On the Use of Auditory Representations for Sparsity-Based Sound Source Separation

Juan José Burred and Thomas Sikora
Communication Systems Group
Technical University of Berlin, Germany
{burred,sikora}@nue.tu-berlin.de

***Abstract*— Sparsity-based source separation algorithms often rely on a transformation into a sparse domain to improve mixture disjointness and therefore facilitate separation. To this end, the most commonly used time-frequency representation has been the Short Time Fourier Transform (STFT). The purpose of this paper is to study the use of auditory-based representations instead of the STFT. We first evaluate the STFT disjointness properties for the case of speech and music signals, and show that auditory representations based on the Equal Rectangular Bandwidth (ERB) and Bark frequency scales can improve the disjointness of the transformed mixtures.**

***Keywords*—source separation, auditory scales, sparse signals, mixture disjointness.**

I. Introduction

Most approaches towards sound separation can be classified into two broad types: Computational Auditory Scene Analysis (CASA) and Blind Source Separation (BSS).

The CASA approach consists of studying and imitating the human hearing system and its ability to perceive and understand sound entities that are present in perceived sound complexes, which in this context are called *auditory scenes*. To this end, computational models have been developed that mimic the several stages of auditory perception from the acoustical processing in the ear to the neural and cognitive processes in the brain.

BSS approaches take into account statistical properties of the sources, the mixtures or the mixing processes and attempt to solve the separation problem from a purely mathematical point of view. As the term *blind* denotes, the mixing process and the sources are unknown. However, it is possible (and in fact absolutely necessary) to make assumptions about their statistical nature.

Both kind of approaches have been successful in different, specific separation scenarios [1] but up to now very little work has been done in combining their respective advantages. Hybrid CASA/BSS systems are believed to have the potential to improve separation reliability with a wider range of mixture types.

Assuming that the sources are mixed linearly and that no reverberation and no noise are present in the mixing process, the problem formulation is to find the vector of N sources $\mathbf{s} = (s_1[n], \ldots, s_N[n])^T$ from the observation of the vector of M mixtures $\mathbf{x} = (x_1[n], \ldots, x_M[n])^T$ given by

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

where $\mathbf{A}$ is the $M \times N$ mixing matrix that describes the contributions of each source to each mixture. When there are the same number of sources than mixtures ($M = N$), the separation is said to be *determined* or *complete*. If the sources are statistically independent and non-gaussian, a determined mixture can be successfully unmixed by Independent Component Analysis (ICA), which consists of estimating the mixing matrix and then inverting it to directly solve eq. 1.

However, in many practical situations, the *underdetermined* or *overcomplete* case ($M < N$) is more common. In this case, the mixing matrix is not invertible and therefore common ICA algorithms are not applicable; furthermore, estimating $\mathbf{A}$ and $\mathbf{s}$ become two different problems.

II. Sparsity and Disjointness

Most of the underdetermined BSS algorithms for the estimation of $\mathbf{A}$ and $\mathbf{s}$ are based on the assumption that the involved signals have some degree of sparsity. A signal is said to be sparse if most of its components are zero or near to zero. Sparsity is desirable for source separation because, the more sparse a signal is, the less it will overlap with the other signals in a mixture (unless the signals follow identical probability distributions). The *disjointness* of a mixture can be defined as the degree of non-overlapping of the mixed signals. In most cases a higher sparsity will result in a higher disjointness and in an easier separation.

Most audio mixtures are not sufficiently disjoint in the time domain, and therefore it is often required to transform the signals into a sparser or more disjoint domain, such as a time-frequency (t-f) representation, perform separation in the transformed domain, and resynthesize back the estimated sources into the time domain. This approach is followed in previous works [2], [3] using the Short Time Fourier Transform (STFT) as t-f representation. In this paper we show that it is possible to further improve disjointness replacing the STFT with a t-f representation in which the frequency axis has been warped following an auditory frequency scale. This is due to the fact that auditory scales emphasize resolution in the mid-low frequency range, where most sound energy is usually concentrated [4]. As representatives of auditory

0-7803-9282-5/05/$20.00 ©2005 IEEE

frequency warpings, the Equal Rectangular Bandwidth (ERB) and the Bark scales have been chosen.

III. Measuring disjointness

To measure the disjointness D of a mixture we use the *W-disjoint orthogonality* (WDO) criterion for t-f representations [2], which relies on the concept of unmixing by binary masking. If a mixture is sufficiently disjoint in some t-f domain, it can be used to estimate a set of unmixing masks, one for each source, that will approximately extract the desired source when applied on the mixture representation. The key idea behind the WDO-based measurement method is that the unmixing capabilities of a set of ideal masks computed from the knowledge of the sources can be also interpreted as the intrinsic disjointness of the mixture.

To measure the WDO of a set of N sources $s_j[n]$ we first define $y_j[n]$ as the sum of all signals interfering with source j:

$$y_j[n] = \sum_{\substack{i=1 \\ i \neq j}}^{N} s_i[n] \tag{2}$$

Let $S_j[n,k]$ denote a discrete t-f representation of signal $s_j[n]$, with n the time and k the frequency index. The ideal binary unmixing mask for source j is defined as

$$M_j[n,k] = \begin{cases} 1, & 20\log\frac{|S_j[n,k]|}{|Y_j[n,k]|} \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

i.e., $M_j[n,k]$ is the indicator function of the t-f bins to which source j contributes more than all its interfering sources. The *preserved-signal ratio* (PSR) measures the energy loss of the desired signal after unmixing:

$$\mathrm{PSR}_j = \frac{\|M_j[n,k] \circ S_j[n,k]\|^2}{\|S_j[n,k]\|^2} \tag{4}$$

where $\|\cdot\|^2$ denotes the squared $_2$ norm (energy) and $\circ$ denotes the Hadamard product (element-wise product). The *signal-to-interference ratio* (SIR) measures the energy difference between the desired signal and its interference after applying the mask:

$$\mathrm{SIR}_j = \frac{\|M_j[n,k] \circ S_j[n,k]\|^2}{\|M_j[n,k] \circ Y_j[n,k]\|^2} \tag{5}$$

The WDO for that particular source is defined as

$$\mathrm{WDO}_j = \mathrm{PSR}_j - \frac{\mathrm{PSR}_j}{\mathrm{SIR}_j} \tag{6}$$

The disjointness of the mixture of the sources can then be measured as their averaged WDO: $D = \overline{\mathrm{WDO}}$. A perfect disjointness (each bin is contributed only by one source) corresponds to $\overline{\mathrm{PSR}} = 1$, $\overline{\mathrm{SIR}} = \infty$ and $\overline{\mathrm{WDO}} = 1$ and would result in perfect separation.

WDO can also be used as a BSS performance measure when based on masks estimated from the mixtures without knowing the sources. However, it should be noted that in the case studied here, the above masks are derived with the sources being known, which implies that the definition of WDO used here can be interpreted as the upper bound in unmixing performance by binary masking.

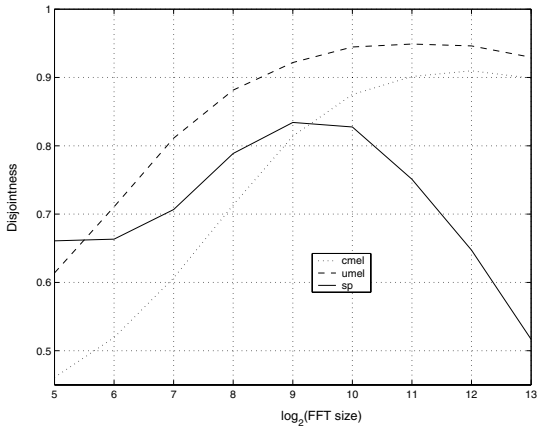


Fig. 1. Disjointness against FFT size for 3-source mixtures and 8 kHz sampling rate.

IV. Disjointness properties of the STFT

We first performed several experiments to evaluate the behavior of the STFT regarding disjointness with speech and music signals. Because of their different time and spectral characteristics, speech and music signals are expected to behave differently, and the validity of the sparseness/disjointness assumption will vary depending on the signal type.

Throughout our experiments, we used three different audio datasets, each containing 1 second fragments sampled at 8 kHz. Dataset SP contains a collection of 50 speech utterances. Dataset UMEL contains 50 fragments of instrumental solos playing uncorrelated melodies, i.e., melodies randomly drawn from an instrumental database which are not intended to be musically coherent when mixed. To evaluate disjointness, 50 different combinations of 3 sources were randomly extracted from each of these two databases and mixed. Dataset CMEL (for "correlated melodies") contains 50 sets of 3 instrumental fragments extracted from a real multitrack recording, in such a way that the resulting mixtures constitute excerpts from a coherent musical performance (in this case a saxophone quintet). Although uncorrelated musical mixtures are often employed for evaluating the performance of source separators, a correlated dataset such as CMEL simulates closer the requirements of a practical musical unmixing application.

Fig. 1 shows the disjointness as defined above averaged over all samples from each database as a function of the FFT size for 3-source mixtures. For the computation of the STFT, a Hann window was used.

It can be seen that a higher disjointness is possible for music signals than for speech signals. This can be explained by the fact that speech signals concentrate their energy in a narrower part of the spectrum, and therefore spectral overlappings are more likely to occur. Also, speech disjointness suffers from

	SP	UMEL	CMEL
% D time	70.8	55.3	51.1
% D_{max} STFT	83.4	94.8	90.9
optimal FFT size	512	2048	4096

TABLE I

COMPARISON OF DISJOINTNESS IN THE TIME AND IN THE STFT DOMAIN, AND OPTIMAL FFT SIZES AT 8 KHZ SAMPLE RATE.

the reduced temporal resolution when increasing the size of the window. It turns out that for speech signals, a compromise should be taken to balance temporal and frequency disjointness by choosing a moderate window size (512 for 8 kHz sampling frequency), whereas for music signals, frequency disjointness plays a more important role than time disjointness and so frequency resolution should be favored.

Mixtures of correlated melodies are less disjoint than uncorrelated ones because of the higher amount of spectral and temporal overlapping. Their disjointness is expected to vary strongly according to music type. It will be lower for tonal and homophonic than for atonal and contrapuntal styles. The saxophone quintet used for our experiments is tonal and highly homophonic (the voices in a homophonic musical texture change notes at the same time, whereas in a contrapuntal style, each voice behaves more independently).

The disjointness in the time domain was also measured. Table I compares the averaged time-domain disjointness with the maximum achievable t-f domain disjointness, and shows the benefits of transforming into the t-f domain. The table also lists the FFT size for which these maximum disjointness are reached when working with 8 kHz sample rate.

V. AUDITORY TIME-FREQUENCY REPRESENTATIONS

Any time-frequency representation with N frequency bands $S[n, k]$ can be viewed as the output of an N-channel filter bank. In the case of the STFT, the center frequencies of the filters are linearly spaced along the frequency axis at the positions $f_{k,STFT} = \frac{k f_s}{N}$, where f_s is the sampling rate, and their frequency responses are modulated versions of the frequency response of the analysis window. Thus, the STFT provides an equal resolution for all frequencies.

Auditory time-frequency representations are obtained by using non-uniform filter banks in which the center frequencies follow a non-linear scale defined to simulate the frequency selectivity of the inner ear. Their resolution is approximately linear at low frequencies and increasingly logarithmic at high frequencies. For our experiments we chose two well-known auditory frequency scales: the ERB and the Bark scale.

The ERB scale is considered one of the most accurate models of the frequency resolution of the basilar membrane [5]. It defines the Equal Rectangular Bandwidth of the auditory filters as a function of frequency as

$$\Delta f_{ERB} = B_{min} + \frac{f}{Q_a} \quad [Hz] \tag{7}$$

where B_{min} is the minimum bandwidth for low frequency channels and Q_a is the asymptotic quality factor to which the high frequency filters tend. The mapping between the frequency in Hz and an ERB scale in which the filters are linearly spaced is given by

$$x_{ERB} = \int \frac{1}{\Delta f_{ERB}}\, df = Q_a \ln\left(\frac{1}{Q_a B_{min}} f + 1\right) \tag{8}$$

In order to compute the center frequencies for a N-channel ERB filter bank, we divide the desired range in equal ERB intervals centered at $x_{k,ERB}$ and then apply the inverse mapping:

$$f_{k,ERB} = Q_a B_{min}(e^{x_{k,ERB}/Q_a} - 1) \quad [Hz] \tag{9}$$

Recommended values for the parameters are $Q_a = 9.26$ and $B_{min} = 24.7$ Hz.

The Bark scale, also called *critical band rate*, is defined by the following bandwidth function [6]:

$$\Delta f_{Bark} = 25 + 75\left(1 + 1.4\left(\frac{f}{1000}\right)^2\right)^{0.69} \quad [Hz] \tag{10}$$

Suitable approximations for the corresponding direct and inverse frequency warpings [7] are given by

$$x_{Bark} = 7 \operatorname{arcsinh}\left(\frac{f}{650}\right) \tag{11}$$

and

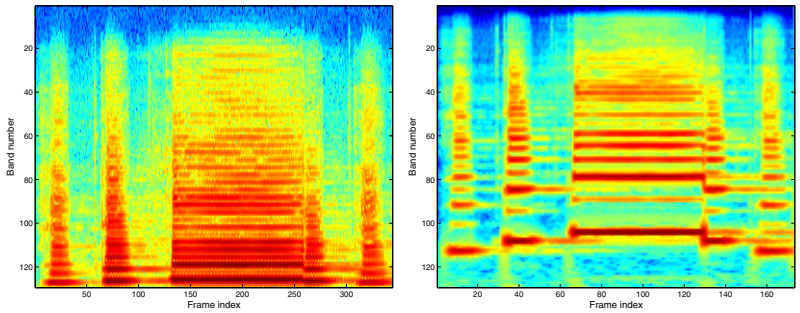
$$f_{k,Bark} = 650 \sinh\left(\frac{x_{k,Bark}}{7}\right) \quad [Hz], \tag{12}$$

where x_k is the center of the k-th Bark interval.

In this work we concentrate on studying the effects of the frequency-warping stage of auditory modeling, motivated by the previous results that show that spectral resolution is crucial in improving disjointness. For these reasons, auditory filter shapes have not been used. Instead, we use a Hann window as the prototype impulse response, as in the case of the STFT.

It should be noted that, in order for a transformation to be useful in the context of source separation, it must be invertible, so that the extracted sources can be synthesized back. Unlike the STFT, non-linear auditory filter banks cannot generally be reconstructed perfectly. However, perfect reconstruction is not critical in source separation, since the by far strongest signal distortions are introduced by the separation algorithm itself, and not by the transform inversion. If the overlapping between frequency responses of adjacent bands is moderate, it is possible to invert a non-uniform filter bank by weighting the bands in the synthesis stage accordingly, obtaining synthesized signals with inaudible error. Nevertheless, care must be taken in adjusting the channel spacing in order to minimize reconstruction error.

Fig. 2 shows the effect of the ERB frequency warping on the time-frequency representation of an excerpt of a clarinet playing a 5-note melody. Comparing the spectrogram (magnitude of the STFT) and an ERB representation it can be observed that, for the same number of bands, the resolution has been enhanced in the low frequency range, where most of the signal



(a) Spectrogram (b) ERB representation

Fig. 2. Comparison of 129-band spectrogram and ERB spectral representation of a clarinet melody.

energy is concentrated, and that the harmonic lines are more clearly visible and separated. This is the reason why auditory warpings have the potential to improve disjointness, a fact that will be confirmed in the evaluation experiments described in the next section.

The major drawback of filter bank based t-f representations is their high computational requirements. For a high number of bands, the needed computation power is much higher than using FFT-based spectrograms, even after applying downsampling according to the bandwidth of each channel.

VI. Disjointness properties of auditory representations

Although the WDO procedure outlined in sec. III was originally defined for the STFT, it can readily be applied with other t-f transformations, such as auditory transformations. However, it should be noted that it is only possible to compute the PSR, SIR and WDO values in the time-frequency domain if the corresponding transform obeys Parseval's theorem, i.e., if the signal energy in the frequency domain is proportional to the energy in the time domain. This is not the case for auditory transformations, which distribute signal energy unequally across the spectral bands, depending on the bandwidth and eventually amplitude weighting of each band. Therefore it is mandatory to invert the transform and compute the above quantities in the time domain.

We repeated the WDO experiments with the same datasets, averaging the results of 50 mixtures per dataset. The results are shown in figs. 3, 4 and 5. The disjointness of the ERB and Bark-transformed mixtures is plotted against the number of bands of the t-f representation and compared to the corresponding STFT curve of fig. 1. Note that an N-point STFT window corresponds to a $N/2 + 1$ band spectrogram representation.

It can be observed that both the Bark and ERB scales improve disjointness in all cases, with the improvement being more significant with speech data than with music data. As before, this can be explained by the higher energy concentration of speech in the spectral area where the highest resolution gain is achieved with the auditory warping. For speech signals, the disjointness is higher with auditory scales than with the STFT for all band resolutions. In the case of music signals, the gain in disjointness is higher when few bands are used and decreases as the number of bands increases. In the particular case of correlated melodies, the performance of large-window

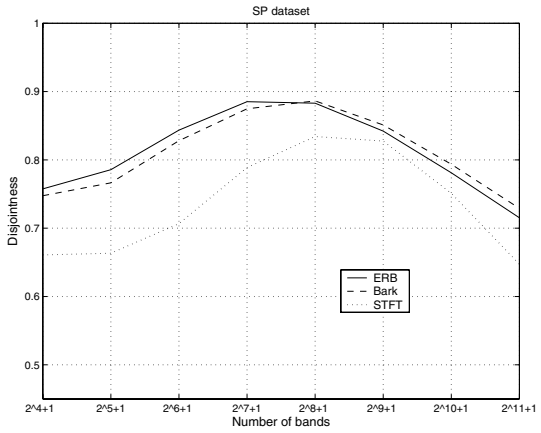


Fig. 3. Disjointness against number of bands for ERB, Bark and STFT representations for speech data, 3-source mixtures and 8 kHz sampling rate.

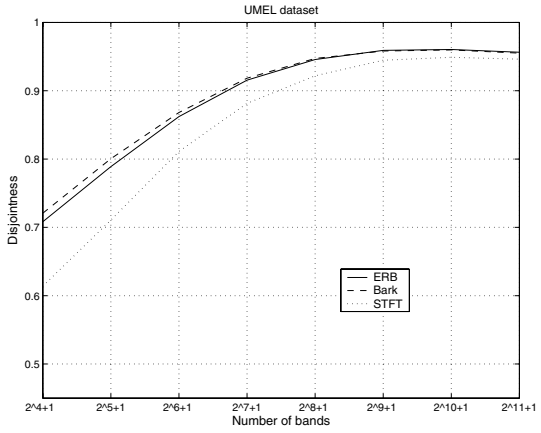


Fig. 4. Disjointness against number of bands for ERB, Bark and STFT representations for uncorrelated music data, 3-source mixtures and 8 kHz sampling rate.

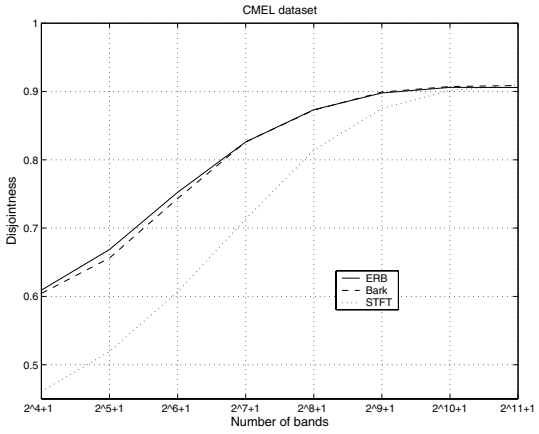


Fig. 5. Disjointness against number of bands for ERB, Bark and STFT representations for correlated music data, 3-source mixtures and 8 kHz sampling rate.

STFT and auditory scales is very similar.

VII. Conclusions

Combining principles from both the BSS and CASA separation methods, we have shown that the use of auditory-based time-frequency representations improves the mixture disjointness in comparison with usual equal-resolution methods like the STFT and thus can facilitate separation by time-frequency based algorithms. The improvement is particularly significant for speech signals. For music signals, and especially for mixtures of correlated melodies, further investigation is needed in order to obtain higher disjointness gains.

Acknowledgment

We thank Emmanuel Vincent for the helpful comments and Jan-Mark Batke for providing the multitrack recording used in the experiments.

The work presented was developed within VISNET, a European Network of Excellence, funded under the European Commission IST FP6 programme.

References

[1] A. van der Kouwe, D. Wang, and G. Brown, "A comparison of auditory and blind separation techniques for speech segregation," *IEEE Trans. on Speech and Audio Processing*, vol. 9, No. 3, March 2001.

[2] 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.

[3] P. Bofill and M. Zibulevsky, "Underdetermined blind source separation using sparse representations," *Signal Processing*, vol. 81, 2001.

[4] 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.

[5] B. Moore and B. Glasberg, "A revision of Zwicker's loudness model," *Acta Acustica*, vol. 82, 1996.

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

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