MISSING DATA IMPUTATION FOR SPECTRAL AUDIO SIGNALS

Paris Smaragdis
Adobe Systems
Newton, MA, USA

Bhiksha Raj
Carnegie Mellon University
Pittsburgh, PA, USA

Madhusudana Shashanka
Mars Inc.
Hackettstown, NJ, USA

ABSTRACT

With the recent attention to audio processing in the time - frequency domain we increasingly encounter the problem of missing data. In this paper we present an approach that allows for imputing missing values in the time-frequency domain of audio signals. The presented approach is able to deal with real-world polyphonic signals by performing imputation even in the presence of complex mixtures. We show that this approach outperforms generic imputation approaches, and we present a variety of situations that highlight its utility.

1. INTRODUCTION

In this paper we address the problem of estimating missing regions of time-frequency representations of audio signals. The problem of missing data in the time-frequency domain occurs in several scenarios. The problem is common in speech processing algorithms that employ computational auditory scene analysis or related methods to mask out time-frequency components for recognition, denoising or signal separation [1, 2]. An increasing number of audio processing tools allow interactive spectral editing of audio signals, which can often result in the excision of time-frequency regions of sound. In yet other scenarios, such as in signals that have passed through a telephone or have encountered other linear or non-linear filtering operations, excision of time-frequency regions of the signal can occur naturally.

In a majority of these scenarios the goal is to resynthesize the audio from the incomplete time-frequency characterizations. To do so, the "missing" regions of the time-frequency representations must first be "filled in" somehow, in order to effect the transform from the time-frequency representation to a time-domain signal. In certain cases the missing values can be set to zero and the resulting reconstructions do not suffer heavily from perceptible artifacts. In most cases however a moderate to severe distortion is easily noticeable. This distortion can render speech recordings unintelligible, or severely reduce the quality of a music recording thereby distracting the listener.

Although existing generic imputation algorithms [3] can be used to infer the values of the missing data they are often ill-suited for use with audio signals and result in audible distortions. Other algorithms such as those in [1, 4] are suitable for imputation of missing time-frequency components in speech. However, these algorithms typically exploit continuity in spectral structures and are implicitly aided by the fact that speech recordings usually contain signals from a single source (the voice) and exhibit spectro-temporal patterns from that source alone. General audio or music recordings however include a variety of sounds, many of which are concurrently active at any time, and each of which has its own typical patterns, and are hence much harder to model.

The algorithm proposed in this paper characterizes time-frequency representations as histograms of draws from a mixture model as proposed in [5]. Being a mixture, the model is implicitly capable of representing multiple concurrent spectral patterns. The process of imputing missing values then becomes one of learning from incomplete data. Experimental evaluations show that the imputed spectral constructions obtained with our algorithm result in distinctly better resynthesized signals than those obtained from other common imputation methods.

2. MISSING DATA IN THE TIME-FREQUENCY DOMAIN

Before we proceed, we briefly describe the time-frequency representations used to characterize the signal, and present some examples with missing data. We will assume that the time-frequency representations are derived through short-time Fourier transformation (STFT) of the signal [6]. The short-time Fourier transform converts an audio signal into a sequence of vectors, each of which represents the Fourier spectrum of a short (typically 20-60ms wide) segment or *frame* of the signal. The STFT of a signal can be inverted to the original time-domain signal by an inverse short-time Fourier transform. Being complex, the STFT of the signal has both a magnitude and a phase. However most sound processing algorithms for denoising, recognition, synthesis and editing operate primarily on the magnitude since it is known to represent most of the perceptual information in the signal – the phase contributes mainly to the perceived quality of the signal rather than its intelligibility. Recovery of full or partial missing phase information can be done with good results using the technique in [7]. Since the phase reconstruction is highly dependent on the magnitude values we will primarily examine the reconstruction of the magnitudes of the missing time-frequency terms in this paper.

Figure 1a shows an example of the magnitude of the STFT of a speech signal that is a mixture of a spoken utterance and a phone ring. We will refer to matrix-like representations of the magnitudes of STFTs of a signal, such as the one in Figure 1a as "magnitude spectrograms", or simply as "spectrograms" in this paper. Spectral magnitudes have the property that when multiple sounds co-occur, the spectral magnitudes of the mixed signal are approximately (but not exactly) equal to the sum of the spectral magnitudes of the component sounds[1]. This is apparent in the spectrogram in Figure 1a which shows the distinctive spectral patterns of both the speech and phone ring.

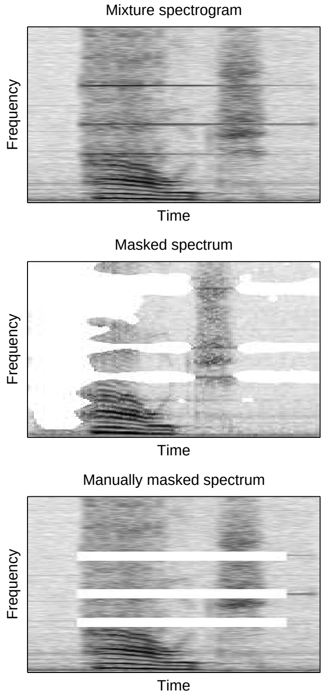

Fig. 1. **a**. Magnitude spectrogram of a mixture of a spoken utterance and a phone ring. The spectrogram clearly shows the spectral patterns for both signals – the wavy lines and the clouds of data are derived from the spoken word and the phone ring is represented by the three horizontal, parallel and thin lines. **b**. An incomplete spectrogram with missing regions. In this example time-frequency components deemed not to belong to the main speech signal have been erased by an automated algorithm. **c**. Another example of an incomplete spectrogram. Here time-frequency regions dominated by the phone ring have been manually edited out.

A spectrogram with "missing" data is one where some time-frequency components have been lost or erased due to some reason. Figures 1b and 1c show examples of spectrograms with missing data. In these examples time-frequency regions of the spectrogram have been erased to eliminate the phone ring from the signal, automatically in one case and by manual editing in the other. Missing data may also occur for other reasons, such as systematic filtering etc. In order to reconstruct a time-domain signal from these incomplete spectrograms the values in the missing regions must be imputed somehow. A simple technique is to simply floor these terms to some threshold value; however time-domain signals reconstructed from such spectrograms will often contain audible and often unacceptable artefacts. A more principled approach is required to reconstruct the missing regions in an acceptable manner.

3. MODELING THE SPECTROGRAM

At the outset, we would like to specify the terminology and notation we will use. We will denote the (magnitude) spectrogram of any signal as $\mathbf{S}$. The spectrogram consists of a sequence of (magnitude) spectral vectors $S_t,\ 0 \le t < T$, each of which in turn consists of a number of frequency components $S_t(f),\ 0 \le f < F$. All of these terms, being magnitudes are non-negative.

In our model we view $\mathbf{S}$ as a scaled version of a histogram $\hat{\mathbf{S}}$ comprising purely integer valued components, such that $\mathbf{S} = \mathcal{C}^{-1}\hat{\mathbf{S}}$; however the scaling factor $\mathcal{C}$ cancels out of all equations in our formulation and is thus not required to be known. In the rest of the paper, we therefore treat $\mathbf{S}$ itself as a histogram and do not explicitly invoke the scaling factor, besides assuming that it is very large so that for any $\hat{S}_t(f)$ such that $S_t(f) = \mathcal{C}^{-1}\hat{S}_t(f)$ the following holds:

$$\mathcal{C}^{-1}\hat{S}_t(f) \approx \mathcal{C}^{-1}(\hat{S}_t(f) + 1). \quad (1)$$

4. MODEL DEFINITION

We model each spectral vector S_t as a scaled histogram of draws from a mixture multinomial distribution. Per our model, S_t is generated by repeated draws from a distribution $P_t(f)$, where f is a random variable over the frequencies $\{1, \ldots, F\}$, and $P_t(f)$ is a mixture multinomial given by

$$P_t(f) = \sum_z P_t(z)P(f|z) \quad (2)$$

Here z represents the identity of the multinomial components in the mixture. $P(f|z)$ are the multinomial components or "bases" that compose the mixture. Note that the component multinomials $P(f|z)$ are not specific to any given S_t but are the same at all t. $P(f|z)$ are thus assumed to be characteristic to the entire data set of which $\mathbf{S}$ is representative. The only parameter that is specific to t are the mixture weights $P_t(z)$.

The model essentially characterizes the spectral vectors themselves as additive combinations of of histograms drawn

[1] Although in theory this statement holds true for spectral *power* when the component signals are uncorrelated, in practice phase cancelations in finite analysis windows make this true for spectral magnitudes raised to a power that is closer to 1.0 than 2.0.

from each of the multinomial bases. Consequently, it is inherently able to model complex sounds such as music that are additively composed by several component sounds. This is in contrast to conventional models used for data imputation *e.g.* [1, 2, 4], that model the spectral components as the outcome of a single draw from a distribution (although the distribution itself might be a mixture) and cannot model the additive nature of the data.

The model of Equation 2 is nearly identical to the one-sided model for Probabilistic Latent Semantic Analysis, introduced by Hoffman [8], with the distinction that whereas the original PLSA model characterizes random variables as documents and words, we refer instead to time and frequencies. Also, while the one-sided PLSA specifies a probability distribution over documents, in our model we do not have a similar probability distribution over the time variable t.

4.1. Learning the model parameters

We can estimate the parameters of the generative model in Equation 2 for $\mathbf{S}$ using the Expectation Maximization algorithm [8]. In the expectation step of the EM algorithm at each time t we estimate the *a posteriori* probabilities for the multinomial bases conditioned on the observation of a frequency f as:

$$P_t(z|f) = \frac{P_t(z)P(f|z)}{\sum_{z'} P_t(z')P(f|z')} \tag{3}$$

In the maximization step we update the spectral-vector-specific mixture weights $P_t(z)$ and data-set characteristic multinomial bases $P(f|z)$ as:

$$P_t(z) = \frac{\sum_f P_t(z|f)S_t(f)}{\sum_{z'} \sum_f P_t(z'|f)S_t(f)} \tag{4}$$

$$P(f|z) = \frac{\sum_t P_t(z|f)S_t(f)}{\sum_{f'} \sum_t P_t(z|f')S_t(f)} \tag{5}$$

5. ESTIMATION WITH INCOMPLETE DATA

The algorithm of Section 4.1 assumes that the entire spectrogram $\mathbf{S}$ is available to learn model parameters. When the spectrograms are incomplete, several of the $S_t(f)$ terms in Equation 4 will be missing or otherwise unknown. Our objective in this paper is to estimate the missing components on the data. Along the way we will also estimate the model parameters themselves as necessary.

In the rest of the paper we use the following notation: we will denote the *observed* regions of any spectrogram $\mathbf{S}$ as $\mathbf{S}_o$ and the *missing* regions as $\mathbf{S}_m$. Within any spectral vector S_t of $\mathbf{S}$, we will represent the set of observed components as S_t^o and the missing components as S_t^m. $S_t^o(f)$ and $S_t^m(f)$ will refer to specific frequency components of S_t^o and S_t^m respectively. $\mathcal{F}_t^o$ will refer to the set of frequencies for which the values of S_t are known, *i.e.* the set of frequencies in S_t^o. $\mathcal{F}_t^m$ will similarly refer to the set of frequencies for which the values of S_t are missing, *i.e.* the set of frequencies in S_t^m.

5.1. The Conditional Distribution of Missing Terms

In the first step we obtain the conditional probability distribution of the missing terms $S_t^m(f)$ given the observed terms S_t^o and the probability distribution $P_t(f)$ from which S^t was drawn. Let $N_t^o = \sum_{f \in \mathcal{F}_t^o} S_t^o(f)$. N_t^o is the total value of all observed spectral frequencies at time t. Let $P_{o,t} = \sum_{f \in \mathcal{F}_t^o} P_t(f)$ be the total probability of all observed frequencies at t. The probability distribution of S_t^m, given that the frequencies in $\mathcal{F}_t^o$ are known to have been drawn exactly N_t^o times cumulatively is easily shown to be given by the *negative multinomial* distribution [9]:

$$P(S_t^m) = \frac{\Gamma(N_t^o + \sum_{f \in \mathcal{F}_t^m} S_t^m(f))}{\Gamma(N_t^o)\prod_{f \in \mathcal{F}_t^m} \Gamma(S_t^m(f)+1)} P_{o,t}^{N_t^o} \prod_{f \in \mathcal{F}_t^m} P_t(f)^{S_t^m(f)} \tag{6}$$

where $\mathcal{F}_t^m$ is the set of all frequency components in S_t^m. The expected value of any term $S_t^m(f)$ whose probability is specified by Equation 6 is given by[2]:

$$E[S_t^m(f)] = N_t^o \frac{P_t(f)}{P_{o,t}} \tag{7}$$

We now describe the actual learning procedures to estimate model parameters and missing spectral components. We identify two situations, stated in order of increasing complexity as (a) where the multinomial bases for the data $P(f|z)$ are known *a priori* and only the mixture weights $P_t(z)$ are unknown, and (b) where none of the model parameters are known. We address them in reverse order below to simplify the presentation.

5.2. Learning the Model Parameters from Incomplete Data

Let Λ be the set of all parameters $P_t(z)$ and $P(f|z)$ of the model defined by Equation 2. We derive a set of likelihood-maximizing rules to estimate Λ from $\mathbf{S}^o$ using the Expectation Maximization algorithm as follows.

We denote the set of draws that resulted in the the generation of $\mathbf{S}$ as $\mathbf{z}$. The *complete data* specification required by EM is thus given by $(\mathbf{S}^o, \mathbf{S}^m, \mathbf{z}) = (\mathbf{S}, \mathbf{z})$, where $\mathbf{S}^m$ and $\mathbf{z}$ are unseen. The EM algorithm iteratively estimates the values of Λ that maximizes the expected value of the log likelihood of the complete data with respect to the unseen variables [10], i.e. it optimizes:

$$\begin{aligned} Q(\Lambda, \hat{\Lambda}) &= E_{\mathbf{S}^m, \mathbf{z}|\mathbf{S}^o, \hat{\Lambda}} logP(\mathbf{S}^o, \mathbf{S}^m, \mathbf{z}|\Lambda) \\ &= E_{\mathbf{S}^m|\mathbf{S}^o, \hat{\Lambda}} E_{\mathbf{z}|\mathbf{S}, \hat{\Lambda}} logP(\mathbf{S}^o, \mathbf{S}^m, \mathbf{z}|\Lambda) \end{aligned} \tag{8}$$

[2] To be precise Equations 6 and 7 must actually be specified in terms of $\mathcal{C}^{-1} N_t^o + 1$; however, given the assumption in Equation 1, Equation 7, which is the primary equation of interest remains valid.

where $\hat{\Lambda}$ is the current estimate of Λ. Since the draws that compose any spectrum S_t are independent of those that compose any other $S_{t'}$, Equation 8 simplifies to:.

$$Q(\Lambda, \hat{\Lambda}) = \sum_t E_{S_t^m|S_t^o,\hat{\Lambda}} E_{z_t|S_t,\hat{\Lambda}} logP(S_t^o S_t^m, z_t|\Lambda) \quad (9)$$

where z_t is the set of draws that composed S_t. Optimizing Equation 9 with respect to Λ, and invoking Equation 7 leads us to the following update rules:

$$P_t(z|f) = \frac{P_t(z)P(f|z)}{\sum_{z'} P_t(z')P(f|z')} \quad (10)$$

$$N_t = \sum_{f \in \mathcal{F}_t^o} \frac{S_t^o(f)}{P_t(f)} \quad (11)$$

$$\begin{aligned} \bar{S}_t(f) &= S_t(f) \text{ if } f \in \mathcal{F}_t^o \\ & N_t P_t(f) \text{ if } f \in \mathcal{F}_t^m \end{aligned} \quad (12)$$

$$P_t(z) = \frac{\sum_f P_t(z|f)\bar{S}_t(f)}{\sum_{z'} \sum_f P_t(z'|f)\bar{S}_t(f)} \quad (13)$$

$$P(f|z) = \frac{\sum_t P_t(z|f)\bar{S}_t(f)}{\sum_{f'} \sum_t P_t(z|f')\bar{S}_t(f)} \quad (14)$$

Note that $\bar{S}_t(f)$ are also the minimum mean-squared estimates of the terms in $\mathbf{S}^m$. The above update rules thus also implicitly impute the missing values of the data.

In some situations the multinomial bases $P(f|z)$ may be available, for instance when they have been learned separately from regions of the data that have no missing time-frequency components. In such situations, only the mixture weights $P_t(z)$ need to be learned for any incomplete spectral vector S_t in order to estimate $\bar{S}_t(f)$. This can be achieved simply by iterations of Equations 10, 11, 12 and 13.

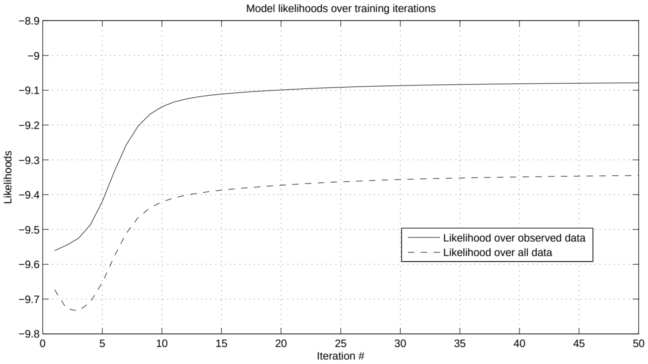

Fig. 2. Likelihood of observed data $\mathbf{S}^o$ (solid line) and complete spectrogram $\mathbf{S}$ (dashed line) as a function of iteration for the data in Figure 6.

The likelihood of the model for the observed data is guaranteed to converge monotonically, and that has been validated with multiple experiments. The model likelihood over both the observed and the imputed data is got guaranteed to increase all the time however it always converges, and does that monotonically so after the first few iterations once the imputation becomes increasingly plausible. Figure 2 shows the convergence patters for the experiment in the next section.

6. EXPERIMENTAL EVALUATION

In this section we will evaluate the algorithms of Section 5 on several examples of spectrograms with missing time-frequency regions, showing both their convergence and effectiveness at imputation. In our examples the spectrograms are from complex musical recordings with multiple, additive concurrent spectral patterns. Such data are particularly difficult to impute using conventional algorithms.

6.1. Illustrative example

We first evaluate our proposed approach with an illustrative example. The input data in this case consisted of a synthetic piano recording of some isolated notes and subsequently of a mixture of these notes. We removed a triangular time-frequency section of the part where the multiple notes took place. We trained our model using the isolated note sections and extracted 60 multinomial bases which we then used to impute the missing values. Figure 3 shows both the original spectrogram (with the missing region marked by the dotted line) and the reconstructed spectrogram. As a comparator we also show the complete spectrograms obtained by imputation of the missing regions using SVD and K-nearest neighbors, two algorithms which are two commonly used models that often perform very well in imputation problems [3, 11]. SVD in particular also models the data additively and may be expected to capture additive patterns.

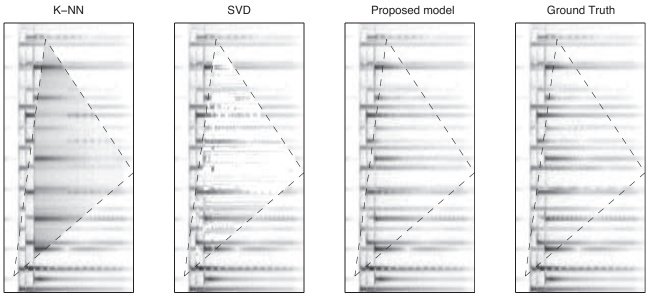

Fig. 3. Comparison of three data imputation approaches on a simple problem. The missing data area is denoted by the dashed black line in all plots. The first plot from the left shows the results from K-NN imputation, the second from SVD imputation, the third using our proposed model and the fourth is the ground truth.

We note that even in this simple problem the K-NN approach does a poor job of modeling all three notes and instead averages out the imputed data thus creating a visibly and audibly incorrect reconstruction. The SVD imputation is more successful, appropriately borrowing elements to reconstruct the coinciding notes. However this approach cannot guarantee that the the imputed output will be non-negative (as required for magnitude spectra) and can potentially return negative values which must be either set to zero, or be rectified, resulting in musical noise. The proposed method properly layers elements of multiple notes to impute the missing data and does not suffer from producing negative values. The result is almost indistinguishable from the ground truth.

Although our approach is not as fast as the k-nearest neighbors approach, it is an order of magnitude faster than the SVD approach since each iteration involves the evaluation of a small number of inner products as opposed to the computation of an entire SVD. The computation times for the above example were 0.7 seconds for k-nearest neighbors, 90 seconds for the SVD and 8 seconds for our proposed approach. Simulations were run on an ordinary laptop computer.

6.2. Real-world examples

In this section we will consider some complex cases derived out of real-world recordings with challenging missing data cases. We will first examine a more complex case of the above example where we attempt to fill a large continuous gap in the spectrogram for a complex musical piece. The section with the gap was a five second real piano recording of Bach's three-part piano invention #3 in D - BWV 879 [12]. The size of the gap was 4.3 seconds by 3 kHz at its widest extent. We also used 10 seconds of data from another piano piece (two-part invention #3 in D -BWV 774) which provided the needed information to impute the missing data. We extracted 60 multinomial bases while training on both the complete and the incomplete data. As a comparator, we also reconstructed the spectrogram using a rank-60 SVD. The sampling rate was 14700Hz and the spectrum size was 1024 points The imputation results are shown in figure 4. One can find some visual inconsistencies using the SVD most noticeably in the rougher texture of the reconstructed area. In contrast our approach results into visually more plausible results. We also invert the spectrograms back to the time domain in order to perform a listening evaluation. In order to accurately compare the artifacts caused by the imputation we used the phase values from the original signal. The proposed method resulted in virtually inaudible reconstruction artifacts whereas the SVD approach introduced audible noise[3].

In the next example we attempt to perform *bandwidth expansion*. Audio signals can often be bandlimited by having been passed through a restrictive channel, such as a telephone. The spectral excision here is very systematic and consistent. We learn a set of 120 multinomial bases from other wideband training data, and use these to infer the missing spectral content of the bandlimited signal. This presumes that the training data is similar in nature to the testing data (i.e. if we need to resample piano music we should train on piano music). An example of this is shown in figure 5. We removed 80% of the upper frequencies of a ten second rock recording of the song ”Back in Black” by the band AC/DC and we trained on an eight second recording by the same band playing a different song (”Highway to Hell”). The sampling rate was once again 14700Hz and the spectrum size was 1024 points. As a comparator we have also reconstructed the spectrogram using rank-120 SVD. The SVD clearly underperforms in both the audible and the visual reconstruction in this experiment, whereas our proposed method results in a plausible, although as expected non-exact, reconstruction.

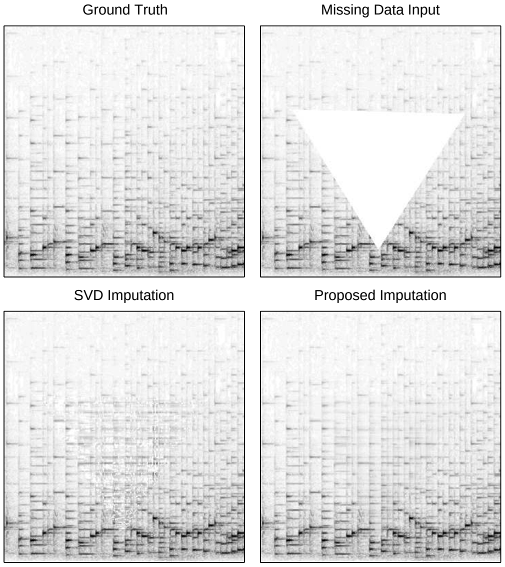

Fig. 4. Example reconstruction of a gap filling experiment. The leftmost plot shows the actual data, the second plot shows the input with the large gap removing about 15% of the data, we have zoomed into the region of the gap by not plotting some of the higher frequency content. the third plot is the SVD reconstruction and the fourth plot is our proposed method.

Finally we present a very challenging case where the missing data was evenly and randomly distributed across the input. For this test a smoothed random binary mask was applied to an input sound so that about 60% of the data was removed. The input sound was sampled as 22050Hz and the time-frequency values were obtained using a 1024 point short-time Fourier transform. We modeled the data with a mixture of 60 multinomials. As a comparator, a rank-60 SVD was also used to reconstruct the spectrogram. The results of

[3] The soundfiles for all the examples in the remainder of this section are attached in the original PDF file of this paper. They can be played using any PDF reading software that allows the viewing of PDF attachments.

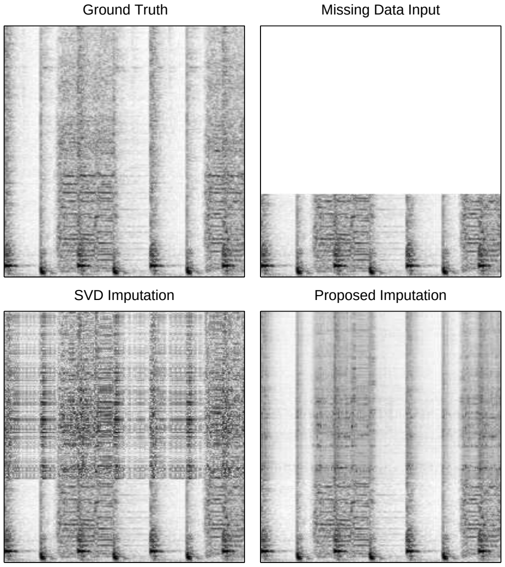

Fig. 5. Example reconstruction of a bandwidth expansion. In the leftmost plot the original signal is shown. The second plot displays the bandlimited input we used where 80% of the top frequencies were removed. The third plot is the SVD reconstruction and the fourth plot is the reconstruction using our model.

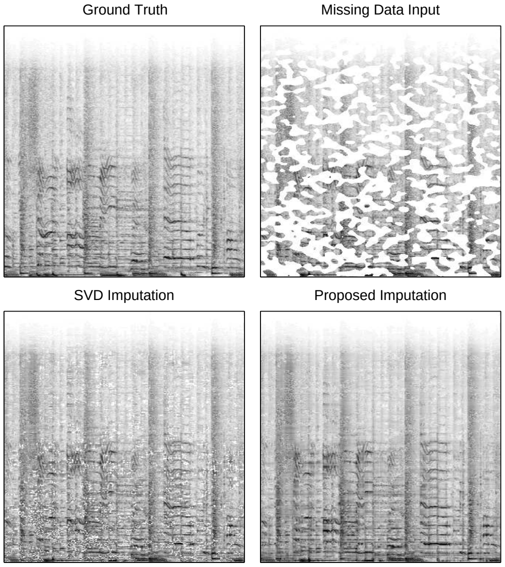

Fig. 6. Example reconstruction of a music signal with a binary mask occluding roughly 60% of the samples. The leftmost plot shows the original signal, the second plot shows the masked input we used for the reconstruction, the third plot shows the reconstruction using the SVD and the fourth one shows the reconstruction using our model.

this experiment are shown in figure 6.

We note that the SVD model results in grainier reconstruction in the missing parts, whereas the proposed model results in a smoother output. Our proposed model results in reconstructed audio signals of significantly improved quality. The processing artifacts are virtually imperceptible unless the sound is compared to the original input.

7. REFERENCES

[1] B. Raj, Reconstruction of Incomplete Spectrograms for Robust Speech Recognition, Ph.D. Dissertation, Carnegie Mellon University, May 2000.

[2] Sam T. Roweis. One Microphone Source Separation. NIPS 2000: 793-799.

[3] M. E. Brand. Incremental Singular Value Decomposition of Uncertain Data with Missing Values, European Conference on Computer Vision (ECCV), Vol 2350, pps 707-720, May 2002.

[4] M.J. Reyes-Gomez,N. Jojic and D.P.W. Ellis. Detailed graphical models for source separation and missing data interpolation in audio. 2004 Snowbird Learning Workshop Snowbird, Utah, March 2004

[5] M. Shashanka, B. Raj, P. Smaragdis. Sparse Overcomplete Latent Variable Decomposition of Counts Data. NIPS 2000.

[6] Smith, J. O. Spectral Audio Signal Processing, March 2007 Draft, `http://ccrma.stanford.edu/~jos/sasp/`, online book, accessed June 2008.

[7] J. Bouvrie and T. Ezzat. ”An Incremental Algorithm for Signal Reconstruction from Short-Time Fourier Transform Magnitude”, in Proc. Ninth International Conference on Spoken Language Processing (Interspeech), Pittsburgh, 2006.

[8] T. Hofmann and J. Puzicha. Unsupervised learning from dyadic data. TR 98-042, ICSI, Berkeley, CA, 1998.

[9] M. Hazewinkel. Encyclopedia of Mathematics. http://eom.springer.de/

[10] A. P. Dempster, N. M. Laird, D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, B, 39, 1-38. 1977.

[11] Hastie, T., Tibshirani, R., Sherlock, G., Eisen, M., Brown, P. and Botstein, D. Imputing Missing Data for Gene Expression Arrays. Technical report (1999), Stanford Statistics Department.

[12] Gould, G. Bach: The Two and Three Part Inventions - The Glenn Gould Edition, by SONY classics, ASIN B000GF2YZ8, 1994.