Audio Engineering Society

Convention Paper

Presented at the 120th Convention
2006 May 20–23 Paris, France

Adaptive Time-Frequency Resolution for Analysis and Processing of Audio

Alexey Lukin[1], *AES Student Member*, Jeremy Todd[2], *AES Member*

[1] Moscow State University, Moscow, Russia
lukin@graphics.cs.msu.ru

[2] iZotope, Inc., Cambridge, MA
jeremy@izotope.com

ABSTRACT

Filter banks with fixed time-frequency resolution, such as the Short-Time Fourier Transform (STFT), are a common tool for many audio analysis and processing applications allowing effective implementation via the Fast Fourier Transform (FFT). The fixed time-frequency resolution of the STFT can lead to the undesirable smearing of events in both time and frequency. In this paper, we suggest adaptively varying STFT time-frequency resolution in order to reduce filter bank-specific artifacts while retaining adequate frequency resolution. Several strategies for systematic adaptation of time-frequency resolution are proposed. The introduced approach is demonstrated as applied to spectrogram displays, noise reduction, and spectral effects processing.

1. INTRODUCTION

It is well known that signal processing algorithms dealing with multimedia information should account for properties of human perception in order to achieve better processing quality. There exist multiple studies of human auditory and visual perception which are extensively employed in image and audio compression algorithms. In this paper, we consider the time-frequency resolution of filter banks commonly used in audio analysis and processing, and we propose a multiresolution approach that improves several existing algorithms.

2. SHORTCOMINGS OF STFT

The Short Time Fourier Transform (STFT) is a filter bank which is widely used in audio analysis and processing. The STFT can also be plotted on a 2D graph as a function of both time and frequency, with color representing magnitude, to form a spectrogram display. Spectrograms are becoming a popular tool among audio engineers as they are much more perceptually-oriented than a traditional waveform display. Filter banks based on the STFT are used in algorithms for noise reduction and various spectral effects such as multiband delays, vocoders, and center channel extraction.

It is known from psychoacoustics that the frequency resolution of human hearing is not uniform. Instead, it follows a mel-scale which is approximately linear below 500 Hz and logarithmic above it [1]. The fixed time-frequency resolution of the STFT is purely linear, so it is not ideal from a perceptual standpoint. The artifacts specific to STFT-based processing are pre-echoes (time smearing of transient events) and insufficient frequency resolution at stationary parts (especially at low frequencies), leading to perceptually inadequate modeling of audio.

Pre-echoes are artifacts resulting from the fact that any modification of time-frequency coefficients of a signal spreads its effect along the entire window length of the filter bank in the time domain. For example, processing of transformed coefficients that capture an onset of a transient event will result in smearing of transient energy in time within the filter bank window, both in the forward and backward directions. The backward spreading (pre-echo) is typically much more audible due to properties of temporal masking of human hearing and the fact that ongoing transient energy will probably mask the post-echo. Audibly this results in "swishy", "non-focused" sounding of transients which include drums, percussion and other instruments with sharp attacks.

Insufficient frequency resolution manifests itself differently in different processing algorithms. Generally, it prevents algorithms from separating closely spaced tones. For example, in noise reduction this may lead to weaker suppression of noise. In center channel extraction or time stretching of audio this may lead to unwanted modulations in low frequencies.

In this paper, we will propose a method for reducing these artifacts simultaneously.

3. ADAPTIVE TIME-FREQUENCY RESOLUTION

There have been many attempts to build filter banks with variable time-frequency resolution for audio compression purposes [2]. However such attempts are limited by the fact that compression requires the critical sampling property of filter banks, to keep the amount of data in the signal at a minimum. This significantly restricts the freedom to vary time-frequency resolution. On the other hand, image and audio processing methods allow redundancy in oversampled filter banks which leads to the multiresolution framework for signal processing algorithms depicted in Fig. 1.

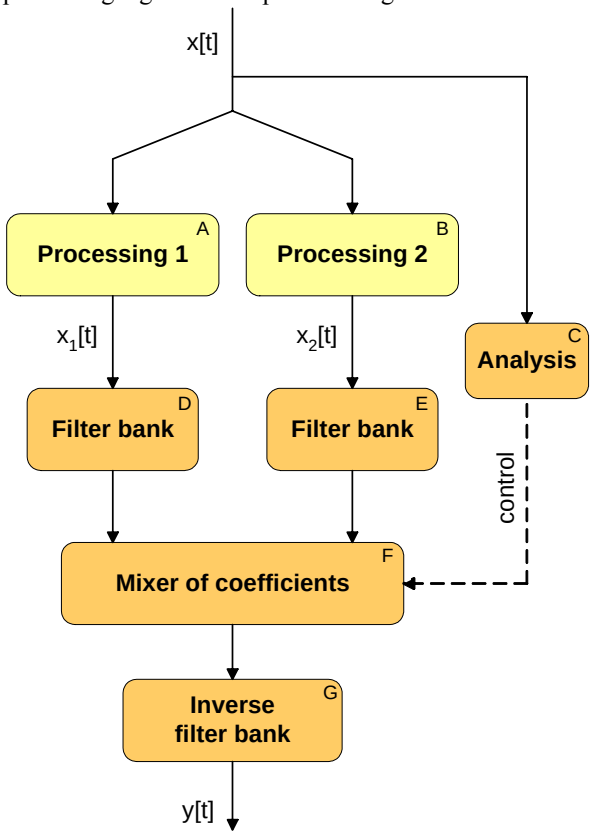

Figure 1. General scheme for signal processing with adaptive time-frequency resolution.

The same processing algorithm is running several instances (only two are depicted, but the framework can be generalized to any number of instances), labeled A and B above, with different fixed time-frequency resolutions that work in parallel on the same input data stream. The resulting signals $x_1[t]$ and $x_2[t]$ are processed signals which were processed with different time-frequency resolutions. Our goal is to combine them in order to achieve the desired resolution in every area of the time-frequency plane. This combination is performed by additional filter banks D and E, both with a single fixed time-frequency resolution that transforms these resulting signals into time-frequency coefficients on the same time-frequency grid. The resulting time-frequency coefficients can be adaptively mixed by the mixer F to select desired coefficients in each area of the time-frequency plane. The process of mixing can be controlled by some prior strategy (reflecting properties of human perception) and/or depending on local signal

features (e.g. on its stationarity) determined by analysis at C of the original input signal. Finally, the inverse filter bank G returns the processed signal to the time domain, forming the output y[t].

Since mixing of processed signals $x_1[t]$ and $x_2[t]$ is performed in the transform domain, the proposed framework allows arbitrary time-frequency resolution in arbitrary areas of the time-frequency plane. The number of individual processors operating at different time-frequency resolutions controls the smoothness of variation of resolution in the combined signal. Also, by mixing together coefficients from several resolutions, we can interpolate between given discrete resolutions.

The proposed framework can also be modified to perform signal *analysis* with arbitrary time-frequency resolution. As shown in Fig. 2, the input signal x[t] is fed directly to filter banks H and I operating with different fixed time-frequency resolutions. To simplify the mixer K, filter banks H and I should produce outputs $a_{f,t,1}$ and $a_{f,t,2}$ at the same grid of time-frequency locations. This can be accomplished in the case of the STFT by using different time-domain window lengths in H and I but using the same analysis hop and FFT size (zero-padding windowed data as necessary). The mixer K combines the outputs of the filter banks according to analysis performed in J, producing outputs $a_{f,t}$ in the transform domain.

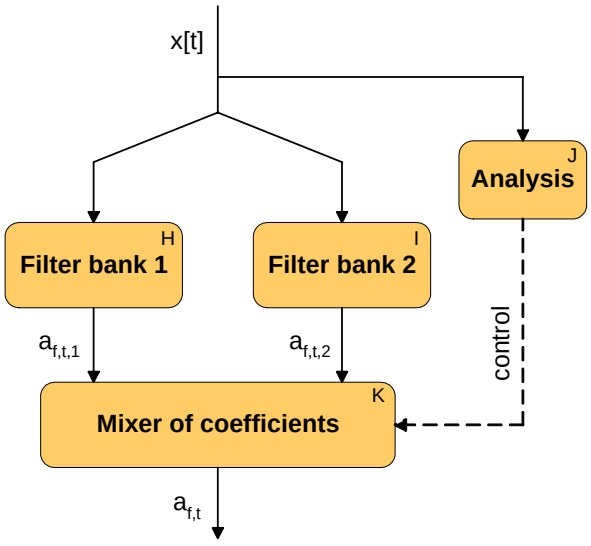

Figure 2. General scheme for signal analysis with adaptive time-frequency resolution.

4. ADAPTATION STRATEGIES

In this section, we describe two strategies for varying the time-frequency resolution of an STFT filter bank. Both of them incorporate prior knowledge about the frequency resolution of our hearing and adapt to the time-varying properties of a signal.

The first strategy is based on a signal transience estimator. We describe the estimator based on analysis of energy evolution in critical bands. Then we describe the strategy of varying time-frequency resolution in order to reduce pre-echoes in transient regions of the time-frequency plane and increase the frequency resolution in stationary regions.

The second strategy is based on the principle of minimal description length (MDL) [3]. It estimates the "optimality" of different time-frequency resolutions and selects the one that is locally optimal. Optimality is defined as minimal possible energy smearing both in time and frequency directions. This method is analogous to the general MDL paradigm of finding the transform with the most compact support for transformed energy.

4.1. Transience adaptation

One approach to adaptive time-frequency resolution of a filter bank is to explicitly account for signal stationarity. Stationarity means preservation of signal properties across time, including power and spectral shape. We define transience as the opposite of stationarity: variance of signal properties in time.

To reduce the time smearing of transients we will increase the temporal resolution of the filter bank at transient signal segments. During stationary segments, we will use higher frequency resolution.

Some simple detectors of transience are described in [4]; they estimate the variance of a short-time spectrum in adjacent time frames. Such a spectral similarity measure is susceptible to false detections of transients at stationary noisy parts of a signal resulting from statistical variance of short time spectral estimates of noise. In this work, we are using an algorithm which integrates signal energy in critical bands [2] and detects fast energy onsets on a per-band basis.

The signal is transformed into the STFT domain with a window size of 12 ms and an analysis hop of 6 ms. For each frame the signal power is integrated inside 24

critical bands covering the entire audible spectrum. The integrated energy is raised to the power of 1/8 to provide better sensitivity to relatively high energy onsets at small absolute levels. Then we detect variation of energy in time within each critical band by cross-correlating energies $e[b, t]$ with a filter $h[t]$ = {-1, -1, -1, 0, 1, 1, 1} (here b is the critical band number, t is the index of the STFT frame):

$$v[b,t] = e[b,t] * h[-t]$$

The transience $T[b,t]$ of the signal in each critical band is estimated as

$$T[b,t] = \begin{cases} v[b,t], & v[b,t] \geq 0 \\ \dfrac{|v[b,t]|}{10}, & v[b,t] < 0 \end{cases}$$

This provides 10 times better sensitivity to energy onsets than to energy decays.

When the transience of a signal in each critical band is estimated, we can use it to control the time-frequency resolution of a filter bank by reducing frequency resolution around transients. This reduces the smearing of transients in time while keeping good frequency resolution at stationary parts of the signal.

The default behavior of the mixer of coefficients can reflect the perceptual property of better low-frequency resolution. At the same time, the suggested transience detector can alter the default mixing strategy towards better time resolution around transients (see section 6.1 for details).

4.2. Maximal energy compaction principle

When plotting spectrograms, the main problem with fixed time-frequency resolutions is the smearing of signal energy. Smearing in frequency causes harmonics to appear as thick lines and can prevent distinguishing closely spaced harmonics. Smearing in time means loss of time resolution and can negatively affect estimation of positions and durations of transient events in the signal. It would be desirable to jointly reduce smearing of energy in both directions. However this is not possible due to the uncertainty principle.

What we propose in this section is to estimate the amount of energy smearing for different fixed time-frequency resolutions and select the resolution that minimizes such smearing in both temporal and frequency directions.

Let's consider a small rectangular area Ω of the time-frequency plane and short-time Fourier transforms with different time-frequency resolutions of the same signal in this area. STFT coefficients for different resolutions can be obtained by calculating the STFT with time-domain windows of varying length. Analysis hops of windows and frequency grids should be equal for all STFT resolutions, just as they are in the general analysis framework depicted in Fig. 2. This ensures that squared STFT magnitudes $a_{f,t,r}$ at different resolutions are calculated in the same grid of time-frequency locations. Here t and f are time and frequency indices of STFT coefficients, and r indexes available STFT resolutions. Our task is to select the r that minimizes energy smearing in the area Ω.

To achieve this, we sort the $a_{f,t,r}$ inside Ω by their magnitudes, in descending order, for each resolution r. We name the sorted results $a_{i,r}$. Next we define energy smearing of every particular STFT resolution as:

$$S_r = \frac{\sum_i i \cdot a_{i,r}}{\sqrt{\sum_i a_{i,r}} + \varepsilon}$$

The numerator of the fraction evaluates the first moment of the statistical distribution of squared magnitudes. The denominator normalizes the numerator by the total energy of the signal in area Ω. The square root in the denominator assigns higher smearing to resolutions with higher overall energy in the area Ω. Since the energy in Ω varies only due to differences in amount of leakage (smearing) from adjacent regions at various resolutions, this penalizes resolutions where excessive smeared energy comes from surrounding areas of the time-frequency plane. The small constant ε prevents division by zero.

When energy smearing measures S_r are calculated for every resolution, we select the resolution r_0 which minimizes energy smearing:

$$r_0 = \arg\min_r S_r$$

This resolution is selected as "optimal" in the area Ω and is used to build a spectrogram or process the audio signal.

5. APPLICATION TO DISPLAY OF SPECTROGRAMS

The audio spectrogram has become an important tool in audio engineering. Many common operations, such as content analysis, removal of artifacts, and basic editing operations are supported by spectrograms in popular sound editors. The main factors limiting the usefulness of a typical STFT-based spectrogram view are a linear frequency scale obscuring many low-frequency details, and the fixed time-frequency resolution of the STFT leading to time or frequency smearing of audio events.

If a typical STFT spectrogram with fixed time-frequency resolution is displayed with a perceptually meaningful frequency scale (e.g. the mel-scale) the lack of low-frequency resolution becomes obvious (Fig. 3). However increasing the frequency resolution of the STFT will produce time smearing of transients (Fig. 4). It is possible to combine spectrograms taking low-frequency spectrogram data from the STFT with high frequency resolution and high-frequency data from the STFT with better temporal resolution (Fig. 5).

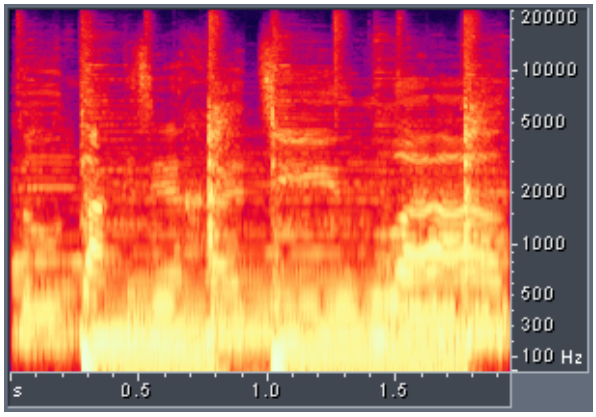

Figure 3. STFT spectrogram, window size is 12 ms.

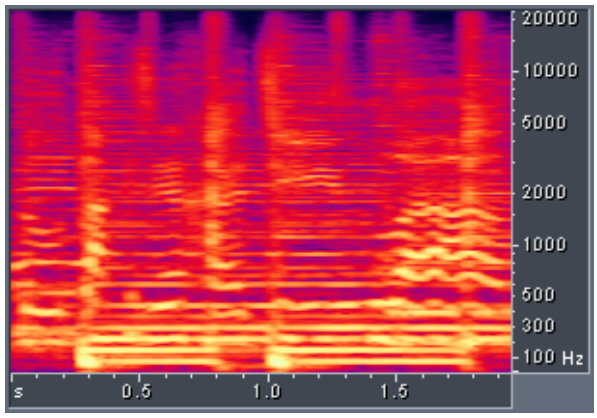

Figure 4. STFT spectrogram, window size is 93 ms.

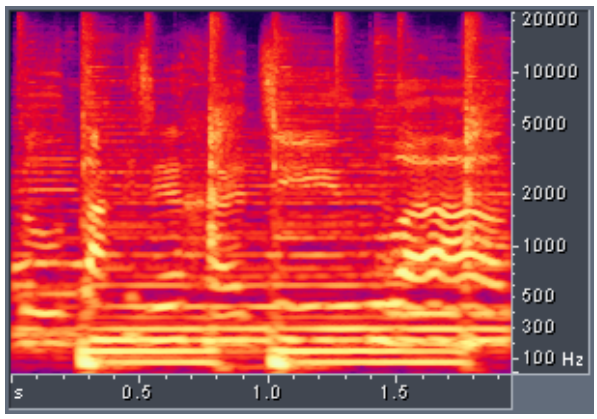

Figure 5. Spectrogram with combined STFT resolutions.

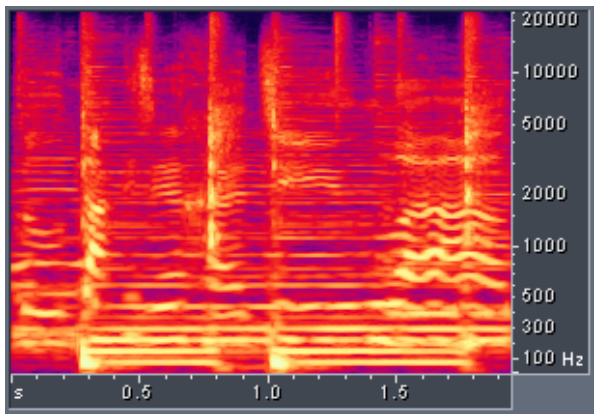

Figure 6. Spectrogram with adaptive resolution.

In section 5.1 we introduce an adaptive spectrogram that provides further improvements in the adaptation of time-frequency resolution (Fig. 6).

5.1. Resolution selection algorithm

To further reduce smearing of energy in spectrograms, we apply the resolution selection strategy described in section 4.2. We calculate the STFT with 4 different window sizes: 12, 24, 48, and 96 ms. Time-frequency magnitudes are calculated on the same grid by zero padding windowed signals and using equal STFT analysis hops for every resolution.

To obtain the optimal resolution at every point (f, t) of the time-frequency plane we consider a rectangular area Ω around this point that is 1 critical band wide and 48 ms long. The tradeoff here is that small Ωs will not allow us to form a robust estimate of energy smearing as there will be too few STFT coefficients inside, and large Ωs will not be local enough for fine control of resolution. We select critical bands because they have some perceptual meaningfulness, and also because they project to equal-height areas on our mel-scale spectrogram. Our 48 ms width is chosen after experimental evaluation of the look and meaningfulness of spectrograms with various widths for Ω.

Next, according to section 4.2, we calculate the best resolution choice r_0 of the 4 available resolutions. The STFT magnitude coefficient a_{f,t,r_0} is used to form the spectrogram view at point (f, t).

In order to prevent hard switching from one resolution to another we have updated the algorithm to *mix* magnitude coefficients instead of switching between them. In this manner we are able to "interpolate" between 4 available frequency resolutions. The mixing is performed according to respective energy smearing measures of each resolution:

$$a_{f,t} = \sum_{r=1}^{4} w_r \cdot a_{f,t,r}$$

Mixing weights w_r are calculated as follows:

$$w_r = \frac{k}{S_r^8 + \varepsilon}$$

Here k is the normalization constant selected so that the sum of all w_r is 1, and ε is a small constant preventing division by 0.

5.2. Simulation results

We have conducted simulations to compare the look and usefulness of conventional STFT spectrograms and adaptive-resolution spectrograms. The first test signal consisted of an artificially generated 1 kHz tone with a sharp fade-in lasting 2 ms and a smooth decay lasting 600 ms (Fig. 7). Since the tone onset is abrupt, it contains a transient energy burst spreading outside of the 1 kHz band, as shown by the conventional spectrogram (Fig. 8). By varying the resolution of the spectrogram, we can make either the horizontal line (the decaying tone) or the vertical line (the transient attack) thinner, but not both at once. However our adaptive spectrogram is able to locally select the time-frequency resolution which minimizes smear both in time and frequency (Fig. 9) leading to less frequency spreading of the slow tone decay and better time localization of the tone onset.

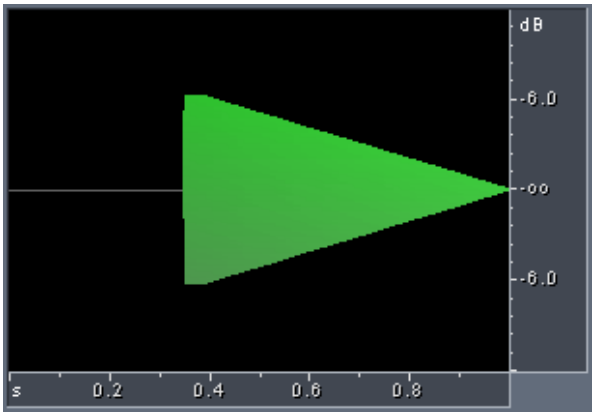

Figure 7. Waveform of a tone onset.

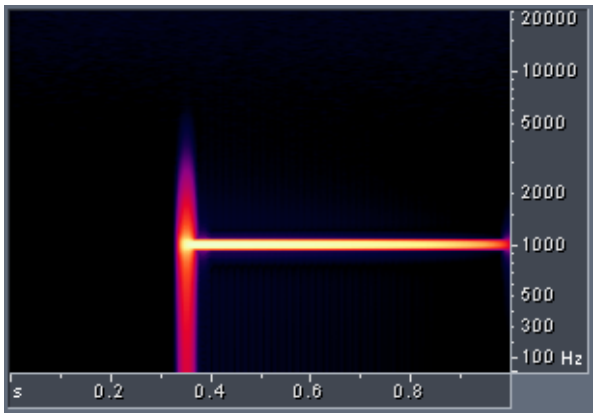

Figure 8. STFT spectrogram of a tone onset, window size is 46 ms.

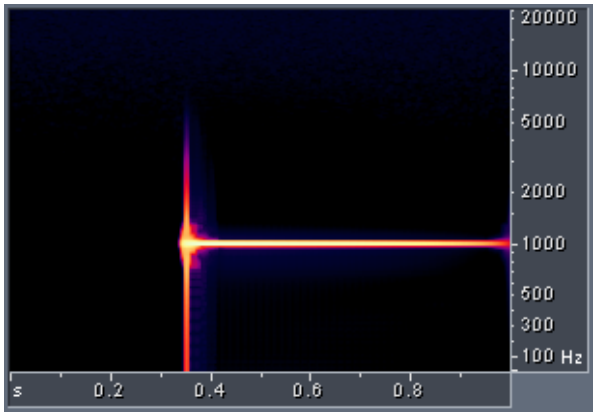

Figure 9. Adaptive resolution spectrogram of a tone onset.

Our next example is a piece of folk music with flute, cello, guitar and percussive drums. The conventional STFT spectrogram (Fig. 10) lacks low-frequency resolution and is unable to separate bass notes of the cello and guitar. At the same time the temporal resolution at high frequencies is not enough to sharply localize onsets of the drums. Our adaptive spectrogram (Fig. 11) fixes both of these problems: the low-frequency resolution is increased, and drum onsets are displayed more sharply due to better local time resolution. The adaptiveness of the time-frequency resolution also enables us to use better frequency resolution at high frequencies to resolve closely spaced guitar overtones above 3 kHz.

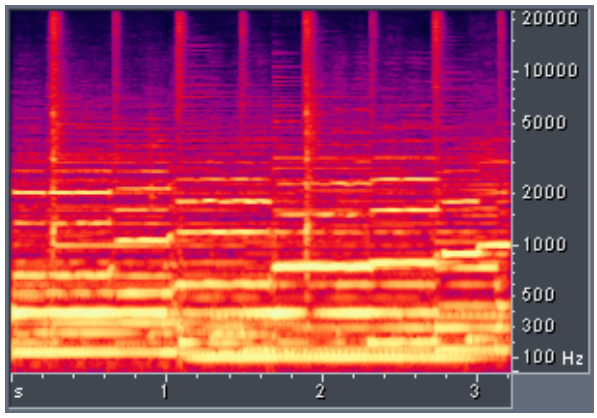

Figure 10. STFT spectrogram of folk music, window size is 46 ms.

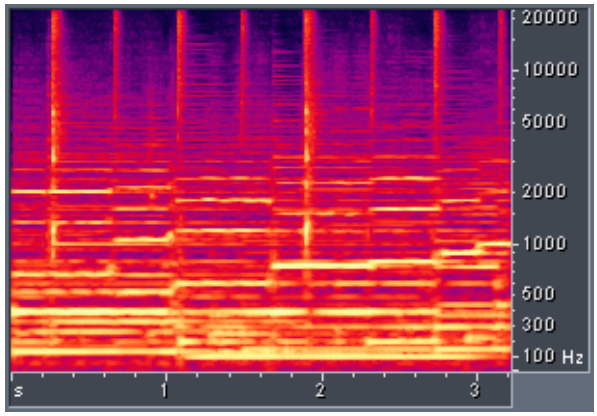

Figure 11. Adaptive resolution spectrogram of folk music.

Another example of our adaptive spectrograms is given in Fig. 6 displaying a piece of rock music with vocal, bass, drums, guitars, flute and violin. Again, the adaptive spectrogram is able to resolve low-frequency harmonics, and it avoids the smearing of bass drum hits (at 0.3 and 1.0 seconds). In the high-frequency area we are able to preserve the sharpness of drum onsets and resolve closely spaced guitar harmonics.

These tests show that the proposed adaptive approach to the calculation of spectrograms allows a spectrogram to display more useful details and musical events with better precision.

6. APPLICATION TO AUDIO PROCESSING

In this section, we show how filter banks with adaptive time-frequency resolution can be applied to improve the quality of several audio processing algorithms: the spectral subtraction algorithm for noise reduction [5] and the center channel extraction algorithm. We run several instances of single-resolution processors and adaptively combine their results in the time-frequency plane using one of the suggested strategies. The resulting audio signal shows significant reduction of time smearing of transients and at the same time good frequency resolution allowing effective suppression of tonal noise or extraction of the center channel.

In spite of the increased computational complexity compared to a single-resolution STFT, both algorithms allow real-time implementation.

6.1. Noise reduction

Most noise reduction methods for additive stationary noises in audio are based upon the spectral subtraction algorithm [5, 6]. This algorithm transforms the noisy signal with a filter bank and attenuates coefficients that are supposedly part of the noise, using a-priori knowledge of the noise spectrum. Then the inverse filter bank reconstructs the cleaned signal. In this paper, we will not discuss details of spectral subtraction methods, but rather show how modification of a filter bank can improve the quality of the result by reducing artifacts specific to filter banks.

A typical filter bank for spectral subtraction is based on the STFT. Good frequency resolution of the STFT filter bank allows separation of closely spaced noise and signal harmonics. Good frequency resolution also leads to stronger possible noise attenuation due to lower noise power per STFT bin. However good frequency resolution requires long STFT windows leading to poor time resolution. Spectral subtraction with poor time resolution is not able to suppress noise before transient onsets since the part of the transient falls into the window and raises the coefficient magnitude preventing attenuation (Figs. 12, 13). Another problem is general pre-echo associated with the modification of STFT coefficients at poor time resolution.

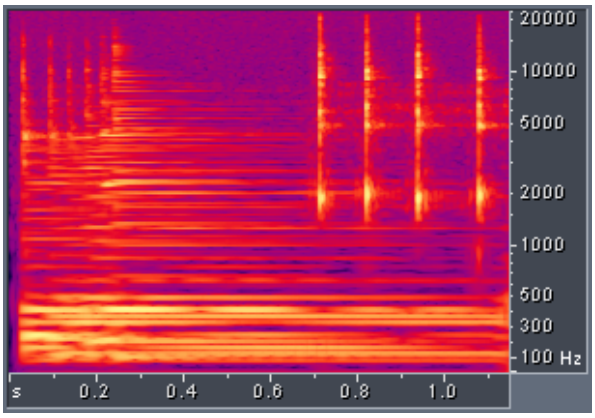

Figure 12. Noisy sample of guitar and castanets.

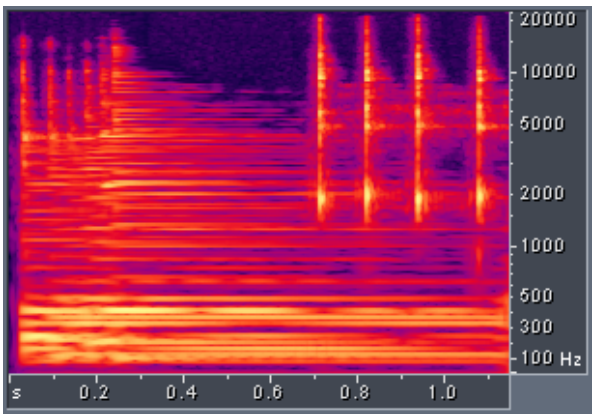

Figure 13. Result of noise reduction with a 46 ms STFT window.

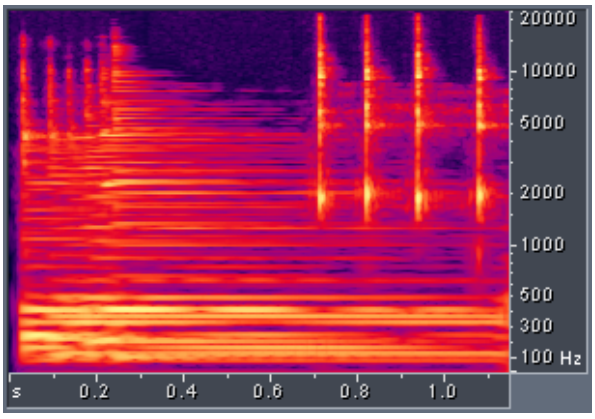

Figure 14. Result of noise reduction with adaptive time-frequency resolution.

We suggest using one of the strategies described in section 4 to adapt the time-frequency resolution of the filter bank. Proper selection of time-frequency resolution will result in better energy compaction in the transform domain, which is always desirable for noise reduction. Adaptive resolution will also allow good frequency resolution in stationary signal parts and good time resolution around transients leading to less time smearing artifacts.

To test our approach, we selected the transient detection strategy described in section 4.1. We used spectral subtraction with 3 STFT filter banks with window sizes of 24, 48, and 96 ms and combined their results using another STFT filter bank with a window size of 12 ms (we require good time resolution when combining results, but the frequency resolution is not as important since all of the noise reduction processing has already been done). The transience detector also operates with a window size of 12 ms.

The combination of results is performed according to the following formula:

$$X_{f,t} = \begin{cases} \alpha X_{f,t,2} + (1-\alpha) X_{f,t,3}, & f \le 4000Hz \\ \alpha X_{f,t,1} + (1-\alpha) X_{f,t,2}, & f > 4000Hz \end{cases}$$

Here α depends on transience for a given bin of the STFT:

$$\alpha = \begin{cases} 0, & T[f,t] < T_1 \\ \dfrac{T[f,t]-T_1}{T_2-T_1}, & T_1 \le T[f,t] < T_2 \\ 1, & T[f,t] \ge T_2 \end{cases}$$

Here T_1 and T_2 are user-defined thresholds, and we have selected $T_2 = 2T_1$.

Such a mixing strategy uses 2 times better frequency resolution below 4 kHz (approximating the property of better low-frequency resolution of our hearing) and adapts the resolution to the local transience of the signal inside each critical band.

As a result of such adaptation of resolution (Fig. 14), we have achieved reduction of time-smearing artifacts without compromise in depth of noise reduction. The conclusions of our informal listening tests were also confirmed by an increase of S/N ratio for recordings restored with adaptive time-frequency resolution (Table 1).

Filter bank algorithm	**S/N ratio, dB**
Noisy recording	48.17
STFT, window 12 ms	50.87
STFT, window 25 ms	50.90
STFT, window 50 ms	50.74
Adaptive resolution	**51.14**

Table 1. S/N ratios after noise reduction of "guitar and castanets" sample.

6.2. Center channel extraction

A multiband center channel extraction algorithm is an improvement upon a widely used "karaoke" feature that subtracts left and right stereo channels to cancel the in-phase signals comprising the center of a stereo field. Some sound editors adopt an STFT-based algorithm for this task that attenuates those STFT coefficients whose magnitudes and phases are close in the left and right channels. This allows for a stereo result, which is impossible with a single-band algorithm.

Again, without going into details of the attenuation of STFT coefficients, we will describe the effects of STFT resolution on the resulting sound and propose an improvement with adaptive time-frequency resolution.

Good frequency resolution allows deeper suppression of harmonic signals (such as vocals) in the center channel. However poor time resolution results in smearing of transients in the reconstructed waveform after modification of STFT coefficients.

We propose that the time-frequency resolution is adapted according to transience of the signal as described in section 6.1. We use 3 parallel STFT-based center channel extractors and combine their results using a local transience estimate in critical bands.

As a result, we are able to get good frequency resolution on harmonic parts of a signal (which are typically to be

suppressed in the center channel) and good time resolution around transients (which prevents time-smearing). Our informal listening experiments verify the reduction of these artifacts.

7. CONCLUSION

We have demonstrated a general framework for effective multiresolution signal processing and analysis. This framework avoids several undesirable side effects of the STFT's fixed time-frequency resolution such as the smearing of events in both time and frequency. It allows signal processing and analysis to adapt its resolution according to a predetermined strategy or the analysis of local signal features. We have shown how this framework can be applied to the display of spectrograms, spectral subtraction algorithms for noise reduction and center channel extraction algorithms.

For more examples of these algorithms and applications, please see the demo web page established for this paper [7].

8. ACKNOWLEDGEMENTS

The authors would like to thank iZotope, Inc. for supporting the project, and Dr. Y.M. Bayakovski, supervisor of Alexey's research in the Graphics & Media Lab of Moscow State University.

9. REFERENCES

[1] B. Logan "Mel Frequency Cepstral Coefficients for Music Modeling" // Proceedings of International Symposium on Music Information Retrieval, 2000.

[2] T. Painter, A. Spanias "A Review of Algorithms for Perceptual Coding of Digital Audio Signals" // Proceedings of 13th International Conference on Digital Signal Processing, 1997, vol. 1, 2-4 July 1997, pp. 179-208.

[3] P. Grunwald "A Tutorial Introduction to the Minimum Description Length Principle" // Chapters 1 and 2 of "Advances in Minimum Description Length: Theory and Applications", MIT Press, April 2005, ISBN 0-262-07262-9.

[4] J. Bonada "Audio Time-Scale Modification in the Context of Professional Audio Post-production" // Research work for PhD program, Universitat Pompeu Fabra, Barcelona, 2002.

[5] J. Thiemann "Acoustic Noise Suppression for Speech Signals Using Auditory Masking Effects" // Ph.D. thesis, Department of Electrical & Computer Engineering, McGill University, Mont-real, Canada, July 2001.

[6] S. Canazza, G. De Poli, G.A. Mian, A. Scarpa "Real Time Comparison Of Audio Restoration Methods Based On Short Time Spectral Attenuation" // Proceedings of Conference on Digital Audio Effects (DAFx01), December 6-8 2001, Limerick, Ireland.

[7] Demo web page for proposed algorithms: http://www.izotope.com/tech/aes_adapt/