EXAMPLE-DRIVEN BANDWIDTH EXPANSION

Paris Smaragdis

Mitsubishi Electric Research Laboratories
201 Broadway, Cambridge, MA 02139
paris@merl.com

Bhiksha Raj

Mitsubishi Electric Research Laboratories
201 Broadway, Cambridge, MA 02139
bhiksha@merl.com

ABSTRACT

In this paper we present an example-driven algorithm that allows the recovery of wide regions of lost spectral components in bandlimited signals. We present a generative spectral model which allows the extraction of salient information from audio snippets, and then apply this information to enhance the bandwidth of bandlimited signals.

1. INTRODUCTION

With the advent of network-enabled audio systems we have also entered a new era of poor audio quality! We now frequently experience devices like cell phones, or chat programs that due to networking limitations transfer audio with suboptimal bandwidths. Although undoubtedly this limitation will be resolved in time there is currently a desire for audio enhancing post-processing. In this paper we will present a novel way of bandwidth expansion and contrast its abilities with known and obvious approaches.

The paper is structured as follows. We first present the problem at hand and then proceed to explain some methods that have been used to resolve it in the past. We then present the computational core that we will use and how to apply it in this particular problem. We conclude by presenting results and performance comparisons and discussion some of the elements regarding future work in this subject.

2. BANDWIDTH EXPANSION

2.1. Problem definition

Audio signals such as music are best appreciated in full bandwidth. A low frequency response and the presence of high frequencies are universally understood to be elements of high quality audio. Quite often though this wide frequency content is not available. We often find signals sampled at a low sampling rate (thereby losing high frequency information), or signals that undergo some processing or distortion which removes certain frequency regions. The goal of bandwidth expansion is to recover the missing frequency band information. Although is isn't the only case, the most common one is that of recovering missing high frequencies (usually when we attempt resampling from a low sample rate to a higher one). As one might expect coming up with the missing frequency data is not straightforward. This is information which is lost and cannot be inferred.

The problem of bandwidth expansion has hitherto been considered chiefly in the context of speech signals. Telephone-bandwidth speech signals typically only contain frequency components between 300Hz and about 3500Hz (the exact frequencies vary for landlines and cellphones, but remain below 4kHz in all cases). The aim of bandwidth expansion has been to somehow fill in frequency components, both below the lower cutoff and above the upper cutoff, in order to deliver a fuller-bodied sounding signal to the listener. The goal has been primarily that of enriching the perceptual quality of the signal, and not so much high-fidelity reconstruction of the missing frequency bands. We briefly mention some techniques that are commonly employed below.

2.2. Data agnostic methods

The simplest methods for expanding the spectrum of a signal involve applying a memoryless non-linearity which has some desirable side effects in the frequency domain. For example consider a non-linearity such as:

$$y(t) = x(t) + \alpha x^{\beta}(t) \tag{1}$$

where $x(t)$ is the bandlimited input and $y(t)$ the result of the bandwidth expansion. This operation's result in the frequency domain is:

$$Y(\omega) = X(\omega) + \alpha(X * X * \cdots * X)(\omega) \tag{2}$$

The β convolutions of $X(\omega)$ have the effect of extending the bandwidth of $x(t)$ by a factor of β. Other non-linearities such as applying a sigmoid or a rectifier on $x(t)$ will have the effect of expanding the bandwidth infinitely [1].

Because of their simplicity these techniques are only appropriate for the simplest cases of bandwidth expansion and are not useful for complex problems where the missing frequencies can come from multiple bands at any position on the frequency spectrum.

2.3. Example-driven methods

The *example-driven* approach attempts to derive unseen frequencies in the signal from their statistical dependencies on the observed frequencies. These dependencies are variously captured through codebooks [2], coupled HMM structures [3], Gaussian mixture models [4], etc, the parameters of which are typically learned from parallel broadband and narrow-band corpora. In order to maintain harmonic structures, the signal itself is typically represented through linear predictive models that, that can be extended into unseen frequencies and excited with the excitation of the original signal itself.

These methods are directed primarily towards speech signals, and have an implicit reliance on the fact that the recorded signal has been generated by a single source and can be expected to exhibit consistency of spectral structures within any analysis frame. For instance, the signal in any frame of voiced speech includes the contributions of the harmonics of a single pitch frequency, which can hence be extended into unseen frequencies. Similarly formant structures evident in spectral envelopes are relatively consistent across multiple instances of the same sound. However, on more complex signals such as music that may contain multiple independent spectral structures from multiple sources, these methods are usually less effective. For instance, frames of music audio often contain multiple possibly independent harmonic structures. Data agnostic methods that blindly extend harmonic structures will not only extend individual harmonics but also introduce spurious spectral structures at the beat frequencies. Further, they are ineffective for broadband sounds such as fricated speech or drums, whose spectral textures at high frequencies different from those at lower ones. Similarly, since the spectral patterns from the multiple sources can co-occur in a nearly unlimited number of ways, example-based methods are either unable to learn all of these combinations from any finite corpus of data, or require models of high complexity to represent all of them.

3. BANDWIDTH EXPANSION USING A LATENT COMPONENT ANALYSIS

In this section we will first introduce a spectral decomposition model which is appropriate for inferring missing spectral data and then we will demonstrate how this model can be used to solve the problem at hand.

3.1. Latent component analysis

The model of latent component analysis can be seen as a multi-state generalization of the magnitude spectrum. Let us assume that we have a time series $x(t)$ with a corresponding time-frequency decomposition $X(\omega, t)$. We do not need to make any strong assumptions on the exact nature of X, but for simplicity we can assume that it is a Short Time Fourier Transform (STFT). The magnitude of that transform $|X(\omega, t)|$ can be interpreted as a two-dimensional probability distribution describing the allocation of frequencies across time. Following that interpretation we can obtain the marginals of this distribution, by averaging across either dimension, which will respectively correspond to the magnitude spectrum and the energy envelope of $x(t)$. We can generalize this idea to the following latent variable model which is a weighted mixture of multiple marginal distributions from each dimension:

$$P(\omega, t) = \sum_z P(z)P(\omega|z)P(t|z) \tag{3}$$

$P(\omega, t)$ can be an arbitrary distribution, but for the purposes of this paper it will be a magnitude STFT. In the simplest case where $z = \{1\}$ then $P(z) = 1$, $P(\omega|z)$ is the magnitude spectrum, and $P(t|z)$ the energy envelope.

Estimation of all the unknowns in equation (3) is straightforward using an Expectation-Maximization approach. During the E-step we estimate:

$$R(\omega, t, z) = \frac{P(z)P(\omega|z)P(t|z)}{\sum_{z'} P(z')P(\omega|z')P(t|z')} \tag{4}$$

and during the M-step we obtain a refined set of estimates:

$$P(z) = \sum_{\forall\omega}\sum_{\forall t} P(\omega, t)R(\omega, t, z) \tag{5}$$

$$P(\omega|z) = \frac{\sum_{\forall t} P(\omega, t)R(\omega, t, z)}{P(z)} \tag{6}$$

$$P(t|z) = \frac{\sum_{\forall\omega} P(\omega, t)R(\omega, t, z)}{P(z)} \tag{7}$$

After iterating between these two steps for a few iterations we can obtain good estimates of all the unknown quantities. It should be noted that this algorithm can be seen as a probabilistic specialization of the well known SVD decomposition in the case where the inputs are 2-D histograms or distributions instead of matrices. It is also numerically equivalent to the Non-negative Matrix Factorization approach.

Now let us examine what these new multiple marginals can represent and where the strength of this algorithm lies when it comes to spectral analysis. Consider the spectrogram in figure 1. It is a spectrogram from a piece of music which includes multiple piano notes performed at the same time. On the left we display the frequency marginals $P(\omega|z)$ we extracted from this input. One can see that the marginals are a set of magnitude spectra that characterize the various harmonic series in that signal. This type of analysis effectively creates a set of additive dictionary elements

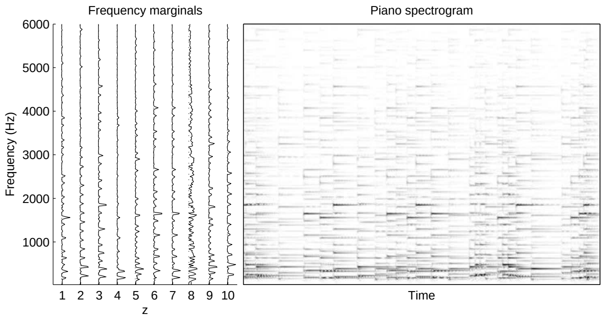

Figure 1: Latent variable frequency marginals extracted from a piano spectrogram.

that can describe the analyzed signal. The time marginals $P(t|z)$ describe how these dictionary elements are to be combined to make up the input, and the priors $P(z)$ provide a measure of the usage distribution of each marginal pair.

3.2. Bandwidth expansion procedure

As demonstrated in the previous section a latent component analysis can be very useful in encapsulating the structure of a complex input. We will now use this property to help us perform bandwidth expansion using an example-based approach. The outline of our process is as follows:

1. Given a signal $x(t)$ with arbitrary missing frequency bands, obtain a high-quality signal $g(t)$ which is spectrally close to the desired result.
2. Compute $|G(\omega, t)|$, a magnitude time-frequency representation of $g(t)$, and estimate from it a set of frequency marginals $P_G(\omega|z)$.
3. Compute $|X(\omega, t)|$, a magnitude time-frequency representation of $x(t)$, and using the already known frequency marginals $P_G(\omega|z)$ try to estimate an appropriate $P_X(z)$ and $P_X(t|z)$. Perform the estimation using only the ω 's where $|X(\omega, t)|$ is significant.
4. Now perform $|\hat{Y}(\omega, t)| = \sum_z P_X(z)P_G(\omega|z)P_X(t|z)$ which will reconstruct $|X(\omega, t)|$ using the high-quality frequency marginals from the high-quality example.
5. Transform $|\hat{Y}(\omega, t)|$ back to the time domain to obtain $\hat{y}(t)$, a high-quality version of $x(t)$ according to $g(t)$.

Now let us examine at these steps in more detail. For a given input $x(t)$ which has missing frequency bands we need to obtain a signal $g(t)$ which will serve as an example of what the output should sound like (in terms of quality). For example in the case of speech it would help to use a high-quality recording of the speaker at hand, in the case of music one should use examples of high-quality recordings of music with similar instrumentation, etc. There is no right or wrong choice of an example sound, however a careful selection will result in better results than otherwise. This is a point which is true for all example-driven bandwidth expansion schemes and due to space limitations we cannot elaborate further, but will stress its importance.

We denote the magnitude STFT of the low and high quality signals by $|X(\omega, t)|$ and $|G(\omega, t)|$ respectively. Using the aforementioned algorithm we perform a latent variable analysis of $|G(\omega, t)|$ and extract a set of frequency marginals $P_G(\omega|z)$. We use a sufficiently large number of states for z (usually around 300 states) to ensure we have an extensive frequency marginal 'dictionary' for this type of recording. $P_G(\omega|z)$ can be seen as a set of spectra that additively compose high-quality recordings of the type expressed in $g(t)$.

We now have to use the known high-quality frequency marginals $P_G(\omega|z)$ to improve the quality of $x(t)$. The assumption is that the unobserved high-quality version of $x(t)$ (let us call this $y(t)$) is composed out of very similar dictionary elements as $g(t)$. That is we can assume that:

$$|Y(\omega, t)| \approx \sum_z P_Y(z)P_G(\omega|z)P_Y(t|z) \tag{8}$$

Under this assumption we can also assume that:

$$|X(\omega, t)| \approx \sum_z P_X(z)P_G(\omega|z)P_X(t|z), \forall\omega \in \Omega \tag{9}$$

Where Ω is the set of available frequency bands of $x(t)$. In the above equation it is easy to estimate $P_X(z)$ and $P_X(t|z)$ using the update equations (5,7) and keeping $P_G(\omega|z)$ fixed to the already known values. Since $P_X(z)$ and $P_X(t|z)$ are not frequency specific we can estimate them using only a small subset of the available frequencies.

Once $P_X(z)$ and $P_X(t|z)$ are estimated we can perform a full-bandwidth reconstruction of our high-quality magnitude spectrogram estimate:

$$|\hat{Y}(\omega, t)| = \sum_z P_X(z)P_G(\omega|z)P_X(t|z) \tag{10}$$

The final step is to obtain the time series $\hat{y}(t)$ from $|\hat{Y}(\omega, t)|$. This can be done in a variety of ways which we have not compared conclusively. The most direct method is that of using the estimated high-quality magnitude spectrum $|\hat{Y}(\omega, t)|$ to modulate the original low-quality phase spectrum $\angle X(\omega, t)$ and performing an inverse STFT. In our experiments this approach has worked adequately well albeit with some minor phase artifacts. A more careful approach is to appropriately manipulate $\angle X(\omega, t)$ or synthesize entirely a phase spectrum to minimize any phase artifacts. Due to space and time constraints we postpone the investigation of these issues in future publications.

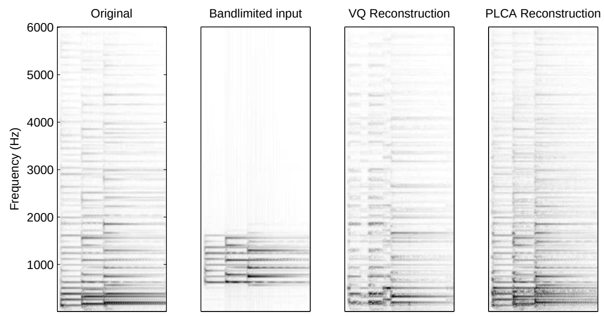

Figure 2: Comparison of VQ and latent variable methods with polyphonic sources. Note how the VQ cannot perform as well since it cannot use multiple elements to describe the additive mixture. It instead alternates between spectra of individual notes from the training data.

Finally it should be noted that there are additional options for reconstructing $\hat{y}(t)$. After equation (10) we can perform $|\hat{Y}(\omega, t)| = |X(\omega, t)|, \forall \omega \in \Omega$ so that we use the original data, or we can even use a weighted average of $\hat{y}(t)$ and $x(t)$ to obtain the final result. Again, these are open questions which warrant further investigation which is however outside the scope of this paper.

4. RESULTS

To highlight the advantages of this technique for bandwidth expansion consider the example shown in figure 2. The leftmost plot displays the original signal, a set of three piano notes which overlap in time. This sound was then bandlimited so that it only had energy in the $650Hz - 1600Hz$ region (second plot from the left). As an example high-bandwidth sound we used a recording of the same piano playing various notes. We extracted a dictionary of 300 elements using both a VQ and a latent variable model. The two right plots display the results of fitting each dictionary to the input. One can see that the VQ model is having trouble dealing with the overlapping notes since the fitting operation uses a nearest neighbor approach which cannot combine dictionary elements to approximate the input. On the other hand the latent variable model is very effective at picking multiple dictionary elements to approximate the areas with concurrent notes. Comparing the final results once can easily see that the latent variable model has produced a superior reconstruction as compared to a more standard VQ model.

This ability of the latent variable model to deal with overlapping dictionary elements is what makes this an appropriate model for complex sources such as music. Traditionally bandwidth expansion is evaluated on speech which is a monophonic source where dictionary elements can be used in succession. In more complex sources the dictionary elements are not present in isolation. This complicates the extraction of an accurate dictionary and the subsequent fitting for the reconstruction. The latent model being a linear additive model doesn't exhibit any problems in extracting or fitting overlapping dictionary elements and it thus better suited for these problems.

To evaluate the performance of this approach we run multiple examples and performed subjective listening tests on a variety of bandwidth expansion cases. It was generally agreed upon listeners that this approach performed well for most cases. Since the results cannot be accuretely represented on paper they are available as soundfiles at:
http://www.merl.com/people/paris/be.html

5. CONCLUSIONS

In this paper we presented an example-based process to create high-bandwidth versions of low bandwidth recordings. We introduced the idea of a latent variable model for spectral analysis and demonstrated its value for extracting and fitting spectral dictionaries from time-frequency distributions. We showed how these dictionaries can be used to map high-bandwidth elements to bandlimited recordings and how to create bandwide reconstructions. As compared to the predominantly monophonic approaches to this problem we've shown how this technique performs well with complex signals such as music where dictionary elements are often linearly added.

The authors would like to acknowledge the assistance of Madhusudana Shashanka in preparing this work.

6. REFERENCES

[1] H. Yasukawa, "Signal Restoration of Broadband Speech using Non-linear Processing", in Proceedings of the European Signal Processing Conference (EUSIPCO), PP. 987-990, 1996.

[2] N. Enbom and W.B. Kleijn, "Bandwidth Expansion of Speech based on Vector Quantization of Mel Frequency Cepstral Coefficients", In Proceedings IEEE Workshop on Speech Coding, Porvoo, Finland, PP. 171-173, 1999.

[3] Y.M.Cheng, D.O'Shaugnessey, and P. Mermelstein, "Statistical Recovery of Wideband Speech from Narrowband Speech", IEEE Trans. on Speech and Audio Processing, Vol. 2, PP. 544-548, Oct 1994.

[4] K.-Y. Park and H. S. Kim, "Narrowband to Wideband Conversion of Speech using GMM Based Transformation", In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1843-1846, 2000.