MODIFYING SIGNALS IN TRANSFORM DOMAIN: A FRAME-BASED INVERSE PROBLEM.

Roswitha Bammer *

NuHAG, Faculty of Mathematics,
University of Vienna
Oskar-Morgenstern-Platz 1
1090 Vienna, Austria
roswitha.bammer@univie.ac.at

Monika Dörfler *

NuHAG, Faculty of Mathematics,
University of Vienna
Oskar-Morgenstern-Platz 1
1090 Vienna, Austria
monika.doerfler@univie.ac.at

ABSTRACT

Within this paper a method for morphing audio signals is presented. The theory is based on general frames and the modification of the signals is done via frame multiplier. Searching this frame multiplier with given input and output signal, an inverse problem occurs and a priori information is added with regularization terms. A closed-form solution is obtained by a diagonal approximation, i.e. using only the diagonal entries in the signal transformations. The proposed solutions for different regularization terms are applied to Gabor frames and to the constant-Q transform, based on non-stationary Gabor frames.

1. INTRODUCTION AND MOTIVATION

What does it mean to convert one signal into another? In this paper a sound-signal modification is performed by morphing one sound into another, i.e. it is assumed that there exist sounds "in between" two given, distinct sounds. This morphing enables to interpolate between two sounds with sufficient similarity, i.e. in the case of instrument morphing, the same fundamental frequency.
Existing methods are based on parametric models based on parameter interpolation [1, 2].Our method allows to observe the modification necessary for morphing directly in the time-frequency domain. In our task the input and output signals are given and the transfer function which is modeled as a frame multiplier has to be estimated. Hence, the preferred output is given and we would like to compute the cause for this output. Reformulating the problem into a minimization of a functional, the estimation is transformed into a linear inverse problem. In order to add some a priori information to the minimization problem, we add regularization terms. Such an inverse problem normally can be solved by iterative shrinkage methods [3, 4] among others, because otherwise a huge matrix system must be inverted. One possible simplification, to gain a better understanding and to obtain a closed-form solution of satisfactory quality, is to perform a diagonal approximation, i.e. considering only the diagonal entries of the matrix from the signal transformations. Stating the exact solution for several regularization terms is the main result of this paper and can be found in Theorem 3.1 in Section 3.1.
Moreover we perform some numerical experiments using the obtained solutions in MATLAB. In these experiments Gabor frames and non-stationary Gabor frames leading to a constant-Q transform, as described in Section 4, are considered. Additional experiments as well as the MATLAB code and corresponding sounds can be found on the website [5]. This paper is a generalization of the first author's master thesis [6]. The basic principles have been developed, in the context of Gabor multipliers, in [7, 8, 9, 10].

* This work was supported by the Vienna Science and Technology Fund (WWTF) project SALSA (MA14-018).

2. BASICS

Frames generalize the concept of a basis, in the sense that the frame functions need not be linearly independent. The resulting *redundancy* leads to increased stability against noise or data loss. In the following we consider a general Hilbert space $\mathcal{H}$, for example $L^2(\mathbb{R}^d)$ or $\mathbb{C}^L$.

Definition 2.1 ([11]). **(Frame, Frame Bounds, Tight Frame)** *A sequence $\{e_j : j \in J\} \subseteq \mathcal{H}$ is called a frame if there exist $A, B > 0$ such that $\forall f \in \mathcal{H}$*

$$A\|f\|^2 \leq \sum_{j\in J} |\langle f, e_j\rangle|^2 \leq B\|f\|^2. \tag{1}$$

Any two constants A, B satisfying equation (1) are called frame bounds. If $A = B$, then we call $\{e_j : j \in J\}$ a tight frame.

If $\mathcal{H} = \mathbb{C}^L$, the coefficient space is also finite dimensional, i.e. $|J| = K < \infty$.
The following important operators included in a signal processing procedure will help to develop the theory of our problem.

Definition 2.2 ([11]). **(Analysis-, Synthesis- and Frame operator)** *Let $\{e_j : j \in J\}$ be a sequence in a Hilbert space $\mathcal{H}$ and $f \in \mathcal{H}$, then the coefficient operator or analysis operator $\mathcal{T} : \mathcal{H} \to \ell^2(J)$ is defined as*

$$(\mathcal{T}f)_j = \langle f, e_j\rangle = c_j, \qquad j \in J \tag{2}$$

The adjoint of the analysis operator $\mathcal{T}^* : \ell^2(J) \to \mathcal{H}$ is the synthesis operator or reconstruction operator and is defined for a finite sequence $\tilde{c} = (\tilde{c}_j)_{j\in J} \in \ell^2(J)$ by

$$\mathcal{T}^*\tilde{c} = \sum_{j\in J} \tilde{c}_j e_j \in \mathcal{H}. \tag{3}$$

Combining these two operators leads to the definition of the frame operator $S : \mathcal{H} \to \mathcal{H}$

$$Sf = \mathcal{T}^*\mathcal{T}f = \sum_{j\in J} \langle f, e_j\rangle e_j. \tag{4}$$

In the following proposition we introduce dual frames which yield a reconstruction formula.

Proposition 2.3 ([11]). **(Dual frame)**
If $\{e_j : j \in J\}$ is a frame with frame bounds $A, B > 0$, then $\{S^{-1}e_j : j \in J\}$ is a frame with frame bounds $B^{-1}, A^{-1} > 0$, the so-called dual frame. Every $f \in \mathcal{H}$ has non-orthogonal expansions

$$f = \sum_{j\in J} \langle f, S^{-1}e_j\rangle e_j = \sum_{j\in J} \langle f, e_j\rangle S^{-1}e_j,$$

where both sums converge unconditionally in $\mathcal{H}$.

A signal processing step between the analysis and the synthesis operator in Definition 2.2, where the coefficients are multiplied by weights $w_j, j \in J$, can be performed. Thus

$$\tilde{c_j} = w_j \cdot c_j.$$

This leads to the the following definition:

Definition 2.4 ([12]). **(Frame multiplier)**
Let $\mathcal{H}_1, \mathcal{H}_2$ be Hilbert spaces, let $(g_j)_{j\in J} \subseteq \mathcal{H}_1$ and $(\gamma_j)_{j\in J} \subseteq \mathcal{H}_2$ be frames. Fix a sequence $m = (m_j)_{j\in J} \in l^\infty$, then we define the frame multiplier

$$\mathbb{M}_{m;g,\gamma} : \mathcal{H}_1 \to \mathcal{H}_2$$

for the frames (g_j) and (γ_j), as

$$\mathbb{M}_{m;g,\gamma}(f) = \sum_j m_j \langle f, g_j\rangle \gamma_j.$$

The sequence $(m_j)_{j\in J}$ mentioned in this definition is called the *symbol mask* of $\mathbb{M}$ and can be interpreted as a *time-frequency transfer function.*
In the following section, we are going to introduce the inverse problem which leads to the estimation of a set of masks, in dependence on a regularization parameter λ, for the controlled modification of one given sound towards a given target sound.

3. ESTIMATION OF THE FRAME MULTIPLIER

In this section, we consider a normed tight frame (i.e. $A = B = 1$) $(g_j)_{j\in J} \subseteq \mathcal{H}$. Assume the input and output signal $s, z \in \mathcal{H}$ to be given, hence the relation

$$z = \mathbb{M}_{m;g}s$$

to be valid. Now we want to identify the linear system, where the system is treated as a frame multiplier. Let $\mathcal{T}$ and $\mathcal{T}^*$ be fixed, we can reformulate optimality as the minimization of a functional and its estimation can therefore be transformed into a linear inverse problem:

$$\tilde{m} = \arg\min_{\mathrm{m}} \|z - \mathcal{T}_g^* m \mathcal{T}_g s\|_2^2.$$

To gain more stability in order to solve the inverse problem, we add a regularization term. We therefore have to minimize the expression

$$\Phi(m) = \|z - \mathcal{T}_g^* m \mathcal{T}_g s\|_2^2 + \lambda r(m), \tag{5}$$

where $r(m) : \ell^\infty \to \mathbb{R}^+$ is a regularization term and $\lambda \in \mathbb{R}^+$ is a regularization parameter. The choice of regularization term is discussed in the next section,Theorem 3.1.

3.1. Diagonal Approximation

Since the frames used in analysis and synthesis usually lack orthogonality iterative methods need to be employed to obtain an exact solution of (5), cp. [3]. In the special case of Gabor frames it has been shown [7, 8] that an approximate solution can be achieved by reducing the term $\mathcal{T}_g^* m \mathcal{T}_g s$ to its diagonal entries. We will address a different example, namely non-stationary Gabor frames leading to a constant-Q transform. The diagonal case brings us to closed-form solutions. These solutions lead to satisfactory quality for example in experiments on audio signals as has been observed in the experimental Section 4.
Using these solutions, iterative algorithms can be applied to achieve exact solutions of equation 5. However the difference in perception is marginal, but the computational effort increases.
For some further information we refer to [13], [7], [8] and [14]. In order to achieve a diagonal approximation, we reformulate (5) in the transform domain

$$\Phi(m) = \|\mathcal{T}_g^*(\mathcal{T}_g z - m\mathcal{T}_g s)\|_2^2 + \lambda r(m).$$

Reducing to the diagonal and writing $S = \mathcal{T}_g s$ and $Z = \mathcal{T}_g z$, leads to

$$\Phi(m) = \|Z - m \cdot S\|_2^2 + \lambda r(m). \tag{6}$$

If the source equals the target, the mask m should be equal to 1 and the regularization term should vanish. This motivates the choice of regularization terms with entries $m - \vec{1}$. If Z and S are different, we can use different terms of regularization. The regularization term helps us to indicate some a priori information in the shape of the solution (the transformed signal). The choice of $r(m)$ is discussed in Remark 3.2. The parameter λ helps balancing between these a priori information of the form and the properties of reconstructing the mask [7]. We are now going to present different choices of regularization terms by stating the following theorem.

Theorem 3.1. Let $\Phi : \mathbb{C}^L \to \mathbb{R}$ be a functional of the form

$$\Phi(m) = \|Z - m \cdot S\|_2^2 + \lambda r(m), \tag{7}$$

where $\lambda \in \mathbb{R}^+$ and $r : \mathbb{C}^L \to \mathbb{R}$ is a regularization term. Minimizing this functional for different solutions with respect to different regularization terms as follows:

a) $r(m) = \|m - 1\|_2^2$ leads to the solution

$$\tilde{m_\ell} = \frac{\overline{S_\ell} Z_\ell + \lambda}{|S_\ell|^2 + \lambda} \qquad \forall \ell \in \{0, ..., L\}.$$

b) $r(m) = \| \, |m| - 1\|_2^2$ leads to the solution

$$\tilde{m_\ell} = \frac{|Z_\ell S_\ell| + \lambda}{|S_\ell|^2 + \lambda} \cdot e^{i \arg(S_\ell \overline{Z_\ell})} \qquad \forall \ell \in \{0, ..., L\}.$$

c) $r(m) = \|m - 1\|_1$ leads to the solution

$$\tilde{m_\ell} = \begin{cases} \frac{|S_\ell||Z_\ell - S_\ell| - \frac{\lambda}{2}}{|S_\ell|^2} \cdot e^{i\varphi} + 1 & \text{if } |S_\ell|\,|T_\ell - S_\ell| > \frac{\lambda}{2} \\ 1 & \text{else} \end{cases},$$

where $\varphi = arg(\overline{S_\ell}(Z_\ell - S_\ell)) \quad \forall \ell \in \{0, ..., L\}$.

d) $r(m) = \| \, |m| - 1\|_1$ leads to the solution

$$\tilde{m_\ell} = \begin{cases} \frac{|Z_\ell S_\ell| - \frac{\lambda}{2}}{|S_\ell|^2} e^{i \arg(S_\ell \overline{Z_\ell})} & \text{if } \frac{|Z_\ell S_\ell|}{|S_\ell|^2} > 1 + \frac{\lambda}{2|S_\ell|^2} \\ \frac{|Z_\ell S_\ell| + \frac{\lambda}{2}}{|S_\ell|^2} e^{i \arg(S_\ell \overline{Z_\ell})} & \text{if } \frac{|Z_\ell S_\ell|}{|S_\ell|^2} < 1 - \frac{\lambda}{2|S_\ell|^2} \\ 1 & \text{else} \end{cases}$$

$\forall \ell \in \{0, ..., L\}$.

Note that some of these solution formulas can be found for the case of Gabor multipliers in [7, 8]. Since we found them to be useful in the general case of frame multipliers [12], we include their proof in the appendix, Section 7.

Remark 3.2. *Let $\Phi : \mathbb{C}^L \to \mathbb{R}$ be as in (7). Then the different regularization terms have the following properties:*

a) $r(m) = \|m-1\|_2^2$ helps to control the total energy. Moreover, if we use normed tight frame bounds, i.e. $A = B = 1$, we favor a multiplier close to the identity operator. This regularization term produces spurious oscillations in the mask $\tilde{m}$, caused by a bad estimation of the phase. A simple calculation shows the reason of the oscillations. Let (j,k) be a point of the time-frequency plane and let the input and the output signal have a phase difference of π, i.e. $Z = Se^{i\pi}$. Then $\tilde{m}$ at the point (j,k) is given by

$$\tilde{m} = \frac{\overline{S}Z + \lambda}{|S|^2 + \lambda} = \frac{|S|^2 e^{i\pi} + \lambda}{|S|^2 + \lambda}.$$

This short calculation shows the presence of amplitude modulations of the mask due to the diagonal approximation, cp. [7, p. 43 et seq.].

b) $r(m) = \| |m| - 1\|_2^2$ gives us the possibility of avoiding spurious oscillations of the amplitude of $\tilde{m}$, apart from that fact it has the same properties as the previous regularization term in a).

c) $r(m) = \|m-1\|_1$ yields sparsity, where the mask is forced to stay close to 1 which corresponds to "no transformation". This regularization term also produces spurious oscillations.

d) $r(m) = \| |m| - 1\|_1$ forces $\tilde{m}$ to sparsity of the deviation from the absolute value 1 and also avoids the oscillations of the previous regularization term in c). For some more information on this regularization term consider [14].

In the next section we will visualize these properties by analyzing examples for diagonal approximation with different regularization terms.

4. NUMERICAL EXPERIMENTS

4.1. Two examples of frames used in audio processing

The results in Section 3.1 hold for general frames and in particular also in higher dimensions, that is, for frames for $L^2(\mathbb{R}^d)$ with $d > 1$. This can be interesting for image or video processing. In the current work, however, we focus on audio signals and present some numerical simulations using classical Gabor frames [11] on the one hand and a constant-Q transform based on non-stationary time-frequency Gabor frames, [15, 16] on the other hand. We briefly introduce the necessary notions next.

The frame elements of Gabor frame are given by time-frequency shifted versions of a non-zero window function $g \in \mathcal{H}$, i.e. $\mathcal{G}(g,a,b) = \{g_{n,k} = T_{ak}M_{bn}g : k,n, \in \mathbb{Z}\}$. Here, $T_{ak}g(t) = g(t-ak)$ and $M_{bn}g(t) = g(t)\cdot e^{2\pi ibn}$ denote time- and frequency shift, respectively.

For the non-stationary Gabor frame-based constant-Q transform, the construction of Gabor frames is generalized as to allow for windows with adaptive bandwidth. To this end, the frame elements are given by $\{g_{n,k} = T_{na_k}g_k : k,n, \in \mathbb{Z}\}$. Thus, while the time-shifts are carried out along a regular lattice as in the Gabor case, the frequency shifts are replaced by choosing a separate window for each desired frequency band. Accordingly, the time-shift parameter a_k can be chosen separately for each band. For all details, in particular regarding the precise choice of parameters for the constant-Q transform, we refer to [15, 16]. We note that for both Gabor frames and and non-stationary Gabor frames, careful choice of windows and sampling parameters a, b leads to the situation of *painless non-orthogonal expansions*, [17], for which straight-forward inversion is possible.

Implementations along with excellent documentation for both Gabor frames and the constant-Q transform can be found in the LTFAT-toolbox, [18, 19] and [16].

4.2. Experimental setup

In this section we describe the setup for the subsequent numerical experiments. All mentioned MATLAB routines are found on the website [5]. We want to find the best fitting multiplier of the inverse problem (5) using different regularization terms introduced in Theorem 3.1. To do so, we use an input and an output signal with sufficient similarity. In the following two experiments we use the sound of a flute and a violin of the VSL [20] playing the same fundamental frequency and vowels sang by a man [21] and a woman [22], also using the same fundamental frequency for sufficient similarity. The sound files are sampled with a rate of 44100 Hertz. To make the sound files the same length, we use the MATLAB code `samesize_power2.m` which fades out the signals with exponential decrease. Moreover, to display the spectrogram of our sounds we use a logarithmic scale, cp. [6, p.24].

In order to show that different classes of frames can be used, we will consider Gabor frames and non-stationary Gabor frames as introduced in Section 4.1, within the numerical examples. The MATLAB code `diagaprox.m` is used to do the numerical calculations with Gabor frames by using the closed-form solutions stated in Theorem 3.1 and the code `diagaprox_cq.m` does this by using a constant-Q transform, based on non-stationary Gabor frames. Note that in the finite discrete case underlying the numerical implementations $\mathcal{H} = \mathbb{C}^L$.

The algorithm basically uses the following steps:

- Input: s (source signal), z (target signal) of the same length (here 1 second) due to `samesize_power2.m`, a preferred norm and λ.
- Tranformation done with:
  - Gabor transform [`dgt.m`] of $s \to S$ and $z \to Z$ with a Hann-window and the parameters $a = 256, M = \frac{L}{b} = 1024$.
  - Constant-Q transform [`cqt.m`] of $s \to S$ and $z \to Z$ with frequency region $[100\text{Hz}; 22050\text{Hz}]$, 64 bin per octave and 1000 time channels per second.
- Obtain the mask m as in Theorem 3.1 corresponding to the respective norm used for regularization.
- Inverse transform [`idgt.m`] or [`icqt.m`], respectively, of $m * S$ to obtain a $\tilde{z}$, the target signal.
- Output: m and $\tilde{z}$.

4.3. Sound morphing

4.3.1. Using musical instruments

For the first experiment we use the sound of a flute and a violin from the VSL [20]. Since sufficient similarity is required, we con-

sider the same fundamental fequency of the instruments for each morphing procedure. Morphed sounds can be obtained by varying λ in Formula (5). A high value of λ puts heavy weight on the regularization term $r(m)$, hence forces the mask to be close to one, i.e. "no transformation" and the signal reconstructed from $\tilde{m} * S$ is similar to the source/input signal. A small λ does not take the regularization term that much into account. This leads to a mask which yield a reconstructed signal close to the target signal.
The choice of these two instruments is due to their different timbres and harmonics. The violins sound is rich in overtones, whereas the flute has less overtones.
The following experiment is done with constant-Q transform. Similar results can be achieved using the Gabor transform, cp. [5]. In Figure 1 we show stepwise morphing, i.e. using different λ $(= 10^{-2}, 10^{-6})$ starting from the original flute sound as source going to the violin as target. This case is very interesting, because it is more difficult for the mask to 'generate' overtones, since the violin has more overtones than the flute, than suppressing them, as it would be the other way round. As common fundamental frequency we use B5; the original sounds and sounds resulting from morphing steps can be found online [5].
In Figure 1 one can see how the noise, coming from the violin increases from step to step. For $\lambda = 10^{-2}$ the sound is a mixture of flute and violin, but for $\lambda = 10^{-6}$ we can verify the sound as a violin.

4.3.2. Using spoken vowels

The second experiment considers German vowels, sung by a professional singer. We use a female [22] as well as a male [21] voice. Both have the fundamental frequency E4. This morphing task is interesting because vowels build different formants, i.e. acoustic resonance of the human vocal tract, where certain harmonics are stronger than others. Within this experiment it is visible that similar vowels, like the german spoken "e" and "i" which sound very similar, also morph with comparably bigger λ into each other. Several tables summarizing which λ has to be used to get reconstruction of the target signal can be found in [5]. In Figure 2, we perform again a morphing procedure using constant-Q transform. The morphing is performed stepwise starting from the vowel "a", with steps in between with $\lambda = 10^{-4}$ and $\lambda = 10^{-8}$, reaching target vowel "i". The noise level goes down in the range between $[600\text{Hz}, 2000\text{Hz}]$. Hence the harmonics are better visible and focus on 300Hz which is the characteristic formant for the vowel "i" [23].
Another thing that can be observed, concerning the mask of the morphing steps are spurious oscillations which are mentioned in Remark 3.2 a). Since the observation is almost only visible using the Gabor transform, we use this transformation to generate Figure 3. Nevertheless there was no audible difference within the morphing experiments. The morphing is, as mentioned above, performed going from vowel "a" to "i" and to show the oscillations, we take the mask corresponding to $\lambda = 10^{-2}$. In Figure 3 the upper image shows the mask obtained with the regularization term $r(m) = \|m - 1\|_2^2$. Here oscillations are visible between 3000Hz and 4000Hz. The lower image neglects the phase by using the modulus of the mask, i.e. $r(m) = \| \, |m| - 1\|_2^2$, hence no oscillations are visible.

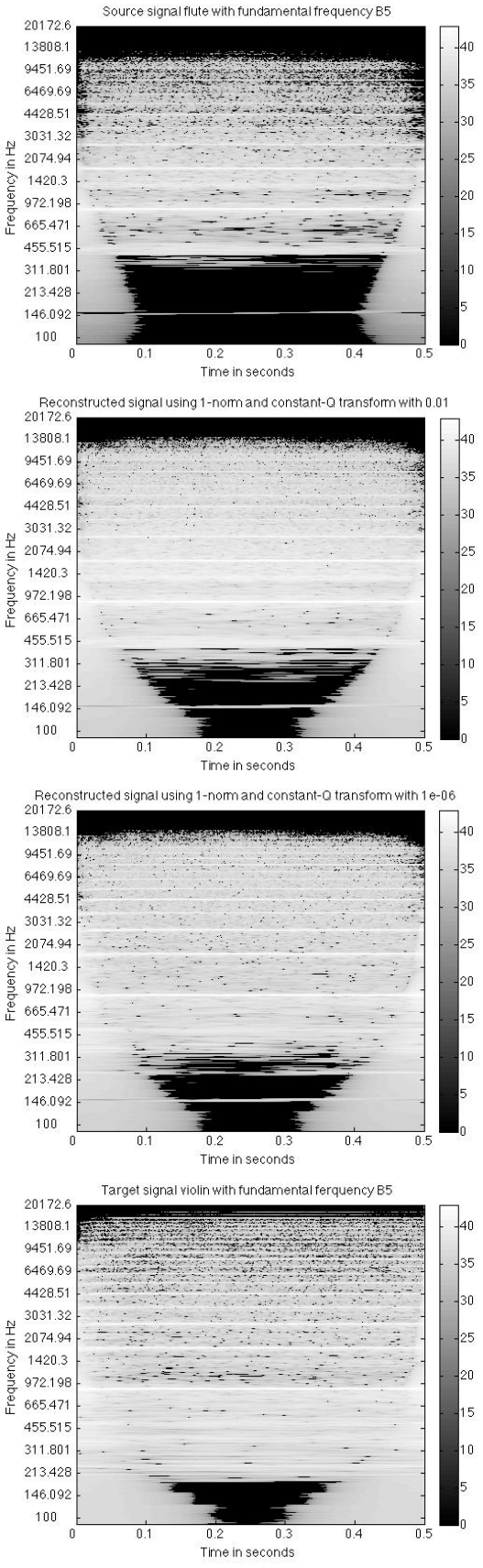

Figure 1: Stepwise morphing from flute to violin, with steps in between at $\lambda = 10^{-2}$ and $\lambda = 10^{-6}$.

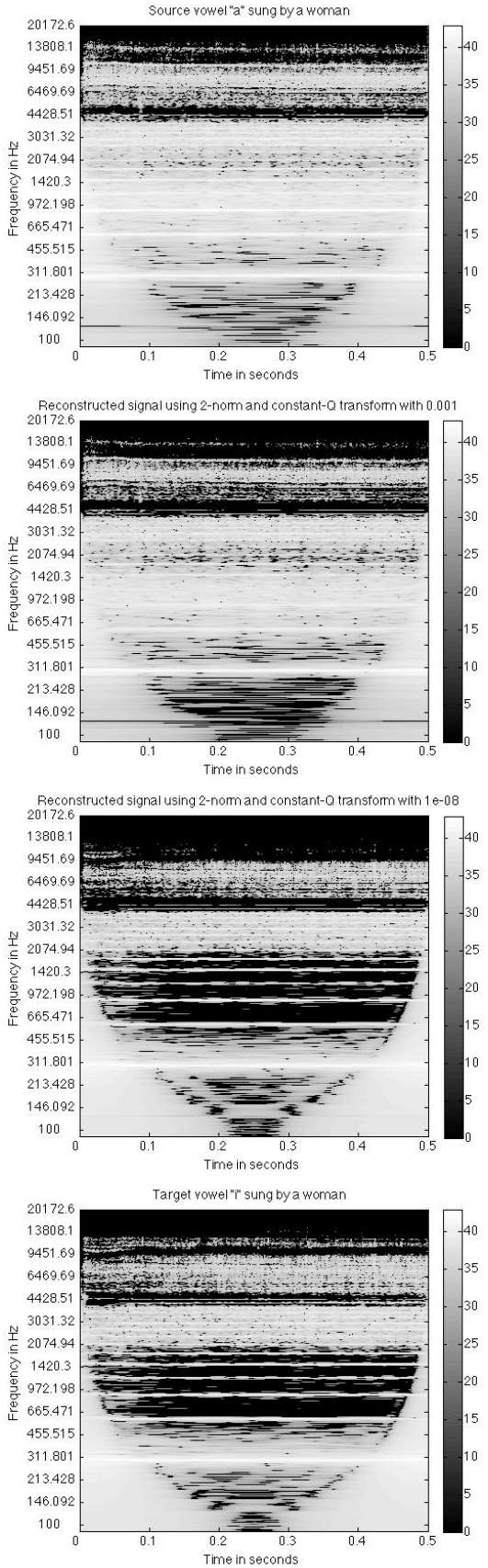

Figure 2: Stepwise morphing from vowel "a" to "i", with steps in between at $\lambda = 10^{-4}$ and $\lambda = 10^{-8}$.

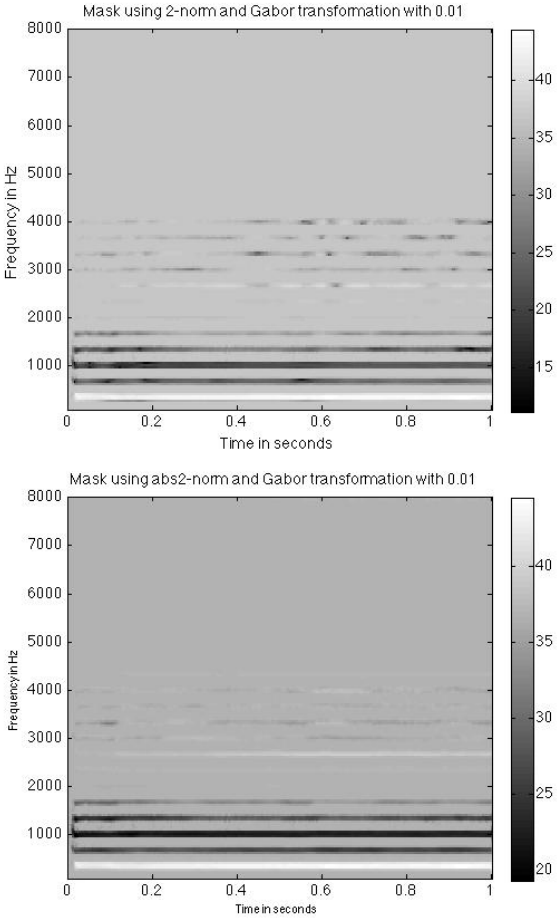

Figure 3: Spurious oscillations for $\lambda = 10^{-2}$, obtained by morphing vowel "a" into "i".

5. CONCLUSION AND PERSPECTIVES

In this paper a morphing procedure was proposed, using frame multipliers in order to morph one audio sound into another. A diagonal approximation was performed in order to get a closed form solution of a regularized inverse problem.
Comparing the solution of the diagonal approximation with the solution computed by iterative shrinkage threshold algorithms (ISTA) yield only a marginal difference. For the monotone fast ISTA the solution was better audible. Here it would be interesting to figure out, why this is the case. Nevertheless the computational effort increases strongly. The obtained solutions were used in the experimental chapter, to show that this approximation also leads to satisfactory perceptive quality. The general concept of frames allowed to state different examples. We focused on the usage of Gabor frames and non-stationary Gabor frame based constant-Q transform.
Extensions of the presented work will include the usage of other frames, for example wavelet frames in the context of image morphing and other non-stationary time-frequency frames. Furthermore, the class of coefficient priors will be extended to mixed-norm and neighborhood-based priors, [24, 25], which will lead to a structure-based signal modification.
Another strand of research will investigate, why spurious oscillations, mentioned in Remark 3.2 and rather prominent if the morphing is based on Gabor frames, are barely visible for constant-Q transform, while the perceptual difference seems marginal. To this end, we will set up an evaluation framework based on perceptual criteria, cp. [26], since comprehensive listening experiments are costly. In parallel, we will isolate the oscillations and use subsequent synthesis to understand their behaviour and consequence on perceptual outcome. Finally, the variety of sounds obtained by con- trolled morphing can be used for data augmentation, [27], within machine learning tasks for audio signals.

6. REFERENCES

[1] A. Chadha, B. Savardekar, and J. Padhya, "Analysis of a modern voice morphing approach using gaussian mixture models for laryngectomees," *CoRR*, vol. abs/1208.1418, 2012.

[2] M.F. Caetano and X. Rodet, "Musical instrument sound morphing guided by perceptually motivated features," *IEEE Trans. Audio, Speech & Language Processing*, vol. 21, no. 8, pp. 1666–1675, 2013.

[3] I. Daubechies, M. Defrise, and C. De Mol, "An Iterative Thresholding Algorithm for Linear Inverse Problems with a Sparsity Constraint," *Comm. Pure Apl. Math.* , vol. 57, no. 11, pp. 1413–1457, 2004.

[4] A. Beck and M. Teboulle, "A Fast Iterative Shrinkage-Threshold Algorithm for Linear Inverse Problems," *SIAM J.Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[5] R. Bammer and M. Dörfler, "Modifying Signals in Transform Domain," `http://projektservice-mathematik.univie.ac.at/morph`.

[6] R. Bammer, "Signaltransfomation via Gabor Multiplier," M.S. thesis, University of Vienna, 2015, `http://projektservice-mathematik.univie.ac.at/morph`.

[7] A. Olivero, *Les Multiplicateurs Temps-Fréquence. Application á l'Analyse et la Synthése de Signaux Sonores et Musicaux*, Ph.D. thesis, University of AIX-Marseille, 2012.

[8] A. Olivero and B. Torrésani and R. Kronland-Martinet, "A Class of Algorithms for Time-Frequency Multiplier Estimation.," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 8, pp. 1550–1559, 2013.

[9] A. Olivero and B. Torrésani and P. Depalle and R. Kronland-Martinet, "Sound morphing strategies based on alterations of time-frequency representations by Gabor multipliers," in *AES 45th International Conference on Applications of Time-Frequency Processing in Audio*, Helsinki, Finland, 2012, p. 17.

[10] A. Olivero and B. Torrésani and R. Kronland-Martinet, "A new method for Gabor multipliers estimation : application to sound morphing," in *European Signal Processing Conference (EUSIPCO 2010)*, Aalborg, Denmark, 2010, pp. 507–511.

[11] K. Gröchenig, *Foundations of Time-Frequency Analysis*, Appl. Numer. Harmon. Anal. Birkhäuser, 2001.

[12] P. Balazs, "Basic definition and properties of Bessel multipliers ," *Journal of Mathematical Analysis and Applications* , vol. 325, no. 1, pp. 571 – 585, 2007.

[13] P. Depalle and R. Kronland-Martinet and B. Torrésani, "Time-Frequency multipliers for sound synthesis," in *SPIE annual Symposium Wavelet XII*, San Diego, United States, 2007, vol. 6701, pp. 670118–1 – 670118–15.

[14] M. Dörfler and E. Matusiak, "Sparse Gabor multiplier estimation for identification of sound objects in texture sound.," in *Sound, Music, and Motion*, M. Aramaki et al., Ed., vol. LNCS 8905, pp. 443–462. Springer International Publishing Switzerland, 2014.

[15] G. A. Velasco, N. Holighaus, M. Dörfler, and T. Grill, "Constructing an invertible constant-Q transform with non-stationary Gabor frames," *Proceedings of DAFX11*, 2011.

[16] N. Holighaus, M. Dörfler, G. A. Velasco, and T. Grill, "A framework for invertible, real-time constant-Q transforms," *IEEE Trans. Audio Speech Lang. Process.*, vol. 21, no. 4, pp. 775 –785, 2013, `http://www.univie.ac.at/nonstatgab`.

[17] I. Daubechies, A. Grossmann, and Y. Meyer, "Painless nonorthogonal expansions," *J. Math. Phys.*, vol. 27, no. 5, pp. 1271–1283, May 1986.

[18] P. L. Søndergaard, B. Torrésani, and P. Balazs, "The Linear Time Frequency Analysis Toolbox," *International Journal of Wavelets, Multiresolution Analysis and Information Processing*, vol. 10, no. 4, 2012, `http://ltfat.sourceforge.net/`.

[19] Z. Průša, P. L. Søndergaard, N. Holighaus, C. Wiesmeyr, and P. Balazs, "The Large Time-Frequency Analysis Toolbox 2.0," in *Sound, Music, and Motion*, M. Aramaki et al., Ed., vol. LNCS 8905, pp. 419–442. Springer International Publishing Switzerland, 2014.

[20] "Vienna Symphonic Library: Vienna Super Package," 2014.

[21] E. Tarilonte, "Altus: The Voice of Renaissance," Software: `http://www.bestservice.de/altus.html`.

[22] E. Tarilonte, "Shevannai: The Voice of Elves," Software: `http://www.bestservice.de/shevannai.html`.

[23] K. Johnson, *Acoustic and auditory phonetics*, Blackwell Cambridge, Mass. a.o., 2003.

[24] B. Torrésani and M. Kowalski, "Sparsity and Persistence: mixed norms provide simple signal models with dependent coefficients," *Signal, Image and Video Processing*, to appear, 2008.

[25] M. Kowalski, K. Siedenburg, and M. Dörfler, "Social Sparsity! Neighborhood Systems Enrich Structured Shrinkage Operators," *IEEE Trans. Signal Process.*, vol. 61, no. 10, pp. 2498 – 2511, 2013.

[26] P. Kabal, *An examination and interpretation of itu- r bs. 1387: Perceptual evaluation of audio quality*, Ph.D. thesis, tech. rep., Dep. of Electrical Engineering and Com- puter Engineering, McGill University, 2003.

[27] J. Schlüter and T. Grill, "Exploring Data Augmentation for Improved Singing Voice Detection with Neural Networks," in *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR 2015)*, Malaga, Spain, 2015.

[28] L.V. Ahlfors, *Complex analysis : an introduction to the theory of analytic functions of one complex variable*, International series in pure and applied mathematics. McGraw-Hill, New York, 1979.

7. APPENDIX

Proof of Theorem 3.1. To find a minimum of (7), we have to find zeros of its first derivative. We first rewrite $\|Z - m \cdot S\|_2^2$ as the sum of its entries squared

$$\begin{aligned}\Phi(m) &= \sum_{\ell\in\Lambda} \Phi^\ell(m_\ell) \\ &= \sum_{\ell\in\Lambda} \big(|Z_\ell - m_\ell S_\ell|^2 + \lambda r(m_\ell)\big) \\ &= \sum_{\ell\in\Lambda} \big((Z_\ell - m_\ell S_\ell)\overline{(Z_\ell - m_\ell S_\ell)} + \lambda r(m_\ell)\big).\end{aligned}$$

Writing the complex vector m as $m = m^r + i\,m^i$, where $m^r, m^i \in \mathbb{R}^L$ we obtain

$$= \sum_{\ell\in\Lambda} \big((Z_\ell - S_\ell(m^r + i\,m^i))\overline{(Z_\ell - S_\ell(m^r + i\,m^i))} + \lambda r(m_\ell)\big).$$

The derivative of Φ can now be understood as a derivative in two variables. Using the formula

$$\frac{\partial\Phi(m)}{\partial m} = \frac{1}{2}\left(\frac{\partial\Phi(m^r, m^i)}{\partial m^r} - i\frac{\partial\Phi(m^r, m^i)}{\partial m^i}\right) \tag{8}$$

we obtain the derivative for holomorphic functions [28]. Next we fix one ℓ since, if we take the derivative component-wise, the other components will vanish. The derivative with respect to the first variable m_ℓ^r is

$$\begin{aligned}\frac{\partial\Phi^\ell(m_\ell^r, m_\ell^i)}{\partial m_\ell^r} =& - S_\ell\overline{(Z_\ell - S_\ell(m^r + i\,m^i))} \\ &+ (Z_\ell - S_\ell(m^r + i\,m^i))\overline{(-S_\ell)} + \lambda\frac{\partial r(m_\ell)}{\partial m_\ell^r}.\end{aligned}$$

Using $\frac{z+\overline{z}}{2} = \mathfrak{Re}(z)$ we get

$$= -2\mathfrak{Re}\big(S_\ell\overline{Z_\ell}\big) + 2\,|S_\ell|^2\, m_\ell^r + \lambda\frac{\partial r(m_\ell)}{\partial m_\ell^r}.$$

Similarly we compute the derivative with respect to m_ℓ^i

$$\frac{\partial\Phi^\ell(m_\ell^r, m_\ell^i)}{\partial m_\ell^i} = 2\cdot\mathfrak{Im}\big(S_\ell\overline{Z_\ell}\big) + 2\,|S_\ell|^2\, m_\ell^i + \lambda\frac{\partial r(m_\ell)}{\partial m_\ell^i}.$$

Using Equation (8) we obtain

$$\begin{aligned}\frac{\partial\Phi^\ell(m_\ell)}{\partial m_\ell} =& \frac{1}{2}\Big(- 2\mathfrak{Re}\big(S_\ell\overline{Z_\ell}\big) + 2\,|S_\ell|^2\, m_\ell^r + \lambda\frac{\partial r(m_\ell)}{\partial m_\ell^r} \\ &- 2i\mathfrak{Im}\big(S_\ell\overline{Z_\ell}\big) - 2i\,|S_\ell|^2\, m_\ell^i - i\lambda\frac{\partial r(m_\ell)}{\partial m_\ell^i}\Big).\end{aligned}$$

To obtain a minimum, we have to set the following equation to zero:

$$\frac{\partial\Phi^\ell(m_\ell)}{\partial m_\ell} = -S_\ell\overline{Z_\ell} + |S_\ell|^2\,\overline{m_\ell} + \underbrace{\frac{\lambda}{2}\Big(\frac{\partial r(m_\ell)}{\partial m_\ell^r} - i\frac{\partial r(m_\ell)}{\partial m_\ell^i}\Big)}_{\rho(m_\ell)} = 0. \tag{9}$$

Now we have to distinguish according to the different regularization terms.

a) Considering $r(m) = \|m-1\|_2^2$ as the first regularization term, we have

$$\rho(m_\ell) = \frac{\lambda}{2}(2m_\ell^r - 2 - 2im_\ell^i) = \lambda\overline{m_\ell} - \lambda.$$

Thus, solving Equation (9) with respect to $\overline{m_\ell}$ and taking the conjugate, we obtain for every ℓ

$$\tilde{m_\ell} = \frac{\overline{S_\ell}Z_\ell + \lambda}{|S_\ell|^2 + \lambda}.$$

b) For the regularization term $r(m) = \|\,|m| - 1\|_2^2$ we have

$$\rho(m_\ell) = \frac{\lambda}{2}\left(2\lambda m_\ell^r - 2\frac{\lambda m_\ell^r}{|m_\ell|} - 2\lambda m_\ell^i + 2\frac{\lambda m_\ell^i}{|m_\ell|}\right) = \lambda\overline{m_\ell} - \frac{\lambda\overline{m_\ell}}{|m_\ell|}$$

If we plug this term into Formula (9), we obtain

$$\overline{m_\ell}\big(\,|S_\ell|^2 + \lambda - \frac{\lambda}{|m_\ell|}\big) = S_\ell\overline{Z_\ell}. \tag{10}$$

Since there is a term containing $|m_\ell|$ in the brackets of this solution, we multiply $\overline{m_\ell}$ with its conjugate and obtain

$$\overline{m_\ell}m_\ell = |m_\ell|^2 = \frac{|Z_\ell S_\ell|^2}{\big(\,|S_\ell|^2 + \lambda - \frac{\lambda}{|m_\ell|}\big)^2}.$$

Solving this equation with respect to the modulus of m_ℓ and using the formula $z = |z|\,e^{i\arg(z)}$, the phase is only given by $S_\ell\overline{Z_\ell}$ in Equation (10), because the term in the brackets is real. The solution of our functional for every ℓ is

$$\tilde{m_\ell} = \frac{|Z_\ell S_\ell| + \lambda}{|S_\ell|^2 + \lambda}\cdot e^{i\arg(S_\ell\overline{Z_\ell})}.$$

c) For the regularization term $r(m) = \|m - 1\|_1$ we apply a substitution $m - 1 = \mu$, hence

$$\rho(\mu_\ell) = \frac{\lambda}{2}\left(\frac{\mu_\ell^r}{|\mu_\ell|} - i\frac{\mu_\ell^i}{|\mu_\ell|}\right) = \frac{\lambda}{2}\frac{\overline{\mu_\ell}}{|\mu_\ell|}.$$

Again we have to multiply with the conjugate in a similar manner as in case b) and we again use the formula $\mu = |\mu|\, e^{i\arg(\mu)}$. Undoing the substitution and applying a threshold argument obtained due to $|\mu_\ell| > 0$, we get

$$\tilde{m_\ell} = \frac{\left|\overline{S_\ell}\right| |Z_\ell - S_\ell| - \frac{\lambda}{2}}{|S_\ell|^2} \cdot e^{i\arg(\overline{S_\ell}(Z_\ell - S_\ell))} + 1 \qquad (11)$$

as long as

$$|S_\ell|\,|T_\ell - S_\ell| > \frac{\lambda}{2}.$$

d) For $r(m) = \|\,|m| - 1\|_1$ we have to make a distinction in two cases $|m_\ell| > 1$ and $|m_\ell| < 1$, from which

$$\rho(m_\ell) = \frac{\lambda}{2}\left(\pm\frac{m_\ell^r}{|m_\ell|} \mp \frac{m_\ell^i}{|m_\ell|}\right) = \pm\frac{\lambda}{2}\frac{\overline{m_\ell}}{|m_\ell|}$$

is obtained.
Using $|m_\ell|^2 = \overline{m_\ell} \cdot m_\ell$ and $z = |z|\, e^{i\arg(z)}$ we obtain

$$\tilde{m_\ell} = \begin{cases} \frac{|Z_\ell S_\ell| - \frac{\lambda}{2}}{|S_\ell|^2} e^{i\arg(S_\ell \overline{Z_\ell})} & \text{if } \frac{|Z_\ell S_\ell|}{|S_\ell|^2} > 1 + \frac{\lambda}{2|S_\ell|^2} \\ \frac{|Z_\ell S_\ell| + \frac{\lambda}{2}}{|S_\ell|^2} e^{i\arg(S_\ell \overline{Z_\ell})} & \text{if } \frac{|Z_\ell S_\ell|}{|S_\ell|^2} < 1 - \frac{\lambda}{2|S_\ell|^2} \\ 1 & \text{else.} \end{cases}$$

□