IEEE International Conference on Acoustics Speech and Signal Processing 2001 ICASSP 2001; Salt Lake City, Utah; May 7-11, 2001

OPTIMAL ESTIMATION OF SUBBAND SPEECH FROM NONUNIFORM NON-RECURRENT SIGNAL-DRIVEN SPARSE SAMPLES

Penio S. Penev[*]

The Rockefeller University
1230 York Avenue, New York, NY 10021
PenevPS@IEEE.org
http://venezia.rockefeller.edu/

Liubomire G. Iordanov

Department of Computer Science
University at Albany
State University of New York
1400 Washington Avenue Albany, NY 12222
lou@cs.albany.edu

ABSTRACT

Speech signals are comprised of auditory objects that are localized in time, but can appear anywhere in the record. We introduce a strategy for non-recurrent irregular signal-driven sampling and subsequent maximum likelihood interpolation of speech subbands that achieves object constancy—the representation of an auditory object is precisely locked to the timing of its features, but is otherwise constant. Moreover, the reconstruction fidelity can be traded flexibly for sampling rate, over a broad range of signal-to-noise ratios and application requirements. In an experiment with wide-band speech, we find a regime in the rate/distortion curve that has almost perfect reconstruction at a rate substantially lower than the respective Nyquist rate.

1. INTRODUCTION

In the *filter bank* approach, when band limited speech is filtered by M linear time-invariant (LTI) systems it can be exactly reconstructed from the uniformly spaced samples of their outputs at $1/M$-th of the Nyquist rate [13]. The *block transform* approach is a special case in which the temporal support of the filters—the block length— is small enough so that there is one sample per block per filter. In both approaches, a given auditory object can have very different representations, depending of its phase relative to the start of the respective sampling period.

Here we consider *nonuniform sampling* that is locked to the features of the signal; hence, it has the *object constancy* property—the "what" and "when" parts of the infomation in the signal, which are presumably independent, are decoupled as early as the sampling stage.

Several extensions to the uniform sampling theorem are well known [10]. Specifically, it is established that a band-limited signal can be recovered from its nonuniform samples, provided that the average sampling rate exceeds the Nyquist rate [2]. Also, an explicit reconstruction formula for the general case is available [20].

Nevertheless, reconstruction from both generalized uniform [13], and nonuniform [20], samples is computationally complex and typically involves iterative algorithms [11, 4]. An efficient filter bank implementation has recently become available for the special case of nonuniform, but *recurrent* sampling [12]. Here we provide a non-iterative reconstruction algorithm for a class of *non-recurrent* nonuniform sampling strategies.

Another consideration taken into account is that the requirement for exact reconstruction is seldom enforced in practice. Indeed, real physical signals are always embedded in noise, which renders exact reconstruction meaningless. Also, the sampling process itself introduces both amplitude quantization noise and temporal jitter. Most importantly, sampling of speech is typically followed by quantization schemes which deliberately loose fidelity [*e.g.*, 7, 9, 18]. Therefore, here we consider a practically useful approach that provides a flexible tradeoff between sampling rate and reconstruction quality, for which perfect reconstruction is but one of the operating regimes.

In this paper, we construct the filter bank as follows: first we define a probability model in terms of block-transform coefficients to calculate the *entropy* of the signal within any given window, and then we group sets of those coefficients in a finite number of *subbands;* for each one, we use the block-transform bases in the respective set to construct an *analysis filter* such that the subband *projector* is given by the convolution with the autocorrelation function of the ensemble of filtered signals.

Further, for any signal, given any irregular set of samples from the subband output, the respective projector is used to construct *maximum likelihood interpolators.* The samples are selected with a *greedy projection pursuit* algorithm, which locks them to the features of the signal, and also ensures stability and locality.

This irregular sampling method is applied to an ensemble of wide-band speech to build *rate/distortion curves* for a given subband. In one regime, 99% of the information in the subband is reconstructed at $1/4$-th of the respective Nyquist rate.

2. BLOCK-TRANSFORM CODES: A GAUSSIAN MODEL

A long record of monaural sound of duration $\tilde{T}$ will be denoted by $s(t)$.[1] The part of the record that falls in the window of dura-

[*] corresponding author, current address: NEC Research Institute, 4 Independence Way, Princeton, NJ 08540.

The major part of this research was made possible by the William O' Baker Fellowship, so generously extended to, and gratefully accepted by, PSP. He is also indebted to M. J. Feigenbaum for his hospitality and support—scientific and otherwise. We are also thankful to A. J. Libchaber for useful discussions and comments, and to M. Sondhi, A. V. Oppenheim, B. Kleijn, F. A. Marvasti, and D. Donoho, for their encouraging remarks.

[1] For this study, a news broadcast was recorded trough a cable box by a monaural VCR on VHS tape; the audio channel was later digitized without clippings to 16 bits/sample at 16 kSa/sec on a *SGI Indy* workstation with

tion V which starts at time $0 \leq t < T \triangleq \tilde{T} - V$, will be denoted by $\phi^t(\tau) \triangleq s(t+\tau)$, $\tau \in V$.[2]

For a given dimensionality N, the *analysis phase* of block-transform coding utilizes a set of *global bases* $\{\psi_r(\tau)\}_{r=1}^N$; global here means that they are non-vanishing throughout the entire window. For an arbitrary $\phi^t(\tau)$, the N *transform coefficients* are calculated by the linear projections

$$a_r^t \triangleq \sigma_r^{-1} \frac{1}{V} \sum_{\tau} \psi_r(\tau)\phi^t(\tau) \tag{1}$$

where $\{\sigma_r\}$ are chosen so that $\{a_r\}$ have *unit variance.* When the filters are *orthonormal,* they are also used in the *synthesis phase;* then the respective *reconstruction* is

$$\phi_N^{rec}(\tau) \triangleq \sum_{r=1}^{N} a_r \sigma_r \psi_r(\tau). \tag{2}$$

When there is no overlap between adjacent windows, $V = N$ must hold for *perfect reconstruction* [19].

An important special case is the Karhunen-Loève transform (KLT) [see *e.g.,* 3, 14] , which utilizes the *spectral decomposition* of the *covariance matrix* of the ensemble

$$R(\tau,\tau') \triangleq \frac{1}{T} \sum_{t} \phi^t(\tau)\phi^t(\tau') = \sum_{r=1}^{V} \psi_r(\tau)\sigma_r^2\psi_r(\tau'). \tag{3}$$

KLT is *optimal* in the sense that the transform coefficients (1) are *decorrelated.*[3] Then, under the standard *multidimensional Gaussian* model for the probability density $\mathcal{P}[\phi]$, the *entropy* of the reconstruction (2), which is also the *optimal code length* [17], is

$$-\log \mathcal{P}[\phi_N^{rec}] \propto \sum_{r=1}^{N} |a_r|^2. \tag{4}$$

Notably, because of the normalization by σ_r in (1), often called *whitening,* the model (4) is *spherically symmetric.* With an appropriate model of the noise, such whitening has been found to account for the psychophysically-measured contrast sensitivity of human vision in all signal-to-noise regimes [1]. In speech applications, σ_r is often heuristically chosen as the basis for the respective quantizer step [19].

the *Iris Audio Processor: version A2, revision 4.1.0.* The record was segmented manually and the 24 segments with the voice of the anchor woman were reassembled, for a total duration of $\tilde{T} = 12,330,080$ Sa ≈ 13 min. With a single affine transformation, the DC component was eliminated, and the samples were converted to IEEE floats in the $[-1, 1]$ range. No attempt was made to apply additional analog pre-filters, or correct for the (colored) noise of the equipment and the occasional fade-ins of background music.

[2]It is well known that processing of a band-limited signal with a continuous-time filter bank is equivalent to first sampling at the Nyquist rate, and then processing with a suitable discrete-time bank [see *e.g.,* 12]; hence, without loss of generality, t can be considered to be a discrete-time index. In practical implementations however, it might be advantageous to do some or most of the pre-filtering in the analog domain.

[3]Notably, for time-invariant ensembles, $R(\tau,\tau')$ is a *Toeplitz* matrix whose diagonals are given by the *autocorrelation function* of the ensemble, $R(\tau,\tau') \equiv \rho(\tau - \tau')$; thus, sines are asymptotically close to its eigenvectors [6], and the DFT and DCT are both asymptotically optimal [3]. Since DFT and DCT are far from optimal for small windows sizes, V, the calculations here were carried out by KLT. Nevertheless, in the limit of infinitely large windows, $V \to \infty$, the presented results are invariant under the choice of any asymptotically optimal block-transform basis.

3. LOCAL ENTROPY DENSITY WITHIN SUBBANDS

Two serious problems with the block-based coding are the global processing, which leads to large latency and complexity, and the edge effects—the transform coefficients (1) of a localized speech object will change in a non-trivial, but purely artifactual, manner as a function of the distance to the block boundary. In order to cope with essentially the same problem in the context of object vision, the method of *Local Feature Analysis (LFA)* has been developed [16, 14]. LFA utilizes a set of *local analysis filters,* $K(\tau,\tau')$, whose outputs are indexed with τ (cf. eq. 1) and are optimally decorrelated

$$O^t(\tau) \triangleq \frac{1}{V} \sum_{\tau'} K(\tau,\tau')\phi^t(\tau'). \tag{5}$$

When the set of indices $\{r\}$ is broken up into *non-overlapping bands,*[4] $B_\alpha \cap B_\beta = \emptyset$ for $\alpha \neq \beta$, *maximal decorrelation* can be achieved in any given band B with $K(\tau,\tau') = K_B^{(1)}(\tau,\tau')$ from the following family of filters [16]

$$K_B^{(n)}(\tau,\tau') \triangleq \sum_{r\in B} \psi_r(\tau)\sigma_r^{-n}\psi_r(\tau'). \tag{6}$$

Then, the *synthesis filters,* $K_B^{(-1)}(\tau,\tau')$, provide reconstructions

$$\phi_B^{rec}(\tau) = \sum_{r\in B} a_r\sigma_r\psi_r(\tau) = \frac{1}{V}\sum_{\tau'} K_B^{(-1)}(\tau,\tau')O_B(\tau') \tag{7}$$

that are exactly equal to the respective global ones (2).

Evidently in Fig. 1, the filters from the family (6) are *local*—their support is substantially non-vanishing only in a small neighborhood around their centers.[5]

Notably, the subband outputs $O_B(\tau)$ (5) are essentially an *orthogonal embedding* of the transform coefficient (1) in the time domain, and provide the *local entropy density* in the respective subband (cf. eq. 4)

$$-\log \mathcal{P}[\phi_B^{rec}] \propto \sum_{r\in B} |a_r|^2 = \frac{1}{V}\sum_{\tau} |O_B(\tau)|^2. \tag{8}$$

4. A FILTER-BANK IMPLEMENTATION

The filters (6), shown in Fig. 1, although operating within a window, have one desirable property—their support is essentially *local,* and they don't feel the window boundaries. Moreover, evidently in Fig. 2, the filters centered over different places within

[4]The typical reason for the breaking up of speech in subbands is that they have different properties, and judicious coding can take advantage of this fact [*e.g.,* 19]. Also, the, tonotopically organized [5], human cochlea analyzes the signal in bands of neighboring *frequencies.* For the experiments here, we build the subbands by *energy*—we group together indices with similar σ_r (3). Part of the motivation is that the dropping of the weaker half of the the bands does not produce any noticeable distortion for the wide-band speech signals considered here; another 20% can be dropped, which results in *just noticeable distortion* [8]. Such *global dimensionality reduction* is not a subject of this paper; what we study in Section 5 is *local* dimensionality reduction—within a subband that needs to be preserved.

[5]Notably, $K^{(-2)}(\tau,\tau') \equiv R_B(\tau,\tau')$, the projection of the covariance matrix to the subband. $K^{(2)}(\tau,\tau') \equiv R_B^{(-1)}(\tau,\tau')$, used as the predictor in *Differential PCM (DPCM)* [7], is conspicuously absent; for reasons that will become apparent in Section 5, $K^{(0)}(\tau,\tau') \equiv P_B(\tau,\tau')$ (9), the *projector* to the subband, is used as the predictor there.

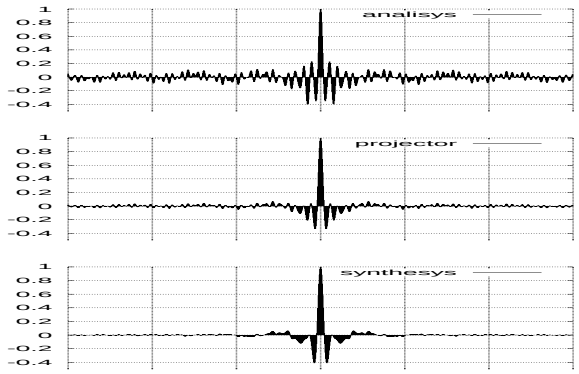

Fig. 1. The filters (6) in the band with the most energy (3) are shown for a window size of $V = 60$ ms and dimensionality rate $N/V = 2000$ Nyquist coefficients/sec; the center, τ, is in the middle of the window; τ' runs horizontally throughout the full window. *(top)* analysis filter, $K(\tau, \tau')$; *(middle)* autocorrelation, $P(\tau, \tau')$ (9); *(bottom)* synthesis filter, $K^{(-1)}(\tau, \tau')$. P is in its natural scale. K and $K^{(-1)}$ convert from energy to entropy and back, respectively, and have some physical units; hence, they are normalized by their central values. Notably, K is a derivative, $K^{(-1)}$ is an integration, and P is a projection operator.

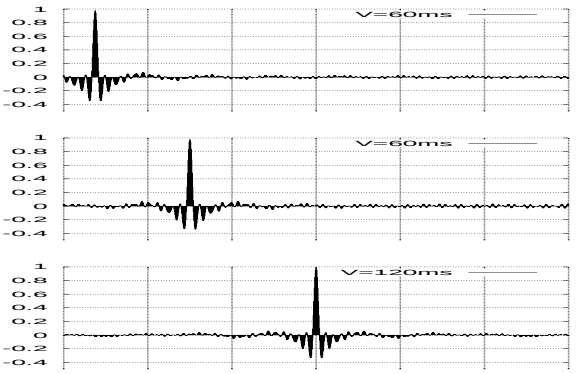

Fig. 2. Projectors $P(\tau, \tau')$ for the band in Fig. 1 for two values of τ that are placed asymetrically in the window *(top, middle)*. The central 60 ms are shown of the projector at the center of the window for a band with $V = 120$ ms and the same dimensionality rate *(bottom)*; cf. Fig. 1*(middle)*.

the window are, essentially, translated versions of each other—$P(\tau, \tau') \approx \pi(\tau - \tau')$. Hence, the output due to some localized structure will be independent of its placement within the window.

Although the time invariance of the filters is only approximate—especially evident close to the boundaries, shown in Fig. 2*(top)*—when the window size, V, is increased at a fixed *dimensionality rate*, N/V, evidently from Fig. 2*(bottom)*, the structure of the filters persists in the center, and their support gets more localized. Hence, the fraction of the filters that feel boundary effects decreases with increasing window size.

In the limit $V \to \infty$, (5) is equivalent to the subband processing in a filter bank, whereby the analysis and synthesis are carried out by convolutions with the respective limits of the filters (6), with τ taken at the center of the window.[6]

5. SPARSIFICATION OF SUBBAND SPEECH

Within any subband B, with dimensionality $N = |B|$, the subband entropy density $O_B(\tau)$ is optimal in the sense that its *residual covariance* (cf. eq. 3), shown in Fig. 1*(middle)*

$$P_B(\tau, \tau') \stackrel{\triangle}{=} \frac{1}{T} \sum_t O_B^t(\tau) O_B^t(\tau') = K_B^{(0)}(\tau, \tau') \qquad (9)$$

is as close to $\delta(\tau, \tau')$ as possible [16]. Nevertheless, when $N < V$, the subband outputs can not be completely decorrelated.

Moreover, since there are only N *degrees of freedom* in (6), any $N + 1$ variables are *linearly dependent*. Hence, $O(\tau)$ can be *subsampled* over a limited set of points within the window, $\{\tau_m\} \equiv \mathcal{M}$, and subsequently *linearly reconstructed* from its $|\mathcal{M}| \leq N$ samples $\{O_m \equiv O(\tau_m)\}_{\tau_m \in \mathcal{M}}$ by

$$O^{rec}(\tau) = \sum_{m=1}^{|\mathcal{M}|} O_m a_m(\tau) \qquad (10)$$

where $\{a_m(\tau)\}$ depend on the choice of $\mathcal{M}$ [16]. When they satisfy $a_n(\tau_m) \equiv \delta_{nm}$, they are *interpolating reconstructors*. In this case, with the probability model (8), it has been shown [14, Section 4.4] that the *maximum likelihood* estimate is

$$a_m(\tau) = \sum_{n=1}^{|\mathcal{M}|} \mathbf{Q}^{-1}{}_{mn} P_n(\tau) \qquad (11)$$

where $P_m(\tau) \equiv P(\tau_m, \tau)$, and $\mathbf{Q} \equiv \mathbf{P}|_{\mathcal{M}}$ is the *restriction* of $\mathbf{P}$ on the set of sampling points, with $\mathbf{Q}_{nm} = P_n(\tau_m)$. In the representation (10–11), both the values of the set of $|\mathcal{M}| \leq N$ samples, $\{O(\tau_m)\}_{\tau_m \in \mathcal{M}}$, and their locations, $\mathcal{M}$, code for $O^{rec}(\tau)$ and, through (7), for $\phi^{rec}(\tau)$.

For the special case when the subband is comprised of all frequencies within a range symmetric around the origin, the projector $\pi(\tau - \tau')$ is the familiar sync function. Additionally, when the sampling $\mathcal{M}$ is determined by *decimation*, $\mathbf{Q} \equiv \delta$ and the reconstructors are all identical, again to the sync function. In general, the selection of $\mathcal{M}$ should be driven by the signal itself. We call this process *sparsification*.

The optimal sparsification of sound, as well as images, is an open question [14]. Here, we apply to sound a strategy from object vision that was found to be extremely efficient—window-based *greedy sparsification* [16]. Iteratively within any given window, we start with the empty set $\mathcal{M}^{(0)} = \emptyset$ and, at step $n + 1$, add to $\mathcal{M}^{(n)}$ one sampling point, τ_{n+1}, until a termination criterion is met. To decide where to sample, given the current set $\mathcal{M} = \mathcal{M}^{(n)}$, we calculate the current maximum-likelihood estimate $O_n^{rec}(\tau)$ (10–11) and error

$$O_n^{err}(\tau) = O(\tau) - O_n^{rec}(\tau). \qquad (12)$$

[6] In practice, there is a finite signal-to-noise ratio; also, the outputs are typically quantized for subsequent transmission and/or storage. Hence, a true limit is not required, as long as the error due to the finite window size does not dominate the other errors. A measure of the *information leakage* from the subband is $1 - (\pi, \pi * \pi)/|\pi|^2$, where $\pi * \pi$ is the convolution of the projector with itself. For all filters here, this leakage was $< 0.4\%$.

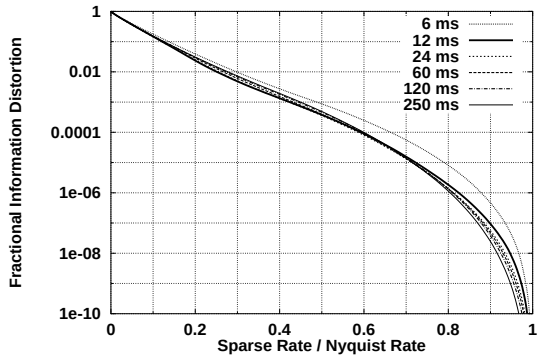

Fig. 3. For a given information fidelity (12), the ratio of the number of Nyquist, global (2), samples and sparse, local (10), samples necessary to achieve that fidelity—is shown for bands of a fixed Nyquist rate of 500 Sa/sec and window sizes $V \in \{6, 12, 24, 60, 120, 250\}$ ms.

We find the grid point τ_{n+1} such that the value of $|O_n^{err}(\tau_{n+1})|^2$ is maximal and terminate the sparsification whenever it is below a predetermined threshold. Otherwise, we define $\mathcal{M}^{(n+1)} \equiv \mathcal{M}^{(n)} \cup \tau_{n+1}$ and continue the iteration. This algorithm terminates—it samples only at places where the entropy is not predicted well; hence its samples are linearly independent, and when $|\mathcal{M}| = N$, the interpolation (10) is exact.

The results of such sparsification of speech are shown in Fig. 3.[7] Remarkably, the *information fidelity* of the reconstruction increases rapidly with the number sparse samples in the subband: for 90% fidelity, only 12% of the Nyquist rate is needed; 25%, for 99%. Notably, what is captured here is the *information*, rather than the energy of signal; although formal listening tests are needed to establish its relationship to *perception*, it has been shown in the context of object vision that $\approx 95\%$ of the information is needed at the just-noticeable distortion threshold [14, 15].

6. DISCUSSION

We have shown here that, when the analysis filters in a filter bank are specially designed to calculate, in a time-invariant fashion, the local entropy density, and when its local structure is used to drive the sub-sampling process, the resulting sparse-distributed representation is of very low dimensionality. This is due to the sparse structure of the signal itself.

There are a number of open questions that arise in this irregular sampling framework, both practical and theoretical, that merit further investigation: the link between perception, and both the number of local terms and their quantization; the influence of the signal-to-noise ratio in the subband on both sparseness and perception; the possible use of the unitary degree of freedom implicit in the orthogonal embedding, for minimization of latency and/or maximization of sparseness.

[7] Although the sparsification here was carried out within each window independently, evidently from Fig. 3, the final result depends only weakly on the window size V, as long as the *dimensionality rate*, N/V is kept constant, and V is "large enough." We expect this result to also hold for truly windowless algorithms, such as *successive sparsification* [14, Section 5.5].

A very interesting possibility is to use an orthogonal embedding that introduces residual correlations *between* adjacent subbands [14, Section 5.4], which is a strategy, used by the human cochlea [see, *e.g.*, 8, 5, for reviews]. Although such *spectral leakage* has been postulated to be undesirable [19], the predictive step of the sparsification (10) can work across subband boundaries and suppress the leakage on the basis of the correlations (9). Then, a large part of the information can be accounted for with a very small number of in-band samples, say 90%, with 10% (Fig. 3) and the rest will be accounted for by the adjacent subbands.

7. REFERENCES

[1] J. J. Atick and A. N. Redlich. What does the retina know about natural scenes? *Neural Comput.*, 4(2):196–210, 1992.

[2] F. J. Beutler. Error-free recovery of signals from irregularly spaced samples. *SIAM Review*, 8(3):328–335, July 1966.

[3] S. J. Campanella and G. S. Robinson. A comparison of orthogonal transformations for digital speech processing. *IEEE Trans. Commum. Technol.*, 19(6):1045–1050, December 1971.

[4] C. Cenker, H. G. Feichtinger, and M. Herrmann. Iterative algorithms in irregular sampling: A first comparison of methods. In *Proc. ICCCP*, pages 483–489, Phoenix, AZ, 1991. IEEE.

[5] R. Fettiplace and P. A. Fuchs. Mechanisms of hair cell tuning. *Annual Review of Physiology*, 61:809–834, 1999.

[6] U. Grenander and G. Szegö. *Toelplitz Forms and Their Applications*. Chelsea Publishing, New York, 2nd edition, 1984.

[7] N. S. Jayant. Digital coding of waveforms: PCM, DPCM, and DM quantizers. *Proc. IEEE*, 62:611–632, May 1974.

[8] N. S. Jayant, J. D. Johnston, and R. Safranek. Signal compression based on models of human perception. *Proc. IEEE*, 81:1385–1422, October 1993.

[9] N. S. Jayant and P. Noll. *Digital Coding of Waveforms: Principles and Applications to Speech and Video*. Prentice Hall, Englewood Cliffs, NJ, 1984.

[10] A. J. Jerri. The Shannon sampling theorem—its various extensions and applications: A tutorial review. *Proc. IEEE*, 65(11):1565–1598, November 1977.

[11] F. Marvasti, M. Analoui, and M. Gamshadzahi. Recovery of signals from nonuniform samples using iterative methods. *IEEE Trans. Sig. Proc.*, 39:872–877, Apr. 1991.

[12] A. V. Oppenheim and Y. C. Eldar. Filterbank reconstruction of bandlimited signals from nonuniform and generalized samples. *IEEE Trans. Sig. Proc.*, 48(10):2864–2875, 2000.

[13] A. Papoulis. Generalized sampling expansion. *IEEE Trans. Circ. Syst.*, 24(11):652–654, Nov. 1977.

[14] P. S. Penev. *Local Feature Analysis: A Statistical Theory for Information Representation and Transmission*. PhD thesis, The Rockefeller University, New York, NY, May 1998. available at http://venezia.rockefeller.edu/penev/thesis/.

[15] P. S. Penev. Redundancy and dimensionality reduction in sparse-distributed representations of natural objects in terms of their local features. In T. K. Leen and et al, editors, *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 2001. to appear.

[16] P. S. Penev and J. J. Atick. Local Feature Analysis: A general statistical theory for object representation. *Network: Comput. Neural Syst.*, 7(3):477–500, 1996.

[17] C. E. Shannon. A mathematical theory of communication. *Bell System Tech. J.*, 27:379–423, 623–656, 1948.

[18] A. S. Spanias. Speech coding: A tutorial review. *Proc. IEEE*, 82(10): 1541–1582, October 1994.

[19] J. M. Tribolet and R. E. Crochiere. Frequency domain coding of speech. *IEEE Trans. Acoust., Speech, and Signal Processing*, 27(5): 512–530, October 1979.

[20] K. Yao and J. B. Thomas. On some stability and interpolation properties of nonuniform sampling expansions. *IEEE Trans. Circ. Theory*, 14(4):404–408, Dec. 1967.