THE RESTORATION OF SINGLE CHANNEL AUDIO RECORDINGS BASED ON NON-NEGATIVE MATRIX FACTORIZATION AND PERCEPTUAL SUPPRESSION RULE

Giuseppe Cabras

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

Sergio Canazza

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

Pier Luca Montessoro

Dep. of Electrical, Management
and Mechanical Engineering
University of Udine, Italy
montessoro@uniud.it

Roberto Rinaldo

Telecommunication and Signal Processing group
Dep. of Electrical, Management
and Mechanical Engineering
University of Udine, Italy
rinaldo@uniud.it

ABSTRACT

In this paper, we focus on the signal-to-noise ratio (SNR) improvement in single channel audio recordings. Many approaches have been reported in the literature. The most popular method, with many variants, is Short Time Spectral Attenuation (STSA). Although this method reduces the noise and improves the SNR, it mostly tends to introduce signal distortion and a perceptually annoying residual noise usually called musical noise. In this paper we investigate the use of Non-negative Matrix Factorization (NMF) as an alternative to the STSA for the *digital curation* of musical heritage. NMF is 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, we introduce a perceptual suppression rule to determine how the perceptual domain is competitive compared to the acoustic domain. 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 wider applicability in audio separation.

1. INTRODUCTION

Noise reduction, aiming at estimating the desired clean speech signal from noisy observations, is a very important problem and has attracted a significant amount of research and engineering attention over the past few decades. In particular, the enhancement of audio sources corrupted by non stationary noise in a monaural recording is a challenging task, only partially addressed by classical methods in Speech Enhancement [1], also adopted in Digital Audio Restoration [2]. A different method is followed by Perceptually motivated approaches, like Computational Auditory Scene Analysis (CASA), where the main idea is to simulate the human auditory system and the perceptual processes there involved [3]. A more 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 addiction, 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 ethnic music noisy musical recordings on Shellac 78 rpm phonographic discs[1]. There are many reasons to believe that these audio documents of ethnic music are the most complex in terms of restoration. Ethnic music refers to the music recordings of non-Western cultures since the beginning of the 20th century and the problems of multiple formats and carriers (e.g., in analogue domain: wax cylinders, sonofils, discs, tapes, and cassettes; in digital domain: magnetic tapes and optical disks) involved with the documenting, researching, and preservation of this music. The analysis of Western mu-

[1] The Shellac disc is a common audio mechanical carrier. 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 ever re-recorded (R&B, Jazz, Ethnic, Western classical, etc.).

sic has been developed almost exclusively on the basis of written scores, which represent musical performance models, rather than the performance itself. Since ethnic music recordings were often made with non-professional systems (low-quality, poorly aligned and maintained – often by technically unskilled researchers – without generally accepted standards and recording practices), the audio carriers – almost obsolete – show risk of deterioration. In this sense, the ethnic-musical heritage is in danger of disappearing, of being forgotten in some public archive or, in most cases, a private collection, because of the poor quality of the material on which the audio documents were recorded and the rapid evolution of the recording formats – that make obsolete and scarcely readable many old recordings. Obviously, this has not been the destiny of music repertoires of wider interest, such as classical western music, rock/pop, or jazz. In these cases, the recording companies have re-recorded most of the audio documents, particularly those of high commercial values. Unfortunately, the same has not happened to ethnic and traditional repertoires. In this sense, it is particularly important to restore the ethnic audio documents that are often the only testimonial of disappeared oral cultures.

The rest of this paper is organized as follows. Sec. 2 details the proposed audio restoration method: in particular, Sec. 2.5 introduces the perceptual suppression rule used. Sec. 3 presents extensive objective quality measures. Since we developed a psychoacoustic technique, a natural choice to measure quality was the Perceptual Evaluation of Speech Quality (PESQ), defined by ITU-T recommendation P.862. 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. 4). Final conclusions are drawn in Sec. 5.

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 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.

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:

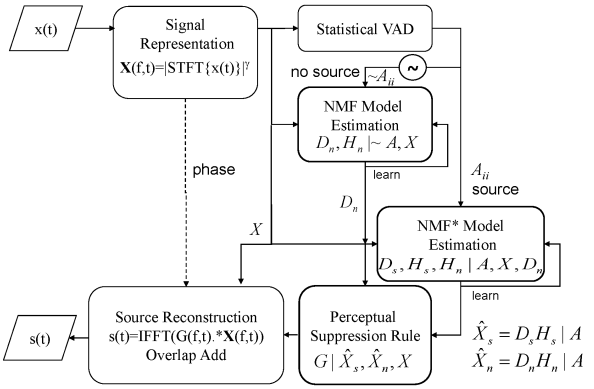

Figure 1: *General scheme of the proposed audio enhancement framework.*

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

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:

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

and its complementary:

$$\bar{A}(t,t) = \begin{cases} 1, & \text{if target source is absent in frame t} \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

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

$$Z(f,t) = X(f,t)\bar{A}(t,t) \tag{6}$$

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

$$Y(f,t) = X(f,t)A(t,t) \tag{7}$$

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:

$$H_0 : X_f = N_f : \quad \text{target source absent,} \tag{8}$$
$$H_1 : X_f = S_f + N_f : \quad \text{target source present.} \tag{9}$$

The following Likelihood Ratio Test (LRT) based VAD decision rule was used:

$$\frac{1}{L}\sum_{f=1}^{L-1} log\Lambda_f \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{10}$$

where the likelihood ratio for the fth bin is:

$$\Lambda_f = \frac{p(X_f|H_1)}{p(X_f|H_0)} = \frac{1}{1+\xi_f} exp\left(\frac{\gamma_f \xi_f}{1+\xi_f}\right) \tag{11}$$

In the previous equation, ξ_f and γ_f define the *a priori* and *a posteriori* SNRs. In particular, ξ_f is estimated using the decision directed approach (with $\alpha = 0.99$) as in [10], L is the size of the FFT and η is a user defined threshold. An Hidden Markov Model (HMM) hang-over algorithm extends and smoothes the VAD decision in order to recover target periods that are masked by the undesired component. Fig. 2 shows an example of VAD applied to a real-world noisy recording (ethnic music, 78 rpm Shellac disc) with singing voice and music accompaniment[2].

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. 6 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 as follows the complete multiplicative iterative algorithm:

1. Initialize $D_n(f,k)$ and $H_n(k,t)$ with random values between 0 and 1, multiply $H_n(k,t)$ by $\bar{A}(t,t)$ to suppress target-present frames.
2. Define Euclidean column-wise normalization of the dictionary to prevent joint numerical drifts in H_n and D_n:

$$\bar{D}_n(f,k) = \frac{D_n(f,k)}{\sqrt{\sum_f D_n(f,k)^2}} = \frac{D_n(f,k)}{||D_n(k)||_2} \tag{12}$$

3. Calculate the reconstruction according to:

$$\hat{X}_n = \bar{D}_n H_n. \tag{13}$$

4. Update the sparse code according to the rule:

$$H_n \leftarrow H_n \bullet \frac{\bar{D}_n^T Z}{\bar{D}_n^T \hat{X}_n + \lambda_n}. \tag{14}$$

5. Calculate the reconstruction according the Eq. 13.
6. Update the non-normalized dictionary according to the rule:

$$D_n \leftarrow \bar{D}_n \bullet \frac{Z H_n^T + \bar{D}_n \bullet (\mathbf{1}(\hat{X}_n H_n^T \bullet \bar{D}_n))}{\hat{X}_n H_n^T + \bar{D}_n \bullet (\mathbf{1}(Z H_n^T \bullet \bar{D}_n))} \tag{15}$$

7. Repeat from step 2 until convergence to a local minimum of the Euclidean Cost function:

$$C^i = \frac{1}{2}\sum_{f,t}(Z(f,t) - \hat{X}_n(f,t))^2 + \lambda_n \sum_{k,t} H_n(k,t) \tag{16}$$

We stop the algorithm at iteration i when $|C^i - C^{i-1}| < \varepsilon C^i$.

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:

$$X = X_s + X_n = \begin{bmatrix} D_s & D_n \end{bmatrix} \begin{bmatrix} H_s \\ H_n \end{bmatrix} + E = DH + E \tag{17}$$

In the previous equation, E is an unknown matrix representing approximation errors. We can not solve Eq. 17 directly with NMF, due to a permutation ambiguity. In fact, we can write

$$DH = (DP)(P^{-1}H) \tag{18}$$

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 [11] 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. 7 (i.e. the observed signal in the target-present frames). Similarly to the previous section, we describe here the complete one-dictionary constrained (D_n^*) algorithm as:

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:

$$\bar{D}_s(f,m) = \frac{D_s(f,m)}{\sqrt{\sum_f D_s(f,m)^2}} = \frac{D_s(f,m)}{||D_s(m)||_2}. \tag{19}$$

[2] *Sta terra nun fa pi mia* (This land is not for me), by R. Gioiosa, arr. R. Romani – 78 rpm 10" Brunswick 58073B (E 26621/2), rec. in New York, February, 23, 1928, length 3'22".

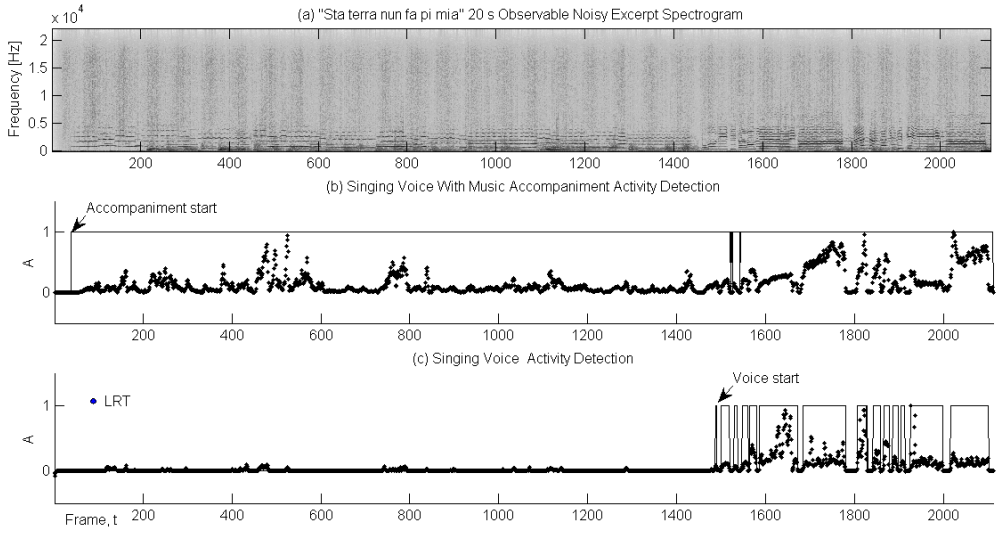

Figure 2: *Statistical VAD at work. (a) In the spectrogram of real-world de-clicked registration excerpt, we recognize harmonic musical structure embedded in the period stationary wide-band noise. (b) Musical content classification providing to the VAD* a priori *information of initial target-absent (i.e. noise only) time extension* $= 0.46$ *s and threshold* $\eta = 0$*. (c) Singing Voice classification providing to the VAD* a priori *information of initial target-absent (i.e. unvoiced) time extension* $= 17$ *s and threshold* $\eta = 0.1$*.*

3. Calculate the overall reconstruction according to:

$$\hat{X} = \bar{D}_s H_s + \bar{D}_n H_n. \tag{20}$$

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

$$H_s \leftarrow H_s \bullet \frac{\bar{D}_s^T Y}{\bar{D}_s^T \hat{X} + \ell_s}. \tag{21}$$

5. Calculate the overall reconstruction as in Eq. 20.

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

$$H_n \leftarrow H_n \bullet \frac{\bar{D}_n^T Y}{\bar{D}_n^T \hat{X} + \ell_n}. \tag{22}$$

7. Calculate the overall reconstruction as in Eq. 20.

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

$$D_s \leftarrow \bar{D}_s \bullet \frac{Y H_s^T + \bar{D}_s \bullet (\mathbf{1}(\hat{X} H_s^T \bullet \bar{D}_s))}{\hat{X} H_s^T + \bar{D}_s \bullet (\mathbf{1}(Y H_s^T \bullet \bar{D}_s))}. \tag{23}$$

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

$$C^{(i)} = \frac{1}{2} \sum_{f,t} (Y(f,t) - \hat{X}(f,t))^2 + \ell_n \sum_{k,t} H_n(k,t) + \ell_s \sum_{m,t} H_s(m,t). \tag{24}$$

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. [11].

2.5. Perceptual Suppression Rule

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:

$$\hat{X}_s = D_s H_s \tag{25}$$

$$\hat{X}_n = D_n H_n \tag{26}$$

The target waveform could be reconstructed by means of STFT^{-1} of $\hat{X}_s$, using the phase information of the observed signal. Unfortunately, this techniques return a non-flexible, poor quality audio waveform. A more flexible and better quality result can be obtained 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:

$$\hat{S}(f,t) = G(f,t) \bullet Y(f,t) \text{ with } 0 \leq G(f,t) \leq 1 \tag{27}$$

In the previous equation, $G(f,t) = G[SNR_{Prior}(f,t)]$ depends on the *a priori* SNR, i.e., the true (but unknown) target to inter-

ference ratio. We have a good estimation of both target and interference sources, provided by Eq. 25 and 26. The function that minimizes the mean squared error of the estimate's time domain reconstruction, when the additive interference is modeled as an uncorrelated, Gaussian random variable, is implemented by the following Wiener-type function:

$$G(f,t) = \frac{\hat{X}_s^{\nu}(f,t)}{\hat{X}_s^{\nu}(f,t) + \hat{X}_n^{\nu}(f,t)} \tag{28}$$

In the previous equation, ν is a positive time-frequency parameter that controls the suppression rate at very low SNR, In particular, a great attenuation is applied when ν is high and $SNR \ll 0dB$. For $\nu = 2$, we obtain the classical Wiener filter, for $\nu = 1$ we have the square-root Wiener filter. 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 [12]. In this 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. A recent technique [13], shows that forward temporal masking models outperform the classical and simultaneous masking-based technique in audio enanchement PESQ objective evaluation, whereas, in subjective evaluation, its performance is aligned to other audio enhancement techniques.

An interesting suppression rule based on Johnston' simultaneous masking threshold $T(f,t)$ is the *Audible Noise Suppression* algorithm proposed by Tsoukalas, Mourjopoulos and Kokkinakis [14]. The rule is based on a psychoacoustic derived quantity, named *audible noise spectrum* $A_n(f,t)$, and defined as the difference between the audible spectrum of the observed signal and the audible spectrum of the target source:

$$A_n(f,t) = A_x(f,t) - A_s(f,t) \tag{29}$$

In particular, $A_n(f,t)$, the noise spectral components lying above the masking threshold, will be suppressed by the algorithm.

Let us express $A_n(f,t)$ in relation with the power spectrum of the target source $P_s(f,t)$, the observed signal $P_x(f,t)$, and the masking threshold $T(f,t)$. We can write:

$$A_n(f,t) = \begin{cases} P_x(f,t) - P_s(f,t) \\ \text{if } P_x(f,t) \geq T(f,t) \text{ and } P_s(f,t) \geq T(f,t) \text{ (I)} \\ P_x(f,t) - T(f,t) \\ \text{if } P_x(f,t) \geq T(f,t) \text{ and } P_s(f,t) < T(f,t) \text{ (II)} \\ T(f,t) - P_s(f,t) \\ \text{if } P_x(f,t) < T(f,t) \text{ and } P_s(f,t) \geq T(f,t) \text{ (III)} \\ 0 \\ \text{if } P_x(f,t) < T(f,t) \text{ and } P_s(f,t) < T(f,t) \text{ (IV)} \end{cases} \tag{30}$$

In cases III and IV, there is no audible noise present because $P_x(f,t) < T(f,t)$. In this case we do not need to enhance $P_x(f,t)$. In case II, $A_n(f,t)$ is always positive or zero and the audible noise needs to be suppressed. In case I, $A_n(f,t)$ may be positive, zero or negative; therefore, the objective of this algorithm is to modify $P_x(f,t)$ when cases I and II happen, so that the modified audible noise spectrum, denoted with $\tilde{A}_n$, must satisfy the condition:

$$\tilde{A}_n(f,t) \leq 0. \tag{31}$$

The audible noise spectrum is minimized according to the following parametric nonlinear Wiener-like function, which allows great flexibility in the gain control:

$$G(f,t) = \frac{X^{\nu}(f,t)}{a^{\nu}(f,t) + X^{\nu}(f,t)} \tag{32}$$

In the previous equation, $X(f,t)$ is the observed signal spectrum, $a(f,t)$ is a positive time-frequency varying threshold, below which all frequency components are heavily suppressed. Suppression remains relatively constant at low SNR values. In contrast, both Spectral Subtraction and Wiener filtering algorithms provide progressively heavy attenuation at very low SNRs. This is an important fact in favor of this method, since loss of intelligibility of speech or naturalness of audio after noise removing is mainly due to aggressive suppression of target source components. In [14], the authors rigorously derived the estimation of the most flexible parameter $a(f,t)$, while suggested to put $\nu = 1$ for simplicity. We report here only the final equation implemented by the algorithm as used in our framework:

$$a(f,t) = \max\{a_I(f,t), a_{II}(f,t)\} \tag{33}$$

where

$$a_I(f,t) = P_x(f,t)\left[\frac{P_x(f,t)}{\hat{P}_s(f,t)} - 1\right]^{1/\nu} \tag{34}$$

$$a_{II}(f,t) = P_x(f,t)\left[\frac{P_x(f,t)}{T(f,t)} - 1\right]^{1/\nu}. \tag{35}$$

Worth to note that it is not desirable to estimate the parameter $a(f,t)$ for every spectral component f, since the estimation will be very sensitive to specific spectral value. Moreover, the Critical Bands are perceptually meaningful frequency regions and will be used to compute the optimum psychoacoustic solution satisfying Eq. 31.

Also note that, in Eq. 32, we used the observed signal spectrum $X(f,t)$ to compute the suppression rule, as documented in [15]. In [14], the same authors suggest to use the power spectrum of the observed signal $P_x(f,t)$, which allows for an improved interference suppression, causing more target source distortion that may reduce the naturalness and fidelity of the enhanced audio or

intelligibility of speech. 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. OBJECTIVE QUALITY MEASURE IN SPEECH ENHANCEMENT

During software development, extensive measures were conducted, in particular, on speech enhancement test sentences, where the original source sentences are also available. Since we developed a psychoacoustic technique, a natural speech quality measure choice was the Perceptual Evaluation of Speech Quality (PESQ), defined by ITU-T recommendation P.862 [16]. This measure has been recently widely used to predict a Mean Opinion Score, rated between 0.5 (bad overall quality or very annoying distortion) to 4.5 (excellent overall quality or imperceptible distortion). We report here only a recent test result, showing the performance of four algorithms:

1. GTM: Gunawan Temporal Masking approach [13],
2. lMMSE: Ephraim and Malah approach [10],
3. SM: our algorithm with Tsoukalas Spectral Magnitude approach,
4. PSM: our algorithm with Tsoukalas Power Spectral Magnitude approach.

applied to four English sentences (courtesy of T.S. Gunawan): FS10 (Female Speech corrupted with Subway Noise @10dB SNR), FS05 (Female Speech corrupted with Subway Noise @5db SNR), MB10 (Male Speech corrupted with Babble Noise @10dB SNR), MB05 (Male Speech corrupted with Babble Noise @5dB SNR).

GTM was chosen because it is a new promising perceptually motivated forward temporal masking method [13], which reaches a high score in PESQ tests, while lMMSE performs consistently best among classical speech enhancement algorithms as shown in a subjective quality comparison [17] .

Results of our objective PESQ scoring are shown in Tab. 1, while audio samples are available in `http://dialogo.fisica.uniud.it/BASS/ComparisionWithGunawan09`. Indeed, PESQ rates correctly the de-noise quality of enhanced speech. Unfortunately, the relation between speech quality and intelligibility is not well understood. While one intuitively would state that a better quality would imply a better intelligibility, the contrary can also be true. In our case, for instance, informal listening tests show that the high quality score of PSM is penalized in naturalness and intelligibility, confirming that PESQ measure can not substitute a formal subjective evaluation test (see Sec. 4, where the results of a perceptual test are shown).

Table 1: *Objective evaluation*

Sentences	**GTM**	**lMMSE**	**SM**	**PSM**
FS10	2.61	2.06	2.39	2.75
FS05	2.30	1.74	2.07	2.43
MB10	2.70	2.53	2.66	2.81
MB05	2.46	2.18	2.28	2.50

4. ASSESSMENT

To validate the system, a listening test was organized. 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.

Restoration of the noisy stimuli was performed by means of the algorithm, based on the Extended Kalman Filter, detailed in [18] (in de-click mode), then using the our SM algorithm, as well as the following commercial products:

1. X-Noise of Waves Restoration bundle (Waves V6 Update 2);
2. Denoiser (enable its *Musical noise suppression* filter) of iZotope RX v1.06;
3. Auto Dehiss of CEDAR Tools;
4. Adobe Audition 3.0;
5. Audacity 1.3.6 (an open source software for recording and editing audio signals).

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 24 restored stimuli were produced.

Test method. The tests were conducted using the EBU MUSHRA test method [19], which is a recommended evaluation method adopted by ITU [20]. 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): 6 researchers of the University of Padova and 12 students of the Conservatorio of Music "Cesare Pollini" of Padova.
2. Musically untrained (MU): 9 students in Multimedia Communication (University of Udine) and 9 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 iMac Intel Core 2 Duo with 2 GB 800 MHz DDR2 SDRAM (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 1 hour, 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. 2 as mean grades. The results from three listeners (all of them belong to MU group) were removed because the mean of their rates (in absolute value) on hidden references is greater than $+/-0.5$.

In *Il funerale di Rodolfo Valentino* (a speech signal) CEDAR and our tool achieved the best scores in both perceptual experiments (with MT as well as MU listeners). In general the only three systems with a score > 2 are our Tool, CEDAR and iZotope RX. The quality range between the best and worst restoration system is only 1.21.

Discussion. It is possible to make some important comments:

- All the restoration algorithms work quite well (i.e., the user's evaluation is good enough) with speech signals: see the scores achieved with the *Il funerale di Rodolfo Valentino*. In the authors' experience, in this case the listeners put their attention on the speech intelligibility, and this is enhanced with the restoration process. On the contrary, if there is singing voice or music, the focus is on the *naturalness* of the signal; typically, the noise reduction systems reduce the non harmonic part of the source and this fact can be interpreted as an artifact.
- The behaviors of the two listener groups are very different. The scores achieved by the restoration tools are very high in the case of musically untrained subjects (3.05 higher for our tools). Moreover, the anchors achieved scores 0.92 (7 kHz) and -0.83 (3.5 kHz) by the MU group; -2.69 (7 kHz) and -4.92 (3.5 kHz) by MT group. Probably, the musicians put attention on the *noise* (non harmonic part in the music signal and/or natural carrier noise): on the contrary, the untrained prefer a perfect clean signal, in which the harmonic information is evident.

  This result explains the importance to develop different tools, in relation to the aim considered.

Table 2: *Mean for restored stimuli and anchors, 34 subjects. Stimuli: S1* = Chi campa deritto campo aflitto*; S2* = Il funerale di Rodolfo Valentino*; S3* = La signorina sfinciusa*; S4* = Sta terra nun fa pi mia*. MT = Musically trained; MU = Musically untrained.*

[Grand averages]

Restoration system	**MT group**	**MU group**	**Average**
Our Tool	+0.80	+3.85	+2.33
CEDAR Tools	+1.97	+3.97	+2.98
iZotope RX	+1.10	+3.42	+2.26
Waves	+0.88	+2.93	+1.90
Audacity	+0.12	+2.43	+1.27
Adobe Audition	+0.12	+2.13	+1.12
Anchor 7 kHz	−2.69	+0.92	−1.02
Anchor 3.5 kHz	−4.92	−0.83	−2.87

[Musically trained: 18 subjects]

Restoration system	**S1**	**S2**	**S3**	**S4**
Our Tool	−0.30	+2.85	+0.70	−0.05
CEDAR Tools	+1.20	+2.70	+2.25	+1.72
iZotope RX	+0.50	+1.70	+1.65	+0.55
Waves	+0.40	+1.55	+1.40	+0.15
Audacity	−0.10	+1.20	+0.20	−0.80
Adobe Audition	−0.10	+1.35	+0.05	−0.80
Anchor 7 kHz	−3.50	−1.25	−3.00	−3.00
Anchor 3.5 kHz	−5.00	−4.67	−5.00	−5.00

[Musically untrained: 15 subjects]

Restoration system	**S1**	**S2**	**S3**	**S4**
Our Tool	+3.10	+4.85	+4.30	+3.15
CEDAR Tools	+4.20	+3.70	+4.25	+3.72
iZotope RX	+3.40	+3.70	+3.40	+3.17
Waves	+2.30	+3.55	+3.45	+2.43
Audacity	+2.60	+2.25	+2.20	+2.68
Adobe Audition	+1.60	+2.55	+2.25	+2.13
Anchor 7 kHz	+0.50	+1.00	+1.50	+0.67
Anchor 3.5 kHz	−1.00	+0.00	−1.00	−1.33

5. CONCLUSIONS

This study is focused on the restoration of single channel audio recordings of ethnic music. The problem of enhancing audio degraded by noise remains largely open, even though many significant techniques have been introduced over the past decades. This problem is severe when no independent information on the nature of noise degradation is available, in which case the enhancement

technique must utilize only the specific properties of given audio and noise signals. This is the most difficult task, because the noise and the speech are in the same channel. Many approaches have been reported in the literature: the most popular method, with many variants, is Short Time Spectral Attenuation (STSA). Although this method reduces the noise and improves the SNR, it mostly tends to introduce signal distortion and a perceptually annoying residual noise usually called musical noise (a special term for short sinusoids – tones – randomly distributed over time and frequency). It occurs due to imperfections in the original spectral subtraction technique and statistical inaccuracy in noise magnitude spectrum estimation.

In this paper we investigate the use of the Non-negative Matrix Factorization (NMF) method as an alternative to the STSA for the *digital curation* of musical heritage. 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, we introduce a perceptual suppression rule to determine how competitive is the perceptual domain, compared to the acoustic domain. To evaluate the proposed approach, both objective and subjective audio enhancement experiments were carried out (see Sections 3 and 4). The results of these 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 (i) will investigate the use of other advanced psychoacoustic models and (ii) will carry out an intensive application of this audio restoration environment on a real archive of ethnic music phonographic discs.

6. REFERENCES

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice (Signal Processing and Communications)*, CRC Press, 1 edition, June 2007.

[2] P. J. Wolfe and S. J. Godsill, "Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement," *EURASIP Journal on Appl. Sig. Processing*, vol. 10, no. 1, pp. 1043–1051, 2003.

[3] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, Wiley-IEEE Press, 2006.

[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. Körner, "Sparse coding and nmf," in *IEEE International Conference on Neural Networks*. 2004, pp. 2529–2533, IEEE.

[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, no. 4, 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*, 2007, pp. 70–74.

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

[10] 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.

[11] 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*, Aug 2007, pp. 431–436.

[12] 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.

[13] T. S. Gunawan, E. Ambikairajah, and J. Epps, "Perceptual speech enhancement exploiting temporal masking properties of human auditory system," *Speech Communication*, vol. 52, no. 5, pp. 381–393, December 2009.

[14] D. Tsoukalas, J. Mourjopoulos, and G. Kokkinakis, "Speech enhancement based on audible noise suppression," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 6, pp. 497–514, Nov. 1997.

[15] D. Tsoukalas, M. Paraskevas, and J. Mourjopoulos, "Speech enhancement using psychoacoustic criteria," in *IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 1993, pp. 359–362.

[16] ITU-T, "Recommendation p.862. perceptual evaluation of speech quality (pesq), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," Tech. Rep., 2001.

[17] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Communication*, vol. 49, no. 7-8, pp. 588–601, 2007.

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

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

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