coustics'08
Paris
June 29-July 4, 2008
www.acoustics08-paris.org

On Perceptual Distortion Measures and Parametric Modeling

Mads Christensen

Aalborg University, Niels Jernes Vej 12 A, DK-9220 Aalborg, Denmark
mgc@es.aau.dk

Over the past two decades, there has been much interest in incorporating human sound perception in signal processing algorithms. An example of this is MDCT-based audio coding where a good quality can be achieved at very low bit-rates by taking masking effects into account. More recently, the same principles have been applied to parametric modeling and coding of audio signals. We discuss the inherent tradeoffs in choosing a perceptual distortion measure and a parametric model, and the pros and cons of various ways of implementing such perceptual distortion measures are discussed. An important question that we seek to answer is whether perception should be taken into account in the estimation of model parameters or this should be done in a separate step.

1 Introduction

Ever since the “13 dB Miracle” at AT&T Bell Labs in 1990 by K. Brandenburg and J. J. Johnston laid the ground for what we now know as MPEG-1 Layer 3 (mp3) and MPEG-2/4 AAC, much effort has been devoted to developing models of the human auditory system (also referred to as perceptual models) for signal processing purposes. The use of a perceptually motivated shaping of quantization errors in signal compression of course dates further back than that, with an example being speech coding [1], but that was based on a much simpler model by comparison. Since then, more sophisticated models have been constructed by those knowledgeable in the areas of psychoacoustics and signal processing. For example, a model of the human auditory system was presented in [2] that aimed at predicting the outcomes of masking experiments. The model is based on signal processing and optimal detection theory and contains nonlinear processing stages. This model has since then been simplified for use in coding in [3] and linearized under high-rate assumptions in [4]. The rate-distortion theory for such distortion measures has also recently been developed [5, 6]. Some examples of the use of perceptual models in signal processing, aside from the obvious examples of mp3 and AAC, are analysis-by-synthesis speech coding [1], audio coding based on pre- and post-filtering [7], sinusoidal modeling using pre-filtering [8], weighted matching pursuit [9, 10], and the psycho-acoustic adaptive matching pursuit [11], pre-filtered auto-regressive noise modeling [12], post-filtering in speech coding [13], perceptual shaping in subspace-based speech enhancement [14], and pre-processing in pitch estimation [15]. Note that the matching pursuit methods mentioned above have been shown to be equivalent under certain conditions in [16]. The application to coding is obvious: a better quality can be obtained at the same bit-rate if the distortion measure used in the encoding process is replaced by one that better resembles the human auditory system, in fact the difference can be quite dramatic, hence the miracle. Some other applications are less obvious. What purpose does it, for example, serve in modeling of speech and audio signals where parameters are extracted for further analysis? It may be used for reducing the complexity by eliminating inaudible or less audible signal components so these are not processed in parametric processing applications, like time- and pitch-scale modification, enhancement, etc. Or, if only a limited number of CPU cycles are available, these are used such that the perceptually best quality is achieved by processing the perceptually most important parts of the signal. There is of course also the anthropomorphic argument that since the human auditory system in combination with the human brain seem to be able to solve some particular problem, maybe by mimicking the process, we can make a computer do it too.

In this paper, we will investigate the use of perceptual models in signal processing further, mainly from two points of view, namely estimation theory and signal compression based on parametric models. The applications of pre- and post-processing techniques, a common way of incorporating perception, to estimation problems is discussed, and we seek to quantify the benefits and costs of incorporating perception in estimators based on examples from the literature. In particular, we discuss the value and meaning that parameters obtained in various ways can be attributed.

The paper is organized as follows. First, in Section 2 we will pose the coding problem and the estimation problem and their associated optimal solutions, similarities and dissimilarities. In Section 3, we illustrate some of the points made in the previous section by discussing some estimators incorporating perception that have appeared in the literature recently. We then proceed to discuss the ramifications of our observations before concluding on the paper in Section 4.

2 Estimation and Coding

In this section, we will relate the coding and estimation problems and investigate in what respects there are differences and similarities between the two problems. This investigation is motived by the fact that statistically motived estimators have been applied to so-called parametric coding problems in the literature, e.g., in [8]. Let $\mathbf{A}(\boldsymbol{\theta}) \in \mathbb{C}^{N\times L}$ be a matrix whose columns are nonlinear functions of the parameter vector[1] $\boldsymbol{\theta} \in \mathbb{R}^L$ and the vector $\mathbf{b} \in \mathbb{C}^L$ be a vector of coefficients (or amplitudes) that scale the columns of $\mathbf{A}(\boldsymbol{\theta})$. We then define the parametric coding problem as follows. Given an input vector $\mathbf{x} \in \mathbb{C}^N$ we seek to find an reconstruction vector $\mathbf{A}(\boldsymbol{\theta})\mathbf{b}$ such that the perceptual error, here parameterized by a weighting or sensitivity matrix $\mathbf{W}$, is minimized, i.e.,

$$(\hat{\boldsymbol{\theta}}, \hat{\mathbf{b}}) = \arg\max_{\boldsymbol{\theta},\mathbf{b}} \|\mathbf{W}(\mathbf{x} - \mathbf{A}(\boldsymbol{\theta})\mathbf{b})\|_2^2 . \qquad (1)$$

An example of such a coding problem is sinusoidal coding, where the columns of $\mathbf{A}(\boldsymbol{\theta})$ are complex sinusoids having frequencies that are elements of the vector $\boldsymbol{\theta}$ and complex amplitudes in $\mathbf{b}$. From an estimation theoretical point of view the vector $\mathbf{x}$ is said to consist of an unknown but deterministic component $\mathbf{A}(\boldsymbol{\theta})\mathbf{b}$ and

[1] Each of the L columns in $\mathbf{A}(\boldsymbol{\theta})$ may of course be functions of a number of parameters, like damped sinusoids depending on frequencies and damping factors, but for simplicity we will here assume that they are functions of only one parameter each.

some stochastic component $\mathbf{e}$ (referred to as observation noise), which for the Gaussian case can be fully parameterized by its possibly non-diagonal covariance matrix $\mathbf{R}$. In that case, the maximum likelihood estimates, i.e., the estimates that are most likely to explain the observed signal $\mathbf{x}$ are

$$(\hat{\boldsymbol{\theta}}, \hat{\mathbf{b}}) = \arg \max_{\boldsymbol{\theta}, \mathbf{b}} \left\| \mathbf{C} \left(\mathbf{x} - \mathbf{A}(\boldsymbol{\theta}) \mathbf{b} \right) \right\|_2^2, \qquad (2)$$

where $\mathbf{C}^H \mathbf{C} = \mathbf{R}^{-1}$ is the Cholesky factorization of $\mathbf{R}^{-1}$. The goal of an estimator is to provide us with some useful information about the observed signal, in this case about the parameter vectors $\boldsymbol{\theta}$ and $\mathbf{b}$. This should be seen in contrast to (1) where these parameters do not necessarily have any particular meaning to them; our only concern is to obtain a reconstruction vector $\mathbf{A}(\boldsymbol{\theta})$ that is perceptually close to the input signal $\mathbf{x}$. We note that we will mostly concern ourselves with the problem of estimating the nonlinear parameters $\boldsymbol{\theta}$ and not the linear parameters in $\mathbf{b}$ since the latter problem is easy compared to the former.

From equations (1) and (2) we see that for Gaussian observation noise, the estimation and coding problems are identical in form, the difference being the weighting by either the perceptual weighting matrix $\mathbf{W}$ or the Cholesky factor $\mathbf{C}$. It is clear that the problems are not identical in general and the resulting estimates may be quite different. For other norms or other kinds of noise, the related coding and estimation problems may not be related at all. For example, if the observation noise $\mathbf{e}$ is instead Laplacian, the 2-norm in (2) should be replaced by the 1-norm, if the maximum likelihood estimates are desired, but the relevant coding problem may still be that of (1). We can, however, also identify a special case where they are the same. For a regular 2-norm and white Gaussian noise, i.e., $\mathbf{W} = \mathbf{I}$ and $\mathbf{C} = \mathbf{I}$, the problems and their solutions are the same. There are, however, also other cases where the solutions will be the same. For example, if the matrix $\mathbf{A}(\boldsymbol{\theta})$ is composed of complex sinusoids characterized by their frequencies, it has been shown that the nonlinear least-squares estimates of $\boldsymbol{\theta}$ obtained from (2) with $\mathbf{C} = \mathbf{I}$ is still asymptotically efficient, i.e. optimal, even if in reality the noise is not white [17]. This means that the most common case of parametric coding is in fact an example where there is a relation between finding the most likely parameters and minimizing the squared error. But what of the perceptual weighting matrix $\mathbf{W}$? It turns out, that for this asymptotic equivalence to hold, $\mathbf{W}$ has to have a certain structure. Such a structure is induced by the perceptual distortion measure in [3] which leads to $\mathbf{W}$ being a filtering matrix [8, 16]. It is quite clear that this is a rather special case. Indeed, it comprises a certain model, a certain distortion measure, and a certain kind of stochastic signal component! What an odd coincidence indeed.

So far, we have not alluded to the importance of the number of nonlinear and linear parameters L. In the coding problem in (1), these quantities are "user parameters" that along with the statistical properties of $\boldsymbol{\theta}$ and $\mathbf{b}$ determine the resulting bit-rate. However, in the estimation problem in (2) their meaning and importance are quite different. The estimates obtained using (2) are only the maximum likelihood estimates if L' is equal to the true number of parameters L or the problem is separable (which is asymptotically the case for sinusoids). In other words, estimates obtained using (2) for $L' \leq L$ are not generally a subset of the parameters obtained for the true order. Similarly, the maximum likelihood estimators obtained for the true order are not the parameters that minimize the perceptually weighted 2-norm for $L' < L$ even if the problems are equivalent for $L' = L$. Again, the conclusion is that the coding and estimation problems and their solutions are quite different and not too much meaning should be attributed to parameters obtained using (1) since the statistical properties of such an estimate may be quite poor.

3 Perception in Estimators

We will now discuss some specific estimators incorporating perception that have appeared in the literature in recent years. All the specific estimators are based on sinusoidal models, i.e., the matrix $\mathbf{A}(\boldsymbol{\theta})$ is composed of complex sinusoids characterized by frequencies in $\boldsymbol{\theta}$ and complex amplitudes $\mathbf{b}$, save on that is based on damped sinusoids including also a damping factor. It is important to stress, though, that the general problems and considerations extend beyond this model.

The coding and estimation problems in (1) and (2), respectively, are multi-dimensional nonlinear optimization problems that are difficult to solve in practice. Therefore, suboptimal schemes are often used instead. A common way of doing this is by splitting the problem into a number of one-dimensional optimization problems that are then solved iteratively. This is also the case for the estimation problem where an example of such an approach is the Expectation Maximization (EM) algorithm as applied to finding the parameters of superpositions of signals in [18]. Quite a few methods that aim at finding approximate solutions to (1) based on this methodology have been proposed. Some, like [11, 19], use an exact cost function while others use only an approximation, e.g., [9]. A common trait of these methods is that they explicitly aim at solving (1). There are other methods, where the objective is less clear with some examples being the methods proposed in [8] and [10]. The methods [9, 11, 19] and others find sinusoids one at the time by extracting the sinusoid that greedily minimizes a perceptual distortion measure. These methods thus result in a perceptual ranking of the found components. Therefore, if a low-order model is desired for low-complexity processing or transmission using a low bit-rate, such an approach is very convenient. Also, computation time is not wasted on extracting sinusoids that are not used anyway. As explained in [16], some of these methods are also asymptotically optimal (for a large number of samples N) from a statistical point of view.

The method of [8] is based on the idea that a perceptual ranking of damped sinusoidal components can be achieved by applying pre-filtering, i.e., a kind of pre-processing, to the input signal. Many examples of pre-processing of a signal before some other kind of processing is performed exist in the literature, with an example being the use of a perceptual model before some estimation task. These can generally be described as mapping the signal to another domain that simplifies the processing. This mapping is sometimes referred to as a compressor while the inverse is referred to as an expander and the combination as a compander. For the sake of this discussion, we here assume that this map has been linearized such that the pre-processing can be written as

$$\mathbf{y} = \mathbf{U}\mathbf{x}. \tag{3}$$

The vector $\mathbf{y}$ is then processed instead of $\mathbf{x}$ where after the modified signal $\hat{\mathbf{y}}$ is mapped back by post-processing as

$$\hat{\mathbf{x}} = \mathbf{U}^{-1}\hat{\mathbf{y}}, \tag{4}$$

where we have assumed that the inverse exists. Sometimes the reconstruction vector $\hat{\mathbf{y}}$ is not of interest, but rather some parameters are extracted from $\mathbf{y}$ in which case only the pre-processing is applied. One may wonder what purpose pre-processing serves. Suppose we have an estimation problem and an optimal estimator, say a maximum likelihood estimator and a large number of samples. Then no pre-processing can improve on the estimator since it is already optimal, in fact the pre-processing at best makes it no worse. On the other hand, suppose the estimator is suboptimal, then it is possible that we can modify the signal somehow such that the estimator performs better. An example of this is pre-whitening in subspace methods. Subspace methods are based on stochastic signal parts being white but not having any particular distribution. If the stochastic signal parts are colored, the estimator may fail completely, and pre-whitening may solve that problem. Thus, such pre-processing can been seen to simplify the estimator in some cases. Similar arguments can be made for coding where compressors are sometimes used to simplify the quantization process in signal compression. After quantization, the inverse function is applied to retrieve the reconstruction vector. The high-rate theory for this kind of coding has been developed in [5,6] and an example of a coder using this principle is [7]. Returning to our discussion of the approach in [8], the question is what exactly the purpose of the pre-filtering is? As mentioned, a pre-requisite for the subspace methods is that the noise is white, but the perceptual pre-filter used in [8] serves not to whiten the signal, but to shape it according to its perceptual importance, and the filter characteristics can be quite extreme. There are in fact several other problems with this approach. The most glaring is that for the problem statement to make sense, only a subset of the sinusoids are of interest. However, an orthonormal basis for a subset of the sinusoids cannot generally be obtained from a subset of covariance matrix eigenvectors, only for some special cases and not for the damped sinusoids considered in [8]. Therefore, the model order is not a “user parameter” that can be chosen arbitrarily. There is also the problem that the estimates obtained using ESPRIT are not ranked by amplitude, and the minor problem that one cannot expect to get correct estimates of the damping factors when such pre-filtering is applied. Similar problems would apply if one was to use pre-filtering in the MUSIC algorithm. The subspace ranks remain unchanged by the filtering process and the algorithm measures orthogonality, not amplitude, and it is therefore not possible to retrieve the perceptually most important sinusoids this way.

It should be clear from this discussion that the pre- and post-processing should be chosen in accordance with the subsequent processing; pre-filtering in subspace methods should whiten the signal, not color it. Since the problem of optimal estimation has little to do with the human auditory system, except for the anthropomorphic argument, one can reasonably wonder why one should ever choose pre-processing based on the properties of the human auditory system when the goal is to find accurate estimates. It is of course possible that the human auditory system has evolved such that it is optimized for certain kinds of signals, in which case the processing in the human auditory system may be relevant from an estimation theoretical point of view. A simple counter example to this point is the use of masking curves in audio coding. Audio coders seek to shape the coding error according to the masking threshold, not such that it is white, and from this point of view, applying a perceptual model possibly even makes the estimation problem harder and degrades the performance of subspace methods. For nonlinear models, it becomes even more complicated to determine the properties of the signal and such models should of course then be applied to estimation problems with care.

The application of pre-processing to estimation problems is also problematic in another way. Suppose we have an estimation problem where we seek to estimate the parameters of the signal $\mathbf{A}(\boldsymbol{\theta})\mathbf{b}$ from the pre-processed signal $\mathbf{y}$. The signal model may be valid for the input signal $\mathbf{x}$, but how do we know that the signal model also applies to $\mathbf{y}$ so that we may ignore the transformation $\mathbf{U}$ in the process? This imposes a certain relationship between the columns of $\mathbf{A}(\boldsymbol{\theta})$ and $\mathbf{U}$. For example, if we seek to estimate the frequencies of a set of sinusoids in $\mathbf{x}$ from $\mathbf{y}$, then the sinusoids have to be invariant to the transformation $\mathbf{U}$. Or, in other words, the vectors of $\mathbf{A}(\boldsymbol{\theta})$ have to be eigenvectors, or good approximations thereof, of $\mathbf{U}$ [16]. For the distortion measure in [3], this is asymptotically the case for sinusoids, because it induces a certain structure on $\mathbf{U}$. This again underlines that arbitrary transformations may not result in meaningful estimation problems for the transformed signal, and we have not even touched upon the implications of nonlinear pre-processing.

4 Conclusion

We have investigated the application of perceptual models to estimation, coding, and modeling problems and we have discussed the similarities between these problems for the case of parametric coding. We have argued that the use of perceptual models is well-founded for coding applications but less so for estimation and modeling where the goal is to extract meaningful parameters whose statistical properties are important. In fact, we have argued that perception may result in worse estimates and we have given examples of this from the literature. There are however certain classes of models, problems and estimators where their use makes sense.

For the case of sinusoidal coding where the input signal is considered to consist of a sum of sinusoids in colored Gaussian noise, there exists asymptotic equivalences between maximum likelihood estimation and finding the perceptually most important sinusoids as defined by a certain structured distortion measure. This means that weighted least-squares methods and approximations thereof, like the matching pursuit, result in a signal model that minimizes the perceptual distortion and accurate parameter estimates at the same time. The conclusion is that perceptual models should be applied to estimation problems with great care and the properties of the signal model, estimators and the perceptual model have to be considered jointly to result in meaningful estimates.

5 Acknowledgments

The author is supported by the Parametric Audio Processing project, Danish Research Council for Technology and Production Sciences grant no. 274–06–0521.

References

[1] B. S. Atal and J. R. Remde, "A new model of LPC excitation for producing natural sounding speech at low bit rates," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1982.

[2] T. Dau, D. Püschel, and A. Kohlrausch, "A quantitative model of the effective signal processing in the auditory system. i. model structure," *J. Acoust. Soc. Am.*, vol. 99(6), pp. 3615–3622, June 1996.

[3] S. van de Par, A. Kohlrausch, R. Heusdens, J. Jensen, and S. H. Jensen, "A perceptual model for sinusoidal audio coding based on spectral integration," *EURASIP J. on Applied Signal Processing*, vol. 9, pp. 1292–1304, 2004.

[4] J. Plasberg, D. Zhao, and W. B. Kleijn, "Sensitivity matrix for a spectro-temporal auditory model," in *Proc. European Signal Processing Conf.*, 2004, pp. 1673–1676.

[5] T. Linder, R. Zamir, and K. Zeger, "High-resolution source coding for non-difference distortion measures: Multidimensional companding," *IEEE Trans. Information Theory*, vol. 45(2), pp. 548–561, 1999.

[6] T. Linder and R. Zamir, "High-resolution source coding for non-difference distortion measures: The rate-distortion function," *IEEE Trans. Information Theory*, vol. 45(2), pp. 533–547, 1999.

[7] G. D. T. Schuller, B. Yu, D. Huang, and B. Edler, "Perceptual audio coding using adaptive pre- and post-filters and lossless compression," in *IEEE Trans. Speech and Audio Processing*, Sept. 2002, vol. 10(6), pp. 379–390.

[8] J. Jensen, R. Heusdens, and S. H. Jensen, "A perceptual subspace approach for modeling of speech and audio signals with damped sinusoids," *IEEE Trans. Speech and Audio Processing*, vol. 12(2), pp. 121–132, Mar. 2004.

[9] T. S. Verma and T. H. Y. Meng, "Sinusoidal modeling using frame-based perceptually weighted matching pursuits," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1999, vol. 2, pp. 981–984.

[10] K. Vos, R. Vafin, R. Heusdens, and W. B. Kleijn, "High-quality consistent analysis-synthesis in sinusoidal coding," in *Proc. Aud. Eng. Soc. 17th Conf.*, 1999, pp. 244–250.

[11] R. Heusdens, R. Vafin, and W. B. Kleijn, "Sinusoidal modeling using psychoacoustic-adaptive matching pursuits," *IEEE Signal Processing Lett.*, vol. 9(8), pp. 262–265, Aug. 2002.

[12] R. Hendriks, R. Heusdens, and J. Jensen, "Perceptual linear predictive noise modelling for sinusoid-plus-noise audio coding," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2004, vol. 4, pp. 189–192.

[13] V. Grancharov, J. H. J. H. Plasberg, J. J. Samuelsson, and W. B. Kleijn, "Generalized postfilter for speech quality enhancement," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 16(1), pp. 57–64, 2008.

[14] F. Jabloun and B. Champagne, "Incorporating the human hearing properties in the signal subspace approach for speech enhancement," *IEEE Trans. Speech and Audio Processing*, vol. 11(6), pp. 700–708, Nov. 2003.

[15] S.C. Sekhar, S. Pilli, L. C, and T.V. Sreenivas, "Novel auditory motivated subband temporal envelope based fundamental frequency estimation algorithm," in *Proc. European Signal Processing Conf.*, 2006.

[16] M. G. Christensen and S. H. Jensen, "On perceptual distortion minimization and nonlinear least-squares frequency estimation," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14(1), pp. 99–109, Jan. 2006.

[17] P. Stoica, A. Jakobsson, and J. Li, "Cisiod parameter estimation in the coloured noise case: Asymptotic Cramér-Rao bound, maximum likelihood, and nonlinear least-squares," in *IEEE Trans. Signal Processing*, Aug. 1997, vol. 45(8), pp. 2048–2059.

[18] M. Feder and E. Weinstein, "Parameter estimation of superimposed signals using the EM algorithm," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36(4), pp. 477–489, Apr. 1988.

[19] M. G. Christensen and S. H. Jensen, "The cyclic matching pursuit and its application to audio modeling and coding," in *Rec. Asilomar Conf. Signals, Systems, and Computers*, 2007, pp. 550–554.