A Super-Resolution Spectrogram Using Coupled PLCA

Juhan Nam[1]*, Gautham J. Mysore*[1]*, Joachim Ganseman*[2]*, Kyogu Lee*[3]*, Jonathan S. Abel* [1]

[1]Center for Computer Research in Music and Acoustics, Stanford University
[2]IBBT-Vision Lab, Department of Physics, University of Antwerp
[3]Department of Digital Contents Convergence, Seoul National University
juhan@ccrma.stanford.edu, gautham@ccrma.stanford.edu, joachim.ganseman@ua.ac.be,
kglee@snu.ac.kr, abel@ccrma.stanford.edu

Abstract

The short-time Fourier transform (STFT) based spectrogram is commonly used to analyze the time-frequency content of a signal. Depending on window size, the STFT provides a trade-off between time and frequency resolutions. This paper presents a novel method that achieves high resolution simultaneously in both time and frequency. We extend Probabilistic Latent Component Analysis (PLCA) to jointly decompose two spectrograms, one with a high time resolution and one with a high frequency resolution. Using this decomposition, a new spectrogram, maintaining high resolution in both time and frequency, is constructed. Termed the "super-resolution spectrogram", it can be particularly useful for speech as it can simultaneously resolve both glottal pulses and individual harmonics.

Index Terms: STFT, spectrogram decompositions, spectral analysis, signal representations, time-frequency distribution

1. Introduction

The short-time Fourier transform (STFT) is a general-purpose tool to represent a signal in a time-frequency domain [1]. The spectrogram, which is the magnitude display of the STFT, is particularly useful for speech and musical signals in that it provides visualized information of sound sources, for example, temporal evolution of harmonic and noise components. A downside of the STFT is that high resolution cannot be achieved simultaneously in both time and frequency. The STFT of a signal is computed from a series of signal segments over a sliding window. When the window is short, temporal changes of the signal can be well observed while harmonic peaks are blurred on the frequency axis. When the window is long, on the other hand, harmonic peaks become sharp while temporal changes are smeared on the time axis.

To overcome the inherent tradeoff between time and frequency resolutions of the STFT, alternative time-frequency representations have been suggested. One approach is using non-uniform time-frequency resolution. The wavelet transform and constant-Q transform fall into this category [2] [3]. Others include quadratic transforms known as Cohen's class, such as, the Wigner-Ville and Choi-Williams distributions [4] [5]. They are known to provide high resolution, particularly for non-stationary signals. Another method is the reassigned spectrogram, which also provides highly sharpened spectral distribution by remapping the regular STFT to instantaneous time and frequency domains [6].

In this paper, we present a novel time-frequency representation that can achieve high resolution simultaneously in both time and frequency domains. The basic idea is to first jointly decompose two spectrograms, one formed with a high time resolution and the other with a high frequency resolution. We propose a method termed here Coupled Probabilistic Latent Component Analysis (PLCA) to perform this decomposition. Coupled PLCA is an extension to PLCA [7] that is used to decompose individual spectrograms. Using the result of these decompositions, we construct a spectrogram that has a high resolution in both time and frequency. This resulting spectrogram effectively resolves temporal changes and sharp harmonic peaks simultaneously.

2. Proposed Method

In this section, we first describe PLCA applied to spectrogram decomposition. We then describe the proposed method, Coupled PLCA, which jointly decomposes two spectrograms derived from the same signal source. Finally, we describe the construction of the super-resolution spectrogram.

2.1. PLCA

PLCA is a technique that is used to decompose a spectrogram into a sum of outer products of non-negative spectral and temporal components. The PLCA model is given by:

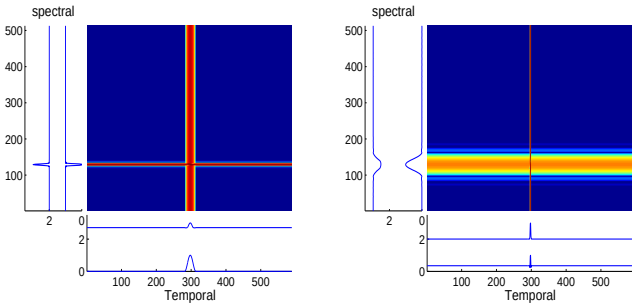
$$V_{ft} \approx \gamma \sum_z P(z)P(f|z)P(t|z) \quad (1)$$

where V_{ft} is spectrogram evaluated at f, t. $P(f|z)$ are spectral components, which can be interpreted as basis vectors. $P(t|z)$ are temporal components that indicate the occurrences of the corresponding spectral components in time. $P(z)$ is a distribution of weights. γ is a scaling factor. All of these distributions are multinomial distributions.

A given spectrogram can be decomposed by estimating the parameters of the multinomial distributions, $P(f|z)$, $P(t|z)$, and $P(z)$. This parameter estimation is done using the Expectation-Maximization (EM) algorithm as follows:

E Step:

$$P(z|f,t) = \frac{P(z)P(f|z)P(t|z)}{\sum_z P(z)P(f|z)P(t|z)} \quad (2)$$



(a) (b)

Figure 1: *Two spectrograms of a mixture of sine and impulse, one with high frequency resolution (left) and the other with high time resolution (right). Only either one of them is displayed in high resolution. Note that shapes of the temporal and spectral components also follow time and frequency resolutions of the spectrograms.*

M Step:

$$P(f|z) = \frac{\sum_t V_{ft} P(z|f,t)}{\sum_f \sum_t V_{ft} P(z|f,t)} \quad (3)$$

$$P(t|z) = \frac{\sum_f V_{ft} P(z|f,t)}{\sum_t \sum_f V_{ft} P(z|f,t)} \quad (4)$$

$$P(z) = \frac{\sum_f \sum_t V_{ft} P(z|f,t)}{\sum_z \sum_f \sum_t V_{ft} P(z|f,t)} \quad (5)$$

2.2. Coupled PLCA

PLCA can be used to decompose an individual spectrogram. However, the spectrogram can have high resolution in either time or frequency, but not both. For a given signal, we can compute a spectrogram with high frequency resolution, $V^{(F)}$ using a long window, and a spectrogram with high time resolution, $V^{(T)}$ using a short window. We can perform a separate decomposition on each of these spectrograms using PLCA. This will yield two separate sets of component distributions:

1. High frequency resolution distributions — $P_F(f|z)$, $P_F(t|z)$, and $P_F(z)$
2. High time resolution distributions — $P_T(f|z)$, $P_T(t|z)$, and $P_T(z)$

$P_F(f|z)$ will describe the spectral structure of the spectrogram with a high frequency resolution. $P_F(t|z)$ will indicate the occurrences of these spectral components in time. However, the resolution of these occurrences in time will be poor. On the other hand, $P_T(t|z)$ will indicate the occurrences with a much higher resolution while the distribution of $P_T(f|z)$ is smeared. Fig. 1 illustrates an example of the two spectrograms and their corresponding spectral and temporal components applied to the sum of a sinusoid and an impulse.

This decomposition gives rise to the idea that a spectrogram with high resolution in both time and frequency can be constructed by using the spectral resolution of $P_F(f|z)$ and the temporal resolution of $P_T(t|z)$. The problem is that $P_F(f|z)$ and $P_T(t|z)$ are computed independently. This means that they have no direct correspondence. For example, there is no guarantee that $P_T(t|Z=1)$ will indicate the occurrences of $P_F(f|Z=1)$. When a large number of components are used, there is very little chance of having a correspondence between the components.

In order to force a correspondence, we couple the two independent problems by assuming that the two STFT spectrograms were derived from the same high resolution parent spectrogram. In this way, the components $P_T(f|z)$ and $P_F(t|z)$ can be interpreted as blurred versions of $P_F(f|z)$ and $P_T(t|z)$, respectively. Using a frequency blurring function $B_F()$ and a time blurring function $B_T()$, we have the following equations that establish the coupling between the two independent problems:

$$P_T(f|z) = B_F\left(P_F(f|z)\right) \quad (6)$$

$$P_F(t|z) = B_T\left(P_T(t|z)\right) \quad (7)$$

Using this correspondence, we eliminate $P_T(f|z)$ and $P_F(t|z)$ from the independent problems and propose a coupled estimation procedure that we call Coupled PLCA. The estimation is done using the EM algorithm as follows:

E Step:

$$P_F(z|f,t) = \frac{P_F(z)P_F(f|z)B_T\left(P_T(t|z)\right)}{\sum_z P_F(z)P_F(f|z)B_T\left(P_T(t|z)\right)} \quad (8)$$

$$P_T(z|f,t) = \frac{P_T(z)B_F\left(P_F(f|z)\right)P_T(t|z)}{\sum_z P_T(z)B_F\left(P_F(f|z)\right)P_T(t|z)} \quad (9)$$

M Step:

$$P_F(f|z) = \frac{\sum_t V_{ft}^{(F)} P_F(z|f,t)}{\sum_f \sum_t V_{ft}^{(F)} P_F(z|f,t)} \quad (10)$$

$$P_T(t|z) = \frac{\sum_f V_{ft}^{(T)} P_T(z|f,t)}{\sum_t \sum_f V_{ft}^{(T)} P_T(z|f,t)} \quad (11)$$

$$P_F(z) = \frac{\sum_f \sum_t V_{ft}^{(F)} P_F(z|f,t)}{\sum_z \sum_f \sum_t V_{ft}^{(F)} P_F(z|f,t)} \quad (12)$$

$$P_T(z) = \frac{\sum_f \sum_t V_{ft}^{(T)} P_T(z|f,t)}{\sum_z \sum_f \sum_t V_{ft}^{(T)} P_T(z|f,t)} \quad (13)$$

2.3. Blurring Function

The blurring functions are derived from the relationship between long and short windows used in forming the two spectrograms. Since the spectrogram is given as the magnitude of the STFT, the temporal and spectral blurring functions are expected to only approximate a Fourier transform pair. Furthermore, as the temporal and spectral components are non-negative, no-sidelobe or low-sidelobe windows [8] should be considered. We approach the solution with elementary signals, assuming that the result can be generalized to complicated signals.

In Fig. 1, the shapes of temporal components corresponding to the impulse are determined by the windows used in the spectrograms because windowing the impulse results in sampling the window every hop size (which is defined as the window length minus the overlap length in STFT). Likewise, from duality between the impulse and sine in time and frequency domains, the shapes of frequency components corresponding to the sine frequency are determined by the magnitude response of the window. Therefore, the blurring function can be viewed as a linear filter that converts a short window to a long window in the

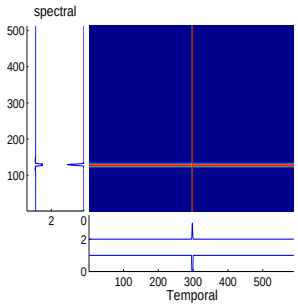


Figure 2: *Super-resolution spectrogram of a mixture of sine and impulse. It is constructed from high time and frequency resolution components obtained by coupled PLCA.*

time domain or the sharp magnitude response to the wide one in the frequency domain. These can be expressed as follows:

$$w_F(n) = B_F(w_T(n)) = \sum_i b_T(i) w_T(n - ih) + e_T(n) \quad (14)$$

$$W_T(k) = B_T(W_F(k)) = \sum_j b_F(j) W_F(k - j) + e_F(k) \quad (15)$$

where h is the hop size of the STFT, $w_T(n)$ and $w_F(n)$ are short and long windows, $W_T(k)$ and $W_F(k)$ are their magnitude responses, and $b_T(n)$ and $b_F(k)$ are coefficients of the time and frequency blurring functions, respectively. Note that the time index i spans the width of the long window, whereas the frequency index j is limited to the main lobe of the wide magnitude response assuming that the sidelobe level is low enough. The unknown blurring filter coefficients $b_T(i)$ and $b_F(j)$ can be computed by least squares, that is, minimizing the squared error noted in the Eq. (14) and (15).

The blurring functions that we derived above only approximate the time or frequency blurring that occurs in the hypothesized conversion from the super-resolution spectrogram to the high frequency resolution spectrogram or the high time resolution spectrogram. Eq. (7) and (6) are therefore approximations and do not hold with equality. We are therefore effectively performing an approximation to Coupled PLCA (and to the EM algorithm).

2.4. Super-Resolution Spectrogram

Using Coupled PLCA and blurring functions, we compute spectral components, $P_F(f|z)$, with high frequency resolution and temporal components, $P_T(t|z)$, with high time resolution. Using these distributions along with the weights, we can construct a spectrogram with high resolution in both time and frequency. Since the problem is coupled, the mixture weights $P_F(z)$ and $P_T(z)$ tend to be almost identical. We can therefore use either one. The super-resolution spectrogram is constructed as follows:

$$V_{ft}^{(S)} = \sum_z P_F(z) P_F(f|z) P_T(t|z) \quad (16)$$

Fig. 2 shows the super-resolution spectrogram for the sine and impulse mixture. Compared to Fig. 1, both sine and impulse are displayed with a high resolution.

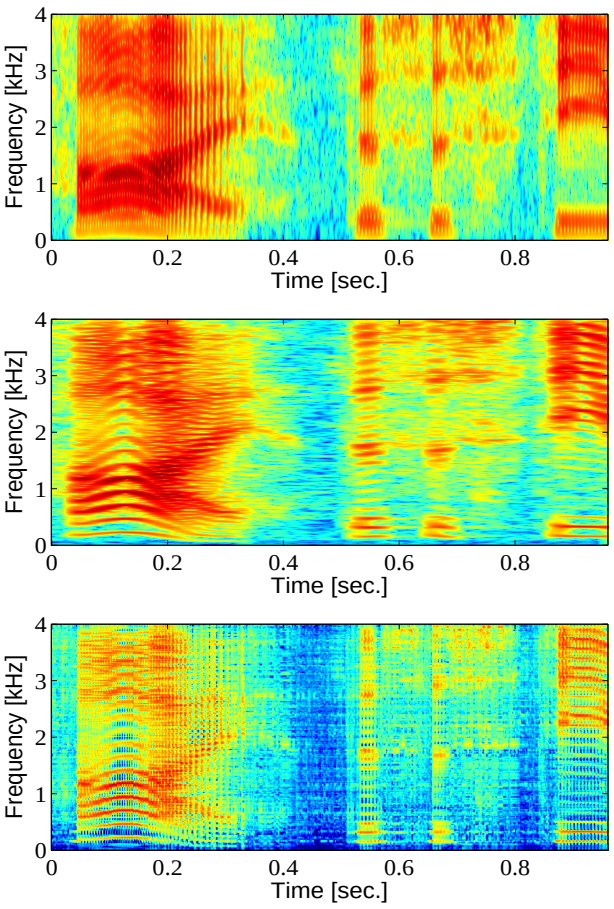


Figure 3: *Three spectrograms of a male voice: high time resolution spectrogram (top), high frequency resolution spectrogram (middle), and super-resolution spectrogram (bottom). Hann windows of 8 and 64 ms were used for the high time resolution and high frequency resolution spectrograms respectively, with 1 ms hop size in both cases.*

3. Results

The proposed method has been applied to speech. Speech signals are typically analyzed using both wideband (high time resolution) and narrowband (high frequency resolution) spectrograms because they provide different types of information. A wideband spectrogram effectively displays the variation of formants and periodic excitation of glottal pulses for voiced speech, while a narrowband spectrogram shows trajectories of individual harmonics, which are associated with the pitch of voiced speech.

Fig. 3 compares wideband, narrowband, and the proposed super-resolution spectrograms of a speech example. To compute the super-resolution spectrogram, Coupled PLCA was performed on the wideband and narrowband spectrograms derived from the same speech signal using 100 components and running the EM algorithm for 100 iterations. The resulting super-resolution spectrogram presents well-resolved occurrences of glottal pulses as well as sharp harmonic curves, preserving the locality in both time or frequency. Note that the voiced part of the super-resolution spectrogram has a waffle pattern due to simultaneous appearance of the harmonics and periodic glottal

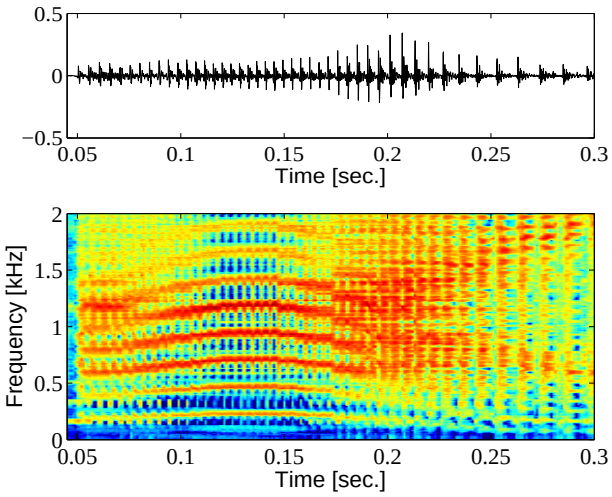


Figure 4: *Magnified portion of the super-resolution spectrogram in Fig. 3 and its time-domain waveform. Note that the glottal pulses in the waveform correspond the vertical lines in the super-resolution spectrogram.*

pulses. This can help distinguishing voiced speech from unvoiced speech, which has a rather ragged and sparse distribution.

Fig. 4 magnifies a voiced portion of the super-resolution spectrogram and its corresponding waveform in the time domain. The periodic occurrences of glottal pulses seen in the waveform exactly match the vertical lines in the super-resolution spectrogram. This demonstrates the surprising fact that pitch of the speech can be observed from not only the harmonic relation on the frequency axis but also the periodic series of vertical lines on the time axis in the super-resolution spectrogram. Since the period and frequency are reciprocal to each other, the rectangular grid becomes taller and thinner as pitch increases (0.05 to 0.14 ms), whereas it turns shorter and broader as the pitch decreases (0.14 to 0.30 ms).

4. Discussion

While experimenting with the proposed method, we found that the following should be taken into account in order to obtain a satisfying super-resolution spectrogram. First, STFT parameters, in particular, window size and hop size, should be appropriately chosen so that high-time and high-frequency resolution spectrograms $V^{(T)}$ and $V^{(F)}$ contain desirable properties in each time and frequency domain. We have seen that the super-resolution spectrogram does not produce a successful result without such conditions. Second, the short and long window of the two STFTs should be center-aligned over the waveform for the same numbered frame of the two spectrograms. When they are not correctly aligned, the time blurring functions can distort the result. Third, the super-resolution spectrogram generally works well for signals with both sharp attack and periodicity. Thus, speech signals are seen to be good examples to demonstrate the effectiveness of the proposed method.

Our first experiment toward the super-resolution spectrogram also revealed two issues. The EM algorithm is guaranteed to increase the log-likelihood and get closer to a local optimum in every iteration. We can therefore assume that the results can only get better with more iterations. Since we are only using an approximation to the EM algorithm (due to inexact blurring functions) as described in sec. 2.3, we lose these convergence guarantees. In practice, we therefore observe that the super-resolution spectrogram occasionally degrades beyond about 150 iterations. In order to avoid the problem, more investigation will be needed to figure out the blurring process for a given window. Furthermore, an optimized window, which is not only robust to the degradation but also minimizes the error in estimating time and frequency blurring functions, should be designed to improve the result. Another caveat is the execution time. We performed the coupled PLCA using Matlab running on a computer with 2.16 GHz Intel Core Duo and 2GB RAM. It took more than a minute for the speech example in Fig. 3, which is one second long at 8kHz sampling frequency, with 100 components and 100 EM iterations. In order to improve the execution time, the minimum iteration and the minimum number of components necessary to produce a reasonably well-resolved spectrogram in both time and frequency, will need to be studied.

5. Conclusion

In this paper, we present a method for constructing a spectrogram with high locality in both time and frequency domains, starting from two STFT-based spectrograms, one with a high resolution in the time domain, and the other with a high resolution in the frequency domain. We extended the PLCA method for spectrogram modeling, and proposed Coupled PLCA. Here we link different spectrograms of the same sound source through blurring functions, which enables us to estimate parameters for a single super-resolution spectrogram. Successful initial testing on speech signals validates our approach. Future works include designing more exact blurring functions to ensure the convergence of the EM algorithm, finding optimal PLCA parameters, and qualitative evaluation of speech signal analysis. In addition, the coupling method can be extended to incorporate three or more spectrograms.

6. References

[1] Smith, J. O., *Spectral Audio Signal Processing*, October 2008 Draft, http://ccrma.stanford.edu/~jos/sasp/, online book, accessed 29-Apr-2010

[2] Mallat, S., *A Wavelet Tour of Signal Processing*, Academic Press, 2008.

[3] Brown, J.C.,"Calculation of a constant Q spectral transform", J. Acoust. Soc. Am., 89(1):425434, 1991.

[4] Cohen, L.,"Time-frequency distributions - A review", Proc. IEEE, vol. 77, no. 7, July 1989

[5] Choi, H. I. and Williams, W. J.,"Improved time-frequency representation of multicomponent signals using exponential kernels", IEEE. Trans. Acoustics, Speech, Signal Processing, vol. 37, no. 6, pp. 862871, June 1989.

[6] Fulop, S. A. and Fitz, K., "Algorithms for computing the time-corrected instantaneous frequency (reassigned) spectrogram, with applications," J. Acoust. Soc. Am. 119, 360371, 2006.

[7] Smaragdis, P., Raj, B. and Shashanka, M.V., "A Probabilistic Latent Variable Model for Acoustic Modeling", Advances in models for acoustic processing workshop, Neural Information Processing Systems (NIPS), December 2006

[8] Depalle, P. and Hélie, T., "Extraction of Spectral Peak Parameters Using a Short-time Fourier Transform and no Sidelobe Windows, IEEE 1997 Workshop on Applications of Signal Processing to Audio and Acoustics, New York, 1997.