Signal Interaction and the Devil Function

John R. Hershey, Peder A. Olsen, Steven J. Rennie

IBM T. J. Watson Research Center
{jrhershey,pederao,sjrennie}@us.ibm.com

Abstract

It is common in signal processing to model signals in the log power spectrum domain. In this domain, when multiple signals are present, they combine in a nonlinear way. If the phases of the signals are independent, then we can analyze the interaction in terms of a probability density we call the “devil function,” after its treacherous form. This paper derives an analytical expression for the devil function, and discusses its properties with respect to model-based signal enhancement. Exact inference in this problem requires integrals involving the devil function that are intractable. Previous methods have used approximations to derive closed-form solutions. However it is unknown how these approximations differ from the true interaction function in terms of performance. We propose Monte-Carlo methods for approximating the required integrals. Tests are conducted on a speech separation and recognition problem to compare these methods with past approximations.

1. Introduction

Signals are often analyzed and classified using models of their log power spectra. In the context of noise, or any other interfering signal, such model-based classifiers must compensate for the noise in some way. Model-based noise compensation requires a signal interaction model, which describes the effect of adding two signals on the resulting acoustic features. Traditionally the influence of phase has either been ignored through the use of approximate interaction models, or has been diminished by averaging, especially when working in the log spectrum domain.

We describe and illustrate the signal interaction model, which we call the “devil function.” Exact inference using this function is difficult because the required integrals are intractable. Even efficient approximate inference can be elusive. Just how important it is to accurately model this signal interaction is an empirical question. To address this question we perform experiments using accelerated Monte Carlo simulations that can approximate inference arbitrarily well if enough samples are used, and compare the results to simpler approximations on a speech separation and recognition task.

2. Interaction models

The relationship between log power spectra \(x, n\), with phases \(\theta_x, \theta_n\), and the log power of their mixture \(y\) is

\[
y = \log((e^{\frac{x}{2}} + e^{i\theta_x} + e^{\frac{n}{2}} + i\theta_n)^2) \quad (1)
\]

\[
= \log(e^x + e^n + 2e^{\frac{x+n}{2}} \cos(\theta)), \quad (2)
\]

where \(\theta \equiv |\theta_x - \theta_n|\) is the phase difference.

For convenience we define the function \(v(x, n, \theta) \equiv \log(e^x + e^n + 2e^{\frac{x+n}{2}} \cos(\theta))\). Note that \(v\) obeys a shift property, \(v(x + s, n + s, \theta) = v(x, n, \theta) + s\), so adding log gain to both sources is equivalent to adding log gain to the mixture. There is also a symmetry property due to the cosine: \(v(x, n, \theta) = v(x, n, 2\pi - \theta)\).

We can use this relation to define an interaction function,

\[
p(y|x, n, \theta) \equiv \delta(y - v(x, n, \theta)), \quad (3)
\]

where \(\delta(.)\) is the Dirac delta function. The surface defined by \(v(x, n, \theta) = y\), for fixed \(y\), we call the unicorn function after its single-pronged shape (see Figure 1). To perform inference in this model, given priors on \(x\) and \(n\), we have to compute

\[
p(y) = \int p(x)p(n)p(\theta)\delta(y - v(x, n, \theta)) \, d\theta \, dx \, dn
\]

\[
= \int p(x)p(n)p(y|x, n) \, dx \, dn, \quad (4)
\]

where \(p(y|x, n) = \int p(\theta)\delta(y - v(x, n, \theta)) \, d\theta\). It is also useful to compute the posterior expected value.

\[
E(x|y) = \int x \frac{p(x)p(n)p(y|x, n)}{p(y)} \, dx \, dn \quad (5)
\]

An empirical histogram of \(p(y|x, n)\) is shown in Figure 2, for mixed speech signals, along with the histogram, \(p(x, n)\), of the original signals prior to mixing.

Figure 1: The unicorn function, defined by \(v(x, n, \theta) = y\), for \(y = 0\). Axes (counter clockwise from top left) are \(n\), \(x\) and \(\theta\).

Define the inverse function of \(v(x, n, \theta)\) that yields \(\theta\) as a function of \(y\) in the range \([y_{\text{min}} = v(x, n, \theta = \pi), \ y_{\text{max}} = v(x, n, \theta = 0)]\), for fixed \(x\) and \(n\):

\[
\Theta(y|x, n) \equiv \arccos \left( \frac{e^y - e^x - e^n}{2e^{\frac{x+n}{2}}} \right) \in [0, \pi]. \quad (6)
\]

\(\Theta(y|x, n)\) is a monotonic function ranging in value from \(\pi\) at \(y_{\text{min}}\) down to 0 at \(y_{\text{max}}\). Since \(\theta\) is uniformly distributed, and considering the symmetry of \(v\) over the ranges \(\theta \in [0, \pi]\), and
\[ \theta \in [\pi, 2\pi], \text{ the cumulative distribution function (cdf) of } y \in [y_{\min}, y_{\max}] \text{ is } P(y \leq y|x,n) = 2P(\theta \leq \Theta(y,x,n)) = 1 - \frac{\Theta(y,x,n)}{\pi}, \]

where the factor of 2 accounts for the symmetry in \( \theta \). The probability density function (pdf) \( p(y|x,n) \) can be found by taking the derivative with respect to \( y \),

\[ p(y|x,n) = \frac{\partial}{\partial y} P(y \leq y|x,n) = \frac{e^{y-c}}{\pi \sqrt{1 - \frac{1}{4} (e^{y-c} - e^c - e^{-c})^2}}, \]

where \( c = (x+n)/2 \) and \( z = (x-n)/2 \).

Note that the allowable region for \( y, x, \) and \( n \) is such that \( y = v(x,n,\theta) \) for some \( \theta \). Outside the allowable region, \( p(y|x,n) = 0 \). Along the edges of this region, defined by \( \theta \in \{0, \pi, 2\pi\} \), \( p(y|x,n) = \infty \). These singularities make the distribution somewhat tricky to work with. Because of these hazards, we call \( p(y|x,n) \) the devil function (see Figure 3). Like the uniform function, the devil function obeys a shift property, \( p(y-s|x,n) = p(y|x+s,n+s) \). A similar distribution can be derived in the amplitude domain, and is known in the wireless communications literature as the two-wave envelope pdf [1].

![Figure 2: The empirical histogram \( P(y|x,n) \), obtained from 120 mixtures of utterances from two speakers. Also shown are the max and log-sum approximations](image)

![Figure 3: The devil function, \( p(y|x,n) \). Horizontal axes are \( x \) and \( n \) with density on the vertical axis.](image)

### 3. Approximate interaction functions

The devil function is traditionally avoided via various approximations. The log-sum approximation, used in [2, 3, 4], uses the log of the expected value in the power domain \( E(e^y) = e^x + e^z \).

\[ p(y|x,n) \approx N(y \mid \log(e^x + e^z), \psi) \overset{\text{def}}{=} p_{\logsum}(y|x,n), \]

where \( \psi \) is a variance designed to compensate for the effects of phase. The log-sum approximation can be written \( \log(e^x + e^z) = \max(x,n) + \log(1 + e^{-(x-n)}) \), which motivates the max approximation:

\[ p(y|x,n) \approx \delta(y - \max(x,n)) \overset{\text{def}}{=} p_{\max}(y|x,n), \]

where \( \delta(\cdot) \) is the Dirac delta function. It was shown in [5] that if the phase difference is uniformly distributed, then the expected value of the log power of the sum of the signals is \( E_p(y|x,n) = \max(x,n) \). The max model was first used in [6] for noise adaptation, and in [7], to compute state likelihoods. Other approximate interaction functions have been proposed to handle the phase term for log averages over power spectrum bins, such as [8].

### 4. Naive Monte Carlo

We can instead estimate \( p(y) \) directly using Monte Carlo methods. It is tempting to directly use the devil function, but the fact that it goes to infinity in some places and is zero over much of the space is a potential pitfall. Instead, here we use the uniform function, replacing the delta function with a Gaussian with variance \( \psi \). Taking samples \( \{x_i, n_i, \theta_i\} \) from a proposal distribution \( h(x,n,\theta) \), the standard importance sampling approximation [9] is

\[ p(y) \approx \frac{1}{N} \sum p(x_i)p(n_i)p(\theta_i)N(y\mid v(x_i,n_i,\theta_i),\psi) \]

This sampling scheme can suffer from the fact that the samples may be far from the true posterior, and far from the set of allowable points, \( x, n \) given the observation. In that case, the estimated posterior mean \( E(x|y) \) may be inaccurate.

### 5. Sampling the devil function

Let us consider the devil function more carefully. We have

\[ p(y) = \int p_x(c)p_n(n)p(\theta)\delta(y - v(x,n,\theta)) \, dx \, dn \, d\theta \] (14)

We introduce the change of variables used in (10), via a transformation matrix,

\[ \begin{bmatrix} z \\ c \end{bmatrix} = A \begin{bmatrix} x \\ n \end{bmatrix} = \begin{bmatrix} x+n \\ \frac{x+n}{2} \end{bmatrix}, \quad A = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \]

(15)

and a new function,

\[ u(z,\theta) \overset{\text{def}}{=} \log(e^z + e^{-z} + 2 \cos \theta) \]

\[ = \log(2 \cosh(z) + 2 \cos \theta), \]

(16)

so that \( v(x,n,\theta) = c + u(z,\theta) \). Then

\[ p(y) = \frac{1}{2\pi} \int p_x(c+z)p_n(c-z)\delta(y - c - u(z,\theta)) \, dc \, dz \, d\theta \]

(17)

\[ = \frac{1}{\pi} \int p_x(y - u(z,\theta) + z) p_n(y - u(z,\theta) - z) \, dz \, d\theta \]

(18)
Figure 4: Max model: a) the prior normal density \( p(x, n) \), with max interaction shown in red. b) the posterior \( p(x, n|y) \).

where \( x \equiv y - u(z, \theta) + z \) and \( n \equiv y - u(z, \theta) - z \). Note that \( u(z, \theta) \approx |z| \) when \( |z| \) is large, and \( c + |z| = \max(x, n) \); that is, the max approximation is accurate for large \( |z| \).

We can then approximate \( p(y) \) by sampling \( \theta_i \sim U(0, 2\pi) \) and \( z_i \sim h(z) \), and computing \( \tilde{x}_i = y - u(z_i, \theta_i) + z_i \) and \( \tilde{n}_i = y - u(z_i, \theta_i) - z_i \).

\[
p(y) \approx \frac{2}{N} \sum_i p_x(y - u(z_i, \theta_i) + z_i) p_n(y - u(z_i, \theta_i) - z_i)
\]

\[
= \frac{2}{N} \sum_i p_x(\tilde{x}_i) p_n(\tilde{n}_i) \frac{h(z_i)}{h(z)}
\]

The \( \tilde{x}_i \) and \( \tilde{n}_i \) we have thus derived can be seen as samples from the devil function multiplied by \( h(z) \).

6.1. Max model importance sampling

The question remains of what proposal distribution \( h(z) \) to use. We could again start with the priors, this time projecting points such that they satisfy the constraints imposed by \( y \). However, such samples still may not fall predominantly in the region of the posterior, which is necessary for a good approximation. An easy option is to sample from a wide range of \( z \) values, but this may be expensive in the number of samples. A better choice is to sample \( z \) values from an approximation to the posterior given the observation and the prior, such as provided by inference in either the log-sum or the max model. We call this approach "refined importance sampling," because it makes an educated guess about the region of interest.

6.2. Algonquin importance sampling

In Algonquin, \[10, 11\] the log-sum interaction function is used. To handle the intractable likelihood and expected value integrals, the log-sum function is linearized, which yields a Gaussian approximation to the posterior. This posterior in turn gives a better linearization point and the process is iterated. The Gaussian posterior,

\[
p \left( \begin{bmatrix} x \\ n \end{bmatrix} \bigg| y \right) \approx N \left( \begin{bmatrix} x \\ n \end{bmatrix} \bigg| \begin{bmatrix} \eta_x \\ \eta_n \end{bmatrix}, \begin{bmatrix} \phi_{xx} & \phi_{xn} \\ \phi_{nx} & \phi_{nn} \end{bmatrix} \right),
\]

can be used to obtain a sampling distribution in \( z \):

\[
h(z) = N(z|\eta_x - \eta_n)/2, (\phi_{xx} + \phi_{nn} - 2\phi_{xn})/4
\]

Figure 5: Integration using control variates: a) (solid) the function to be integrated, b) (dashed) the control variate approximation, c) (blue) the known integral of the approximation, and d) (green) the remaining area to be estimated.

When the two signals \( x \) and \( n \) are conditioned on state variables in a larger model, we typically have to estimate the likelihood under all combinations of states. This presents one drawback of Algonquin-based importance sampling: it requires independent samples under all combinations of states.

7. Control variates

The method of control variates [9] gives us another means to improve the sampling efficiency. In this method we use an integrable approximation to the function of interest, and use sampling to compute the difference between the exact function and its approximation, as shown in Figure 5. This results in a faster convergence to the true integral.

Let us consider, for example, using the max model as a control variate for sampling from the devil function. The max model likelihood integral, given in (20) can be transformed according to (15), noting that \( \max(x, n) = c + |z| \):

\[
p_{\max}(y) = \int p_x(x)p_n(n) \delta(y - \max(x, n)) \, dx \, dn
\]

\[
= \int p_x(c + z)p_n(c - z) \delta(y - c - |z|) \, dx \, dz
\]

\[
= 2 \int p_n(y - |z| + z) p_x(\eta_x - \eta_n) \, dx \, dz
\]

\[
= p_n(y) \Phi_n(y) + p_n(y) \Phi_n(y)
\]

\[
E_{\max}(x|y) = \pi_x y + \pi_n E_{\max}(x|x < y).
\]
Table 1: Percent word error rate (WER) on a speech separation task using different Monte Carlo methods to estimate likelihoods and expected values, as a function of the number of samples used per Gaussian combination. Baseline for the max model is 23.9% and for Algonquin is 24.9%.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Number of Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10</td>
</tr>
<tr>
<td>Naive Monte Carlo</td>
<td>30.8</td>
</tr>
<tr>
<td>Devil Function MC</td>
<td>33.3</td>
</tr>
<tr>
<td>Max Control Variate</td>
<td>24.6</td>
</tr>
</tbody>
</table>

For Gaussian $p_k(x)$, with mean $\mu_k$, and variance $\sigma_k^2$, it can be shown that the expected value, $E_{\text{max}}(x|x < y)$, is $\mu_x - \sigma_x^2 p_k(y)/\Phi_k(y)$ [12]. Defining $x' = y - |z_i| + z_i$, and $n' = y - |z_i| - z_i$, we can thus use the max model as a control variate for the devil function:

$$p(y) \approx p_{cv_{\text{max}}}(y) = \alpha \left( p_k(y) \Phi_k(y) + p_n(y) \Phi_n(y) \right)$$

$$+ \frac{2}{N} \sum_i \left( p_k(\overline{x}_i) p_n(\overline{n}_i) - \alpha p_k(x'_i) p_n(n'_i) \right) / h(z_i),$$

where $\alpha$ is a free parameter that can be chosen to reduce the variance of the estimator, or to keep the estimate positive. The latter term corrects for inaccuracies in the max model by measuring the difference in likelihood between samples from the devil function, $x'_i, n'_i$, and the projections $x'_i, n'_i$ of those samples onto the max function. The expected value is given by

$$E(x|y) \approx E_{cv_{\text{max}}}(x|y) = \alpha E_{\text{max}}(x|y)$$

$$+ \frac{2}{N} \sum_i \left( \frac{x'_i p_k(\overline{x}_i) p_n(\overline{n}_i) - \alpha x'_i p_k(x'_i) p_n(n'_i)}{h(z_i)p_{cv_{\text{max}}}(y)} \right).$$

Similarly Algonquin and many other methods may be used as control variates.

8. Experiments

Experiments were conducted on a speech separation and recognition task [13] using the model-based approach described in [14]. The model consists of a hidden Markov model for each speaker with Gaussian observation densities in the log spectrum domain. The front-end of the system in [14] was modified to eliminate spectral averaging that would alter the attention function. Inference involves computing the likelihoods (4) and expected values (5). The likelihood values are computed for each of 319 frequency bins, for every combination of 256 acoustic Gaussians for the pair of speakers, $x$ and $n$, and band quantization is used to speed up this computation as described in [14]. The 2-D Viterbi algorithm is used to estimate state sequences through the task grammar. The time-domain signal is then reconstructed using the expected values of the hidden signals $x$ and $n$, conditioned on the state sequences, and the result is fed to a conventional recognizer.

Results for different Monte Carlo algorithms and sample numbers are presented in Table 1 for the 0 dB signal-to-noise ratio (SNR) condition, which contains 600 two-speaker mixtures of utterances from a six-word grammar. The naive Monte Carlo algorithm, which sampled from the prior Gaussians fared the worst, whereas the devil-function sampling, using a Gaussian importance-sampling distribution derived from the mean and variance speech model, worked better. The control variate method, using the max model and the same Gaussian importance sampling distribution, worked best, and the 22.7% WER compares favorably with the Algonquin baseline of 24.9% and the max model baseline of 23.9%.

Although the improvements relative to baseline are not dramatic, all of the methods outperform the human listener scores of 28.8% in [13]. The speech recognition task is a relatively complicated test of these methods, and may be dominated by other factors such as the constraints of the task grammar. Future work is underway to directly measure the convergence of the likelihood and expected value estimates in isolation to further validate the proposed Monte Carlo methods. It should be noted that the Monte Carlo methods are quite slow in comparison with the max model and log-sum methods: the devil function is so far mainly of theoretical value. It is also still unknown whether the integrals were approximated well enough to give a clear comparison with the max model and log-sum methods. However, early indications are that there is room for improvement to these models.

9. References