# SUPER-RESOLUTION OF MUSICAL SIGNALS USING APPROXIMATE MATCHING PURSUIT

*Brennan P. Keegan, Steven K. Tjoa, and K. J. Ray Liu*

Signals and Information Group, Department of Electrical and Computer Engineering
University of Maryland, College Park – College Park, MD 20742 USA
{bpkeegan, kiemyang, kjrliu}@umd.edu

## ABSTRACT

Super-resolution (SR) is a well-studied problem in signal processing, particularly with regard to image and video applications. SR techniques are useful because unlike simple interpolation, they create a high-resolution signal from a low-resolution input by generating new information that was not previously present. A growing body of research shows progress in development of SR techniques using dictionary learning. We propose a method for SR of musical signals through matrix factorization and use of a known musical dictionary. The approximate matching pursuit (AMP) algorithm is used to query the dictionary and perform factorization, making the overall process efficient and scalable. By approximating the frequency information of a low-resolution input spectrogram as the linear combination of entries in the musical dictionary, we are able to closely match the input signal's missing high frequency information and thereby create a high-resolution output signal.

***Index Terms***— Nonnegative matrix factorization, sparse coding, signal restoration and enhancement.

## 1. INTRODUCTION

Super-resolution (SR) is the problem of creating a high-resolution (HR) output signal from a low-resolution (LR) input. In order to achieve this goal, SR techniques often make use of certain assumptions or outside knowledge about the original signal to acheive their goal. Within audio processing, SR is often useful for overcoming limitations related to encoding, such as low bandwidth. For example, in cell phone and other speech transmission applications, encoding often needs to be performed at a low bitrate to ensure the signal can be sent within the limitations created by a small bandwidth. In these situations, the audio may be recorded with a sampling rate as low as 4 kHz, which creates a very low-quality recording, with much of the high-frequency information missing from the signal.

In the field of music, SR has a variety of practical benefits. In the case of old musical recordings with a low sampling rate, SR techniques can help reproduce a recording more true to what was heard when the piece was originally performed. Just as record companies have expended much time and effort in remastering famous recordings to reproduce original pieces more exactly, SR methods would be able to provide the listener with a higher quality listening experience. SR techniques are also useful in situations for which recording quality is limited by hardware or computational power, again allowing listeners an improved listening experience.

In audio signal processing, little work has been done looking into the specific problem of SR for music, but research has examined the more general idea of reconstructing portions of audio from signals with incomplete or damaged regions in the time-frequency domain. Specifically, [1], [2], and [3] propose methods of filling in missing time-frequency information, while [4] examines the problem of reconstructing damaged time-frequency features. Similar to our proposed method, [5] suggests a means of performing missing data imputation through factorization of non-negative data. Their method is much more general than ours, however, as we choose to focus solely on factorization of musical audio signals.

In the fields of image and video signal processing, considerable success has been found in using dictionaries of spectral events to perform SR. These methods generally rely on first learning a dictionary of spectral "patches" from a set of sample images. The patches in the dictionary exist in pairs, such that for each LR patch in the dictionary, a corresponding HR patch also exists. Next, for a given section LR input image, LR patches of the dictionary are overlapped with different weighting factors applied such that their linear combination closely approximates the section of the image. Finally, these LR patches are replaced by their HR counterparts in the dictionary. With the same linear combination applied to the HR patches, a good approximation of the HR equivalent can be generated, thus achieving SR. This approach is suggested in [6] for image processing. In video processing, as in the case of [7] and [8], we see very similar methods, often centering on dictionary learning based on select HR keyframes. Finally, in [9] and [10] we see examples of mixed methods which take advantage of cameras that are able to take HR still images while recording continuous video.

Our proposed method parallels the above-mentioned techniques in many ways, though instead of learning a dictionary from sample media, we assume a reliable musical dictionary is already known. In order to query our dictionary and find coeffients to appropriately weight individual dictionary entries and recreate the LR input, a variation of the matching pursuit algorithm is used. This algorithm, approximate matching pursuit (AMP), makes the process of searching the dictionary fast, efficient, and scalable – desirable traits for producing practical music applications. Overall, we want to help bridge the gap in SR for audio by showing how dictionary-based methods, which have seen much success in other fields of signal processing, can be useful in musical applications as well.

## 2. PROBLEM FORMULATION

In practice, audio SR involves input and output signals in the time domain. Our proposed method deals exclusively with processing a LR input spectrogram matrix to generate a HR output spectrogram. Therefore, for the purposes of this study we focus our attention on the central processing portion of the SR system, disregarding the time-domain components at both ends of the system.

978-1-4577-0693-6/11/$26.00 ©2011 IEEE

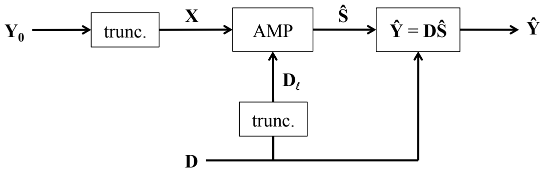

Figure 1: System block diagram.

The input matrix is the LR spectrogram $\mathbf{X} \in \mathbb{R}^{M_{\text{low}} \times N}$, formed by truncating the ground truth HR spectrogram $\mathbf{Y}_0 \in \mathbb{R}^{M_{\text{high}} \times N}$ in frequency, by removing the top $M_{\text{high}} - M_{\text{low}}$ frequency bins. The known, high-resolution dictionary matrix is $\mathbf{D} \in \mathbb{R}^{M_{\text{high}} \times K}$. By truncating the columns of matrix $\mathbf{D}$ in frequency, we create a low-resolution dictionary matrix $\mathbf{D}_\ell \in \mathbb{R}^{M_{\text{low}} \times K}$. The central problem is to generate a sparse matrix of coefficients, $\hat{\mathbf{S}} \in \mathbb{R}^{K \times N}$, such that we minimize the difference between the input $\mathbf{X}$, and $\mathbf{D}_\ell \hat{\mathbf{S}}$, i.e.:

$$\min_{\hat{\mathbf{S}}} ||\mathbf{X} - \mathbf{D}_\ell \hat{\mathbf{S}}||_F, \tag{1}$$

where $||\cdot||_F$ denotes the Frobenius norm. In this expression, matrix $\hat{\mathbf{S}}$ contains coefficients that represent the relative weighting of the dictionary atoms in $\mathbf{D}_\ell$ necessary to approximate $\mathbf{X}$.

One key assumption made in this study is that for musical spectra, there is high correlation between low frequency and high frequency information. In a practical sense, this means that the middle C played on any piano shares great similarity with the middle C of other pianos not only in its lower frequencies, but in its high frequencies as well. If this were not the case, it would be impossible for a representative dictionary of piano spectra to correctly fill in the high frequency information in a piano recording, as our method proposes to do. From a technical perspective, this assumption means that, given the HR ground truth spectrogram $\mathbf{Y}_0$, we assume that finding $\hat{\mathbf{S}}$ such that Eq. (1) is satisfied will also satisfy the expression

$$\min_{\hat{\mathbf{S}}} ||\mathbf{Y}_0 - \mathbf{D}\hat{\mathbf{S}}||_F. \tag{2}$$

## 3. PROPOSED METHOD

Our system model, shown in Figure 1, begins with the HR ground truth spectrogram $\mathbf{Y}_0$. $\mathbf{Y}_0$ is truncated in frequency to form the LR input spectrogram $\mathbf{X}$. Similarly, the known HR dictionary $\mathbf{D}$ is truncated in frequency to generate the LR dictionary $\mathbf{D}_\ell$. Our proposed matching pursuit algorithm, AMP, then takes in $\mathbf{X}$ and $\mathbf{D}_\ell$ and returns the sparse coefficient matrix $\hat{\mathbf{S}}$, which is constructed to minimize $||\mathbf{X} - \mathbf{D}_\ell \hat{\mathbf{S}}||_F$. On a smaller scale, AMP takes in $\mathbf{x}_i$, the $i^{th}$ column of $\mathbf{X}$, and generates a sparse column matrix $\hat{\mathbf{s}}_i$ such that $||\mathbf{x}_i - \mathbf{D}\hat{\mathbf{s}}_i||_2$ is minimized. $\hat{\mathbf{s}}_i$ then becomes the $i^{th}$ column in coefficient matrix $\hat{\mathbf{S}}$. Lastly, the HR dictionary $\mathbf{D}$ and the output coefficient matrix $\hat{\mathbf{S}}$ are multiplied to create the HR output spectrogram $\hat{\mathbf{Y}} = \mathbf{D}\hat{\mathbf{S}}$. To summarize the system model:

1. Input: Ground truth HR spectrogram $\mathbf{Y}_0$, HR dictionary $\mathbf{D}$
2. $\mathbf{Y}_0$ and $\mathbf{D}$ truncated in frequency to form LR spectrogram $\mathbf{X}$ and LR dictionary $\mathbf{D}_\ell$
3. AMP algorithm uses $\mathbf{X}$ and $\mathbf{D}_\ell$ to generate sparse coefficient matrix $\hat{\mathbf{S}}$
   (a) For each column $\mathbf{x}_i$ in $\mathbf{X}$, finds column matrix $\hat{\mathbf{s}}_i$ such that $\mathbf{x}_i - \mathbf{D}_\ell \hat{\mathbf{s}}_i$ is minimized.
   (b) The $N$ generated column matrices $\hat{\mathbf{s}}_i$ are joined to form $\hat{\mathbf{S}}$, $\mathbf{X}$ is approximated by $\mathbf{D}_\ell \hat{\mathbf{S}}$.
4. HR spectrogram $\hat{\mathbf{Y}}$ generated using the HR dictionary and coefficient matrix, $\hat{\mathbf{Y}} = \mathbf{D}\hat{\mathbf{S}}$.
5. Output: HR spectrogram $\hat{\mathbf{Y}}$, sparse coefficient matrix $\hat{\mathbf{S}}$

We can intuitively see this output spectrogram should correspond to an audio signal very close to that of the original input that formed $\mathbf{X}$, since $\hat{\mathbf{S}}$ has been chosen to minimize the distance between $\mathbf{X}$ and $\mathbf{D}_\ell \hat{\mathbf{S}}$. Assuming the entries in $\mathbf{D}$ contain good high-resolution representations of the musical spectra originally used to produce $\mathbf{X}$, we hypothesize $\hat{\mathbf{Y}}$ will effectively create a high-resolution spectrogram very similar to that which would be produced if $\mathbf{X}$ had originally been recorded at a higher sampling rate.

## 4. EXPERIMENTS

For each each stage of experimentation, we vary certain system parameters and examine how well the proposed method is able to produce a HR output spectrogram $\hat{\mathbf{Y}}$ that accurately represents the ground truth HR spectrogram $\mathbf{Y}_0$. The metric we use to evaluate system performance is the expression:

$$F = ||\mathbf{Y}_0 - \hat{\mathbf{Y}}||_F \tag{3}$$

This difference metric, which we will refer to simply as $F$ from this point on, provides a quantifiable assessment of how accurate $\hat{\mathbf{Y}}$ is as a reconstruction of $\mathbf{Y}_0$.

The experiments are divided into two stages: testing using a synthetically generated input and real musical dictionary, and testing using both real musical signals for the input and dictionary. For both stages, we examine $F$ with respect to controlled change in two parameters: the number of entries in the dictionary, $K$, and the number of low frequency bins $M_{\text{low}}$, for some constant $M_{\text{high}}$. We chose these parameters because the dictionary size and sampling rate of the input signal are likely to vary in application. By examining changes in our method's performance as we vary $K$ and $M_{\text{low}}$, we are able to make conjectures as to the advantages and limitations of the system.

In order to generate the real musical dictionary, we used piano spectra found in the Iowa State Musical Recordings. This is a catagorized database which contains high-quality, single-pitch recordings of every pitch in an instrument's regular range, for a variety of common classical instruments. For our experiments, we used the database's 259 recordings of piano, which consisted of a short audio recording for nearly every note on the piano, at three different dynamic levels (fortissimo, metzo forte, and pianissimo). The recordings of different dynamic levels are important in ensuring our dictionary contains a full representation of a piano's timbre.

To generate frequency spectra from these audio recordings, we sectioned each into smaller time frames and took the Fourier transform of this frame. Generated spectra with very low energy were disregarded and all others were returned as viable spectra for a specific piano note of a given dynamic level. We then generated our testing dictionary by randomly selecting a set number of these viable spectra for each note. For example, for a given iteration of

| $K$ | 1036 | 3108 | 5180 | 7252 | 9324 |
|---|---|---|---|---|---|
| $F$ | 1.919 | 1.877 | 1.809 | 1.812 | 1.798 |

Table 1: Reconstruction error, $F$, as a function of dictionary size, $K$.

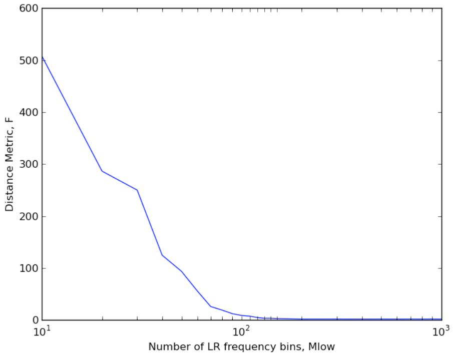

Figure 2: Distance Metric, $F$ vs. Number of LR Frequency Bins, $M_{\text{low}}$

testing we could randomly select 4 viable spectra from each of the 259 files in the database, yielding a dictionary of 1036 spectra.

### 4.1. Testing with Synthetic Input, Real Musical Dictionary

First, we use a dictionary matrix of real musical spectra. The input spectrogram $\mathbf{Y}_0$ is generated from the dictionary $\mathbf{D}$ and sparsely generated matrix of coefficients $\mathbf{S}_0$, such that 5 dictionary atoms are active for any frame in time. This method of input generation means that increasing our dictionary size $K$ will not likely have a significant effect on the distance metric $F$, since whether we have thousands of dictionary entries or just ten, we know there enough entries to closely approximate $\mathbf{Y}_0$. Our data in Table 1 confirms this expectation, with little change in $F$ as we increase the dictionary size $K$ from 1036 (4 spectra per file) to 9324 (36 spectra per file).

For the test varying $M_{\text{low}}$, we expect a sharp drop in $F$ for small values of $M_{\text{low}}$, with $F$ leveling out as $M_{\text{low}}$ becomes larger. This expectation comes from the idea that relative to the full 44.1 kHz, most of the energy in piano or other musical tones exists in relatively low frequencies, with the highest note on a piano having a fundamental frequency of approximately 4 kHz. The AMP algorithm as used by our method looks for the largest spectral peaks, so as long as the high energy portions of $\mathbf{Y}_0$ are included in $\mathbf{X}$, we should be able to closely approximate $\mathbf{Y}_0$ from $\mathbf{X}$. As $M_{\text{low}}$ increases over very small values, each step up will significantly increase the amount of high energy information from $\mathbf{Y}_0$ that will be present in $\mathbf{X}$. The increase in high energy information will allow us to more accurately recreate $\mathbf{Y}_0$ using $\mathbf{X}$, resulting in this sharp drop in $F$ for small values of $M_{\text{low}}$. After $M_{\text{low}}$ reaches a certain threshold, however, all the major high energy information from $\mathbf{Y}_0$ will be contained in $\mathbf{X}$. Therefore, further increases in $M_{\text{low}}$ will not substantially change how well we can approximate $\mathbf{Y}_0$, and $F$ will level out to become approximately constant. This general trend is reflected by our data, shown in Figure 2.

The above reasoning and results signify two important characteristics of our proposed method. First, we see that $M_{low}$ need not be very large in order for the system to perform SR. In a practical sense, this means that we can theoretically produce a high quality output signal even if the original input was recorded at a very low sampling rate. The one major requirement is that the original sampling rate needs to be high enough to include the frequencies in the original analog signal that have the most energy. Second, we see that after a certain threshold, improving the quality of the input signal does little to improve the resulting HR output. For example, if the same analog signal was sampled at 8 and 16 kHz, our proposed method would likely return very similar HR outputs for the two input signals, despite the 16 kHz input containing significantly more frequency information about the original signal.

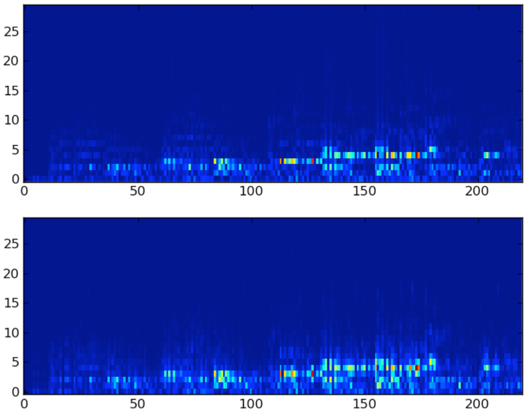

Figure 3: Time vs Frequency plots of spectrograms $\mathbf{X}$ (top) and $\mathbf{D}\hat{\mathbf{S}}$. (bottom), taking the first five measures of Mozart's "Rondo Alla Turca" as the input signal.

### 4.2. Testing with Real Musical Input and Dictionary

In the second stage of experimentation, we examine the performance of our proposed method using a real musical input signal to generate the input spectrogram. To visually demostrate our method's ability to reproduce a real musical spectrogram, we look to the example in Figure 3, which shows the LR input spectrogram $\mathbf{X}$ and our output $\mathbf{D}\hat{\mathbf{S}}$. For this example, we are using the first five measures of the famous piano piece, "Rondo Alla Turca," from Mozart's Sonata No. 11 in A major.

For testing, our input audio file is the first ten seconds of Prelude No. 1 in C major, from Bach's the Well Tempered Clavier, Book 1. Once again we examine change in $F$ as $M_{low}$ and $K$ are adjusted. As in the previous section of experiments, as we vary $M_{low}$ we expect $F$ to drop sharply and then level out, as $M_{low}$ becomes large. This is confirmed by our results in Figure 4, with a plot showing the same trend as was seen in Figure 2.

Now that we are using a real musical spectrogram as the input, we can no longer guarantee that the entries in the dictionary are sufficient to reproduce the original signal. This means that as we increase the size of our dictionary, we expect $F$ to decrease, as $\hat{\mathbf{Y}}$ should more accurately reproduce $\mathbf{Y}_0$. As shown in Figure 5, this expectation of a steadily decreasing $F$ was not confirmed by our data. As $K$ ranges from approximately 1,000 to 11,000, we actually see $F$ level out, rather than continue decreasing. We suspect this is

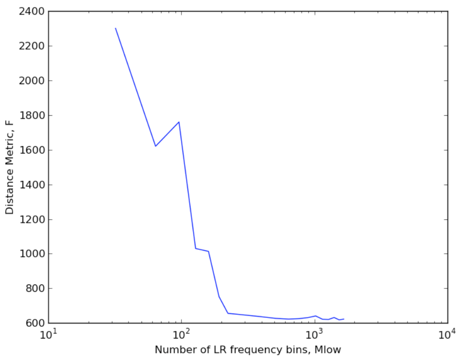

Figure 4: Distance Metric, $F$ vs. Number of LR Frequency Bins, $M_{\text{low}}$

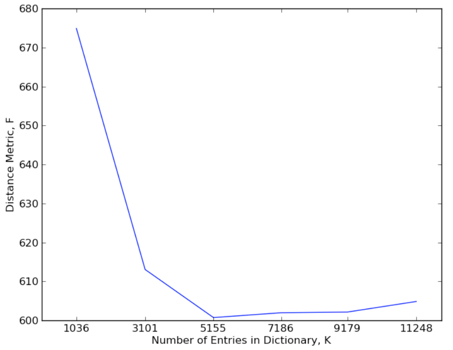

Figure 5: Distance Metric, $F$ vs. Dictionary Size, $K$

related to our method for generating the dictionary, which returns a limited number of viable spectra per file. Spectra are deemed viable if they contain energy above a certain threshold, which allows fortissimo recordings to have many more viable spectra returned than their pianissimo counterparts. For small values of $K$, this has no effect on the dictionary, since there are plenty of viable spectra for each file. As $K$ reaches very large values however, even though the dictionary attempts to return 64 spectra per file, it may only be able to return the 20 or fewer viable spectra for a given pianissimo file. This effect means that subsequent increases in $K$ may not improve the dictionary for certain categories of notes, which could cause $F$ to remain nearly constant, even though the dictionary size $K$ is increasing.

## 5. CONCLUSION

As shown by the results of the last experiments section, SR of real musical signals is possible with the existence of a reliable dictionary of musical spectra. We also saw that this method works well for very LR input; as long as the input signal contains the high energy frequency information of the original analog signal. For input which contains this high energy information, our method performs about the same regardless of input resolution, underscoring the importance of the dictionary in our proposed system.

While our focus was solely in reference to musical application, the proposed method of using a pursuit algorithm and known dictionary to perform SR could theoretically be performed in a wide variety of contexts, as long as there is high correlation between low and high frequency information of the dictionary entries. A logical follow-on to this work would be to take our approach and apply it to image and video processing as well.

For future study, we would like to further examine how our method performs under conditions in which instruments of many different timbre and range are active simultaneously. The results of this study were based on a dictionary created from just one musical database of less than 1 million spectral atoms, so a logical next step would be to examine system performance using the combined information of multiple musical databases. Comparisons could also be drawn between individual databases, to show their effectiveness in forming spectral dictionaries for SR.

Finally, this work examines only a small part of the overarching SR system. We recommend further study with a more holistic approach, looking at performance of the system when the time domain aspects are also included. This would provide a clearer picture of how our method operates in practice, and allow opportunity for evaluation of the true output of the system, the HR audio signal.

## 6. REFERENCES

[1] P. Smaragdis, B. Raj, and M. Shashanka, "Missing data imputation for spectral audio signals," in *Proc. IEEE Workshop Machine Learn. Signal Process.(MLSP)*. Citeseer, 2009.

[2] B. Raj, M. Seltzer, and R. Stern, "Reconstruction of missing features for robust speech recognition," *Speech Communication*, vol. 43, no. 4, pp. 275–296, 2004.

[3] J. Le Roux, H. Kameoka, N. Ono, A. de Cheveigné, and S. Sagayama, "Computational auditory induction by missing-data non-negative matrix factorization," *Proc. SAPA*, 2008.

[4] B. Raj, M. Seltzer, and R. Stern, "Reconstruction of damaged spectrographic features for robust speech recognition," in *Proc. ICSLP*. Citeseer, 2000.

[5] P. Smaragdis, M. Shashanka, B. Raj, and G. Mysore, "Probabilistic factorization of non-negative data with entropic co-occurrence constraints," *Independent Component Analysis and Signal Separation*, pp. 330–337, 2009.

[6] J. Yang, J. Wright, T. Huang, and Y. Ma, "Image super-resolution via sparse representation," *Image Processing, IEEE Transactions on*, vol. 19, no. 11, pp. 2861–2873, 2010.

[7] C. Bishop, A. Blake, and B. Marthi, "Super-resolution enhancement of video," in *Proc. Artificial Intelligence and Statistics*, vol. 2. Citeseer, 2003.

[8] B. Song, S. Jeong, and Y. Choi, "Video super-resolution algorithm using bi-directional overlapped block motion compensation and on-the-fly dictionary training," *Circuits and Systems for Video Technology, IEEE Transactions on*, no. 99, pp. 1–1, 2011.

[9] D. Kong, M. Han, W. Xu, H. Tao, and Y. Gong, "A conditional random field model for video super-resolution," *Pattern Recognition*, vol. 3, pp. 619–622, 2006.

[10] A. Gupta, P. Bhat, M. Dontcheva, O. Deussen, B. Curless, and M. Cohen, "Enhancing and experiencing spacetime resolution with videos and stills," in *Computational Photography (ICCP), 2009 IEEE International Conference on*. IEEE, 2009, pp. 1–9.