Algorithms for audio inpainting based on probabilistic nonnegative matrix factorization

Ondřej Mokrý, Paul Magron, Thomas Oberlin, Cédric Févotte

Algorithms for audio inpainting based on probabilistic nonnegative matrix factorization

Ondřej Mokrý[a,*], Paul Magron[b], Thomas Oberlin[c], Cédric Févotte[d]

[a] *Brno University of Technology, Faculty of Electrical Engineering and Communications, Department of Telecommunications, Technická 12, 616 00 Brno, Czech Republic*
[b] *Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France*
[c] *ISAE-SUPAERO, Université de Toulouse, France*
[d] *IRIT, Université de Toulouse, CNRS, Toulouse, France*

Abstract

Audio inpainting, i.e., the task of restoring missing or occluded audio signal samples, usually relies on sparse representations or autoregressive modeling. In this paper, we propose to structure the spectrogram with nonnegative matrix factorization (NMF) in a probabilistic framework. First, we treat the missing samples as latent variables, and derive two expectation–maximization algorithms for estimating the parameters of the model, depending on whether we formulate the problem in the time- or time-frequency domain. Then, we treat the missing samples as parameters, and we address this novel problem by deriving an alternating minimization scheme. We assess the potential of these algorithms for the task of restoring short- to middle-length gaps in music signals. Experiments reveal great convergence properties of the proposed methods, as well as competitive performance when compared to state-of-the-art audio inpainting techniques.

Keywords: alternating minimization, audio inpainting, expectation–maximization, nonnegative matrix factorization

1. Introduction

Audio inpainting [1] is an inverse problem aiming at restoring audio signals degraded by sample loss. Such a problem typically occurs as a result of packet loss during transmission (packet loss concealment [2, 3]) or in digitization of physically degraded media. Inpainting can also be used to restore signal samples subject to a degradation so heavy that the information about the samples can be considered lost. Formally, let $\mathbf{y} \in \mathbb{C}^L$ denote the original time-domain signal prior to the degradation (we consider complex-valued signals for the sake of generality). The goal of inpainting is to estimate $\hat{\mathbf{y}} \in \mathbb{C}^L$, given an incomplete observation of $\mathbf{y}$. This problem is ill-posed because of

*The work was supported by the Czech Science Foundation (GAČR) Project No. 20-29009S, the French ANITI Project No. ANR-19-PI3A-0004 and the European Research Council (ERC) Project FACTORY No. 6681839. Part of the work was realized during the stay of O. Mokrý at IRIT, co-financed from the Erasmus+ mobility program. The authors would like to thank Pavel Rajmic and Pierre-Hugo Vial for their inputs concerning both the development of the methods and the preparation of the manuscript.

*Corresponding author

Email addresses: ondrej.mokry@vut.cz (Ondřej Mokrý), paul.magron@inria.fr (Paul Magron), thomas.oberlin@isae-supaero.fr (Thomas Oberlin), cedric.fevotte@irit.fr (Cédric Févotte)

the missing samples, and even the observed samples are prone to some measurement error or noise. Restoring the signal thus requires some structure assumptions about the original signal, in order to guide the estimation towards the most desirable solution.

One of the earliest, yet most successful approaches to audio inpainting is to assume the underlying autoregressive (AR) nature of clean audio signals: Janssen's iterative method [4] is still among state-of-the-art methods to date. Great performance is also achieved by Etter's extrapolation-based technique [5], however, it is limited by the need for a sufficiently long clean context to reconstruct each individual gap. Recently, the class of sparsity-based methods methods has emerged [1, 6, 7, 8]. Generally speaking, these methods solve a regularized inverse problem, where the solution is assumed to have a sparse time-frequency (TF) spectrum, while fitting the observed temporal samples.

A disadvantage of most methods is the local nature of the regularizers. For example, sparsity-based methods are effective for inpainting of short- to medium-length gaps, typically up to 50 ms, or for restoring randomly subsampled signals [9, 10]. Even for gaps of length of tens of milliseconds, there is a need for strong regularization that not only exploits the local TF sparsity but also the global structure of audio signals. A step towards using global properties of audio signals is the so-called social sparsity [11, 12, 13, 14], where the significant TF coefficients are expected to occur in particular transient or temporal patterns, or recent approaches based on generative deep neural networks [15, 16].

In the present paper, we focus on audio inpainting methods that leverage the low-rank structure of audio signals in the TF domain. Among low-rank models, nonnegative matrix factorization (NMF) has been intensively used for analysis and decomposition, both in machine learning and signal processing [17, 18, 19]. Concerning audio, NMF can be used to decompose the signal's power or magnitude spectrogram as the product of nonnegative matrices $\mathbf{WH}$, where $\mathbf{W}$ is a dictionary of spectral patterns, and $\mathbf{H}$ contains the temporal activations of these patterns. Such a decomposition provides a semantically reasonable generative model for audio signals, which is one of the reasons why NMF is among the classical approaches to source separation [20, 21]. It also benefits from being an unsupervised and interpretable method while being computationally cheap. NMF has also been used as a prior for the restoration of degraded signals [22, 23], thus being successfully applied to audio declipping [22, 24].

The main idea of NMF-based signal reconstruction can be summarized as building an estimation problem where the NMF is used as a prior, which requires to estimate parameters given the incomplete data, usually via the expectation–maximization (EM) algorithm [25]. Formally, the structure of the spectrogram is introduced via the probabilistic Gaussian composite model [26], which results in applying NMF with the Itakura–Saito divergence. However, the estimation of the parameters – the matrices $\mathbf{W}$ and $\mathbf{H}$ – is a non-trivial problem which can be approached in different ways. One possibility is to derive a generalized EM algorithm to estimate the factorization of the power spectrogram of the clean signal in the maximum likelihood (ML) sense. This has been previously presented by Bilen, Ozerov and Pérez [22, 23] with application to audio declipping, where it reaches state-of-the-art performance. However, the method is known to be computationally very demanding [24, Sec. V.D], and it represents only one of several possible approaches to treating the missing samples (specifically, they are treated as latent variables). Furthermore, it has not been compared to the state-of-the-art methods in the audio inpainting setting.

Drawing on the work of Bilen et al., we provide a novel generalization of their EM-based approach, where the missing samples are formally treated as latent variables. We formulate the estimation problem in both time and TF domains and therefore derive two algorithms, among which one is novel. As a novel approach, we also propose to treat the missing samples as parameters.

This leads to a new estimation problem, for which we derive an alternating minimization (AM) scheme. Even though we built upon the application in audio inpainting, the core of the work is the conceptual development of novel approaches to the NMF-based modeling in signal restoration. We conduct experiments on the task of restoring short- to middle-length gaps in music signals. These reveal great convergence properties of the proposed methods, as well as competitive performance when compared to state-of-the-art audio inpainting techniques.

The rest of the paper is organized as follows. We formulate audio inpainting with NMF as an estimation problem in Section 2. In Section 3 we review and extend the approach by Bilen et al. [22]. In Section 4 we consider the missing samples as parameters, for which we derive a novel AM estimation approach. Section 5 is devoted to experiments and evaluation of the proposed methods. Finally, Section 6 concludes the paper.

2. Problem formulation

For the whole derivation of the methods, it is convenient to divide the signal into windowed temporal frames $\mathbf{x}_n \in \mathbb{C}^M$, $n = 1, \ldots, N$, potentially arranged in a matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$. In the case of inpainting, we observe the samples $\mathbf{x}_n^{\text{obs}} = \mathbf{M}_n \mathbf{x}_n$ in each frame n and we aim at obtaining an estimate $\hat{\mathbf{x}}_n$ of the whole frame given these observed samples. The matrix $\mathbf{M}_n$ is constructed from the identity matrix by omitting the rows corresponding to the missing samples, thus the multiplication with $\mathbf{M}_n$ shortens the vector, selecting only the observed samples. Note that in the present work, the *indices* of the missing samples are assumed to be known, thus the matrices $\mathbf{M}_n$, $n = 1, \ldots, N$ are known. The whole signal estimate $\hat{\mathbf{y}} \in \mathbb{C}^L$ is then obtained by folding all $\hat{\mathbf{x}}_n$ together by the overlap-add procedure.

To propose a probabilistic formulation of the inpainting problem, it is necessary to postulate a statistical model for the data at hand. Since we aim at promoting the low-rank structure of the TF coefficients $\mathbf{S} = [\mathbf{s}_1, \ldots, \mathbf{s}_N] = [s_{fn}] \in \mathbb{C}^{F \times N}$ of the audio signal, we first define the following synthesis model:

$$\mathbf{x}_n = \mathbf{T}\mathbf{s}_n, \quad n = 1, \ldots, N, \tag{1}$$

or, in matrix form, $\mathbf{X} = \mathbf{T}\mathbf{S}$, where $\mathbf{T} \in \mathbb{C}^{M \times F}$ is a linear reconstruction operator. The matrix $\mathbf{T}$ typically represents the inverse discrete Fourier transform (DFT). If the temporal frames are weighted by a windowing function, $\mathbf{S}$ represents the short-time Fourier transform (STFT) coefficients of the whole signal $\mathbf{y}$. We will discuss the effect of particular choices of $\mathbf{T}$ later in 3.3.

The low-rank structure of the TF coefficients can be formalized within the following assumptions.

Assumption 1 (Gaussian coefficients)**.** *The time-frequency coefficients are treated as conditionally mutually independent,*[1] *and each coefficient follows a complex circular zero-mean Gaussian distribution:*

$$s_{fn} \sim \mathcal{N}(0, v_{fn}). \tag{2}$$

Due to the assumed independence of the individual TF coefficients, we can rewrite the distribution of the spectrum of the n-th frame as:

$$\mathbf{s}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{D}_n), \quad \mathbf{D}_n = \operatorname{diag}\left([v_{fn}]_{f=1,\ldots,F}\right), \tag{3}$$

[1] Note that temporal Markov NMF models such as in [27] could readily be considered but we use independence as a working assumption for ease of presentation.

where the symbol $\mathcal{N}$ from now on denotes the multivariate complex Gaussian distribution.

Remark 1. *Note that Assumption 1 is stated for the case of a generic linear transform* $\mathbf{T}$ *interconnecting the TF coefficients and temporal samples. In the common case where the TF coefficients are computed from the temporal samples by the DFT (i.e.,* $\mathbf{T}$ *represents the inverse DFT operator), and the temporal signal is real-valued, the independence assumption needs to be relaxed: The TF coefficients of real-valued signals are independent only in half of the frequency spectrum, the other half being determined via the Hermitian property of the transform. However, for the sake of generality, we will assume no particular structure of* $\mathbf{T}$ *throughout the article.*

Assumption 2 (NMF structure of the variances)**.** *The variance matrix* $\mathbf{V} = [v_{fn}]$ *has the low-rank NMF structure:*

$$v_{fn} = \sum_{k=1}^{K} w_{fk} h_{kn}, \tag{4}$$

where K *is small and all parameters are nonnegative. This model amounts to* $\mathbf{V} = \mathbf{W}\mathbf{H}$ *with* $\mathbf{W}$ *and* $\mathbf{H}$ *being* $F \times K$ *and* $K \times N$ *nonnegative matrices, respectively [22, Sec. 2.2].*

To estimate the parameters $\mathbf{W}, \mathbf{H}$ of the variance matrix, given the observed samples $\mathbf{X}^{\text{obs}} = \{\mathbf{x}_1^{\text{obs}}, \dots, \mathbf{x}_N^{\text{obs}}\}$, we employ maximum likelihood (ML) estimation. The audio inpainting *per se*, i.e., the computation of the missing samples, is then performed explicitly given the estimated parameters and the reliable samples in a way similar to Wiener filtering.

However, there is not a unique way to formulate the ML problem. The following sections present two approaches that differ in whether the missing samples are treated as latent variables (Section 3) or explicitly treated as parameters (Section 4). Note that subsection 3.1 reformulates the method proposed by Bilen et al. in [23], whereas the rest of Section 3 and the whole Section 4 are new contributions.

3. ML estimation by treating the missing samples as latent variables

We first present the ML formulation, where the goal is to estimate the parameters $\mathbf{W}, \mathbf{H}$ of the distribution of the restored signal in the TF domain, by minimizing the negative log-likelihood of the observed samples $\mathbf{X}^{\text{obs}} = \{\mathbf{x}_1^{\text{obs}}, \dots, \mathbf{x}_N^{\text{obs}}\}$, as presented by Bilen et al. in [23]. The problem can be formalized as

$$\hat{\mathbf{W}}, \hat{\mathbf{H}} = \operatorname*{arg\,min}_{\mathbf{W},\mathbf{H}} - \log p(\, \mathbf{X}^{\text{obs}} \mid \mathbf{W}, \mathbf{H}\,). \tag{5}$$

This expression can be broken down for a single frame n, where the probability p is given by the distribution of $\mathbf{s}_n$ from Eq. (3) and by the linear observation model as

$$\mathbf{x}_n^{\text{obs}} = \mathbf{M}_n \mathbf{x}_n = \mathbf{M}_n \mathbf{T} \mathbf{s}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{M}_n \mathbf{T} \mathbf{D}_n \mathbf{T}^{\mathsf{H}} \mathbf{M}_n^{\mathsf{T}}), \quad n = 1, \dots, N, \tag{6}$$

where the dependence on $\mathbf{W}, \mathbf{H}$ is contained within the definition of the matrices $\mathbf{D}_n$. We resort to the EM algorithm [25], in line with Bilen et al. [23]. However, we broaden their work by considering two different settings of the EM algorithm, depending on the domain of the complete data to be estimated.

3.1. EM-tf

First, we briefly describe the EM algorithm for the problem (5). The setting is that the *incomplete data* corresponds to the observed reliable signal, i.e., $\mathbf{X}^{\text{obs}} = \{\mathbf{x}_1^{\text{obs}}, \dots, \mathbf{x}_N^{\text{obs}}\}$ in the framed time domain. The *complete data* corresponds to the TF spectrum $\mathbf{S} = [\mathbf{s}_1, \dots, \mathbf{s}_N] \in \mathbb{C}^{F \times N}$ of the original signal (which is reflected by the abbreviation EM-*tf*). Finally, the parameters to be estimated are $\boldsymbol{\theta} = \{\mathbf{W}, \mathbf{H}\}$, and $\tilde{\boldsymbol{\theta}}$ is the current value of the parameters.

Using this setting, the EM algorithm aims at minimizing the functional:

$$Q(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) = -\int \log p(\mathbf{S} \mid \boldsymbol{\theta}) p(\mathbf{S} \mid \mathbf{X}^{\text{obs}}, \tilde{\boldsymbol{\theta}}) \, \mathrm{d}\mathbf{S} \tag{7}$$

by iterating two steps:

1. **E-step**: compute $Q(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}})$,
2. **M-step**: update $\tilde{\boldsymbol{\theta}}$ by minimizing $Q(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}})$ with respect to $\boldsymbol{\theta}$.

For the E-step, we obtain

$$p(\mathbf{S} \mid \mathbf{X}^{\text{obs}}, \tilde{\boldsymbol{\theta}}) = \prod_{n=1}^{N} p(\mathbf{s}_n \mid \mathbf{x}_n^{\text{obs}}, \tilde{\boldsymbol{\theta}}) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{s}_n \mid \hat{\mathbf{s}}_n, \hat{\boldsymbol{\Sigma}}_n), \tag{8}$$

where

$$\hat{\mathbf{s}}_n = \mathbf{D}_n \mathbf{T}^{\mathsf{H}} \mathbf{M}_n^{\mathsf{T}} \left(\mathbf{M}_n \mathbf{T} \mathbf{D}_n \mathbf{T}^{\mathsf{H}} \mathbf{M}_n^{\mathsf{T}} \right)^{-1} \mathbf{x}_n^{\text{obs}}, \tag{9a}$$

$$\hat{\boldsymbol{\Sigma}}_n = \mathbf{D}_n - \mathbf{D}_n \mathbf{T}^{\mathsf{H}} \mathbf{M}_n^{\mathsf{T}} \left(\mathbf{M}_n \mathbf{T} \mathbf{D}_n \mathbf{T}^{\mathsf{H}} \mathbf{M}_n^{\mathsf{T}} \right)^{-1} \mathbf{M}_n \mathbf{T} \mathbf{D}_n, \tag{9b}$$

and the matrices $\mathbf{D}_n$ are computed using the current value of the parameters $\tilde{\boldsymbol{\theta}}$. The formulas (9) can be derived from the assumed distribution of $\mathbf{s}_n$ in (3) and from the linear observation model $\mathbf{x}_n^{\text{obs}} = (\mathbf{M}_n \mathbf{T}) \mathbf{s}_n$ (see e.g. [28, Theorem 10.3]). Note that these formulas can be seen as Wiener filtering given estimated posterior covariance [28, Chapter 12] to recover the missing samples.

Now, from (3) and from the independence assumption, it holds that

$$p(\mathbf{S} \mid \boldsymbol{\theta}) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{s}_n \mid \mathbf{0}, \mathbf{D}_n), \quad \mathbf{D}_n = \operatorname{diag}\left([v_{fn}]_{f=1,\dots,F} \right). \tag{10}$$

The M-step, i.e., the minimization of (7) with respect to the parameters $\boldsymbol{\theta} = \{\mathbf{W}, \mathbf{H}\}$ is equivalent to the minimization of the Itakura–Saito divergence $D_{\text{IS}}(\mathbf{P} \mid \mathbf{WH})$ [22, 26], where the divergence is defined for matrices $\mathbf{A} = [a_{ij}], \mathbf{B} = [b_{ij}]$ as

$$D_{\text{IS}}(\mathbf{A} \mid \mathbf{B}) = \sum_{i,j} d_{\text{IS}}(a_{ij} \mid b_{ij}) = \sum_{i,j} \left(\frac{a_{ij}}{b_{ij}} - \log \frac{a_{ij}}{b_{ij}} - 1 \right), \tag{11}$$

and the matrix $\mathbf{P} = [p_{fn}]$ is the posterior power spectrum given by:

$$p_{fn} = \mathbb{E}\left(|s_{fn}|^2 \mid \mathbf{x}_n^{\text{obs}}, \mathbf{W}, \mathbf{H} \right) = |(\hat{\mathbf{s}}_n)_f|^2 + (\hat{\boldsymbol{\Sigma}}_n)_{ff}. \tag{12}$$

The M-step can thus be performed by applying (and iterating) the following multiplicative rules [26, Alg. 1]:

$$\mathbf{W} \leftarrow \mathbf{W} \odot \frac{\left((\mathbf{WH})^{\odot[-2]} \odot \mathbf{P}\right) \mathbf{H}^{\mathsf{T}}}{(\mathbf{WH})^{\odot[-1]} \mathbf{H}^{\mathsf{T}}}, \tag{13a}$$

$$\mathbf{H} \leftarrow \mathbf{H} \odot \frac{\mathbf{W}^{\mathsf{T}} \left((\mathbf{WH})^{\odot[-2]} \odot \mathbf{P}\right)}{\mathbf{W}^{\mathsf{T}} (\mathbf{WH})^{\odot[-1]}}, \tag{13b}$$

where $\frac{\mathbf{A}}{\mathbf{B}}$ denotes the matrix $\mathbf{A} \odot \mathbf{B}^{\odot[-1]}$ and the symbol $\odot$ is used to denote entry-wise multiplication or power. Note that in the algorithm, the $\mathbf{H}$ update (13b) uses the already updated value of $\mathbf{W}$ from (13a). In practice, the two updates are followed by a normalization step: The columns of $\mathbf{W}$ are scaled to have unit norm and the rows of $\mathbf{H}$ are inversely scaled by the same factor, such that the product $\mathbf{WH}$ does not change.

The whole EM-*tf* algorithm is summarized in Alg. 1. Its output is the estimate of the complete data, i.e., the full TF spectrum $\hat{\mathbf{S}} \in \mathbb{C}^{F \times N}$, together with the estimate of the parameters $\mathbf{W}$ and $\mathbf{H}$. Then, the framed time-domain signal $\hat{\mathbf{X}}$ is synthesized from $\hat{\mathbf{S}}$ using the operator $\mathbf{T}$. Finally, the whole signal estimate $\hat{\mathbf{y}}$ is obtained from $\hat{\mathbf{X}}$ using the overlap-add procedure.

3.2. EM-t

As an alternative to EM-*tf*, we build an EM algorithm that uses a complete dataset in the time domain (hence the abbreviation EM-*t*). As a result, we directly get the posterior distribution of the missing data as a side-product of the algorithm.

This novel algorithm addresses the original problem (5) by minimizing the functional:

$$Q(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) = -\int \log p(\mathbf{X} \mid \boldsymbol{\theta}) p(\mathbf{X}^{\text{miss}} \mid \mathbf{X}^{\text{obs}}, \tilde{\boldsymbol{\theta}}) \, \mathrm{d}\mathbf{X}^{\text{miss}}, \tag{14}$$

where $\mathbf{X}^{\text{miss}} = \{\mathbf{x}_1^{\text{miss}}, \ldots, \mathbf{x}_N^{\text{miss}}\}$ represents the missing samples. These are identified as $\mathbf{x}_n^{\text{miss}} = \overline{\mathbf{M}}_n \mathbf{x}_n$, where $\overline{\mathbf{M}}_n$ is the complementary selection matrix to $\mathbf{M}_n$, i.e., $\overline{\mathbf{M}}_n$ selects the unreliable (missing) samples of frame n.

It is clear that in some cases, this approach will be equivalent to EM-*tf*, such as in the case of one-to-one correspondence between the temporal frame samples and their frequency coefficients. However, this part discusses a general situation with no other assumptions on the transforms involved, and particular cases will be examined later in Section 3.3.

To derive the steps of the EM algorithm, observe that the relation $\mathbf{x}_n = \mathbf{T}\mathbf{s}_n$, together with Assumption 1 (Gaussian coefficients), directly leads to

$$p(\mathbf{X} \mid \boldsymbol{\theta}) = \prod_{n=1}^{N} \mathcal{N}\left(\mathbf{x}_n \mid \mathbf{0}, \mathbf{T}\mathbf{D}_n\mathbf{T}^{\mathsf{H}}\right). \tag{15}$$

In a similar manner to the E-step of EM-*tf*, we can derive the following

$$p(\mathbf{X}^{\text{miss}} \mid \mathbf{X}^{\text{obs}}, \tilde{\boldsymbol{\theta}}) = \prod_{n=1}^{N} p(\mathbf{x}_n^{\text{miss}} \mid \mathbf{x}_n^{\text{obs}}, \tilde{\boldsymbol{\theta}}) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{x}_n^{\text{miss}} \mid \overline{\mathbf{M}}_n \mathbf{T}\hat{\mathbf{s}}_n, \, \overline{\mathbf{M}}_n \mathbf{T}\hat{\boldsymbol{\Sigma}}_n \mathbf{T}^{\mathsf{H}} \overline{\mathbf{M}}_n^{\mathsf{T}}), \tag{16}$$

with $\hat{\mathbf{s}}_n, \hat{\boldsymbol{\Sigma}}_n$ defined in Eq. (9).

Algorithm 1: Audio inpainting via EM-*tf*.

Input: reliable samples $\{\mathbf{x}_n^{\text{obs}}\}_{n=1,\dots,N}$, respective selection matrices $\{\mathbf{M}_n\}_{n=1,\dots,N}$, linear transform $\mathbf{T} \in \mathbb{C}^{M \times F}$

initialize $\mathbf{W} \in \mathbb{R}^{F \times K}$, $\mathbf{H} \in \mathbb{R}^{K \times N}$ non-negative

repeat

`// E-step:`

for $n = 1, \dots, N$ **do**

$\mathbf{D}_n \leftarrow \text{diag}\left([v_{fn}]_{f=1,\dots,F}\right)$ with $[v_{fn}]_{f=1,\dots,F}$ being the n-th column of the matrix $\mathbf{V} = \mathbf{WH}$

$\hat{\mathbf{s}}_n \leftarrow \mathbf{D}_n \mathbf{T}^{\mathsf{H}} \mathbf{M}_n^{\mathsf{T}} \left(\mathbf{M}_n \mathbf{T} \mathbf{D}_n \mathbf{T}^{\mathsf{H}} \mathbf{M}_n^{\mathsf{T}}\right)^{-1} \mathbf{x}_n^{\text{obs}}$

$\hat{\mathbf{\Sigma}}_n \leftarrow \mathbf{D}_n - \mathbf{D}_n \mathbf{T}^{\mathsf{H}} \mathbf{M}_n^{\mathsf{T}} \left(\mathbf{M}_n \mathbf{T} \mathbf{D}_n \mathbf{T}^{\mathsf{H}} \mathbf{M}_n^{\mathsf{T}}\right)^{-1} \mathbf{M}_n \mathbf{T} \mathbf{D}_n$

$p_{fn} \leftarrow |(\hat{\mathbf{s}}_n)_f|^2 + (\hat{\mathbf{\Sigma}}_n)_{ff}, \ f = 1, \dots, F$

end

`// M-step:`

repeat

$\mathbf{W} \leftarrow \mathbf{W} \odot \dfrac{\left((\mathbf{WH})^{\odot[-2]} \odot \mathbf{P}\right)\mathbf{H}^{\mathsf{T}}}{(\mathbf{WH})^{\odot[-1]}\mathbf{H}^{\mathsf{T}}}$ with $\mathbf{P} = [p_{fn}]$

$\mathbf{H} \leftarrow \mathbf{H} \odot \dfrac{\mathbf{W}^{\mathsf{T}}\left((\mathbf{WH})^{\odot[-2]} \odot \mathbf{P}\right)}{\mathbf{W}^{\mathsf{T}}(\mathbf{WH})^{\odot[-1]}}$ with $\mathbf{P} = [p_{fn}]$

normalize columns of $\mathbf{W}$, scale rows of $\mathbf{H}$

until *satisfied with the factorization*

until *convergence criteria met*

Output: $\hat{\mathbf{S}} = [\hat{\mathbf{s}}_1, \dots, \hat{\mathbf{s}}_N], \hat{\mathbf{W}} = \mathbf{W}, \hat{\mathbf{H}} = \mathbf{H}$

The crucial part of the algorithm is the ML estimation of the parameters $\mathbf{W}, \mathbf{H}$ in the M-step, given the posterior distribution of the missing temporal samples. Even though the closed form of $Q(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}})$ is available due to the expressions in Eq. (15) and (16), it is expensive to compute and optimize directly. Thus, we proceed to re-estimate the spectrum corresponding to the signal estimated by Eq. (16) and update $\mathbf{W}$ and $\mathbf{H}$ as the factorization of this spectrum.

To do this, we introduce an analysis operator, represented by the matrix $\mathbf{U} \in \mathbb{C}^{F \times M}$, associated to the synthesis operator $\mathbf{T}$. So far, the only assumption about the analysis operator is linearity – more details are given below. Using $\mathbf{U}$, it is straightforward to derive the posterior distribution of the TF coefficients $\mathbf{S}^{\text{alt}} = \mathbf{UX}$ associated to the posterior time-domain samples:

$$p(\mathbf{S}^{\text{alt}} \mid \mathbf{X}^{\text{obs}}, \tilde{\boldsymbol{\theta}}) = \prod_{n=1}^{N} p(\mathbf{s}_n^{\text{alt}} \mid \mathbf{x}_n^{\text{obs}}, \tilde{\boldsymbol{\theta}}) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{s}_n^{\text{alt}} \mid \hat{\mathbf{s}}_n^{\text{alt}}, \hat{\mathbf{\Sigma}}_n^{\text{alt}}) \tag{17}$$

Algorithm 2: Audio inpainting via EM-*t*.

Input: reliable samples $\{\mathbf{x}_n^{\text{obs}}\}_{n=1,\dots,N}$, respective selection matrices $\{\mathbf{M}_n\}_{n=1,\dots,N}$, linear transforms $\mathbf{T} \in \mathbb{C}^{M\times F}, \mathbf{U} \in \mathbb{C}^{F\times M}$

initialize $\mathbf{W} \in \mathbb{R}^{F\times K}$, $\mathbf{H} \in \mathbb{R}^{K\times N}$ non-negative

repeat

`// E-step:`

for $n = 1, \dots, N$ **do**

$\mathbf{D}_n \leftarrow \operatorname{diag}\left([v_{fn}]_{f=1,\dots,F}\right)$ with $[v_{fn}]_{f=1,\dots,F}$ being the n-th column of the matrix $\mathbf{V} = \mathbf{WH}$

$\hat{\mathbf{s}}_n^{\text{alt}} \leftarrow \mathbf{UTD}_n\mathbf{T}^{\mathsf{H}}\mathbf{M}_n^{\mathsf{T}}\left(\mathbf{M}_n\mathbf{TD}_n\mathbf{T}^{\mathsf{H}}\mathbf{M}_n^{\mathsf{T}}\right)^{-1}\mathbf{x}_n^{\text{obs}}$

$\hat{\mathbf{\Sigma}}_n^{\text{alt}} \leftarrow \mathbf{UT}\left(\mathbf{D}_n - \mathbf{D}_n\mathbf{T}^{\mathsf{H}}\mathbf{M}_n^{\mathsf{T}}\left(\mathbf{M}_n\mathbf{TD}_n\mathbf{T}^{\mathsf{H}}\mathbf{M}_n^{\mathsf{T}}\right)^{-1}\mathbf{M}_n\mathbf{TD}_n\right)\mathbf{T}^{\mathsf{H}}\mathbf{U}^{\mathsf{H}}$

$p_{fn} \leftarrow \left|(\hat{\mathbf{s}}_n^{\text{alt}})_f\right|^2 + (\hat{\mathbf{\Sigma}}_n^{\text{alt}})_{ff},\ f = 1, \dots, F$

end

`// M-step:`

repeat

$\mathbf{W} \leftarrow \mathbf{W} \odot \dfrac{\left((\mathbf{WH})^{\odot[-2]} \odot \mathbf{P}\right)\mathbf{H}^{\mathsf{T}}}{(\mathbf{WH})^{\odot[-1]}\mathbf{H}^{\mathsf{T}}}$ with $\mathbf{P} = [p_{fn}]$

$\mathbf{H} \leftarrow \mathbf{H} \odot \dfrac{\mathbf{W}^{\mathsf{T}}\left((\mathbf{WH})^{\odot[-2]} \odot \mathbf{P}\right)}{\mathbf{W}^{\mathsf{T}}(\mathbf{WH})^{\odot[-1]}}$ with $\mathbf{P} = [p_{fn}]$

normalize columns of $\mathbf{W}$, scale rows of $\mathbf{H}$

until *satisfied with the factorization*

until *convergence criteria met*

Output: $\hat{\mathbf{S}}^{\text{alt}} = \left[\hat{\mathbf{s}}_1^{\text{alt}}, \dots, \hat{\mathbf{s}}_N^{\text{alt}}\right], \hat{\mathbf{W}} = \mathbf{W}, \hat{\mathbf{H}} = \mathbf{H}$

with

$$\hat{\mathbf{s}}_n^{\text{alt}} = \mathbf{UT}\hat{\mathbf{s}}_n, \quad \hat{\mathbf{\Sigma}}_n^{\text{alt}} = \mathbf{UT}\hat{\mathbf{\Sigma}}_n\mathbf{T}^{\mathsf{H}}\mathbf{U}^{\mathsf{H}}. \tag{18}$$

Finally, the M-step is equivalent to the M-step defined by the updates in Eq. (13), with the alternative posterior power spectrum (computed from $\hat{\mathbf{s}}_n^{\text{alt}}$ and $\hat{\mathbf{\Sigma}}_n^{\text{alt}}$). The whole algorithm is summarized in Alg. 2.

3.3. On the relations between EM-tf and EM-t

As previously mentioned, there are some natural choices of the pair $\{\mathbf{T}, \mathbf{U}\}$ which result in the equivalence of the algorithms EM-*tf* and EM-*t*. Several alternatives are discussed in what follows.

1. $\mathbf{T}$ is invertible, $\mathbf{U} = \mathbf{T}^{-1}$.
   It follows from (18) that $\hat{\mathbf{s}}_n = \hat{\mathbf{s}}_n^{\text{alt}}$ and $\hat{\mathbf{\Sigma}}_n = \hat{\mathbf{\Sigma}}_n^{\text{alt}}$, therefore EM-*tf* and EM-*t* are identical algorithms. One noticeable special case is when $\mathbf{T}$ is unitary, i.e., $\mathbf{U} = \mathbf{T}^{-1} = \mathbf{T}^{\mathsf{H}}$. A popular example is the case of a properly scaled DFT realized formally by multiplication with the unitary matrix $\mathbf{U} \in \mathbb{C}^{M\times M}$.

2. $\mathbf{T}$ is the synthesis operator of a tight frame [29, Ch. 1], $F > M$, $\mathbf{U} = \mathbf{T}^{\mathsf{H}}$, $\mathbf{TU} = \mathbf{I}$, where $\mathbf{I}$ denotes the identity matrix of appropriate size.
   The two algorithms are no longer equivalent, since $\mathbf{UT}$ represents the projection operator on the range space of $\mathbf{U}$, which is in general different from identity. For instance, this is the case of a redundant DFT, e.g., with $F = 2M$ (twice more frequency bins than the frame length).[2]
3. $\mathbf{T}$ is the analysis operator of a tight frame, $F < M$, $\mathbf{U} = \mathbf{T}^{\mathsf{H}}$, i.e., $\mathbf{UT} = \mathbf{I}$.
   In this case, we do not have enough frequency coefficients to reconstruct *any* signal in the framed time domain. This means that the time-domain solution (in each frame) is restricted to the range space of $\mathbf{T}$. However, the estimation EM-*tf* and EM-*t* are once again equivalent in this case.
4. $\mathbf{T}$ is arbitrary, $\mathbf{U} = \mathbf{T}^{+}$.
   In this case, the matrix $\mathbf{UT} = \mathbf{T}^{+}\mathbf{T}$ used in (18) represents the orthogonal projection onto the range space of $\mathbf{T}^{\mathsf{H}}$ (which equals the orthogonal complement of the kernel of $\mathbf{T}$). This is in general different from identity, unless the range space of $\mathbf{T}^{\mathsf{H}}$ is the whole coefficient space $\mathbb{C}^F$.
5. $\mathbf{U}$ is arbitrary, $\mathbf{T} = \mathbf{U}^{+}$.
   Similarly to the previous option, $\mathbf{UT} = \mathbf{UU}^{+}$ represents the orthogonal projection onto the range space of $\mathbf{U}$ (which equals the orthogonal complement of the kernel of $\mathbf{U}^{\mathsf{H}}$). This is in general different from identity, unless the range space of $\mathbf{U}$ is the whole coefficient space $\mathbb{C}^F$.

Even though the list is far from being exhaustive, it illustrates that there are commonly used settings (e.g., the redundant DFT used in some sparsity-based reconstruction algorithms [30, 7, 8]) where EM-*tf* and EM-*t* are not just conceptually but also practically different. This will be further detailed in the experiments in Section 5.

4. ML estimation by treating the missing samples as parameters: the AM algorithm

As a novel approach, we propose to treat the missing samples as parameters and include them explicitly into the estimation problem, which results in:

$$\hat{\mathbf{W}}, \hat{\mathbf{H}}, \hat{\mathbf{X}}^{\text{miss}} = \underset{\mathbf{W},\mathbf{H},\mathbf{X}^{\text{miss}}}{\arg\min} \; -\log p\big(\mathbf{X}^{\text{obs}}, \mathbf{X}^{\text{miss}} \mid \mathbf{W}, \mathbf{H}\big). \tag{19}$$

Note that the objective of this novel estimator is a function different from the likelihood in (5), since the parameter space is extended by the inclusion of the missing samples in the problem. We propose to approach it via AM. This approach consists of two steps – minimization of (19) with respect to the missing samples (signal update) and minimization with respect to the NMF parameters (model update).

4.1. Signal update

Performing the signal update means minimizing the objective in (19) with respect to $\mathbf{x}_n^{\text{miss}}$ while the current estimates of $\mathbf{W}$, $\mathbf{H}$ are fixed. This is equivalent to finding the mode of the conditional distribution of $\mathbf{x}_n^{\text{miss}}$ given $\mathbf{x}_n^{\text{obs}}, \mathbf{W}, \mathbf{H}$, which, due to the Gaussian assumption, equals its expectation:

$$\hat{\mathbf{x}}_n^{\text{miss}} = \mathbb{E}\left(\mathbf{x}_n^{\text{miss}} \mid \mathbf{x}_n^{\text{obs}}, \mathbf{W}, \mathbf{H}\right). \tag{20}$$

[2] In practice, this can be implemented by zero-padding the signal to twice its length and then computing the DFT. The backward transform is the inverse DFT, followed by cropping the result to the original length.

Using Eq. (16) and Eq. (9a), we can compute this expectation, which yields:

$$\hat{\mathbf{x}}_n^{\text{miss}} = \overline{\mathbf{M}}_n \mathbf{T}\mathbf{D}_n\mathbf{T}^{\mathsf{H}}\mathbf{M}_n^{\mathsf{T}} \left(\mathbf{M}_n\mathbf{T}\mathbf{D}_n\mathbf{T}^{\mathsf{H}}\mathbf{M}_n^{\mathsf{T}}\right)^{-1} \mathbf{x}_n^{\text{obs}}. \tag{21}$$

The whole signal frame, including both the estimated missing samples $\hat{\mathbf{x}}_n^{\text{miss}}$ and the observed samples $\mathbf{x}_n^{\text{obs}}$, can be merged together as:

$$\hat{\mathbf{x}}_n = \mathbf{T}\mathbf{D}_n\mathbf{T}^{\mathsf{H}}\mathbf{M}_n^{\mathsf{T}} \left(\mathbf{M}_n\mathbf{T}\mathbf{D}_n\mathbf{T}^{\mathsf{H}}\mathbf{M}_n^{\mathsf{T}}\right)^{-1} \mathbf{x}_n^{\text{obs}}. \tag{22}$$

4.2. NMF parameters update

For the NMF model update, we aim at deriving the computation directly from the optimization problem (19). Since not only the observed samples but also the (estimated) missing ones are fixed in this step, this is equivalent to minimizing

$$-\log p\left(\mathbf{x}_n \mid \mathbf{W}, \mathbf{H}\right) = \log\det\left(\pi\mathbf{T}\mathbf{D}_n\mathbf{T}^{\mathsf{H}}\right) + \mathbf{x}_n^{\mathsf{T}}\left(\mathbf{T}\mathbf{D}_n\mathbf{T}^{\mathsf{H}}\right)^{-1}\mathbf{x}_n \tag{23}$$

with respect to $\mathbf{W}, \mathbf{H}$. To simplify the development of the method, let us pose the following assumption.

Assumption 3 (invertibility of the synthesis)**.** *The synthesis operator* $\mathbf{T}$ *is invertible and the analysis operator is* $\mathbf{U} = \mathbf{T}^{-1}$*. In particular, this means that* $\mathbf{T}$ *is square, i.e.,* $F = M$*.*

Under assumption 3, we see that $\det\left(\pi\mathbf{T}\mathbf{D}_n\mathbf{T}^{\mathsf{H}}\right) = \pi^M \det\left(\mathbf{T}\right)^2 \det\left(\mathbf{D}_n\right)$ and $(\mathbf{T}\mathbf{D}_n\mathbf{T}^{\mathsf{H}})^{-1} = (\mathbf{T}^{-1})^{\mathsf{H}}\mathbf{D}_n^{-1}\mathbf{T}^{-1}$. The optimization problem then reduces to:

$$\underset{\mathbf{W},\mathbf{H}}{\arg\min}\ \log\det\left(\mathbf{D}_n\right) + \left(\mathbf{T}^{-1}\mathbf{x}_n\right)^{\mathsf{H}}\mathbf{D}_n^{-1}\left(\mathbf{T}^{-1}\mathbf{x}_n\right). \tag{24}$$

Now recall that $\mathbf{D}_n = \operatorname{diag}\left([v_{fn}]_{f=1,\dots,F}\right)$, thus we can rewrite the objective function as:

$$\log\prod_{f=1}^{F} v_{fn} + \sum_{f=1}^{F}\left(\mathbf{T}^{-1}\mathbf{x}_n\right)_f^* \frac{1}{v_{fn}}\left(\mathbf{T}^{-1}\mathbf{x}_n\right)_f = \sum_{f=1}^{F}\log v_{fn} + \sum_{f=1}^{F}\frac{\left|\left(\mathbf{T}^{-1}\mathbf{x}_n\right)_f\right|^2}{v_{fn}}. \tag{25}$$

It is now straightforward to show that the minimization of (25) is equivalent to the minimization of the Itakura–Saito divergence:

$$\underset{v_{fn}}{\arg\min}\sum_{f=1}^{F} d_{\text{IS}}\left(\left|\left(\mathbf{T}^{-1}\mathbf{x}_n\right)_f\right|^2 \mid v_{fn}\right), \quad v_{fn} = \sum_k w_{fk}h_{kn}. \tag{26}$$

Taking into account all the frames finally leads to the desired result that $\mathbf{W}, \mathbf{H}$ are obtained by minimizing $D_{\text{IS}}(\mathbf{P} \mid \mathbf{W}\mathbf{H})$ where $p_{fn} = \left|(\mathbf{T}^{-1}\hat{\mathbf{x}}_n)_f\right|^2$ and $\hat{\mathbf{x}}_n$ is the signal estimate from Eq. (22). The procedure is summarized in Alg. 3.

Remark 2. *A simple heuristic possibility for the case of non-invertible* $\mathbf{T}$ *is to compute the spectrum of* $\hat{\mathbf{x}}_n$*, defined in Eq. (22), as*

$$\hat{\mathbf{s}}_n = \mathbf{U}\hat{\mathbf{x}}_n = \mathbf{U}\mathbf{T}\mathbf{D}_n\mathbf{T}^{\mathsf{H}}\mathbf{M}_n^{\mathsf{T}}\left(\mathbf{M}_n\mathbf{T}\mathbf{D}_n\mathbf{T}^{\mathsf{H}}\mathbf{M}_n^{\mathsf{T}}\right)^{-1}\mathbf{x}_n^{\text{obs}} \tag{27}$$

Algorithm 3: Audio inpainting via AM.

Input: reliable samples $\{\mathbf{x}_n^{\text{obs}}\}_{n=1,\dots,N}$, respective selection matrices $\{\mathbf{M}_n\}_{n=1,\dots,N}$, invertible linear transform $\mathbf{T} \in \mathbb{C}^{M\times F}$

initialize $\mathbf{W} \in \mathbb{R}^{F\times K}$, $\mathbf{H} \in \mathbb{R}^{K\times N}$ non-negative

repeat

`// Signal update:`

for $n = 1, \dots, N$ **do**

$\mathbf{D}_n \leftarrow \operatorname{diag}\left([v_{fn}]_{f=1,\dots,F}\right)$ with $[v_{fn}]_{f=1,\dots,F}$ being the n-th column of the matrix $\mathbf{V} = \mathbf{WH}$

$\hat{\mathbf{s}}_n \leftarrow \mathbf{D}_n\mathbf{T}^{\mathsf{H}}\mathbf{M}_n^{\mathsf{T}}\left(\mathbf{M}_n\mathbf{T}\mathbf{D}_n\mathbf{T}^{\mathsf{H}}\mathbf{M}_n^{\mathsf{T}}\right)^{-1}\mathbf{x}_n^{\text{obs}}$

$\hat{\mathbf{x}}_n \leftarrow \mathbf{T}^{-1}\hat{\mathbf{s}}_n$

$p_{fn} \leftarrow |(\hat{\mathbf{s}}_n)_f|^2\,,\ f = 1, \dots, F$

end

`// Model update:`

repeat

$\mathbf{W} \leftarrow \mathbf{W} \odot \dfrac{\left((\mathbf{WH})^{\odot[-2]} \odot \mathbf{P}\right)\mathbf{H}^{\mathsf{T}}}{(\mathbf{WH})^{\odot[-1]}\mathbf{H}^{\mathsf{T}}}$ with $\mathbf{P} = [p_{fn}]$

$\mathbf{H} \leftarrow \mathbf{H} \odot \dfrac{\mathbf{W}^{\mathsf{T}}\left((\mathbf{WH})^{\odot[-2]} \odot \mathbf{P}\right)}{\mathbf{W}^{\mathsf{T}}(\mathbf{WH})^{\odot[-1]}}$ with $\mathbf{P} = [p_{fn}]$

normalize columns of $\mathbf{W}$, scale rows of $\mathbf{H}$

until *satisfied with the factorization*

until *convergence criteria met*

Output: $\hat{\mathbf{X}} = [\hat{\mathbf{x}}_1, \dots, \hat{\mathbf{x}}_N]\,, \hat{\mathbf{W}} = \mathbf{W}, \hat{\mathbf{H}} = \mathbf{H}$

and the power spectrogram $p_{fn} = |(\hat{\mathbf{s}}_n)_f|^2$. *Then, we apply the multiplicative rules to minimize* $D_{\text{IS}}(\mathbf{P} \mid \mathbf{WH})$. *However, this approach is not justified by the minimization of* (25) *with respect to* $\mathbf{W}, \mathbf{H}$. *The problem is that if we cannot compute the inversion* $(\mathbf{T}\mathbf{D}_n\mathbf{T}^{\mathsf{H}})^{-1}$ *as* $\mathbf{A}\mathbf{D}_n^{-1}\mathbf{B}$ *for some matrices* $\mathbf{A}, \mathbf{B}$, *we cannot separate the individual diagonal entries of* $\mathbf{D}_n$ *to fit it to the IS-NMF problem.*

Remark 3. *Note that in the setting imposed in Assumption 3 (invertibility of the synthesis), EM-tf is equivalent to EM-t, but the alternating minimization produces a different algorithm, because we do not include any covariance matrix in the power spectra. Thus, AM might provide a different inpainting solution, as demonstrated by the numerical experiments in Section 5.*

Remark 4 (computational complexity)**.** *It is intricate to express the computational complexity of the algorithms EM-tf, EM-t and AM. However, several claims are evident:*

1. *In EM-tf on line 6 of Alg. 1, we do not in fact need to compute the whole matrix* $\hat{\mathbf{\Sigma}}_n$, *but solely its diagonal entries. Thus, if the complicated term* $\mathbf{D}_n\mathbf{T}^{\mathsf{H}}\mathbf{M}_n^{\mathsf{T}}\left(\mathbf{M}_n\mathbf{T}\mathbf{D}_n\mathbf{T}^{\mathsf{H}}\mathbf{M}_n^{\mathsf{T}}\right)^{-1}$ *is*

computed and saved on line 5 (e.g., using Matlab's mrdivide without the need to perform the matrix inversion), then no more matrix multiplications are needed on line 6 since the diagonal entries can be extracted by computing only F scalar products.

2. *In AM, the matrix* $\mathbf{D}_n\mathbf{T}^{\mathsf{H}}\mathbf{M}_n^{\mathsf{T}}\left(\mathbf{M}_n\mathbf{T}\mathbf{D}_n\mathbf{T}^{\mathsf{H}}\mathbf{M}_n^{\mathsf{T}}\right)^{-1}$ *does not need to be calculated at all – we can first compute* $\left(\mathbf{M}_n\mathbf{T}\mathbf{D}_n\mathbf{T}^{\mathsf{H}}\mathbf{M}_n^{\mathsf{T}}\right)^{-1}\mathbf{x}_n^{\text{obs}}$ *efficiently and then multiply the resulting vector with the matrix* $\mathbf{D}_n\mathbf{T}^{\mathsf{H}}\mathbf{M}_n^{\mathsf{T}}$*. The difference to EM-tf is that the inversion operates with a vector, not a matrix, which makes AM less computationally demanding per iteration than EM-tf.*
3. *Although a similar strategy as in EM-tf can be applied to EM-t, computing even the diagonal entries of* $\hat{\mathbf{\Sigma}}_n^{\text{alt}}$ *as derived in Eq. (18) is more complicated due to the multiplication with* $\mathbf{UT}$*. As a result, EM-t has computationally the most expensive iteration from the three algorithms, even in settings when the operations can be implemented efficiently using FFT.*

5. Experiments

In this section, we evaluate the performance of the proposed estimators for the task of inpainting noise-less musical recordings. The implementation is done in Matlab with the use of the LTFAT toolbox [31]. For the sake of reproducibility, the source code is published online.[3]

5.1. Protocol

The experiments focus on audio inpainting of compact gaps, which is a challenging setup. However, we start with a preliminary experiment dedicated to the restoration of missing samples at random location, which is mainly motivated by the high computational cost of the algorithms (see Section 5.2.1).

For the gap-filling experiment, we use the set of 10 musical recordings from the EBU SQAM dataset [32, 33], sampled at 44.1 kHz and shortened to 7 seconds, as used commonly in recent related publications [6, 8]. The proposed estimators are first evaluated against each other (Section 5.2.2), and then compared to state-of-the-art baselines (Section 5.3).

As a measure of audio quality, we consider the commonly used signal-to-noise ratio (SNR) defined as:

$$\text{SNR}(\mathbf{y},\hat{\mathbf{y}}) = 10\log_{10}\frac{\|\mathbf{y}\|^2}{\|\mathbf{y}-\hat{\mathbf{y}}\|^2}, \tag{28}$$

and expressed in dB, as well as the perceptually motivated objective measures PEMO-Q [34] and PEAQ [35, 36], since no other standard for perceptual similarity is established for the evaluation of audio inpainting. All the metrics measure the similarity of the ground truth signal $\mathbf{y}$ and its estimate $\hat{\mathbf{y}}$; SNR represents the sample-wise similarity of the waveforms, whereas PEMO-Q and PEAQ estimate the perceived difference (objective difference grade, ODG) on a scale ranging from -4 (very annoying) to 0 (imperceptible). Note that since we consider a noise-less scenario and all the methods considered fit the reliable samples perfectly, we measure the SNR only on the inpainted segments.

The relative solution change, used as a measure of convergence, is computed as $\|\hat{\mathbf{y}}^{(i+1)}-\hat{\mathbf{y}}^{(i)}\|/\|\hat{\mathbf{y}}^{(i)}\|$, where $\hat{\mathbf{y}}^{(i)}$ is the estimate at the i-th iteration folded together from

[3] https://github.com/ondrejmokry/InpaintingNMF

$\mathbf{T}\hat{\mathbf{S}} = \mathbf{T}[\hat{\mathbf{s}}_1, \ldots, \hat{\mathbf{s}}_N] = [\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_N]$, where $\hat{\mathbf{S}}$ is the output of the proposed algorithms. Similarly, we compute the relative objective change, where the objective is the negative log-likelihood of Eq. (5) for EM-*tf* and EM-*t* and of Eq. (19) for AM, evaluated using the current iterates of $\mathbf{W}$, $\mathbf{H}$ and the corresponding solution.

5.2. Comparison of the proposed estimators

We start the experiments with the comparison of the behavior of the three estimators, EM-*tf*, EM-*t* and AM. In 5.2.1, we perform a preliminary comparison including the very demanding calculation of the objective function. Then, we validate the results in 5.2.2 for the gap-filling task.

5.2.1. Preliminary comparison with random missing samples

First, we illustrate the capabilities of the three estimators EM-*tf*, EM-*t* and AM on a demonstrative example of an inpainting problem. The original signal is an excerpt of the first 12 seconds of the song *Mamavatu* by Susheela Raman, containing acoustic guitar and drums, sampled at 16 kHz. We discard 60 % of the signal samples (chosen randomly) and perform inpainting using EM-*tf*, EM-*t* and AM. The temporal frames are extracted using sine window of length $M = 1024$ samples (64 ms), the hop length is 512 samples and we use $K = 10$ components of the NMF. To distinguish between the individual algorithms, the number of frequency channels F in this experiment varies between M and $2M$, which corresponds to examples 1. and 2. of subsection 3.3, respectively.

The comparison is visualized in Fig. 1 by means of several quantities: the negative log-likelihood (i.e., the objective function of the estimation problem (5) or (19)), SNR, relative objective change and relative solution change. Our key finding is the difference in the observed convergence speed with respect to iteration count. It is visible in all the plots of Fig. 1 that AM approaches its solution faster than EM-*tf*, which is even more pronounced with respect to CPU time according to remark 4 (computational complexity). However, the quality of the solution may decrease after the peak is reached. A similar behavior is observed for the case of redundant transform $\mathbf{T}$. This redundancy causes that:

1. the convergence of EM-*tf* for $F = 2M$ is slower than with $F = M$ while reaching similar reconstruction quality,
2. the convergence of EM-*t* for $F = 2M$ is faster than both cases of EM-*tf*, but as in the case of AM, it reaches worse reconstruction quality.

Based on these first observations, a natural question arises: Can we combine the convergence properties of AM with the performance of EM-*tf*? To answer it, we consider a combined algorithm AM-to-EM-*tf*, which consists in initializing EM-*tf* with 5 iterations of AM. As observed in Fig. 1, the initialization with AM does improve the algorithmic behavior in the first few iterations and at the same time, the resulting restoration quality is the same as with EM-*tf*. However, the number of iterations needed to reach the peak performance is not reduced by the switching strategy.

5.2.2. Missing gaps

To validate the preliminary results concerning EM-*tf* and AM, we perform a larger experiment with a set of signals and for the more demanding problem of inpainting short to middle-length gaps (instead of random subsampling). In each of the 10 signals from the EBU SQAM dataset, 10 gaps of given length are artificially introduced, with the gap length ranging from 20 to 80 ms. We do not track the objective function in this case, because its computation is computationally demanding. We also focus on the practical case of $F = M$ and invertible transform $\mathbf{T}$ representing the DFT

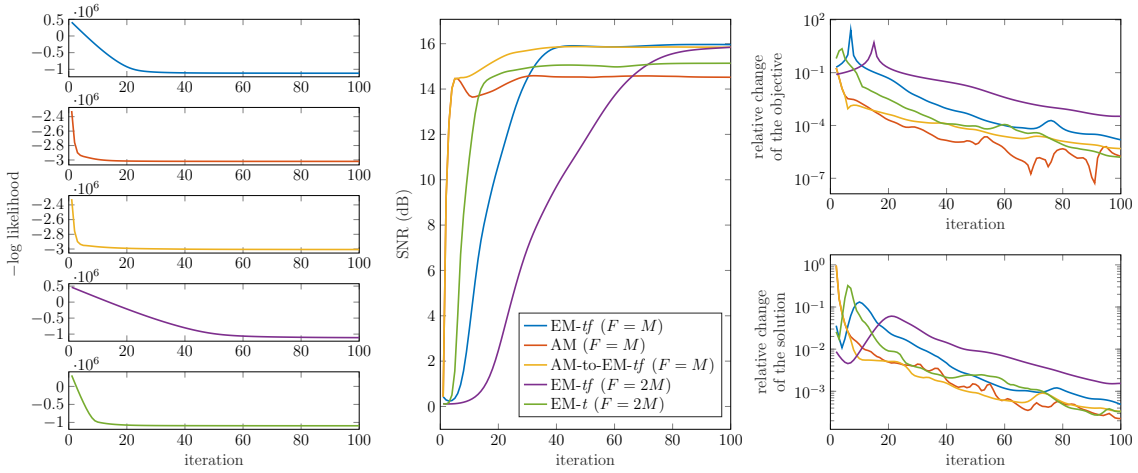

Figure 1: Comparison of the performance and convergence properties of EM-*tf*, EM-*t* and AM, including the switching variant AM-to-EM-*tf*. The legend in the middle plot is common for the whole figure. Note that the first column thus shows three different quantities, since the objective depends on the choice of F and also on the algorithm. Especially, the formula for log likelihood switches after the initializing iterations of AM-to-EM-*tf*, which is however disregard on purpose in the plot.

in each temporal frame, thus EM-t is omitted. The frame length is $M = 4096$ samples (approx. 92 ms), and the temporal frames are extracted using sine window with 50% overlap.

As in the previous experiment, we observe that the performance difference between AM and EM-*tf* is not significant, but AM reaches its peak faster, as shown in the plots for SNR in Fig. 2. The relative solution change supports this observation, as it has been demonstrated above (see Fig. 1) that this measure mostly corresponds to the convergence of the algorithm with respect to its objective value. A new observation is that this phenomenon depends on the gap length – the longer the gap, the slower the convergence of EM-*tf* is, compared to AM.

5.3. Comparison with the state of the art

Finally, we compare our NMF-based methods to the following state-of-the-art techniques for the task of inpainting middle-length audio gaps:

- Janssen's AR approach [4] (denoted **Janssen**): This iterative algorithm builds upon a frame-wise AR nature of the clean signal. At each iteration, it estimates the AR coefficients of the signal estimate (starting from the observed signal with the missing samples initialized as zeros), and then recompute the missing samples using the reliable samples and the current model parameters' estimates.

- Modified variant of Janssen's algorithm (**Janssenmod**): This method is similar to the AR approach, but instead of processing the signal in overlapping segments, it treats each gap with its context individually.

- Cross-faded extrapolation based on AR modeling [5] (**LR**): This is an efficient method for inpainting of compact gaps with sufficiently long reliable contexts. Per each gap it consists of estimating two AR models for the left and right contexts, extrapolating the contexts into the gap using the fitted model and cross-fading the two candidate solutions.

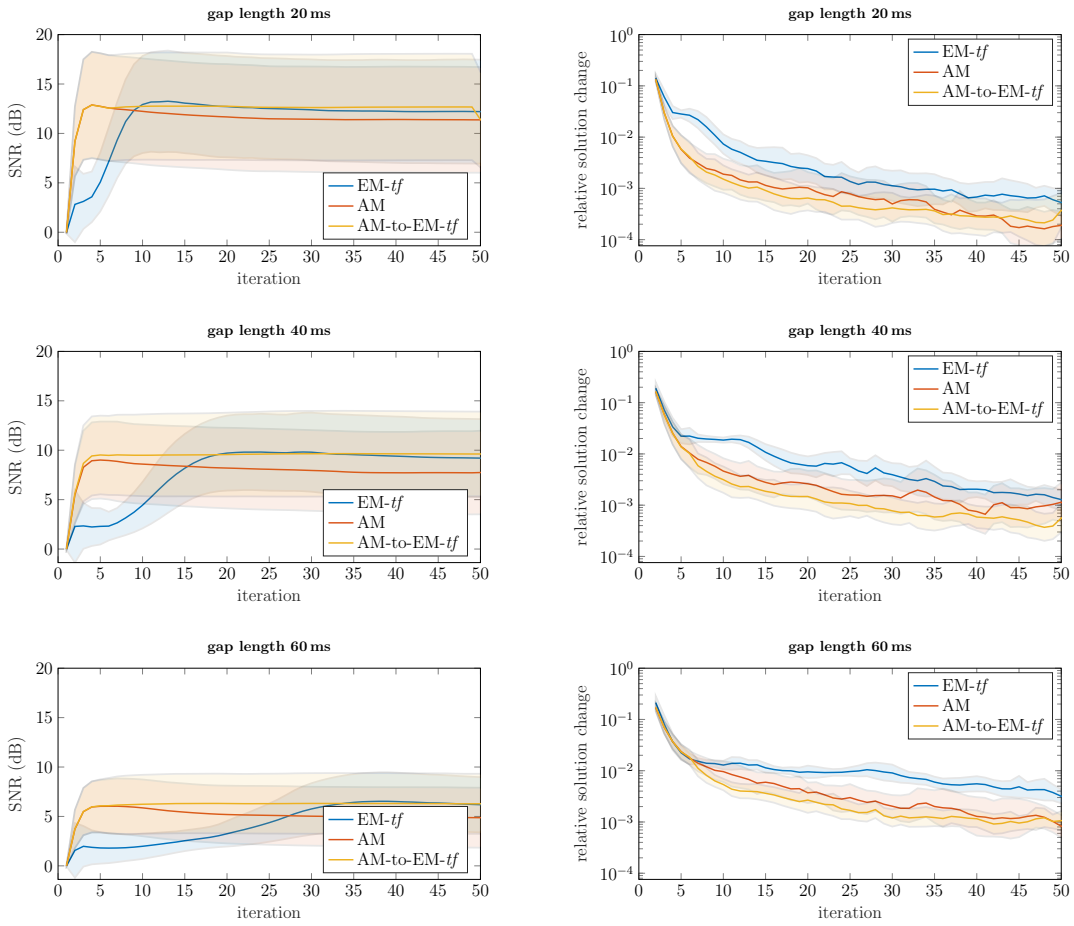

Figure 2: Comparison of the performance of EM-*tf*, AM, and the switching variant AM-to-EM-*tf* (switching after 5 iterations). The left column shows the evolution of the SNR over iterations, the right column shows the relative solution change. Both the metrics are averaged over the dataset and plotted together with 95% confidence interval represented by the light colored areas (note that for the relative solution change plot, this confidence interval is computed from the decimal logarithm of the data).

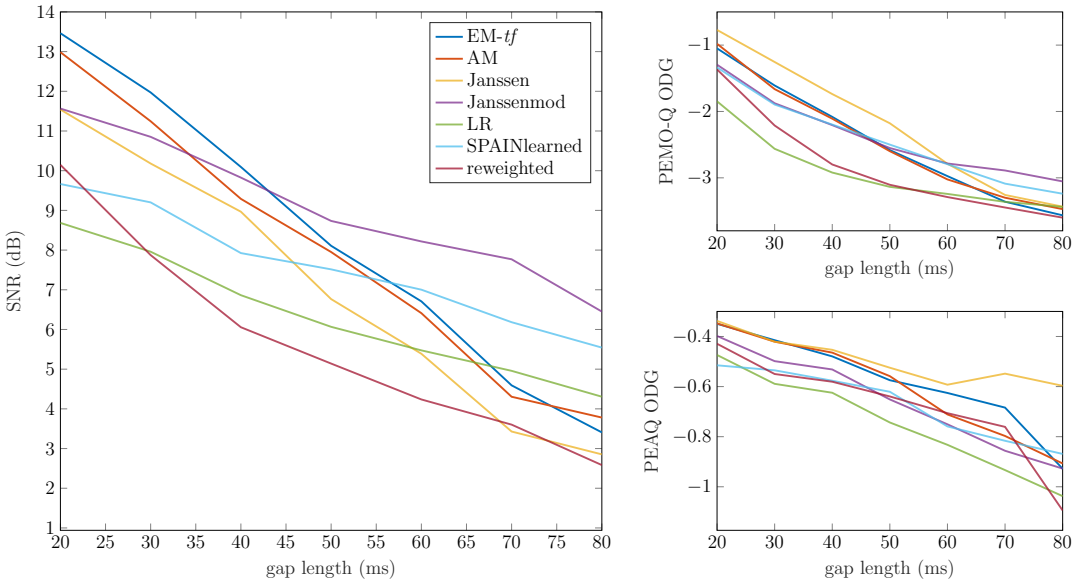

Figure 3: Comparison with the state-of-the-art algorithms for inpainting short to middle-length gaps. The legend is common for all the plots.

- **SPAINlearned** [8], a variant of the sparsity-based non-convex approach SPAIN [7], which treats individual gaps and their contexts (instead of overlapping frames as in SPAIN). The performance is further enhanced by a dictionary learning step to deform the STFT based on the particular signal such that it allows for a sparser representation than using the STFT.

- A convex alternative is the (weighted) ℓ_1 minimization as a relaxation of the non-convex sparsity [6] (**reweighted**).

For all methods which use STFT or segmentation (**Janssen**, **SPAINlearned**, **reweighted**, and the NMF-based methods), the window is a sine window of length 4096 samples (approx. 92 ms) with 50% overlap. The NMF-based methods are applied with $M = F$, thus EM-t is equivalent to EM-*tf* and is omitted for brevity. The AR-based methods use a model of order 512. The context of **Janssenmod** and LR is set to 4096 samples, while **SPAINlearned** use a longer context (8192 samples) for the sake of the dictionary learning. These values are chosen based on the corresponding studies, where they have shown good performance. For particular choices of all the parameters of the individual methods, please refer to the published source code.

The performance results averaged over the 10 test signals are shown in Fig. 3. We observe that in terms of SNR the proposed NMF-based methods outperform the state-of-the-art for short gaps (up to 35 ms) while their performance drops for longer gaps. The perceptually-motivated comparison includes PEMO-Q and PEAQ ODG. However, based on the range of the PEAQ ODG values, we do not find these results very informative, and rather include them for the sake of completeness. On the other hand, we observe from the PEMO-Q ODG values that EM-*tf* and AM are among the top three methods for short- to middle-length gaps.

6. Conclusion

In this paper, we derived new estimators for NMF-based audio inpainting. We formulated inpainting as an optimization problem where the goal is to estimate the signal's power spectrum, which is structured using NMF. To that end, we have derived three algorithms, among which two are new, which encompasses and extends previous related works [22, 23]. Even though the proposed estimators build upon the same low-rank assumption about the signal's TF spectrum, we have shown that there are both theoretical and practical differences between them. In particular, they all exhibit a different behavior and lead to different solutions to the audio inpainting problem. Importantly, the novel approaches (EM-tf and AM) improve the convergence rate compared to EM-tf, while reaching similar reconstruction quality. They have also been demonstrated competitive with state-of-the-art audio inpainting methods.

Throughout the derivation, we assumed independence of the temporal frames. A natural extension of the model would be to employ temporal Markov NMF models, as presented e.g., in [27]. Future research will also study the possibility to leverage psychoacoustics in audio inpainting, e.g., by using perceptually-motivated TF representations [9, 37].

References

[1] A. Adler, V. Emiya, M. Jafari, M. Elad, R. Gribonval, M. Plumbley, Audio Inpainting, IEEE Transactions on Audio, Speech, and Language Processing 20 (3) (2012) 922–932. doi:10.1109/TASL.2011.2168211.

[2] J. Lindblom, P. Hedelin, Packet loss concealment based on sinusoidal modeling, in: Speech Coding, 2002, IEEE Workshop Proceedings., IEEE, 2002. doi:10.1109/scw.2002.1215725.

[3] C. Rodbro, M. Murthi, S. Andersen, S. Jensen, Hidden Markov model-based packet loss concealment for voice over IP, IEEE Transactions on Audio, Speech, and Language Processing 14 (5) (2006) 1609 –1623. doi:10.1109/TSA.2005.858561.

[4] A. J. E. M. Janssen, R. N. J. Veldhuis, L. B. Vries, Adaptive interpolation of discrete-time signals that can be modeled as autoregressive processes, IEEE Trans. Acoustics, Speech and Signal Processing 34 (2) (1986) 317–330. doi:10.1109/TASSP.1986.1164824.

[5] W. Etter, Restoration of a discrete-time signal segment by interpolation based on the left-sided and right-sided autoregressive parameters, IEEE Transactions on Signal Processing 44 (5) (1996) 1124–1135. doi:10.1109/78.502326.

[6] O. Mokrý, P. Rajmic, Audio inpainting: Revisited and reweighted, IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (2020) 2906–2918. doi:10.1109/taslp.2020.3030486.

[7] O. Mokrý, P. Záviška, P. Rajmic, V. Veselý, Introducing SPAIN (SParse Audio INpainter), in: 2019 27th European Signal Processing Conference (EUSIPCO), IEEE, 2019.

[8] G. Taubock, S. Rajbamshi, P. Balazs, Dictionary learning for sparse audio inpainting, IEEE Journal of Selected Topics in Signal Processing 15 (1) (2021) 104–119. doi:10.1109/jstsp.2020.3046422.

[9] F. Lieb, H.-G. Stark, Audio inpainting: Evaluation of time-frequency representations and structured sparsity approaches, Signal Processing 153 (2018) 291–299. doi:10.1016/j.sigpro.2018.07.012.

[10] O. Mokrý, P. Rajmic, Approximal operator with application to audio inpainting, Signal Processing 179 (2021) 107807. doi:https://doi.org/10.1016/j.sigpro.2020.107807.

[11] M. Kowalski, K. Siedenburg, M. Dörfler, Social sparsity! neighborhood systems enrich structured shrinkage operators, Signal Processing, IEEE Transactions on 61 (10) (2013) 2498–2511. doi:10.1109/TSP.2013.2250967.

[12] K. Siedenburg, M. Kowalski, M. Dorfler, Audio declipping with social sparsity, in: Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, IEEE, 2014, pp. 1577–1581.

[13] C. Gaultier, S. Kitić, R. Gribonval, N. Bertin, Sparsity-based audio declipping methods: Selected overview, new algorithms, and large-scale evaluation, IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2021) 1174–1187. doi:10.1109/TASLP.2021.3059264.

[14] P. Záviška, P. Rajmic, Analysis social sparsity audio declipper (May 2022). arXiv:2205.10215, doi:10.48550/ARXIV.2205.10215.

[15] A. Marafioti, N. Holighaus, P. Majdak, N. Perraudin, Audio inpainting of music by means of neural networks, in: Audio Engineering Society Convention 146, 2019.
URL http://www.aes.org/e-lib/browse.cfm?elib=20303

[16] A. Marafioti, P. Majdak, N. Holighaus, N. Perraudin, GACELA: A generative adversarial context encoder for long audio inpainting of music, IEEE Journal of Selected Topics in Signal Processing 15 (1) (2021) 120–131. arXiv:2005.05032v1, doi:10.1109/JSTSP.2020.3037506.

[17] D. Lee, H. S. Seung, Algorithms for non-negative matrix factorization, in: T. Leen, T. Dietterich, V. Tresp (Eds.), Advances in Neural Information Processing Systems, Vol. 13, MIT Press, 2000.

[18] Y.-X. Wang, Y.-J. Zhang, Nonnegative matrix factorization: A comprehensive review, IEEE Transactions on Knowledge and Data Engineering 25 (6) (2013) 1336–1353. doi:10.1109/tkde.2012.51.

[19] Z. Huang, A. Zhou, G. Zhang, Non-negative matrix factorization: A short survey on methods and applications, in: Communications in Computer and Information Science, Springer Berlin Heidelberg, 2012, pp. 331–340. doi:10.1007/978-3-642-34289-9_37.

[20] T. Virtanen, Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria, Audio, Speech, and Language Processing, IEEE Transactions on 15 (3) (2007) 1066–1074. doi:10.1109/TASL.2006.885253.

[21] A. Ozerov, C. Févotte, Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation, IEEE Transactions on Audio, Speech, and Language Processing 18 (3) (2010) 550–563. doi:10.1109/tasl.2009.2031510.

[22] Ç. Bilen, A. Ozerov, P. Pérez, Audio declipping via nonnegative matrix factorization, in: 2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2015, pp. 1–5. doi:10.1109/WASPAA.2015.7336948.

[23] Ç. Bilen, A. Ozerov, P. Pérez, Solving time-domain audio inverse problems using nonnegative tensor factorization, IEEE Transactions on Signal Processing 66 (21) (2018) 5604–5617. doi:10.1109/TSP.2018.2869113.

[24] P. Záviška, P. Rajmic, A. Ozerov, L. Rencker, A survey and an extensive evaluation of popular audio declipping methods, IEEE Journal of Selected Topics in Signal Processing 15 (1) (2021) 5–24. doi:10.1109/JSTSP.2020.3042071.

[25] A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society: Series B (Methodological) 39 (1) (1977) 1–22.

[26] C. Févotte, N. Bertin, J.-L. Durrieu, Nonnegative matrix factorization with the Itakura–Saito divergence: With application to music analysis, Neural computation 21 (3) (2009) 793–830.

[27] P. Smaragdis, C. Fevotte, G. J. Mysore, N. Mohammadiha, M. Hoffman, Static and dynamic source separation using nonnegative factorizations: A unified view, IEEE Signal Processing Magazine 31 (3) (2014) 66–75. doi:10.1109/msp.2013.2297715.

[28] S. M. Kay, Fundamentals of Statistical Processing, Volume I: Estimation Theory, Prentice Hall, 1993.

[29] O. Christensen, Frames and Bases, An Introductory Course, Birkhäuser, Boston, 2008.

[30] P. Záviška, P. Rajmic, O. Mokrý, Z. Průša, A proper version of synthesis-based sparse audio declipper, in: 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom, 2019, pp. 591–595. doi:10.1109/ICASSP.2019.8682348.

[31] Z. Průša, P. L. Søndergaard, N. Holighaus, C. Wiesmeyr, P. Balazs, The large time-frequency analysis toolbox 2.0, in: Sound, Music, and Motion, Lecture Notes in Computer Science, Springer International Publishing, 2014, pp. 419–442. doi:{10.1007/978-3-319-12976-1_25}.
URL http://dx.doi.org/10.1007/978-3-319-12976-1_25

[32] EBU SQAM CD: Sound quality assessment material recordings for subjective tests, online (2008).
URL https://tech.ebu.ch/publications/sqamcd

[33] European Broadcasting Union, Geneva, Sound Quality Assessment Material recordings for subjective tests, eBU – TECH 3253 (Sep. 2008).
URL https://tech.ebu.ch/docs/tech/tech3253.pdf

[34] R. Huber, B. Kollmeier, PEMO-Q—A new method for objective audio quality assessment using a model of auditory perception, IEEE Trans. Audio Speech Language Proc. 14 (6) (2006) 1902–1911. doi:10.1109/TASL.2006.883259.

[35] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerends, C. Colomes, M. Keyhl, G. Stoll, K. Brandenburg, B. Feiten, PEAQ – The ITU standard for objective measurement of perceived audio quality, The Journal of the Audio Engineering Society 48 (1/2) (2000) 3–29.
URL http://www.aes.org/e-lib/browse.cfm?elib=12078

[36] P. Kabal, An examination and interpretation of ITU-R BS.1387: Perceptual evaluation of audio quality, Tech. rep., MMSP Lab Technical Report, Dept. Electrical & Computer Engineering, McGill University (May 2002).

[37] T. Necciari, P. Balazs, N. Holighaus, P. L. Søndergaard, The ERBlet transform: An auditory-based time-frequency representation with perfect reconstruction, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013, pp. 498–502. doi:10.1109/ICASSP.2013.6637697.