Signal Modeling and Psychoacoustic Data Collection

Ray Migneco
rmigneco@drexel.edu

January 13th, 2009
## Contents

1 Introduction 1

2 Perceptual Audio Modeling with Damped Sinusoids 2
   2.1 Main Objectives ................................................. 2
   2.2 ESM Parameter Extraction via HTLS ............................ 2
       2.2.1 HTLS Algorithm ........................................... 2
   2.3 Perceptual Distortion Measure and Prefiltered HTLS ............. 3
       2.3.1 Perceptual Distortion Measure ............................. 3
       2.3.2 Prefiltered HTLS ........................................ 3
       2.3.3 Estimating complex amplitudes ........................... 4
   2.4 Experimental Results .......................................... 4
       2.4.1 Modeling a Synthetic Signal ............................... 4
       2.4.2 Performance vs. Model Order ............................... 5
       2.4.3 Subjective Evaluation .................................... 5
   2.5 Possible Extensions ............................................ 5

3 Speech Modeling via the MVDR Spectrum 6
   3.1 Main Objectives ................................................. 6
   3.2 Limitations of All-Pole Modeling via Linear Prediction ...... 6
       3.2.1 LP Development ............................................ 6
       3.2.2 Spectral Modeling via LP .................................. 6
   3.3 MVDR Spectrum for Speech Modeling ............................ 6
       3.3.1 MVDR Development ........................................ 6
       3.3.2 MVDR All-Pole Model ..................................... 7
   3.4 Experimental Results .......................................... 8
       3.4.1 MVDR Modeling of Synthetic Spectra ....................... 8
       3.4.2 MVDR Modeling of Voiced Speech ........................ 8
   3.5 Possible Extensions ............................................ 9

4 Activities for Psychoacoustic Data Collection and Education 9
   4.1 Objectives ..................................................... 9
   4.2 Data Collection via Online Activities ........................ 9
   4.3 Tone Bender Activity Description ............................... 10
   4.4 Cocktail Party Game Description ............................... 10
   4.5 Pilot Study .................................................... 11

5 Broad Connections 12
   5.1 Objectives ..................................................... 12
   5.2 System Overview .............................................. 12
   5.3 Technical Implementation ..................................... 12
       5.3.1 Signal Modeling ......................................... 12
       5.3.2 Incorporating MVDR ...................................... 13
       5.3.3 Incorporating ESM ....................................... 13
   5.4 Data Collection .............................................. 14
       5.4.1 Collaborative Activities ................................. 14
       5.4.2 Evaluation Metrics ...................................... 14
       5.4.3 Data Collection and Educational Outreach ................ 15
   5.5 Conclusions ................................................... 15

A Spectral Modeling with MVDR Spectra 16
   A.1 Objectives .................................................... 16
       A.1.1 Spectral Threshold via MVDR ............................. 16
       A.1.2 Frequency Extraction with MVDR and LP Spectra .......... 17
1 Introduction

Research in psychoacoustics and signal processing has led to the development of signal modeling schemes that exploit properties of the human auditory system in order to represent perceptually salient features of audio signals. These combined research areas have facilitated the advancement of several important applications including perceptual codecs, which provide bandwidth compression for the storage and transmission of audio signals (i.e. mp3, MPEG Advanced Audio Coding, Windows Media Audio). This study presents an overview of selected methods in audio signal modeling and experiments in psychoacoustic data collection to propose a research platform that allows for continuous collection and evaluation of psychoacoustic data in order to develop improved signal modeling techniques.

The presented papers in signal modeling provide theory, methods and experimental results in order to support analysis schemes that offer perceptual benefits over existing methods. Specifically, the use of perceptual measures in these papers aims to incorporate preexisting knowledge and assumptions about the human auditory system so that signals are modeled with respect to the way humans hear. With regard to the experimental aspect of the research platform, the paper on data collection presents methodology for developing web-based activities concerning psychoacoustic problems and results from a pilot study utilizing these activities. This work also demonstrates that web-based activities are suitable for hosting experiments and collecting data for psychoacoustic evaluation on a large scale. Perceptually motivated signal modeling schemes and methods of collecting large-scale data from a diverse subject population are required in order to understand how the auditory system responds to complex mixtures. Throughout this overview, the main concepts from the selected literature will be discussed in relation to how they contribute towards the research platform’s goal of developing improved audio signal models.

A signal modeling technique based on psychoacoustic masking and the exponential sinusoidal model (ESM) is presented by Jensen et al. in A Perceptual Subspace Approach for Modeling of Speech and Audio With Damped Sinusoids. In contrast to other modeling schemes, the proposed method extracts parameters for audio signals in accordance with an auditory model provided by a psychoacoustic mask. The effect of the mask is to bias component selection so that resulting signal satisfies the auditory model it represents, thus employing assumed knowledge on human perception. The performance of the proposed technique (P-ESM) is compared against a traditional technique (I^2-ESM) by computing a perceptual SNR between original and modeled signals. Experimental results show that P-ESM yields an improved SNR over I^2-ESM as model order is increased for harmonic and noise-like signals. The subjective quality of the two ESM schemes is evaluated by modeling and synthesizing a variety of real audio signals with each algorithm and evaluating listeners’ preferences for the scheme that produces the best result.

In All-Pole Modeling of Speech Based on the Minimum Variance Distortionless Response Spectrum [1], Murthi et al. present a speech modeling technique that provides improved spectral envelope and power estimation over existing linear prediction (LP) methods. The proposed method relies on computing the MVDR spectrum for speech signals so that a parametric representation of speech spectra is obtained that maintains the power at the harmonic frequencies. This feature makes MVDR desirable since the resulting spectra have smooth contours near the signal harmonics thus resulting in a better perceptual approximation of the signal. This paper presents the limitations of LP speech models and the design methodology for an all-pole model of speech using MVDR. Additionally, simulation results are provided which compare the modeling performance of LP and MVDR for synthetic and non-synthetic speech.

In Collaborative Online Activities for Acoustics Education and Psychoacoustic Data Collection [2], Kim et al. present prototype activities suitable for large scale collection and evaluation of psychoacoustic data. The objective of these activities is to determine if users can identify a target signal in an audio mixture that has been obscured in some way by another user. The results of these activities provide evaluation data which can be used to assess the factors influencing human perception in response to particular problems in psychoacoustics (i.e. speaker identification and musical instrument identification). The paper presents the objectives, design and implementation of each activity and the quantitative results from a pilot study in which the activities were deployed in two middle-school math and science classes.

In order to gain an understanding of how the auditory system perceives complex mixtures with regard to problems in psychoacoustics, signal modeling techniques are required that capture relevant perceptual features. Additionally, data must be collected over many individuals so that individual variations in perception are accounted for. As will be discussed throughout this study, the main concepts from the presented papers contribute towards developing improved audio modeling schemes for the proposed research platform.
2 Perceptual Audio Modeling with Damped Sinusoids

2.1 Main Objectives

Jensen et al. present a signal modeling algorithm for speech and audio based on the exponential sinusoidal model (ESM) and psychoacoustic masking phenomena [3]. In contrast to traditional ESM schemes where parameters are estimated based on a least squares minimization of the model error ($l^2$-ESM), the proposed technique extracts parameters based on minimizing a weighted distortion measure, which incorporates psychoacoustic information. This paper presents the traditional $l^2$-ESM technique and the formulation of the weighted distortion measure which forms the basis of the proposed P-ESM technique. As will be explained, the intended effect of the weighted distortion measure is to bias parameter extraction in order to satisfy the perceptual model imposed by the mask. Additionally, this paper presents simulation results that compare both techniques with regard to modeling synthetic speech and evaluating performance vs. model order. The quality of the two schemes is assessed via a subjective analysis using human test subjects.

2.2 ESM Parameter Extraction via HTLS

Jensen et al. employ P-ESM in order to solve for parameters that satisfy the exponential sinusoidal model, which represents a signal as a sum of exponentially damped sinusoids. This model can be expressed as

$$\tilde{x}_n = \sum_k a_k \exp(-d_k n) \cos(w_k n + \phi_k), n = 0, \ldots, N - 1,$$

where $a_k, d_k, w_k, \phi_k$ are the amplitude, damping, frequency and phase parameters to be estimated. However, in order to fully understand P-ESM, it is necessary to discuss the Hankel Total Least Squares (HTLS) algorithm which is the basis for extracting parameters in the $l^2$-ESM method. Jensen et al. employ the so-called pre-filtered HTLS algorithm to extract P-ESM parameters but they do not explicitly describe HTLS. This section will provide a brief overview of the HTLS algorithm as it pertains to obtaining signal poles that fit the model in (1). The concept of “signal poles” can be realized by expressing (1) as

$$\tilde{x}_n = \sum_{k=1}^{K} c_k z_k^n,$$

where $\exp(-d_k + j\omega_k)$ indicates the signal poles represented by $z_k$ and $c_k$ indicates the complex amplitudes formed by $a_k \exp(j\phi_k)$.

2.2.1 HTLS Algorithm

Van Huffel et al. developed the HTLS algorithm in order to model time domain Nuclear Magnetic Resonance (NMR) data using ESM [4]. This is done by observing that $N$ time domain data points can be represented by a state space model

$$\tilde{x} = CA^n i_0$$

where $i$ is the initial state vector, $A$ is the system matrix containing the poles of $\tilde{x}$ and $C$ is the output matrix. This model can be related to ESM by equating (2) and (3) and thus the signal poles and complex amplitudes can be found by solving for $A$ and $C$, respectively. Assuming that ESM is a physically reasonable representation for an observed signal frame $x$ and that the exact model order $K$ is known, the steps to solve for the signal poles can be summarized as [4]:

**Step 1:** Formulate the Hankel data matrix $X$ from the observed signal frame $x$. $X$ should have dimensions $L \times M$ with $L + M - 1 = N$, where $N$ is the frame length of the data in samples.

**Step 2:** Compute the singular value decomposition (SVD) on $X$ in order to obtain $X = U \sum V^T$ where $U$ and $V^T$ are the left and right singular vectors of $X$, respectively, and $\sum$ contains the singular values of $X$ along the main diagonal.

**Step 3:** Invoke the shift invariance property of $U$ in order to solve $U^\dagger Z^{(u)} = U^\dagger$ for $Z^{(u)}$ using total least squares. $\dagger$ and $\downarrow$ indicate the deletion of the top and bottom rows of $U$, respectively.
**Step 4:** Obtain the signal poles $z_k$ by eigenvalue decomposition of $Z^{(u)}$, which is a $K \times K$ matrix.

In practice, the HTLS algorithm must be modified because the exact model order $K$ that represents $x$ is not known. To account for this, $X$ is truncated to a matrix $X_k$, which has rank $K$ and the SVD from Step 2 is performed on $X_k$. In Step 3, $U_k$, which contains the first $K$ columns of $U$, is used to solve for $Z^{(u)}$ using total least squares and hence the signal poles can be derived by eigenvalue decomposition as in Step 4. Essentially, the truncation ensures that the rank of $X$ is equal to the number of desired signal poles, or the model order $K$, so that HTLS yields a unique solution.

### 2.3 Perceptual Distortion Measure and Prefiltered HTLS

#### 2.3.1 Perceptual Distortion Measure

In order to extract ESM parameters that satisfy an auditory model, Jensen et al. formulate a perceptually weighted distortion measure that incorporates psychoacoustic masking information. The purpose of the mask is to exploit properties of the human auditory system that affect the audible frequencies in a complex mixture. In particular, humans cannot discriminate between closely spaced frequencies if a neighboring frequency component is significantly stronger. In order for a component to be audible its magnitude must exceed a masking threshold imposed by the stronger, masking frequency component. Jensen et al. use this information so that the modeled signal components satisfy the psychoacoustic mask they specify.

To bias signal modeling in accordance with the chosen mask, Jensen et al. define a distortion measure

$$D = \int_0^1 \hat{a}(f) |(\hat{\omega} \epsilon)|^2 df,$$

where $\epsilon$ is the modeling error, $x - \tilde{x}$, $w$ is the analysis window and $a$ is the reciprocal of the chosen auditory mask. The $\hat{a}$ operator indicates the Fourier transform. The integral in (4) can be expressed as a time domain convolution that defines a $l^2$-norm by invoking Parseval’s Theorem and choosing $\hat{a}$ as the inverse Fourier transform of $\sqrt{\hat{a}}$.

$$D = \sum_n |(h \ast w \epsilon)(n)|^2 = ||h \ast w \epsilon||^2_2.$$

Furthermore, the convolution in (5) can be expressed as a matrix multiplication

$$\min_{a_k,d_k,w_k,\varphi_k} ||HW(x - \tilde{x})||^2_2$$

where $W$ is a diagonal matrix matrix that implements $w$ and $H$ is a filter matrix containing the elements of $h$. The problem at hand is finding the ESM parameters that best minimize the norm in (6).

#### 2.3.2 Prefiltered HTLS

The basis for P-ESM is indicated by the weighted minimization in (6), which incorporates the psychoacoustic mask through the filter matrix. The use of this perceptual filter necessitates that parameters be extracted via the pre-filtered HTLS algorithm used by Jensen et al. To understand why HTLS, as described in the previous section, is not immediately suitable for treating (6), the complexities introduced by the filtering matrix must be discussed. It may seem appropriate to filter the observed signal frame with the perceptual filter and formulate the result into a Hankel matrix, which is fed into the HTLS algorithm. However, the convolution operation changes the length of the output sequence which alters the rank when $X$ is formed. To maintain the length of $x$, the middle of the filtered sequence can be retained by truncating the first and last values introduced by filtering. However, this approach wastes information that could be used to find optimal ESM parameters.

In order to account for the complexities introduced by the filter matrix, Jensen et al. utilize a prefiltered HTLS algorithm that implements the convolution in (6) through a full rank convolution matrix. To formulate this matrix, Jensen et al. note that the signals of interest are discrete time sequences so the integral in (4) should be replaced with a summation. This amounts to a point-wise multiplication in the frequency domain which is equivalent to circular convolution. A Toeplitz circular convolution matrix $F$ containing shifted
versions of \( h \) along its columns is used to form \( \bar{X} = XF \). As a result, \( \bar{X} \) retains the original rank of \( X \) and the signal poles can be obtained using the HTLS algorithm.

### 2.3.3 Estimating complex amplitudes

Although careful consideration was required to make HTLS suitable for filtered signals, estimating the complex amplitudes is a straightforward process requiring that (6) be reformulated to include the estimated signal poles. To solve for the complex amplitudes, (6) is rewritten as a minimization over a complex amplitude vector \( c = [c_1, \ldots, c_k] \)

\[
\min_c \ = \ ||HW(x - Vc)||_2^2,
\]

where \( V \) indicates the Vandermonde matrix containing the estimated signal poles \( z_k \) [3]. The Vandermonde matrix implements a polynomial interpolation in order to find \( c \) that best minimizes the norm in (7).

### 2.4 Experimental Results

With the framework established for P-ESM, Jensen et al. conduct several experiments in order to compare the performance between \( l^2 \)-ESM and P-ESM. This section will discuss the simulations which compare the two schemes in terms of modeling synthetic signals, performance versus model order and their ability to faithfully reproduce non-synthetic signals based on subjective evaluation.

It should be noted that metric used to compare the performance of the modeling schemes requires computing a perceptual SNR developed by the authors. The perceptual SNR is defined as

\[
\text{SNR}_p = 10 \log_{10} \left( \frac{||HWx||_2^2}{||HW(x - \hat{x})||_2^2} \right) \text{ [dB]},
\]

where \( x \) is the observed signal frame and \( \hat{x} \) is the synthesized frame resulting from P or \( l^2 \)-ESM parameters. Essentially, \( \text{SNR}_p \) evaluates the ability of the modeling schemes to satisfy the perceptual model imposed by the psychoacoustic mask. The main limitation with this metric is that it requires \( l^2 \)-ESM to perform in accordance with constraints that were used to derive the P-ESM scheme. In this case, one would naturally expect P-ESM to yield better performance since \( l^2 \)-ESM is based on error minimization in a least squares sense. Furthermore, the authors consider \( \text{SNR}_p \) an objective measure by assuming the mask they use is a reasonable model of the auditory system. However, a single auditory model cannot encapsulate all variations amongst the general population thereby making their model reasonable only for the assumptions it incorporates.

### 2.4.1 Modeling a Synthetic Signal

In order to demonstrate the perceptual bias of P-ESM towards the chosen psychoacoustic mask, both schemes are used to model a synthetic signal consisting of two low energy, closely spaced sinusoids and one high energy, high frequency sinusoid. The results are shown in Figure 1, where the top row indicates the original signal for both schemes and the proceeding rows indicate the results from increasing model order for \( K = 2, 4, 6 \) so that suboptimal and optimal performance is demonstrated. The figure indicates that when model order is suboptimal (\( K < 6 \)), P-ESM prefers the model components that lie above the masking curve (dotted line) while \( l^2 \)-ESM prefers the higher energy components. As the last row of Figure 1 shows, both schemes model the same components in the optimal case where \( K = 6 \).

Although the synthetic signal was a trivial case for analysis, the result of this simulation reflects how the formulation of the norm in (6) biases the modeled compo-

![Figure 1](image)
nents for each scheme. In the absence of perceptual filtering, the goal is to minimize modeling error in a purely least squares sense, which causes $l^2$-ESM to model the highest energy frequency components. However, when the mask is incorporated, P-ESM selects the components that lie above the masking threshold.

### 2.4.2 Performance vs. Model Order

The next simulation between $l^2$-ESM and P-ESM involves evaluating the perceptual SNR incurred from each method as a function of the model order $K$. The signals used for this evaluation consisted of a voiced and unvoiced speech segment. For each ESM approach, the model order was varied from $K = 2, \ldots, 80$ and $\text{SNR}_p$ was calculated at each value of $K$.

The performance results on the voiced speech segment indicate interesting results, particularly at low model orders. For $K < 20$, there was little difference in $\text{SNR}_p$ for each method, indicating no clear benefit for P-ESM. Jensen et al. attribute this result to the spectral structure of voiced speech since the lower frequency components often have the highest energy and are the most perceptually relevant. In this case both schemes would model essentially the same components for $K < 20$. While this result seems logical, it indicates a deficiency in P-ESM since additional computation is required to incorporate the auditory mask, but it provides no performance benefit over $l^2$-ESM for lower model orders. This additional computation is essentially wasted since $l^2$-ESM satisfies the perceptual metric just as well in these cases. As $K$ is increased from 20 to 80, the performance gap grows in favor of P-ESM by about 2 dB.

When considering the unvoiced speech segment, P-ESM yielded better $\text{SNR}_p$ for all model orders $K > 2$ and the performance gap peaked to about 3 dB at $K = 80$. Unlike voiced speech, the low frequencies are not necessarily the most perceptually relevant, which causes each method to perform differently as $K$ is varied. In this case P-ESM outperforms $l^2$-ESM in satisfying the constraints imposed by $\text{SNR}_p$.

### 2.4.3 Subjective Evaluation

In order to gauge the quality of $l^2$-ESM and P-ESM techniques, each method was used to model real audio signals which were resynthesized and evaluated by 9 test subjects, which excluded the authors. 7 test signals consisting of simple and complex mixtures were modeled using each ESM technique for $K = 50$ and resynthesized for listening. The subjects were asked to indicate which modeling scheme produced a signal that was closest to the original. Despite the small subject size used, the result indicate that P-ESM was preferred by at least 69% and as much as 94% for all test signals. The authors claim their own subjective listening tests reveal that P-ESM has a considerably higher subjective quality than $l^2$-ESM for $K > 50$ due to P-ESM’s ability to model noise like components better than $l^2$-ESM. This observation was substantiated by noticing $l^2$-ESM’s tendency to cluster frequency components near the center of the spectrum which creates a narrow-band noise effect. On the other hand, P-ESM tends to model noise more uniformly across the spectrum thus minimizing narrow-band effects.

### 2.5 Possible Extensions

Jensen et al present a signal modeling approach that extracts ESM parameters based on a human auditory model provided by a psychoacoustic mask. While the experimental results indicate that P-ESM is largely more successful at satisfying the perceptual measure chosen by the authors, further work is required to demonstrate the usefulness of P-ESM at low model orders where it does not outperform $l^2$-ESM. This deficiency suggests using a dual modeling approach in applications where low model order is required and computational resources are limited. The subjective evaluation could also be improved by having a larger sample size of subjects and test audio signals. In this sense, it would be useful to establish thresholds for acceptable $\text{SNR}_p$ and thus a minimum model order for resynthesis based on the preference of a large subject population. This information would be beneficial in coding schemes where the precise model order for a particular application must be known a priori.
3 Speech Modeling via the MVDR Spectrum

3.1 Main Objectives

Murthi et al. present an all-pole modeling scheme for speech based on the minimum variance distortionless response (MVDR) spectrum in order to improve spectral envelope modeling and power estimation [1]. The need for improved spectral modeling in speech arises from the deficiencies in linear prediction (LP) modeling techniques, which often does not provide smooth spectral estimates for medium and high pitched voice. With regard to maintaining the perceptual quality of the signal, smooth spectral envelopes are an attractive feature for signal modeling schemes since they provide a better estimation of the power at the input signal’s component frequencies. This paper discusses the problems incurred with LP speech modeling and the use of the MVDR spectrum in order to improve on LP’s shortcomings. Experimental results will be discussed that compare the modeling performance of LP and MVDR on various speech types.

3.2 Limitations of All-Pole Modeling via Linear Prediction

3.2.1 LP Development

The LP method of speech modeling involves obtaining a set of parameters that define a prediction error filter, so that the mean squared error (MSE) between a speech sample $s[n]$ and its predicted value $\hat{s}[n]$ is minimized based on a linear combination of $M$ past samples [5]. This prediction error filter can be described by a finite-impulse response (FIR) filter $A(z) = 1 - \sum_{k=1}^{M} a_k z^{-k}$ where $a_k$ indicates the $k^{th}$ prediction coefficient and $M$ is the desired order. The inverse filter is obtained by taking $\frac{1}{A(z)}$, which provides the all-pole model of the vocal tract. When the inverse filter exactly models the vocal tract, the speech sequence $s[n]$ can be reconstructed by filtering it with a periodic impulse train [5].

3.2.2 Spectral Modeling via LP

The problem with the spectral models provided by LP is that they can lead to sharp contours which overestimate the perceptually important frequencies. This effect is inherent in the minimization of the MSE used to formulate the prediction error filter. To understand this, an input consisting of voiced speech with $L$ harmonics is considered. The resulting prediction error filter will attempt to null the input’s harmonics (or poles) by placing the zeros of $A(z)$ such that the MSE is minimized. In the case where $M \geq 2L$, $A(z)$ can completely suppress the input signal’s harmonics leading to a prediction error variance $P_e = 0$. Although this is the desired effect of LP, the problems with spectral modeling become apparent by noticing that the zeros of $A(z)$ become the poles of $\frac{1}{A(z)}$. When the all-pole model is sampled at the signal’s harmonic frequencies, the filter’s response will feature sharp contours that overestimate the corresponding spectral powers.

Murthi et al. discuss how the problems associated with LP spectral modeling present challenges for speech coding applications. Specifically, these applications often utilize a fixed sampling frequency $f_s$, which dictates the number of observable harmonics in an input signal with a particular pitch frequency $f_0$. This number of harmonics is given by $L = \frac{f_s}{2f_0}$. In the case where the coding scheme uses a fixed order, the spectral envelopes produced by LP will worsen as the pitch frequency is increased. This occurs because the prediction error filter will be able to suppress fewer harmonics more effectively thus resulting in an all-pole model that overestimates the power at the input signal’s harmonics.

3.3 MVDR Spectrum for Speech Modeling

3.3.1 MVDR Development

As discussed in the previous section, LP yields an all-pole model that provides a high resolution indication of the harmonics contained in an input signal as the model order is increased. However, Murthi et al. desire an all-pole model that yields smoother contours and improved spectral power estimation as the model order is increased. These features are essential in providing a good representation of the perceptual qualities contained in a signal’s spectrum.
In order to meet the desired signal modeling objectives, Murthi et al. utilize a modeling approach based on the minimum variance distortionless response (MVDR) spectrum. This spectrum is defined as

$$P_{MV}^{(M)}(\omega) = \frac{1}{\mathbf{v}^H(\omega) \mathbf{R}_{M+1}^{-1} \mathbf{v}(\omega)}, \quad (9)$$

where $$\mathbf{R}_{M+1}$$ is the $$(M+1) \times (M+1)$$ autocorrelation matrix of observed speech data and $$\mathbf{v}(\omega)$$ is a signal vector with $$\mathbf{v}(\omega) = [1, e^{j\omega}, e^{2j\omega}, \ldots, e^{jM\omega}]$$.

The MVDR spectrum can be viewed as a filterbank spectral analysis method similar to the Fast Fourier Transform (FFT) based periodogram. In the case of the FFT periodogram, the power at a particular frequency of an input signal can be observed at the output of a bandpass filter corresponding to that frequency. The bandpass filters comprising the periodogram are equally spaced and thus data and frequency independent. The MVDR spectrum is similar to the periodogram in the sense that the spectrum of a signal can be viewed at the output of a filterbank. However, as the autocorrelation term in (9) indicates, the MVDR spectrum is data and frequency dependent unlike the periodogram.

Conceptually, the design methodology for the MVDR spectra consists of designing a FIR bandpass filter $$h_l(n)$$ for a frequency of interest $$\omega_l$$ such that the signal is passed through the filter without any distortion at that frequency. Murthi et al. refer to this as the “distortionless constraint” and it is defined as

$$h_l(e^{j\omega_l}) = \sum_{k=0}^{M} h_l(k) e^{-j\omega_l k} = \mathbf{v}^H(\omega_l) \mathbf{h}_l = 1, \quad (10)$$

where $$\mathbf{h}_l$$ is a vector containing $$M+1$$ filter coefficients. Solving for the optimal distortionless filter involves finding an $$\mathbf{h}_l$$ that minimizes $$\mathbf{h}_l^H \mathbf{R}_{M+1} \mathbf{h}_l$$, or the total output power of the filter, subject to the constraint in (10). The power minimization subject to the distortionless constraint ensures that $$\omega_l$$ is passed without distortion and all other interfering exponentials are suppressed. This filter design process is repeated for every frequency component in the input signal.

The MVDR filter design methodology significantly differs from LP for speech modeling. In the LP approach, an all-pole model of the vocal tract is obtained by deriving a prediction error filter that minimizes the error for predicted speech samples. The prediction error filter cannot be used to obtain meaningful power estimates at the signal’s frequencies since the filter is designed to null out their influence. On the other hand, the MVDR methodology is based on designing a filter that can pass all the frequencies contained in the signal without distortion provided the model order is high enough. This leads to a filter that yields meaningful power estimates at the signal’s harmonic frequencies. Furthermore, since maintaining the power at the harmonics is a priority, the resulting spectral envelope provided by the all-pole filter will have smoother contours.

### 3.3.2 MVDR All-Pole Model

In order to obtain a parametric representation of a signal’s MVDR spectrum, Murthi et al. utilize a method developed by Musicus that computes the spectrum based on correlating samples of an observed signal [6]. This method relates an observed signal’s MVDR coefficients $$\mu(k)$$ to its LP coefficients. This relationship is indicated by

$$\mu(k) = \begin{cases} \frac{1}{P_e} \sum_{i=0}^{M-k} (M+1-k-2i)a_i a_i^* & \text{for } k = 0, \ldots, M \\ \mu^*(-k) & \text{for } k = -M, \ldots, -1 \end{cases} \quad (11)$$

where $$a_k$$ is the $$k$$th order linear prediction coefficient and $$P_e$$ is the prediction error variance obtained from an $$M$$ order LP analysis. Using the coefficients obtained for $$\mu(k)$$, the MVDR spectrum in (9) can be written as

$$P_{MV}^{(M)}(\omega) = \frac{1}{\sum_{k=-M}^{M} \mu(k) e^{-j\omega_k}}, \quad (12)$$

The MVDR spectrum can be viewed as an “averaging” of several low order LP spectra [1]. In particular, Murthi et al. indicate that an order $$p$$ MVDR spectrum is equivalent to the harmonic mean of the LP spectra computed from orders 0 to $$p$$. This feature makes MVDR useful in spectral modeling applications as will be discussed in the experimental results.
3.4 Experimental Results

Murthi et al. present several simulations in order to compare the performance of MVDR and LP in modeling speech spectra and demonstrate the ability of MVDR to provide smooth spectral envelopes. In these simulations, spectral modeling performance is considered using synthetic speech and real speech samples by varying the model order in order to demonstrate optimal and suboptimal cases.

3.4.1 MVDR Modeling of Synthetic Spectra

In order to demonstrate MVDR’s ability to provide smooth spectral envelopes and improved power estimations over LP, Murthi et al. generate LP and MVDR spectra for a synthetic signal consisting of 10, non-uniformly spaced spectral samples. The purpose of this evaluation is to gauge the performance of MVDR in speech coding applications based on modeling spectral samples. The modeled spectra were generated from the spectral samples using $M = 20$ for LP and $M = 19$ for MVDR. The results of this evaluation are shown in Figure 2 and indicate that the MVDR spectrum (solid line) exactly models the spectral samples when the model order satisfies $M = 2L - 1$.

However, Figure 2 indicates that the peaks of the LP spectrum are close to the frequencies of the spectral samples, but the power at the spectral samples is significantly overestimated.

Murthi et al. conduct an additional simulation to compare distortion incurred with LP and MVDR when modeling spectral samples. In particular, LP and MVDR spectra are generated for model orders, $M = 10, \ldots , 27$, and a log spectral distortion (LSD) is computed based on the power of the known spectral samples and the power estimated by the LP and MVDR spectra. An interesting result from these simulations is that MVDR actually has a high initial LSD when $M = 10$, but decreases monotonically as the order is increased. The LSD incurred with LP is initially low but increases with model order as the power estimations worsen.

These results provide an interesting comparison on the spectral resolution afforded by each method as model order is varied. As the LP model order is increased, its ability to resolve the frequencies contained in the signal improves at the expense of worse power estimation. On the other hand, MVDR provides better power estimates at the signal’s component frequencies, but it actually loses spectral resolution since the spectrum consists of averaging lower order LP spectra.

3.4.2 MVDR Modeling of Voiced Speech

Murthi et al. consider the performance of MVDR on voiced, synthetic speech samples obtained from a database. In the first simulation, the model order is chosen as a function of the number of voiced harmonics $L$ contained in the signal ranging from $1.5L$ to $3L$ in increments of $0.1L$. For each frame of voiced speech in the database, MVDR and LP spectra were generated for the specified $M$ and the LSD was computed and averaged over all frames. The results of this simulation indicate that LP only outperforms MVDR up until $1.6L$ and MVDR yields significantly lower LSD when $M > 2L$. The next simulation involved analyzing the speech samples in the database with each method utilizing a fixed model order ranging from $M = 10, \ldots , 60$ in increments of 5. Again, the LSD was computed and averaged over all frames in the database for each model order. The results indicate that LP outperforms MVDR from $M = 10$ until $M = 35$. However, as $M$ was increased the LSD associated with MVDR monotonically decreased while it stayed relatively constant for LP.

These simulations indicate that the error incurred with spectral power estimation via MVDR monotonically decreases as model order is increased. This is a desirable feature of signal modeling since the additional computation required to obtain higher order models is rewarded with improved performance. Murthi et al. do not explain the reasons why MVDR was outperformed by LP in the previously described situations. One possible reason may be attributed to MVDR requiring a minimum model order so that enough lower order LP spectra are averaged in to the resulting MVDR spectrum. That is, MVDR spectra generated below a certain model order may be too coarse of an approximation for the spectral envelope of speech. Therefore, the
model error resulting from using suboptimal MVDR yields a distortion that is worse than the corresponding distortion with LP.

3.5 Possible Extensions

Murthi et al. present an all-pole speech modeling technique that utilizes the MVDR spectrum in order to improve upon the shortcomings of LP spectral modeling. The MVDR design methodology entails designing a filter that can pass $L$ harmonics of the input frequency without distortion provided the model order is sufficiently high, which results in smooth spectral envelope models. However, the work could be expanded to provide evidence that substantiates MVDR’s ability to yield models that preserve the perceptual qualities of speech or other audio signals. The assumption made by the authors is that the MVDR spectrum provides an improved spectral model since the powers are maintained near the lower frequency harmonics. In particular, subjective tests involving audio synthesis using LP and MVDR models could help determine if MVDR modeling has an advantage by assessing listeners’ preferences. As discussed previously, low order MVDR models can be outperformed by LP models when dealing with voiced speech. Although this may seem like a serious disadvantage, modeling approaches that utilize both schemes in a complimentary manner could be employed since MVDR coefficients are derived from LP parameters.

4 Activities for Psychoacoustic Data Collection and Education

4.1 Objectives

Kim et al. present activities developed in order to collect psychoacoustic data from a diverse subject population with the additional intent of educating middle and secondary school students [2]. The presented activities consist of Tone Bender and the Cocktail Party Game, which both deal with research problems in psychoacoustics. In the former activity, the problem of identifying musical instruments with altered timbre is considered while the latter consists of sound source isolation within a complex mixture. Each activity hosts a creation interface where users can contribute data to the activity that reflects their own perception on these problems in psychoacoustics. This information is evaluated by other users in a listening interface, which is also provided by both activities. The web-based, collaborative architecture employed by Kim allows for a large number of evaluation samples to be collected with a nearly unlimited number of parameter variations. This feature makes the activities useful in helping to understand the perceptual boundaries that dictate human performance in response to the psychoacoustic problems under consideration. An overview of each activity is presented along with quantitative results from a pilot study where the activities were field tested with a group of middle school students. These activities are currently available online and may be found at http://schubert.ece.drexel.edu/~raym/ToneBender/ToneBender.php and http://schubert.ece.drexel.edu/~travis/webpage/WebPage.html.

4.2 Data Collection via Online Activities

The web-based activities developed by Kim et al. draw inspiration from other activities designed to solve complex problems through large-scale, collaborative gameplay. Specifically, the activities presented by von Ahn demonstrate how human participants can be considered part of a distributed system that helps solve computationally intense problems such as labeling and identifying objects in images for web search [7]. von Ahn considers the users of the activities as individual processors in the network all performing a computation by their participation in the activity to help solve the problem at hand. Provided that the entertainment value afforded by the activities encourages participation, the combined efforts of many “human computations” has the capacity to solve difficult problems.

The principles of “human computation” are utilized by the activities developed by Kim et al. in order to facilitate large-scale data collection and solve problems in psychoacoustics. In particular, these activities provide users with a competitive scoring system, instant feedback and the ability to track peer performance so that users are encouraged to keep playing and contributing data. However, the Tone Bender and Cocktail Party Game differ from the platforms presented by von Ahn since they were designed to provide specific educational objectives to benefit users and allows them to contribute their own subjective data to the activities. This feature is significant because it utilizes test data that incorporates many parameter variations based on
the perception of many different users. The entertainment value provided by activities increases their utility as research tools as well since more data will be collected and available for analysis as more people collaborate to participate. This feature makes data collection via web-based activities superior to traditional experiments in psychoacoustics, which rely on a limited number of test subjects and data.

4.3 Tone Bender Activity Description

As the developer of Tone Bender, I designed an activity that illustrates the effects of altering the timbre of musical instruments for the user and collects data on the perception of instruments for the researcher. The player’s objective in the creation interface is to explore various time and frequency parameter variations in order to create an instrument sound that has different timbral qualities but is still identifiable. The manipulated sounds are submitted to the database so that they can be evaluated by players in a listening interface.

The activity’s target audience consists of the general population so no prior knowledge of music, acoustics or engineering is assumed. This necessitates a simple graphical user interface (GUI) to help users understand the objectives while making the activity enjoyable. The parameters chosen to represent timbre consist of the amplitude and spectral envelope representing the audio from a single note played by a particular musical instrument. The amplitude envelope is extracted by rectifying the signal and selecting the largest amplitude peaks over short time durations. For the spectral envelope, the time averaged spectrum of the signal is computed in order to extract the first twenty harmonics that trace the signal’s spectral contour. The player is presented with the amplitude and spectral envelopes in two separate ‘XY’ plots for visualization and modification.

In order to play the activity, the players must first listen to the note produced from the real musical instrument. They can then modify the instrument’s timbre by redrawing the amplitude envelope and manipulating the strengths of the harmonics. The original spectral and amplitude envelopes are always provided by the GUI so the player can gauge the similarity between the signals visually. The player submits their sound parameters to the database when they are certain they’ve produced a recognizable, yet altered instrument. They also receive a potential score based on the SNR calculated between the original and created signal, which indicates the relative difficulty of their sound. In the listening interface, the player must listen to an instrument that was created by another player. Their objective is to identify the instrument they think produced the sound and they receive a score based on the correctness of their response and the perceived difficulty indicated by the SNR. The players’ responses are recorded and submitted to the database.

4.4 Cocktail Party Game Description

The Cocktail Party Game is an activity developed to demonstrate the effect of manipulating physical parameters of acoustic space on the resulting sound mixture. These concepts are demonstrated in the activity by changing the number of speakers and their positions in a simulated room in order to obscure a target speaker, thus emulating the well known “cocktail party problem.” These created rooms are evaluated by players in a listening interface in order to determine if a target speaker is present.

Similar to the Tone Bender activity, the target audience consists of the general population and therefore a simple GUI is required. In the creation interface, the player’s goal is to configure a simulated room so that a target speaker’s voice is obscured, but still identifiable within the mixture. They can achieve this goal by adding interfering speakers, changing speaker locations or adjusting the amount of reverb in the room. When the player is satisfied with the room configuration and sure the target is still identifiable, they submit the room to the database and receive a projected score based on the signal-to-interferers plus noise ratio (SINR). As in Tone Bender, the score is based on difficulty so players are encouraged to create room mixtures that are challenging to identify so they can earn more points if correctly identified. In the listening component of the activity, the player is presented with a target voice and a room mixture, which may or may not contain the target voice. Their objective is to determine if the target voice is contained in the mixture. If their response is correct, they receive a score proportional to the SINR of the associated mixture.
4.5 Pilot Study

The Tone Bender and Cocktail Party activities were field tested with a group of 56 eigth grade students attending a magnet school specializing in music performance. The pilot study persisted over a two day period with each activity featured on separate days. On each day of the trial, the students were divided up into 6 groups consisting of about 10 students per group and each group was allotted 20 minutes to play the creation and listening interfaces in the activities.

As a result of the pilot study, over 800 listening trials and 300 creation samples were collected from each activity. The Results were examined quantitatively by plotting the detection rates (for instruments or speakers) as a function of SNR.

![Graph 1](image1)

![Graph 2](image2)

Figure 3: Detection rates for Tone Bender (left) and Cocktail Party Game (right) [2].

With respect to Tone Bender, the results show that family identification is better than instrument identification over all SNR values. This is a reasonable result since, in general, it is easier to identify the general family of possible instruments that produced an ambiguous sound. An interesting observation form this figure is that the relationship between detection and SNR is not strictly monotonic for either activity. While the players’ performance in both activities indicate an upward trend in detection with increasing SNR, there are variations which suggest that SNR is not completely indicative of how the auditory system perceives complex mixtures. In the case of the Tone Bender activity, there may be cases where an instrument with a low SNR is better identified than another instrument with the same SNR. An example of this may be an instrument that has had significant distortion imparted to the amplitude envelope but little or no change to the spectral characteristics, thus maintaining its identifiability. Such cases in either activity where SNR does not provide enough perceptual information, necessitates further research to determine the salient factors in the auditory system regarding identification of musical sounds and speech mixtures.
5 Broad Connections

5.1 Objectives

The main concepts presented in each of the papers can be unified into a research platform that allows for large-scale collection of psychoacoustic data to develop improved signal modeling schemes that employ knowledge of the human auditory system. Specifically, this will be accomplished by utilizing the “human computation” aspect of the web-based activities so that psychoacoustic data can be analyzed across many samples on a continuous basis. The link between the proposed research platform and data collection arises from the activities presented by Kim et al., while the technical link is provided by the signal modeling schemes proposed by Murthi and Jensen.

5.2 System Overview

An overview of the proposed platform is shown in Figure 4. The technical connections with respect to signal modeling schemes used in the presented papers are provided by the sinusoidal, MVDR and P-ESM blocks. The data collection methods from the psychoacoustics paper are incorporated by the collaborative activities block. The collaborative activities will utilize parametric representations of the audio provided by the signal model blocks so that activity users can manipulate the audio in various ways, thus demonstrating the interconnection between signal modeling and data collection. In the spirit of the activities developed by Kim et al., the goal of the activities will entail stretching the perceptual boundaries of speech and music signals by manipulating the model parameters chosen to represent the signals. Specifically, users will interact with the activities through a GUI, which will abstract the technical details of the audio manipulation to maintain accessibility. Users will receive a score based on their performance in the creation and evaluation aspects of the activities so that a competitive incentive is provided. As individual evaluations are collected over a large sample size, the data will be analyzed by the researcher to establish correlations between the model schemes used and listener performance. This information will be used in the bottom of the pyramid in Figure 4 in order to facilitate research for developing detailed auditory models that reflects the evaluation data collected. These new perceptual models can be used in the development of perceptual codecs (data compression and improved coding) and machine listening applications (automatic sound identification and auditory scene analysis).

![Figure 4: Proposed platform for research in psychoacoustic data collection and evaluation](image)

5.3 Technical Implementation

5.3.1 Signal Modeling

The pyramid in Figure 4 organizes the signal modeling schemes used in the activities into a “perceptual hierarchy” so that the audio signals under analysis may be represented with various degrees of perceptual detail. The sinusoidal model used in the Tone Bender activity does not directly utilize any perceptual information so it is placed outside of the pyramid. The top of the pyramid contains model schemes that represent signals with only a coarse amount of perceptual detail. For example, the MVDR all-pole model developed by Murthi...
et al. is mainly concerned with preserving the signal’s spectral resonances and maintaining the power at low frequency components. This representation is satisfactory for some signals, such as voiced speech, but will not necessarily handle all audio signals well. Models that incorporate more auditory information are placed closer to the bottom of the pyramid. The P-ESM method exemplifies such a method since it represents frequency components of a signal in accordance with a psychoacoustic mask. Although this approach utilizes more auditory information, the resulting models are limited to satisfying a chosen psychoacoustic mask. The bottom of the pyramid in Figure 4 is reserved for highly detailed perceptual models, which are developed using the evaluation results from the activities along with existing models.

The proceeding discussion will describe how the signal models presented by Murthi and Jensen can be used to obtain parametric representations of the audio suitable for use in the activities developed by Kim et al. These schemes may be useful in enhancing the methods used by Kim’s activities, especially with regard to the Tone Bender activity. While this study is limited to MVDR and P-ESM models, the signal pyramid can incorporate additional signal modeling methods as well.

5.3.2 Incorporating MVDR

The all-pole MVDR modeling scheme presented by Murthi et al. offers attractive spectral modeling features that may be useful in extracting spectral components in the Tone Bender activity. The method employed by Kim et al. involves computing a time averaged spectrum and using a fixed order LP envelope to approximate the contours of the time averaged spectrum. This LP envelope is lowered and used as a threshold in order to select the dominant harmonics contained in the signal. This is shown in the figure below.

![Figure 5: Example of harmonic extraction from a musical spectrum (blue) using an LP envelope (green). The extracted components are shown in red.][2]

The problems associated with fixed order LP modeling, as explained by Murthi, is that the LP envelope may overestimate the power at the harmonics. This problem is exacerbated when the signals under consideration are musical and contain a wide range of pitch frequencies. Although the LP model order used in Tone Bender is selected as a “best fit” for the musical signals under analysis, this extraction technique will inevitably miss some important frequencies. The MVDR spectrum provides a smoother spectral envelope that is tolerant to a wide range of pitch frequencies when the model order is sufficiently high. This approach would be useful in Tone Bender since the pitch frequency of the input signal is unknown a priori. Although using high model orders for MVDR necessitates more computation, the MVDR spectra can be generated from LP parameters, which have efficient algorithms such as the Levinson Recursion [5].

5.3.3 Incorporating ESM

The signal analysis schemes used in Tone Bender represent the spectral content of the signal by computing a time averaged spectrum and extracting the most dominant harmonics in that spectrum. This approach provides a reasonable representation of single tones produced by musical instruments, but they cannot be used to parameterize dynamic audio mixtures such as speech or music. Specifically, incorporating noise and the natural flux of the signal’s spectra are essential in modeling the time varying nature of real signals. The subspace modeling approach presented in [3] offers a flexible alternative method of obtaining a parametric
representation of audio. Rather than using a fixed set of parameters, the relevant signal components can be extracted from the audio signal across short time segments so signal fluctuations are captured. By using the exponential sinusoidal model (ESM), signal poles and complex amplitudes can be extracted regardless of the type of signal contained in a particular frame (i.e. harmonic, noise, or mixed). Furthermore, the formulation of the least squares problem to derive the ESM parameters can incorporate perceptual information from an auditory model if desired.

5.4 Data Collection

5.4.1 Collaborative Activities

As demonstrated by Kim et al., collaborative activities have the potential to collect large-scale data that can help solve problems in psychacoustics. This potential is realized by the web-based architecture and entertainment value built into the activities, which allows many different users to participate in the activity while engaging them with competitive incentives. To understand why these features are attractive for the research platform in Figure 4, it is necessary to acknowledge that auditory characteristics vary among individuals. Since the goal of the research platform is to obtain improved signal modeling schemes that incorporate perceptual features, data must be collected across many individuals in order to encapsulate the vast number of variations.

The limitations of traditional experiments in psychoacoustics can be attributed to a fixed number of subjects and test data. This is problematic when studying psychoacoustic problems since the number of parameter variations that can be used for analysis is limited. The web-based activities can overcome these limitations since the activities are made public via the internet thereby allowing any number of participants to join. Furthermore, since the users are required to submit data in order to play the activities, the amount of test data available for evaluation will continue to increase as long as the activities are being used. In order to encourage continuous participation on the activities, they are designed to be enjoyable by providing educational and competitive incentives. Through my own experiences utilizing the Tone Bender activity in the Discovery K-12 program, I have observed that students have an interest in musical instruments thus making it an engaging activity especially for educational use. The entertainment value afforded by these activities also overcomes the problems associated with obtaining volunteers for experiments and negates the need to pay subjects as well. When these features are properly incorporated into the activity component of the research platform, large-scale data collection is possible.

5.4.2 Evaluation Metrics

A suitable evaluation metric, or set of metrics, is required in order to determine the correlations between user performance and perception for the activities in Figure 4. The Tone Bender and Cocktail Party Game activities can potentially gather large amounts of data, but the evaluation metrics used are limited in providing information about the auditory system. These activities utilize SNR (or SINR), which calculate the differences between the original and modified signals on a per sample basis. While SNR indicates the amount of relative difference between two mixtures, it does not provide detailed insight on the perceptual similarities or differences that may exist. As the pilot study from Kim's activities indicates, user detection rates generally increase with SNR, but the relationship is not strictly monotonic, thus suggesting that other factors influence listener perception. For example, it is possible that a user can modify the time-varying amplitude characteristics of a sound in Tone Bender while leaving the spectral characteristic unchanged. This change may still retain the instrument's identifiability despite yielding a high SNR. Likewise, a user could obscure a target voice in the Cocktail Party Game by adding many interfering speakers to the room thus resulting in a high SINR. However, the identifiability of the target is dependent on the timbre of the voices, speaker position, room reverb level and possibly more features which are not indicated by SINR.

The “perceptual hierarchy” imposed by the pyramid in Figure 4 is intended to incorporate several signal modeling schemes in order to help determine the salient factors in perception of manipulated audio. By employing aspects of multiple modeling schemes, it is possible to develop evaluation metrics that provide a better indication of the perceptual differences between two mixtures. For example, when considering the spectral characteristics of an audio signal, all-pole schemes can provide a spectral envelope and a psychoacoustic mask can be selected to determine the perceptually important frequency components that lie above the masking
threshold. These features could be combined into a “weighted” spectral envelope where variations imparted on the original signal by the user will correlate to differences based on the most sensitive regions of the spectrum. The power of such metrics can be assessed by considering user performance in the evaluation component of the collaborative activities. Since the perceptually sensitive regions are known a priori, the evaluation results could verify the salience of this model and possibly provide insight into other salient features. As more modeling schemes are incorporated into the signal pyramid in Figure 4, additional evaluation metrics can be developed to drive the development of signal models which incorporate detailed auditory information.

5.4.3 Data Collection and Educational Outreach

Although the primary objective of the research platform is to develop detailed signal models, the component containing collaborative activities has the potential to complement math and science coursework in K-12 education. As part of the pilot study conducted by Kim et al. and through my participation in the Discovery K-12 (DK-12) program, I have observed that middle and high school students enjoy using the activities because they allow learning through experimentation. For example, students can connect the visual representation of timbre with instant audio feedback when manipulating audio in the Tone Bender activity. Similarly, students can explore the factors influencing speaker identification in the Cocktail Party Game as well. The web-based architecture of the activities benefits the course instructor since the activities are readily available to supplement math and science coursework. In terms of evaluation, the activities record each user’s sessions so the instructor can track the students’ performance to determine if the activities are helping them learn about the chosen topics.

5.5 Conclusions

The proposed research platform discussed in this section demonstrates how signal modeling and data collection schemes can facilitate research in the development of perceptual audio models. The research platform unifies existing signal modeling schemes and data collection methods so that evaluation data can be collected and analyzed by the researcher in order to develop detailed evaluation metrics and signal models. This section discussed how the signal modeling schemes presented in the study could be used within the research platform although it is flexible enough to handle additional schemes, which are organized into the signal pyramid based on the amount of perceptual detail they contain. Additionally, the feedback between the collaborative activities and the pyramid allows the researcher to continuously refine the models based on the evaluation data collected. By unifying audio signal modeling and data collection techniques, this platform promotes a continuous cycle of research that can ultimately lead to improved signal models and the development of applications which utilize such models.
A Spectral Modeling with MVDR Spectra

A.1 Objectives

As discussed in Section 2, the MVDR spectrum is a useful tool for generating smooth spectral envelopes for speech signals. Consequently, this feature is attractive for extracting frequency components from spectra since the MVDR spectral envelope provides a better indication of the spectral powers especially at the lower and perceptually important frequencies. This section will demonstrate how MVDR can be used to parameterize musical signals by extracting component frequencies from the spectrum.

A.1.1 Spectral Threshold via MVDR

One way frequencies can be extracted via the MVDR spectrum is very similar to the method used by Kim et al. to model musical tones in the Tone Bender activity. In this approach, a linear prediction (LP) envelope is computed that provides a spectral envelope approximation of the signal’s time averaged spectrum. This envelope is lowered to provide a threshold for selecting the prominent peaks, which represent the dominant frequencies in the signal. However, as shown in Figure 5 from the previous section, the LP envelope does not always yield a good approximation of the signal’s spectral contour. This is problematic when selecting peaks since certain harmonics can be skipped over, or spurious peaks arising from noise can be mistaken for frequency components.

Utilizing the MVDR spectrum as a threshold can help alleviate the problems associated with LP methods by utilizing smooth spectral envelopes that better approximate the power at the signal’s frequencies. Not only does this improve the spectral extraction process, but it also improves modeling of the perceptually important frequencies. To illustrate these advantages, a musical signal consisting of a bassoon playing a pitch frequency $f_0 = 130$ Hz and sampled at $f_s = 22050$ Hz will be considered. The signal’s frequency response is indicated in Figure 6 along with the associated LP and MVDR computed using the techniques described by Murthi.

![Frequency Response of Musical Signal with LP and MVDR Spectra. M = 80](image)

Figure 6: Musical signal spectrum plotted with associated LP and MVDR spectra. $M = 80 - 2$

The LP and MVDR spectra in Figure 5 were both generated using model order $M = 80$. Since the analyzed signal contains a pitch frequency $f_0 = 130$ Hz, the number of observable harmonics can be computed by $L = \frac{f_s}{f_0}$ or 84 for this example. The order required for MVDR to perfectly model $L$ harmonics in a signal is given by Murthi as $M \geq 2L - 1$. As Figure 6 indicates, the MVDR spectrum approximates the contours of the spectrum fairly well despite using suboptimal model order. To illustrate the optimal case where $M \geq 2L - 1$, the LP and MVDR spectra are generated with $M = 168$. As shown in Figure 7, the resulting MVDR spectrum features improved modeling of the signal’s harmonics, especially near the perceptually important frequencies, which are characterized by larger amplitudes. On the other hand, the effect of increasing the model order results in the LP spectrum overestimating some of these frequencies. This consequence of LP modeling results from the minimization of the MSE as discussed in Section 2, which results in sharp contours from trying to null out the signal’s poles. It should be noted that although the model order was chosen to represent the...
optimal case, the power at all the component frequencies is not exactly modeled by MVDR spectrum in Figure 7. This is due to the non-ideal nature of the signal since it is of finite data length and contains some degree of noise.

To demonstrate how harmonics can be extracted from a spectrum using the LP or MVDR spectra, the bassoon signal used in Figures 6 and 7 is again considered. As was demonstrated in the previous discussion, a sufficiently high order MVDR spectrum approximates the contours of the signal’s spectrum, yielding a close indication of the signal’s true magnitude near the pole frequencies. Also, as discussed in the review of the Tone Bender activity, the LP spectrum does not yield a good spectral envelope as order is increased thus making it difficult to extract frequency components. LP and MVDR spectra are used as thresholds for frequency component extraction in Figure 8. Each spectra was generated with \( M = 168 \) and lowered by 8 dB’s in order to isolate the signal’s spectral peaks. By isolating spectral regions above the threshold, a local maxima search can yield the signal’s largest magnitude pole frequencies.

By comparing the plots in Figure 8, it is clear that the smooth contour of the MVDR spectrum does a better job of isolating the large amplitude peaks in the signal’s spectra while ignoring the spurious peaks from the signal’s frequency response. On the other hand, the LPC extraction approach is sensitive to the spurious peaks that are produced by noise and side-lobe interaction in the frequency response.

A.1.2 Frequency Extraction with MVDR and LP Spectra

As the previous simulations and the work by Murthi et al. indicate, LP and MVDR spectra have complementary signal modeling features. When the model order is sufficiently high, the LP spectrum provides a strong indication of the frequency component locations within the signal, although the power at these frequencies is not well represented. On the other hand, the MVDR spectra provides a better estimate of the spectral
powers. By using both spectral modeling techniques, it is possible to represent the frequency locations and power estimates of a voiced (or harmonic) signal using only LP and MVDR parameters.

In [8], Dubnov demonstrated that MVDR and LP spectra can be used together to isolate frequency components in audio mixtures. Dubnov’s methodology relies on seeking large differences between the MVDR and LP spectra. In particular, if the LP spectra at a particular frequency possessed a significantly larger amplitude than the corresponding frequency in the MVDR spectra, it is likely that this component represented a partial from the signal. This assumption draws on the LP design methodology, which causes the LP spectra to overestimate the signal’s true spectral powers in order to minimize the error.

The component extraction method discussed in this section utilizes both the MVDR and LP spectra as in Dubnov’s approach, but also shows that the signal can be compactly represented by computing its LP parameters and the associated MVDR spectrum. The bassoon signal from the previous simulations is again considered to demonstrate the frequency extraction technique. The LP parameters are computed for the signal using $M = 160$. The poles of the signal are obtained by calculating the zeros of the resulting prediction error filter, which is parameterized by the LP coefficients. By computing the angle of the poles, the frequencies in the signal can be obtained. The corresponding MVDR spectrum with $M = 160$ is derived using equations (11) and (12) from Section 3 in order to obtain the coefficients and spectrum, respectively. The power estimates of the signal’s frequencies are obtained by evaluating the MVDR spectrum at the pole frequencies associated with the LP spectrum. In order to ensure that the perceptually important, large amplitude signals are modeled, the LP and MVDR magnitudes at the pole frequencies are required to be within a certain threshold so that relatively low amplitude components are ignored. This process is indicated in Figure 9 where the LP and MVDR spectra are plotted along with the extracted peaks indicated by the red x’s.

![Figure 9: Frequency component extraction using LP pole frequencies and MVDR spectrum. $M = 160$](image)

The potential of the approach demonstrated by Figure 9 lies in that only the LP and MVDR parameters of a signal are required to represent the dominant frequencies in a harmonic signal. Since a high order LP spectra provides an indication of the line frequencies contained in a signal and MVDR provides a good estimate of the power, the signal’s spectrum is not required for thresholding in order to extract the harmonics. This approach is very powerful for component extraction and modeling especially in coding applications where bandwidth concerns make transmitting the entire signal unreasonable. Furthermore, the representation is compact since the MVDR parameters are derived directly from LP parameters of the same order.
References


