BRNO UNIVERSITY OF TECHNOLOGY

**FACULTY OF ELECTRICAL ENGINEERING
AND COMMUNICATION**

DEPARTMENT OF TELECOMMUNICATIONS

INPAINTING OF MISSING AUDIO SIGNAL SAMPLES

DOPLŇOVÁNÍ CHYBĚJÍCÍCH VZORKŮ V AUDIO SIGNÁLU

SHORT VERSION OF PH.D. THESIS

Ing. Václav MACH

Specialization: Teleinformatics
Supervisor: doc. Mgr. Pavel Rajmic, Ph.D.
Opponents:
Date of defence:

KEYWORDS

Sparse Representations, Audio Inpainting, Proximal Algorithms, Audio Restoration, Denoising.

KLÍČOVÁ SLOVA

Řídké reprezentace, interpolace signálů, proximální algoritmy, restaurace zvuku, odšumování.

Doctoral thesis is available at the Science Department of Dean’s Office FEEC, Brno University of Technology, Technická 10, Brno, 616 00

ISBN 80-214-
ISSN 1213-4198

OBSAH

1 Introduction

Historical sound recordings usually suffer from imperfections. The history of sound recordings started at the end of nineteenth century with wax cylinders, followed by shellac, acetate, vinyl disc recordings, stainless steel wire recordings and magnetic tapes. Typical distortions like a hiss, impulse noise, crackle, wow and flutter, background noise or power line hum are a natural part of such archive audio sources. The high quality digital audio brought by Digital Audio Tape (DAT) or Compact Disc (CD) caused an enormous increase of sound quality demands. Meanwhile, the interest in nostalgic and historical material was still retained. Therefore, requirement for the audio restoration of degraded recordings grew.

Audio restoration is a generalized term for the process of removing faultiness from sound recordings. The perfect restoration would reconstruct the original audio exactly as captured by the transducer (e.g. microphone, reproducer, horn). Original pure quality can, of course, never be reached in practice. However, there are methods which can come close according to some suitable error criterion ideally based on the perceptual characteristics of the human hearing.

Recently, after more than one hundred years, the wax cylinders captured at the end of 19th century by the famous music composer Leoš Janáček and his collaborators were re-recorded by the digital systems, digital restoration was performed and they are available together with detailed study for a large audience in [26].

The theoretical interest in the process of audio restoration was enforced by cooperation with The Institute of Ethnology of the Academy of Sciences of the Czech Republic, v.v.i. and the National Institute of Folk Culture which led in joint research interests in audio signal processing and the ethnological research. The restoration of ancient audio recordings was performed in a sensitive way to the music genre. The experience of treating a lot of sound impairments using a commercial software resulted in experiments using novel restoration methods based on the sparse representations.

Audio restoration is not only the matter of historical recordings. Nowadays, professional audio/video production has human and technical resources that are mostly aware of audio processing failures. However, people who are not experienced in working with recording equipment can cause signal degradation which is usually discovered in the post-processing or mastering stage.

All of the algorithms and software tools are a great assistants for audio mixing, post-processing, mastering and publishing. However, they should never be used for audio archiving, since the original information is lost during the restoration process.

Condition of the material is clearly affected by the age of the sound information carrier. The description of common types of damages is based on a detailed study of the collection of wax cylinders [33]. Typical damages of audio carriers are: scratches, fissures and weld, scratching off the groove, mould, pitch fluctuation. The consequent audio signal artifacts coming out of these damages are: clicks, crackles, noise, hum and rumbling or clipping.

2 State of the Art

The early methods of interpolation of the missing samples portions performed an *extrapolation*, i.e. expanding the signal from only one side of the gap. Extrapolation algorithms were based on a signal periodicity and samples repetition supposed to fill in reliably only a stationary signal [17], [19], [39].

The extrapolation from both sides of the gap is more advantageous since there is less restriction on the signal stationarity. Filling in the signal based on knowledge of the samples from both sides of the gap is called an *interpolation*.

One of simpler methods of the interpolation is to repeat the most recent M samples

$$\{x[l-M], x[l-M+1], \ldots, x[l-1]\}, \tag{1}$$

where x is a signal vector and l is an index of the first missing sample. Having an estimate from both sides of the gap, the resulting signal could be enhanced by the linear weighted combination of both periodicity-based substitutions.

The signal interpolation based on modeling of an autoregressive process improves the results of interpolation in contrast with samples repetition [21]. Recovering of the missing samples is performed by minimizing the sum of squares of the residual errors which result in the estimates of the autoregressive parameters. The interpolation based on autoregressive modeling from both sides of the gap is introduced in [17]. Recalculation of missing samples in both sides of the gap is utilized according to the linear prediction. Among several block-based methods for calculating the autoregressive parameters two of them were chosen: Yule-Walker and Burg method.

A more advantageous approach based on interpolation of the signal parameters (amplitude, frequency) by the autoregressive modeling. Pure time domain interpolation methods presented in previous section often fail if the length of the signal gap is longer than $10\,\mathrm{ms}$. Bringing into account a two-dimensional time-frequency signal structure promises more space to the process of restoration of degraded recordings [25]. Sinusoidal model [29] of the signal interprets the audio signal as a sum of harmonic and non-harmonic components usually called the *partials*. These components are described by an amplitude, frequency and phase in time. Using these parameters the resulting signal is obtained as

$$y(t) = \sum_{p=1}^{P} A_p(t) \cos(\phi_p(t)), \tag{2}$$

where P is the number of partials, A_p is an instantaneous amplitude and ϕ_p is an instantaneous phase of the pth partial. This group of three parameters (f_p, A_p, ϕ_p) of the additive model represents particular samples of partials [24].

The proposed method in [24] models only the tonal part of the signal while the noisy part is not considered at all. An extension was brought in [25] adding the interpolation of the noisy residual signal and minor harmonics. Since most of real world signals contain noise with significant amount of energy, the residuum should not be avoided.

Presented state-of-the-art methods are supposed to be comparative approach to the novel methods based on underdetermined systems of linear equations which will be described in the next chapter. Regarding the results in papers utilized as the references in this chapter, the sinusoidal modeling should be the most promising state-of-the-art method for audio signal interpolation.

3 Sparse representations

3.1 Frames

The representation of the signal $\mathbf{x}$ is not unique considering the number of generators of a vector space $\mathbb{V}$ is greater than the dimension n of the space. The signal vector can be represented using various linear combinations. This property is called the *underdetermination* and the group of linearly dependent basis vectors is called the *frame*. Frames are generally less constrained than the basis and are widely utilized because of their flexibility. According to a mathematical definition a frame is formed by a countable set of vectors $\{\mathbf{\Phi}_k\}_{k\in J}$ in a vector space $\mathbb{V}$ if there are two positive constants $0 < A \leq B < \infty$ such that

$$A\|\mathbf{x}\|^2 \leq \sum_{k\in J} |\left\langle \mathbf{x}, \mathbf{\Phi}_k \right\rangle|^2 \leq B\|\mathbf{x}\|^2,\ \forall\mathbf{x}\in\mathbb{V}, \tag{3}$$

where constants A, B are representing *the frame bounds*. Each element (column) of a frame $\mathbf{\Phi}_k$ is called the *atom* [32], [10].

The *frame operator* $\mathbf{S}$ is defined as

$$\mathbf{S}\mathbf{x} = \sum_k \langle \mathbf{x}, \mathbf{\Phi}_k \rangle \mathbf{\Phi}_k. \tag{4}$$

The upper bound A is equal to the smallest eigenvalue of a frame operator. Likewise, lower frame bound B is equal to the biggest eigenvalue. If $A = B$ the frame is called *tight*. Moreover, if $A = B = 1$ the frame is called *normalized* or *Parseval tight*.

Dual frame is an important element in searching for coordinates c_k for a vector representation as $\mathbf{x} = \sum_k c_k\mathbf{e}_k$ using frame $\mathbf{E} = \{\mathbf{e}_1, \dots, \mathbf{e}_m\}$ in vector space $\mathbb{V}$ of dimension $n < m$. Frame $\mathbf{F} = \{\mathbf{f}_1, \dots, \mathbf{f}_m\}$ is called the dual frame of $\mathbf{E}$ if each $\mathrm{x} \in \mathbb{V}$ is defined as

$$\mathbf{x} = \sum_k \langle \mathbf{x}, \mathbf{f}_k \rangle\, \mathbf{e}_k = \sum_k \langle \mathbf{x}, \mathbf{e}_k \rangle\, \mathbf{f}_k. \tag{5}$$

Searching for coordinates c_k of vector $\mathbf{x}$ in a primary frame is performed using the dual frame by equation $c_k = \langle \mathbf{x}, \mathbf{f}_k \rangle$. Matrix representation of the same problem is $\mathbf{c} = \mathbf{F}^*\mathbf{x}$. This operation is called the *analysis*. A backwards reconstruction of the signal by superposition of atoms is called the *synthesis* [32].

Each primary frame has got an infinite number of dual frames in general. However, usually the task is to find a *canonical* dual frame. This frame results in a set of coefficients $\{c_k\}_{k\in J}$ with minimal energy (norm). Such frame is defined as

$$\mathbf{f}_k = \mathbf{S}^{-1}\mathbf{e}_k \tag{6}$$

If matrix $\mathbf{E}$ has a full row rank, using the Moore-Penrose pseudoinverse a canonical dual frame can be obtained by

$$\mathbf{E}^{+} = \mathbf{E}^{*}(\mathbf{E}\mathbf{E}^{*})^{-1}. \tag{7}$$

A special family of frames are the *Gabor frames* [32], [3]. The construction is based on the translation and modulation operators. The signal $\mathbf{x}$ is represented as a superposition if translated and modulated version of the basic function $g \in L^2(\mathbb{R})$ which are generated as

$$g_{\tau,\omega}(t) = g(t-\tau)\mathrm{e}^{2\pi \mathrm{i} t\omega}. \tag{8}$$

3.2 Sparse representations

During last decade, the attention of researchers in the field of signal processing increasingly focused on mathematical methods searching so-called *sparse representations*. Mathematical fields such as linear algebra, functional analysis, convex optimization or statistics provide a basis for finding sparse solutions of systems of linear equations which is usable in various fields. This thesis focuses on the use of sparse representations in the field of signal theory and systems, which were one of the first areas of sparse solutions applications.

The basic task is to find a solution of a linear system

$$\min_{\mathbf{x}} \|\mathbf{x}\|_0 \qquad \text{subject to} \qquad \mathbf{y} = \mathbf{D}\mathbf{x}, \tag{9}$$

where $\mathbf{D}$ is a transformation matrix called the *dictionary*, $\mathbf{x}$ is a coefficient vector, which we are looking for and $\mathbf{y}$ is a vector of input signal samples. Solution is called sparse if the resultant vector of the coefficients $\mathbf{x}$ contains only a few non-zero elements in comparison with the size of the system of linear equations. Because of uncertainty of the solution there is a possibility of adaptivity of the solution, which is amongst others the most appropriate for the particular purpose. Graphical demonstration of the problem is in Fig. 1

Sparse solutions offer adaptation of the dictionary for a particular purpose which benefits in terms of information compression, analysis, interpretation and numerical stability. Their advantage is the ability to represent the signal with a few important coefficients.

Because the selected dictionaries are overdetermined (the matrix contains more columns than rows), searching for such solutions can be computatively intensive, depending on the choice of algorithm for signal decomposition. Moreover, improperly assembled dictionaries can lead to system instability.

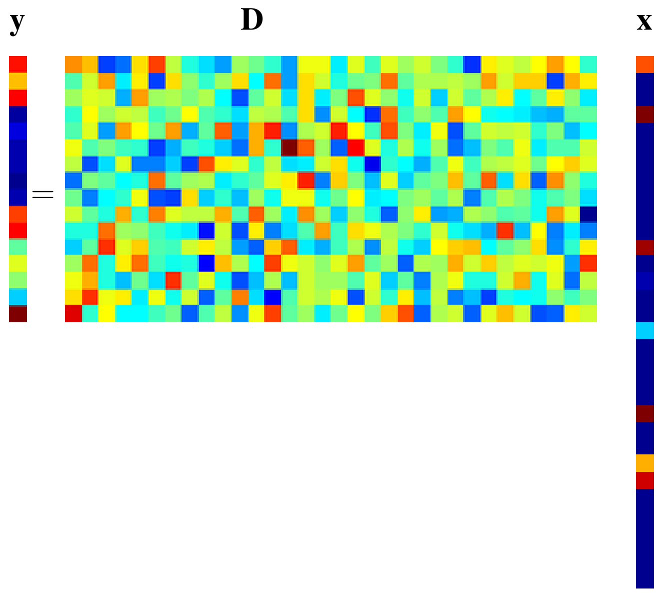

Figure 1: Graphical illustration of the sparse representation

3.3 Audio Inpainting

Sparse signal representations for inpainting problems were first used in image signal processing [16] and a few years later the Audio Inpainting algorithm was introduced in [1].

The Audio Inpainting is based on the approximation of missing or distorted information with a linear combination of atoms $\{\mathbf{d}_j\}$ from the dictionary. Input signal samples are classified into two parts: reliable samples with support vector $\mathcal{I}^{\mathrm{r}}$ and distorted samples with support vector $\mathcal{I}^{\mathrm{m}}$. Longer signals (typically audio) can be segmented into portions of defined length equal to the atom length N and defined overlap, while information about signal support is preserved. Segments are formed in a matrix $\mathbf{Y}$ as well as their support vectors form the measurement matrix $\mathbf{M}^{\mathrm{r}}$ with reliable samples support and $\mathbf{M}^{\mathrm{m}}$ with missing signals support. For those segments where wrong samples are detected an individual inpainting process is performed. Wrong samples are identified in a measurement matrix $\mathbf{M}^{\mathrm{m}} \in \{0, 1\}$ as ones as well as reliable samples in matrix $\mathbf{M}^{\mathrm{r}} \in \{0, 1\}$ with values complementing the $\mathbf{M}^{\mathrm{m}}$ matrix. Using the measurement matrix reliable data are obtained as

$$\mathbf{y}^{\mathrm{r}} = \mathbf{M}^{\mathrm{r}}\mathbf{y} \tag{10}$$

and the main goal is to reconstruct the non-reliable samples $\mathbf{y}(\mathcal{I}^{\mathrm{m}})$, where $\mathcal{I}^{\mathrm{m}} = \{1, 2, \ldots, L\}$ is a vector of missing signal support from the whole input signal of length L. Regarding sparse representations of signals, each segment of reliable samples can be obtained by the input signal approximation through dictionary atoms and appropriate coefficients

$$\mathbf{y}_i^r = \mathbf{M}_i^{\mathrm{r}}\mathbf{D}\mathbf{x}_i. \tag{11}$$

Coefficients $\mathbf{x}_i$ are computed from reliable samples using any greedy or relaxation algorithm. The process of analysis is performed by the dictionary $\mathbf{D}$ with atom length restricted to the length of the reliable samples of input signal. Recovering unknown samples $\widehat{\mathbf{y}}(\mathcal{I}^{\mathrm{m}})$ can be performed by estimating as $\widehat{\mathbf{x}}_i$ a sparse vector of each segment

$$\widehat{\mathbf{y}}_i(\mathcal{I}_i^{\mathrm{m}}) = \mathbf{M}_i^{\mathrm{m}}\mathbf{D}\widehat{\mathbf{x}}_i. \tag{12}$$

For reconstruction of missing samples with obtained coefficients $\mathbf{x}_i$ we use the original dictionary $\mathbf{D}$ with original atoms length. Only samples at missing positions are replaced by the reconstructed samples

$$\widehat{\mathbf{y}} = \mathbf{y}(\mathcal{I}^{\mathrm{r}}) + \widehat{\mathbf{y}}(\mathcal{I}^{\mathrm{m}}). \tag{13}$$

3.4 Dictionaries

The most frequent dictionaries are trivial dictionaries (Heaviside disctionary), frequency dictionaries (Fourier, cosine dictionary), time-scale dictionaries (wavelet dictionary) and time-frequency dictionaries [9].

Time-frequency analysis is a projection of signal $\mathbf{y}$ onto the atoms $\mathbf{g}_{\tau,\omega}$. In fact, we are dealing with a redundant STFT known as the *Gabor analysis* [18]. The distribution of sampling points in the time-frequency plane and the corresponding frame construction is defined as

$$\mathbf{G}(g, a, b) = \mathbf{M}_{bm}\mathbf{T}_{an}g, \tag{14}$$

where $\mathbf{M}_{bm}$ is a modulation operator, $\mathbf{T}_{an}$ is a translation operator and $a, b, m, n \in \mathbb{Z}$ are sampling parameters of the STFT. The basic question is how to choose parameters a, b and a function $g \in L^2(\mathbb{R})$ to form a frame in space $L^2(\mathbb{R})$. Limits of the time-frequency resolution are described by the *Heisenberg's principle of uncertainty* saying that there is no signal well concentrated in both time and frequency [14]. Function g is called the *window function*. In theory the Gaussian function $g(x) = \exp(-x^2/2)$ is the only function with optimal time-frequency concentration according to the Heisenberg's principle [32]. In praxis, the disadvantage is its support of infinite length. Therefore, more feasible window functions are utilized such as Hamming's, Hann's, Nuttall's etc. An example of modulated and translated Hann's window is in Fig. 2. If the Gabor system forms a frame, then the reconstruction of the signal from time-frequency coefficients could be performed.

One possible way how to represent the signal more sparsely is to adapt the dictionary to a specific signal. This process is called the *dictionary learning*. The process of adaptation of the dictionary to the specific signal comprises two main steps: the first step is to isolate a group of training samples from the reliable part of the signal. This means that we can not train dictionary from noisy or missing sections of the signal, which will later be reconstructed. The range and number of samples, from which the dictionary will be *learned*, specifies the user when setting parameters for the selected algorithm. In the second step the adaptation itself will be processed on the dictionary data gathered in the first step.

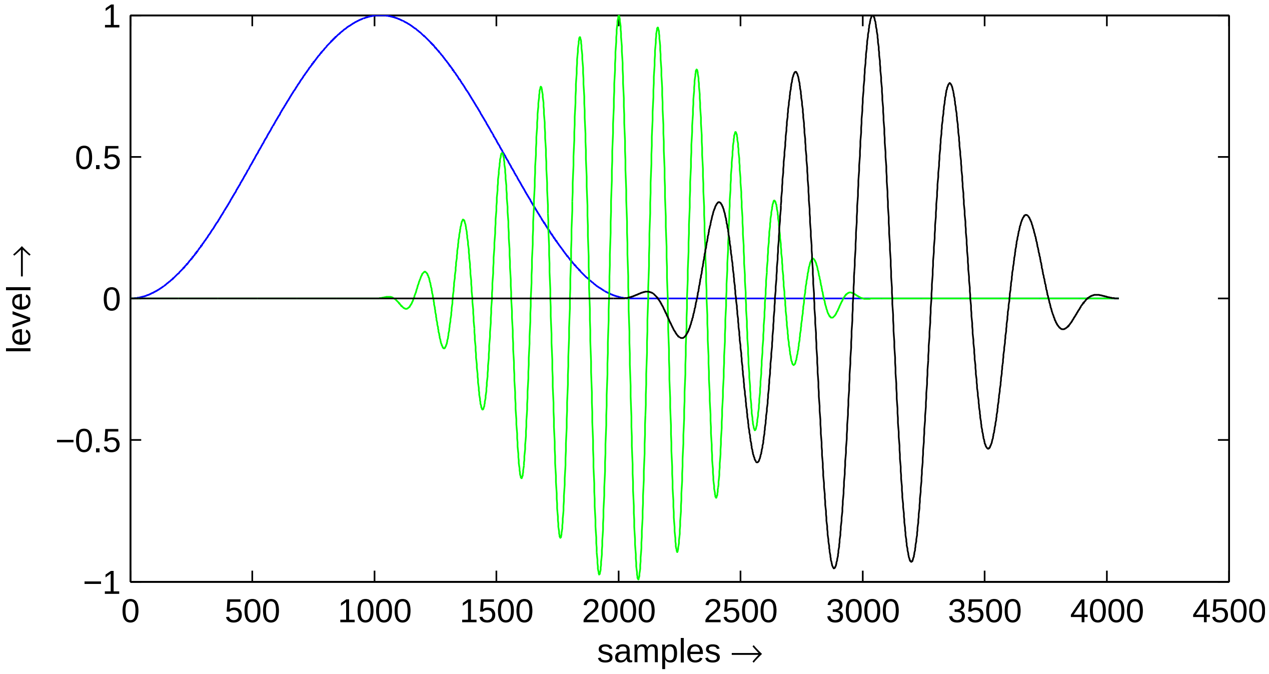

Figure 2: Hann window function (blue), modulated, translated and windowed cosine functions (green, black).

Preparing such adapted dictionary brings a more computationally demanding task, which certainly prolong the total time required for data reconstruction. However, output may result in sparser representation and thus more accurate reconstruction of missing data [4]. K-SVD is a non-convex dictionary learning algorithm The goal of the algorithm is to adapt the dictionary $\mathbf{D}$ in order to achieve a higher degree of sparsity in the representation of input signal $\mathbf{y}_k$, using any algorithm which approximates the optimization problem.

4 Algorithms for sparse approximations

Traditional methods of synthesis use transformation matrix Φ which is orthogonal (i.e. that the dimensions are $n \times n$) and forms the basis. The analysis coefficients are obtained by simply inverting the operator Φ. However, since the tasks that we deal with use overdetermined dictionaries that produce infinitely many solutions it is necessary to apply the algorithms that find optimal solutions in some way. Searching for sparse solutions is a problem of ℓ_0 norm minimization. Unfortunately, this norm is not a convex function and it is not possible to use any currently existing algorithm solving convex optimization. It constitutes an NP-hard problem and the exact solution can not be found in polynomial time [8].

4.1 Greedy algorithms

Greedy algorithms are selecting one (or more) of the most important atoms in each iteration. An important feature is that once the coefficient of the respective atom is found it does not change during the calculation of other coefficients nevermore. Algorithms like Matching Pursuit [28] or the currently popular modification called Orthogonal Matching Pursuit (OMP) [31] have relatively low complexity, unfortunately, achieving the global optimum is not guaranteed.

OMP algorithm performs the following steps in each iteration: first calculates the correlation of all atoms of the dictionary with the current input signal segment. The highest coefficient which is found in this step is saved, backward synthesis of the coefficients and the dictionary is performed and the signal is then compared with the original input signal. The difference between these two signals is stored as a residue. The reconstruction error for the residual is computed as the square of ℓ_2 norm and every following iteration correlates atoms (except those already used) with the current residuum. Therefore, new coefficients are being obtained repeatedly. The reconstruction error decreases with increasing number of coefficients.

OMP algorithm iterates over and over again until one or more stopping criterion is fulfilled [15].

4.2 Relaxation algorithms

The closest convex norm that can be utilized for approximation of sparse solutions is ℓ_1. In most cases results of minimization of the coefficients of ℓ_0 and ℓ_1 norm coincide. This group of algorithms assumes that under certain conditions we get to accurate or at least approximate solution. The algorithms are based on ℓ_1 relaxation.

Basis Pursuit is an optimization problem which decomposes the signal into a superposition of atoms in an optimal way. Optimality is reached by having the smallest ℓ_1 norm of coefficients

$$\min_{\mathbf{c}} \|\mathbf{c}\|_1 \quad \text{subject to} \quad \mathbf{Dc} = \mathbf{y}, \tag{15}$$

among all considered decompositions where $\mathbf{c}$ is the coefficient vector, $\mathbf{D}$ is the dictionary and $\mathbf{y}$ is the resulting signal.

Proximal algorithms are methods from the optimization theory including relaxation tasks with ℓ_1 norm. Proximal algorithms are splitting the problem of sparse regression into separate problems as

$$\operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^N} \left(f_1(\mathbf{x}) + f_2(\mathbf{x})\right), \tag{16}$$

where $\mathbf{x}$ is the input (observed) signal, which are solved iteratively whereas the conditions of convergence of the algorithm are known. These algorithms are not very fast, however, the flexibility is advantageous [12].

The constrained form of the optimization problem (not corresponding to the task 16) is

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1 \qquad \text{subject to} \qquad \|\mathbf{D}\mathbf{x} - \mathbf{y}\|_2 \leq \delta, \tag{17}$$

where δ is an allowed error from the true solution.

The unconstrained task which is going to be solved is called the LASSO.

$$\widehat{\mathbf{y}} = \mathbf{D}\text{argmin}_{\mathbf{x}\in\mathbb{R}^N} \frac{1}{2} \|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_1 \tag{18}$$

where $\lambda \|\mathbf{x}\|_1$ is a regularization term which penalizes certain types of solutions and λ is a weighting coefficient controlling strength of the term. The higher the λ the more penalized the non-sparse solutions are [38].

In an overcomplete case the solution is different for analysis and synthesis model. Analysis model is referred as a co-sparse analysis [30]. The point is to enforce sparsity of the analysis coefficients $\mathbf{A}\mathbf{y}$ instead of synthesis coefficients $\mathbf{c}$. The definition of solving and audio inpainting problem in an analysis sense is

$$\widehat{\mathbf{y}} = \text{argmin}_{\mathbf{y}\in\mathbb{R}^N} \|\mathbf{M}^{\text{r}}\mathbf{x} - \mathbf{M}^{\text{r}}\mathbf{y}\|_2^2 + \lambda \|\mathbf{A}\mathbf{y}\|_1 \tag{19}$$

where $\mathbf{x}$ is the observed signal, $\mathbf{y}$ is the unknown signal and $\mathbf{A}$ is an analysis operator.

The definition of solving and audio inpainting problem in an synthesis sense is

$$\widehat{\mathbf{y}}^{\text{m}} = \mathbf{M}^{\text{m}}\mathbf{D} \cdot \text{argmin}_{\mathbf{c}\in\mathbb{R}^N} \left(\frac{1}{2} \|\mathbf{M}^{\text{r}}\mathbf{D}\mathbf{c} - \mathbf{M}^{\text{r}}\mathbf{x}\|_2^2 + \lambda \|\mathbf{c}\|_1 \right). \tag{20}$$

If both of the functions f_1 and f_2 from Eq. 16 are non-smooth, we are not able to compute the gradient (first derivative) and the problem is solved by *Douglas-Rachford* algorithm [11]. On the other hand, if at least one of the functions f_1, f_2 is smooth (eg. ℓ_2 norm), the gradient for this function is defined and we can utilize the $Forward - Backward$ algorithm to solve the sparse regression problem.

Minimization of the function is performed using the *proximity operator* which minimizes the function without getting too far from the initialization point. The proximity operator is a generalization of the projection [13].

Some of methods known from other areas of signal processing which can be formulated as proximal are (F)ISTA (Fast Iterative Schrinkage/Thresholding Algorithm) [6] or ADMM [7]. The synthesis model for solving an Audio Inpainting problem defined in 20 is solved by proximal splitting.

Signal atoms which are affected by the signal gap do not fulfill the condition of $\|d_j\|_2 = 1$. The computation of weights was formulated as

$$\mathbf{w}_i = \frac{\|\mathbf{g}\|_2}{\sqrt{\|\mathbf{g}\|_2^2 - |\mathbf{m}^{\text{m}} * \mathbf{g}^2|_i}}, \tag{21}$$

where $\mathbf{m}^{\text{m}}$ is a missing mask of a corresponding atom of values $\{0, 1\}$.

4.3 Structured sparsity

Previous methods of ℓ_1-relaxation treated the coefficients independently regarding no correspondence with the neighbourhood. Nevertheless, every typical spectrogram of a musical signal is naturally structured. Considering this fact, the algorithm for sparse signal modelling incorporating information about a structure in an analysis stage of processing would be an advantage compared to the regular sparse modelling where coefficients are treated independently [23]. Keeping or discarding particular coefficient under consideration is decided up to certain neighbourhood of the coefficient.

The convex optimization problem for Audio Inpainting with mixed norms is reformulated as

$$\widehat{\mathbf{y}}^{\mathrm{m}} = \mathbf{M}^{\mathrm{m}}\mathbf{D} \cdot \operatorname{argmin}_{\mathbf{c}\in\mathbb{C}^N} \left(\frac{1}{2} \|\mathbf{M}^{\mathrm{r}}\mathbf{y} - \mathbf{M}^{\mathrm{r}}\mathbf{D}\mathbf{c}\|_2^2 + \lambda \|\mathbf{c}\|_{p,q} \right) \tag{22}$$

where p represents a within-group penalty and q is an across-group penalty.

Due to the non-stationarity of sound signals, windowing with overlapping and weighting is incorporated. Generally, there are two set of mixed norms which are widely used in literature Windowed-Group-Lasso and Persistent-Elitist-Lasso.

5 Experiments and evaluation of audio restoration

Experiments are evaluated using objective evaluation methods SNR and PEMO-Q. The SNR is defined as

$$\mathrm{SNR}(\mathbf{y}, \widehat{\mathbf{y}}) = 10 \log \frac{\|\mathbf{y}(\mathbf{I}^{\mathrm{m}})\|_2^2}{\|\mathbf{y}(\mathbf{I}^{\mathrm{m}}) - \widehat{\mathbf{y}}(\mathbf{I}^{\mathrm{m}})\|_2^2}, \tag{23}$$

where $\mathbf{y}$ is the original signal, $\widehat{\mathbf{y}}$ is the reconstructed signal and $\mathbf{I}^{\mathrm{m}}$ is the vector of indices of the corrupted signal. PEMO-Q evaluates the audio quality of a given distorted signal in a relation to the corresponding high-quality reference signal whereas the auditory model is employed to compute so-called internal representations adjusted to the cognitive aspects [20].

All of the experiments have got following common input values: input signal samples, index of the first missing sample and length of the signal gap. Each approach to audio interpolation uses individual method to trim the input signal according the the transformation/modeling parameters.

An inpainting toolbox (Brno-Wien Inpainting Toolbox) including all of the presented methods was developed and all the following experiments were performed using this MATLAB toolbox. Inner computations of inpainting algorithms in the toolbox are dependent on some other toolboxes. All of the files are single channel (mono) with the sampling frequency of 16 kHz and 16 bit depth. Example files consist of harmonic and non-harmonic musical samples and speech signals.

5.1 State-of-the-Art interpolation methods

Regarding samples repetititon method of interpolation, the best SNR $= 1.28\,\mathrm{dB}$ was reached for $q_\mathrm{u} = 260$ samples. q_u specifies the area surrounding the gap supposed to be the model for signal estimation. The subjective listening of the result is not satisfying and the listener can clearly recognise the gap position in the sound. Regarding the speed of computations, this method is one of the fastest with the time duration of less than a second. The speed is definitely shorter than the gap length, therefore the algorithm is feasible also for real-time applications.

Utilizing the AR modeling of signal samples (Least-Squares Residual Predictor method) on an example gap, the best SNR $= 0.67\,\mathrm{dB}$ is reached with the maximum AR model order. Second method, the Weighted Forward-Backward Predictor resulted in the maximum SNR $= 0.84\,\mathrm{dB}$ was at AR model order of 270.

Results of batch testing of the first method, the Least-Squares Residual Predictor was performed for gap size of $\{10, 20,\ldots, 100\}$ ms which corresponds to {160,320, ...,1600} samples. The AR model order was selected from the range of $\{0.5, 1, 2, \ldots, 5, 6, 8, \ldots, 12\}$ times of the gap size. Every combination of gap size and AR model order was used ten times for different gap positions in music file *music11_16-kHz.wav* and the resulting SNR was computed as an average of these experiments. A lot of experiments were not performed at all because of the very long processing time ($> 1000\,\mathrm{s}$ for single interpolation experiment). The best SNR $= 5.78\,\mathrm{dB}$ was reached for gap size of 360 samples and AR model order of 10 times the gap length. However, the variance of $14\,\mathrm{dB}$ makes the results very unstable. Therefore, there is no general recommendation for interpolation parameters selection.

Furthermore, the same experiment was made with the *Weighted Forward-Backward Predictor*. Since the computational load of this method is higher than the previous method, batch experiment was performed with much more restricted range of parameters. The AR model order was selected from the range of $\{0.5, 1, 2, 3, 4\}$ times the gap size. The best result (SNR $= 5.06\,\mathrm{dB}$) was reached for gap size of 160 samples and model order of 400 samples. Compared to LSRI method, the variance of the results was much lower ($4.65\,\mathrm{dB}$).

Previous methods needed only one parameter to be set up: the model order or neighbourhood size. The sinusoidal modeling uses four parameters that influence the power of the algorithm:

- Frequency difference threshold,
- amplitude difference threshold,
- length of the vector for amplitude mean value computation,
- order of AR model.

The frequency threshold in the following experiments will be set to value of 3 while this value should keep SNR higher and bring a little variability in the matching decision. Further experiments indicate the best amplitude difference threshold thrA $= 0.5$ Batch testing for the optimal settings of length of the vector for amplitude mean value

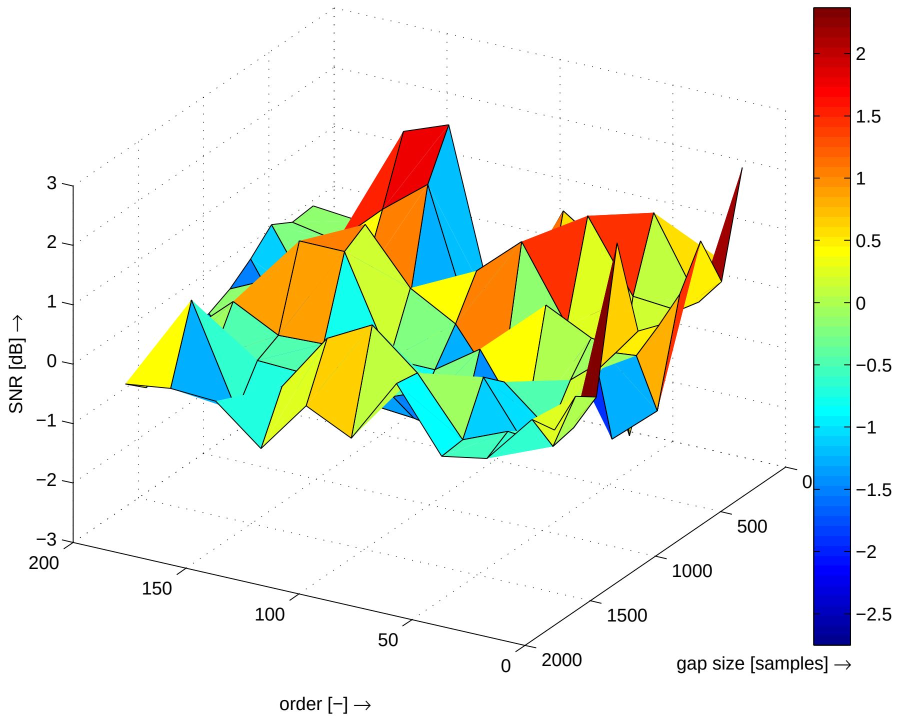

Figure 3: SNR of audio interpolation by sinusoidal modeling with various gap length and AR model order.

computation (M) show that there are very little differences between the mean amplitude range of $\langle 10; 100\rangle$. Therefore, following experiments will be performed with $M = 60$.

The most important parameter is the model order. It is obvious that there is no conspicuous evolution of the SNR with respect to gap length in Fig. 3. Therefore, any general recommendations for the selection of the model order according to the gap length can not be defined. The interpolation algorithm has to be tuned for each gap size and position individually to reach the best results. Regarding the speed of computations, the time consumption of the interpolation algorithm increases with the increasing model order value.

5.2 Greedy algorithm

The core algorithm for solving the sparse approximation by greedy methods is the Orthogonal Matching Pursuit (OMP). The size of the segment (neighbourhood size) S is computed from the gap length M and neighbourhood multiplier N such that $S = N \times M$. There are several parameters of the OMP algorithm which influence the result of audio inpainting:

- no. of coefficients obtained in each iteration
- maximum number of iterations,
- maximum error,
- dictionary redundancy,

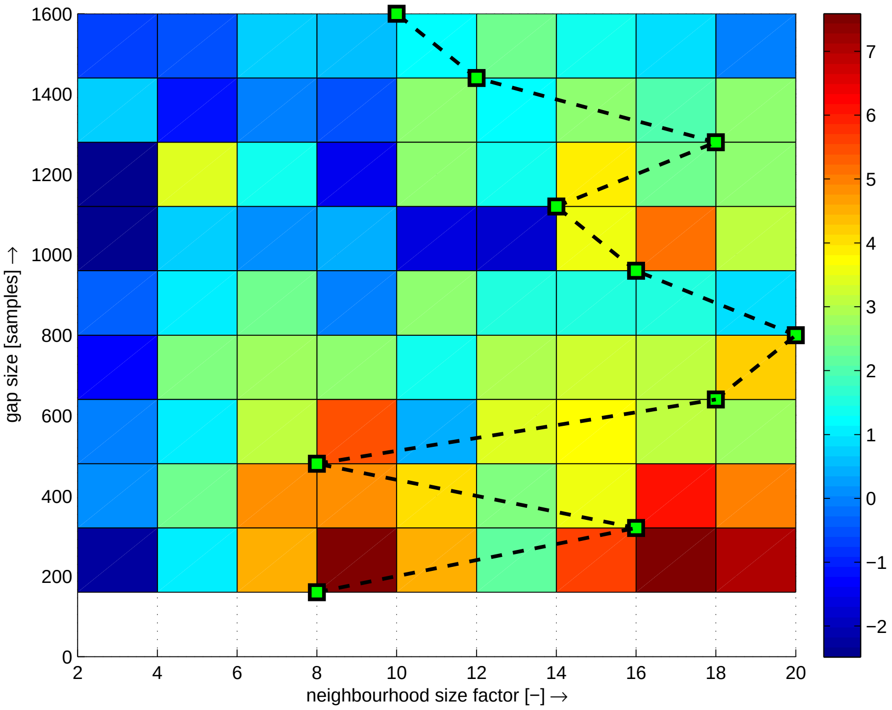

Figure 4: SNR of audio inpainting by OMP with various neighbourhood size according to gap length with highlighted best results by green points.

- neighbourhood size for inpainting.

Analyzing the number of coefficients, better SNR was reached obtaining only 1 coefficient in a single iteration. In contrast, the processing time increases enormously when 5 coefficients are obtained simultaneously. Therefore, for the following experiments only 1 coefficient per an iteration will be obtained.

In another experiment the redundancy factor *red* is set up in the range of $red = \{1, 2, \ldots, 10\}$. Redundancy of 1 and 2 produce poor results in terms of SNR, sometimes with SNR below zero. Higher values of redundancy factor produce better results with the best reached value of SNR $= 15,64$ dB for red $= 9$ and neighbourhood size of 1280 samples. However, the processing time of higher values of the redundancy factor is not efficient either for experimental or practical purpose. Therefore, for the following experiments redundancy factor of 3 will be used since it provides a good trade-off between the interpolation results and the processing time.

The most extensive batch testing was performed on music file *music11_16kHz.wav*. The objective of the experiment was to find an optimal neighbourhood size according to the various gap length. The size of the gap was from the range of $\{10, 20, \ldots, 100\}$ ms which corresponds to $\{160, 320, \ldots, 1600\}$ samples. Neighbourhood size factor is a multiplier of the gap size resulting in the full neighbourhood size and was selected from the range of $\{2, 4, \ldots, 20\}$.

The best results were reached for the shortest gap size. The highest average SNR $=$ 7.59 dB was obtained for gap size of 10 ms and neighbourhood size factor of 8 which results in the neighbourhood size of 1280 samples (80 ms). Larger neighbourhood size produces better inpainting results in the terms of SNR. However, the processing time increases dramatically with increasing neighbourhood and gap size.

As a conclusion of this evaluation, the optimal trade-off between the speed of computation and the resulting inpainting performance using OMP algorithm should be the neighbourhood size factor of at least 4 (see Fig. 4).

5.3 Dictionary learning

Among several dictionary learning algorithms the K-SVD method implemented within the SMALLbox[1] was chosen for experiments. Several parameters were examined during the tests. Setting the number of iterations to 4 reaches a satisfying value of RMSE (Root Mean Squeare Error) and after 10 iterations the RMSE is stabilized at its minimum. Other experiments were focused on minimizing RMSE according to space between segments obtained from reliable samples to get the training data.

Using the audio file of length $80\,000$ samples, the segment length of 256 samples and redundancy factor of 3, the dictionary **D** has got a size of 256×768 samples. With these parameters the shift of segments could be set to a value from interval $\{1, 2, \ldots, 100\}$. Increasing the segment shift value results in smaller RMSE during the dictionary learning process.

Another parameter of the dictionary learning explored further was the maximum number of non-zero coefficients $S_{\max}$. For $S_{\max} \in \{1, 2, 3, 4, 5\}$ dictionary learning experiments were performed with focus on the lowest RMSE depending on different $S_{\max}$ and therefore reaching the minimal error. After six iterations the minimal RMSE was reached by $S_{\max} = 3$ and remains minimal with very little change.

The number of iterations has to be set up deliberately. The number can be small and RMSE will remain high (the dictionary is not adapted as much as it can be) or the number can be too high and after reaching the minimum RMSE the algorithm can waste the time with new iterations or worse the RMSE can raise up. As a result, satisfying RMSE can be obtained with three or four iterations.

Figure 5 shows that for gap length of 40 to 110 samples dictionary trained by the K-SVD algorithm overcomes static dictionaries by about $10\,\mathrm{dB}$. Results of reconstruction by static dictionaries (Gabor and DCT) produce almost the same results. Presented results were published in [27].

5.4 l_1–relaxation algorithm

Most of the contribution was focused on the implementation of the analysis and the synthesis model for their comparison. Since all of the papers dealing with the audio inpainting use the synthesis model ([1],[22],[5]) there was a natural interest in an implementation of the analysis model and its evaluation.

[1] http://small-project.eu/software-data/smallbox/

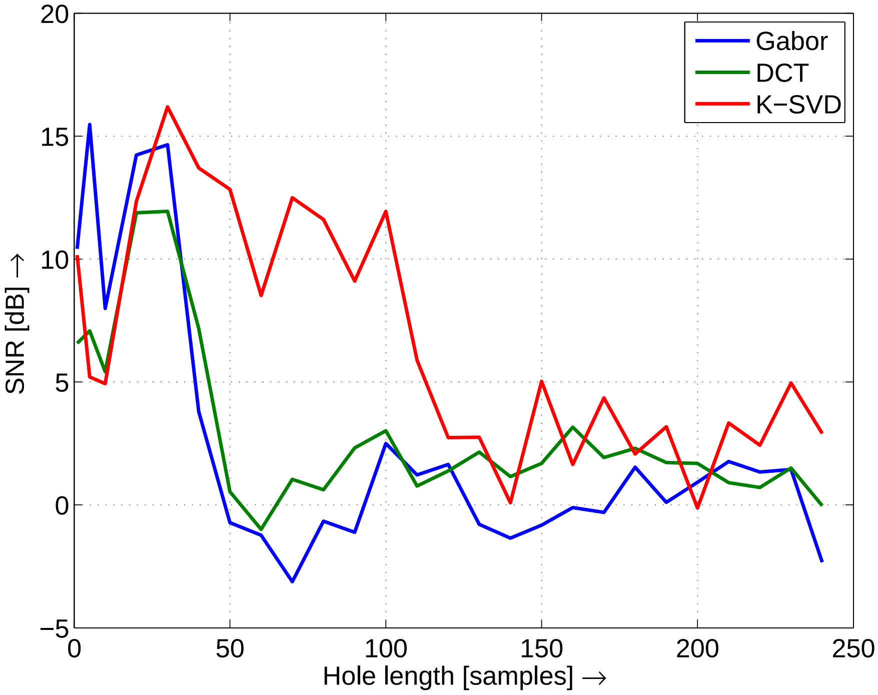

Figure 5: Signal reconstruction of a guitar audio sample using the K-SVD algorithm and greedy solver.

The time-frequency transform is built up from time and frequency localized windows and their translations [34]. There were two varying parameters during these tests: redundancy value from the ranges of $\{2, 3, \ldots, 8\}$ and the length of the window of $\{513, 1026, 2052, 3078, 4104, 5130, 6156, 7182, 8208, 9234, 10260, 11286\}$ samples. The evolution of SNR for an increasing window length seems to be more or less the same for all redundancy values. From this time the redundancy will be fixed to the value of 3 for all of the following experiments. The weights of the atoms are applied on resulting coefficients.

For each combination of gap size and window length there were 10 independent inpainting experiments with different gap position for each analysis and synthesis model. Very small (window length)/(gap size) ratio the window is not able to find enough of a reliable pattern for inpainting such a gap. Regarding the analysis model for harmonic signals, the best restoration in the terms of objective SNR evaluation was obtained in the case of short gaps and a size of window length from 2000 to 9000 samples (from 125 to 560 ms). Likewise in the synthesis model, the satisfying SNR for gap length of 1120 samples is reached using window length in the range from 5130 to 7182 samples.

As a conclusion, there are somehow optimal parameters for a particular gap size and gap position, nevertheless, they can be very easily missed by a slight modification of a window length.

Audio Inpainting for the reconstruction of non-harmonic signals failed in all com-

Table 1: Results of inpainting experiments over all of the sound files (analysis model, sorted by the average SNR)

Filename	Average SNR [dB]	Average STD [dB]	Character
music11_16kHz	8.1	4.4	harmonic, guitar
music03_16kHz	7.9	4.6	harmonic, guitar
music02_16kHz	5.0	3.8	harmonic, double bass
music04_16kHz	4.5	4.7	harmonic, woman singing
music10_16kHz	4.2	1.7	harmonic, orchestra
music08_16kHz	3.4	2.0	harmonic, pop music
music09_16kHz	1.7	1.7	speech, rap
music12_16kHz	1.1	1.5	speech, DJ show
music07_16kHz	0.1	0.6	non-harmonic, drums

Table 2: Results of inpainting experiments over all of the sound files (synthesis model, sorted by the average SNR)

Filename	Average SNR [dB]	Average STD [dB]	Character
music11_16kHz	7.3	4.3	harmonic, guitar
music03_16kHz	6.7	4.7	harmonic, guitar
music02_16kHz	4.4	4.0	harmonic, double bass
music04_16kHz	4.0	5.2	harmonic, woman singing
music10_16kHz	3.8	1.6	harmonic, orchestra
music08_16kHz	2.8	1.9	harmonic, pop music
music09_16kHz	1.4	1.7	speech, rap
music12_16kHz	0.8	1.6	speech, DJ show
music07_16kHz	−0.3	1.5	non-harmonic, drums

binations of parameters. Therefore, the Audio Inpainting of such non-harmonic signal should be performed by some other technique.

Another notable remark is that the order of the analysis and synthesis results is the same, however, results of the analysis model are reaching better SNR values then the synthesis. Further, the standard deviation of the synthesis model is the same or higher in all of the experiments compared to the analysis model standard deviation. Results of average SNR for various kinds of audio files are in Tab. 1 and 2

Considering the speed of the computation, there is quite a big difference between the analysis and the synthesis model especially using large window sizes. The processing time of the analysis model is significantly lower than the processing time of the synthesis model. The average processing time of the inpainting process using a proximity algorithm is not dependent on the gap size.

Results of inpainting of the example gap are in Tab. 3.

Table 3: Complex evaluation of inpainting/interpolation algorithms

Method name	SNR [dB]	PSM [-]	PSM$_t$ [-]
Samples repetition	-3.9338	0.9839	0.5118
AR samples modeling (LSRI)	3.9738	0.9958	0.8312
AR samples modeling (WFBI)	3.9526	0.9971	0.9108
AR sinusoidal modeling	0.7873	0.9953	0.7330
Greedy algorithm (OMP)	13.7471	0.9980	0.9202
ℓ_1-relaxation (analysis model)	25.5870	0.9999	0.9961
ℓ_1-relaxation (synthesis model)	23.6777	0.9998	0.9964
ℓ_1-relaxation (structured sparsity)	21.1735	0.9997	0.9873

5.5 Structured Sparsity

The core problem being solved is described in Eq. 22. The unconstrained version of the model (see Eq. 18) for solving the inpainting problem utilizes the penalty term λ which and has to be experimentally set to produce the smallest possible error. Using the example gap position and size like in all other experiments the first parameter under investigation was the empirical Wiener exponent. It is clear that the exponent set to 2 produces better results of inpainting. Using $\lambda = 0.012$ and expo $= 2$ the following batch experiment was performed.

From other experiments is is clear that the inpainting performance decreases with increasing gap length independently of the neighbourhood size. Center of the neighbourhood was always set to the middle sample. On the other hand, the processing time increases with increasing neighbourhood size. For better illustration of power of the algorithm the example gap inpainting is in Fig. 6 in time and time-frequency representation.

5.6 Audio Denoising

Another successful application of the structured sparsity is audio denoising [37] [35] [36]. As described in [26] very old recordings on wax cylinders contain periodical short-time distortions caused by fissures and mould of the material. Applying structured sparsity denoising on the digitized recordings of wax cylinders brought both successful elimination of short time distortions and reduction of the broadband noise. The preprocessing had to be performed on the original file to suppress the strongest interferences such as crackles or low frequency rumbling.

Signal restoration shown in Fig. 7 was performed with the structured sparsity framework. All the parameters were setup according to Tab. 4 and the value $\lambda = 0.02$ was chosen experimentally. The consequently created Gabor frame is a Parseval tight frame. Higher value of λ produces stronger denoising, lower value preserves a lot of short-time harmonics in the whole spectrum. The subjective evaluations of Structured Sparsity for Audio Denoising outperformed denoising using the professional software.

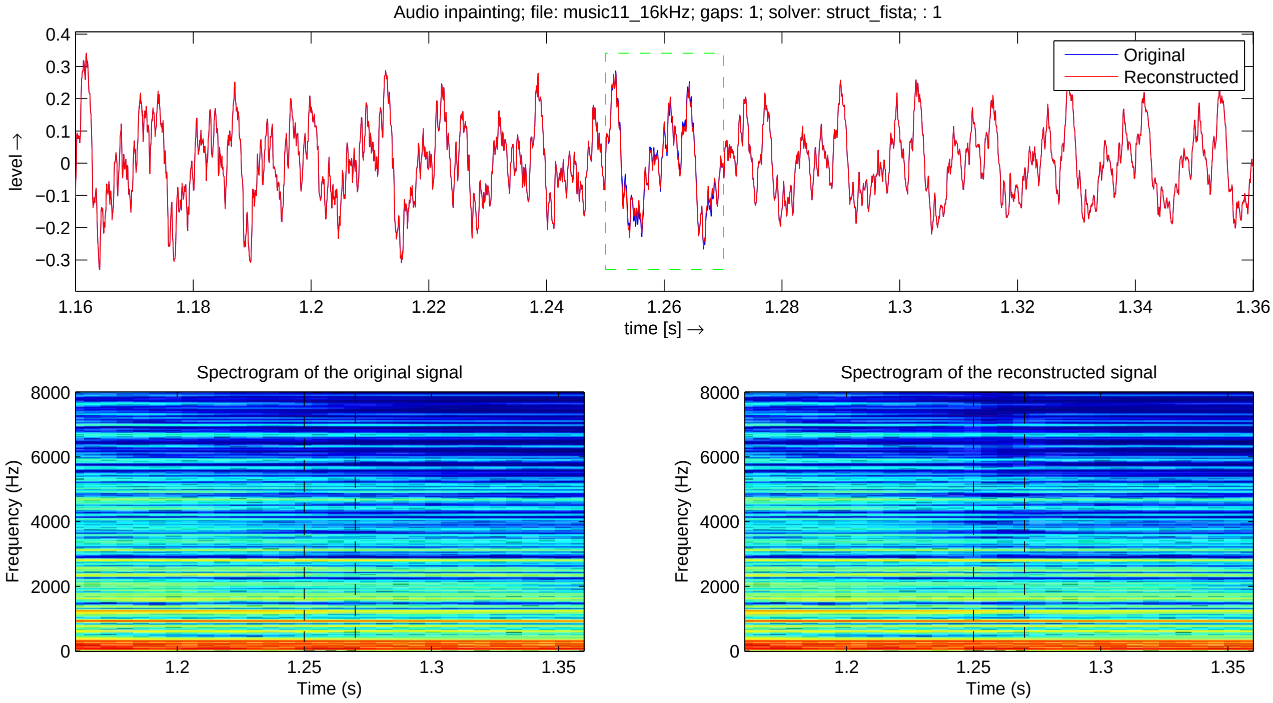

Figure 6: Time and spectral representation of inpainting using the structured sparsity.

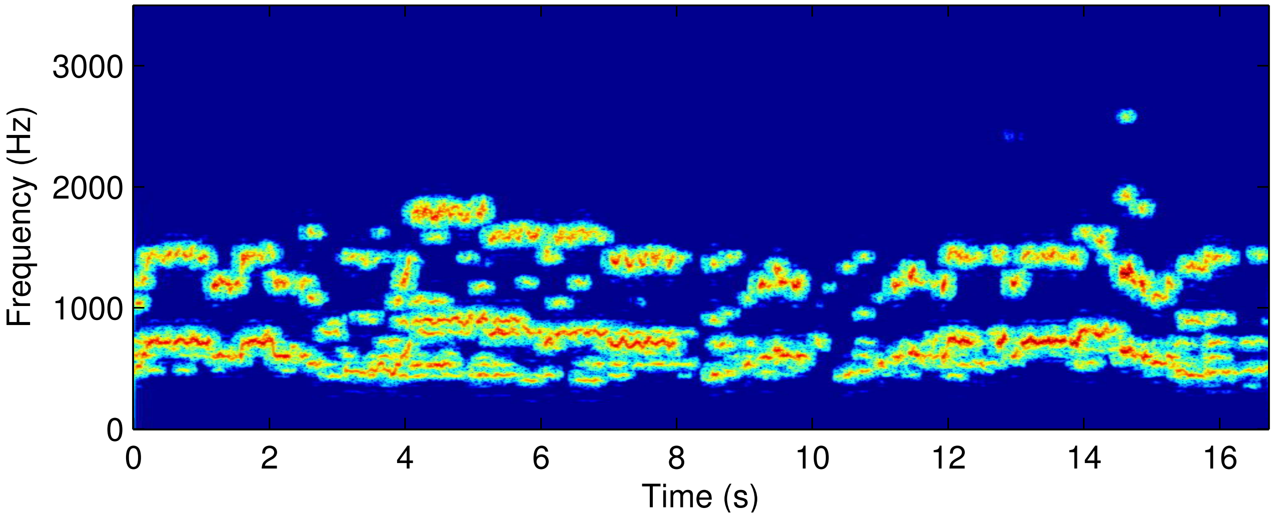

Figure 7: Reconstruction by the structured sparsity

Table 4: Structured sparsity parameters

Dictionary	Gabor
Window length	2304 samples
Window overlap	75 % of window length
Frequency channels	2304
Shrinkage type	WGL
λ	variable
Neighbourhood	5×50 coefs. $\approx 0.32\,\mathrm{s}$
Center sample	[3,25]
Exponent α	$2 \approx$ Wiener

6 Conclusion

This thesis dealt with the process of audio restoration, especially the interpolation of missing data segments. Short time interruptions like clicks, crackles or signal gaps are nowadays successfully treated by interpolation methods based on samples repetition or autoregressive modeling of either signal samples or signal parameters. Recently, sparse representations of signals brought novel approaches of signal analysis and synthesis and naturally penetrated into the field of audio processing. The process of signal interpolation using overcomplete dictionaries was termed the Inpainting and the main goal of this thesis was to explore these new techniques, find possible ways of improvement and compare them to the state-of-the-art methods.

In the beginning of this research, the presumptions were that the process of Audio Inpainting will be feasible especially for harmonic signals and the restoration process will be most efficient in shorter gap sizes. As will be described further, these prepositions were proven by experiments.

Regarding the simplest method for interpolation, the Weighted Repetitive Substitution, both objective evaluation and subjective listening to the results is not satisfying and the listener can clearly recognise the gap position in the sound. A naturally arising question is whether this kind of interpolation method is really useful because if the signal gap was filled with zero samples the result sounds more naturally compared to the interpolation result.

Interpolation using the AR modeling of signal samples brought improvement of the restoration in terms of SNR. The best average result of SNR $= 5.78\,$dB was reached for a gap of length 360 samples using an AR model of order 3600. However, the standard deviation of 14 dB makes the results very unstable. The computational time of larger model orders (larger than 10 times the signal gap) makes this method unusable in real experiments because reconstruction of a single gap takes more than 1000 seconds.

Exploring the most important parameters of sinusoidal modeling resulted in an optimal setting of frequency threshold thrF $= 3$, amplitude threshold thrA $= 0.5$, length of the vector for amplitude mean value computation $M = 60$. There are no general recommendations for selection of the core parameter, the model order, according to the gap length since the values did not show any regular evolution using various gap length and model order. Maximum average SNR did not exceed the value of 3 dB.

The Orthogonal Matching Pursuit (greedy algorithm) reached the best values of SNR $= 15.64\,$dB for redundancy of dictionary of 9. However, the processing time of such redundancy factor is quite long, therefore, the optimal redundancy factor is red $= 3$. Batch experiments resulted in the highest average SNR $= 7.59\,$dB obtained for gap size of $10\,$ms and neighbourhood size factor of 8 which results in the neighbourhood size of $1280\,$samples ($80\,$ms). The optimal trade-off between the speed of computation and the resulting inpainting performance using OMP algorithm should be the neighbourhood size factor of at least 4.

Main contribution of this thesis is the experimental verification of audio inpainting utilizing ℓ_1-relaxation algorithms. Considering single coefficient sparsity (without relation to its neighbourhood), both synthesis and analysis model were implemented. Note that until now there was no scientific contribution on analysis model implementation for Audio Inpainting.

Fixing the atom length of the dictionary, results of the analysis model reached slightly higher SNR values than the synthesis, both with best average SNR higher than 20 dB. Further, the standard deviation of the synthesis model is the same or higher in all of the experiments compared to the analysis model standard deviation. There are some optimal parameters for a particular gap size and gap position, nevertheless, they can be very easily missed by a slight modification of a window length.

Regarding the speed of computation, the analysis model is about 2 to 4 times faster than the synthesis whereas the number of iterations of the synthesis model was from 4 to 8 times higher than of the analysis model. Another remark from this experiment is that the average computational time of the inpainting process using a proximity algorithm is not dependent on the gap size. On the other hand, the standard deviation of the computational time is slightly increasing with the growing gap size. The number of iterations is dependent on the window length, especially in the synthesis model.

The highest SNR has been reached in files containing a harmonic signal, especially when only a single instrument is playing. The worst results were obtained with rather non-harmonic records containing speech or completely non-harmonic signals.

The structured sparsity for audio inpainting evaluated by SNR produced restoration results comparable to the ℓ_1-relaxation without structure. Looking at the resulting time plot of the signal, there is almost no visible difference between the original and reconstructed signal. The restoration could be called perfect for small gap sizes up to 500 samples. Moreover, such comparable results were reached in a shorter time period. Finally, denoising using the structured sparsity outperformed professional software and was successfully utilized for denoising of recently found wax cylinders recordings.

This thesis proves that audio restoration could profit from sparse representations in terms of restoration quality. However, there is a long way from the theory to the real audio engineering field mainly because of the efficient implementation and optimization. Further research in this field could be focused on content based audio inpainting [2].

Like in all fields of research, new unanswered questions are arising from every answered query. There is great opportunity for the success of new methods. The ideas and results presented in this thesis are a step to contributing in this never ending journey.

References

[1] ADLER, A., EMIYA, V., JAFARI, M., ELAD, M., GRIBONVAL, R., AND PLUMBLEY, M. Audio Inpainting. *IEEE Transactions on Audio, Speech, and Language Processing 20*, 3 (March 2012), 922–932.

[2] BAHAT, Y., SCHECHNER, Y. Y., AND ELAD, M. Self-content-based audio inpainting. *Signal Processing 111*, 0 (2015), 61–72.

[3] BALAZS, P., DÖRFLER, M., JAILLET, F., HOLIGHAUS, N., AND VELASCO, G. Theory, implementation and applications of nonstationary Gabor frames. *Journal of computational and applied mathematics 236*, 6 (2011), 1481–1496.

[4] BARCHIESI, D. *Sparse Approximation and Dictionary Learning with Applications to Audio Signals.* PhD thesis, Queen Mary University of London, 2013.

[5] BAYRAM, I., AND KAMASAK, M. A simple prior for audio signals. *IEEE Transactions on Acoustics Speech and Signal Processing 21*, 6 (2013), 1190–1200.

[6] BECK, A., AND TEBOULLE, M. A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM Journal on Imaging Sciences 2*, 1 (2009), 183–202.

[7] BOYD, S., PARIKH, N., CHU, E., PELEATO, B., AND ECKSTEIN, J. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning 3*, 1 (2011), 1–122.

[8] BRUCKSTEIN, A. M., DONOHO, D. L., AND ELAD, M. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM Review 51*, 1 (2009), 34–81.

[9] CHEN, S., DONOHO, D., AND SAUNDERS, M. *Atomic decomposition by basis pursuit.* SIAM J. Sci Comput. 20 (1998), no.1, reprinted in SIAM Review, 2001.

[10] CHRISTENSEN, O. *Frames and Bases, An Introductory Course.* Birkhäuser, Boston, 2008.

[11] COMBETTES, P., AND PESQUET, J. A Douglas–Rachford splitting approach to nonsmooth convex variational signal recovery. *IEEE Journal of Selected Topics in Signal Processing 1*, 4 (2007), 564–574.

[12] COMBETTES, P., AND PESQUET, J. Proximal splitting methods in signal processing. *Fixed-Point Algorithms for Inverse Problems in Science and Engineering* (2011), 185–212.

[13] COMBETTES, P., AND WAJS, V. Signal recovery by proximal forward-backward splitting. *Multiscale Modeling & Simulation 4*, 4 (2005), 1168–1200.

[14] DONOHO, D. L., AND STARK, P. B. Uncertainty principles and signal recovery. *SIAM J. Appl. Math. 48*, 3 (1989), 906–931.

[15] ELAD, M. *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. Springer, 2010.

[16] ELAD, M., STARCK, J., QUERRE, P., AND DONOHO, D. Simultaneous cartoon and texture image inpainting using morphological component analysis (mca). *Applied and Computational Harmonic Analysis 19*, 3 (2005), 340–358.

[17] ETTER, W. 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.

[18] FEICHTINGER, H. G., AND STROHMER, T. *Advances in Gabor Analysis*. Birhäuser, 2001.

[19] GOODMAN, D., LOCKHART, G. B., WASEM, O., AND WONG, W.-C. Waveform substitution techniques for recovering missing speech segments in packet voice communications. *IEEE Transactions on Acoustics, Speech and Signal Processing 34*, 6 (Dec 1986), 1440–1448.

[20] HUBER, R., AND KOLLMEIER, B. PEMO-Q—A new method for objective audio quality assessment usinga model of auditory perception. *IEEE Trans. Audio Speech Language Proc. 14*, 6 (November 2006), 1902–1911.

[21] JANSSEN, A. J. E. M., VELDHUIS, R. N. J., AND VRIES, L. B. Adaptive interpolation of discrete-time signals that can be modeled as autoregressive processes. *IEEE Trans. Acoustics, Speech and Signal Processing 34*, 2 (4 1986), 317–330.

[22] KERELIUK, C., DEPALLE, P., AND PASQUIER, P. Audio interpolation and morphing via structured-sparse linear regression. In *Proceedings of the Sound and Music Computing Conference 2013* (Stockholm, 2013), pp. 546–552.

[23] KOWALSKI, M., AND TORRÉSANI, B. Structured Sparsity: from Mixed Norms to Structured Shrinkage. In *SPARS'09 – Signal Processing with Adaptive Sparse Structured Representations* (2009), R. Gribonval, Ed., Inria Rennes – Bretagne Atlantique, pp. 1–6.

[24] LAGRANGE, M., MARCHAND, S., AND RAULT, J.-B. Long interpolation of audio signals using linear prediction in sinusoidal modeling. *J. Audio Eng. Soc 53*, 10 (2005), 891–905.

[25] LUKIN, A., AND TODD, J. Parametric interpolation of gaps in audio signals. In *Audio Engineering Society Convention 125* (Oct 2008), pp. 3–6.

[26] MACH, V. Digital restoration of recordings from the phonograph cylinders and their copies. In *As recorded by the phonograph: Slovak and Moravian songs recorded by Hynek Bím, Leoš Janáček and Františka Kyselková in 1909–1912* (Brno, 2012), The Institute of Ethnology of the Academy of Sciences of the Czech Republic, v.v.i., pp. 165–176.

[27] MACH, V., AND OZDOBINSKI, R. Optimizing dictionary learning parameters for solving audio inpainting problem. *International Journal of Advances in Telecommunications, Electrotechnics, Signals and Systems 2*, 1 (2013), 40–45.

[28] MALLAT, S., AND ZHANG, Z. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing 41*, 12 (1993), 3397–3415.

[29] MCAULAY, R., AND QUATIERI, T. Speech analysis/synthesis based on a sinusoidal representation. *Acoustics, Speech and Signal Processing, IEEE Transactions on 34*, 4 (Aug 1986), 744–754.

[30] NAM, S., DAVIES, M., ELAD, M., AND GRIBONVAL, R. The cosparse analysis model and algorithms. *Applied and Computational Harmonic Analysis 34*, 1 (2013), 30 – 56.

[31] PATI, Y., REZAIIFAR, R., AND KRISHNAPRASAD, P. Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition. In *Conference Record of The Twenty-Seventh Asilomar Conference on Signals, Systems and Computers* (1993), pp. 40 –44.

[32] ŠPIŘÍK, J., RAJMIC, P., AND VESELÝ, V. Representation of signals: from bases to frames. *Elektrorevue – the online journal* (2010), 1–11.

[33] PROCHÁZKOVÁ, J., AND MACH, V. The Editing of Sound Recording from Phonograph (Wax) Cylinders at the Brno Branch of the Institute of Ethnology of the ASCR. In *As Recorded by the Phonograph* (Brno, 2012), The Institute of Ethnology of the Academy of Sciences of the Czech Republic, v.v.i., pp. 199–206.

[34] Rajmic, P., Bartlova, H., Prusa, Z., and Holighaus, N. Acceleration of Audio Inpainting by Support Restriction. In *Proceedings of the 7th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)* (2015), Brno University of Technology, pp. 325–329.

[35] Siedenburg, K. Persistent Empirical Wiener Estimation With Adaptive Threshold Selection for Audio Denoising. In *Proceedings of the 9th Sound and Music Computing Conference* (Copenhagen, Denmark, 2012), pp. 426–433.

[36] Siedenburg, K., and Doerfler, M. Persistent time-frequency shrinkage for audio denoising. *J. Audio Eng. Soc 61*, 1/2 (2013), 29–38.

[37] Siedenburg, Kai; Dörfler, M. Structured sparsity for audio signals. In *Proc. of the 14th Int. Conference on Digital Audio Effects (DAFx-11)* (September 2011), pp. 23–26.

[38] Tibshirani, R. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society. Series B (Methodological) 58*, 1 (1996), 267–288.

[39] Veldhuis, R. N. J. A method for the restoration of burst errors in speech signals. In *Signal Processing 3: Theories and Applications*. North-Holland, Amsterdam, 1986, pp. 403–406.

CURRICULUM VITÆ

Name:	**Ing. Václav Mach**
Born:	June 10, 1987 in Uherské Hradiště
Contact:	mach.vasek@gmail.com

Education

1998 – 2006	Secondary grammar school in Uherské Hradiště
2006 – 2009	Brno University of Technology, FEEC Bachelor of Teleinformatics
2009 – 2011	Brno University of Technology, FEEC Master of Communications and Informatics
2011 – 2016	Brno University of Technology, FEEC Ph.D. student of Telecommunications

Experience

2010 – 2014	Freelance Audio Engineer
2013 – 2015	Assistant, Brno University of Technology
2014 – 2015	C/C++ Developer, Team leader, Artisys, s r.o. (Brno, CZ)
2016 – now	Software Developer - Analyst, Ramet a.s. (Kunovice, CZ)

Research Grants

2010 – 2012	Cultural Identity and Cultural Regionalism in the Process of Forming an Ethnic Image of Europe. Academy of Sciences of the Czech republic, v.v.i. project no. AV0Z90580513
2013	Applications of sparse solutions in multidimensional data processing. Project leader, junior interfaculty research grant of BUT. Project no. FEKT/FSI-J-13-1903.
2013 – 2015	Center of Sensor, Information and Communication Systems (SIX). Project no. ED2.1.00/03.0072.
2013 – 2016	Novel methods for missing audio signal inpainting. Bilateral project with NuHAG and ARI Austria. Project no. 7AMB13AT021 and 7AMB15AT033.
2014 – 2015	Applications of digital audio restoration methods in the process of digitization of the audio records on magnetic tapes. Nation Institute of Folk Culture (Strážnice, CZ).
2014 – 2016	Cognitive multimedia analysis of audio and image signals. BUT project no. FEKT-S-14-2335.

Research Internships

May 2012	NuHAG, Faculty of Mathematics, University of Vienna, Austria
October 2013	NuHAG, Faculty of Mathematics, University of Vienna, Austria

Reviewer

- Certified Methodology for Digitization and Online Access of Gramophone Records and Other Sound Documents for Memory Institutions. Moravian Library (Brno), 2013

Awards

- 2nd place in EEICT Student Competition, category Signal processing, Cybernetics and Automation. April 2014.

Publications

- Book Chapters: 4
- Papers published in journals: 1
- Conference proceedings papers: 4
- Papers indexed in ISI WoS: 1
- Papers indexed in Scopus: 1
- Software: 1

Teaching

2011 – 2015	Occasional lectures on Audio Restoration for Audio Engineers and Ethnologists
2013 – 2014	Introduction to Computer Typography and Graphics, computer exercises
2012 – 2014	Electroacoustics, laboratory exercises
2011	Basics of Programming, computer exercises

7 Abstract

Recently, sparse representations of signals became very popular in the field of signal processing. Sparse representation mean that the signal is represented exactly or very well approximated by a linear combination of only a few vectors from the specific representation system.

This thesis deals with the utilization of sparse representations of signals for the process of audio restoration, either historical or recent records. Primarily old audio recordings suffer from defects like crackles or noise. Until now, short gaps in audio signals were repaired by interpolation techniques, especially autoregressive modeling. Few years ago, an algorithm termed the Audio Inpainting was introduced. This algorithm solves the missing audio signal samples inpainting using sparse representations through the greedy algorithm for sparse approximation.

This thesis aims to compare the state-of-the-art interpolation methods with the Audio Inpainting. Besides this, ℓ_1-relaxation methods are utilized for sparse approximation, while both analysis and synthesis models are incorporated. Algorithms used for the sparse approximation are called the proximal algorithms. These algorithms treat the coefficients either separately or with relations to their neighbourhood (structured sparsity). Further, structured sparsity is used for audio denoising.

In the experimental part of the thesis, parameters of each algorithm are evaluated in terms of optimal restoration efficiency vs. processing time efficiency. All of the algorithms described in the thesis are compared using objective evaluation methods Signal-to-Noise ratio (SNR) and PEMO-Q. Finally, the overall conclusion and discussion on the restoration results is presented.