (SNR)[3] is used to measure the outputs of both imputation methods. During the experiments, we found existing signal measurements do not always correspond well to the perceptual quality of the audio. Examples of the experimental results are available at the authors' website [8]. These show the perceptual quality of the reconstructed signals.

We first examine two examples that favor the proposed approach against the PLCA method. The first one is a 5.5-second audio clip from "Born to be

[3] $SNR = 10log_{10}\frac{\sum_t s(t)^2}{\sum_t(\bar{s}(t)-s(t))^2}$ where $s(t)$ and $\bar{s(t)}$ are the original and the reconstructed signals respectively.

wild" by "Steppenwolf". The spectrogram of the original audio, corrupted audio, output of the proposed method and PLCA are illustrated in Fig.5. The proposed method produces an output with a higher SNR than PLCA.

The next example is a 5.4-second audio clip from "Scar Tissue" by "Red Hot Chili Peppers". In this example, both PLCA and the proposed method improve the SNR of the corrupted audio by about 7 dB. The proposed method has a lower SNR measurement, however the output of the proposed method is perceptually better than the output of the PLCA method. This difference is also shown in the spectrogram plot in Fig.6. The spectrogram reconstructed by PLCA has more random energy scatted in the high frequency region, while the proposed method only reconstructs the signal in the region where it should have been.

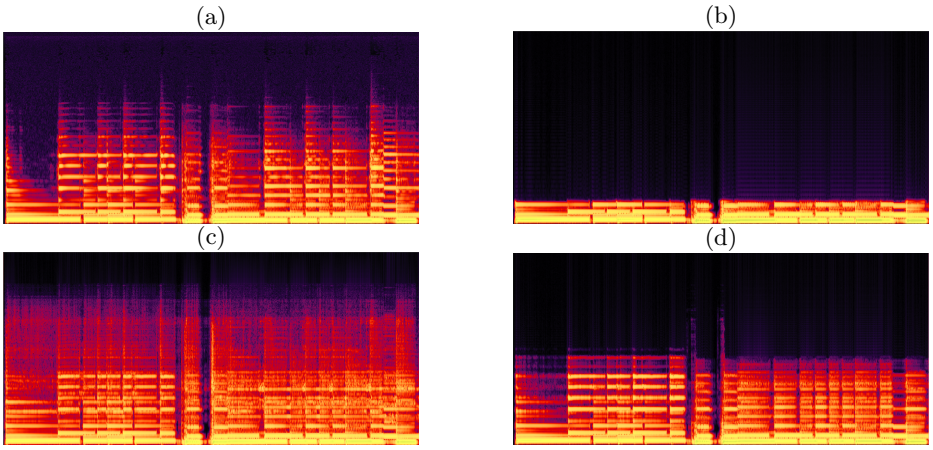

Fig. 6. 5.4-second audio clip from "Scar Tissue" by "Red Hot Chili Peppers". The x-axis represents time and the y-axis represents frequency. a) Original Audio; b) Corrupted Audio Input (7.46 dB SNR); c) Imputation result by PLCA (15.56 dB SNR); d) Imputation result by proposed algorithm (14.34 dB SNR).

Table 1 presents the performance of PLCA and the proposed algorithm on 12 clips of real-world music recordings using the SNR measurement. The average performance of the proposed method is 15.32 dB SNR, improving 5.67 dB from the corrupted audio and 1.8 dB from the output of the PLCA. The proposed method has better SNR measurement than PLCA on 9 out of 12 song clips. For the audio where the proposed method does not have better SNR measurement, as shown by the example in Fig.6, the proposed method may still produce an audio signal with equivalent or better perceptual quality. We encourage the readers to compare the results of both methods by listening to examples at the authors' website [8].

Table 1. Performances of the Imputation results by the proposed method and PLCA

	SNR (dB)			Audio length (Second)	
Song name	Input	Proposed	PLCA	Testing	Training
Better together	11.23	**22.48**	19.5	4.5	10.3
1979	14.43	**19.72**	18.07	5.7	11.3
Born to be wild	1.05	**4.89**	1.41	5.5	20.4
Scar tissue	7.46	14.34	**15.56**	5.4	10
Bad day	6.48	**13.84**	12.55	6.3	11.5
Wonderwall	-2.21	**8.36**	5.28	5.8	5.5
Here I go again	11.49	**15.95**	14.5	5.1	9.5
Every breath you take	7.46	14.34	**15.65**	6.9	10
Viva La Vida	7.6	11.66	**11.77**	6.2	10.1
She will be loved	17.66	**18.46**	15.2	5.7	11.9
Making memories of us	18.06	**21.3**	18.11	9.8	12.8
Daughters	15.11	**18.47**	14.56	8.2	16.2
Average measurement	9.65	**15.32**	13.52	6.29	11.63

5 Conclusions

In this paper we present an approach that allows us to estimate the missing values in the time-frequency domain of audio signals. The proposed approach is based on the N-HMM, which enables us to learn the spectral information as well as the temporal dynamics of the audio signal. Initial experimental results showed that this approach is quite effective in reconstructing missing values from corrupted spectrograms and has advantages over performing imputation using PLCA. Future work includes developing techniques for missing phase recovery.

References

1. Brand, M.: Incremental singular value decomposition of uncertain data with missing values. ECCV pp. 707–720 (2002)
2. Han, J., Pardo, B.: Reconstructing completely overlapped notes from musical mixtures. In: ICASSP (2011)
3. Le Roux, J., Kameoka, H., Ono, N., de Cheveigné, A., Sagayama, S.: Computational auditory induction as a missing-data model-fitting problem with bregman divergence. Speech Communication (2010)
4. Mysore, G.J.: A Non-negative Framework for Joint Modeling of Spectral Structure and Temporal Dynamics in Sound Mixtures. Ph.d. dissertation, Stanford University (2010)
5. Nawab, S., Quatieri, T., Lim, J.: Signal reconstruction from short-time fourier transform magnitude. Acoustics, Speech & Signal Processing, IEEE Trans. 31, 986–998 (1983)
6. Raj, B.: Reconstruction of Incomplete Spectrograms for Robust Speech Recognition. Ph.d. dissertation, Carnegie Mellon University (2000)
7. Smaragdis, P., Raj, B., Shashanka, M.: Missing data imputation for time-frequency representations of audio signals. J. Signal Processing Systems (2010)
8. `www.cs.northwestern.edu/~jha222/imputation`