Audio Engineering Society

Convention Paper 6924

Presented at the 121st Convention
2006 October 5–8 San Francisco, CA, USA

Comparison of frequency-warped representations for source separation of stereo mixtures

Juan José Burred[1] and Thomas Sikora[1]

[1] *Communication Systems Group, Technical University of Berlin, Einsteinufer 17, D-10597, Berlin, Germany*

Correspondence should be addressed to Juan José Burred (burred@nue.tu-berlin.de)

ABSTRACT

We evaluate the use of different frequency-warped, nonuniform time-frequency representations for the purpose of sound source separation from stereo mixtures. Such transformations enhance frequency resolution in spectral areas relevant for the discrimination of the different sources, improving sparsity and mixture disjointness. In this paper, we study the effect of using such representations on the localization and detection of the sources, as well as on the quality of the separated signals. Specifically, we evaluate a constant-Q and several auditory warpings in combination with a shortest path separation algorithm and show that they improve detection and separation quality in comparison to using the Short Time Fourier Transform.

1. INTRODUCTION

Underdetermined Blind Source Separation (BSS) aims at extracting N sources $\mathbf{s} = (s_1[n], \ldots, s_N[n])^T$ by observing $M < N$ mixtures $\mathbf{x} = (x_1[n], \ldots, x_M[n])^T$. In the linear, anechoic and noiseless case, and if no delays are considered, the mixtures are described by an instantaneous mixing model, given by:

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{1}$$

where $\mathbf{A}$ is the $M \times N$ mixing matrix. We are particularly interested in the separation of musical signals, in which case most mixtures are available in stereo format ($M = 2$), and usually more than 2 instruments and voices are present ($N > 2$), thus resulting in an underdetermined problem. In this context, an instantaneous model is applicable with recordings using intensity stereo, or artificially mixed using panning. In both cases, the position in the stereo field is determined solely by the amplitude differences between sources.

In the determined case ($M = N$), source separation equals to the problem of estimating an inverse of the mixing matrix. If the sources are sta-

tistically independent and non-Gaussian, this can be achieved with Independent Component Analysis (ICA)[1]. Once $\mathbf{A}^{-1}$ has been estimated, the sources are directly extracted just by inverting eq. 1. However, in the underdetermined case $\mathbf{A}$ is unsquare and therefore not invertible, which means that both $\mathbf{A}$ and the source vector $\mathbf{s}$ must be estimated in order to solve the problem. Since it is hard to estimate $\mathbf{A}$ and $\mathbf{s}$ simultaneously, the problem is usually divided into a mixing matrix estimation stage and a source estimation or resynthesis stage.

Most algorithms for underdetermined separation are based on the assumption that the signals are sparse in some domain [2]. A signal is said to be sparse if most of its samples are zero or near to zero, which corresponds to a supergaussian probability density (such as a Laplacian density). In most cases, the sparser the sources are, the less they will overlap when mixed (i.e., the more disjoint will be their mixture), and consequently the easier will be their separation. Audio signals are not sufficiently sparse in the time domain [3, 4], and therefore must be converted to a sparser domain in order to obtain acceptable results.

To this end, the Short Time Fourier Transform (STFT) has been widely used to transform the mixtures before performing separation in the spectral domain [5, 4, 6]. However, the equal frequency resolution offered by the STFT is disadvantageous for the task of speech or music separation. The reason for this is, on the one hand, that speech and music signals concentrate most of their energy in the mid-lower part of the spectrum, and therefore ovelappings are more likely to occur in this area. On the other hand, musical notes follow a logarithmic frequency relationship that does not correspond with the linearly spaced subbands of a STFT spectrogram. Notes in the lower range often fall into the same subbands and will thus overlap.

To overcome this, the application of multiresolution analysis to source separation has been proposed, in particular through the usage of wavelets [7]. The wavelet transform provides a constant-Q, non-uniform time-frequency representation (often called *scalogram*, in opposition to the STFT *spectrogram*), with high frequency resolution for low frequencies and high time resolution for high frequencies. This decomposition is adequate for music signals, and resembles human auditory perception. In the cited work, it was shown to improve sparsity and therefore separation when compared to the STFT.

A different approach comes from the field of Computational Auditory Scene Analysis (CASA) [8], which imitates more closely the several stages of auditory perception (from the acoustical processing in ear to the neural and cognitive processes in the brain) in order to characterize mixtures and perform sound separation. Such systems employ more sophisticated, non-constant-Q frequency warpings derived from psychoacoustical scales, usually implemented as nonuniform auditory filter banks. An example of application of such a scale (the Equal Rectangular Bandwidth, ERB, scale) in the context of blind music source separation can be found in [9].

Our goal in the present work is to perform a thorough evaluation of a constant-Q and three auditory time-frequency representations (ERB, Bark and Mel) as front-ends for a specific underdetermined BSS algorithm. In a previous article [3], we showed that auditory representations increase the disjointness of the mixtures and are therefore appropriate for sparsity-based algorithms. In the present contribution, we measure the effects of such frequency warpings on the quality of the separated and resynthesized signals, as well as on the accuracy of the mixing matrix estimation. We use objective quality measures to compare their performance with the STFT.

2. SEPARATION ALGORITHM

Denoting the columns of the mixing matrix $\mathbf{A}$ by $\mathbf{a}_j$, we can rewrite eq. 1 as

$$\mathbf{x} = \sum_{j=1}^{N} \mathbf{a}_j s_j \tag{2}$$

This equation is valid for each sample of the time domain signals or for each time-frequency bin of the transformed signals. It becomes apparent that, if each mixture sample or bin is contributed only by one signal (i.e., $s_k \neq 0$ and $s_j = 0$ for all $j \neq k$), the point $\mathbf{x}$ will lie on the direction defined by vector $\mathbf{a}_k$ in the complex mixture space $\mathbb{C}^M$. In the more realistic case in which each mixture bin contains contributions from all sources, $\mathbf{x}$ will lie near the direction

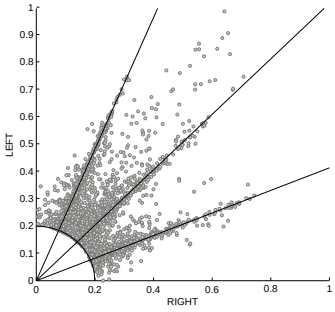

Fig. 1: Projected scatter plot for a 3-source, 2-channel mixture of musical instruments, transformed using a Constant-Q scaling.

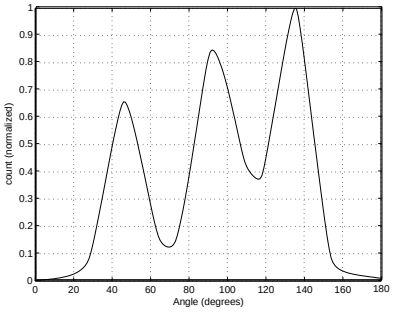

Fig. 2: Smoothed histogram for the scatter plot of Fig. 1.

$\mathbf{a}_k$ corresponding to the source s_k that predominates in that particular bin. If the mixture is sufficiently disjoint (i.e., if the bins of the individual sources are sufficiently non-overlapping), a scatter plot of all B samples/bins, defined by $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_B)$, will show the bins corresponding to a particular source concentrating around its direction. As an example, Fig. 1 shows the scatter plot of a 3-source, 2-channel musical mixture transformed into a Constant-Q domain. The direction clustering is more clear the more super-Gaussian (sparser) the probability distributions of the sources are. It should be noted that, in order for the direction vectors $\mathbf{a}_j$ to remain unchanged, the applied transformation must be linear in amplitude. All transformations used in the present work fulfill this condition.

In the case of instantaneous mixtures of sparse sources, the clustering phenomenon on the scatter plot is clear enough to allow the estimation of $\mathbf{A}$ by simple histogram analysis. Fig. 2 shows a smoothed histogram corresponding to Fig. 1. The directions $\mathbf{a}_j$ correspond to the peaks of the histogram, and the number of sources detected can be obtained by setting an appropriate peak threshold. Note that, since both real and imaginary, as well as positive and negative coefficients cluster around the same directions, it suffices to cluster in the first quadrant of $\mathbb{R}^2$, after the concatenation and projection defined by $\mathbf{X}_{proj} = (|Re\{\mathbf{x}_1\}|, \ldots, |Re\{\mathbf{x}_B\}|, |Im\{\mathbf{x}_1\}|, \ldots, |Im\{\mathbf{x}_B\}|)$.

Due to sparsity, most of the bins accumulate near to zero. However, bins with small modules do not add much information when searching for the mixing directions, and so they can be ignored for the histogram analysis. This spares computation time and does not affect the performance significantly. The optimization threshold is denoted in Fig. 1 by the circle around the origin.

For the source resynthesis stage, we use the *shortest path* approach, which has shown a robust performance with stereo anechoic mixtures [4]. For a given mixing matrix, it partitions the mixture space $\mathbb{R}^2$ into regions delimited by the mixing directions $\mathbf{a}_j$. Then, for each bin $\mathbf{x} = (x_1, x_2)$ at direction $\theta_x = \arctan(x_2/x_1)$, a 2×2 reduced mixing matrix $\mathbf{A}_r = [\mathbf{a}_a, \mathbf{a}_b]$ is defined, whose columns are the delimiting directions of the region it belongs to, i.e. $\theta_L = \arctan(a_{a2}/a_{a1})$ and $\theta_R = \arctan(a_{b2}/a_{b1})$ are the closest mixing directions to the left respectively to the right that enclose θ_x. Source estimation is performed inverting the determined 2×2 sub-problem and setting all other $N - M$ sources to zero:

$$\hat{\mathbf{s}}_r = \mathbf{A}_r^{-1}\mathbf{x} \quad (3)$$

$$\hat{s}_j = 0 \quad if\ j \neq a, b \quad (4)$$

It can be shown [4] that, if the sources are independent and assumed to follow a Laplace distribution, defined by

$$p(s_j) = \frac{\lambda}{2} e^{-\lambda |s_j|}, \quad (5)$$

the above method is equivalent to the ℓ_1-norm con-

strained minimization problem:

$$\hat{\mathbf{s}} = \underset{\mathbf{s}}{\operatorname{argmin}} \left\{ \ell_1(\mathbf{s}) = \sum_{j=1}^{N} |s_j| \;\middle|\; \mathbf{x} = \mathbf{A}\mathbf{s} \right\} \quad (6)$$

3. FREQUENCY-WARPED REPRESENTATIONS

A discrete time-frequency representation $X[n,k]$, where n is the time frame and k is the frequency or band index ($k = 0, 1, \ldots, L-1$), can be interpreted as the output of an L-channel bank of bandpass filters. Its frequency resolution is determined by the center frequencies of the filters f_k and by their bandwidths Δf_k. The individual impulse responses $h_k[n]$ of such a filter bank can be obtained by modulating and scaling a prototype impulse response $w_{N_k}[n]$ of length $N_k = f_s/\Delta f_k$, where f_s is the sampling rate:

$$h_k[n] = \frac{1}{N_k} w_{N_k}[-n] e^{j2\pi f_k n/f_s} \quad (7)$$

The discrete-time Short Time Fourier Transform (STFT), defined by

$$X^{stft}[n,k] = \sum_{m=0}^{N-1} x[nR+m] w[m] e^{-j2\pi km/N} \quad (8)$$

where $x[m]$ is the time domain signal, n is the frame index, N is the length of window $w[m]$ and R is the hop size, is equivalent to a bank of N filters equally spaced at the frequencies $f_k^{stft} = kf_s/N$, with constant bandwidth $\Delta f_k^{stft} = f_s/N$, with a prototype impulse response equal to the time-reversed analysis window $w[-n]$ and critically downsampled by a factor N. This follows from the interpretation of eq. 8 as the convolution $X[n,k] = x[n] * h_k[n]$.

If we impose the condition that all filters must have the same quality factor $Q = f_k/\Delta f_k$, we obtain a nonuniform spectral representation with subband center frequencies spaced geometrically according to

$$f_k^{cq} = f_0 2^{\frac{k}{b}} \quad (9)$$

and with bandwidths

$$\Delta f_k^{cq} = f_k^{cq}(2^{\frac{1}{b}} - 1) \quad (10)$$

where f_0 is the lowest central frequency and b is the number of filters per octave. The corresponding window lengths are given by $N_k^{cq} = Qf_s/f_k^{cq}$. In analogy to the STFT, imposing these conditions on eq. 7 results in the definition of the Constant-Q Transform (CQT) [10]:

$$X^{cq}[k] = \frac{1}{N_k^{cq}} \sum_{n=0}^{N_k^{cq}-1} x[n] w_{N_k^{cq}}[n] e^{-j\ 2\pi Qn/N_k^{cq}} \quad (11)$$

In this way, we obtain a logarithmic frequency warping which, for the same number of bands, has more frequency resolution in the low frequencies, and less frequency resolution in the high frequencies than the STFT (the inverse applies to time resolution).

More sophisticated frequency warpings can be defined to simulate more closely the nonuniform frequency resolution in the cochlea. Frequencies are mapped into a linear auditory scale according to experimental measurements. The resulting filters are equally spaced in the auditory scale, but nonuniformly spaced in frequency. In particular, the Bark scale [11] defines an analytical approximation to measurements of the *critical bands* of hearing, which are ranges in the basilar membrane where different frequencies interact:

$$\Delta f^{bark} = 25 + 75 \left[1 + 1.4 \left(\frac{f}{1000} \right)^2 \right]^{0.69} \quad (12)$$

The mapping to the auditory scale ξ^{bark} can be approximated by [12]

$$\xi^{bark} = 7 \operatorname{arcsinh}\left(\frac{f}{650} \right) \quad (13)$$

To obtain the warped filter bank, we sample the previous equation linearly between the values corresponding to DC and $f_s/2$ with the desired number of bands L, and then apply the inverse mapping to obtain the center frequencies:

$$f_k^{bark} = 650 \sinh\left(\frac{\xi_k^{bark}}{7} \right) \quad (14)$$

It should be noted that the original Bark definition of eq. 12 assumes a number of 24 critical bands. For the current application, however, a higher number of bands is needed to obtain an acceptable frequency segregation, and thus the range

corresponding to one Bark unit must be subdivided, and the filter bandwidths accordingly adapted. In this way, the final filter bandwidths are obtained as $\Delta f_k^{bark} = \Delta f^{bark}/B^{bark}$, where B^{bark} is the number of bands per Bark unit. The window lengths are, then, $N_k^{bark} = B^{bark} f_s/\Delta f^{bark}$.

The nominal bandwidth definition of the closely related ERB scale [11] is given by:

$$\Delta f^{erb} = 24.7 + \frac{f}{9.26} \tag{15}$$

The mapping to ERB units is:

$$\xi^{erb} = 9.26 \ln\left(\frac{1}{228.7} f + 1\right) \tag{16}$$

After linear sampling to ξ_k^{erb}, the inverse mapping is

$$f_k^{erb} = 228.7 \exp(^{\xi_k^{erb}/9.26} - 1) \tag{17}$$

and, again, the actual filter bank bandwidths are $\Delta f_k^{erb} = \Delta f^{erb}/B^{erb}$, where B^{erb} is the number of bands per ERB unit.

The Mel scale [11] was derived from the nonlinear perception of pitch ratios and, in contrast to the ERB and Bark scales, it is not defined in terms of bandwidths, but as a direct mapping between frequencies and mel units ξ^{mel}:

$$\xi^{mel} = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{18}$$

The sampled inverse mapping is:

$$f_k^{mel} = 700\left(10^{\xi_k^{mel}/2595} - 1\right) \tag{19}$$

The bandwidth per Mel unit (which is much smaller than an ERB or a Bark unit) can be obtained as $\Delta f^{mel} = df^{mel}/d\xi^{mel}$, which gives the relationship

$$\Delta f^{mel} = \frac{1}{1127}(700 + f) \tag{20}$$

and, finally, $\Delta f_k^{mel} = \Delta f^{mel}/B^{mel}$. The Mel scale is well-known in audio analysis applications as the warping stage of the Mel Frequency Cepstral Coefficients (MFCC) algorithm.

Fig. 3 compares the distribution of center frequencies versus subband number, for all transformations defined above, and for the particular case $f_s = 16$kHz and $L = 257$.

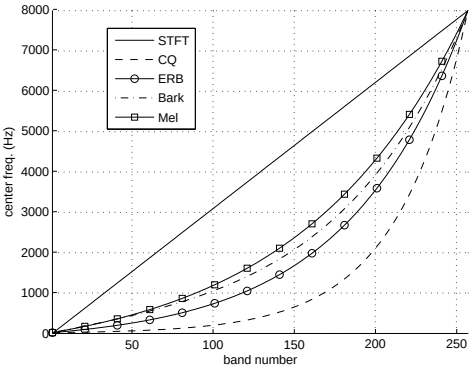

Fig. 3: Filter bank center frequencies as a function of band number, for 16 kHz sample rate and 257 bands.

A warped auditory representation $X[n, k]$ of signal $x[n]$ can thus be obtained by applying one of the previous definitions of f_k and N_k to the filter bank definition of eq. 7. In our experiments, we used a Hann window as the prototype impulse response for all transformations.

It must be noted that, in order to resynthesize the sources separated in a transformed domain, the employed transformation must be invertible. Unlike the STFT, nonuniform filter banks cannot be perfectly inverted. However, perfect reconstruction is not critical in source separation, since the largest reconstruction errors are introduced by the separation algorithm itself and are much more significant than the errors introduced in inverting the transformation [13]. Directly implemented warped filter banks can be approximately inverted by re-filtering with the time-reversed analysis filters and adding the subbands with appropriate weighting according to their bandwidth [14].

We used a direct, downsampled implementation of the filter banks. This method is computationally inefficient in comparison with the STFT. However, there exist more efficient implementations of frequency-warped filter banks using chains of all-pass filters [15], which can be combined with analytical expressions of the all-pass coefficient in such a way that the warping approximates an auditory warping [16].

4. PERFORMANCE EVALUATION

The estimation performance of the mixing matrix was measured by the percentage of experiments in which the correct number of sources were detected (detection rate, DR), and by the angular error e_{ang} between original directions $\mathbf{a}_j$ and their predictions $\hat{\mathbf{a}}_j$, averaged across each source and across each experiment.

To evaluate the quality of the estimated, separated signals $\hat{s}_j$ we used two objective measures: the Source to Distortion Ratio (SDR) and the Source to Artifacts Ratio (SAR).

The SDR is an overall measure of all distortions and errors introduced in the process, including errors by interference with the undesired signals, artifacts introduced by the separation algorithm and distortion due to imperfect transform inversion. For each source, it is given by:

$$\mathrm{SDR}_j = 10\log_{10}\frac{\|s_j\|^2}{\|\hat{s}_j - s_j\|^2} \qquad (21)$$

where $\|\cdot\|$ denotes the ℓ_2 norm (energy). We consider differences in scaling as irrelevant for the evaluation of separation quality, and therefore it must be assured that the signals are normalized prior to evaluation.

The artifacts introduced by the separation algorithm are often the main cause of distortion in source separation. Many algorithms, such as the one used here, introduce many artificial zeros in the spectrum before signal resynthesis (see eq. 4), which causes the so-called *musical noise* or *burbling* artifacts. For this reason, we also used a measure that specifically evaluates the influence of the artifacts on the separation quality, isolating it from the other sources of error, namely the Source to Artifacts Ratio (SAR). The computation of the SAR is more complex and involves orthogonal projections. The process is explained in detail in [17]. For each experiment run, the averaged values of SDR and SAR across all sources will be computed.

5. EXPERIMENTAL SETUP AND RESULTS

For the evaluation experiments, we used 10 stereo mixtures of $N = 3$ sources and 10 stereo mixtures of $N = 4$ sources. The sources to be mixed were randomly extracted from a database of 3 second musical fragments played by melodic instruments and sampled at 8 kHz. For each mixture, the experiment was repeated for each previously described time-frequency representation (STFT, constant Q (CQ), ERB, Bark and Mel), and for a different number of representation bands L_p, ranging from $L_1 = 33$ to $L_P = 4097$. Note that for real signals, the $N/2$ upper spectral bins of the STFT are redundant, and thus an N-points STFT corresponds to a spectrogram representation of $L = N/2 + 1$ bands (positive frequencies plus DC value). For this reason, we choose the values $L_p = N_{min}2^{p-1} + 1$ where $p = 0, 1, \ldots, P-1$ as evaluation points, where N_{min} is a power of two to benefit from an efficient FFT computation (in this case, $N_{min} = 64$ and $P = 8$). This makes a total number of 800 separation experiments.

Representation	$N = 3$ sources		$N = 4$ sources	
	DR (%)	$e_{ang}(^\circ)$	DR (%)	$e_{ang}(^\circ)$
STFT	81.3	1.22	65.0	3.38
CQ	80.0	0.75	67.5	4.82
ERB	82.5	0.76	71.3	0.83
Bark	82.5	0.78	73.8	0.90
Mel	82.5	0.76	71.3	1.50

Table 1: Evaluation of the mixing matrix estimation stage: averaged source detection rate (DR) and angular error (e_{ang}) in degrees, for stereo mixtures of $N = 3$ (left) and $N = 4$ sources (right).

Each source was normalized, artificially panned and mixed. The mixing matrix was defined with equally spaced directions, i.e., $\theta_1 = 3\pi/4$, $\theta_2 = \pi/2$ and $\theta_3 = \pi/4$ for $N = 3$ and $\theta_1 = 4\pi/5$, $\theta_2 = 3\pi/5$, $\theta_3 = 2\pi/5$ and $\theta_4 = \pi/5$ for $N = 4$, where 0 corresponds to hard right and π to hard left. To find the direction clusters, the scatter plot was rastered using a radial grid with 0.5° resolution. The resulting values for DR and e_{ang} are shown on Table 1. For $N = 3$, the DR does not improve significantly, but e_{ang} has been nearly halved. The $N = 4$ problem is more difficult, as expected, but the performance difference to the STFT has been increased. In particular, the angular error has been reduced by a factor of 4 with the ERB and Bark warpings.

Figures 4 and 5 show the results of the evaluation of the source estimation stage, specifically, the SDR

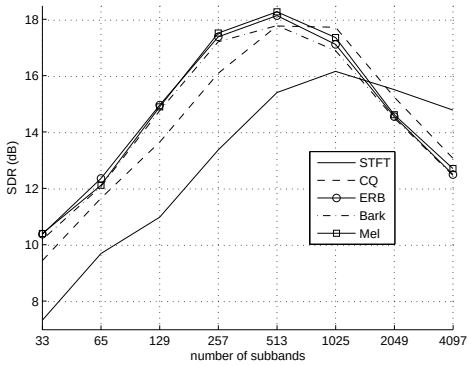

Fig. 4: Evaluation of the source resynthesis stage: Source to Distortion Ratio (SDR) as a function of number of subbands L, for stereo mixtures of $N = 3$ sources.

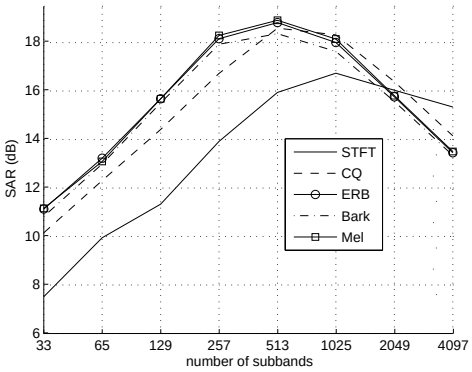

Fig. 5: Evaluation of the source resynthesis stage: Source to Artifacts Ratio (SAR) as a function of number of subbands L, for stereo mixtures of $N = 3$ sources.

Repr.	$N = 3$ sources		$N = 4$ sources	
	SDR (dB)	SAR (dB)	SDR (dB)	SAR (dB)
STFT	16.18	16.74	10.61	10.81
CQ	17.79	18.59	12.77	13.42
ERB	18.16	18.83	13.01	13.53
Bark	17.81	18.36	12.83	13.34
Mel	18.30	18.92	13.08	13.57

Table 2: Evaluation of the source resynthesis stage: maximum achieved SDR and SAR for for stereo mixtures of $N = 3$ (left) and $N = 4$ sources (right).

and SAR as a function of number of filter bank subbands. Table 2 shows the maximal achieved values in the curves. It can be seen that performance has been increased in all cases in comparison to the STFT. Again, the improvement is larger with 4 than with 3 sources. All nonuniform representations reach the highest performance with $L = 513$ frequency bands. In contrast, the STFT has its peak at $L = 1025$. For higher number of bands, all curves begin to decrease. This is due to the fact that global time resolution decreases as L grows, and thus time-domain overlaps are stronger. All three auditory warpings (ERB, Bark and Mel) showed similar behaviors, with Mel obtaining a slightly better performance in both measures, followed by ERB and Bark. However, the difference with CQ is greater. As can be seen in Fig. 3, CQ is the transformation offering the highest frequency resolution in the low frequency area. It turns out that a more equilibrated trade-off between low frequency and high frequency resolution, as offered by the auditory warpings (whose warping curves lie in the middle region between CQ and STFT) is advantageous for our purposes.

6. CONCLUSIONS AND FUTURE WORK

It has been shown that the usage of frequency-warped representations as front-end for underdetermined sound source separation improves the performance of both stages of the algorithm: mixing matrix estimation and ℓ_1-norm minimization-based source estimation, when compared to using the STFT. They improved all objective measures evaluated: source detection rate, angular error, resynthesis error due to artifacts and overall distortion error. Also, auditory warpings like ERB, Bark or Mel perform better than constant-Q warpings and offer an optimal trade-off between resolution in low and high frequencies. These results further support the convenience of combining psychoacoustical knowledge with mathematical separation algorithms in the form of hybrid CASA/BSS systems.

The separation performed by the system treated here is solely based on spatial information and on a broad sparsity assumption on the sources (a Laplacian distribution). In order to further improve performance and robustness, more sophisticated knowledge about the nature of the sources must be added.

This can take the form of source-dependent models of spectral content or temporal structure. This is particularly important in the case of real performances of tonal music, where the overlapping of partials in the same band is very likely to occur, even in a frequency-warped scale. This issue will be the main point in our future research.

Sound examples corresponding to the experiments can be found under `www.nue.tu-berlin.de/research/projects/sourcesep/demo1/`

7. ACKNOWLEDGMENT

This research was supported by the European Commission under contract FP6-027026-K-SPACE.

8. REFERENCES

[1] J. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley & Sons, 2001.

[2] P. D. O'Grady, B. A. Pearlmutter, and S. T. Rickard, "Survey of Sparse and Non-Sparse Methods in Source Separation," *International Journal of Imaging Systems and Technology*, vol. 15, no. 1, 2005.

[3] J.J. Burred and T. Sikora, "On the Use of Auditory Representations for Sparsity-Based Sound Source Separation," *Proc. Int. Conf. on Information, Communications and Signal Processing (ICICS), Bangkok, Thailand*, December 2005.

[4] P. Bofill and M. Zibulevsky, "Underdetermined Blind Source Separation Using Sparse Representations," *Signal Processing*, vol. 81, 2001.

[5] Ö. Yilmaz and S. Rickard, "Blind Separation of Speech Mixtures via Time-Frequency Masking," *IEEE Trans. on Signal Processing*, vol. 52, No. 7, July 2004.

[6] R. Balan and J. Rosca, "Statistical Properties of STFT Ratios for Two Channel Systems and Applications to Blind Source Separation," *Proc. Int. Workshop Independent Component Analysis and Blind Source Separation, Helsinki, Finland*, June 2000.

[7] P. Kisilev, M. Zibulevsky, and Y.Y. Zeevi, "A Multiscale Framework For Blind Separation of Linearly Mixed Signals," *Journal of Machine Learning Research*, vol. 4/7-8, 2003.

[8] D.F. Rosenthal, H.G. Okuno, and (Eds.), *Computational Auditory Scene Analysis*, Lawrence Erlbaum Assoc., May 1998.

[9] E. Vincent and X. Rodet, "Underdetermined Source Separation with Structured Source Priors," *Proc. International Conference on Independent Component Analysis and Blind Signal Separation (ICA), Granada, Spain*, September 2004.

[10] J.C. Brown, "Calculation of a Constant Q Spectral Transform," *J. Acoust. Soc. Am.*, vol. 89(1), January 1991.

[11] E. Zwicker and H. Fastl, *Psychoacoustics. Facts and Models*, Springer, 1990.

[12] M.R. Schroeder, B.S. Atal, and J.L. Hall, "Optimizing digital speech coders by exploiting masking properties of the human ear," *Journal of the Acoustical Society of America*, vol. 66, 1979.

[13] E. Vincent, *Modèles d'Instruments pour la Séparation de Sources et la Transcription d'Enregistrements Musicaux*, PhD Thesis, Université Paris VI, 2004.

[14] M. Slaney, D. Naar, and R. F. Lyon, "Auditory Model Inversion for Sound Separation," *Proc. IEEE ICASSP*, 1994.

[15] A. Härmä, M. Karjalainen, L. Savioja, V. Välimäki, U.K. Laine, and J. Huopaniemi, "Frequency-Warped Signal Processing for Audio Applications," *J. Audio Eng. Soc.*, vol. 48, No. 11, November 2000.

[16] J.O. Smith and J.S. Abel, "Bark and ERB Bilinear Transforms," *IEEE Trans. on Speech and Audio Processing*, pp. 697–708, November 1999.

[17] R. Gribonval, L. Benaroya, E. Vincent, and C. Févotte, "Proposals for Performance Measurement in Source Separation," *4th Int. Symposium on Independent Component Analysis and Blind Signal Separation (ICA), Nara, Japan*, April 2003.