Inversion of Auditory Spectrograms, Traditional Spectrograms, and Other Envelope Representations

Decorsière, Remi Julien Blaise; Søndergaard, Peter Lempel; MacDonald, Ewen; Dau, Torsten

Published in:
I E E E Transactions on Audio, Speech and Language Processing

Publication date:
2015

Document Version
Publisher's PDF, also known as Version of record

Inversion of Auditory Spectrograms, Traditional Spectrograms, and Other Envelope Representations

Rémi Decorsière, Peter L. Søndergaard, Ewen N. MacDonald, and Torsten Dau

***Abstract*—Envelope representations such as the auditory or traditional spectrogram can be defined by the set of envelopes from the outputs of a filterbank. Common envelope extraction methods discard information regarding the fast fluctuations, or phase, of the signal. Thus, it is difficult to invert, or reconstruct a time-domain signal from, an arbitrary envelope representation. To address this problem, a general optimization approach in the time domain is proposed here, which iteratively minimizes the distance between a target envelope representation and that of a reconstructed time-domain signal. Two implementations of this framework are presented for auditory spectrograms, where the filterbank is based on the behavior of the basilar membrane and envelope extraction is modeled on the response of inner hair cells. One implementation is direct while the other is a two-stage approach that is computationally simpler. While both can accurately invert an auditory spectrogram, the two-stage approach performs better on time-domain metrics. The same framework is applied to traditional spectrograms based on the magnitude of the short-time Fourier transform. Inspired by human perception of loudness, a modification to the framework is proposed, which leads to a more accurate inversion of traditional spectrograms.**

***Index Terms*—Spectrogram inversion, short-time Fourier transform, auditory spectrogram, gradient methods.**

I. Introduction

MULTI-CHANNEL envelope representations, also known as spectrograms, form a widely used class of time-frequency representations. However, due to envelope extraction, spectrograms do not explicitly retain all of the characteristics of the signals they represent, such as the very fast fluctuations known as temporal fine structure. Hence, recovering (or reconstructing) a time-domain signal from a given spectrogram representation is a challenging and interesting problem in mathematics and signal processing and has received significant attention (e.g., [1]–[3]). In this class of problem, the particular case of inverting what we call in this study the "traditional" spectrogram, which is given by the squared magnitude of the short-time Fourier transform (STFT), has been studied the most. A fundamental result for that case is presented in [1]. There, the authors provide a mathematical proof that representations obtained from the magnitude of the output coefficients from a linear system (which the STFT is) were injective, given that the linear system was sufficiently redundant (i.e., given that the STFT had sufficient overlap and/or channels). This means that, in principle, a time-domain signal *could* be recovered from its traditional spectrogram only.

Many studies have considered this particular case of the traditional spectrogram, some of them well before the mathematical validation of [1]. Initially, the problem was the reconstruction of a time-domain signal given only the magnitude of its Fourier transform (e.g., [3]). But the landmark study which first proposed a method for traditional spectrogram inversion is described in [4]. In their approach, Griffin & Lim use STFT and inverse STFT to project the signal back and forth between time domain and time-frequency domain in an iterative fashion. At each iteration, the STFT magnitude is constrained to equal the target spectrogram from which the signal is being reconstructed. In [4], the authors proved that the mean squared error of the STFT magnitude of the generated time-domain signal monotonically decreases with each iteration. This efficient and simple approach has now become a key reference of most modern studies and still inspires some state-of-the-art methods, even after three decades. Examples of such modern studies that extend on the work of Griffin & Lim (G&L) are [5] and [6]. In [5], the authors propose two variations on the G&L algorithm, the "real-time iterative spectrogram inversion" (RTISI) and its improved version, the "RTISI with look ahead" (RTISI-LA), which can perform real-time spectrogram inversion, and does so faster (i.e., in fewer iterations) than the original offline algorithm. The RTISI method processes time frames of the spectrogram one at a time and uses information from previous overlapping frames to generate a better initial phase for the current frame, allowing for faster convergence. In addition, the RTISI-LA method allows a number of future frames to be also considered, which was shown to improve accuracy in [5]. Conversely, [6] suggests an approximation to the G&L iteration process that allows much faster computation without impairing the accuracy of the inversion. Other modern methods have also investigated completely different approaches. For example, in [7], the authors suggest a frame-by-frame solution to the problem, where each frame is considered as a root-finding problem. Some mathematical studies (e.g., [8], [9]) have also considered retrieving the rank-one matrix associated with the signal. Instead of looking for the signal s, they look for the matrix ss^* (i.e., the outer

Manuscript received December 12, 2013; revised May 15, 2014; accepted October 21, 2014. Date of publication November 07, 2014; date of current version January 14, 2015. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Laurent Daudet.

R. Decorsière, E. N. MacDonald, and T. Dau are with the Centre for Applied Hearing Research and the Oticon Centre of Excellence for Hearing and Speech Science, Technical University of Denmark, DK-2800 Kongens Lyngby, Denmark (e-mail: remi.decorsiere@gmail.com; emcd@elektro.dtu.dk; tdau@elektro.dtu.dk).

P. L. Søndergaard was with the Center for Applied Hearing Research, Technical University of Denmark, DK-2800 Kongens Lyngby, Denmark. He is now with Oticon A/S, DK-2765 Smørum, Denmark (e-mail: peter@sonderport.dk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2014.2367821

product of s with itself). This results in a convex problem, which is easier to solve, but at the cost of squaring the dimensionality. Thus, this method is limited in practice to signals shorter than a few hundred samples. A further review of traditional spectrogram inversion can be found in [10].

While inversion of a traditional spectrogram is the most studied case, there are many other definitions of the spectrogram that can be considered as "multi-channel" envelopes. Among these are auditory representations based on human perception. In humans, the transduction from mechanical vibrations to electrical impulses in neurons occurs in the cochlea. Mechanical vibrations are transmitted into the cochlea via the middle ear and cause the basilar membrane to vibrate. The mechanical properties of the basilar membrane vary along the length of the cochlea. Conceptually, this can be modeled using a bank of bandpass filters with center frequencies and bandwidths that increase logarithmically. Situated atop the basilar membrane is the organ of Corti which contains inner hair cells (IHCs). These cells have stereocilia, small hair-like projections, which deflect in response to displacement of the basilar membrane. When the stereocilia are deflected in one (but not the other) direction, the IHCs become depolarized, leading to action potentials in afferent neurons. Conceptually, we can model this as a half-wave rectifier. After depolarizing, the IHC and afferent neurons must re-polarize. This imposes an upper limit to the frequency at which action potentials can be generated. This upper limit can be modeled as a low-pass filter, and, when applied after half-wave rectification, performs envelope extraction. Thus, as a first approximation, we can model the transduction in the cochlea as envelope extraction of the outputs from a filterbank. The result is a multi-channel envelope, that we call an "auditory spectrogram" (e.g.,[11]). Previous work has demonstrated that the envelope representation of a signal plays an important role in perception. For example, the envelope information from only a handful of bands can be sufficient for speech intelligibility (e.g., [12], [13]), and some models for predicting speech intelligibility rely on information derived from the envelope representations of the speech and noise signals (e.g., [14], [15]). Furthermore, a faithful representation of the envelope has been shown to be crucial for the perception of complex sounds (e.g., [16]). Hence, there are clear applications for an auditory spectrogram inversion tool in psycho-acoustic research. Unfortunately, the methods used for traditional spectrogram reconstruction cannot be applied to auditory spectrograms. However, a few studies (e.g.[17]) have examined this problem. In [17], the authors investigated the reconstruction from an auditory spectrogram where the filterbank is a bank of cascading low-pass filters and the IHC behavior is modeled by half-wave rectification only. They showed how the half-wave rectification in each channel can be approximately inverted by limiting the bandwidth of said channels, through iterative projections between time and spectro-temporal domain.

The present study, which expands the findings of [18], describes a more general tool for reconstructing a signal from its envelope representation. Here, we propose to address the spectrogram inversion problem as a general optimization problem on a time-domain signal, and use gradient-based methods to solve it. This results in a highly dimensional optimization problem (each sample in the sought signal is a dimension) requiring significant computational power to solve, and might explain why this approach has received little investigation. Here, we demonstrate how results can be obtained for short signals (e.g., a few seconds in length) using a standard modern computer. The applicability of the gradient-based approach relies on the possibility to efficiently compute said gradient. Although an analytical formula for the gradient might not be available for some envelope representations, it can be derived and efficiently computed for both auditory and traditional spectrograms. Inversion of an auditory spectrogram is motivated by the human auditory system, with an envelope extraction model based on IHC activity. Two implementation approaches for this method are presented, a straightforward application of the proposed framework, as well as a two-step process where the low-pass filter of the IHC envelope extraction is inverted and a signal is reconstructed from the half-wave rectified filterbank output, as suggested in [17] but using the present framework. The second case is based on the STFT magnitude representation, the term-by-term square root of the "traditional" spectrogram. By taking advantage of the flexibility of this new approach, adjustments are made in order to improve the accuracy of the reconstruction in this case. Results from both scenarios are evaluated and compared to other methods where possible.

II. General Framework

This section formalizes and suggests a potential solution to the problem of reconstructing a time-domain signal from a given envelope representation. Here, the term envelope representation is used to denote the set of envelopes of narrow-band channels obtained at the output of a filterbank. This representation is therefore dependent on the choice of a given filterbank and a given envelope extraction method. An approach to formalizing the role of a filterbank is to define a filterbank operator $\mathbf{V}$ which is used jointly with a set of analysis windows $\{g_m\}_{1\leq m\leq M}$, forming the analysis operator $\mathbf{V}_g$. It operates on an input signal s with a finite duration of L samples as follows:

$$(\mathbf{V}_g s)_{m,n} = \sum_{k=1}^{L} s\,[k]\, g_m\,[an-k] \tag{1}$$

Here, $s[k]$ denotes the kth sample of signal s. In practice, with this definition, the output $(\mathbf{V}_g s)$ is a matrix. Each row, later denoted by $(\mathbf{V}_g s)_m$, corresponds to a different frequency channel output (i.e., a subchannel) and is indexed by m, with $1 \leq m \leq M$. The columns of this matrix span time and are indexed by n, with $1 \leq n \leq N$. The decimation rate of the filterbank is controlled by the parameter a, which represents the hop-size, in terms of samples of the original signal s, between two consecutive points in any given subchannel. Given that a suitable set of synthesis (or dual) windows $\{g_m^{(d)}\}_{1\leq m\leq M}$ exists, this representation admits an inverse, the synthesis operator $\mathbf{U}$:

$$s\,[k] = \mathbf{U}_{g^{(d)}}\,(\mathbf{V}_g s)\,[k] = \sum_{m,n} (\mathbf{V}_g s)_{m,n} g_m^{(d)}\,[an-k] \tag{2}$$

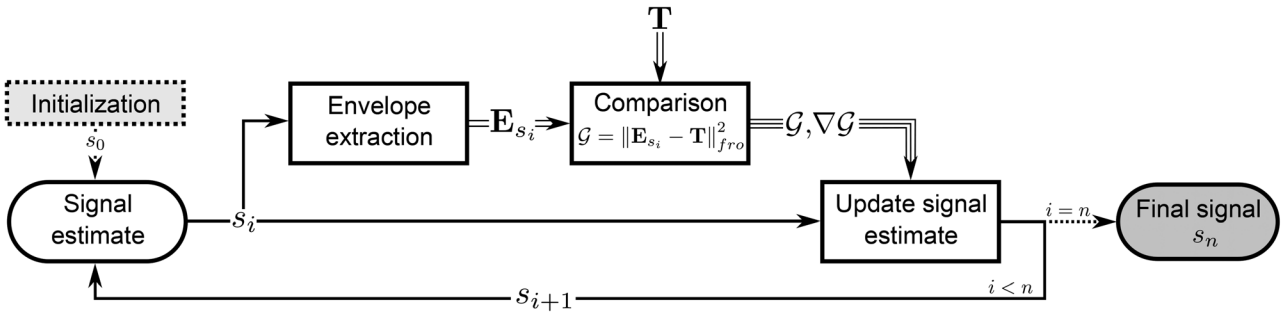


Fig. 1. Block diagram of the processing scheme. At a given iteration i, the envelope representation of the signal estimate is extracted. Its distance to the target envelope $\mathbf{T}$ and the gradient of this distance provide information to update the signal to a new estimate. This process is then repeated for n iterations.

Note that in the following, two different filterbanks will be used, and will have their own notation for the windows $\{g_m\}_{1\leq m\leq M}$ to avoid confusion.

Now, consider an envelope extractor function $E(\cdot)$ that operates on band-limited signals. The envelope representation is the set of envelopes of each subchannel $\{E((\mathbf{V}_g s)_m)\}_{1\leq m\leq M}$. As for the filterbank output, it is then convenient to adopt a matrix notation for the envelope representation, where each line represents a different channel, and the columns are for different samples in time:

$$\mathbf{E}_s = \begin{pmatrix} E\left((\mathbf{V}_g s)_1\right) \\ \vdots \\ E\left((\mathbf{V}_g s)_M\right) \end{pmatrix} \tag{3}$$

The matrix $\mathbf{E}_s$ is the envelope representation of the signal s. Assuming the envelope extraction does not perform any kind of decimation, $\mathbf{E}_s$ is an $M \times N$ matrix of non-negative real coefficients. As common analysis filterbanks usually provide band-limited outputs centered at different frequencies, each line in this matrix representation provides information related to the frequency content of the input signal. Hence, this matrix provides a time-frequency representation of the signal s. For example, in the case of the filterbank being the short-time Fourier transform (STFT), and the envelope extraction being the squared magnitude function, this matrix would correspond to the "traditional" spectrogram of the signal s.

Given a target envelope representation of a signal $\mathbf{T}$, an $M \times N$ matrix of non-negative real coefficients, the reconstruction problem is then stated as follows: Find a signal s such that $\mathbf{E}_s = \mathbf{T}$, or alternatively, such that $\mathbf{E}_s - \mathbf{T} = 0$. It is then convenient to define the real-valued function $\mathcal{G}$ that applies on any signal s with a length of L samples as follows:

$$\mathcal{G}(s) = \|\mathbf{E}_s - \mathbf{T}\|^2_{fro} = \sum_{i=1}^{M}\sum_{j=1}^{N}\left((\mathbf{E}_s)_{i,j} - (\mathbf{T})_{i,j}\right)^2 \tag{4}$$

The Frobenius norm $\|\cdot\|_{fro}$ is a matrix norm; hence $\mathcal{G}$ is the square of a norm-induced distance measure between the envelope of signal s and the target envelope $\mathbf{T}$. Therefore, the function $\mathcal{G}$ is positive-valued and equal to zero if, and only if, the matrices $\mathbf{E}_s$ and $\mathbf{T}$ are equal. Hence, $\mathcal{G}$ reaches a global minimum when the signal s has the required envelope $\mathbf{T}$. This suggests an optimization approach where the problem is restated as follows: find s that minimizes $\mathcal{G}(s)$. In this approach, the function $\mathcal{G}$ is now referred to as the objective function. Using an iterative optimization algorithm, a minimum of this function can be found.

In practice, spectrogram inversion methods are often used on targets $\mathbf{T}$ that are not explicitly derived from a time-domain signal (e.g., spectrograms that were processed in some ways). Some studies (e.g., [6]) denotes such targets as "inconsistent spectrograms", as they were not obtained from (3) but are instead an arbitrary two-dimensional set of coefficients. In such cases, there is no exact solution to the spectrogram inversion problem. The approach suggested here will in theory return the time-domain signal which minimizes $\mathcal{G}$, i.e., the signal s whose spectrogram $\mathbf{E}_s$ is *as close as possible* to $\mathbf{T}$ in terms of the distance $\mathcal{G}$, but may not ever reach $\mathbf{E}_s = \mathbf{T}$.

A block diagram of the general procedure is illustrated in Fig. 1. The optimization process begins with a random initial signal estimate. Each iteration i starts by calculating the envelope representation of the current signal estimate, $\mathbf{E}_{s_i}$. This is compared to the target $\mathbf{T}$ using the objective function $\mathcal{G}$. The value of $\mathcal{G}$ and its gradient $\nabla\mathcal{G}$ are used to update the signal estimate, resulting in a new estimate s_{i+1}. Thus, with each iteration, the optimization procedure generates an updated signal estimate that is "closer" (with relation to the distance $\mathcal{G}$) to the target envelope. The iteration process is terminated after n iterations.

From a practical perspective, it is important to note that numerical optimization methods often require knowledge of the first and sometimes second order derivatives of the minimized function (particularly when many dimensions are involved and "brute-force" search of the minimum becomes impractical). Although some algorithms can numerically estimate these derivatives, it is preferable to have an analytical expression to increase accuracy and reduce computational load. In the context of the present problem, the analytical expression of $\mathcal{G}$ will depend on how the filterbank is constructed and how envelope extraction is defined. The applicability of the method relies on being able to efficiently compute the gradient $\nabla\mathcal{G}$ of the function $\mathcal{G}$. In the following sections, we will present how this gradient can be efficiently computed for two particular cases: an auditory-motivated envelope representation (i.e., auditory filterbank and inner hair-cell envelope), and the traditional spectrogram (i.e., STFT squared magnitude). As an analytic expression for the derivative is needed, there may be some filterbank/envelope combinations that are not compatible with our approach. For example, it might not be possible to analytically derive a gradient expression for

envelope definitions that are based on the estimation of the instantaneous frequency of an associated carrier wave, as is done in some vocoder studies (e.g., in [19]).

To speed up convergence and improve accuracy, information regarding the second-order derivative is helpful. However, for a signal with a length of L samples, the matrix of the second-order derivative of $\mathcal{G}$, the Hessian matrix, contains L^2 elements. Thus, for typical speech signals, computing and storing all the elements of this matrix is often impractical, if not impossible. However, this can be overcome using a specific class of optimization algorithms, a limited memory Broyden-Fletcher-Goldfarb-Shanno (l-BFGS) algorithm [20]. This algorithm only manipulates a sparse representation of the Hessian matrix consisting of a few vectors of size L instead of the full L^2 matrix.

By taking these practical considerations into account, implementing this optimization approach to reconstructing signals from their envelope representation becomes reasonable even on a standard computer, as will be described in Section V.

III. Reconstruction from IHC Inspired Envelope Extraction

At the simplest functional level, two elements are required to generate envelope representations of a signal: an analysis filterbank to generate a time-frequency representation and an envelope extraction operator. In this section, a Gammatone filterbank is used, as it provides a simplified model of the time-frequency analysis conducted by the human cochlea [22]. Similarly, an envelope extraction operator that is based on inner hair-cell (IHC) processing is used.

A. Gammatone Filterbank

The Gammatone filterbank provides a simplified, linear model of basilar membrane motion. Unlike the linear spacing of frequency bins in the STFT, the center frequencies of the Gammatone filterbank are equally spaced on an Equivalent Rectangular Bandwidth (ERB) scale (see [23] for further details). An individual Gammatone filter with center frequency f_c, bandwidth β, amplitude α and order n_f is given by its impulse response as follows:

$$\gamma(t) = \alpha t^{n_f - 1} e^{-2\pi\beta t} \cos(2\pi f_c t) \tag{5}$$

Given a vector of ERB-spaced center frequencies, $\{(f_c)_m\}_{1\le m\le M}$, we obtain a set of filters $\{\gamma_m\}_{1\le m\le M}$ that forms a Gammatone filterbank and applies to a signal according to (1).

B. IHC Inspired Envelope Extraction

As described in the introduction, we have based our envelope extractor on a simplified IHC model that consists of a half-wave rectifier followed by a low-pass filter. While similar envelope extractors are used in electronic circuits, the filter parameters that we have used are based on psychoacoustic data (e.g., [11]). To generate an envelope representation of a signal, the half-wave rectification and low-pass filtering are applied to the output from the Gammatone filterbank. Given the mth channel at the output of the Gammatone filterbank $(\mathbf{V}_\gamma s)_m$, the first step of the envelope extraction is half-wave rectification, which will be denoted by $(\mathbf{V}_\gamma s)^+_m$. This consists of setting all negative valued samples to zero while leaving positive-valued samples unchanged:

$$(\mathbf{V}_\gamma s)^+_{m,n} = \begin{cases} (\mathbf{V}_\gamma s)_{m,n} & \text{if} \quad (\mathbf{V}_\gamma s)_{m,n} \ge 0 \\ 0 & \text{else} \end{cases} \tag{6}$$

Alternatively, we can introduce the Heaviside function $\mathcal{H}\{\cdot\}$. Given an input vector, this function returns a vector of the same size with values of 1 for indices where the input was positive and values of 0 for indices where the input was non-positive:

$$(\mathbf{V}_\gamma s)^+_m = \mathcal{H}\left\{(\mathbf{V}_\gamma s)_m\right\} \cdot (\mathbf{V}_\gamma s)_m \tag{7}$$

In (7) and further equations, $\cdot$ denotes term-by-term multiplication of two vectors or matrices with the same number of elements. The second step of the envelope extraction is low-pass filtering. Assuming a low-pass filter with an impulse response h, the envelope of the mth channel is given by:

$$(\mathbf{E}_s)_m = (\mathbf{V}_\gamma s)^+_m * h \tag{8}$$

Here, $*$ denotes convolution. Given a target envelope representation $\mathbf{T}$ computed according to (8), the reconstruction problem is to recreate the signal having $\mathbf{T}$ as envelope representation.

C. Direct Reconstruction

Using the definition of the envelope from (8), the objective function is given by the following equation:

$$\mathcal{G}(s) = \|\mathbf{E}_s - \mathbf{T}\|^2_{fro} \tag{9}$$

It can be seen from (1) that the derivative of the Gammatone analysis operator with relation to the kth coefficient of the input can be expressed as follows:

$$\frac{\partial}{\partial s[k]}(\mathbf{V}_\gamma s)_{m,n} = \gamma_m[n-k] \tag{10}$$

when there is no decimation in the filterbank (which is the case for this application). Combining (10) with (7), (8) and (9), and assuming the low-pass filter has a finite impulse response (i.e., an FIR filter), it is possible to express the gradient of $\mathcal{G}$ analytically. While typical IHC models do not use FIR filters, truncating the otherwise infinite impulse response is a reasonable approximation. If $h[k] = 0$ for $k > K$, then the gradient of the objective function can be expressed as follows:

$$\nabla\mathcal{G} = 2\sum_{k=0}^{K} h[k]\, \mathbf{U}_{\mathcal{T}^k\{\gamma\}}\left\{(\mathbf{E}_s - \mathbf{T})\, \mathcal{T}^k\{\mathbf{H}\}\right\} \tag{11}$$

Here, $\mathcal{T}$ denotes the time-shift (translation) operator,

$$(\mathcal{T}\{s\})[k] = s[k-1] \tag{12}$$

or similarly,

$$(\mathcal{T}^p\{s\})[k] = s[k-p], \text{ for any integer } \quad p \tag{13}$$

The matrix $\mathbf{H}$ represents the Heaviside function applied to all the channels at the output of the filterbank:

$$\mathbf{H}_{m,n} = \mathcal{H}\left\{(\mathbf{V}_\gamma s)_{m,n}\right\} \tag{14}$$

The gradient in (11) is expressed as a finite sum (of $K+1$ elements) under the assumption of the low-pass filter being an FIR filter. Each element of the sum is expressed using the filterbank synthesis operator $\mathbf{U}$ defined in (2) applied with a time-shifted version of the original filterbank *analysis* window. Importantly, note that direct knowledge of the *synthesis* windows $\{\gamma_m^{(d)}\}_{1\leq m\leq M}$ introduced in Section II is not needed. This will also be the case for the gradient expressions in later sections where different envelope extraction schemes are used. As can be seen from (2), the operator $\mathbf{U}$ can be implemented using the fast Fourier transform (FFT), hence the gradient in (11) can be efficiently computed. Using (9) and (11), $\mathcal{G}(s)$ can be minimized with an iterative optimization procedure (l-BFGS algorithm).

D. Two-Step Reconstruction

The direct approach detailed above attempts to reconstruct a signal directly from a target envelope representation. However, it is also possible to process the envelope representation before applying the iterative optimization algorithm. Here, we propose a two-step reconstruction method inspired by [17]. Conceptually, the approach is straightforward. Recall that the IHC envelope extraction is modeled as half-wave rectification followed by low-pass filtering. Thus, if the inverse of the low-pass filter is applied to the target envelope representation, the result is the half-wave rectified output of the filterbank. The signal can then be reconstructed from this representation using the iterative optimization approach. Under the assumption that each channel has a narrow bandwidth, it was suggested in [17] that a band-pass filter should be used to remove harmonics introduced by the half-wave rectification. However, we propose a more global approach based on the general framework suggested in Section II that takes the interactions between channels into account. This has the advantage of using the information from neighboring channels to recover the information lost in a given channel by the half-wave rectification.

Low-Pass Filter Inversion and Regularization: The low-pass filter impulse response h in (8) results in the filter response H in the frequency domain. The low-pass filtering is inverted by multiplying each channel in the frequency domain with the inverse filter response $1/H$. However, by definition, the response of a low-pass filter at higher frequencies is very small. Thus, a direct application of the inverse filter response would result in a very large unbounded gain being applied at high frequencies. This would introduce instability in the reconstruction, as any errors at high frequencies (e.g., rounding error) would be unreasonably amplified. Hence, it is necessary to regularize the inverse filter response by introducing an upper bound G_{max}, on the maximum gain allowed on the inverse filtering procedure. Given the mth channel of the target $(\mathbf{T})_m$ and the classic Fourier transform operator $\mathcal{F}\{\cdot\}$, the regularized inverse low-pass filtering generates the new target for this channel $(\mathbf{T}^+)_m$ as follows:

$$\left(\mathbf{T}^+\right)_m = \mathcal{F}^{-1}\left\{\mathcal{F}\left\{(\mathbf{T})_m\right\}\cdot\max\left(\frac{1}{|H|},G_{max}\right)e^{-j\angle H}\right\} \tag{15}$$

Here, the function $\max(\cdot)$ operates on individual coefficients of the vector $1/H$. The phase of the inverse response is maintained by multiplying with $e^{-j\angle H}$. The outcome of this step is the new target $\mathbf{T}^+$, where the superscript $(\cdot)^+$ suggests that this target corresponds to a half-wave rectified output of the filterbank. The regularization introduces inaccuracies in the representation, and in practice there is a tradeoff in the choice of the maximal gain G_{max}. Low gain results in good stability of the procedure but large inaccuracies. Alternatively, a high-gain limits the loss of information from the regularization, but at the cost of reduced stability. We have observed that, when increasing G_{max}, transients of large amplitude appear at the beginning and end (i.e., first and last few milliseconds) of the reconstructed signals. Increasing it further will eventually lead to global instability of the reconstruction scheme. This phenomenon can be used in an actual blind scenario, i.e., when the original signal is unknown, to adjust G_{max}, by increasing its value while monitoring these onset and offset transients. We have found that $G_{max}=60$ dB is a suitable compromise for speech signals. However, G_{max} could be increased further for more stationary signals.

Half-Wave Rectification Inversion: For the two-step approach, the reconstruction problem is to estimate a signal whose half-wave rectified output from the filterbank is *as close as possible* to the target $\mathbf{T}^+$. Typically, the low-pass filter inversion described in (15) is not perfect, hence these two representation can only be close but not strictly equal. This can be solved using the optimization approach proposed above, by defining the objective function as follows:

$$\mathcal{G}(s) = \left\|(\mathbf{V}_\gamma s)^+ - \mathbf{T}^+\right\|_{fro}^2 \tag{16}$$

With this formulation, the gradient is expressed as follows:

$$\nabla\mathcal{G} = 2\mathbf{U}_\gamma\left[\left((\mathbf{V}_\gamma s)^+ - \mathbf{T}^+\right)\cdot(\mathbf{V}_\gamma s)^+\right] \tag{17}$$

In comparison to (11), the gradient here has a simpler form and requires approximately K times fewer calculations. Thus, there is a clear advantage of this two-step approach in terms of implementation.

IV. Reconstruction from STFT Magnitude

A. Direct Application of the Framework

Most studies concerning spectrogram reconstruction have been conducted on the "traditional" spectrogram, i.e., an envelope representation given by the squared magnitude of the short-time Fourier transform (STFT) coefficients. The STFT with a window w and a hop-size a can be implemented as a filterbank, where individual filters have as impulse response the original window w modulated by a complex-valued exponential at a given channel frequency:

$$\tilde{g}_m[k] = w[k]e^{2\pi j f_m k} \tag{18}$$

The channel frequencies $\{f_m\}_{1\leq m\leq M}$ are chosen such that they span the Nyquist domain and are linearly spaced. The number of channels, M, is related to the length L in samples of the window w through $M=L/2+1$. Given these filters, the STFT is applied to an input signal using (1).

The envelope extraction in this case is the magnitude function:

$$\mathbf{E}_s = |\mathbf{V}_{\tilde{g}}s| \tag{19}$$

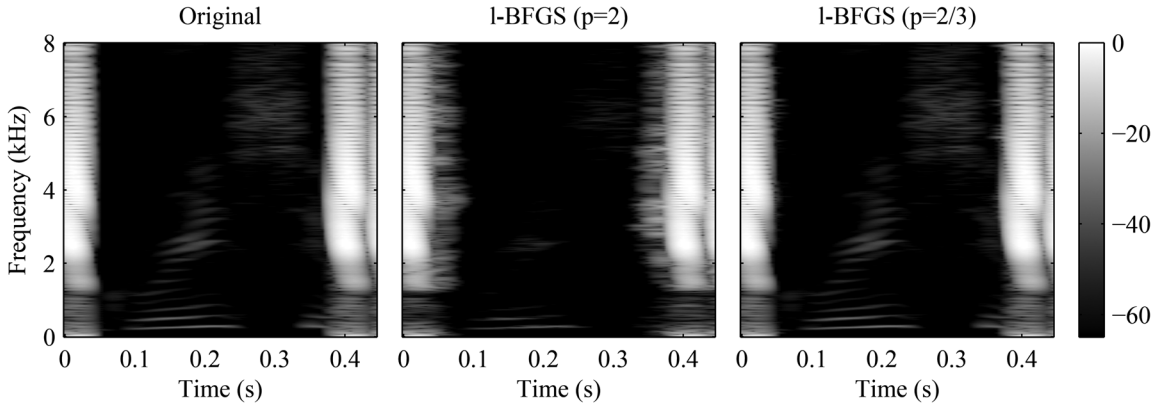


Fig. 2. Spectrogram (in dB) of a quiet, yet audible speech signal (from around 0.05 to 0.35s) embedded between two short bursts (left plot). In a signal reconstructed from this spectrogram with the standard proposed method ($p = 2$, center plot), the speech becomes inaudible. However, when using a compressed objective function ($p = 2/3$, right plot), the speech is audible.

We can then define an objective function from the "traditional" spectrogram, i.e., the squared STFT magnitude:

$$\mathcal{G}(s) = \left\| |\mathbf{V}_{\tilde{g}} s|^2 - \mathbf{T}^2 \right\|_{fro}^2 \tag{20}$$

With this definition, individual coefficients of the envelope contribute to the objective function with regard to their energy (i.e., squared magnitude). A convenient property of this definition is that the derivative of the squared magnitude function can be expressed as follows:

$$\left(|u|^2\right)' = 2\Re\left(\overline{u}u'\right) \tag{21}$$

Here, $(\cdot)'$ is the derivative, $\overline{(\cdot)}$ the complex conjugate, and $\Re(\cdot)$ the real part. Hence, by combining (21), (10) and later (2), the gradient of the objective function can be expressed using the filterbank synthesis operator, but once again applied using the original analysis window:

$$\nabla\mathcal{G} = 4\Re\left(\mathbf{U}_{\tilde{g}}\left[\left(|\mathbf{V}_{\tilde{g}} s|^2 - \mathbf{T}^2\right) \cdot \mathbf{V}_{\tilde{g}} s\right]\right) \tag{22}$$

B. Refining the Objective Function

Optimizing the objective function $\mathcal{G}(s)$ as written in (20) will reduce the average error in the envelope representation of the reconstructed signal. However, we have observed that this approach would offer poor reconstruction of low-energy regions of the spectrogram. This can be a problem, as such regions might still be audible. Fig. 2 (left and center panels) provides an example of such phenomenon. Left panel shows the spectrogram of an attenuated speech signal that is embedded between two short bursts. The speech is relatively quiet, but still audible. If a signal is reconstructed from this spectrogram using (20), the speech is no longer audible. The spectrogram of the reconstructed signal (center panel of Fig. 2) indeed shows that the speech sample has not been reconstructed.

To account better for low-energy regions of the spectrogram, a modified objective function $\mathcal{G}_L(s)$ is proposed:

$$\mathcal{G}_L(s) = \left\| |\mathbf{V}_{\tilde{g}} s|^p - \mathbf{T}^p \right\|_{fro}^2 \tag{23}$$

If $p < 2$ in (23), the dynamic range of individual contributions to the objective function is reduced in comparison to (20), which increases the relative contribution of regions with lower energy. Because of this compressive behavior, we refer to $\mathcal{G}_L$ as the *compressed* objective function. For an arbitrary p, the gradient corresponding to this function is given as follows:

$$\nabla\mathcal{G}_L(s) = 2p\Re\left(\mathbf{U}_{\tilde{g}}\left[(|\mathbf{V}_{\tilde{g}} s|^p - \mathbf{T}^p) \cdot |\mathbf{V}_{\tilde{g}} s|^{\frac{p}{2}-1} \cdot \mathbf{V}_{\tilde{g}} s\right]\right) \tag{24}$$

which simplifies to (22) if $p = 2$.

Choosing a value for p is a matter of compromise. Ideally, p should be as close to 0 as possible to increase the contribution of lower-energy regions. However, the smaller p is, the more "flat" the objective function in (23) is, and, therefore, the harder it becomes to find its minimum. We have found that $p = 2/3$ was a good compromise as it performed slightly better than $p = 1$ and produced more consistent results than $p = 1/2$, where the method would often fail to find a minimum. Additionally, based on Stevens' power law for loudness [24], $p = 2/3$ implies that the contribution to the objective function of individual time-frequency bins is approximately proportional to their loudness (e.g., as modeled in [25], [26]). The right panel of Fig. 2 presents the spectrogram of a signal reconstructed using $p = 2/3$ in (23). Listening to this reconstructed signal reveals that the speech sample is now audible and adequately reconstructed, as can be seen visually on its spectrogram.

V. Evaluation and Comparison of Techniques

A. Implementation, Testing Material and Evaluation of Convergence

To evaluate the proposed techniques, the framework was implemented in Matlab. The Matlab implementation of the l-BFGS optimization algorithm was found in [27] and used with all default settings, except for one. The termination tolerance was reduced to avoid the algorithm stopping prematurely. The reconstruction framework was tested using a speech corpus containing individual recordings of 100 English words, 50 spoken by a female and 50 by a male. Both were native English speakers. This corpus was formed from segmented keywords of the NU-6 WIN test [28]. Results depend on the random initialization of the algorithm. Hence, when different methods for a same representation are compared in the following, they will be initialized in the exact same way.

TABLE I
RESULTS FOR IHC ENVELOPE REPRESENTATIONS

Method	$\mathcal{C}$ (dB)	Time (s)	Iterations	$\mathrm{RMS_n}$ (dB)
l-BFGS (Direct)	-35.9	208	79	-15.1
l-BFGS (Two-step)	-47.3	8.5	32	-38.0

Methods for IHC inspired representations were initialized using $s_0 = \mathbf{U}_{\gamma^{(d)}}[\mathbf{T} \cdot e^{i\phi_0}]$, where ϕ_0 is a randomly generated phase. Similarly, methods for STFT magnitude representations were initialized by using the same realization of a uniformly distributed random STFT phase, except for the RTISI-LA algorithm which involves a particular initialization strategy.

A useful measure to compare algorithms introduced in [10] is the spectral convergence $\mathcal{C}$ (a related measure was earlier proposed by [6]). The spectral convergence measures the distance between target and reconstructed signals, in the time-frequency (STFT-magnitude) domain. It is the normalized Euclidean distance between the target spectrogram and the spectrogram of the reconstructed signal:

$$\mathcal{C} = \frac{\|\mathbf{E}_s - \mathbf{T}\|_{fro}}{\|\mathbf{T}\|_{fro}} \tag{25}$$

The exact expression of $\mathbf{E}_s$ (as well as the way $\mathbf{T}$ is computed) depends on the type of spectrogram, and is given either by (8) or (19). The more accurate the reconstruction (i.e., the closer the spectrogram of the reconstructed signal is to the target), the smaller $\mathcal{C}$ is.

B. Results from IHC Inspired Envelope Representations

Although there are various models of the IHC envelope extraction documented in the literature, most use a similar structure and differ only with regards to the low-pass filter order and cutoff frequency (e.g., [11], [29], [30]). Here, the IHC envelope extraction model from [11] was selected. This model uses half-wave rectification followed by a second order Butterworth low-pass filter with cutoff frequency at 1000 Hz. Time-domain signals were reconstructed from such envelope representations, using both the direct and two-step approaches described in Section III, and for the corpus of 100 words. Results, averaged over the whole corpus, are presented in Table I.

For both methods, the maximum number of iterations was set to 80, but the algorithm often stopped prematurely without being able to find a better solution (particularly for the two-step approach). Hence, the average number of iterations and elapsed time are also presented in Table I, along the averaged spectral convergence $\mathcal{C}$ expressed in dB. In terms of spectral convergence, there is a substantial benefit for the two-step approach. In addition, from a practical point of view, the two-step approach has a clear advantage with much shorter computation time.

In addition to the spectral convergence, and since we have knowledge of the original signal, it is possible to measure the root mean square (RMS) error. The normalized RMS error of the reconstructed signal s_r with relation to the original signal s, is expressed with the Euclidean norm $\|\cdot\|$ as follows:

$$\mathrm{RMS_n} = \frac{\|s - s_r\|}{\|s\|} \tag{26}$$

This assumes that the signals have the same number of samples, which is the case here. The RMS measures reconstruction errors in the signal domain, whereas the spectral convergence measures an error in the spectrogram domain. The $\mathrm{RMS_n}$, averaged over the whole corpus, is presented in Table I. In terms of normalized RMS error, while both perform well, the two-step approach performs better than the direct approach.

C. Results for STFT Magnitude

As with many studies presenting new or improved methods for spectrogram inversion (e.g., [5]–[7], [9], [10], [21]), we use the Griffin and Lim (G&L) algorithm from [4] as one of our baselines to evaluate our method. In addition, the following results will also be compared to the more modern "real-time iterative spectrogram inversion with look ahead" (RTISI-LA) introduced in [5]. The amount of look-ahead was set to 3 frames, which is the configuration adopted in most scenarios presented in [5].

As mentioned earlier, the spectrogram is determined by a given window w, which determines the number of frequency points, and a hop-size a (sometimes given as a percentage of overlap between neighboring windows). As a thorough investigation of all parameters related to traditional spectrograms is beyond the scope of this paper, results from a common parameter set is presented: a 1024-sample Hann window, with $a = 256$ (i.e., 75% overlap). This set was used in [5].

Fig. 3 compares the spectral convergence $\mathcal{C}$ measured in dB as a function of iteration number (left panel) and computation time (right panel) for our suggested method, using an uncompressed ($p = 2$) and compressed ($p = 2/3$) objective function, as well as for the Griffin and Lim (G&L) and the RTISI-LA methods. Spectral convergence and computation time were averaged over the 100 available speech signals. Despite the care taken to force the l-BFGS algorithm to stop at a requested number of iterations, the algorithm with a compressed objective function sometimes failed to find a better solution and stopped prematurely. This occurred for 15 of the 100 utterances and is a consequence of choosing a low p value in (23). The average spectral convergence in the left plot of Fig. 3 includes these outliers. However, they were excluded from the average in the right plot, as they would bias the average computing time towards lower values. The absence of these outliers in the right plot also explains why the spectral convergence for l-BFGS ($p = 2/3$) is lower by about 4 dB in comparison to the left plot. Computation times reported on the right plot are meant to be compared in a relative way, as they are strongly influenced by the implementation (which may not have been optimal) and hardware used. In particular, our implementation of the RTISI-LA is sub-optimal in terms of computation time, due to the limitations of the software used (Matlab). Its computation times are reported for the sake of completeness, but should not be compared in absolute, nor in relation to other methods. However, in [5], the accuracy of the reconstruction is measured in terms of "signal-to-error ratio" (SER), which is simply the opposite (when expressed in dB) of the spectral convergence. Although they were derived from a different speech material, the accuracy of the signals obtained from our implementation of the RTISI-LA is comparable with the results presented in [5].

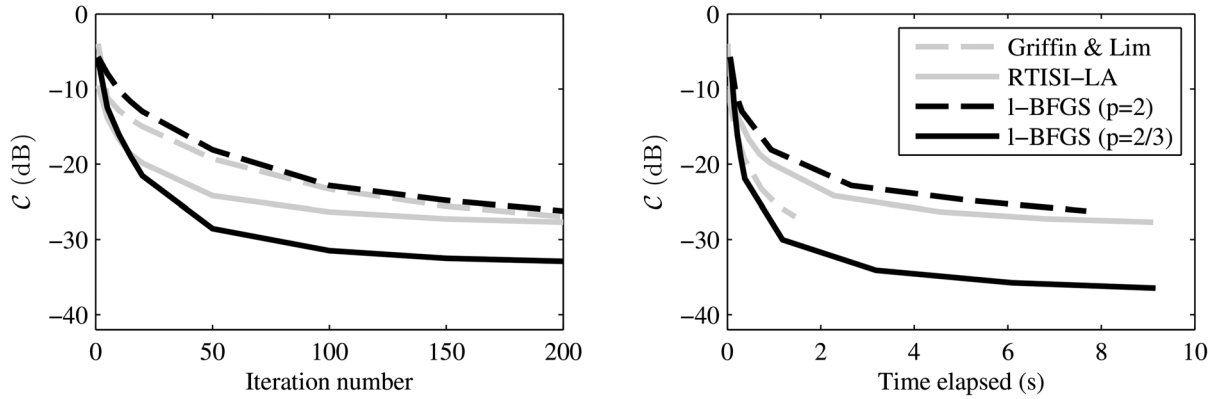


Fig. 3. Spectral convergence $\mathcal{C}$ as a function of the number of iterations (left panel) and computation time (right panel), measured for the Griffin and Lim algorithm, the RTISI-LA algorithm with 3 look-ahead frames, and the proposed algorithm (labeled l-BFGS) when $p = 2$ and $p = 2/3$. These results are averages obtained over 100 individual word tokens, with the same random initialization for the three methods. The l-BGFS method with $p = 2/3$ terminated early for 15 of the speech tokens. These early terminations are included in the left plot, but not in the right plot.

Our proposed method with an uncompressed objective function ($p = 2$) exhibits similar spectral convergence $\mathcal{C}$ to the G&L algorithm. However, the G&L approach is much less computationally intensive. Introducing the compressed objective function (i.e., when $p = 2/3$) substantially changes the results. Our proposed method outperforms G&L from the start and shows a reduction in spectral convergence of around 10 dB at 50 iterations. Because of the above-mentioned outliers in the average, this advantage is reduced slowly as more iterations are considered, down to 5 dB. If the outliers are not considered in the average, the 10 dB advantage of the proposed method over G&L is maintained across iteration number (as seen on the right plot of Fig. 3). The RTISI-LA is slightly better for the first 15 iterations, after which it falls behind our proposed method.

Further analysis was conducted with different degrees of overlap in the spectrogram. Our method produces very similar results for both lower and higher overlap. The G&L method tends to provide more accurate results for lower overlap, though still less accurate than the optimization approach. The RTISI-LA performance is strongly influenced by increasing the amount of overlap while keeping the same number k of look-ahead frames. If the overlap is increased such that more than $k + 1$ frames overlap at a given instant, then the RTISI-LA performance is impaired and the spectral convergence will not decrease further after some iteration (details can be found in [5]). A solution is to increase k, but this significantly increases the computation time.

VI. Discussion

A. Advantages of the Proposed Framework

In this study, a framework for reconstructing time-domain signals from an arbitrary multi-channel envelope representation (i.e., a spectrogram) was presented. The framework is based on minimizing a distance measure (the objective function) between the spectrogram of a signal and a target spectrogram by means of a numerical optimization algorithm. A necessary condition for the framework to be applicable to a given representation is for the gradient of the objective function to be efficiently computed. This was shown to be the case for both "traditional" and auditory spectrograms. In addition, a fundamental advantage of this approach is that it does not rely on having a mathematical inverse for the time-frequency representation. Various methods, including [4], rely on projecting the signal back and forth between time and time-frequency domain. This is made possible for "traditional" spectrogram because an *inverse* STFT is available. But there is no such inverse for the combination of Gammatone filterbank and IHC envelope, and similar iterative projection strategies cannot be applied in that case.

The main advantage of this framework over more specific methods resides in the fact that the objective function can be tailored to a specific problem. The flexibility of this definition allows the use of the same method for different problems (e.g., both traditional and auditory spectrograms). In theory, other combinations of filterbank and envelope extraction schemes could be considered, as long as the gradient of the objective function can be efficiently computed. Further, the definition of the objective function can be fine-tuned to a specific problem. For example, a compressed objective function was shown to improve substantially the reconstruction accuracy. Other scenarios involving adjustments to the objective function could easily be imagined. For example, a frequency-weighted objective function could be used when accurate reconstruction is needed only in a given frequency range, or when the target spectrogram is unknown in some range.

B. Methods Comparison for Traditional Spectrogram Inversion

Three methods for "traditional" spectrogram inversion were compared. Although the framework presented here achieved the lowest overall spectral convergence when using a compressed objective function, the two other methods would have significant advantages in specific scenarios. The Griffin & Lim algorithm [4] was both the fastest and easiest to implement. It is a safe choice to consider when low computation time and ease of use is more important than achieving a more accurate reconstruction. On the other hand, The RTISI-LA method [5] has the clear advantage that it is compatible with real-time processing. It was also the most accurate for a low iteration number (below 15 in Fig. 3), the scenario for which it was designed. However, maintaining its accuracy for representations with higher overlap involves increasing the number of look-ahead frames,

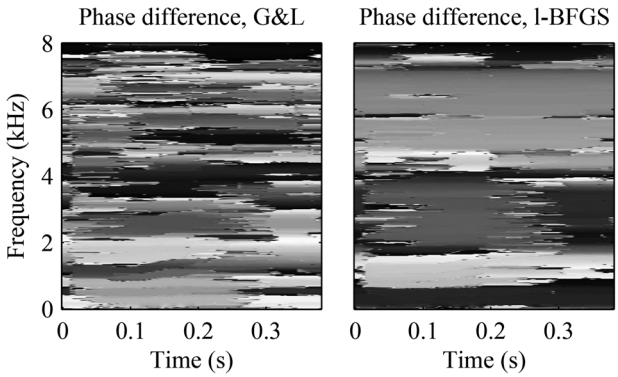


Fig. 4. Illustration of the stagnation phenomenon: phase difference between STFTs of original and reconstructed signal using G&L (left) and when using the proposed method (right) with a compressed objective function ($p = 2/3$), both after a very high number of iteration.

which will reflect on the computation time. Overall, our proposed framework is best suited for offline spectrogram inversion where maximum accuracy is more important than computation time.

C. Reducing the Intrinsic Limitation of Reconstruction from STFT Magnitude

A substantial improvement in the accuracy of the reconstruction was observed when introducing the compressed objective function (see Fig. 3). We relate this improvement to a release in the intrinsic limitation faced when reconstructing traditional spectrograms. This limitation was originally referred to as stagnation in [31]: because the magnitude suppresses all information about the absolute phase, there can be a phase mismatch between local regions of the STFT of the original and reconstructed signals. To improve reconstruction further, this mismatch must be reduced. Thus, phase shifts must be applied to large time-frequency regions of the STFT. In terms of our optimization approach, this means that a slightly better solution exists, but it may be very "far" away in the signal domain. Stagnation causes waveforms of original and reconstructed signals to present strong discrepancies in the time domain, even if very low spectral convergence was achieved. This can limit the utility of $\mathrm{RMS_n}$ as a performance metric.

To illustrate stagnation, Fig. 4 presents the phase difference (modulo 2π) between the STFT of original and reconstructed signals for two methods. In the left panel, the signal was obtained using G&L with 1000 iterations, which is sufficiently many to assume no significant further progression of the algorithm. For this specific case, a spectral convergence of -27 dB was measured. In the right panel, the proposed method with the compressed objective function was used. It stopped after 467 iterations, leading to a spectral convergence of -34 dB. As only relative phase is relevant, the color bar was omitted to save space, but both plots are on the same scale. Although for both methods the magnitude difference between the STFT of original and reconstructed signals would be very small everywhere, the phase differences are far from uniform across the time-frequency plane. However, the phase difference in the right panel is smoother than in the left panel, suggesting that the use of the compressed function leads to signal estimates that are less prone to stagnation. This could explain the significant improvement in spectral convergence that can be observed on Fig. 3.

When the compressed objective function is used, the low-energy regions of the spectrogram contribute more than they would if either the uncompressed function or the G&L algorithm was used. It appears reasonable that, in order to avoid stagnation, one needs to have a good estimate of the phase of the STFT not only in high energy regions, but over the entire time-frequency plane. By increasing the contribution of lower energy regions, the proposed method provides better reconstruction of these regions. This, in turn, provides a more consistent estimate of the phase over the whole time-frequency plane, and limits stagnation.

D. Reconstruction from Auditory Spectrograms

While stagnation is a problem that is intrinsic to the way the "traditional" spectrogram is defined, it might not affect other types of spectrograms. We believe that, despite their higher redundancy in relation to "traditional" spectrograms which are decimated in time, this is the main reason why reconstruction of auditory spectrograms was shown to be more accurate (particularly in terms of a time-domain metric such as $\mathrm{RMS_n}$). Unlike the magnitude function that removes the absolute phase of a signal, the half-wave rectification step in the IHC envelope extraction sets negative portions of channel outputs to zero, and therefore still maintains basic information regarding the absolute phase of the signal. This means that there is no sign indeterminacy and thus no stagnation. Although the reconstruction from IHC envelope provides signals with spectral convergence that is comparable to the one obtained when reconstructing from STFT magnitude, a close match between the waveforms of original and reconstructed signals can still be reached (as seen in the $\mathrm{RMS_n}$ values in Table I).

When comparing the one- vs. two-step approaches, there is a clear advantage of the two-step approach in terms of reconstructed waveform accuracy (i.e., $\mathrm{RMS_n}$). For the direct approach, errors presumably consist of a misestimation of both magnitude and phase, due to inaccuracies or round-off errors in the estimation of the gradient. However, for the two-step approach, the regularized low-pass filter inversion introduced some magnitude errors at very high frequencies but preserved the phase. As the RMS error is more sensitive to errors in phase than in magnitude, this could explain why the RMS error in the two-step approach was much lower than in the direct approach. The lower $\mathrm{RMS_n}$ value, along with a significantly reduced computational load, favors the two-step approach over the direct approach. However, the direct approach has the advantage that it does not require any tuning of parameters, such as G_{max} in (15). This may be advantageous when the original signal is unknown and highly non-stationary, such that the stability of the reconstruction cannot be easily assessed. In addition, there is a monotonic relationship between spectral convergence, (25), and the objective function for the direct approach, (9). Since the two-step approach achieves a lower spectral convergence than the direct approach, this means that it finds a lower minimum than that found by the direct approach objective function. Hence, in optimization terms, the two-step approach appears to be better conditioned.

In [17], the authors presented a technique to recover a time-domain signal from the half-wave rectified output of a filterbank. This is quite similar to what the two-step approach of the present algorithm does after inversion of the low-pass filtering. Interestingly, when the low-pass filtering inversion was bypassed and a "perfect" half-wave rectified filterbank output was used as a target in (16), the original time-domain signal could be recreated perfectly (i.e., with a $\mathrm{RMS}_{\mathrm{n}}$ on the order of the quantization error). This suggests that the main limiting step in the framework we present is the low-pass filter inversion.

E. Application to Inconsistent Spectrograms

Most spectrogram inversion methods, in practice, are used on *inconsistent* spectrograms, i.e., two-dimensional representations that were not directly obtained from (3). These could be the results of processing applied to spectrograms, spectrograms resulting from a binary mask (e.g., for source separation), a grayscale image, or an arbitrary two-dimensional set of coefficients. In such cases, the performance of any algorithm will be very case dependent and, to our knowledge, there is no standard test for the performance of inconsistent spectrogram inversion. Thus, we will limit ourselves to a qualitative discussion. From our experience, the method proposed here achieves similar results as the Griffin & Lim algorithm [4] for processed traditional spectrograms. A more interesting result is observed for inconsistent auditory spectrograms, which appear to be particularly resilient to modification. We have observed that inverting an auditory spectrogram that had been modified in a "naive" way would often yield a signal whose spectrogram would be close to the *unmodified*, original spectrogram. We believe this to be caused by the very high degree of redundancy in auditory spectrograms, more than a flaw in our method. As opposed to "traditional" spectrograms, auditory spectrograms retain fine-structure information in low frequency channels where the center frequency is below the cutoff frequency of the low-pass filter involved in the envelope extraction. This results in auditory spectrograms having a complex, intricate structure, particularly in low-frequency regions. Any naive modification, even if very subtle, will yield a "very" inconsistent spectrogram for which the closest consistent spectrogram will be the original, unmodified one. Therefore, much more care must be taken when processing an auditory spectrogram. Its particular redundancy has to be taken into account in order for the processing to be successful.

F. Implications for Current IHC Models

The method based on IHC envelope representation was capable of reconstructing a time domain signal with high accuracy. This has implications with regard to modeling human auditory processing. While the details vary across current models of auditory processing, they all involve envelope extraction applied to the output of a filterbank. Because of this envelope extraction, it is often assumed that, for high-frequency channels, information regarding the temporal fine structure (i.e., the high frequency carrier fluctuations that are amplitude modulated by the envelope) is lost. However, the reconstruction method presented here suggests that this information could be recovered by processing envelopes across frequency channels. This interpretation is consistent with results from [32]. In [32], the authors provided a theoretical framework for evaluating the neural basis for the perceptual salience of acoustic temporal fine structure and envelope cues. In their framework, temporal fine structure (carrier) information could be retrieved from (across-frequency) envelope information.

VII. Conclusion

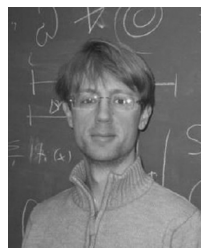
An optimization-based approach using gradient based methods in the time domain to reconstruct signals from multi-channel envelope representations was suggested. This approach offers two main advantages over traditional methods: it can be applied to a wider range of spectrogram classes and it offers more control to improve reconstruction accuracy. Successful implementations to invert both auditory and traditional spectrograms were presented. Successful recovery of time-domain information for auditory spectrograms suggests that temporal fine structure information assumed to be lost during IHC transduction can be recovered through across channel processing. For STFT magnitude envelope representations, the proposed method outperformed the algorithm of Griffin and Lim [4] in a standard STFT configuration. An analysis of the results suggested that this approach reduced the intrinsic limitations usually encountered when performing traditional spectrogram inversion.

References

[1] R. Balan, P. Casazza, and D. Edidin, "On signal reconstruction without phase," *Appl. Comput. Harmon. A.*, vol. 20, no. 3, pp. 345–356, 2006.
[2] M. Slaney, D. Naar, and R. Lyon, "Auditory model inversion for sound separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1994, vol. 2, pp. 77–80.
[3] M. Hayes, J. Lim, and A. Oppenheim, "Signal reconstruction from phase or magnitude," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 6, pp. 672–680, Dec. 1980.
[4] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 2, pp. 236–243, Apr. 1984.
[5] X. Zhu, G. T. Beauregard, and L. Wyse, "Real-time signal estimation from modified short-time Fourier transform magnitude spectra," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 5, pp. 1645–1653, Jul. 2007.
[6] J. Le Roux, H. Kameoka, N. Ono, and S. Sagayama, "Fast signal reconstruction from magnitude STFT spectrogram based on spectrogram consistency," in *Proc. Int. Conf. Digital Audio Effects DAFx*, 2010, vol. 10, pp. 397–403.
[7] J. Bouvrie and T. Ezzat, "An incremental algorithm for signal reconstruction from short-time Fourier transform magnitude," in *Proc. 9th Int. Conf. Spoken Lang. Process.*, 2006.
[8] R. Balan, "On signal reconstruction from its spectrogram," in *Proc. Conf. Inf. Sci. Syst.*, 2010, pp. 1–4.
[9] D. Sun and J. Smith, III, "Estimating a signal from a magnitude spectrogram via convex optimization," in *Proc. 133rd Conv. Audio Eng. Soc*, 2012.
[10] N. Sturmel and L. Daudet, "Signal reconstruction from STFT magnitude: A state of the art," in *Proc. Int. Conf. Digital Audio Effects DAFx*, 2011, 2012, pp. 375–386.
[11] T. Dau, D. Püschel, and A. Kohlrausch, "A quantitative model of the effective signal processing in the auditory system. I. Model structure," *J. Acoust. Soc. Amer.*, vol. 99, no. 6, pp. 3615–3622, 1996.
[12] R. Shannon, F. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, vol. 270, pp. 303–304, 1995.
[13] Z. Smith, B. Delgutte, and A. Oxenham, "Chimaeric sounds reveal dichotomies in auditory perception," *Nature*, vol. 416, no. 6876, pp. 87–90, 2002.

[14] H. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *J. Acoust. Soc. Amer.*, vol. 67, pp. 318–326, 1980.
[15] S. Jørgensen and T. Dau, "Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing," *J. Acoust. Soc. Amer.*, vol. 130, pp. 1475–1487, 2011.
[16] T. Chi, P. Ru, and S. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *J. Acoust. Soc. Amer.*, vol. 118, pp. 887–906, 2005.
[17] M. Slaney, "Pattern playback from 1950to 1995," in *Proc. IEEE Int. Conf. Syst., Man Cybern.*, 1995, vol. 4, pp. 3519–3524.
[18] R. Decorsière, P. L. Søndergaard, J. Buchholz, and T. Dau, "Modulation filtering using an optimization approach to spectrogram reconstruction," in *Proc. Forum Acusticum*, 2011.
[19] J. L. Flanagan, D. Meinhart, R. Golden, and M. Sondhi, "Phase vocoder," *J. Acoust.l Soc. Amer.*, vol. 38, p. 939, 1965.
[20] D. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," *Math. Program.*, vol. 45, no. 1, pp. 503–528, 1989.
[21] J. Le Roux, N. Ono, and S. Sagayama, "Explicit consistency constraints for STFT spectrograms and their application to phase reconstruction," in *Proc. Statist. Percept. Audition*, 2008, pp. 23–28.
[22] R. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," *APU Rep.*, vol. 2341, 1988.
[23] B. R. Glasberg and B. C. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hear. Res.*, vol. 47, no. 1, pp. 103–138, 1990.
[24] S. S. Stevens, "On the psychophysical law," *Psychol. Rev.*, vol. 64, no. 3, p. 153, 1957.
[25] E. Zwicker and B. Scharf, "A model of loudness summation," *Psychol. Rev.*, vol. 72, no. 1, p. 3, 1965.
[26] B. C. Moore and B. R. Glasberg, "A revision of Zwicker's loudness model," *Acta Acust. United Acust.*, vol. 82, no. 2, pp. 335–345, 1996.
[27] M. Schmidt, "L-BFGS algorithm implementation," [Online]. Available: http://www.di.ens.fr/~mschmidt/Software/minFunc.html
[28] "Speech recognition and identification materials. disc 4.0 (cd)," Dept. of Veterans Affairs Medical Center, 2006, Dept. of Veterans Affairs.
[29] W. Lindemann, "Extension of a binaural cross-correlation model by contralateral inhibition. i. simulation of lateralization for stationary signals," *J. Acoust. Soc. Amer.*, vol. 80, p. 1608, 1986.
[30] J. Breebaart, S. Van De Par, and A. Kohlrausch, "Binaural processing model based on contralateral inhibition. I. Model structure," *J. Acoust. Soc. Amer.*, vol. 110, p. 1074, 2001.
[31] J. Fienup and C. Wackerman, "Phase-retrieval stagnation problems and solutions," *J. Opt. Soc. Amer. A*, vol. 3, no. 11, pp. 1897–1907, 1986.
[32] M. G. Heinz and J. Swaminathan, "Quantifying envelope and fine-structure coding in auditory nerve responses to chimaeric speech," *JARO - J. Assoc. Res. Oto.*, vol. 10, no. 3, pp. 407–423, 2009.

Rémi Decorsière received the general engineering degree from the Ecole Centrale in Nantes, France, in 2009 and the M.Sc. and the Ph.D. degrees from the Technical University of Denmark in 2009 and 2013. He is now a Post-Doctoral Fellow at the Centre for Applied Hearing Research at the Technical University of Denmark. His research interests include time–frequency analysis, envelope processing and speech perception.



Peter L. Søndergaard received the M.Sc., and Ph.D. degrees from the Technical University of Denmark in 2002 and 2007. From 2007 to 2012, he was with the Department of Electrical Engineering at the Technical University of Denmark as a Post-Doctoral Fellow, and Associate Professor. From 2012 to 2013, he was the leader of the "Mathematics and Signal Processing in Acoustics" group at the Acoustics Research Institute, Austrian Academy of Sciences. Since 2013, he has been a Senior R&D Engineer with Oticon A/S in Denmark. His research interests are in mathematics for signal processing and acoustics.

Ewen N. MacDonald received the B.A.Sc. (Hons.), M.A.Sc., and Ph.D. degrees from the University of Toronto in 1999, 2002, and 2007, respectively. Since 2011, he has been with the Department of Electrical Engineering at the Technical University of Denmark where he is currently an Associate Professor. His research interests include hearing, speech perception, and speech production.

Torsten Dau received the M.Sc. degree from the University of Göttingen, Germany, in 1992 and the doctoral degree in physics (Dr.rer.nat.) and habilitation degree in applied physics (Dr.rer.nat.habil.) from the university of Oldenburg, Germany, in 1996 and 2003, respectively. Since 2003, he has been with the Department of Electrical Engineering at the Technical University of Denmark, where he is currently Professor, head of the Hearing Systems, Speech, and Communication group and head of the Centre for Applied Hearing Research. His research interests include auditory perception, signal processing, systems neuroscience, and hearing impairment.