Proc SSVM 2015: Scale Space and Variational Methods in Computer Vision, Springer LNCS vol 9087, pages 3-15, 2015.

Scale-space theory for auditory signals

Tony Lindeberg[1] and Anders Friberg[2]

[1]Department of Computational Biology, [2]Department of Speech, Music and Hearing
School of Computer Science and Communication*
KTH Royal Institute of Technology, Stockholm, Sweden

Abstract. We show how the axiomatic structure of scale-space theory can be applied to the auditory domain and be used for deriving idealized models of auditory receptive fields via scale-space principles. For defining a time-frequency transformation of a purely temporal signal, it is shown that the scale-space framework allows for a new way of deriving the Gabor and Gammatone filters as well as a novel family of generalized Gammatone filters with additional degrees of freedom to obtain different trade-offs between the spectral selectivity and the temporal delay of time-causal window functions. Applied to the definition of a second layer of receptive fields from the spectrogram, it is shown that the scale-space framework leads to two canonical families of spectro-temporal receptive fields, using a combination of Gaussian filters over the logspectral domain with either Gaussian filters or a cascade of first-order integrators over the temporal domain. These spectro-temporal receptive fields can be either separable over the time-frequency domain or be adapted to local glissando transformations that represent variations in logarithmic frequencies over time. Such idealized models of auditory receptive fields respect auditory invariances, can be used for computing basic auditory features for audio processing and lead to predictions about auditory receptive fields with good qualitative similarity to biological receptive fields in the inferior colliculus (ICC) and the primary auditory cortex (A1).

1 Introduction

The information in sound is carried by variations in the air pressure over time, which for many sound sources can be modelled as a superposition of sine wave oscillations of different frequencies. To capture this information by auditory perception or signal processing, the sound signal has to be processed over some non-infinitesimal amount of time and in the case of a spectral analysis also over some range of frequencies. Such a region over time or over the spectro-temporal domain is referred to as a temporal or spectro-temporal *receptive field* (Aertsen and Johannesma [1]; Miller et al. [2]).

The subject of this article is to show how a principled theory for auditory receptive fields can be developed based on scale-space theory. Our aim is to

* Support from the Swedish Research Council contracts 2010-4766, 2012-4685 and 2014-4083, a KTH CSC Small Visionary Project and the EU project SkAT-VG FET-Open grant 618067 is gratefully acknowledged.

express auditory operations that (i) are well localized over time and frequencies and (ii) allow for well-founded handling of temporal phenomena that occur at different temporal scales as well as (iii) receptive fields that operate over different ranges of frequencies in such a way that operations over different ranges of frequencies can be related in a well-defined manner.

When applied to the definition of a spectrogram, alternatively to the formulation of an idealized cochlea model, the scale-space approach can be used for deriving the Gabor (Gabor [3]; Wolfe et al. [4]) and Gamma-tone (Johannesma [5]; Patterson et al. [6]) approaches for computing local windowed Fourier transforms as specific cases of a complex-valued scale-space transform over different frequencies. In addition, the scale-space approach to defining spectrograms leads to a new family of *generalized Gamma-tone filters*, where the time constants of the individual first-order integrators coupled in cascade are not equal as for regular Gamma-tone filters but instead distributed logarithmically over temporal scales and allowing for different trade-offs in terms of *e.g.* the frequency selectivity of the spectrogram and the temporal delay of time-causal receptive fields.

When applied to a logarithmic transformation of the spectrogram, as motivated from the desire of handling sound signals of different strength (sound pressure) in an invariant manner and with a logarithmic transformation of the frequencies as motivated by the desire of enabling invariance properties under a frequency shift, such as transposing a musical piece by one octave, the theory also allows for the formulation of spectro-temporal receptive fields at higher levels in the auditory hierarchy in terms of spectro-temporal derivatives of spectro-temporal smoothing operations as obtained from scale-space theory.

Such second-layer receptive fields can be used for (i) *computing basic auditory features* such as onset detection, partial tone enhancement and formants, and (ii) generating *predictions of auditory receptive fields qualitatively similar to biological receptive fields* as measured by cell recordings in the inferior colliculus (ICC) and the primary auditory cortex (A1) (Miller et al. [2]; Qiu et al. [7]; Elhilali et al. [8]; Atencio and Schreiner [9]).

In this concise summary of the theory, we emphasize the scale-space aspects of auditory receptive fields. A more extensive treatment is given in [10].

2 Multi-scale spectrograms

To capture the frequency content in an auditory signal $f : \mathbb{R} \to \mathbb{R}$, the notion of spectrograms or locally windowed Fourier transforms constitutes a natural tool

$$S(t, \omega;\ \tau) = \int_{t'=-\infty}^{\infty} f(t')\, e^{-i\omega t'}\, w(t - t';\ \tau)\, dt'. \tag{1}$$

A basic question in this context concerns how to choose the window function. Would any choice of window function w do? Specifically, how long should the effective integration time τ be? *A priori* there may be no principled reason for preferring a particular duration of the temporal window function for the windowed Fourier transform over some other temporal duration. Specifically,

different temporal durations may be appropriate for different auditory tasks, such as a preference for a short temporal duration for onset detection and a preference for a longer temporal duration to separate sounds with nearby frequencies.

If we apply a scale-space approach to this problem and associate a temporal window scale τ with any spectrogram, let us require that we should be able to relate spectrograms computed for different temporal window sizes between scales. If we assume a continuum of temporal window scales, then a *semi-group structure* $w(\cdot;\ \tau_2) = w(\cdot;\ \tau_2 - \tau_1) * w(\cdot;\ \tau_1)$ on the window functions implies a *cascade property* between the spectrograms

$$S(\cdot, \omega;\ \tau_2) = w(\cdot;\ \tau_2 - \tau_1) * S(\cdot, \omega;\ \tau_1). \tag{2}$$

If we instead assume a discrete set of temporal window scales, with each temporal window function $w(\cdot;\ n)$ at a coarser scale defined as the composition of a set of primitive temporal window functions $(\Delta w)(\cdot;\ k)$ such that $w(\cdot;\ n) = *_{k=1}^{n}(\Delta w)(\cdot;\ k)$, then we obtain a *Markov property* of the following type

$$S(\cdot, \omega;\ \tau_n) = (\Delta w)(\cdot;\ m \mapsto n)\, S(\cdot, \omega;\ \tau_m). \tag{3}$$

For pre-recorded sound signals we may in principle take the liberty of accessing the virtual future in relation to any time moment. For real-time audio processing or when modelling biological auditory perception there is on the other hand no way to access the future. For real-time audio models, the temporal window functions must therefore be *time-causal* such that $w(t;\ \tau) = 0$ for $t < 0$.

In the case of non-causal time and a continuum of temporal window scales, let us assume that the window functions in addition should guarantee non-creation of new structure in the sense of non-enhancement of local extrema in either of the real or purely imaginary channels. Then, it follows from general results in (Lindeberg [11], eq. (45)) that the temporal window function must be Gaussian

$$g(t;\ \tau) = \frac{1}{\sqrt{2\pi \Sigma_\tau}}\, e^{-(t-\delta_\tau)^2/2\tau} \tag{4}$$

with $\Sigma_\tau = \tau\, \Sigma_0$ and $\delta_\tau = \tau\, \delta_0$ where we without loss of generality can set $\Sigma_0 = 1$.

If we in the case of time-causal data and a discrete set of temporal window scales assume that the temporal window functions should guarantee non-creation of new structure in the sense of guaranteeing non-creation of new local extrema in either of the real or purely imaginary channels, then it follows from general results in (Lindeberg and Fagerström [12], eq. (8)) that the temporal window functions should be given by a cascade of truncated exponential functions

$$h_{composed}(t;\ \mu) = *_{k=1}^{K} h_{exp}(t;\ \mu_k) \tag{5}$$

where $\mu = (\mu_1, \ldots, \mu_k)$ and

$$h_{exp}(t;\ \mu_k) = \begin{cases} \frac{1}{\mu_k} e^{-t/\mu_k} & t \geq 0 \\ 0 & t < 0 \end{cases} \tag{6}$$

Thereby the convolution kernels in temporal scale spaces for a general time-varying signal are used as scale-dependent window functions for defining windowed Fourier transforms of different temporal extent. Specifically, this scale-space approach allows for the definition of windowed Fourier transforms for all temporal extents in such a way that a windowed Fourier transform at any coarse temporal scale can be related to a windowed Fourier transform at any finer temporal scale using the cascade property (2) or the Markov property (3) derived from the underlying scale-space kernels. Combined with the additional scale-space properties of non-creation of new structures with increasing scale, this guarantees well-founded theoretical properties between corresponding windowed Fourier transforms at different temporal scales.

Relations to Gabor functions. By rewriting the expressions (1) and (4) for the complex-valued spectrogram based on the Gaussian temporal scale space as

$$S_g(\omega, t;\ \tau) = e^{-i\omega t} \int_{t'=-\infty}^{\infty} g(t - t';\ \tau)\, e^{i\omega(t-t')} f(t')\, dt' \tag{7}$$

it can be seen that up to a phase shift this multi-scale spectrogram can equivalently be interpreted as the convolution of the original auditory signal f by *Gabor functions* [3] of the form

$$G(t, \omega;\ \tau) = g(t;\ \tau)\, e^{i\omega t}. \tag{8}$$

Such Gabor functions have been previously used for analyzing auditory signals by several authors, including Wolfe et al. [4] and Heckmann et al. [13].

Relations to Gammatone filters. In the special case when the time constants of the K truncated exponential filters that are coupled in cascade are all equal $\mu_k = \mu$, then the multi-scale spectrogram defined by (1) and (5) is given by [10]

$$S_h(t, \omega;\ \mu, K) = e^{-i\omega t} \int_{t'=-\infty}^{\infty} \frac{(t - t')^{K-1}\, e^{-(t-t')/\mu}}{\mu^K\, \Gamma(K)}\, e^{i\omega(t-t')} f(t')\, dt' \tag{9}$$

and does up to a phase shift correspond to convolution of the input signal f by filters of the form

$$h_{cos}(t, \omega;\ \mu, K) = \frac{t^{K-1}\, e^{-t/\mu}}{\mu^K\, \Gamma(K)} \cos \omega t, \tag{10}$$

$$h_{sin}(t, \omega;\ \mu, K) = \frac{t^{K-1}\, e^{-t/\mu}}{\mu^K\, \Gamma(K)} \sin \omega t. \tag{11}$$

For comparison, the *Gammatone filter* with parameters a and b and frequency ϕ is defined according to $\gamma(t) = a\, t^{n-1} e^{-2\pi b t} \cos(2\pi\phi\, t + \alpha)$. By identifying the parameters $a = 1/(\mu^K \Gamma(K))$, $b = 1/(2\pi\mu)$ and $\omega = 2\pi\, \phi$, it follows that we can derive the Gammatone filter as a special case of applying a time-causal scale-space representation with discrete scale levels to the projections $f \cos \omega t$ and $f \sin \omega t$ of an auditory signal $f(t)$ onto a complex sine wave $e^{-i\omega t}$.

Gammatone filter banks are also commonly used in audio processing (Johannesma [5]; Patterson et al. [6]; Ngamkham et al. [14]).

Generalized Gammatone filters. By allowing for different time constants in the primitive truncated exponential filters, we obtain *generalized Gammatone filters*

$$h_{cos}(t, \omega;\ \mu) = h_{composed}(t;\ \mu)\ \cos \omega t \tag{12}$$

$$h_{sin}(t, \omega;\ \mu) = h_{composed}(t;\ \mu)\ \sin \omega t \tag{13}$$

with $h_{composed}$ according to (5) and $\mu = (\mu_1, \ldots, \mu_K)$. If we have the freedom of choosing the minimum temporal window scale τ_{min} freely, we can parameterize the intermediate temporal scale levels using a parameter $c > 1$ such that [16]

$$\tau_k = c^{2(k-K)} \tau_{max} \qquad (1 \leq k \leq K) \tag{14}$$

which shares some qualitative similarities to the logarithmic transformation of the past used in the scale-time model proposed by Koenderink [15].

By the additive property of variances (which for a primitive truncated exponential filter (6) with time constant μ_k is given by μ_k^2) under convolution this implies that time constants of the individual first-order integrators will be [16]

$$\mu_1 = c^{1-K} \sqrt{\tau_{max}} \tag{15}$$

$$\mu_k = \sqrt{\tau_k - \tau_{k-1}} = c^{k-K-1} \sqrt{c^2 - 1} \sqrt{\tau_{max}} \qquad (2 \leq k \leq K) \tag{16}$$

By comparing graphs of the underlying temporal scale-space kernels [16], one finds that filters based on truncated exponentials with a logarithmic distribution of the intermediate temporal scales allow for a faster temporal response compared to the corresponding filters based on truncated exponentials with equal time constants. Thereby, these generalized Gammatone filters allow for additional degrees of freedom to obtain different trade-offs between the frequency selectivity and the temporal delay of time-causal window functions by varying the number of levels K and the distribution parameter c for a given τ_{max}.

Frequency-dependent window scale. To guarantee basic covariance properties of the spectrogram under a frequency shift $\omega \mapsto \alpha\, \omega$, it is natural to let the temporal window scale vary with the frequency ω in such a a way that the temporal window scale in units of $\sigma = \sqrt{\tau}$ is *proportional to the wavelength* $\lambda = 2\pi/\omega$

$$\tau = \left(\frac{2\pi\, n}{\omega} \right)^2 \tag{17}$$

where n is a parameter. By such frequency dependent temporal window scale, the spectral selectivity in the spectrogram (the width of a spectral band) will be independent of the frequency ω. This is a prerequisite for the desirable property that a shift by one octave of a musical piece should imply that the corresponding spectrogram should appear similar while shifted by one octave, if the frequency axis of the spectrogram is parameterized on a logarithmic scale.

Additionally, to prevent the temporal window scale from being too short for high frequencies or too long at low frequencies, we also introduce soft lower and upper bounds on the temporal window scale. Thereby, self-similarity will only hold within a limited range of frequencies.

3 Second-layer receptive fields over the spectrogram

Given that a spectrogram has been computed by a first layer of auditory receptive fields, we define a *second layer of receptive fields* by operating on the spectrogram with 2-D spectro-temporal filters in a structurally similar way as visual receptive fields are applied to time-varying visual input (see overview in Lindeberg [17]).

3.1 Invariances by logarithmic transformations of the spectrogram

Prior to the definition of receptive fields from the spectrogram, it is natural to allow for a self-similar *logarithmic transformation of the magnitude values*

$$S_{dB} = 20 \log_{10} \left(\frac{|S|}{S_0} \right). \tag{18}$$

Then, a multiplicative transformation of the sound pressure $f \mapsto a\, f$, corresponding to $|S| \mapsto a\, |S|$, or an inversely proportional reduction in the sound pressure of the signal from a single auditory point source as function of distance $f \mapsto f/R$, corresponding to $|S| \mapsto |S|/R$, are both transformed into a subtraction of the logarithmic magnitude by a constant.

If we operate on the logarithmically transformed spectrogram by a receptive field $\mathcal{A}_\Sigma$ that is based on a combination of a spectro-temporal smoothing operation $\mathcal{T}_\Sigma$ with logspectral and temporal scale parameters as determined by a spectro-temporal covariance matrix Σ, temporal and/or logspectral derivatives $\partial_t^\alpha \partial_\nu^\beta$ of orders α and β with at least one of $\alpha > 0$ or $\beta > 0$

$$\mathcal{A}_\Sigma\, S_{dB} = \partial_t^\alpha \partial_\nu^\beta \mathcal{T}_\Sigma\, S_{dB} \tag{19}$$

then the influence on the receptive field responses of the constants a and R

$$\mathcal{A}_\Sigma S_{dB} = \partial_t^\alpha \partial_\nu^\beta \mathcal{T}_\Sigma\, (S_{dB} + 20 \log_{10} a - 20 \log_{10} R) = \partial_t^\alpha \partial_\nu^\beta \mathcal{T}_\Sigma\, S_{dB} + 0 + 0 \tag{20}$$

will be eliminated if the constants a and R do not depend on time t or the logarithmic frequency ν, implying *invariance of the second-layer receptive field responses to variations in the sound pressure or the distance to a sound source*.

Since logarithmic frequencies constitute a natural metric for relating frequencies of sound and there is an approximately logarithmic distribution of frequencies both on the basilar membrane and in the auditory cortex, it is natural to express these derived receptive fields in terms of *logarithmic frequencies*

$$\nu = \nu_0 + C\, \log \left(\frac{\omega}{\omega_0} \right) \tag{21}$$

for some constants C and ω_0, where specifically $\nu_0 = 69$, $C = 12/\log 2$ and $\omega_0 = 2\pi \cdot 440$ correspond to the MIDI standard.

This logarithmic parameterization implies that a shift in frequency, caused by *e.g.* transposing a piece of music by one octave or varying the fundamental

frequency in singing resulting in a multiplicative transformation of the harmonics (overtones), corresponds to a mere *translation* in logarithmic frequency.

Note, however, that some properties of voice or instruments, such as the formant structure in speech or physical resonances in instruments, are independent of the fundamental frequency and therefore not frequency invariant.

3.2 Structural requirements on second-layer receptive fields

Given such a logarithmically transformed spectrogram, we define a family of *second-layer spectro-temporal receptive fields* $A(t, \omega;\ \Sigma)$ that are to operate on the transformed spectrogram $S_{dB}(t, \nu; \tau)$ and be parameterized by some multi-dimensional spectro-temporal scale parameter Σ comprising smoothing over time t and logarithmic frequencies ν, and obeying:

(i) *linearity* over the logarithmic spectrogram to ensure that (a) the multiplicative relations of the magnitude of the spectrogram that are mapped to linear relations by the logarithmic transformation (18) are preserved as linear relations over the receptive field responses and (b) the scale-space properties imposed to ensure non-creation of new structures in smoothed spectrograms as defined by spectro-temporal smoothing kernels do also transfer to spectro-temporal derivatives of these.

(ii) *shift-invariance* with respect to translations over time $t \mapsto t + \Delta t$ and logarithmic frequencies $\nu \mapsto \nu + \Delta\nu$ such that all temporal moments and all logarithmic frequencies are treated in a similar manner. Temporal shift invariance implies that an auditory stimulus should be perceived in a similar manner irrespective of when it occurs. Shift-invariance in the logarithmic frequency domain implies that, for example, a piece of music should be perceived in a similar manner if it is transposed by *e.g.* one octave.

(iii.a) For pre-recorded sound signals, for which we can take the freedom of accessing data from the virtual future in relation to any time moment, we impose a *continuous semi-group structure over spectro-temporal scales* on the second-layer receptive fields $T(\cdot, \cdot;\ \Sigma_2) = T(\cdot, \cdot;\ \Sigma_2 - \Sigma_1)\, T(\cdot, \cdot;\ \Sigma_1)$ corresponding to an additive structure over the multi-dimensional scale parameter Σ.

(iii.b) For time-causal signals, we require a *continuous semi-group structure over logspectral scales* s, $T(\cdot;\ s_2) = T(\cdot;\ s_2 - s_1)\, T(\cdot;\ s_1)$, and a *Markov property between adjacent temporal scales* τ, $T(\cdot;\ \tau_{k+1}) = (\Delta T)(\cdot;\ k)\, T(\cdot;\ \tau_k)$.

(iv.a) For the non-causal spectrogram (7) we require *non-enhancement of local extrema* in the sense that if for some scale Σ_0 the point (t_0, ν_0) is a local maximum (minimum) for the mapping $(t, \nu) \mapsto (\mathcal{A}_\Sigma S_{dB})(t, \nu;\ \tau, \Sigma_0)$ then the value at this point must not increase (decrease) with increasing scale Σ.

(iv.b) For the time-causal spectrogram generated by (10)–(11) or (12)–(13) we require: (iv.b1) the smoothing operation over the logspectral domain to satisfy *non-enhancement of local extrema* in the sense that if at some logspectral scale s_0 a point ν_0 is a local maximum (minimum) of the mapping $\nu \mapsto (\mathcal{A}_\Sigma S_{dB})(\nu;\ \tau, s_0)$ obtained by disregarding the temporal variations, then the value at this point must not increase (decrease) with increasing

logspectral scale s, and (iv.b2) the purely temporal smoothing operation to be a time-causal scale-space kernel guaranteeing *non-creation of new local extrema* under an increase of the temporal scale parameter τ.

(v) *glissando covariance* in the sense that if two local patches of two spectrograms are related by a local glissando transformation $S' = \mathcal{G}_v\, S$ of the form $\nu' = \nu + v\, t$ and corresponding to frequencies that vary smoothly over time, such as during singing or for instruments with continuous pitch control, then it should be possible to relate the local spectro-temporal receptive field responses such that $\mathcal{A}_{G_v(\Sigma)}\, \mathcal{G}_v\, S = \mathcal{G}_v\, \mathcal{A}_\Sigma\, S$ for some transformation $\Sigma' = G_v(\Sigma)$ of the spectro-temporal scale parameters Σ.

3.3 Idealized models for spectro-temporal receptive fields

Given these structural requirements, it follows from derivations similar to those that are used for constraining visual receptive fields given structural requirements on a visual front-end (Lindeberg [17]) that the second layer of auditory receptive fields should be based on spectro-temporal receptive fields of the form

$$A(t, \nu;\ \Sigma) = \partial_t^\alpha \partial_\nu^\beta \left(g(\nu - vt;\ s)\, T(t;\ \tau)\right) \tag{22}$$

where

- ∂_t^α represents a *temporal derivative operator* of order α with respect to time t which could alternatively be replaced by a glissando-adapted temporal derivative of the form $\partial_{\bar{t}} = \partial_t + v\, \partial_\nu$,
- ∂_ν^β represents a *logspectral derivative operator* of order β with respect to logarithmic frequency ν,
- $T(t;\ \tau)$ represents a *temporal smoothing kernel* with temporal scale parameter τ, which should either be (i) a temporal Gaussian kernel $g(t;\ \tau)$ (4) or (ii) the equivalent kernel $h_{composed}(t;\ \mu)$ according to (5) and corresponding to a set of truncated exponential kernels coupled in cascade, and
- $g(\nu - vt;\ s)$ represents a Gaussian *spectral smoothing kernel* over logarithmic frequencies ν with logspectral scale parameter s and v representing a glissando parameter making it possible to adapt the receptive fields to variations in frequency $\nu' = \nu + vt$ over time and
- the spectro-temporal covariance matrix Σ in the left hand expression for spectro-temporal receptive fields comprises both the temporal scale parameter τ, the logspectral scale parameter s and the glissando parameter v.

Thereby, the spectro-temporal receptive fields (22) constitute a combination of a Gaussian scale-space concept over the logspectral dimension with purely temporal receptive fields obtained by either a non-causal Gaussian temporal scale space or a time-causal scale space obtained by coupling truncated exponential kernels/first-order integrators in cascade (see figure 2, columns 2-3).

The proofs concerning spectro-temporal receptive fields are similar to those regarding spatio-temporal receptive fields over a 1+1-D spatio-temporal domain with the spatial dimension replaced by a logspectral dimension.

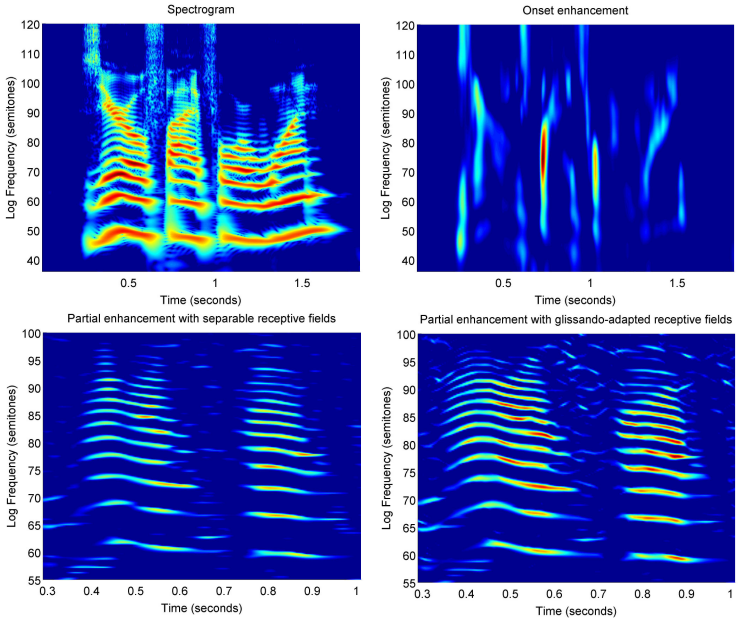

Fig. 1. (top left) Spectrogram of a male voice that reads "zero five four one" (from the TIDigits database) computed with generalized Gammatone functions. (top right) Onset enhancement by first-order temporal derivatives. (bottom left) Enhancement of partial tones by second-order logspectral derivatives using separable receptive fields. (bottom right) Enhancement of partial tones by the maximum of second-order logspectral derivatives over a filter bank of glissando-adapted receptive fields. Note the better ability of the glissando-adapted receptive fields to capture rapid frequency variations.

3.4 Auditory features from second-layer receptive fields

In the following, we will show examples of auditory features that can be defined from a second layer of auditory receptive fields of this form:

Onset enhancement. Computation of first-order temporal derivatives $\mathcal{D}_t(t, \nu;\ \tau, s) = \sqrt{\tau}\, \partial_t T(t, \nu;\ \tau, s)$ where $\sqrt{\tau}$ is a scale normalization factor to approximate *scale-normalized derivatives* (Lindeberg [18]). To select receptive field responses that correspond to onsets only, we add the non-linear logical operation $D_t > 0$ such that $\mathcal{A}_{onset}\, S_{dB} = \mathcal{D}_t\, S_{dB}$ if $\mathcal{D}_t\, S_{dB} > 0$ and 0 otherwise (see figure 1, top right).

Enhancement of partials. Computation of second-order logspectral derivatives $\mathcal{D}_{\nu\nu}(t, \nu;\ \tau, s) = s\, \partial_{\nu\nu} T(t, \nu;\ \tau, s)$ where the factor s is a scale normalization factor for scale-normalized derivatives in the Gaussian scale space (Lindeberg [18]). Depending on the value of the logspectral scale parameter s, this operation may either enhance partial tones or formants. This operation is naturally combined with the (non-linear) logical operation $\mathcal{D}_{\nu\nu} < 0$ such that $\mathcal{A}_{band}\, S_{dB} = -\mathcal{D}_{\nu\nu}\, S_{dB}$ if $\mathcal{D}_{\nu\nu}\, S_{dB} < 0$ and 0 otherwise (see figure 1, bottom left).

Enhancement of partials using filter bank of glissando-adapted receptive fields. To more accurately capture the harmonic components in sound for which the frequencies vary rapidly over time, we use a filter bank of receptive fields that are adapted to different glissando values v, which are combined by taking the maximum over all glissando-adapted filter responses (see figure 1, bottom right).

4 Relations to biological receptive fields

In the central nucleus of the inferior colliculus (ICC) of cats, Qiu et al. [7] report that about 60 % of the neurons can be described as separable in the time-frequency domain (see figure 2, top row), whereas the remaining neurons are either obliquely oriented (see figure 2, second row) or contain multiple excitatory/inhibitory subfields. This overall structure is nicely compatible with the treatment in section 3.4, where the second-layer receptive fields are expressed in terms of spectro-temporal derivatives of either time-frequency separable spectro-temporal smoothing operations or corresponding glissando-adapted features as motivated by the structural requirements in section 3.2.

Qualitatively similar shapes of receptive fields can be measured from neurons in the primary auditory cortex (see figure 2, third row, as well as Miller et al. [2] regarding binaural receptive fields). Specifically, the use of multiple temporal and spectral scales as a main component in the model is in good agreement with biological receptive fields having different degrees of spectral tuning ranging from narrow to broad and different temporal extent (see figure 2, rows 4-5).

5 Summary and discussion

We have presented a theory for how idealized models of auditory receptive fields can be derived from structural constraints (scale-space axioms) on the first stages of auditory processing. The theory includes (i) the definition of multi-scale spectrograms at different temporal scales in such a way that a spectrogram at any coarser temporal scale can be related to a corresponding spectrogram at any finer temporal scale using theoretically well-defined scale-space operations, and additionally (ii) how a second layer of spectro-temporal receptive fields can be defined over a logarithmically transformed spectrogram in such a way that the resulting spectro-temporal receptive fields obey invariance or covariance properties under natural sound transformations including temporal shifts, variations in the sound pressure, the distance between the sound source and the observer, a shift in the frequencies of auditory stimuli or glissando transformations. Specifically, theoretical arguments have been presented showing how these idealized receptive fields are constrained to the presented forms from symmetry properties of the environment in combination with assumptions about the internal structure of auditory operations as motivated from requirements of handling different temporal and spectral scales in a theoretically well-founded manner.

We propose that this theory should be of wide general interest for the audio processing community by providing theoretically well-founded and provably

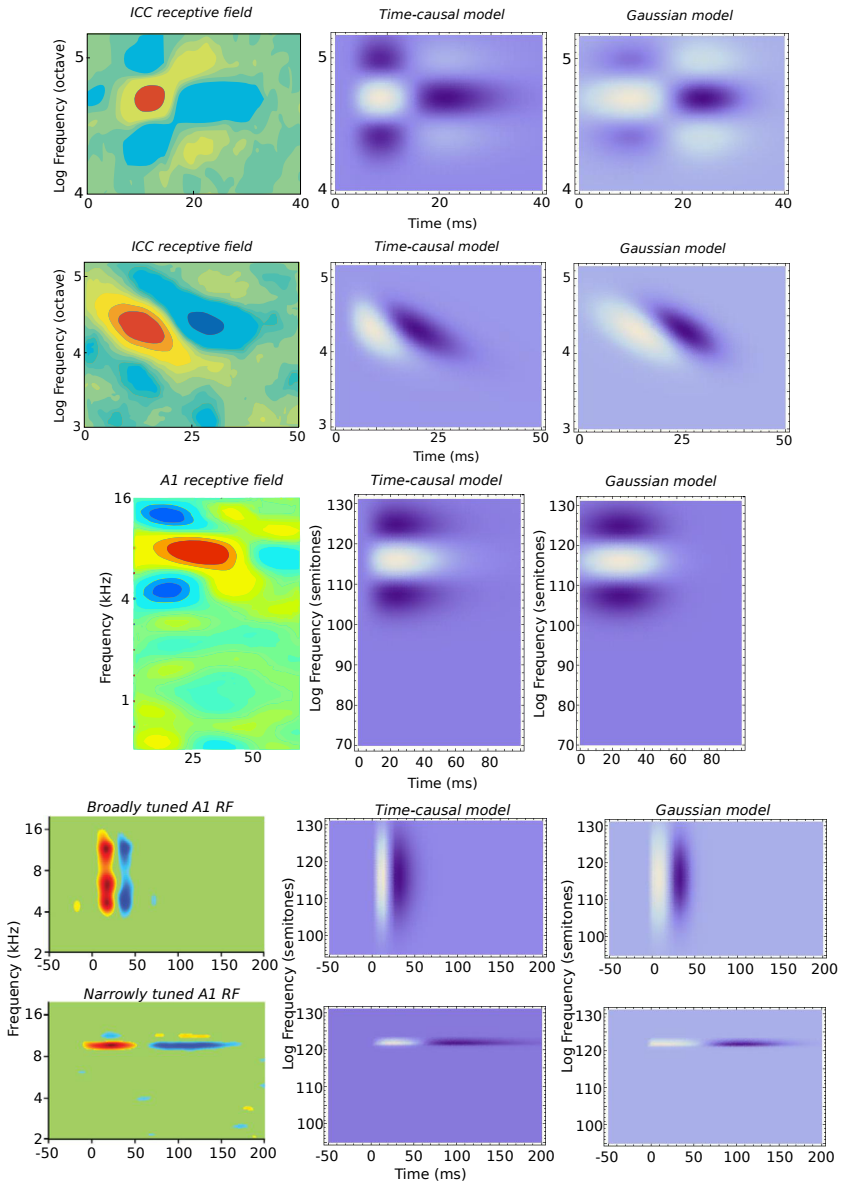

Fig. 2. (top row left) A separable monaural spectro-temporal receptive field in the central nucleus of the inferior colliculus (ICC) of cat as reported by Qiu et al. [7]. (second row left) A non-separable spectro-temporal receptive field in the central nucleus of the inferior colliculus (ICC) of cat as reported by Qiu et al. [7]. (third row left) A separable spectro-temporal receptive fields in the primary auditory cortex (A1) of ferret as reported by Elhilali et al. [8]. (fourth and bottom rows left) Spectro-temporal receptive fields of broadly and narrowly tuned neurons in the primary auditory cortex (A1) of cats as reported by Atencio and Schreiner [9]. (middle and right columns) Time-causal and non-causal receptive field models according to eq. (22). (Figures reprinted from [10] with permission.)

invariant/covariant audio operations for processing sound signals and for computational modelling or measurements of receptive fields, auditory invariances, theoretical biology and psychophysics, by serving as a general theoretical foundation and understanding of how receptive fields in ICC and A1 support invariant visual processes at higher levels in the auditory hierarchy.

References

1. Aertsen, A.M.H.J., Johannesma, P.I.M.: The spectro-temporal receptive field: A functional characterization of auditory neurons. Biol. Cyb. **42** (1981) 133–143
2. Miller, L.M., Escabi, N.A., Read, H.L., Schreiner, C.: Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. J. Neurophys. **87** (2001) 516–527
3. Gabor, D.: Theory of communication. J. of the IEE **93** (1946) 429–457
4. Wolfe, P.J., Godsill, S.J., Dorfler, M.: Multi-Gabor dictionaries for audio time-frequency analysis. Appl. of Signal Proc. to Audio and Acoustics. (2001) 43–46
5. Johannesma, P.I.M.: The pre-response stimulus ensemble of neurons in the cochlear nucleus. In: IPO Symposium on Hearing Theory, Eindhoven, (1972) 58–69
6. Patterson, R.D., Nimmo-Smith, I., Holdsworth, J., Rice, P.: An efficient auditory filterbank based on the gammatone function. In: A meeting of the IOC Speech Group on Auditory Modelling at RSRE. Volume 2:7. (1987)
7. Qiu, A., Schreiner, C.E., Escabi, M.A.: Gabor analysis of auditory midbrain receptive fields: Spectro-temporal and binaural composition. J. of Neurophysiology **90** (2003) 456–476
8. Elhilali, M., Fritz, J., Chi, T.S., Shamma, S.: Auditory cortical receptive fields: Stable entities with plastic abilities. J. of Neuroscience **27** (2007) 10372–10382
9. Atencio, C.A., Schreiner, C.E.: Spectrotemporal processing in spectral tuning modules of cat primary auditory cortex. PLOS ONE **7** (2012) e31537
10. Lindeberg, T., Friberg, A.: Idealized computational models of auditory receptive fields. PLOS ONE 10(3):e0119032 (2015) 1–58, preprint at arXiv:1404.2037.
11. Lindeberg, T.: Generalized Gaussian scale-space axiomatics comprising linear scale-space, affine scale-space and spatio-temporal scale-space. J. of Mathematical Imaging and Vision **40** (2011) 36–81
12. Lindeberg, T., Fagerström, D.: Scale-space with causal time direction. In: European Conf. on Computer Vision, Springer LNCS Vol. 1064 (1996) 229–240
13. Heckmann, M., Domont, X., Joublin, F., Goerick, C.: A hierarchical framework for spectro-temporal feature extraction. Speech Communication **53** (2011) 736–752
14. Ngamkham, W., Sawigun, C., Hiseni, S., Serdijn, W.A.: Analog complex gammatone filter for cochlear implant channels. In: ISCAS (2010) 969–972
15. Koenderink, J.J.: Scale-time. Biological Cybernetics **58** (1988) 159–162
16. Lindeberg, T.: Separable time-causal and time-recursive receptive fields. In: Scale Space and Variational Methods in Computer Vision, Springer LNCS Vol. 9087 (2015) 90–102
17. Lindeberg, T.: A computational theory of visual receptive fields. Biological Cybernetics **107** (2013) 589–635
18. Lindeberg, T.: Feature detection with automatic scale selection. Int. J. of Computer Vision **30** (1998) 77–116