ARTICLE IN PRESS

Computers and Electrical Engineering xxx (xxxx) xxx

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

Tensor completion algorithms for estimating missing values in multi-channel audio signals☆

Wenjian Ding [a], Zhe Sun [b,*], Xingxing Wu [c,*], Zhenglu Yang [a,*], Jordi Solé-Casals [d,e,*], Cesar F. Caiafa [f,d,*]

[a] College of Computer Science, Nankai University, China
[b] Computational Engineering Applications Unit, Head Office for Information Systems and Cybersecurity, RIKEN, 351-0198 Wako-Shi, Japan
[c] School of Statistics and Data Science, Nankai University, China
[d] College of Artificial Intelligence, Nankai University, Jinnan, Tianjin, China
[e] Data and Signal Processing Research Group, University of Vic-Central University of Catalonia, 08500 Vic, Catalonia, Spain
[f] Instituto Argentino de Radioastronomía - CCT La Plata, CONICET/CIC-PBA/UNLP, 1894 V. Elisa, Argentina

ARTICLE INFO

Keywords:
Audio inpainting
Tensor completion
Signal reconstruction
Multi-channel signals

ABSTRACT

Audio inpainting is a widely used technology in the real world since audio signals with missing data are pervasive in many scenarios. The majority of existing works address the time gaps in single-channel audio signals, while completing multi-channel audio signals is rarely investigated.

In this work, we tackle this issue using four different tensor completion algorithms and we evaluate them on speech audio datasets with gaps in the time domain. Based on extensive quantitative and qualitative experiments, the tensor completion algorithms generally achieve a superior predictive performance, including when the gap duration of the signals reaches values of up to 200 ms. Specifically, the experimental results illustrate that all of the applied tensor completion algorithms yield at least 56% improvement in signal restoration performance compared with single-channel based methods. Therefore, the tensor based approaches can capture the underlying latent structure over different channels to reconstruct incomplete multi-channel data.

1. Introduction

Audio signals are ubiquitous in the real world. Examples include speech, music and other acoustic types of sound. An originally normal audio signal is likely to be corrupted with missing or noisy data. The problem can occur during its acquisition, compression, transmission or decompression. Poor-quality audio can sound unnatural, distracting or even unintelligible to human ears. As a result, restoring incomplete audio signals has drawn extensive attention in recent years. This task is termed as audio inpainting [1].

The most addressed research topic in the field is the single-channel signal restoration. For instance, Janssen et al. presented an autoregressive interpolator method that interpolated missing audio samples in the time domain [2]. Kitic et al. proposed a sparse and cosparse decomposition approach [3] and SPAIN (SParse Audio INpainter) [4]. Nonnegative matrix factorization (NMF)-based approaches have also achieved great success in audio inpainting, including one of the *state-of-the-art* restoration

☆ This paper is for regular issues of CAEE. Reviews processed and approved for publication by the co-Editor-in-Chief Huimin Lu.

* Corresponding authors.
E-mail addresses: zhe.sun.vk@riken.jp (Z. Sun), wuxx@nankai.edu.cn (X. Wu), yangzl@nankai.edu.cn (Z. Yang), jordi.sole@uvic.cat (J. Solé-Casals), ccaiafa@fi.uba.ar (C.F. Caiafa).

https://doi.org/10.1016/j.compeleceng.2021.107561
Received 25 May 2021; Received in revised form 30 September 2021; Accepted 16 October 2021

methods developed in [5], which sought to solve time domain problems using a probabilistic Gaussian model. Recently, due to the impressive achievements made in the field of deep learning, several graph convolutional frameworks have been investigated for matrix completion in general [6,7] and some researchers have applied deep learning approaches for audio restoration tasks [8–10].

The most current approaches to audio inpainting were designed for single-channel audio signals. However, nowadays many audio recordings are comprised of multi-channel signals. In this work, we consider different situations regarding multi-channel audio signals, with gap ranges going from one millisecond to hundreds of milliseconds, with different missing ratios and number of channels. Our goal is to recover the missing parts in the audio, and we resort to four multi-channel based tensor completion algorithms for this purpose.

The main contributions of this work are listed as follows: **(i)** Four characteristically different tensor completion algorithms are applied for audio inpainting in multi-channel audio signals. Previous studies have mostly considered single-channel audio arranged into a matrix, for which matrix completion methods have then been applied to. To the best of our knowledge, tensor completion algorithms have rarely been explored for audio inpainting tasks before. **(ii)** Tensor completion algorithms and matrix factorization based methods are compared under three different scenarios: different gap times, different masking ratios and different number of channels. **(iii)** Extensive experiments have shown that tensor completion algorithms significantly outperform methods based on matrix factorization, regardless of whether the evaluation of the reconstruction or the quality of the signal are considered.

The rest of this paper is organized as follows. The related works concerning tensor completion are reviewed in Section 2. In Section 3, notations and the main ideas for the four tensor completion algorithms are introduced. Section 4 describes the datasets, evaluation methods, simulations of corrupted data and the extensive experimental results. Finally, conclusions are discussed in Section 5.

2. Related work

In real world contexts, sets of data often have multi-modal representations. Compared to matrices and vectors, tensors can represent multi-modal data with complex properties, such as videos, audio signals and images, in a more accurate way. During the past fifteen years, researchers have been investigating the theory of tensor completion algorithms to capture the underlying relationships between latent factors [11–13].

The task of tensor completion involves filling in missing entries in a partially observed tensor. A variety of tensor completion approaches are based on tensor decomposition, which can capture the structural properties of multidimensional data. Two popular tensor factorizations are the Tucker model [14] and the CANDECOMP/PARAFAC (CP) model, also known as the Canonical polyadic decomposition (CPD) [15,16]. It is well-known that tensor factorization can be used to capture multiple latent factors from partially known data. For example, Zhao et al. formulated a fully Bayesian CP factorization, which could automatically determine the CP rank by incorporating a sparsity-inducing prior over all unknown parameters [17]. Zheng et al. introduced a matrix factorization method under smoothness constraints [18]. Cai et al. proposed a two-stage nonconvex algorithm based on CP [19].

Some of the completion methods presented above have been proven to be superior when compared to simple interpolation methods [20]. Considering this, we employ them to estimate the missing values in multi-channel audio signals. This is a newly presented and more challenging learning task than the single-channel signal reconstruction.

3. Materials and methods

3.1. Notation and definitions

A tensor is a multiway array whose order is the number of dimensions. For example, vectors are first-order tensors and matrices are second-order tensors. A third-order tensor is denoted as $X = \left(x_{i_1 i_2 i_3}\right) \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ and its Frobenius norm is denoted as $\|X\|_F := (\sum_{i_1,i_2,i_3} |x_{i_1 i_2 i_3}|^2)^{\frac{1}{2}}$.

The unfolded mode-1 matrix for the tensor X is defined as $X_{(1)} \in \mathbb{R}^{I_1 \times I_2 I_3}$, consisting of all the mode-1 vectors as columns. The mode-1 vectors are obtained by fixing every index except for the one in mode 1. Likewise, the mode-2 and mode-3 vectors are the columns of the unfolded mode-2 and mode-3 matrices, respectively. The operation of vectorizing the tensor X is defined as $vec(X) := vec(X_{(1)}) \in \mathbb{R}^{I_1 I_2 I_3}$, and is done by stacking all the mode-1 vectors.

Audio signals are stored as third-order tensors before tensor completion algorithms are performed. A corrupted audio segment is represented by a third-order tensor $X \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ with I_1 channels, I_3 frames in each channel and I_2 samples or representations for each frame. The tensor X has incomplete entries due to the missing audio samples in the audio clips. The data corruption is represented by two collections of triplets Ω and $\backslash\Omega$ along with a tensor $W = (w_{i_1 i_2 i_3}) \in \mathbb{R}^{I_1 \times I_2 \times I_3}$. If $x_{i_1 i_2 i_3}$ is known, then $(i_1, i_2, i_3) \in \Omega$ and $w_{i_1 i_2 i_3} = 1$; otherwise, if $x_{i_1 i_2 i_3}$ is missing, then $(i_1, i_2, i_3) \in \backslash\Omega$ and $w_{i_1 i_2 i_3} = 0$. The corrupted tensor X and the corruption tensor W are fed to a tensor completion algorithm, whose goal is to recover the missing entries and get a complete tensor Y that satisfies $y_{i_1 i_2 i_3} = x_{i_1 i_2 i_3}$ if $(i_1, i_2, i_3) \in \Omega$.

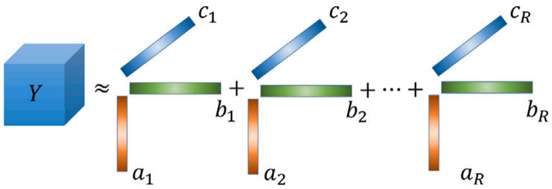

Fig. 1. An R-component CP model for a third-order tensor Y.

3.2. Tensor completion algorithms

3.2.1. CANDECOMP/PARAFAC weighted optimization algorithm

CANDECOMP/PARAFAC (CP) is one of the most well-known tensor factorization technique that captures the multi-linear structure of the tensor. The CP weighted optimization (CP-WOPT) algorithm [21] uses a first-order optimization approach to solve the weighted least squares formulation of the CP problem.

Let $Y = (y_{i_1 i_2 i_3}) \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ be a complete three-way tensor whose rank is R. For CP decomposition, Y can be written as

$$y_{i_1 i_2 i_3} = \sum_{r=1}^{R} a_{i_1 r} b_{i_2 r} c_{i_3 r}, \tag{1}$$

where $Y = (y_{i_1 i_2 i_3}) \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ is the complete tensor, $A = (a_{i_1 r}) \in \mathbb{R}^{I_1 \times R}$, $B = (b_{i_2 r}) \in \mathbb{R}^{I_2 \times R}$ and $C = (c_{i_3 r}) \in \mathbb{R}^{I_3 \times R}$ are factor matrices. Considering the presence of missing entries, Eq. (1) cannot reach absolute equivalence for every triplet. Instead, CP decomposition is formulated to minimize the error function:

$$f(A, B, C) = \frac{1}{2} \sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} \sum_{i_3=1}^{I_3} \left(y_{i_1 i_2 i_3} - \sum_{r=1}^{R} a_{i_1 r} b_{i_2 r} c_{i_3 r} \right)^2. \tag{2}$$

Fig. 1 shows an intuitive illustration of this. As can be observed, the third-order tensor Y can be approximated by a sum of rank-1 tensors.

Nevertheless, CP-based algorithms face the risk of converging to a sub-optimal factorization solution with the increase of missing data. CP-WOPT merely focuses on the known entries and uses a weighted error function to bypass the missing data. In this situation, the goal of CP-WOPT is to find the factor matrices that minimize the weighted function as follows:

$$f(A, B, C) = \frac{1}{2} \sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} \sum_{i_3=1}^{I_3} \left[w_{i_1 i_2 i_3} \left(y_{i_1 i_2 i_3} - \sum_{r=1}^{R} a_{i_1 r} b_{i_2 r} c_{i_3 r} \right) \right]^2, \tag{3}$$

where $w_{i_1 i_2 i_3}$ is defined as

$$w_{i_1 i_2 i_3} = \begin{cases} 1 & \text{if } y_{i_1 i_2 i_3} \text{ is known} \\ 0 & \text{if } y_{i_1 i_2 i_3} \text{ is missing.} \end{cases}$$

Users need to pre-define the tensor rank parameter R. In Section 4, experiments carried out to find the best rank R corresponding to the weighted function will be described.

3.2.2. 3D patch-based tensor completion algorithm

A principle of the compressed sensing theory is that a sparse dictionary-based representation can be used to recover signals from their incomplete observations. Though the sparse representation has proven to be helpful for recovering two-dimensional signals with missing entries, its extension to three dimensional signals is more computationally demanding. To tackle this problem, Caiafa et al. generalized the theory of sparse representations of vectors to tensors [22] and proposed a sparse Tucker decomposition model. When applied to three-dimensional patches of an image, it is referred to as 3D patch-based tensor completion (3DPB-TC) algorithm [23], aiming at approximating the tensor Y by using a large number of small overlapped 3D patches (subtensors): Y_1, Y_2, …, Y_N. The Tucker model decomposes the 3-mode subtensor Y as follows (note that we abbreviate the subtensors Y_n to Y in the following part for simplification, so all the mentioned Y below refer to subtensors rather than the full tensor):

$$y_{i_1 i_2 i_3} = \sum_{r_1=1}^{R_1} \sum_{r_2=1}^{R_2} \sum_{r_3=1}^{R_3} \overline{y}_{r_1 r_2 r_3} d^{(1)}_{i_1 r_1} d^{(2)}_{i_2 r_2} d^{(3)}_{i_3 r_3}, \tag{4}$$

which can be written in the following form:

$$Y = \overline{Y} \times_1 D_1 \times_2 D_2 \times_3 D_3. \tag{5}$$

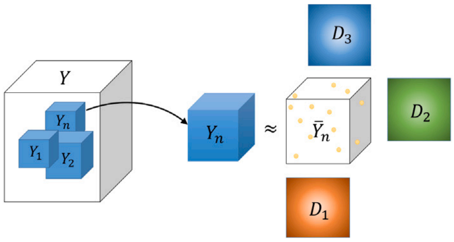

Fig. 2. The 3DPB-TC algorithm: (left) A third-order tensor Y is covered by a large set of overlapped small subtensors (3D-patches). (right) Every 3D-patch Y_n is decomposed using a Sparse Tucker model in which $\overline{Y}_n$ is a "larger" sparse core tensor than Y_n. Finally, missing entries approximations given by different subtensors are averaged to provide final reconstructions of said missing entries.

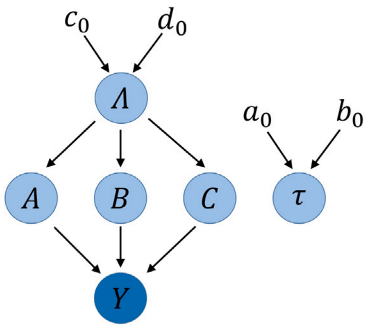

Fig. 3. Probabilistic graphical model of BCPF for a third-order tensor Y.

where $\overline{Y} = (\overline{y}_{r_1 r_2 r_3}) \in \mathbb{R}^{R_1 \times R_2 \times R_3}$ ($R_1 \geq I_1$, $R_2 \geq I_2$ and $R_3 \geq I_3$) is the sparse core tensor, and the factor matrices $D_1 = \left(d^{(1)}_{i_1 r_1}\right) \in \mathbb{R}^{I_1 \times R_1}$, $D_2 = \left(d^{(2)}_{i_2 r_2}\right) \in \mathbb{R}^{I_2 \times R_2}$ and $D_3 = \left(d^{(3)}_{i_3 r_3}\right) \in \mathbb{R}^{I_3 \times R_3}$ are dictionary matrices associated to each mode. We obtain the following expression by vectorizing Eq. (5) as

$$vec(Y) = Dvec(\overline{Y}), D = D_3 \otimes D_2 \otimes D_1. \tag{6}$$

where $vec(Y)$ is a long vector concatenating all the entries of a 3D-patch. The Kronecker product $D \in \mathbb{R}^{I_1 I_2 I_3 \times R_1 R_2 R_3}$ is a global dictionary (matrix) containing "atoms" in its columns. $vec(\overline{Y})$ is vectorized from the sparse core tensor and the non-zero entries of this vector indicate which "atom" is linearly combined to obtain an approximation of Y. In the standard Tucker model, the size of the core tensor $\overline{Y}$ is much smaller than that of Y, while the 3DPB-TC algorithm uses a large and sparse core tensor, as shown in Fig. 2. Once all sparse Tucker decomposition models have been fitted to all available subtensors (3D-patches), every missing entry is reconstructed. Since every tensor decomposition provides an approximation to the overlapped 3D-patch entries, the final estimation is obtained by just averaging all available approximations for every entry.

When dealing with tensors with missing entries, the dictionary D is assumed to be known. This dictionary can be chosen from classical sparsifying transforms, such as wavelets and the cosine transform, or can obtained by applying a dictionary learning algorithm to a complete training dataset, which is what was done in this work. Following [23], an alternate least squares algorithm was used to learn the optimal set of dictionaries from an available collection of 3D-patches. A large number of small 3D-patches were obtained, vectorized and expressed as a sparse representation over a dictionary via Tucker decomposition, where the patch size and the sparsity value needed to be pre-defined. The patch size is associated with the number of channels, and experiments conducted to achieve the optimal sparsity will be described in Section 4. Please refer to [22] for more details on computing sparse representation of signals on a known dictionary.

3.2.3. Bayesian CP factorization

The Bayesian CP factorization (BCPF) algorithm [17] combines a fully Bayesian treatment and CP factorization to infer the rank of the true complete tensor and the underlying multilinear factors. This method can automatically infer all of its parameters without the need for cross validations or likelihood maximization, which are computationally costly and imperative for other tensor completion methods. Fig. 3 illustrates the procedure of BCPF.

The tensor Y can be represented by a CP model with the factor matrices A, B and C. Given a factor matrix $D = [D_1, D_2, \ldots, D_{I_1}]^T$ ($D = A, B$, or C) and a noisy observation Y with i.i.d. noises following Gaussian distributions, the probability of Y_Ω is given as

$$p(Y_\Omega | A, B, C, \tau) = \prod_{i_1=1}^{I1} \prod_{i_2=1}^{I2} \prod_{i_3=1}^{I3} \mathcal{N}(y_{i_1 i_2 i_3} | < A_{i_1}, B_{i_2}, C_{i_3} >, \tau^{-1})^{w_{i_1 i_2 i_3}}, \tag{7}$$

where Ω denotes a set of 3-tuple indices, $(i_1, i_2, i_3) \in \Omega$ if $y_{i_1 i_2 i_3}$ is observed, $W = (w_{i_1 i_2 i_3}) \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ is defined as in the previous section and represents the observed entries, τ is the noise parameter, A_{i_1}, B_{i_2} and C_{i_3} are the row-wise vectors of the corresponding matrices, $< A_{i_1}, B_{i_2}, C_{i_3} >= \sum_r a_{i_1 r} b_{i_2 r} c_{i_3 r}$ denotes a generalized inner-product of the three vectors, and finally $\mathcal{N}(x|x_0, \tau^{-1})$ denotes a Gaussian distribution with mean x_0 and precision τ. Given the hyperparameters $\lambda_i (i = 1, 2, \ldots, R)$, the prior distribution over the factor matrices is given as

$$p(D|\Lambda) = \prod_{i=1}^{I_1} \mathcal{N}(D_i|0, diag(\Lambda)^{-1}), \tag{8}$$

where $\Lambda = [\lambda_1, \lambda_2, \ldots, \lambda_R]$ and $D = A, B, C$. The prior distributions of Λ and τ are given as

$$p(\Lambda) = \prod_{r=1}^{R} G(\lambda_r|c_0^r, d_0^r), \quad p(\tau) = G(\tau|a_0, b_0), \tag{9}$$

where $G(x|a, b) = \frac{b^a x^{a-1} e^{-bx}}{\Gamma(a)}$ is the Gamma distribution with the Gamma function $\Gamma(a)$.

When all of the unknowns are denoted as $\Theta = A, B, C, \Lambda, \tau$, the joint distribution can be written as

$$p(Y_\Omega, \Theta) = p(Y_\Omega|A, B, C, \tau) p(A|\Lambda) p(B|\Lambda) p(C|\Lambda) p(\Lambda) p(\tau). \tag{10}$$

Given the observed data, the probability distribution of Θ is

$$p(\Theta|Y_\Omega) = \frac{p(Y_\Omega, \Theta)}{\int p(Y_\Omega, \Theta) d\Theta}. \tag{11}$$

Finally, the predictive distribution over missing entries $Y_{\backslash\Omega}$ is given as

$$p(Y_{\backslash\Omega}|Y_\Omega) = \int p(Y_{\backslash\Omega}|\Theta) p(\Theta|Y_\Omega) d\Theta. \tag{12}$$

3.2.4. High-accuracy low-rank tensor completion

The ranks of the matrices can be used to estimate the missing data. To calculate them, nonconvex optimization is often used. This can be approximated by calculating the trace norm of the matrices, which is deemed to be the tightest convex approximation. To extend this from matrices to higher-order tensors, Liu et al. defined the tensor trace norm and presented three low-rank tensor completion algorithms, namely simple low-rank tensor completion (SiLRTC), fast low-rank tensor completion (FaLRTC) and high accuracy low rank tensor completion (HaLRTC) [24].

The trace norm of the tensor Y is defined as

$$\|Y\|_* := \sum_{n=1}^{3} \alpha_n \|Y^{(n)}\|_*, \tag{13}$$

where $Y^{(n)}$ is the mode-n unfolded matrix, $Y^{(1)} \in \mathbb{R}^{I_1 \times I_2 I_3}$, $Y^{(2)} \in \mathbb{R}^{I_2 \times I_1 I_3}$, $Y^{(3)} \in \mathbb{R}^{I_3 \times I_1 I_2}$, and α_n satisfies $\alpha_n \geqslant 0$ and $\sum_{n=1}^{3} \alpha_n = 1$.

According to our pilot study, HaLRTC can typically achieve a faster and better performance for visual data completion than SiLRTC. Thus, HaLRTC is employed in this study to solve the optimization problem:

$$min_Y \; : \|Y\|_*, \; s.t. \; Y_\Omega = X_\Omega, \tag{14}$$

where $X, Y \in \mathbb{R}^{I_1 \times I_2 \times I_3}$. Ω is a set containing the index of observed entries in X, which was defined in Section 3.1, and X_Ω is a tensor which has the same entries as X when the index of X_Ω is in Ω. Y is determined such that the trace norm of Y is minimized. HaLRTC uses the alternating direction method of multipliers (ADMM) framework [25] by taking into account the efficiency issue to solve large scale optimization problems.

4. Experiments

4.1. Datasets

The dataset was built from a corpus called voiceHome-2 (found in https://zenodo.org/record/1252143). The corpus consists of audio recordings of speeches made by native French speakers. The recordings have a sample rate of 16 kHz and were done via an 8-microphone device. The eight microphones in the device correspond to eight audio channels among which there is a considerable inherent redundancy and a high correlation, allowing for the recovery of the missing data.

Three 120s-long audio files were selected from the folder named “spontaneous”. The files were recorded in the same home (home 1), room (room 1) and the positions and orientation of the microphones were the same (arrayGeo 1 and arrayPos 1). On the other hand, there were differences in the speakers, the speaker’s positions and the noises. The first file is a recording of spontaneous speech by the speaker F1 at the speakerPos 2 in a noisy condition of noiseCond 2. The second file contains speech articulated by the speaker M1 at the speakerPos 3 in a noisy condition of noiseCond 3. Finally, the third file contains speech articulated by the speaker M2 at the speakerPos 4 in a noisy condition of noiseCond 4. Three 10-second audio segments are extracted from each file, and therefore a total of nine audio segments are used for evaluation. Three additional 10-second audio segments are extracted from each audio file for tuning the parameters. The original eight channels are reduced to six for simplification purposes.

4.2. Performance evaluations

Given the recovered audio $\hat{S}$ and the original audio S, the signal-to-noise ratio $\mathrm{SNR_f}$ is defined as

$$SNR_f = 10\log\frac{\|S\|_F^2}{\|\hat{S}-S\|_F^2}, \tag{15}$$

similarly, $\mathrm{SNR_m}$ is defined as

$$SNR_m = 10\log\frac{\|S_{\setminus\Omega}\|_F^2}{\|\hat{S}_{\setminus\Omega}-S_{\setminus\Omega}\|_F^2}. \tag{16}$$

$\mathrm{SNR_f}$ reflects the global restoration quality, while $\mathrm{SNR_m}$ represents the average reconstruction performance for each sample over the missing part of the audio. In fact, $\mathrm{SNR_f}$ is the sum of $\mathrm{SNR_m}$ with a bias that is irrelevant to the completion algorithm but indicates the audio degradation rate [1].

The relative standard error (RSE) is a measure of the difference between the original signal S and the reconstructed signal $\hat{S}$, and is defined as follows:

$$RSE(\hat{S}_{\setminus\Omega}, S_{\setminus\Omega}) = \frac{\|\hat{S}_{\setminus\Omega}-S_{\setminus\Omega}\|_F}{\|S_{\setminus\Omega}\|_F}. \tag{17}$$

Hence, $\mathrm{SNR_m}$ can be expressed as

$$SNR_m = -20\log RSE(\hat{S}_{\setminus\Omega}, S_{\setminus\Omega}). \tag{18}$$

The $\mathrm{SNR_m}$ for each segment of the testing dataset is computed and the average is found. This average $\mathrm{SNR_m}$ is then used to evaluate the reconstruction performance of the algorithms (the term is abbreviated to SNR for simplification purposes from now on).

When it comes to audio quality assessment, it is well known that SNR does not perform well in audio analysis compared to mean opinion score (MOS) assessments. Perceptual Objective Listening Quality Analysis (POLQA), Perceptual Evaluation of Speech Quality (PESQ) or Perceptual Evaluation of Audio Quality (PEAQ) are examples of perceptual audio quality assessment measurements. To assess speech quality in the experiments, PESQ will be used.

4.3. Simulation of corrupted data

The objective of this work is to evaluate the recovery ability of the tensor completion algorithms on audio signals with missing data. In the parameter tuning part, the gap duration of the training data was fixed to 10 ms, the missing ratio was set to 10% and all of the six channels were selected. The average SNR of the three training audio samples was calculated in order to select the optimal parameters.

In order to comprehensively evaluate the four tensor-based methods, three types of missing data were artificially generated to be used as the testing data in the following forms. For each experiment, we averaged the results across nine testing audio segments. The detailed experimental configurations are set as follows:

- The gap duration was in the range from 1 ms to 200 ms. For each audio segment, the number of missing samples was fixed to 10% with all of the six channels.
- The missing ratio of audio signals was in the range from 1% to 70%, and all of the six channels and a fixed 10 ms gap duration for all samples were selected for the experiments.
- The number of channels was set to 1, 2, 3, and 6, respectively. The gap duration and missing ratio of each audio signal were fixed to 10 ms and 10%, respectively.

Extensive quantitative and qualitative experiments were conducted on the three types of datasets using the four tensor completion algorithms to evaluate their reconstructing performances of the lost time domain audio samples.

4.4. Results

In this section, we first test the CP-WOPT and 3DPB-TC algorithms on the training audio segments to find the optimal parameters under the SNR. BCPF and HaLRTC are the parameter-free methods, and therefore this part is skipped. Then, these four tensor completion algorithms are run on different gap times, missing ratios and number of channels and their predictive performance is evaluated. In each of the evaluation stages, extensive experiments are carried out on the audio samples to compare the tensor completion algorithms.

4.4.1. Parameters tuning

For the CP-WOPT tensor completion algorithm, users need to pre-define the tensor rank parameter R. The algorithm was tested using the training audio segments with the rank R being in the range of {10, 20, 50, 75, 100, 150, 200}. As shown in Fig. 4, CP-WOPT is first enhanced when R increases and obtains its best restoration results when R is 75. However, its predictive performance begins to decline when the rank R exceeds 75.

The algorithm 3DPB-TC uses a global dictionary associated with a patch size and a sparsity value, both of which are user-defined. As shown in Fig. 5, the best sparsity was found to be $\rho = 0.2$.

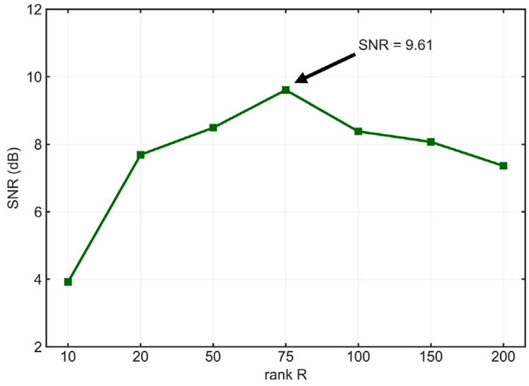

Fig. 4. SNR evaluation as a function of rank R for the CP-WOPT algorithm. The algorithm is performed on the training audio segments with the gap duration set at 10 ms, the missing ratio set at 10% and the number of channels set at 6. The CP-WOPT algorithm achieves its best performance when $R = 75$.

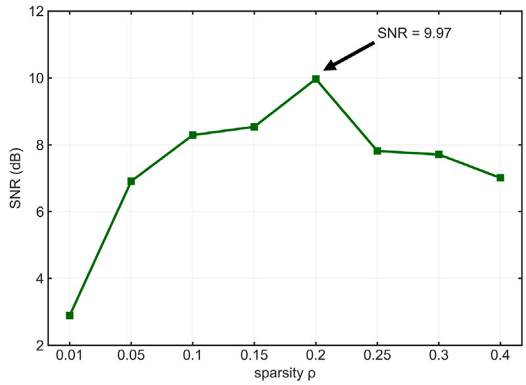

Fig. 5. Determination of sparsity ρ for the 3DPB-TC algorithm. The experiments were conducted on the training audio segments with the gap duration set at 10 ms, the missing ratio set at 10% and the number of channels set at 6. The 3DPB-TC algorithm achieves its maximum SNR when $\rho = 0.2$.

4.4.2. Gap time evaluation

The restoration performance of the tensor completion algorithms was evaluated for a variable duration of missing intervals of samples. The ratio of the missing samples was fixed to 10% for each audio segment and the number of channels was set to 6. The results were obtained by averaging the predictive performance across the nine testing audio segments with a fixed missing ratio and number of channels. The missing gap duration was in the range from 1 ms to 200 ms.

For an extensive comparison, we evaluated the performance of the four multi-channel based tensor completion algorithms with four single-channel based methods: Janssen's method [2], Analysis SParse Audio INpainter (A-SPAIN), Synthesis SParse Audio INpainter with hard thresholding (S-SPAIN H) and Synthesis SParse Audio INpainter Orthogonal Matching Pursuit (S-SPAIN OMP) [4]. Janssen's method is an interpolation-based method, and A-SPAIN, S-SPAIN H and S-SPAIN OMP are matrix factorization based methods. Note that these four methods used for comparison are applied frame-wise and are developed for single channel audio. The reconstruction is executed frame by frame for each channel, and the SNR is calculated by concatenating every channel to a tensor. The results are presented in Fig. 6, and the detailed information including SNR, PESQ and the running time is recorded in Table 1.

Running the four tensor completion algorithms and the four competitor algorithms results in various SNR values for each different gap time (see Fig. 6). In general, the tensor completion algorithms achieve better recovery results than the single-channel based methods. Furthermore, all of the four tensor completion algorithms applied have the ability to recover the missing data even for gap times of up to 200 ms, while the recovery performances of the single-channel based strategies generally get worse as the gap time becomes larger, indicating that the single-channel based methods are not robust with regards to gap duration. As a result, the tensor based (multi-channel based) methods outperform the single-channel based methods. The three matrix factorization based methods (A-SPAIN, S-SPAIN H and S-SPAIN OMP) perform better than the interpolation-based method (Janssen's method). Meanwhile, as can be seen from Table 1, even though the tensor-based (multi-channel based) methods have a better reconstruction performance, they also require more running time for convergence than the single-channel based strategies. To be more specific, the running times of 3DPB-TC and BCPF are more than 1000 s, but in turn they can obtain an excellent performance particularly when the gap duration is in the 5 ms to 20 ms range. HaLRTC is a more time-friendly tensor completion method, but it only performs particularly well when the gap time is 1 ms or 2 ms. CP-WOPT offers a balanced solution among the multi-channel based algorithms. To summarize,

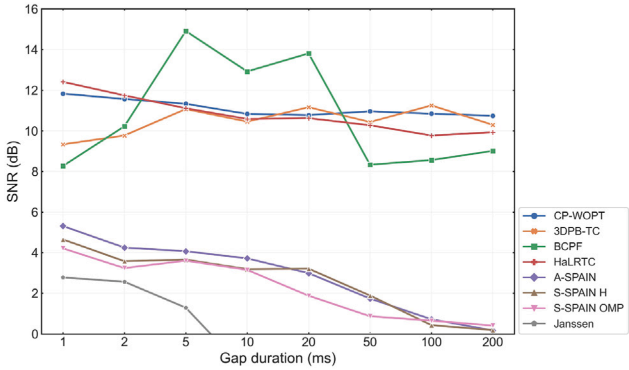

Fig. 6. Audio inpainting results with varying gap durations. Four tensor completion methods and their competitors are performed on the nine audio segments with the gap duration set to 1 ms, 2 ms, 5 ms, 10 ms, 20 ms, 50 ms, 100 ms and 200 ms. Only results with a positive SNR were visualized in the figure.

Table 1
The recovery performance SNR, PESQ and running time (seconds) for the eight methods with different gap times.

		CP-WOPT	3DPB-TC	BCPF	HaLRTC	Janssen	A-SPAIN	S-SPAIN H	S-SPAIN OMP
1 ms	SNR	11.83	9.33	8.27	**12.41**	2.78	5.31	4.65	4.21
	PESQ	**3.91**	3.86	3.79	3.88	2.96	3.15	3.23	3.31
	Runtime	208.72	10318.46	1380.63	53.66	**19.36**	42.81	33.28	470.76
2 ms	SNR	11.56	9.77	10.22	**11.74**	2.57	4.25	3.58	3.24
	PESQ	3.85	3.88	3.86	**3.89**	2.73	3.22	3.20	3.15
	Runtime	203.21	9219.88	1022.56	49.01	**15.28**	44.38	34.41	490.55
5 ms	SNR	11.33	11.07	**14.91**	11.11	1.29	4.07	3.66	3.61
	PESQ	3.90	3.91	**3.95**	3.84	2.41	3.07	3.14	3.22
	Runtime	227.81	9237.68	1027.23	41.23	**9.74**	33.96	26.87	361.91
10 ms	SNR	10.83	10.45	**12.92**	10.58	−1.90	3.72	3.19	3.15
	PESQ	3.86	3.75	**3.87**	3.71	2.08	2.95	3.05	3.00
	Runtime	192.82	9663.41	1182.17	40.68	**13.49**	25.49	19.05	272.82
20 ms	SNR	10.77	11.16	**13.81**	10.62	−0.83	2.99	3.21	1.88
	PESQ	**3.83**	3.73	3.81	3.78	1.64	3.11	2.98	3.02
	Runtime	204.57	10394.81	1114.68	39.46	17.29	13.51	**9.45**	111.03
50 ms	SNR	**10.96**	10.43	8.33	10.27	−0.09	1.74	1.89	0.87
	PESQ	3.79	**3.82**	3.62	3.77	1.76	2.88	2.74	2.91
	Runtime	196.56	9141.71	1053.43	38.78	21.08	6.87	**4.87**	62.12
100 ms	SNR	10.84	**11.25**	8.56	9.77	−0.22	0.72	0.43	0.65
	PESQ	3.81	3.80	3.69	**3.82**	1.55	2.91	2.82	2.75
	Runtime	186.88	9953.36	1184.60	36.85	16.45	3.09	**2.76**	89.26
200 ms	SNR	**10.74**	10.29	9.01	9.93	−6.91	0.17	0.19	0.41
	PESQ	3.74	3.77	3.71	**3.80**	Nan	2.44	2.63	2.69
	Runtime	198.61	10147.09	1059.83	37.04	21.62	**1.02**	1.52	97.54

the single-channel based methods require less running time than the multi-channel based methods, except for S-SPAIN OMP, and less running time is needed as the gap time becomes larger.

4.4.3. Masking ratios

In this experiment, the four tensor completion algorithms with different masking ratios of the audio signals are compared with the four single-channel based methods acting as the baseline in order to measure the effects of the missing ratios in audio inpainting. The number of missing samples of all the gaps is in the range from 1% to 70%. The gap time is fixed to 10 ms for each audio segment, and the number of channels is set to 6.

As shown in Fig. 7, the tensor completion algorithms significantly outperform the single-channel based methods for all the missing ratios of the audio signals. It can also be observed that all of the algorithms, including the comparison methods, perform worse as the missing ratios increase. When the missing part of the audio signals reaches 70%, the matrix factorization based methods perform competitively compared with the CP-WOPT, 3DPB-TC and HaLRTC tensor completion methods. This result illustrates that the matrix factorization based methods offer a more stable performance as the missing ratio increases. The BCPF method achieves the best performance out of all the other algorithms in most of the missing ratio cases. From Tables 1 and 2, we can conclude that with an increase of the gap time, the running time of the matrix factorization based methods continues to decrease, while that of the other methods remains essentially unchanged. As the missing ratio continues to increase, the running time of the Janssen, A-SPAIN and S-SPAIN H algorithms increases while that of the other methods remains more stable.

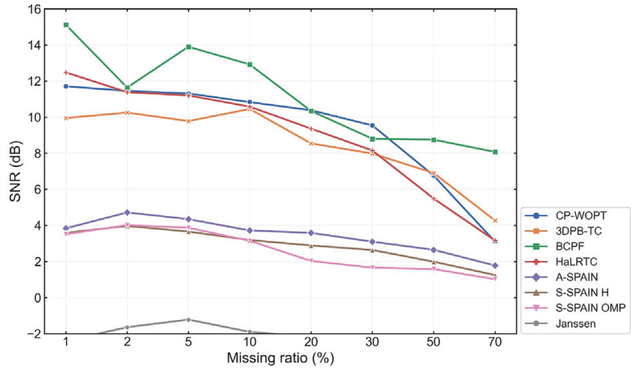

Fig. 7. Predictive performance with varying missing ratios. The algorithms are performed on the nine audio segments with missing ratios of 1%, 2%, 5%, 10%, 20%, 30%, 50% and 70%. Only results with a positive SNR were visualized in the figure.

Table 2
The recovery performance SNR, PESQ and running time (seconds) for the eight methods with different missing ratios. In the table, "Nan" means that the algorithm did not converge.

		CP-WOPT	3DPB-TC	BCPF	HaLRTC	Janssen	A-SPAIN	S-SPAIN H	S-SPAIN OMP
1%	SNR	11.71	9.95	**15.12**	12.48	−2.41	3.84	3.59	3.51
	PESQ	3.95	3.87	**4.02**	3.74	2.56	3.12	3.37	3.26
	Runtime	207.50	9532.11	1262.29	31.54	**3.06**	4.65	3.29	79.39
2%	SNR	11.45	10.25	**11.64**	11.38	−1.64	4.72	3.97	4.02
	PESQ	3.89	3.92	**3.98**	3.81	2.35	3.16	3.25	3.18
	Runtime	242.29	9689.54	1197.55	31.16	**5.61**	9.82	6.84	68.54
5%	SNR	11.31	9.78	**13.90**	11.20	−1.22	4.35	3.66	3.87
	PESQ	3.83	3.80	**3.91**	3.62	2.16	3.09	3.13	3.04
	Runtime	234.57	9040.87	1153.21	31.48	**7.83**	14.74	13.03	143.39
10%	SNR	10.83	10.45	**12.92**	10.58	−1.90	3.72	3.19	3.15
	PESQ	3.86	3.75	**3.87**	3.71	2.08	2.95	3.05	3.00
	Runtime	192.82	9663.41	1182.17	40.68	**13.49**	25.49	19.05	272.82
20%	SNR	**10.38**	8.54	10.34	9.36	−67.39	3.59	2.89	2.03
	PESQ	3.72	3.64	**3.81**	3.73	Nan	2.99	2.96	2.92
	Runtime	205.24	9054.78	1223.19	37.17	25.51	30.45	**24.37**	238.31
30%	SNR	**9.54**	7.98	8.80	8.15	Nan	3.10	2.64	1.66
	PESQ	**3.77**	3.57	3.62	3.69	Nan	2.86	2.89	2.81
	Runtime	239.99	9093.15	1036.63	39.99	65.16	29.86	**21.39**	225.11
50%	SNR	6.76	6.90	**8.75**	5.49	Nan	2.65	1.99	1.58
	PESQ	3.23	3.16	**3.64**	3.41	Nan	2.75	2.86	2.74
	Runtime	178.26	9669.04	1109.97	41.24	134.27	19.98	**12.03**	99.63
70%	SNR	3.11	4.28	**8.07**	3.17	Nan	1.78	1.25	1.02
	PESQ	3.01	3.05	**3.59**	3.25	Nan	2.94	2.71	2.85
	Runtime	159.14	10262.86	1298.65	39.92	186.35	19.87	**15.87**	76.84

4.4.4. *Effects of the number of channels*

In this section, the restoration abilities of the four tensor completion algorithms is evaluated when the number of channels is 1, 2, 3 and 6. These experiments can validate the capacity of the model for factorizing the incomplete dataset with the goal of capturing the latent structure between different channels and possibly reconstructing missing values. The missing ratio is set to 10%, and the gap duration is in the range from 1 ms to 200 ms. Fig. 8 illustrates the performances of the different methods with different numbers of channels.

As can be seen from Figs. 8a, 8b and 8d, which present the results for CP-WOPT, 3DPB-TC and HaLRTC, respectively, a larger number of channels can lead to a higher reconstruction performance. Nevertheless, Fig. 8c shows some intersections between different number of channels, although generally a larger number of channels is much more efficient for obtaining a higher SNR. For the comparison methods, the last four subplots in Fig. 8 present varying results for each different number of channels, illustrating that for these single-channel based methods, an increased number of channels does not necessarily lead to a larger SNR, unlike for the multi-channel based methods. Note that when the number of channels is 1, the SNR of the tensor methods is in the range from 2 dB to 6 dB, which is also similar to the performance of the matrix factorization based methods, because when the number of channels is 1, the multi-channel audio files are single-channel audio signals, and thus the tensor completion algorithms degenerate into matrix factorization based methods. As the number of channels increases, the tensor completion methods outperform the matrix

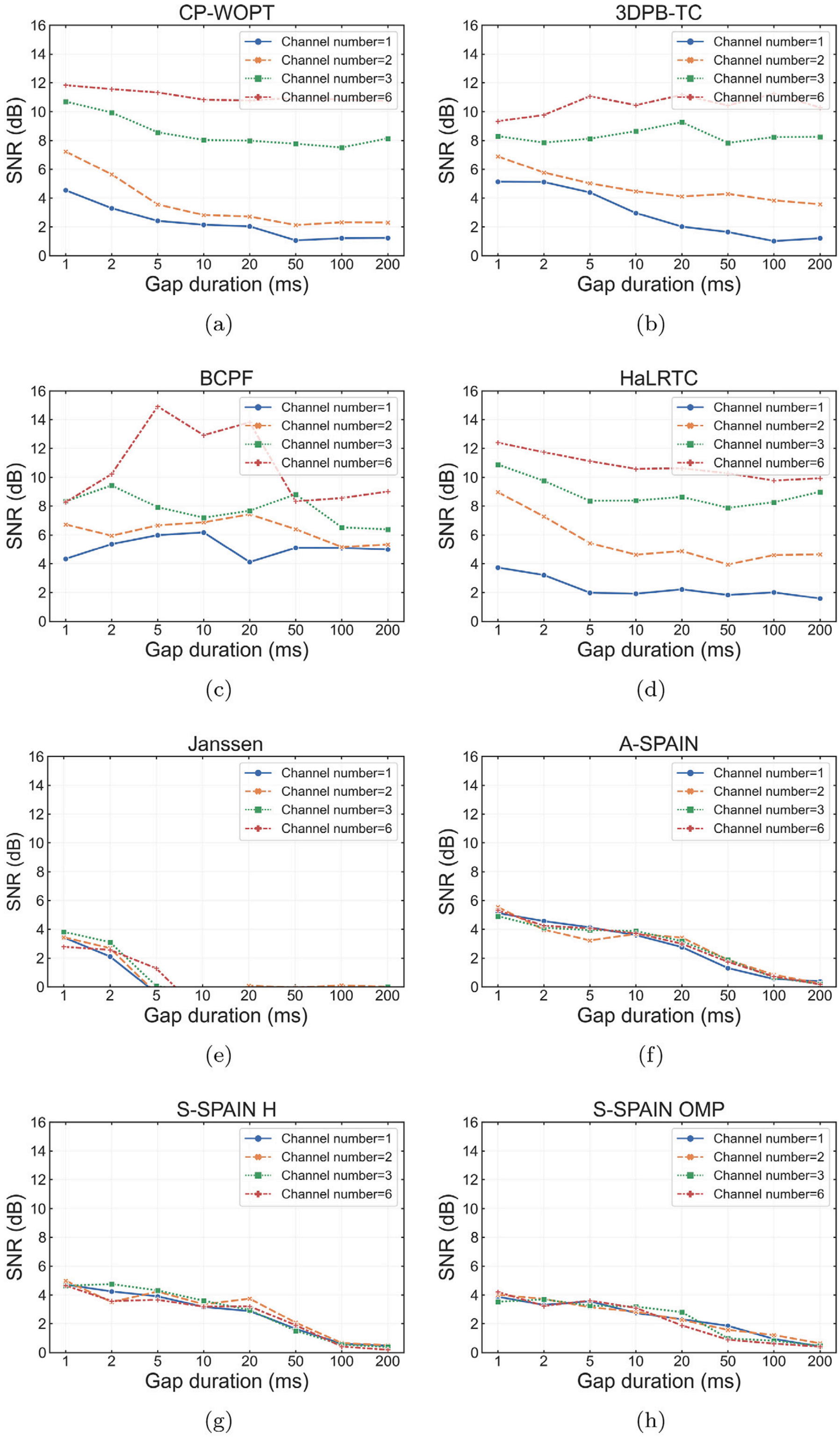

Fig. 8. Reconstruction performance for different gap durations and different number of channels. Only results with a positive SNR were visualized in the figure.

factorization based strategies, which indicates that the four methods can capture the underlying structure between different channels and therefore reconstruct the missing values.

The results from these figures indicate that the recovering abilities of the four tensor completion algorithms are superior when the number of channels is increased. We can conclude that the four tensor-based methods can effectively capture the hidden correlations among the different channels to estimate the missing values in the audio signals.

5. Discussion and conclusions

In this paper, four characteristically different tensor completion algorithms (i.e., CP-WOPT, 3DPB-TC, BCPF, and HaLRTC) were used to estimate missing values in multi-channel audio signals. The experimental results have demonstrated that the tensor completion methods can be used to perform audio inpainting of the signals with some missing data in the time domain. Specifically, the tensor completion algorithms show an improved performance with different gap times and missing ratios when compared with the single-channel based methods. Furthermore, the tensor completion algorithms perform better on multi-channel audio signals than on single-channel audio signals, indicating that there are underlying structures between different channels and the tensor-based methods can capture these latent factors and reconstruct the missing values.

CRediT authorship contribution statement

Wenjian Ding: Conceptualization, Methodology, Data curation, Software, Writing – original draft. **Zhe Sun:** Conceptualization, Methodology, Data curation, Supervision, Writing - review & editing. **Xingxing Wu:** Conceptualization, Visualization. Investigation, Project administration, Writing – original draft. **Zhenglu Yang:** Conceptualization, Methodology, Supervision, Project administration, Writing - review & editing. **Jordi Solé-Casals:** Methodology, Software, Writing - review & editing. **Cesar F. Caiafa:** Methodology, Software, Writing - review & editing.

Declaration of competing interest

No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.compeleceng.2021.107561.

Acknowledgments

CF Caiafa work was partially supported by grants PICT 2017-3208, UBACYT 20020190200305BA and UBACYT 20020170100 192BA (Argentina). This work is also based upon work from COST Action CA18106, supported by COST (European Cooperation in Science and Technology). We thank Pau Solé-Vilaró for checking the grammar, the style and the syntax of this manuscript.

References

[1] Adler A, Emiya V, Jafari MG, Elad M, Gribonval R, Plumbley MD. Audio inpainting. IEEE Tran Audio Speech Lang Process 2011;20(3):922–32.
[2] Janssen A, Veldhuis R, Vries L. Adaptive interpolation of discrete-time signals that can be modeled as autoregressive processes. IEEE Trans Acoust Speech Signal Process 1986;34(2):317–30.
[3] Kitić S, Bertin N, Gribonval R. Sparsity and cosparsity for audio declipping: A flexible non-convex approach. In: International conference on latent variable analysis and signal separation. Springer; 2015, p. 243–50.
[4] Mokrỳ O, Záviška P, Rajmic P, Veselỳ V. Introducing spain (sparse audio inpainter). In: 2019 27th European signal processing conference. IEEE; 2019, p. 1–5.
[5] Bilen Ç, Ozerov A, Pérez P. Solving time-domain audio inverse problems using nonnegative tensor factorization. IEEE Trans Signal Process 2018;66(21):5604–17.
[6] Berg Rvd, Kipf TN, Welling M. Graph convolutional matrix completion. 2017, arXiv preprint arXiv:1706.02263.
[7] Li J, Zhang S, Liu T, Ning C, Zhang Z, Zhou W. Neural inductive matrix completion with graph convolutional networks for miRNA-disease association prediction. Bioinformatics 2020;36(8):2538–46.
[8] Lee MS. Deep learning restoration of signals with additive and convolution noise. In: Pham T, Solomon L, editors. Artificial intelligence and machine learning for multi-domain operations applications III, Vol. 11746. International Society for Optics and Photonics, SPIE; 2021, p. 285–96. http://dx.doi.org/10.1117/12.2585170.
[9] Marafioti A, s, Holighaus N, Majdak P, Perraudin N, l. Audio inpainting of music by means of neural networks. In: Audio engineering society convention, Vol. 146. 2019, URL http://www.aes.org/e-lib/browse.cfm?elib=20303.
[10] Ebner PP, Eltelt A. Audio inpainting with generative adversarial network. 2020, arXiv, arXiv:2003.07704.
[11] Kolda TG, Bader BW. Tensor decompositions and applications. SIAM Rev 2009;51(3):455–500.
[12] Cichocki A, Zdunek R, Phan AH, Amari S-i. Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation. John Wiley & Sons; 2009.
[13] Cichocki A, Mandic D, De Lathauwer L, Zhou G, Zhao Q, Caiafa C, et al. Tensor decompositions for signal processing applications: from two-way to multiway component analysis. IEEE Signal Process Mag 2015.
[14] Tucker LR. Some mathematical notes on three-mode factor analysis. Psychometrika 1966;31(3):279–311.
[15] Carroll JD, Chang J-J. Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. Psychometrika 1970;35(3):283–319.
[16] Harshman RA, et al. Foundations of the PARAFAC procedure: Models and conditions for an" explanatory" multimodal factor analysis. 1970.
[17] Zhao Q, Zhang L, Cichocki A. Bayesian CP factorization of incomplete tensors with automatic rank determination. IEEE Trans Pattern Anal Mach Intell 2015;37(9):1751–63.
[18] Zheng Y-B, Huang T-Z, Ji T-Y, Zhao X-L, Jiang T-X, Ma T-H. Low-rank tensor completion via smooth matrix factorization. Appl Math Model 2019;70:677–95.

[19] Cai C, Li G, Poor HV, Chen Y. Nonconvex low-rank symmetric tensor completion from noisy data. 2019, arXiv preprint arXiv:1911.04436.
[20] Solé-Casals J, Caiafa CF, Zhao Q, Cichocki A. Brain-computer interface with corrupted EEG data: A tensor completion approach. Cogn Comput 2018;10(6):1062–74.
[21] Acar E, Dunlavy DM, Kolda TG, Mørup M. Scalable tensor factorizations for incomplete data. Chemometr Intell Lab Syst 2011;106(1):41–56.
[22] Caiafa CF, Cichocki A. Computing sparse representations of multidimensional signals using kronecker bases. Neural Comput 2013;25(1):186–220.
[23] Caiafa CF, Cichocki A. Multidimensional compressed sensing and their applications. Wiley Interdisciplinary Rev: Data Min. Knowl. Discov. 2013;3(6):355–80.
[24] Liu J, Musialski P, Wonka P, Ye J. Tensor completion for estimating missing values in visual data. IEEE Trans Pattern Anal Mach Intell 2012;35(1):208–20.
[25] Boyd S, Parikh N, Chu E. Distributed optimization and statistical learning via the alternating direction method of multipliers. Now Publishers Inc; 2011.

Wenjian Ding: He received the BS degree from Hunan University, and the MS degree from Fordham University. His research interests include natural language processing and speech recognition.

Zhe Sun: Research Scientist at Computational Engineering Applications Unit, R&D Group, Head Office for Information Systems and Cybersecurity, RIKEN, Japan. His current research interests include large scale brain simulation and neuromorphic engineering, machine learning and brain signal processing.

Xingxing Wu: Research Assistant at the School of Statistics and Data Science in Nankai University, China. His main research interests include computational neuroscience and data mining.

Zhenglu Yang: He received the BS degree from Tsinghua University, and the MS and Ph.D. degrees from the University of Tokyo. He is currently working as a professor at the College of Computer Science and School of Statistics & Data Science in Nankai University, China. His main research interests include natural language processing, data mining and web search.

Jordi Solé-Casals: Full Professor at the University of Vic-Central University of Catalonia (Barcelona, Catalonia), and Visiting Researcher at the Department of Psychiatry (University of Cambridge, UK) and at the College of Artificial Intelligence (Nankai University, China). His research interests include signal processing (especially in the biomedical field), machine learning/deep learning, and statistical modeling for applied sciences.

Cesar F. Caiafa: Independent Researcher at CONICET and Adjunct Professor at Engineering Faculty – University of Buenos Aires, Argentina. Visiting Scientist at the RIKEN Center for Advanced Intelligence Project, Japan; and at the College of Artificial Intelligence (Nankai University, China). He currently works on machine learning algorithms exploiting tensor decompositions and sparsity with applications ranging from Neuroscience to Astronomy.