NOISE-TO-MASK RATIO MINIMIZATION BY WEIGHTED NON-NEGATIVE MATRIX FACTORIZATION

Joonas Nikunen, Tuomas Virtanen

Department of Signal Processing, Tampere University of Technology
Korkeakoulunkatu 1, FI-33720 Tampere, Finland

ABSTRACT

This paper proposes a novel algorithm for minimizing the perceptual distortion in non-negative matrix factorization (NMF) based audio representation. We formulate the noise-to-mask ratio audio quality criterion in a form where it can be used in NMF and propose an algorithm for optimizing the criterion. We also propose a method for compensating the spreading of the representation error in the synthesis filterbank. The objective perceptual quality produced by the proposed method is found to outperform all the reference methods. We also study the trade-off between the window length and the rank of factorization with a fixed data rate, and find that the best performance is obtained with window lengths between 10 and 30 ms.

Index Terms— Non-negative matrix factorization, Noise-to-mask ratio, Audio coding, Signal representations

1. INTRODUCTION

In audio signal processing, an acoustic time-domain signal is often represented using a mid-level representation [1], which allows more efficient analysis or manipulation of the signal. Commonly used mid-level representations include, for example, the time-frequency representations such as the short-time Fourier transform (STFT), and parametric representations such as the sinusoidal model. More advanced models can take into account the structure of the sounds in more detail, for example by using a harmonic model [2]. The latter two can be also viewed as lossy compression, since they reduce the amount of information needed to approximate the original signal. The parameters of a representation can be estimated by using a statistical criterion such as the mean-square error, but also the properties of the human audio perception can be taken into account.

All present-day perceptual audio coders are essentially based on a sub-band bit allocation upon a psychoacoustical masking model. They quantize a time-frequency representation of audio signal in such way that the quantization noise stays below the masking threshold and thus remains inaudible [3]. An objective measure of the perceptual quality of a compressed signal is the noise-to-mask ratio (NMR) [4], which measures the relative level of the quantization noise in comparison with the masking threshold. An alternative approach to audio compression is object-based audio coding, where individual sound sources or objects (e.g. musical instruments, speakers, notes) in an audio recording are represented separately [5]. Object-based coding allows using the most efficient codec for each object, but as well interactive synthesis of the signal.

Recently, non-negative matrix factorization (NMF) has been applied in many audio signal processing tasks, such as sound source separation [6]. Its main advantage is the ability to automatically decompose a mixture signal into a representation where each sound source is represented as an individual object [6]. The NMF decomposition also effectively finds repetitive structures in the signal, thus being able to reduce redundancy and being attractive from signal compression point of view.

This paper proposes a novel algorithm for NMF which minimizes the noise-to-mask ratio of the signal decomposition. The NMR objective is formulated as a cost function for NMF and it is minimized using a weighted NMF algorithm. We also propose to filter the estimated masking patterns in time, which effectively reduces the pre-echo caused by the spreading of errors in the synthesis filterbank. Potential applications of the proposed method include object-based audio coding and analysis of audio signals.

The block diagram of the proposed system is shown in Figure 1. First, the magnitude spectrogram of an input signal is calculated for the NMF algorithm. Masking thresholds are estimated from an input signal, which are then used for NMF weighting. Approximation of original spectrogram is obtained from the weighted NMF algorithm and the signal is reconstructed by assigning the original phases to it and taking the inverse FFT. Frames are finally combined in the synthesis filterbank by overlap-add.

The structure of the paper is as follows: Section 2 gives short review of the noise-to-mask ratio which is the objective of the proposed method. In Section 3 we derive a weighted cost function for NMF corresponding to the NMR. Section 4 presents a synthesis procedure and proposes a technique to reduce the pre-echo effect. The proposed method is compared to conventional NMF algorithms in Section 5. Section 5 also presents results from an experiment on finding the best combination of coding parameters in case of constant data rate.

2. NOISE-TO-MASK RATIO

Human hearing includes a masking phenomenon, which causes low-intensity frequency components to become masked by more intense ones, that occur spatially and temporally close to each other. It means that a loud frequency component can make a fainter component become completely inaudible to our hearing [7, p. 56]. The masking concept can be utilized in audio coding, where it is used to decide, which parts of the audio can be disregarded without perceptual difference.

A quality metric to measure the audibility of distortions is the noise-to-mask ratio, which was introduced by Brandenburg [4]. The metric consists of the following processing steps: 1) The error between a distorted signal and a reference signal is calculated. 2) The masking threshold is estimated from the reference. 3) The noise-to-mask ratio in each time frame is calculated in Bark scale. 4) The final measure is average over all the time-frequency points. Distortions having a NMR value of -10 dB or below can be assumed to be

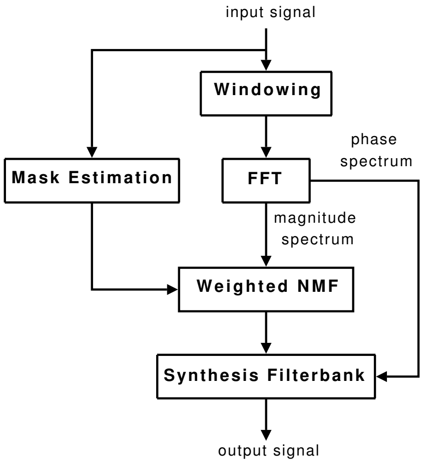

Fig. 1. Block diagram of proposed method

inaudible.

NMR has been included into recommendation BS.1387 [8] for perceptual evaluation of audio quality (PEAQ). The recommendation includes specifications for the auditory model to be used for estimating the masking threshold required for NMR evaluation. PEAQ auditory model (with clarifications from [9]) is used here for masking threshold estimation. The model includes parameter L_p for scaling the mask estimation to correspond to desired listening sound pressure level (SPL). This is due to the fact that spatial and temporal spreading functions are dependent on the energy of the masker component.

The NMR in PEAQ can be described using the equation

$$\mathrm{NMR_B} = 10\log_{10}\Big(\frac{1}{T}\sum_{t=1}^{T}\frac{1}{B}\sum_{b=1}^{B}[\mathbf{M}]_{b,t}[\mathbf{CH}(\mathbf{X}-\hat{\mathbf{X}})^{\cdot 2}]_{b,t}\Big), \quad (1)$$

and it consists of the operations below: 1) Squared difference between the magnitude spectrograms of the original signal $\mathbf{X}$ and the estimated signal $\hat{\mathbf{X}}$ is calculated. $\mathbf{X}^{\cdot 2}$ denotes element-wise power of two. The spectrograms are calculated using a 42.7 ms Hanning window and discrete Fourier transform (DFT). 2) The error is weighted by middle- and outer ear transfer function, which is implemented by multiplying the squared error spectrogram by a diagonal matrix $\mathbf{H}$ having the values of the transfer function on the diagonal. 3) The error is decimated to a bark scale representation, which is implemented by multiplication by matrix $\mathbf{C} \in \mathbb{R}^{\geq 0, B\times K}$, where each row contains the power response of a bark band for all the DFT indices. 4) The error in bark scale is weighted by $\mathbf{M} \in \mathbb{R}^{\geq 0, B\times T}$, which is the element-wise inverse of the masking threshold in each frame t and bark band b. Both the error and masking patterns are having a quarter bark band frequency resolution, which results to 109 bands with 48 kHz sampling frequency. 5) The results are averaged over frequency and time and converted to the dB scale. T is the total number of frames, and the total number of bark bands is B.

3. PROPOSED PERCEPTUALLY WEIGHTED NMF

NMF approximates the observation matrix $\mathbf{X} \in \mathbb{R}^{\geq 0, K\times T}$ as a product of basis matrix $\mathbf{B} \in \mathbb{R}^{\geq 0, K\times R}$ and gain matrix $\mathbf{G} \in \mathbb{R}^{\geq 0, R\times T}$ as $\mathbf{X} \approx \mathbf{BG}$. Matrix $\mathbf{X}$ consists of magnitudes of frame-wise DFTs of the observed audio signal, calculated in frames $t = 1, \dots, T$. Only positive frequencies $k = 1, \dots, K$ of the DFT are used. The rank of the decomposition is denoted by R, which is a free parameter chosen by the user.

Matrices $\mathbf{B}$ and $\mathbf{G}$ are estimated by minimizing the error of the approximation. Measures for the error include, for example, the squared Euclidean distance (EUC), generalized Kullback-Leibler divergence (KLD), and the Itakura-Saito divergence (ISD) [10].

3.1. NMR as cost function for NMF

The masking thresholds in $\mathbf{M}$ for certain observations $\mathbf{X}$ are calculated before the NMF algorithm. The mask estimation and NMR evaluation in PEAQ is defined in bark scale, but due to its lower resolution, we wish to perform the NMF decomposition in a linear frequency scale provided by the DFT. In the following we formulate the NMR objective into a weighted squared error, calculated in a linear frequency scale. Let use denote the squared error in Equation (1) as $\mathbf{E} = (\mathbf{X} - \hat{\mathbf{X}})^{\cdot 2}$. The measure (1) is a monotonic function ($\log_{10}$ and scalar multipliers) of term $\sum_{t=1}^{T}\sum_{b=1}^{B}[\mathbf{M}]_{b,t}[\mathbf{CHE}]_{b,t}$. Thus, minimizing the NMR is equivalent to minimizing the above term. In each frame t, the term can be formulated as

$$\sum_{b=1}^{B}[\mathbf{M}]_{b,t}[\mathbf{CHE}]_{b,t} = \sum_{k=1}^{K}\sum_{b=1}^{B}[\mathbf{M}]_{b,t}[\mathbf{CH}]_{b,k}[\mathbf{E}]_{k,t}$$
$$= \sum_{k=1}^{K}[\mathbf{W}]_{k,t}[\mathbf{E}]_{k,t}\ , \quad \text{where} \quad \mathbf{W} = (\mathbf{CH})^{\mathrm{T}}\mathbf{M}$$

The above formulation can be placed back to Equation (1) and the result is an NMR metric defined for linear frequency scale error

$$\mathrm{NMR_L} = 10\log_{10}\Big(\frac{1}{T}\sum_{t=1}^{T}\frac{1}{B}\sum_{k=1}^{K}[\mathbf{W}]_{k,t}[\mathbf{X}-\hat{\mathbf{X}}]_{k,t}^{2}\Big). \quad (2)$$

When applying the above equation as NMF cost function, we model $\hat{\mathbf{X}}$ using $\mathbf{BG}$. The resulting NMF criterion is the weighted squared Euclidean distance:

$$\mathrm{D}_{\mathrm{WEUC}}(\mathbf{X}, \mathbf{BG}, \mathbf{W}) = \sum_{kt}[\mathbf{W}]_{k,t}([\mathbf{X}]_{k,t} - [\mathbf{BG}]_{k,t})^2, \quad (3)$$

The NMR quality criterion has been also implemented as cost function for NMF by O'Grady in [11]. His method calculated the error between the observed magnitude spectrogram and the model in bark bands, which does not allow modeling the fine spectral structure, that the linear frequency scale models.

3.2. Algorithm for minimizing the NMR

The weighted squared Euclidean distance and thus the proposed cost function can be minimized by the update rules proposed in [12] and applied in [11]. First, the entries of matrices $\mathbf{B}$ and $\mathbf{G}$ are initialized with random values normally distributed between zero and one. The matrices are updated iteratively using the update rules

$$\mathbf{B} \leftarrow \mathbf{B}.\times\frac{(\mathbf{W}.\times\mathbf{X})\mathbf{G}^{\mathrm{T}}}{(\mathbf{W}.\times(\mathbf{BG}))\mathbf{G}^{\mathrm{T}}}$$
$$\mathbf{G} \leftarrow \mathbf{G}.\times\frac{\mathbf{B}^{\mathrm{T}}(\mathbf{W}.\times\mathbf{X})}{\mathbf{B}^{\mathrm{T}}(\mathbf{W}.\times(\mathbf{BG}))}, \quad (4)$$

where operators $.\times$ and $\frac{\mathbf{X}}{\mathbf{Y}}$ denote element-wise multiplication and division, respectively. The update rules are repeated until the algorithm converges.

4. SIGNAL RECONSTRUCTION AND WEIGHT SMOOTHING

The above section described the model parameter estimation stage of the algorithm. In signal analysis the estimated parameters can be used as such, but for example in audio coding applications a signal needs to be reconstructed from the parameters.The synthesis procedure requires generating the phases for the reconstructed magnitude spectrogram $\mathbf{BG}$, applying inverse DFT in each frame, and combining the frames by overlap-add.

An example of an algorithm that can be used to generate the phases has been proposed in [13]. Our main focus in this study is in the magnitude spectrogram modeling and in order to prevent the artefacts caused by the phase reconstruction from affecting the evaluation, we use the phase spectrogram estimated from the original signal, as illustrated in Figure 1.

The NMF cost function derived in the previous section does not take into account the synthesis procedure, i.e., it assumes that the magnitude spectrogram of the synthesized signal equals $\hat{\mathbf{X}}$ in (1). In practice, the overlap-add synthesis procedure affects the quality in the sense that an error produced in a frame is spread to the neighboring frames where it may become audible. Specifically, the phenomenon becomes prominent if a quiet frame is followed by an intense one where fair amount of error is produced. In audio coding the phenomenon is called pre-echo.

We approximate the effect of the synthesis procedure by assuming that the modeling error $[\mathbf{E}]_{k,t}$ of the magnitude spectrograms in frame t is divided into frames $t-1$, t, and $t+1$ by weights h_{-1}, h_0, and h_1, respectively. We use values α, $1-2\alpha$, and α for the weights, where the amount of spreading defined by the parameter α is dependent on the shape of the window function. We also assume that the errors produced in adjacent frames are independent from each other, so that the errors (represented by energies) are additive. In practice the spreading depends on the lengths and relative positions of the windows of the synthesis filter bank and the analysis filter bank in NMR, but for simplicity we restrict ourselves to the above approximation. The spread error is given as

$$\frac{1}{T}\sum_{t=1}^{T}\frac{1}{B}\sum_{\tau=-1}^{1}\sum_{k=1}^{K}[\mathbf{W}]_{k,t}[\mathbf{E}]_{k,t-\tau}h_{\tau},$$

which can be formulated as

$$\frac{1}{T}\sum_{t=1}^{T}\frac{1}{B}\sum_{k=1}^{K}[\mathbf{W}']_{k,t}[\mathbf{E}]_{k,t},$$

where $[\mathbf{W}']_{k,t} = \sum_{\tau=-1}^{1}[\mathbf{W}']_{k,t+\tau}h_{\tau}$. Thus the effect of the synthesis filterbank can be taken into account by filtering the weights $\mathbf{W}$ in time. The simulation results show that the overall quality is slightly improved by the spreading.

5. SIMULATION AND RESULTS

The proposed NMF algorithm was tested by applying it to various styles of audio signals and measuring the NMR of the synthesized signals. The test set consisted of 10-second monaural excerpts from following categories (number of entries in brackets): classical music (16), drum patterns (20), western pop music (24), solo instruments (20), solo singing (10) and speech (10), equaling to total of 100 samples. The speech samples have a 16 kHz sampling frequency, whereas the rest of them have a 44.1 kHz sampling frequency.

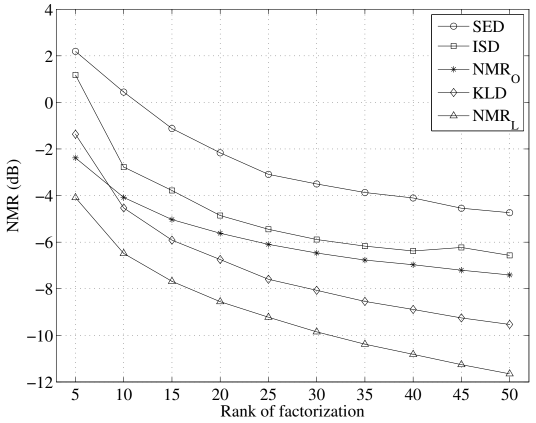

Fig. 2. NMR of the tested NMF algorithms as the function of the rank of factorization

Each test sample was processed using the method illustrated in Figure 1. We used Kaiser-Bessel derived window function [3, p 171] in analysis and synthesis, since it was found to produce the best performance among various tested window functions. We used 50% overlap between adjacent windows. The synthesized signals were evaluated with NMR criterion described in PEAQ and the average NMR over the whole test set was calculated. The scaling parameter L_p was set to 40 dB.

The tested NMF algorithms were EUC, KLD, ISD and proposed $\mathrm{NMR_L}$. The weighting method from [11] is denoted as $\mathrm{NMR_O}$. The masking estimation for NMF was done using a 42.7ms window, but the hop size was set equal to the NMF hop size. The number of iterations was chosen by calculating NMR after each iteration to determine the rate of convergence for a subset of the test signals. The experiments showed that EUC and $\mathrm{NMR_L}$ needed more iterations to converge. The number of iterations was set to 200 for KLD and ISD and 400 for EUC and $\mathrm{NMR_L}$.

The results of different ranks of factorization with a 20 ms window are shown in Figure 2. Results indicate that the proposed method enables on average 1.9 dB better NMR than the best reference method. The test was also repeated for 40ms window and the results were very similar, the advantage of the proposed method being again approximately 1.6 dB. Few demonstrative test signals are available at http://www.cs.tut.fi/sgn/arg/nikunen/demo/icassp2010/.

Increasing the hop size will reduce the amount frames per second. From audio coding point of view this decreases the amount of gains to be represented. The number of frequency indices for each source in $\mathbf{B}$ is half of the window length, since the DFT length equals the window length and only positive frequencies are retained. We restrict the hop size to be 50% of the window size, and therefore longer windows will result to longer DFTs, which need to be encoded as well. We consider each parameter to be represented as a particle, and study the effect of the frame length and the rank of factorization when constraining a fixed amount of particles per second. The total amount of particles per second in a decomposition is $P = (Z + K/S)R$, where Z denotes the number of frames per second, K is the number of positive DFT coefficients, S is the signal length in seconds and R is the rank of factorization.

We fixed the amount particles per second to 3000, and determined the parameters by selecting a certain rank of factorization and searching for the shortest possible window that did not exceed the particle rate. The results with different ranks of factorization are

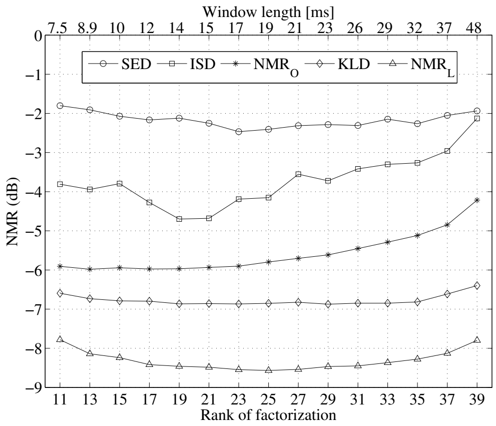

Fig. 3. NMR as the function of the window length and the rank of factorization when 3000 particles per second are used

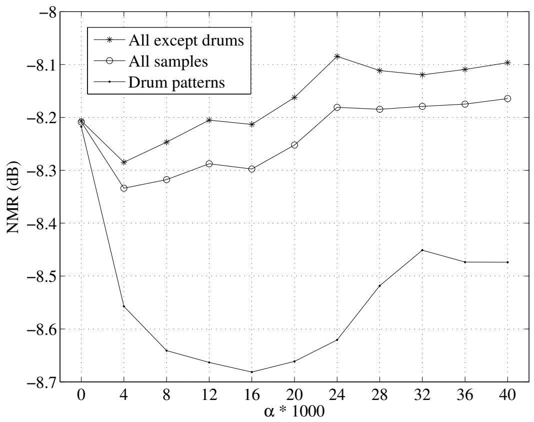

Fig. 4. Weights $\mathbf{W}$ filtered with different time averaging filters, NMR evaluated over whole test set, without drum patterns and only drum patterns

shown in Figure 3. For this test we used 30-second excerpts where the total number of samples was 50. The window lengths depend on the sampling frequency. In the figure they are denoted for the signals with sampling frequency of 44100 Hz. Considering the average quality, the range of equally good parameter combinations seems to be wide for all the NMF algorithms. The quality decreases only when a too short or a too long window is used. By examining the results of individual samples it seems that a good combination depends greatly on the signal to be composed.

Figure 4 illustrates the average NMR as the function of the spreading parameter α. The average is calculated separately for drum signals, which contain lot of transients, and thus the pre-echo phenomenon is assumed to be the largest. It can be seen that with a suitable value of α, the filtering improves the average quality NMR of drums by 0.4 dB. For other signals the filtering does not improve the quality.

6. CONCLUSION

We have proposed a method for minimizing the noise-to-mask ratio using non-negative matrix factorization. We have formulated the noise-to-mask ratio calculated on bark-band signal representation as a cost function for linear-frequency NMF. Simulation experiments show that the proposed method allows better quantitative perceptual quality than the reference methods. The proposed method for spreading the masking patterns in time enables a better quality for signals with plenty of transient sounds. The overall results show improvement of audio quality in benefit for proposed method and it could be plausible for future object-based audio coding applications.

7. REFERENCES

[1] D. Ellis and D.F. Rosenthal, “Mid-level representations for computational auditory scene analysis,” in *Proceedings of International Joint Conference on Artificial Intelligence – Workshop on Computational Auditory Scene Analysis*, Montreal, Canada, 1995.

[2] E. Vincent and M.D. Plumbley, “Low bit-rate object coding of musical audio using Bayesian harmonic models,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1273–1282, 2007.

[3] Andreas Spanias, Ted Painter, and Venkatraman Atti, *Audio Signal Processing and Coding*, John Wiley & Sons, 2007.

[4] K. Brandenburg and T. Sporer, “NMR and Masking Flag: Evaluation of Quality Using Perceptual Criteria,” in *Proceedings of the AES 11th International Conference on Test and Measurement*, Portland, USA, May 1992, pp. 169–179.

[5] E. D. Scheirer, “Structured audio and effects processing in the MPEG-4 multimedia standard,” *Multimedia Systems*, vol. 7, no. 1, pp. 11–22, 1999.

[6] Tuomas Virtanen, “Monoaural Sound Source Separation by Nonnegative Matrix Factorization With Temporal Continuity and Sparseness Criteria,” *IEEE Transactions on Audio, Speech and Language processing*, vol. 15, pp. 1066–1074, 2007.

[7] E. Zwicker and H. Fastl, *Psychoacoustics Facts and Models*, Springer-Verlag, 1990.

[8] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerends, and C. Colomes, “PEAQ - The ITU Standard for Objective Measurement of Perceived Audio Quality,” *Journal of the Audio Engineering Society*, vol. 48, pp. 3–29, 2000.

[9] P. Kabal, “An Examination and Interpretation of ITU-R BS.1387: Perceptual Evaluation of Audio Quality,” Tech. Rep., Department of Electrical & Computer Engineering, McGill University, 2002.

[10] C. Févotte and Cemgil A. T., “Nonnegative matrix factorizations as probabilistic inference incomposite models,” in *17th European Signal Processing Conference*, Scotland, 2009.

[11] Paul D. O’Grady, *Sparse Separation of Under-Determined Speech Mixtures*, Ph.D. thesis, Nui Maynooth, 2007.

[12] Tuomas Virtanen, “Separation of Sound Sources by Convolutive Sparse Coding,” in *ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing*, Jeju, Korea, 2004.

[13] J. Le Roux, N. Ono, and S. Sagayama, “Explicit consistency constraints for STFT spectrograms and their application to phase reconstruction,” in *ISCA Tutorial and Research Workshop on Statistical and Perceptual Audition*, Brisbane, Australia, 2008.