RESTORATION OF AUDIO DOCUMENTS WITH LOW SNR: A NMF PARAMETER ESTIMATION AND PERCEPTUALLY MOTIVATED BAYESIAN SUPPRESSION RULE

Giuseppe Cabras
Dep. of Electrical, Management and Mechanical Engineering
University of Udine, Italy
giuseppe.cabras@uniud.it

Sergio Canazza
Sound and Music Computing Group,
Dep. of Information Engineering
University of Padova, Italy
canazza@dei.unipd.it

Pier Luca Montessoro, Roberto Rinaldo
Dep. of Electrical, Management and Mechanical Engineering
University of Udine, Italy
{montessoro, rinaldo}@uniud.it

ABSTRACT

In the field of audio restoration, the most popular method is the Short Time Spectral Attenuation (STSA). Although this method reduces the noise and improves the SNR, it mostly tends to introduce signal distortion and a residual noise called musical noise (a tonal, random, isolated, time-varying noise). This work presents a new audio restoration algorithm based on Non-negative Matrix Factorization (NMF) with a noise suppression rule that introduce the masking phenomenon of the human hearing to calculate a noise masking threshold from the estimated target source. Extensive test with PESQ measure at low SNR (i.e. < 10dB) show that the method does not introduce musical noise and permits to control the trade-off between undesired component suppression and source attenuation. In particular, we show that NMF is a suitable technique to extract the clean audio signal from undesired non stationary noise in a monaural recording of ethnic music. Moreover, we carry out a listening test in order to compare NMF with the state of the art audio restoration framework using the EBU MUSHRA test method. The encouraging results obtained with this methodology in the presented case study support their applicability in several fields of audio restoration.

1. INTRODUCTION

The ethnic-musical heritage – often the only testimonial of past oral cultures – is in danger of disappearing: the audio documents were usually recorded in non-professional carriers by means of amateur recording system. Thus, for their appropriate fruition and/or for a suitable use of Music Information Retrieval techniques it is necessary to process the signals by means of audio restoration algorithms.

Different strategies can be adopted in a combined way with audio restoration algorithms, in accordance with the final purposes of the access copy:

- Documental approach: in this case, the de-noising algorithms only concern the cases in which the internal evidence of the degradation is unquestionable, without going beyond the technological level of that time.
- Aesthetical approach: it pursues a sound quality that matches the actual user's expectations (for both new commercial editions and to arrange the signal before the use of MIR techniques).
- Sociological approach: it has the purpose of obtaining a historical reconstruction of the recording as it was listened to at the time (see Storm, Type I [1]).
- Reconstructive approach: it has the objective of preserving the intention of the author (see Storm, Type II [1])

In order to reach one or more of the above aims, it is necessary to have at disposal several audio restoration instruments (often in the same audio document there are corruptions with different physical characteristics, that can be attenuated with different de-noise filters). The audio restoration algorithms can be divided into three categories [2]:

1. frequency-domain methods, such as various forms of non-casual Wiener filtering or spectral subtraction schemes and recent algorithms that attempt to incorporate knowledge of the human auditory system; these methods use little a priori information;
2. time-domain restoration by signal models such as Extended Kalman Filtering (EKF): in these methods a lot of a priori information is required in order to estimate the statistical description of the audio events;
3. restoration by source models: only a priori information is used.

The advantage of frequency-domain methods is that they are straightforward and easy to implement. However, the limitations are as follows: musical noise (short sinusoids randomly distributed over time and frequency) is unavoidable; the results depend on a good noise estimation. Restoration by source model is limited to very few cases (e.g. only monophonic recordings) and it is not generalizable. The EKF is able, in principle, to simultaneously solve the problems of filtering, parameter tracking and elimination of the

outliers, but it is very sensitive to parameter setting and ineffective where the Signal-to-Noise Ratio (SNR) is very low (< 10dB), as happen in many ethnic music audio documents.

This work presents a new audio restoration method – that fall within the first category – based on Non-negative Matrix Factorization (NMF), an emerging new technique in the blind extraction of signals recorded in a variety of different fields. The application of NMF to the analysis of monaural recordings is relatively recent. We show that NMF is a suitable technique to extract the clean audio signal from undesired non stationary noise in a monaural recording of ethnic music. More specifically, based on finding by Wolfe and Godsill [3], we develop a perceptually motivated distortion measure as a generalization of the Minimum Mean Square Error (MMSE) cost function that incorporates the masking threshold. Moreover, we carry out a listening test in order to compare NMF with the state of the art audio restoration framework using the EBU MUSHRA test method.

A recent approach to separate an acoustic source is provided by Non-negative Matrix Factorization (NMF). The basic idea is that we can obtain a meaningful *part-based* factor decomposition [4] from a data observation (e.g., the monaural recording) by the only constrain of non-negativity and sparsity, since no cancellation of factors can occur and only additive combinations are permitted. The use of sparse code can favor a factorization where only a few dictionary elements are used to model the source, introducing an ℓ_1 norm penalty term on the coefficients of the code matrix, which explicitly enforces sparseness [5]. However, a further non trivial step is needed to assign the decomposed parts to the source of interest (e.g., the original audio signal) to discard the interference source (e.g., the corrupting noise). The proposed approach tries to solve this problem with a solution based on an extended Non-negative Matrix Factorization algorithm and prior knowledge on interference. In addition, our approach reduces both distortion and perceptually annoying musical noise by taking into account the masking phenomenon of the human hearing, in order to calculate a noise masking threshold from the estimated target source.

We apply this method to improve the quality of noisy recordings of ethnic music on Shellac 78 rpm phonographic discs. The Shellac disc is a common audio mechanical carrier, where the audio information is recorded by means of a groove cut into the surface by a stylus modulated by the sound, either directly in the case of acoustic recordings or by electronic amplifiers. There are more than 1,000,000 Shellac discs in the worldwide audio archives containing music never re-recorded (R&B, Jazz, Ethnic, Western classical, etc.).

The rest of this paper is organized as follows. Sec. 2 details the proposed audio restoration method: in particular, Sec. 2.5 introduces perceptually motivated Bayesian suppression rules used. In order to validate the system, we carry out a listening test – using ethnic music audio documents – in order to compare NMF with the state of the art audio restoration framework using the EBU MUSHRA test method (Sec. 3). Final conclusions are drawn in Sec. 4.

2. AUDIO ENHANCEMENT FRAMEWORK

The objective of the proposed method is to estimate the undesired components, or interference, $n(t)$ and the source of interest, or target, $s(t)$ directly from the observable data mix (i.e. in the time domain), with the minimum *a priori* knowledge. We assume that saturation effects are absent in the mixed observable signal $x(t)$, that can be expressed as:

$$x(t) = s(t) + n(t) \tag{1}$$

We assume that $s(t)$ and $n(t)$ are uncorrelated. This extends linearity in the power spectral domain, and let us to transform the data in a non-negative representation suitable for NMF processing:

$$|X(t,f)|^2 = |S(t,f)|^2 + |N(t,f)|^2 \tag{2}$$

where the observable signal $x(t)$ is transformed in a time-frequency representation $X(t,f)$. Our method is shown in Fig. 1 and functional modules are discussed in the next subsections.

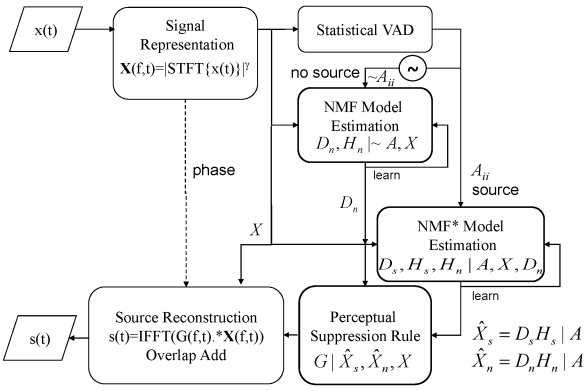

Figure 1. *General scheme of the proposed audio enhancement framework.*

2.1 Signal Representation

A common technique to manipulate audio signals consists of transforming the time-varying observed signal in a time-frequency representation (by means a Short Time Fourier Transform – STFT – analysis) which shows the signal energy variation along time elements (frames) and frequency elements (bins), thus providing a non-negative matrix representation. In the following, we represent the signal in the time-log frequency domain as an element-wise exponentiated STFT:

$$X = |STFT\{x(t)\}|^{\gamma} \tag{3}$$

The linearity expressed by Eq. 2 applies also to Eq. 3 when $\gamma = 2$, but wide experimentation shows that γ is an important parameter to NMF performance. In particular, it turns out that $\gamma = 2$ is a bad choice for component separation, while an optimal choice is $\gamma = 0.67$, which corresponds to the cube root compression of power STFT.

Surprisingly, this is consistent with Stevens' Power Law exponent for the perceived loudness of a sound pressure of 3 kHz tone stimulus. Moreover, Stevens' Power Law was used to model cochlear non-linearities [6] and intensity to loudness conversion in Perceptual Linear Predictive (PLP) speech analysis [7]. More recently, Plourde and Champagne integrated the cochlear compressive nonlinearity in a Bayesian Short Time Spectral Attenuation (STSA) estimation for speech enhancement [8]. This curious coincidence about the exponent value, suggests to follow a perceptually motivated approach to audio de-noising, as we explain in Sec. 2.5.

2.2 Voice Activity Detection

A Voice Activity Detector (VAD) is widely used as a component of speech enhancement methods to update the noise spectrum frame by frame. In our implementation, a statistical-model based VAD [9] is used to construct two diagonal binary square matrices:

$$A(t,t) = \begin{cases} 1, & \text{if target source is present in frame t} \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

and its complementary $\bar{A}(t,t)$.

This allows us to train the undesired components dictionary, computing NMF on the signal:

$$Z(f,t) = X(f,t)\bar{A}(t,t) \quad (5)$$

during target-absent periods, and then separate the target components dictionary, computing a modified NMF^* on the signal:

$$Y(f,t) = X(f,t)A(t,t) \quad (6)$$

during target-present periods. Assuming that the target and the undesired component are additive (as stated in Eq. 1), the VAD module has to decide, for each frame t, in favor of one of the two hypotheses:

$$H_0 : X_f = N_f : \quad \text{target source absent,} \quad (7)$$
$$H_1 : X_f = S_f + N_f : \quad \text{target source present.} \quad (8)$$

2.3 Undesired component training

During training stage, we assume availability of some target-absent frames, computed applying a VAD to the observable signal $X(f,t)$; the resulting signal $Z(f,t)$ of Eq. 5 is equivalent to $X(f,t)$, with target-present frame suppressed. Applying a Regularized Euclidean NMF to $Z(f,t)$, we obtain the strictly positive dictionary $D_n(f,k)$ and sparse code $H_n(k,f)$ matrices, where k is the number of user defined elements of interference. Following the simplification proposed in [5], we define the multiplicative iterative computation of H_n and D_n:

$$\hat{X}_n = \bar{D}_n H_n; H_n \leftarrow H_n \bullet \frac{\bar{D}_n^T Z}{\bar{D}_n^T \hat{X}_n + \lambda_n}; \quad (9)$$

$$D_n \leftarrow \bar{D}_n \bullet \frac{ZH_n^T + \bar{D}_n \bullet (\mathbf{1}(\hat{X}_n H_n^T \bullet \bar{D}_n))}{\hat{X}_n H_n^T + \bar{D}_n \bullet (\mathbf{1}(ZH_n^T \bullet \bar{D}_n))}. \quad (10)$$

Where $\bar{D}_n$ is the Euclidean column-wise normalization of D_n in current iteration (see Sec. 2.4), the $\bullet$ operator indicates element-wise multiplication, the fraction line indicates element-wise division, and $\mathbf{1}$ is a square matrix of ones. The regularization parameter λ_n weights the importance of the sparsity term to the reconstruction.

The final D_n matrix represents the dictionary of the interference learned from data and it will be used by the next module to estimate the two additive sources composing the mixed signal.

2.4 Estimation of undesired source and target source

In order to estimate the sources, we use again a constrained NMF (NMF*) to compute the dictionary of the target source and the sparse code of both sources. Assuming, as usual, the additivity of sources, the dictionary of the mixed signal can be seen as the concatenation of the individual source dictionaries. Moreover, the sparse code of the mixed signal can be seen as the concatenation of the individual source sparse codes:

$$X = X_s + X_n = \begin{bmatrix} D_s \, D_n \end{bmatrix} \begin{bmatrix} H_s \\ H_n \end{bmatrix} + E = DH + E \quad (11)$$

In the previous equation, E is an unknown matrix representing approximation errors. We can not solve Eq. 11 directly with NMF, due to a permutation ambiguity. In fact, we can write

$$DH = (DP)(P^{-1}H) \quad (12)$$

where P is a generalized permutation matrix, i.e., a matrix with only one non-zero positive element in each row and each column.

Schmidt, Larsen and Hsiao [10] suggest to pre-compute D_n, as we have done in the previous section for the interference in the $Z(f,t)$ signal; then learn $D_s(f,m)$, $H_s(m,t)$ and $H_n(k,t)$, where m is the number of user defined elements of the target source, with a modified constrained NMF, which we apply to $Y(t,f)$ in Eq. 6 (i.e. the observed signal in the target-present frames). We describe here the developed one-dictionary constrained (D_n^*) algorithm:

1. Initialize $D_s(f,m)$, $H_s(m,t)$ and $H_n(k,t)$ with random values in the range $[0 \div 1]$; to multiply $H_s(m,t)$ and $H_n(k,t)$ by A to suppress target-absent frames.
2. Define Euclidean column-wise normalization of the target dictionary to prevent joint numerical drifts in H_s and D_s:

$$\bar{D}_s(f,m) = \frac{D_s(f,m)}{\sqrt{\sum_f D_s(f,m)^2}} = \frac{D_s(f,m)}{||D_s(m)||_2}. \quad (13)$$

3. Calculate the overall reconstruction according to:

$$\hat{X} = \bar{D}_s H_s + \bar{D}_n H_n. \quad (14)$$

4. Update the sparse code of target according to the rule:

$$H_s \leftarrow H_s \bullet \frac{\bar{D}_s^T Y}{\bar{D}_s^T \hat{X} + \ell_s}. \quad (15)$$

5. Calculate the overall reconstruction as in Eq. 14.

6. Update the sparse code of interference according to the rule:

$$H_n \leftarrow H_n \bullet \frac{\bar{D}_n^T Y}{\bar{D}_n^T \hat{X} + \ell_n}. \quad (16)$$

7. Calculate the overall reconstruction as in Eq. 14.

8. Update the target non-normalized dictionary according to the rule:

$$D_s \leftarrow \bar{D}_s \bullet \frac{Y H_s^T + \bar{D}_s \bullet (\mathbf{1}(\hat{X} H_s^T \bullet \bar{D}_s))}{\hat{X} H_s^T + \bar{D}_s \bullet (\mathbf{1}(Y H_s^T \bullet \bar{D}_s))}. \quad (17)$$

9. Repeat from step 2 until it reach the convergence of the Euclidean Cost function to minimize:

$$C^{(i)} = \frac{1}{2} \sum_{f,t} (Y(f,t) - \hat{X}(f,t))^2 + \ell_n \sum_{k,t} H_n(k,t) + \ell_s \sum_{m,t} H_s(m,t). \quad (18)$$

We stop the algorithm at iteration i when $|C^i - C^{i-1}| < \varepsilon C^i$. The regularization parameters ℓ_s and ℓ_n determine the degree of sparsity in the activity matrix. D_n, the dictionary of the undesired component, is left unchanged by this algorithm because it is predefined and fixed by the previous training stage; moreover, we do not seek a sparse code for the fixed dictionary, but the code that minimizes the reconstruction error, setting $\ell_n = 0$. In general λ_n, ℓ_s, k and m are depending on unknown sources. In our experimental datasets, good results were obtained for $\lambda_n = 0.2$ and $\ell_s = 0.05$, $k = 256$ and $m = 256$, confirming in a wider field of application the results of Schmidt et al. [10].

2.5 Perceptually motivated Bayesian Suppression Rules

The output of the two previous stages are the estimation of D_s, H_s, D_n and H_n; we can estimate the spectrogram of the target source and interference in target-present frames as:

$$\hat{X}_s = D_s H_s \quad (19)$$

$$\hat{X}_n = D_n H_n \quad (20)$$

Figure 2 shows the result spectrograms of $\hat{X}_s$ and $\hat{X}_n$ where the undesired component is the period stationary wide-band noise present in the observed extract 4 of Sec. 3.

We can reconstruct the target source using a noise suppression rule, a well known technique in speech enhancement and audio denoising in general. A suppression rule may be viewed as a non-negative real-valued time-frequency-varying gain $G(f,t)$, applied to the observable, target-present signal spectrum $Y(f,t)$, in order to estimate the target source spectrum:

$$\hat{S}(f,t) = G(f,t) \bullet Y(f,t) \text{ with } 0 \leq G(f,t) \leq 1 \quad (21)$$

Although in many cases, with high SNR, we can get a good reconstructed target source by means of the Wiener filter, in low SNR we get increasing target distortion and perceptually annoying musical noise (a tonal, random, isolated, time-varying noise). Generally speaking, we can reduce noise suppression in favor of better audio fidelity or speech intelligibility introducing the masking phenomenon of the human hearing model to calculate a noise masking threshold from the estimated target source. A listener tolerates additive interference, as long as its energy remains below the masking threshold defined by the target source energy, and we don't need to suppress this masked interference because it is non-audible. In this sense we suppress only the non-masked excess of interference.

A widely used, simple but effective masking model was proposed by Johnston [11] to mask the distorsion introduced in speech and audio process and adopted with success in speech enhancement. In this psychoacoustic model, a weak interference at a certain frequency is made inaudible by a stronger target occurring simultaneously (i.e., in the same frame) within the same perceptual frequency range, termed *Critical Band*, and across Critical Bands, applying a convolution with a spreading function. The Johnston's masking threshold calculation does not take into account backward or forward temporal masking. According to Wolfe and Godsill [3], we can formulate a perceptually motivated distortion measure as a generalization of the Minimum Mean Square Error (MMSE) cost function that incorporates the masking threshold:

$$C_{WG}(S,\hat{S},T) = \begin{cases} \left(\hat{S} - S - \frac{T}{2}\right)^2 - \left(\frac{T}{2}\right)^2, \\ \qquad \text{if } \left|\hat{S} - S - \frac{T}{2}\right| > \frac{T}{2}; \\ 0, \qquad \text{otherwise.} \end{cases} \quad (22)$$

where S is the true but unknown STFT amplitude of the source, $\hat{S}$ is the STFT amplitude of estimated source and T is the masking threshold; for simplicity, we omit the frequency f and the frame t indices. We can see in Eq. 22 that no cost is assigned if the estimation error is below the masking threshold and a penalty cost is assigned only when the estimation error is above the masking threshold. This prevent unwanted source attenuation when the undesired component is masked and suppress only human ear perceptible undesired component. Unfortunately, the analytical minimization of $E[C_{WG}(S,\hat{S},T)]$ is intractable and a numerical implementation was adopted by the authors [3].

In our perceptually motivated model, depicted in fig. 3, we followed a different approach based on [12], where a perceptually criterion was implicitly implemented by weighting error STFT amplitude with a filter that has the shape of the inverse STFT amplitude of the source, so that less emphasis is placed near the formant peaks (implicit masked) and more emphasis is placed on spectral valleys (implicit unmasked):

$$C_{WE}(S,\hat{S},p) = (\hat{S} - S)^2 \bullet S^p \text{with} - 2 < p \leq 0 \quad (23)$$

where p is a real value time-frequency parameter that

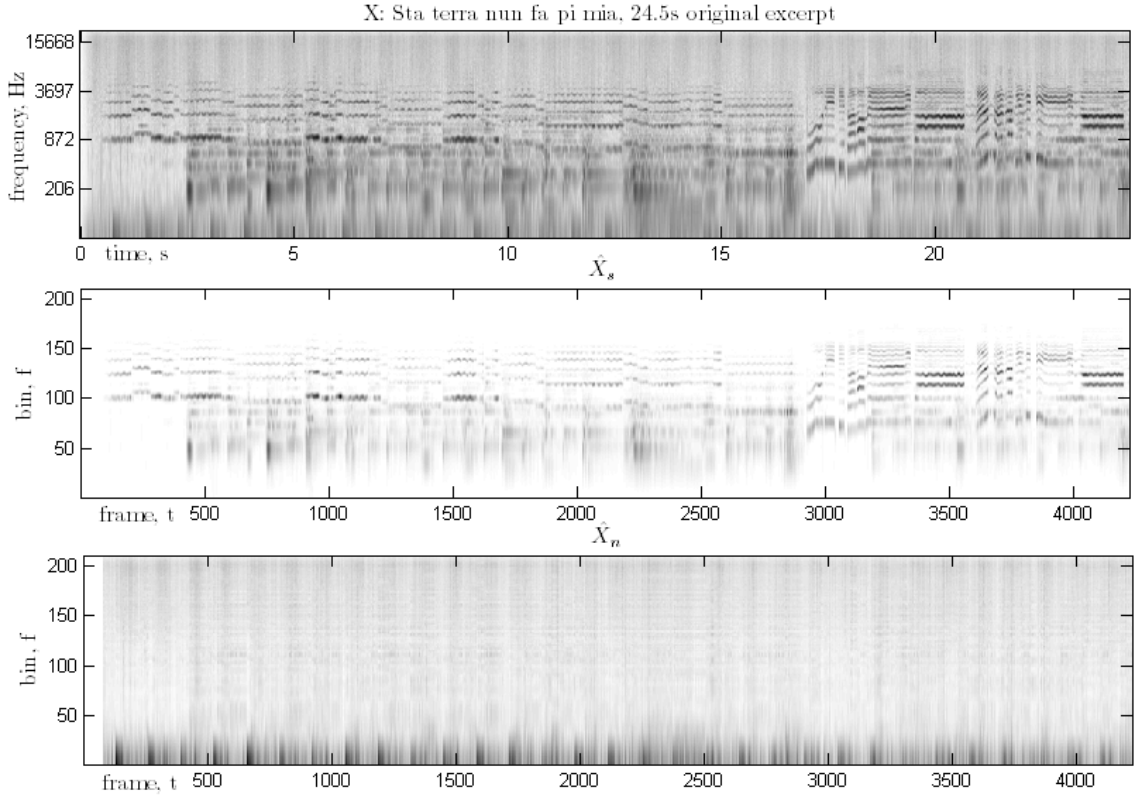

Figure 2. *Spectrograms of original noisy signal X (top), estimated target source $\hat{X}_s$ (center) and estimated interference $\hat{X}_n$ (bottom) of 24.5 seconds excerpt 'Sta terra nun fa pi mia' (see Sec. 3 item 4), spectrograms are in log-frequency representation, from $f_{min} = 50Hz$ and $24bin/octave$ resolution. Audio pattern and period stationary wide-band noise are clearly separated.*

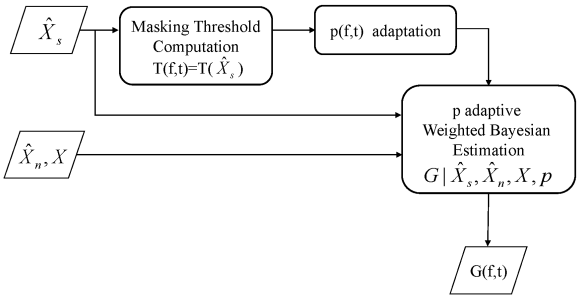

Figure 3. *The proposed perceptual suppression rule scheme based on threshold-mask-adaptive weighted Bayesian estimator.*

emphasizes spectral valleys when negative. This cost function is known as Weighted Euclidean distortion measure and the analytical minimization of $E[C_{WE}(S,\hat{S},p)]$, assuming $S(f,t)$ modelled as statistically independent zero-mean Gaussian random variables, evaluates to the time-frequency gain $G_{WE}(f,t)$:

$$G_{WE} = \frac{\sqrt{\nu}}{\gamma} \bullet \frac{\Gamma\left(\frac{p+1}{2}+1\right) \bullet \Phi\left(-\frac{p+1}{2}, 1; -\nu\right)}{\Gamma\left(\frac{p}{2}+1\right) \bullet \Phi\left(-\frac{p}{2}, 1; -\nu\right)}, \quad p > -2 \tag{24}$$

where:

$$\gamma = \frac{X}{\hat{X}_n};\ \xi = \frac{\hat{X}_s}{\hat{X}_n};\ \nu = \frac{\xi}{1+\xi} \bullet \gamma \tag{25}$$

$\Gamma(\cdot)$ denotes the gamma function and $\Phi(a,b;z)$ denotes the confluent hypergeometric function. When $p = 0$, we get the classical Ephraim and Malha MMSE STSA estimator [13].

We consider now a similar cost function, also proposed in [12], called Weighted Cosh distortion measure:

$$C_{WCOSH}(S,\hat{S},p) = \left(\frac{S}{\hat{S}} + \frac{\hat{S}}{S} - 1\right) \bullet S^p \ \text{ with } -1 < p \leq 0 \tag{26}$$

this cost function, again, emphasize spectral valleys when p is negative and evaluates to the $G_{WCOSH}(f,t)$ gain:

$$G_{WCOSH} = \frac{\sqrt{\nu}}{\gamma} \bullet \sqrt{\frac{\Gamma\left(\frac{p+3}{2}\right) \bullet \Phi\left(-\frac{p+1}{2}, 1; -\nu\right)}{\Gamma\left(\frac{p+1}{2}\right) \bullet \Phi\left(-\frac{p-1}{2}, 1; -\nu\right)}},\ p > -1 \tag{27}$$

The third Bayesian Estimator implemented in the framework is the β-Order MMSE STSA estimator (βSA), proposed in [14], and further developed in [15]. The MMSE-STSA estimator [13] [16] was generalized by the real exponent parameter β (for uniformity with previous estimators, we continue to call p):

$$C_{SA}(S,\hat{S},p) = (\hat{S}^p - S^p)^2 \text{ with } -2 < p < 0 \tag{28}$$

taking $p < 0$, the behavior to penalize the cost function is similar to Weighted Euclidean Distortion Measure of

Eq. 23, and both perform an accurate estimation of source in spectral valleys. The gain function $G_{SA}(f,t)$ derived from the βSA estimator is expressible as:

$$G_{SA} = \frac{\sqrt{\nu}}{\gamma} \bullet \left[\Gamma\left(\frac{p}{2}+1\right) \bullet \Phi\left(-\frac{p}{2}, 1; -\nu\right)\right]^{1/p}, \quad p > -2 \tag{29}$$

When $p \to 0$, the βSA is equivalent to the Ephraim and Malah MMSE log-STSA [16].

Extensive tests of the three perceptually motivated Bayesian estimators, with PESQ measure and informal audio assessment on speech phrases at low SNR (i.e. < 10dB), show that all of them perform equally well, in the sense that they don't introduce musical noise and permit to control the trade-off between undesired component suppression and source attenuation by varying the parameter p. A performance evaluation and comparative audio samples are available in `http://dialogo.fisica.uniud.it/BASS/ComparisionWithGustafsson02`.

Indeed, the optimal choice of the real parameter $p(f,t)$ is an important performance issue. Therefore, we considered to express explicitly the relation with the masking threshold $T(f,t)$, although in an heuristic manner. We have seen that the undesired component can be reduced by decreasing p, however this leads to more source distortion. Therefore, the adaptation is based on the following consideration: if the masking threshold is high, interference will be masked and consequently inaudible. Consequently, there is no need to reduce, which helps to keep distortion as low as possible. In this case the parameter p is kept to his maximal value: $p = p_{max}$. However, if the masking threshold is low, undesired component will be unpleasant or even annoying to the ear and it is necessary to reduce it. This is done by a decrease of p toward his minimum value: $p = p_{min}$. For each frame t, the minimum of the masking threshold $T(f,t)$ corresponds to the minimum of the power parameter $p(f,t)$. In order to avoid discontinuities in the gain function G due to this adaptation, a smoothing operation is applied, controlled by user value x. The adaptation of the parameter $p(f,t)$ is performed with the following relation:

$$p(f,t) = \left(\frac{T(f,t) - T_{min}(t)}{T_{max}(t) - T_{min}(t)}\right)^{x} (p_{max} - p_{min}) + p_{min} \tag{30}$$

where $T_{max}(t)$ and $T_{min}(t)$ are the maximal and minimal values of noise masking threshold $T(f,t)$ at current frame t. In this way, $p(f,t)$ adapts to a minimal interference reduction for the maximal values of the masking threshold (i.e. in correspondence of source formant peaks) and a maximal reduction for the minimal values of the threshold (i.e. in correspondence of spectral valleys). Figure 4 show the simple smoothing curves obtained with Eq. 30 varying the smoothing parameter x.

The minimal and maximal values of p and x determine the tradeoff between residual noise and source distortion. A number of experiments with different noise types and levels have been performed to select the appropriate values for these parameters.

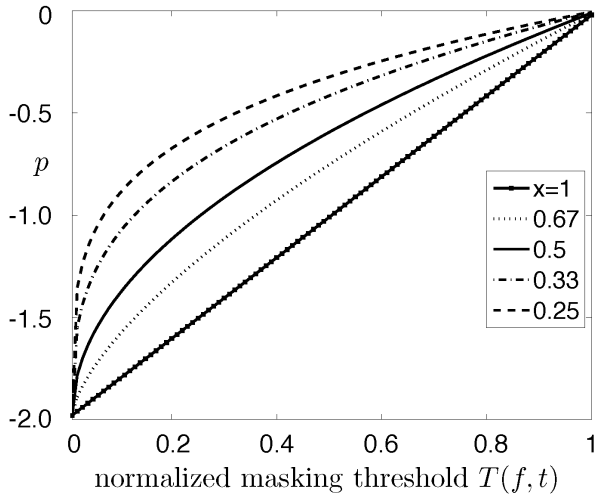

Figure 4. *Parameter p versus the normalized masking threshold T for smoothing parameter $x = (0.25, 0.33, 0.5, 0.67, 1)$, $p_{min} = -1.98$, $p_{max} = 0$.*

For additive interference at SNR proximal to 0 dB, the following values have been chosen in order to obtain a good tradeoff for a human listener for each Bayesian estimator:

1. For Weighted Euclidean: $p_{WEmax} = 0$, $p_{WEmin} = -1.98$;
2. For Weighted Cosh: $p_{WCOSHmax} = 0$, $p_{WCOSHmin} = -0.99$;
3. For βSA: $p_{max} = -0.001$, $p_{min} = -1.98$;

The smoothing exponent is kept fixed for all estimators: $x = 0.5$.

This tradeoff can be easily changed depending on the application; in general, only a regulation of p_{min} is needed to optimal tradeoff.

The use of Johnston' simultaneous masking threshold estimation allows the construction of effective and sophisticated perceptual noise suppression rules. However, if the threshold is not correctly estimated, performance greatly suffers in terms of very annoying musical noise injected in the target source waveform, compromising any noise suppression rule. To properly estimate the threshold, we need an extremely accurate estimation of the target source spectrum $\hat{X}_s(f,t)$ that we obtained with NMF.

3. SUBJECTIVE EVALUATION

To validate the system, a listening test was conducted. As audio material, several sound documents of ethnic music were considered.

Material. Four music pieces recorded in Shellac disc were used. In order to minimize fatigue and maximize attention by the participating subjects, we selected the 20 first seconds of each stimulus. Since the task was more a comparison than an individual analysis, those short extracts seemed to be sufficient.

1. *Chi campa deritto campo aflitto* (Who lives honestly lives poorly, by Perrocato and Canoro), Eduardo Migliaccio (voc) - 78 rpm 10" Victor 14-81712-B (BVE 46692-2), rec. in New York, August, 14, 1928, length 3'36". In the excerpt considered: singing voice and music.

2. *Il funerale di Rodolfo Valentino* (The funeral of Rodolfo Valentino), Compagnia Columbia (2 male singers, 2 female singers, bells and Orchestra) - 78 rpm 10" Columbia 14230-F (w 107117 2), rec. in New York, September, 1926, length 2'55". In the excerpt considered: speech voices.

3. *La signorina sfinciusa* (The funny girl), Leonardo Dia (voc), Alfredo Cibelli (mandolin), unknowns (2 guitars) - 78 rpm 10" Victor V-12067-A (BVE 53944-2), rec. in New York, July, 24, 1929, length 3'20". In the excerpt considered: singing voice and music.

4. *Sta terra nun fa pi mia* (This land is not for me, by R. Gioiosa, arr. R. Romani), Rosina Gioiosa Trubia (voc), Alfredo Cibelli (mandolin), unknowns (2 guitars) - 78 rpm 10" Brunswick 58073B (E 26621/2), rec. in New York, February, 23, 1928, length 3'22". In the excerpt considered: singing voice and music.

Noisy stimuli was pre-processed with the Extended Kalman Filter, detailed in [17] (in de-click mode), then broad band restoration was performed using our framework with the suppression rule detailed in 2.5, as well as the following three commercial products, selected among the most appreciated products in the audio archives and post-processing studios:

1. X-Noise of Waves Restoration bundle (Waves V6 Update 2);

2. Denoiser (with the *Musical noise suppression* filter enabled) of iZotope RX v1.06;

3. Auto Dehiss of CEDAR Tools;

The CEDAR Tools plug-ins are used in a Pro Tools HD system. The parameters used to control the different systems were subjectively set to obtain the best tradeoff between noise removal and music signal preservation. In this way 16 restored stimuli were produced.

Test method. The tests were conducted using the EBU MUSHRA test method [18], which is a recommended evaluation method adopted by ITU [19]. This protocol is based on the "double-blind triple-stimulus with hidden reference" method, which is stable and permits accurate detection of small impairments. An important feature of this method is the inclusion of the hidden reference and of two bandwidth-limited anchors signals (7 kHz and 3.5 kHz).

The noisy stimuli under test are all real-world signals. This implies that we can not compare test enhanced sound with a high quality reference sound (graded 5.0 at the top of the grading scale), but with the noisy reference sound (graded 0.0). Moreover, negative scores are allowed to evaluate test sounds that rate worse then the noisy reference. At least the hidden reference must be graded 0.0 by the evaluator. All the other test stimuli and hidden anchors can be evaluated subjectively to rate the overall quality of sound excerpts.

Training phase. The purpose of the training phase, according to the MUSHRA specification, was to allow each listener: i) to become familiar with all the sound excerpts under test and their quality-level ranges; ii) to learn how to use the test equipment and the grading scale.

Listeners. Two subject groups were selected:

1. Musically trained (MT): 12 researchers (musicologists and/or musicians) of the University of Padova and 12 technicians of different international audio archives.

2. Musically untrained (MU): 16 students in Information Engineering (University of Padova).

Equipment. The audio signals were recorded at 44.1 kHz/24 bit (uncompressed sound files) and played through Apple PowerBook Pro 2.4 GHz Intel Core 2 Duo with 2 GB 1067 MHz DDR3 equipped with a D/A converter RME Fireface 400, and headphones AKG K 501. The listeners could play in any order all the stimuli under test, including the hidden reference and the two bandwidth-limited anchor signals.

Test duration. The training session for each listener took approximately 40', including an explanation about the tests and equipment, and a practice grading session. The grading phase consisted of 4 test sessions (one for each music piece), each one containing 9 test signals (1 noisy signal, 6 restored signals, 2 anchors). Each session took, on average, about 8 minutes. Subjects were allowed a rest period between each session, but not during a session.

Main results. The statistical analysis method described in the MUSHRA specification was used to process the test data. The results are presented in Tab. 1 as mean grades. The results from six listeners (five of them belong to MU group, one to MT) were rejected because the mean of their rates (in absolute value) on hidden references is greater than $+/-0.5$.

The quality range between the best and worst restoration system is only 0.80 (MT group) and $0.40(MUgroup)$. In general there are only two systems with a score > 3.5: our Tool and CEDAR. Our algorithm produces scores similar to CEDAR in both test sessions (better for MU group) and better than the others softwares.

Table 1. *Mean for restored stimuli and anchors, 34 subjects. MT = Musically trained; MU = Musically untrained.*

Restoration system	**MT group**	**MU group**	**Average**
Our Tool	$+3.00$	$+4.20$	$+3.60$
CEDAR Tools	$+3.40$	$+4.00$	$+3.70$
Waves	$+2.80$	$+3.80$	$+3.30$
iZotope RX	$+2.20$	$+3.80$	$+3.00$
Anchor 7 kHz	-2.69	$+0.20$	-1.02
Anchor 3.5 kHz	-5.00	-4.20	-4.60

4. CONCLUSIONS

This study is focused on the restoration of single channel audio recordings of ethnic music: for this purpose, we applied EKF framework to audio signal enhancement problems. In this paper we investigate the use of the Non-negative Matrix Factorization (NMF): we show that NMF is a suitable technique to extract the clean audio signal from undesired non stationary noise in a monaural recording with low SNR. More specifically, we introduce a perceptual suppression rule based on an advanced psychoacoustic models (Sec. 2.5. To evaluate the proposed approach a subjective audio enhancement experiments was carried out (see Section 3). The results of this experiments show that the proposed method results in improved audio quality and that it is a useful alternative to the classical STSA methods.

Future work will carry out an intensive application of this audio restoration environment on two real archives of ethnic music phonographic discs.

5. REFERENCES

[1] W. Storm, "The establishment of international re-recording standards," *Phonographic Bulletin*, vol. 27, pp. 5–12, 1980.

[2] S. Canazza and A. Vidolin, "Preserving electroacoustic music," *Journal of New Music Research*, vol. 30, no. 4, pp. 351–363, 2001.

[3] P. J. Wolfe and S. J. Godsill, "Towards a perceptually optimal spectral amplitude estimator for audio signal enhancement," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. II, (Istanbul, Turkey), pp. 821–824, 2000. ISBN 0-7803-6296-9.

[4] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, no. 401, pp. 788–791, 1999.

[5] J. Eggert and E. Krner, "Sparse coding and nmf," in *IEEE International Conference on Neural Networks*, pp. 2529–2533, IEEE, 2004.

[6] R. Meddis, L. P. O'Mard, and E. A. Lopez Poveda, "A computational algorithm for computing nonlinear auditory frequency selectivity," *Journal of the Acoustical Society of America*, vol. 109, no. 6, pp. 2852–2861, 2001.

[7] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, pp. 1738–1752, Apr 1990.

[8] E. Plourde and B. Champagne, "Integrating the cochleas compressive nonlinearity in the bayesian approach for speech enhancement," in *15th EUSIPCO, Poznan, Poland*, pp. 70–74, 2007.

[9] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, pp. 1–3, 1999.

[10] M. N. Schmidt, J. Larsen, and F. T. Hsiao, "Wind noise reduction using non-negative sparse coding," in *IEEE Workshop on Machine Learning for Signal Processing*, pp. 431–436, Aug 2007.

[11] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 2, pp. 314–323, 1988.

[12] P. C. Loizou, "Speech enhancement based on perceptually motivated bayesian estimators of the magnitude spectrum," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5-2, pp. 857–869, 2005.

[13] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, 1984.

[14] S. H. K. C. H. You and S. Rahardja, "β-order mmse spectral amplitude estimation for speech enhancement," *IEEE Trans. Speech Audio Processing*, vol. 13, no. 4, pp. 475–486, 2005.

[15] E. Plourde and B. Champagne, "Further analysis of the β-order mmse stsa estimator for speech enhancement," in *Canadian Conference on Electrical and Computer Engineering. CCECE 2007*, pp. 1594–1597, 2007.

[16] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, 1985.

[17] S. Canazza, G. De Poli, and G. Mian, "Restoration of audio documents by means of extended kalman filter," *IEEE Transactions on Audio, Speech and Language Processing*, vol. in press, 2010.

[18] ITU-R, "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems," *Recommendation BS.1116-1*, 2000.

[19] EBU Project Group B/AIM, "EBU report on the subjective listening tests of some commercial internet audio codecs," October 2000.