SPARSE TIME-FREQUENCY REPRESENTATIONS IN AUDIO PROCESSING, AS STUDIED THROUGH A SYMMETRIZED LOGNORMAL MODEL

Patrick J. Wolfe

Division of Engineering and Applied Sciences
Department of Statistics, Harvard University
Harvard-MIT Division of Health Sciences & Technology
Oxford Street, Cambridge, MA 02138 USA
patrick@seas.harvard.edu

ABSTRACT

Time-frequency representations are ubiquitous in speech and audio signal processing, their use being motivated by both auditory physiology and the mathematics of Fourier analysis. Nonparametric statistical models (or equivalently transform based signal processing methods) formulated in this space provide a principled way to decompose sounds into their constituent parts, as well as an effective means of exploiting the local correlation present in the time-frequency structure of naturally generated acoustic signals. Here we describe how an appropriate generative statistical model, even under very simple assumptions, provides a means of exploring sparse time-frequency representations in audio. We introduce a symmetrized lognormal model for spectral coefficients, which shows good agreement across a broad range of speech samples taken from the TIMIT database, and demonstrate preliminary speech enhancement results based on a maximum a posteriori shrinkage estimator.

1. INTRODUCTION

Time-frequency representations are ubiquitous in speech and audio signal processing, their use being motivated by both auditory physiology and the mathematics of Fourier analysis. Indeed, information-carrying natural sound signals can often be conveniently characterized as a superposition of simple, well-understood mathematical building blocks.

In combination with so-called nonparametric statistical models formulated in the time-frequency plane, such an approach provides a principled way to decompose sounds into their constituent parts, as well as an effective means of exploiting the local correlation present in the time-frequency structure of naturally generated acoustic signals. At the same time, however, more structured representations offer the hope of truly generative statistical models for audio analysis and synthesis tasks, but at the price of less universal model elicitation and more complex model fitting procedures.

In conjunction with a companion article featured in this EUSIPCO special session, this paper approaches these trade-offs simultaneously from the points of view of signal processing and statistics, yielding a common framework intended to shed light on a variety of techniques in the literature. This framework is then used to describe the ways in which an appropriate generative statistical model, even under very simple assumptions, provides a means of exploiting sparse time-frequency representations in speech and audio signal processing.

2. SHORT-TIME AUDIO SIGNAL PROCESSING

The scenario we consider here stems from the standard method of "short-time" audio signal processing, in which the signal under consideration is decomposed according to the principles of Gabor analysis (i.e., the subsampled short-time Fourier transform) over finite cyclic groups [2]. Simply put, this viewpoint is a formalization of the tried-and-true overlap-add method commonly used for short-time audio signal analysis and synthesis (see, e.g., [3] for details).

To this end, we recall that the standard practice for modification of an audio times series vector $\boldsymbol{x} \in \mathbb{R}^L$ proceeds as follows: first, $\boldsymbol{x}$ is divided into overlapping segments via the multiplicative action of a (typically) smooth, real, and symmetric window $\boldsymbol{g}$ whose effective size l (typically $\ll L$) is chosen as a function of the sampling rate such that the analysis window length lies in the range of 15–40 ms, depending on the time-varying nature of the audio signal class under consideration. The discrete Fourier transform (DFT) is applied on each interval and the resultant spectral coefficients are modified according to the task at hand; the inverse DFT is then taken and a corresponding synthesis window applied to each segment. Finally, the overlapping segments are added together in an appropriately weighted manner in order to reconstitute the modified time series vector $\hat{\boldsymbol{x}}$.

2.1 Gabor Analysis and the Overlap-Add Method

As a prelude to the signal models presented below, it is helpful to understand the overlap-add procedure more formally as follows: using the pair (m,n) to denote modulation and translation indices respectively, and thus to index a (separable) lattice of points in the time-frequency plane, we may think of mapping each windowed segment of $\boldsymbol{x}$ to a corresponding short-time spectral segment, or sampled "slice" of that signal's short-time Fourier transform (STFT). In particular, this operation corresponds to a representation of $\boldsymbol{x}$ in terms of a set of *Gabor transform* coefficients $\{c_{m,n}\}$ representing a sufficiently fine tiling of the time-frequency plane. The Gabor transform is hence a sampled version of the STFT.

The so-called *Gabor analysis coefficients* are calculated as inner products of $\boldsymbol{x}$ and translated, modulated versions of some chosen analysis window as $c_{m,n} = \langle \boldsymbol{x}, \boldsymbol{g}_{m,n} \rangle$, where $\boldsymbol{g}_{m,n}$ denotes a discretized, time-frequency shifted version of a window function $g(t)$:

$$g_{m,n}(t) = g\left(t - \frac{n}{N}L\right) e^{2\pi j \frac{m}{M}t}, \quad t \in \{0,1,\ldots,L-1\}.$$

Here M and N are positive, integer lattice constants chosen according to parameters a and b (representing time and frequency sampling intervals, respectively) such that $Na = Mb = L$, the length of the vector $\boldsymbol{x}$. The corresponding *Gabor expansion* in turn provides a means of reconstructing $\boldsymbol{x}$ from its Gabor coefficients, which act as weights in the sum of translations and modulations of a dual (or *synthesis*) window function $\tilde{g}(t)$:

$$x(t) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} c_{m,n} \tilde{g}_{m,n}(t) = \sum_{m,n} \langle x, g_{m,n} \rangle \tilde{g}_{m,n}(t).$$

In the discrete-time case, we may hence denote the Gabor transform of a vector $\boldsymbol{x}$ as $\boldsymbol{c} = \boldsymbol{G}^*\boldsymbol{x}$, where $\boldsymbol{G}^*$ denotes the Hermitian transpose of the $L \times MN$ Gabor analysis matrix $\boldsymbol{G}$ having the time-frequency atom $\boldsymbol{g}_{m,n}$ as its $(m+nM)$th column, and the Gabor transform coefficients $\{c_{m,n}\}$ are written in the form of a "stacked" column vector $\boldsymbol{c}$ of length MN. Likewise, we may denote the Gabor

Work supported by DARPA under Grant HR0011-07-1-0007. The expository material contained in Section 2 of this article first appeared in [1].

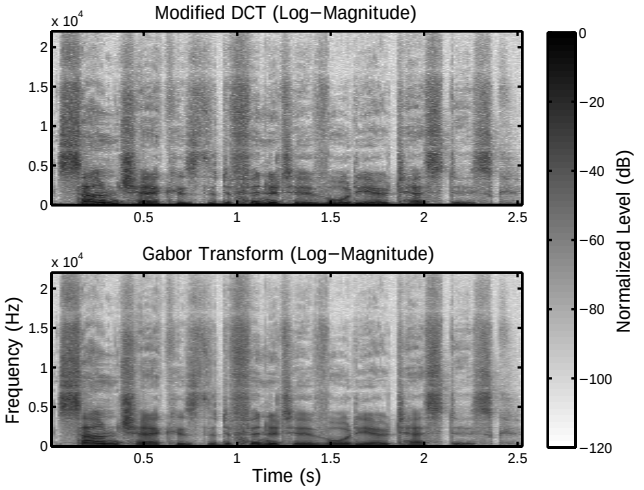

Figure 1: Comparison of the log-magnitudes of the MDCT coefficients (top) and the Gabor analysis coefficients (bottom) corresponding to a wideband speech waveform, computed using identical parameter settings and window functions

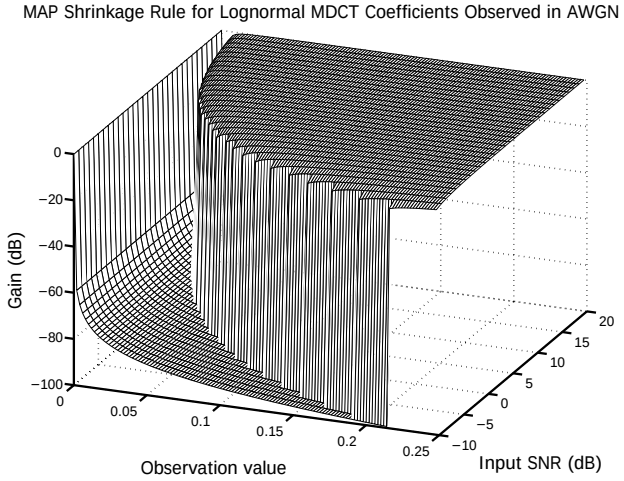

Figure 2: Realization of the maximum a posteriori lognormal shrinkage rule for coefficients of the speech signal of Figure 1, shown as a function of input signal-to-noise ratio (see Section 5)

expansion of $\boldsymbol{x}$ by $\boldsymbol{x} = \tilde{\boldsymbol{G}}\tilde{\boldsymbol{c}}$, where $\tilde{\boldsymbol{G}}$ denotes the $L \times MN$ Gabor synthesis matrix having $\tilde{\boldsymbol{g}}_{m,n}$ as its $(m+nM)$th column, and the vector $\tilde{\boldsymbol{c}}$ represents the corresponding synthesis coefficients.

2.2 Overcompleteness and Gabor Frames

We distinguish between $\boldsymbol{c}$ and $\tilde{\boldsymbol{c}}$ in the above discussion because an *overcomplete* representation admits an entire subspace of perfect-reconstruction synthesis coefficients. Indeed, if the column rank of $\boldsymbol{G}$ is equal to L, then the family $(\boldsymbol{g}, a, b)$ will form a Gabor frame with redundancy MN/L (see, e.g., [4]); i.e., the span of the set $\{\boldsymbol{g}_{m,n}\}$ is $\mathbb{C}^L$:

$$\exists A, B > 0 : \forall \boldsymbol{x} \in \mathbb{C}^L,\ A||\boldsymbol{x}||^2 \leq \sum_{m,n} |\langle \boldsymbol{x}, \boldsymbol{g}_{m,n} \rangle|^2 \leq B||\boldsymbol{x}||^2.$$

Owing to the overlap of the windowed time series segments in the scenario we consider here, we have $MN > L$ vectors in dimension L, and therefore this representation is redundant (typically by a factor of two, corresponding to use of the DFT algorithm as described earlier and a "window overlap" in time of 50%).

3. FROM PARSEVAL FRAMES TO THE MDCT

In contrast, a representation popular in the coding literature employs the *modified discrete cosine transform*, or MDCT. The MDCT is a unitary transformation that may be thought of as 1) a partition of unity applied to the time axis through a sequence of overlapping windows which yield short-time blocks, as in the case of the Gabor transform; followed by 2) a type-IV discrete cosine transform (which implicitly extends the signal at each block's boundary by introducing a symmetric extension to the left followed by an anti-symmetric extension to the right, as a means of periodic extension).

Given the differences between the MDCT and Gabor frame representations, it is natural to ask how the redundancy or "overcompleteness" of the latter affects the signal representation that we obtain. While a detailed discussion of the differences between frames and bases is outside the scope of our present enquiry, key differences center around a lack of unicity (overcomplete Gabor frames), and a lack of translation invariance (bases). Our specific interest here lies in the implications of these differences for *model-building*, and hence we shortly proceed to outline the results of an empirical investigation into differences in sparsity with respect to frames and bases for a variety of speech utterances.

To proceed with our investigation, and indeed to address the underlying questions surrounding overcompleteness, it is necessary to be able to compare frame and basis representations of waveforms in $\mathbb{R}^L$ directly. Note that while the coefficients of a redundancy-two Gabor system for $\mathbb{R}^L$ are complex-valued, we may consider them as elements of $\mathbb{R}^{2L}$ by an appropriate bijection that takes into account the conjugate symmetry of the spectra of real-valued waveforms. This enables us to compare our representations on an equal footing. In particular, one can formulate a specific choice of redundancy-two Gabor system such that the resultant frame is "as close as possible" to a lapped orthogonal transform (of which the MDCT is one example). In this case, the frame bounds A, B will be *tight* ($A = B$) and the frame will also be *normalized* ($A = B = 1$). Together these conditions describe what is known as a *Parseval frame*, suggesting its isometric (but not unitary) properties. In fact, while space constraints preclude a detailed discussion, the necessary conditions for Gabor frame tightness in the redundancy-two case we consider correspond to those which define admissible windows in lapped orthogonal transforms [5], or equivalently enforce the Princen-Bradley conditions for "time-domain aliasing cancellation" (TDAC) [6].

3.1 Comparing the Gabor Transform and MDCT

Under the equivalence conditions described above, we first consider the application of the Gabor transform and the MDCT to a wideband speech signal of duration 2.5 s, sampled at a rate of 44.1 kHz. An identical window was employed in each case, derived from a 1024-sample ($\approxeq$ 23 ms) Hanning window, by computing the closest window (in the ℓ^2 sense) satisfying the Parseval/TDAC conditions.

Comparing the (log-) spectrogram—defined as the (log-) magnitude-squared of the Gabor transform—and equivalent MDCT representation of this utterance, as shown in Figure 1, we may observe the subtle nature of their relationship. While the representations look very similar at a "macro" level (bearing in mind that in each case we have discarded the "phase" information), upon closer inspection we notice that the spectrogram as a time-frequency surface appears less rough than the log-magnitude plot of the MDCT coefficients. In fact, this smoothness may be explained by recalling that the spectrogram may also be viewed as a convolution of the speech signal's Wigner-Ville distribution with a smoothing kernel corresponding to the Wigner-Ville distribution of the chosen window. The relative roughness of the MDCT also suggests a lack of translation invariance, in contrast to the covariant property of the (continuous-time) spectrogram and other objects which admit a loose interpretation as so-called "time-frequency energy densities."

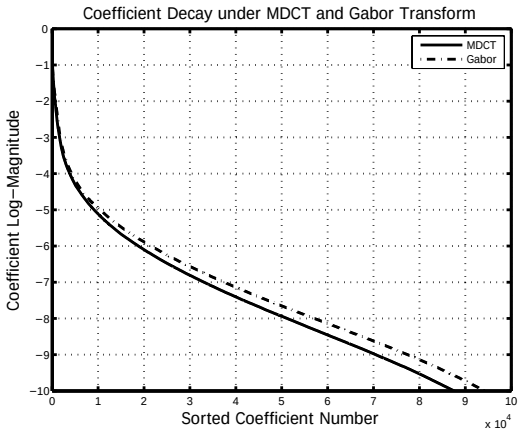

Figure 3: Coefficient log-magnitudes corresponding to Figure 1, sorted in decreasing order to reveal comparative rates of decay

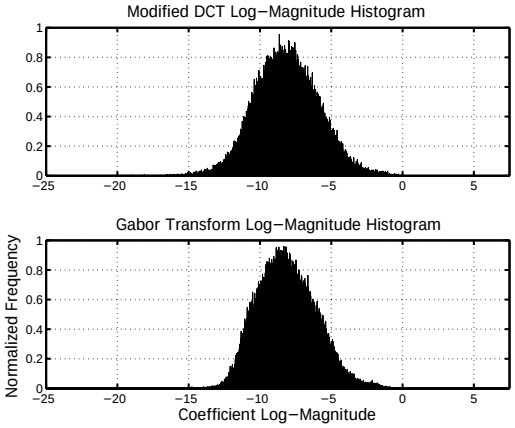

Figure 5: Histograms showing MDCT (top) and Gabor transform (bottom) coefficient log-magnitudes corresponding to Figure 1

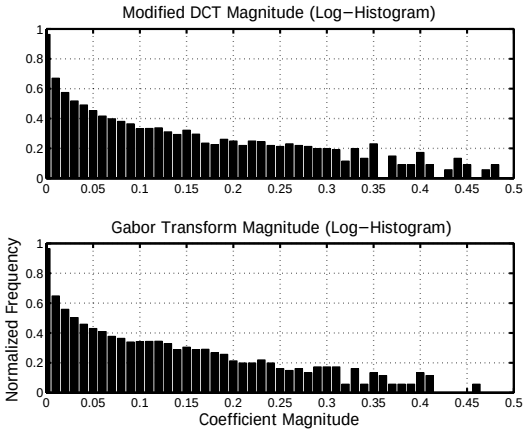

Figure 4: Comparison showing the slightly heavier tail of the empirical distribution of MDCT coefficient magnitudes for the speech signal of Figure 1 (top) compared to the Gabor transform (bottom)

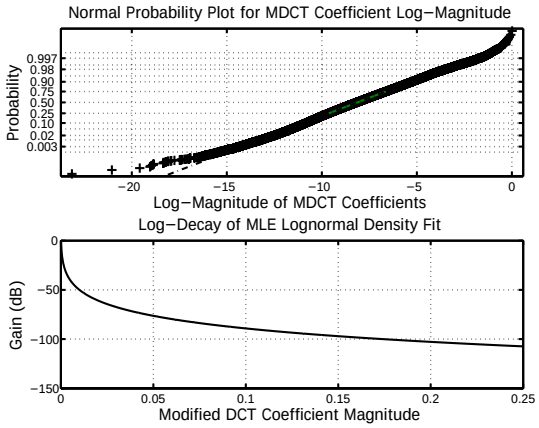

Figure 6: Normal probability plot of MDCT data of Figure 5 (top); log-decay of the corresponding best-fit magnitude density under maximum-likelihood lognormal parameter estimation (bottom)

3.2 Implications for Sparse Signal Models

Considering the above discussion, it is natural to wonder what implications these differences might have for sparse signal models. We formulate the question as follows: If a "sparse" signal model is taken to be one under which the sorted coefficient magnitudes exhibit rapid decay, then which representation—the Gabor transform or the MDCT—will yield the fastest decay rate? Some empirical insight into this question is given by Figures 3 and 4, which characterize the decay of the coefficients that comprise Figure 1. These figures suggest a slightly heavier coefficient "tail" under the MDCT representation, relative to the short-time Fourier transform.

4. MODEL ELICITATION

If we are willing to consider coefficients as random variables and hence imbue them with prior distributions, then a sparse signal model will require that *most* of the coefficients are small ("mass near zero"), while just a few coefficients capture the vast majority of the signal's energy ("heavy tails"). In this sense, the histograms of Figure 4 suggest that an appropriate specification can capture the idea that coefficients will be "sparse" for a given signal class—akin to the characterization of Besov spaces by wavelet coefficient decay, and its use in deriving nonlinear shrinkage rules for wavelet regression. Indeed, the intuition behind this line of reasoning has long been exploited by practitioners in the speech and audio processing community, through simple time-frequency shrinkage estimators such as spectral subtraction and its numerous variants [7].

In fact, by positing an independent and identically distributed (i.i.d.) *generative* model for audio signals (suggesting exchangeability and invariance under permutation in the coefficient domain), we may interpret our histograms as suggesting (at least empirically) a form of prior distribution. In particular, our preliminary investigations with the well-known TIMIT speech utterance database suggest that a so-called *lognormal* model, in which the logarithm of coefficient magnitude is taken to be Normally distributed, fits reasonably well across a wide range of examples, in comparison with the Exponential, Generalized Gaussian, Student's t, log-Rayleigh, and other common "heavy-tailed" or super-Gaussian distributions.

To this end, Figure 5 shows a comparison of log-magnitude coefficient histograms under both the MDCT and the Gabor transform, from which it can be seen that the lognormal model appears qualitatively reasonable. The slightly super-Gaussian tail behavior of these coefficients is captured by the Normal probability plot in the top portion of Figure 6, and the resultant *magnitude* density fitted via maximum likelihood is shown in its lower portion (cf. Figure 4).

Of course, regardless of which distribution and coefficient domain we choose to adopt, this model is only defined on the coefficient *magnitudes*—leaving open the question of how to model the coefficient *phases*. (We may interpret the MDCT coefficients as having a phase restricted to $\theta = 0$ or $\theta = \pi$.) Fortunately, we may specify the magnitude and phase models separately for each short-time spectral (Gabor transform or MDCT) component, and take the joint distribution resulting from our separable prior model to represent our (potentially sparse) coefficient prior model overall.

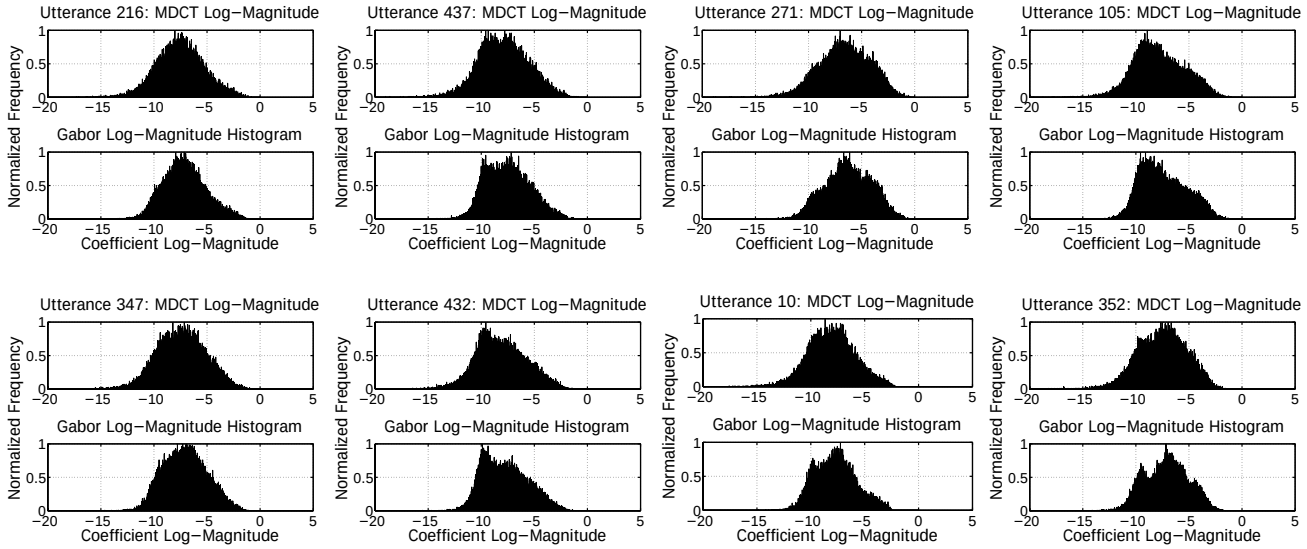

Figure 7: Eight pairs of MDCT (top) and Gabor transform (bottom) log-magnitude coefficient histograms for speech signals selected at random from the database of [8] and employed in the experiments of Section 6. A voice activity detector was employed to eliminate large regions of silence; however, the recurring mode centered near -10 is presumed to be due to time-frequency regions absent of speech energy.

5. MODEL FITTING

The prior models discussed above also lend themselves to fitting procedures. In particular, by equating the (Bayesian) negative log-posterior with an objective function for minimization, we arrive at a family of maximum a posteriori (MAP) shrinkage estimators that admit a correspondence as solutions to variational problems. We consider here the case of the lognormal model described earlier for MDCT coefficient magnitudes, in conjuction with a uniform prior distribution on MDCT "phase" (i.e., sign values). Together these define a lognormal model for MDCT coefficients that is a mixture which has been "symmetrized" about zero.

5.1 Symmetrized Lognormal MDCT Regression

Our interest is in fitting the lognormal model to speech log-magnitude coefficient data in the presence of additive noise. To this end, consider first the standard additive observation model

$$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{n}, \quad \boldsymbol{n} \sim \text{Normal}(\boldsymbol{0}, \sigma_n^2 \boldsymbol{I}),$$

where $\boldsymbol{y} = [y_0 \quad y_1 \quad \cdots \quad y_{L-1}]^{\mathrm{T}}$ is the vector of the observed waveform, $\boldsymbol{x}$ is that of the underlying signal we wish to estimate, and $\boldsymbol{n}$ comprises i.i.d. samples of a continuous Gaussian noise process with variance σ_n^2.

Recall that our model for MDCT coefficients is also i.i.d., allowing us to revert to scalar notation in the sequel. Moreover, as the MDCT is unitary, we will simply use the same notation $y = x + n$ to represent our model for each observed MDCT coefficient. Now note that a lognormal distribution in $|x|$ is specified by parameters μ and σ^2, corresponding respectively to the mean and the variance of the density of $\ln|x|$, which is Normal. When a symmetrized lognormal random variable $|x|$ is observed in additive white Gaussian noise of variance σ_n^2 and the MAP estimator $\widehat{x}_{\text{MAP}}$ is sought, the fitting procedure can be verified to lead to a variational problem of the following form:

$$\widehat{x}_{\text{MAP}} := \underset{x}{\text{argmin}} \left\{ (y-x)^2 + \frac{\sigma_n^2}{\sigma^2} \Big(\ln|x| - (\mu - \sigma^2) \Big)^2 \right\}.$$

Note that under the assumed model (and as reflected by the form of the objective function), the sign of $\widehat{x}_{\text{MAP}}$ will always be equal to that of the observation y. The (implicit) solution to this minimization problem is shown in Figure 2 (see second page of this article), formulated as a shrinkage rule, or "gain" to be applied to each observed MDCT coefficient y as a function of *overall* input signal-to-noise ratio (SNR)—which in turn depends on σ_n^2 and the energy of the speech waveform. (Hence, the rule itself must be computed for each observed waveform of interest, owing to the appearance of μ and σ^2.) Owing to the effect of the prior mean $\mu - \sigma^2$ in the log domain, the minimization procedure can lead to gains greater than unity for observations near zero—since the symmetrized lognormal distribution cannot have mass at negative infinity—though this effect is suppressed in Figure 2 for clarity of presentation.

5.2 Parameter Estimation

For the purposes of modeling and investigation, parameters (μ, σ^2) of the symmetrized lognormal model can easily be fit via maximum likelihood estimation using the "clean" waveform data (the procedure corresponding precisely to maximum likelihood for a Normal, but using the logarithm of the data absolute values). However, any realistic inference procedure must enable the estimation of these parameters from noisy data; i.e., the vector of observations $\boldsymbol{y}$.

To this end, note that the kth moment $\mathbb{E}x^k$ of a lognormal random variable x with parameters (μ, σ^2) is given by the expression $e^{k(\mu + k\sigma^2/2)}$. Since the moments of x (under an "ordinary" lognormal distribution) and $|x|$ (under our symmetrized lognormal distribution) are equal for k even, we can employ the method of moments to estimate our distribution parameters (μ, σ^2) as a function of the (noisy) observed data vector $\boldsymbol{y}$. In particular, it follows that for the case of zero-mean additive white Gaussian noise having variance σ_n^2, the following moment equations may be shown to hold:

$$\mu = \alpha - \frac{1}{4}\beta; \quad \sigma^2 = \frac{1}{4}\beta - \frac{1}{2}\alpha,$$

where parameters

$$\alpha := \ln\Big(\mathbb{E}y^2 - \sigma_n^2\Big), \;\; \beta := \ln\Big(\mathbb{E}y^4 - 6\sigma_n^2\mathbb{E}y^2 + 3\sigma_n^4\Big)$$

can be straightforwardly estimated from noisy data simply by equating the appropriate kth sample moments $L^{-1}\sum_{i=1}^{L} y_i^k$ with the respective kth distribution moments $\mathbb{E}y^k$. (This relation holds as the observations y_i remain i.i.d. under the assumed model.)

(a) Parameters estimated from clean data

(b) Parameters estimated from noisy data

Figure 8: Shrinkage comparisons averaged across the eight utterances whose transform coefficients are shown in Figure 7, with model parameters (μ, σ^2) fitted from clean (a) and noisy data (b) according to the methods of Section 5.2. Results are shown in terms of SNR gain as a function of input SNR; error bars represent approximate 95% confidence intervals based on the eight speech samples considered here.

6. EXPERIMENTAL RESULTS AND CONCLUSION

To test the fitting methods and lognormal shrinkage procedure introduced in Section 5, an investigation was carried out using eight utterances selected at random from the database of [8], which describes a representative subset of 516 samples from the TIMIT corpus. These utterances were analyzed via the MDCT and the Gabor transform using a 256-sample window ($\approxeq$ 16 ms given the TIMIT sampling rate of 16 kHz), derived as described in Section 3.1.

The coefficient log-magnitude histograms of the eight selected examples are shown in Figure 7 on the preceding page, with utterance numbers corresponding to the directory structure of the database of [8]. These histograms include only those short-time segments whose energy exceeded the lower 15th percentile with respect to each utterance overall, though subsequent noise reduction experiments treat the entire recorded utterance in each case. As in the example of Figure 5 we observe that these empirical densities are somewhat super-Gaussian, yet relatively symmetric and unimodal—except for a recurring mode near -10, which we attribute to time-frequency regions absent of speech energy, based on an informal exploration of other utterances in the TIMIT corpus.

While it may be verified that these MDCT coefficient magnitudes decay slightly faster than their Gabor counterparts, the histograms of Figure 7 suggest that a symmetrized lognormal model for the Gabor coefficients may also be of interest. Noting that our fitting procedures can be applied equally well to these data (despite the introduction of coefficient correlations, which violates our original modeling assumptions), we may thus pose the question of which representation—MDCT basis or redundancy-two Gabor frame—yields the best-performing model in terms of mean-squared error. (Though other metrics may be more *perceptually* relevant, our immediate priority is to establish a quantifiable performance baseline.)

As mean-squared error reduction can be re-cast as improvement in SNR, our experimental procedure here consists of first degrading the eight selected utterances by additive white Gaussian noise to yield SNRs in the set $\{-5, 0, 5, 10, 15, 20\}$, and then restoring them according to the lognormal fitting procedures outlined in Section 5.2. For each utterance and SNR considered, this procedure was repeated 25 times, with SNR gains then averaged accordingly. To this end, Figure 8(a) shows the results of a "best-case" lognormal estimation based on the MDCT and the Gabor transform, with parameters fitted from the clean data. It can be seen that the MDCT outperforms the Gabor representation by about 1 dB on average, though the approximate 95% confidence intervals tend to overlap.

Figure 8(b) shows the performance of both methods when the moment estimator of Section 5.2 is applied to the noisy data. While the MDCT representation retains its performance well in low SNR regimes yet degrades somewhat at high SNR, the Gabor representation's performance follows the opposite pattern. As our data in the Gabor case do not correspond to the assumptions of our model, and as we do not yet have quantitative notions of goodness of fit, it is difficult to postulate an immediate explanation for these phenomena. They may well represent an interaction with, or manifestation of, the moment estimator's properties, as the measured shrinkage performance under both representations converges at high SNR; however, a more detailed study is needed to draw firm conclusions.

In perceptual terms, informal listening tests indicate that a stronger suppression is obtained relative to baseline spectral subtraction approaches to shrinkage, leading to a reduction in perceived residual noise but at the loss of some low-level signal detail. These findings are consistent with our initial attempts to approximate the exact MAP estimator analytically, which have yielded simple spectral subtraction schemes in which the noise variance term is artificially inflated—a (typically heuristically motivated) technique known in the speech enhancement literature as "oversubtraction."

In summary, we have presented in this article an overview of sparse time-frequency representations by way of a new symmetrized lognormal model and fitting procedure. Future work will involve a more thorough performance analysis and investigation of approximate estimator solutions, as well as comparisons to other shrinkage estimators in the speech and audio literatures. As a subsequent modeling step, we plan to extend our estimator derivation to model the uncertainty of speech presence, through the inclusion of mixture priors for the coefficients that admit mass at zero.

REFERENCES

[1] P. J. Wolfe and S. J. Godsill, "Interpolation of missing data values for audio signal restoration using a Gabor regression model," in *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, 2005, vol. 5, pp. 517–520.

[2] P. J. Wolfe, S. J. Godsill, and W.-J. Ng, "Bayesian variable selection and regularisation for time-frequency surface estimation (with discussion)," *J. R. Stat. Soc. B.*, vol. 66, no. 3, pp. 575–589, 2004.

[3] M. Dörfler, "Time-frequency analysis for music signals: A mathematical approach," *J. New Mus. Res.*, vol. 30, no. 1, pp. 3–12, 2001.

[4] T. Strohmer, "Numerical algorithms for discrete Gabor expansions," in *Gabor Analysis and Algorithms: Theory and Applications*, H. G. Feichtinger and T. Strohmer, Eds., chapter 8, pp. 267–294. Birkhäuser, Boston, 1998.

[5] H. S. Malvar, *Signal Processing with Lapped Transforms*, Artech House, Norwood, MA, 1992.

[6] J. P. Princen and A. B. Bradley, "Analysis/synthesis filter bank design based on time domain aliasing cancellation," *IEEE Trans. Acoust. Speech Signal Process.*, vol. ASSP-34, no. 5, pp. 1153–1161, 1986.

[7] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proc. IEEE*, vol. 80, no. 10, pp. 1526–1555, 1992.

[8] L. Deng, X. Cui, R. Pruvenok, J. Huang, S. Momen, Y. Chen, and A. Alwan, "A database of vocal tract resonance trajectories for research in speech processing," in *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, 2006, pp. 369–372.