TIME–FREQUENCY STRUCTURED DECORRELATION OF SPEECH SIGNALS VIA NONSEPARABLE GABOR FRAMES

Werner Kozek *and* *Hans G. Feichtinger*

Numerical Harmonic Analysis Group
Department of Mathematics, University of Vienna
Strudlhofgasse 4, A-1090 Vienna, Austria
kozek@tyche.mat.univie.ac.at

ABSTRACT

We present a new approach to the linear representation of speech signals that combines desirable structure, computational efficiency and almost decorrelation. The basic principle is a statistically adapted, group-theoretical modification of the classical Gabor expansion. In contrast to traditional linear time–frequency (TF) representations which always correspond to a separable tiling of the TF plane, we suggest the use of a hexagonal (thus nonseparable) tiling whose parameters are matched to the TF correlation of the speech signal. We estimate the TF correlation via a pitch-adapted Zak-transform motivated by modeling the vocal tract as underspread system. The TF correlation determines both the optimum tiling and the optimum window.

1. INTRODUCTION

Typical modern speech processing algorithms involve a cascade of linear and nonlinear (signal adaptive) transforms. In the first stage of a speech processing algorithm one is endeavored to decrease the considerable redundancy in the locally stationary parts of the speech signal via linear transforms [1]. The theoretic optimum is the Karhunen–Loeve (KL) transform which, however, fails to satisfy indispensable practical side-constraints. The desirable properties of a general purpose, first-stage speech transform are as follows:

Time–Frequency Parametrization. The double index of the transform domain should correspond to an appropriately normalized time scale and frequency scale. This constraint assures modularity with existing higher level speech processing algorithms.

Invertibility. For speech coding or speech enhancement invertibility is of obvious relevance. In applications such as speech recognition (where the transform is used as a front-end to produce a feature vector) invertibility guarantees no loss of information.

Efficiency. A first-stage transform should leave most of the available computational power for the higher level processing.

2. STFT AND GABOR EXPANSION

STFT. The short-time Fourier transform (STFT) and its squared magnitude, the spectrogram, are the fundamental tools for a TF analysis of speech signals. The STFT is defined as [2, 3] (the range of all integrals is the real line)

$$V_g f(x,\xi) := \int f(y)\overline{g(y-x)}e^{-j2\pi y\xi}dy,$$

where f is the signal, g is the analysis window, x is time, ξ is frequency and the bar denotes complex conjugation. (We assume that $f, g \in L^2(\mathbb{R})$.)

Gabor Representation. Mapping a one-dimensional signal onto the two-dimensional TF plane introduces considerable redundancy. While redundancy is appropriate for visualization, it is generally undesired in "black box" signal processing. The natural way to reduce this redundancy is by sampling the STFT on a separable lattice (a, b denote the time/frequency periods):

$$T_g f[k,l] := V_g f(ka, lb).$$

This leads to a linear, discrete TF signal representation $T_g f[k,l]$ known as Gabor representation [4, 3].

Reconstruction. From a modern signal decomposition point of view, the Gabor representation can be interpreted as the coefficients of a nonorthogonal expansion. The set of functions $\{g(x-ak)e^{j2\pi blx}\}_{k,l\in\mathbb{Z}}$ establishes a coherent *frame* which admits a (numerically stable and efficient) reconstruction via a so-called *dual frame* [3]. Indeed, in his classical work [4], Gabor takes a signal synthesis point of view, postulating a signal expansion of the following form

$$f(x) = \sum_{k\in\mathbb{Z}}\sum_{l\in\mathbb{Z}} T_g f[k,l]h(x-ka)e^{j2\pi x\, lb},$$

where the function h can be interpreted as *synthesis window*

One can view the traditional Gabor frame as a rectangular tiling of the time-frequency plane. The reconstruction is numerically stable for sufficiently high density, i.e.,

$$ab < 1, \tag{1}$$

this requirement also implies existence of a "nice" synthesis window h given a usual Gaussian-like analysis window g.

This work was partially supported by project S-7001 of the Austrian Science foundation FWF.

3. UNDERSPREAD ENVIRONMENTS

Modern speech processing is predominantly based on the concept of stationary processes. The profound statistical theory is tied to the true behavior of speech signals via a so-called *quasistationarity* assumption. The resulting choice of a window is largely due to heuristic considerations. A mathematically precise definition of quasistationarity can be based on the theory of underspread environments [5, 6].

Spreading Function. The second order theory of nonstationary random processes and the nonparametric theory of linear time-varying (LTV) systems can be characterized by linear integral operators acting as:

$$Hf(x) = \int k_H(x,y)f(y)dy,$$

where the kernel k corresponds either to the impulse response of an LTV system or to the covariance function of a random process. In order to classify the "degree of nonstationarity" of such environments, it is advantageous to switch to a different I-O-relation :

$$Hf(x) = \int\int S_H(\tau,\nu)f(x-\tau)e^{j2\pi\nu x}d\tau d\nu,$$

i.e., the output signal Hf is formulated as a superposition of TF-shifted versions of the input signal. The (generally complex valued) weight function S_H is in one-to-one correspondence to the kernel:

$$S_H(\tau,\nu) = \int k_H(x,x-\tau)e^{-j2\pi\nu x}dx. \qquad (2)$$

We shall refer to S_H as *spreading function* of the operator H, which is the system theoretic terminology. In the context of random processes, where we have covariance operators defined by a covariance kernel, the spreading function can be interpreted as a TF correlation function (more precisely, an expected ambiguity function). We henceforth assume that all involved processes are zero-mean such that the correlation function is equivalent to the covariance function defined as $(R)(x,y) := \mathrm{E}\{f(x)\overline{f(y)}\}$ (E denotes the expectation operator and f is a nonstationary process).

A fundamental classification of nonstationary environments can be formulated via singular support constraints on the spreading function. Convolution operators (corresponding to a wide-sense stationary process or to a linear time-invariant system) do not introduce frequency shifts, while multiplication operators (corresponding to nonstationary white noise or to modulation systems) do not introduce time-shifts, thus their spreading function is concentrated on one axis of the (τ,ν)-plane, see Figure 3.

Underspread Condition. In view of the support constraints of S_H for stationary and "totally nonstationary" (multiplicative) environments, a canonical definition of quasistationarity is given by restricting S_H to a centered rectangle of the (τ,ν)-plane:

$$\mathrm{supp}S_H \subseteq [-\tau_0,\tau_0] \times [-\nu_0,\nu_0],$$

where $\sigma_H = 4\tau_0\nu_0$ is called *total spread* and *underspread* environments are defined by $\sigma_H \ll 1$.

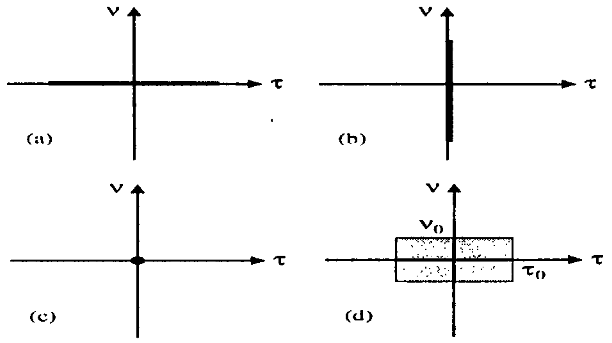

Figure 1: Support of $S_H(\tau,\nu)$: (a) convolution, (b) multiplication, (c) identity, (d) underspread operator ($\tau_0\nu_0 \ll 1$)

Fundamental Property. Underspread operators satisfy a number of interesting mathematical properties. Generally speaking, the total spread σ_H is a fundamental condition number for the applicability of (nonparametric) TF-domain representations, while the spreading ratio $\frac{\tau_0}{\nu_0}$ is the critical design parameter for estimators or identification procedures. Of particular relevance for the present context is the existence of approximate eigenfunctions and eigenvalues in the following sense. Consider a pulse-like prototype function g whose duration is adapted to the spreading constraint by an intuitive appealing matching rule

$$\frac{T_g}{F_g} = \frac{\tau_0}{\nu_0} \qquad (3)$$

where T_g, F_g are the duration/bandwidth of g defined via normalized second-order moments. Then, one can show that any TF-shifted version of g is an approximate eigenfunction of H, and the associated almost-eigenvalue is determined by the transfer function of H at the very same TF location:

$$Hg_{\tau,\nu} = Z_H(\tau,\nu)g_{\tau,\nu} + \mathcal{O}(\sigma_H^2) \qquad (4)$$

where $g_{\tau,\nu}$ denotes the TF-shifted version of g, i.e., $g_{\tau,\nu}(x) := g(x-\tau)e^{j2\pi x\nu}$ and Z_H is Zadeh's time-varying transfer function (Kohn-Nirenberg symbol) of H,

$$Z_H(\tau,\nu) := \int k_H(\tau,\tau-y)e^{-j2\pi\nu y}dy,$$

and $\mathcal{O}(\sigma_H^2)$ is a small (L^2-sense) error term. Note that this property corresponds to the well-known fact that complex sinusoids $g_\nu(x) = e^{-j2\pi\nu x}$ are (generalized) eigenfunctions of convolution operators, $Hg_\nu = \bar{Z}(\nu)g_\nu$ where the generalized eigenvalue distribution $\bar{Z}(\nu)$ is given by the Fourier transform of the convolution kernel (transfer function in case of LTI systems or power spectrum in case of stationary processes). For the present context (where H stands for the covariance of the speech signal as a nonstationary random process) approximate eigenfunction means just that $g_{\tau,\nu}$ is close to a KL basis function for abitrary (τ,ν). Hence, it remains to sample the matched continuous frame $\{g_{\tau,\nu}\}_{(\tau,\nu)\in\mathbf{R}\times\mathbf{R}}$ on a properly chosen subgroup in order to obtain a coherent, KL-like discrete family of functions [6]. Before proceeding in this direction we have to determine the underspread support of the speech signal's covariance.

4. IDENTIFICATION OF THE VOCAL TRACT

Following the classical speech production model [7] we assume periodic excitation of the vocal tract by a glottal pulse train (during vowel sounds):

$$p_a(x) = a \sum_{k \in \mathbb{Z}} \delta(x - ka),$$

where a is the pitch period. More precisely, we merge the glottis pulse with the impulse response of the vocal tract to an overall time–varying impulse response. (We are not really interested to identify the vocal tract per se, rather we want to characterize the TF correlation of the speech signal in a reproducible way.) In contrast to the usual quasistationarity assumption we explicitly consider the vocal tract as a linear time–varying system. One may hope that this system is indeed approximately underspread, because during vowel sounds the velocity of the moving parts of the vocal tract is certainly much smaller than the velocity of the propagating acoustic waves. However, irrespective of the physical relevance, it is clear that to any given signal f one can assign an LTV system H such that

$$f = Hp_a.$$

It can be shown that direct, unbiased estimates of S_H can be obtained via the *Zak transform* of Hp_a:

$$\mathcal{Z}Hp_a(\tau, \nu) = \sum_{k \in \mathbb{Z}} \sum_{l \in \mathbb{Z}} S_H\left(\tau - ka, \nu - \frac{l}{a}\right) e^{j2\pi(\nu - k/a)\tau},$$

with the Zak transform defined as [8]

$$\mathcal{Z}f(\tau, \nu) := \sum_{k \in \mathbb{Z}} f(t + ka) e^{-j2\pi k\nu a}. \qquad (5)$$

While preparing this article, we became aware of [9] where the Zak transform has already been suggested for speech analysis (motivated by the theory of cyclostationary processes).

Assuming noise–free observation of Hp_a one has the following "anti–aliasing" condition for the Zak–based estimate of S_H: $2\tau_0 \leq a, 2\nu_0 \leq \frac{1}{a}$. For general underspread systems one has considerable freedom in the selection of a. In the specific context of speech, we have to assume that the impulse response of the vocal tract does not exceed the pitch period.

In practice, the exact knowledge of the excitation pulse is certainly unrealistic. Detailed analysis shows that a time–shift of p_a results in a time shift of the magnitude of $\mathcal{Z}Hp_a$, while the complex phase is changed in a "twisted" way. But this ambiguity is not really a problem, because we are content with an incomplete characterization of the vocal tract. One can show that the identification of the composite operator HH^* is invariant with respect to an (unknown) time–shift of p_a, hence $|S_{HH^*}|$ characterizes the spreading behavior of the vocal tract in a reproducible way. Note, moreover, that according to the innovations system interpretation (switching to white noise excitation) the spreading function of HH^* is the TF–correlation function (expected ambiguity function) of the speech signal [10].

In summary, we suggest the following estimation procedure for the TF correlation of the speech signal f:

1. Compute the pitch–adapted Zak transform of f (see (5)).
2. Apply the inversion formula of the spreading function to obtain the kernel of H
   $k_H(x, y) = \int \mathcal{Z}f(x - y, \nu) e^{j2\pi\nu x} d\nu.$
3. Build the composite operator HH^* according to
   $k_{HH^*}(x, y) = \int k_H(x, z) \overline{k_H(y, z)} dz.$
4. Compute the spreading function of HH^* (see (2)).

In the following numerical experiment we compare the Zak based estimate of the speech signal's TF correlation with an averaged ambiguity function estimate as proposed in [11]. The speech sample was 0.250msec of voiced speech, sampled at 8Khz. The pitch period was estimated from a standard autocorrelation estimate. The signal was pre–emphasized. Figure 2 shows the two different TF correlation estimates. Averaging the ambiguity functions of successive blocks neglects the quasicyclostationarity of speech, which leads to a temporal broadening of the TF correlation estimate. Hence, we prefer the Zak–based estimate.

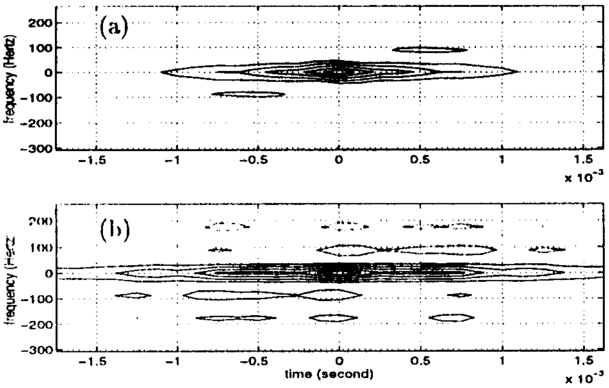

Figure 2: TF correlation estimates of voiced speech: (a) Via Zak–transform, (b) Via averaged ambiguity functions

5. NONSEPARABLE GABOR FRAMES

Sampling reduces the inherent redundancy of the STFT in an obvious, deterministic way. However, since we model the signal as a realization of a nonstationary process, we require that the chosen STFT samples are optimally uncorrelated. The covariance of the STFT $V_g f$ is basically a four–dimensional function:

$$r_V(x, \xi, \tau, \nu) = \mathrm{E}\left\{V_g f(x, \xi) \overline{V_g f(x - \tau, \xi - \nu)}\right\}.$$

However, for intuitive purposes it is sufficient to consider the following "TF–stationary" upper bound of the STFT correlation:

$$|r_V(x, \xi, \tau, \nu)| \leq (|S_R| * |S_{g \otimes g}|)(\tau, \nu). \qquad (6)$$

Here, S_R is the signal process' TF correlation, and $S_{g \otimes g}$ is the ambiguity function of the analysis window (the asterisk denotes double convolution). We emphasize that incomplete a priori knowledge in the form of $|S_R|$ already suggests the use of a TF–coherent frame, because no absolute TF localization is distinguished in a statistical sense. TF–coherence is obtained by sampling the STFT on a lattice (subgroup of $\mathbb{R} \times \mathbb{R}$) which need not be separable [12]. A general lattice can be defined by a 2 × 2 sampling matrix $(\mathbf{u}, \mathbf{v})$ (see Fig.3(b)), the Gabor coefficients are then given by

$$T_g f[k,l] = V_g f(ku_1 + lv_1, ku_2 + lv_2).$$

In view of (6) it may be expected that knowledge of $|S_R|$ determines the optimum grid and the optimum window. For underspread processes it can be shown that this is indeed true. Recall that optimality means that the covariance of the Gabor coefficients, $E\{T_g f[k,l]\overline{T_g f[m,n]}\}$ is close to diagonal. The optimization theory largely parallels that presented for the separable case in [11, 5]. Invertibility is a key side–constraint, because reducing either the sampling density or the norm of the window, reduces the off–diagonal contributions in a trivial way that violates invertibility. The density of a general lattice is given by the inverse determinant of the generating matrix, hence, the side constraints for minimizing the Gabor coefficient correlation are: (i) $\det(\mathbf{u}, \mathbf{v}) < 1$, (ii) $\|g\| = 1$. The reconstruction itself is structurally equivalent to the analysis and can be realized via FFT methods:

$$f(x) = \sum_{k \in \mathbb{Z}} \sum_{l \in \mathbb{Z}} T_g f[k,l] h(x - ku_1 - lv_1) e^{j2\pi x(ku_2 + lv_2)},$$

where h is the synthesis window.

The *joint* optimization of lattice and window is a fairly complicated problem which seems to be analytically intractable. However, a *successive* optimization leads to closed form analysis when we restrict ourselves to elliptical symmetry of $|S_R|$. (Recall that our experiments suggest that the support of $|S_R|$ for speech signals is closer to an ellipse than to a rectangle.) Due to lack of space, we cannot go into more mathematical details, rather we emphasize that both orders of optimization suggest the use of nonseparable lattices for speech signals by the following arguments:

From window to lattice. It can be shown [5], that the statistically optimum STFT window for an underspread process with elliptical spreading constraint is the Gaussian window, adapted according to (3). Numerical experiments (and an intuitive sphere packing consideration) suggest that the nonseparable grid obtains better frame bounds given the Gaussian window and fixed density [12].

From lattice to window. A window independent matched grid can be obtained by the theory of underspread operators [5], for elliptical symmetry of the spreading constraint it is hexagonal.

Numerical Experiment. As an illustrative example we consider a simple numerical experiment based on 1sec speech sampled at 8kHz and pre–emphasized. We computed a Frobenius–type off–diagonal norm of the Gabor coefficient block covariance for various Gabor frame setups. The numerical results are listed in Table 1. Note that this average goes over 1sec speech containing both voiced and unvoiced parts.

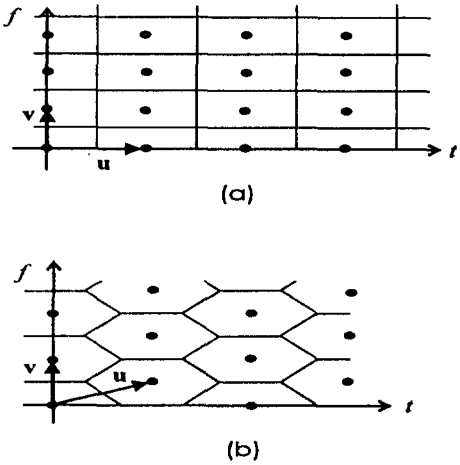

Figure 3: Separable/nonseparable tilings of the TF plane.

Gabor frame setup	Glob.-Corr.
Matched hexagonal grid	1.38
Matched separable grid	1.66
Mismatched separable grid	7.2

Table 1: Numerical comparison of different Gabor frames

6. REFERENCES

[1] P. Noll. Wideband speech and audio coding. *IEEE Communications Magazine*, 31(11):34–44, 1993.

[2] M.R. Portnoff. Time-frequency representation of digital signals and systems based on short-time Fourier analysis. *IEEE Trans. Sign. Proc.*, 28:55–69, 1980.

[3] H.G. Feichtinger and K. Gröchenig. Gabor wavelets and the Heisenberg group: Gabor Expansions and Short Time Fourier transform from the group theoretical point of view. In C.K. Chui, editor, *Wavelets – A Tutorial in Theory and Applications*, pages 359–397. Academic Press, Boston, 1992.

[4] D. Gabor. Theory of Communication. *J. IEE (London)*, 93(III):429–457, November 1946.

[5] W. Kozek. *Matched Weyl-Heisenberg expansions of nonstationary environments.* PhD thesis, Vienna University of Technology, 1996.

[6] W. Kozek. On the transfer function calculus for underspread LTV channels. *to appear in IEEE Trans. Signal Processing*, 45(1), January 1997.

[7] P. Liebermann. *Speech Physiology and Acoustic Phonetics: An Introduction.* Mac Millan Publishing Company, New York, 1977.

[8] A.J.E.M. Janssen. The Zak Transform: A Signal Transform for Sampled Time-Continuous Signals. *Philips J. Res.*, 43(1):23–69, 1988.

[9] G. Kubin. Voice Processing–Beyond the Linear Model. In *Proc. PRORISC/IEEE Workshop on Circuits, Systems and Signal Processing*, pages 393–400, Mierlo (NL), 1996.

[10] G. Matz, F. Hlawatsch, and W. Kozek. Generalized evolutionary spectral analysis and the Weyl spectrum of nonstationary random processes. *to appear in IEEE Trans. Signal Processing*, 1997.

[11] W. Kozek. Matched generalized Gabor expansion of nonstationary processes. In *Proc. IEEE Int.Conf. Signals, Systems and Computers*, pages 499–503, Pacific Grove (CA), 1993.

[12] H.G. Feichtinger, O. Christensen, and T. Strohmer. A group-theoretical approach to Gabor analysis. *Optical Engineering*, 34(6):1697–1704, 1995.