Extracting More Detail from the Spectrum with Phase Distortion Analysis

Paul Masri
Nishan Canagarajah
Digital Music Research Group, University of Bristol
5.01 Merchant Venturers Building, Woodland Road, Bristol BS8 1UB, UK
Paul.Masri@bristol.ac.uk http://www.fen.bris.ac.uk/elec/dmr/

Abstract

In the sinusoidal analysis of sound, using the Short Time Fourier Transform (STFT), there is the assumption that the signal is locally stationary within each FFT frame. If, as in practice, this assumption is violated, the spectrum becomes distorted. Phase Distortion Analysis (PDA) was introduced in 1995 [1] to enhance the analysis of degraded peaks, by using the distortion itself as a source of information about the signal nonstationarity. It was shown that the first order frequency and amplitude modulation could be measured from the degree of phase shift close to the maximum of the mainlobe peak. This paper presents advances with the PDA technique, in particular a neural network implementation that makes estimation robust to noise. The capability to analyse nonstationarities relaxes the restraint on keeping the FFT analysis window short and therefore effectively improves time-frequency resolution. This, in turn, promises greater analysis-synthesis quality through improved identification and tracking of partials during the analysis phase.

1 Introduction

Sinusoidal analysis-synthesis of sound has its basis in Fourier theory, which states that any periodic waveform may be constructed by the superposition of harmonically related sinewaves and that the relationship between a particular waveform and the set of sinusoids is unique. McAulay and Quatieri are credited with first applying this to the analysis and resynthesis of sound [2] and their foundations persist in today's models such as Spectral Modelling Synthesis (SMS) [3].

The Short Time Fourier Transform (STFT) assumes a local approximation to true (i.e. unchanging) periodicity. The spectrum of each FFT frame is the convolution between the ideal local spectrum and the analysis window's spectrum. From the spectral peaks (window function mainlobes), estimation of the frequency, amplitude and phase of sinusoids is straightforward.

In reality, not all sounds change slowly (and even slowly changing sounds may contain rapidly changing higher harmonics). In these cases, the assumption is violated to some extent and the spectral peaks suffer apparent distortion that affects the amplitude and phase shapes.

2 Phase Distortion Analysis

On the basis that the FFT spectrum contains *all* information about a time domain signal, whether the signal is stationary or not, the Phase Distortion Analysis (PDA) method [1] was introduced to make use of the apparent distortion, treating it as a source of information. Observation showed that the degree and direction (rising or falling) of frequency and amplitude modulation affects the phase spectrum uniquely. Hence the PDA method was devised to estimate modulation parameters from measurements of phase shift (with respect to the flat phase response of a stationary sinewave).

Since, in practice, a sound spectrum will contain many peaks, phase measurements are made within the mainlobe peak, close to the maximum, where the magnitude of the sinusoid of interest is maximal. By zeropadding the time domain window, it is possible to take these measurements closer than a single unpadded bin. In this paper, a Hamming window is used, of length 1023 samples and the FFT is zeropadded to 8192 samples. This is loosely called 8× zeropadding. (Note that because PDA is empirical, the actual amount of measured phase shift is specific to the window function shape and duration and the distance from the mainlobe maximum at which measurements are taken, although the method of estimation is generic.)

This work has been supported by funding from the Engineering and Physical Sciences Research Council (EPSRC) of the UK under the ROPA scheme. Project code: GR/K66550.

The notation Φ_{N+} is used to denote the phase shift at a positive offset of N zeropadded bins.

i.e. $$\Phi_{N+} = \Phi_{MAX+N} - \Phi_{MAX} \quad (1)$$

and $$\Phi_{N-} = \Phi_{MAX-N} - \Phi_{MAX} \quad (2)$$

where Φ is phase, its suffix is the frequency location, suffix MAX is the mainlobe maximum, suffix N is a positive integer.

So far, phase distortion analysis has been defined for first order modulation; that is, linear frequency modulation (1FM) and/or exponential amplitude modulation[1] (1AM). In this paper, the symbols *delf* and *dela* are used to represent the degree of modulation, measured in *unpadded* bins per frame and dB per frame, respectively, where the frame is the duration of the time domain window (actually, rounded up to the next power of two, i.e. 1024). See figure 1.

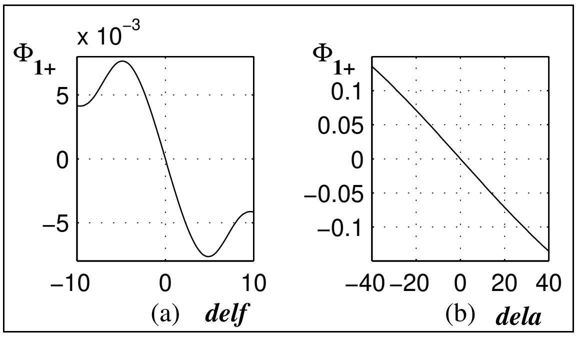

Figure 1. a) phase shift at Φ_{1+} as *delf* varies (*dela*=0); b) phase shift at Φ_{1+} as *dela* varies (*delf*=0). (Φ_{1+} shown in radians.)

In [1], it was suggested that the total phase shift was the sum of phase shifts due to 1FM and 1AM. Further investigation over a wider range of *delf* and *dela* have revealed that this was a good approximation within the range $delf \in [-1,+1]$ & $dela \in [-1,+1]$. However, over wider ranges, the relationship is more complex; see figure 2. Nevertheless, *delf* and *dela* can still each be estimated from just two phase shift measurements.

3 Neural Network Development

For neural network implementation, training and validation data were generated that spanned the (*delf*, *dela*) space in regular intervals. At each location, a modulated sinusoid was generated, windowed, zeropadded, FFT'd, phase measurements were taken and the phase shifts were calculated as per (1) and (2) above.

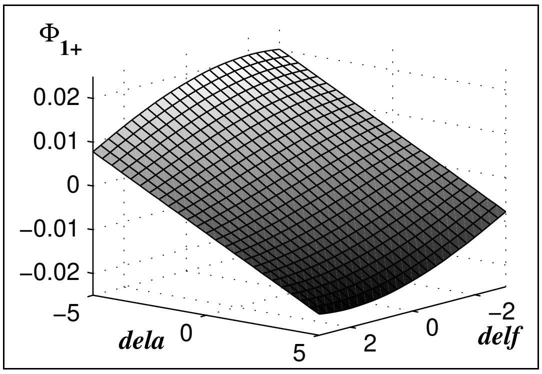

Figure 2. Phase shift as a function of first order modulation (for the example of Φ_{1+}).

All the network configurations described are feedforward, with one or two hidden layers and full connectivity. The activation function of each neuron is a tan-sigmoid (whose output ranges between –1 and +1). Training was by the Levenberg-Marquardt method of backpropagation, with phase shifts as inputs and modulation estimates as targets/outputs. Performance was measured as the standard deviation of the network error (i.e. NN output minus target value). This is denoted σ_{TRAIN}, σ_{VAL} or σ_{TEST} depending respectively upon whether the measure is for training data, validation data or test data.

3.1 Reproducing Established Results with a Neural Network

Initial experiments began using a 2-input 2-output network with a single hidden layer and training data within the range $delf \in [-1,+1]$ & $dela \in [-1,+1]$. This had the purpose of replicating the results in [1] using a NN, in order to verify that this approach can work and to assess its accuracy.

It quickly became apparent that the accuracy of the NN was poor, with either *delf* or *dela* being better approximated to the detriment of the other. The next set of experiments separated the outputs, creating an independent 2-input NN for each. This vastly improved accuracy, suggesting that, although both *delf* and *dela* are estimated from the same input data, their functions are radically different. σ_{TRAIN} and σ_{VAL} were reduced to the order of 0.005, using around 10 hidden layer neurons and sufficiently dense training data. This is probably sufficient accuracy for most applications. Figure 3 (overleaf) shows an example error surface. Hereon examples refer to *delf* only, but are equally applicable to *dela.*

[1] For most magnitude measurements, the decibel (dB) scale is used. On this logarithmic scale, 1AM appears as linear modulation. Hence, for simplicity, in this paper it is referred to as first order modulation.

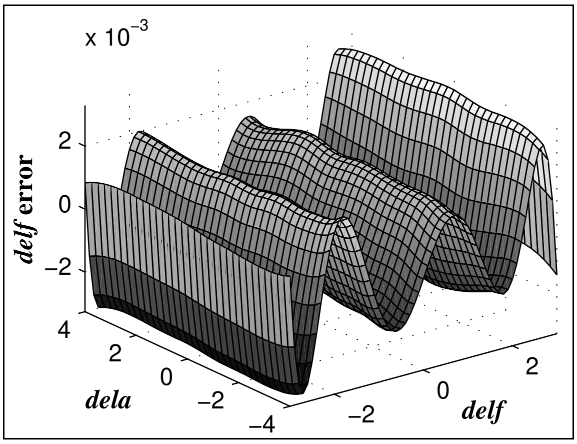

Figure 3. An example error surface for a 2-input neural network, trained to estimate *delf.*

Having achieved the first goal of replicating the results from [1], further experiments with the 2-input 1-output configuration were done to extend the functionality, with the following results:

- The range of the training data was increased up to $delf \in [-4,+4]$ & $dela \in [-5,+5]$, with 41 and 25 data points in the ranges respectively. Because the relationship between *dela* and phase shift function is smoother (see figure 1b), less points were required per unit.
- When using data outside the range $delf \in [-1,+1]$ & $dela \in [-1,+1]$, it was necessary to scale the target and network output data into the range $(-1,+1)$, to match that of the neuron activation function. In practice, target data was actually scaled into the range $[-0.8,+0.8]$, since a large input change is required to achieve a significant output change outside this range, making training more difficult.

3.2 Implementing Noise Resilience

It has already been demonstrated that only two phase shift measurements are necessary in order to estimate *delf* and *dela* to a sufficient accuracy. However, this is the minimal solution and is susceptible to noise in the time domain signal, which is likely in practice. The aim, in this section, is to make the estimator robust in the presence of noise by making use of more phase shift measurements, on the basis that the effects of noise upon each phase shift measurement will be uncorrelated and could therefore be corrected for to some extent.

In the following experiments, training data was contaminated with simulated Additive White Gaussian Noise (AWGN) at various signal-to-noise ratios (SNR), in comparison with no-noise training. Similarly, the performance was assessed with test data at various SNR levels including no-noise data. SNR was calculated as the ratio between the sinusoid amplitude at the centre of the analysis window (equals mean dB amplitude) and the noise amplitude.

The extra phase shift measurements were added into the NN input list in pairs, first adding Φ_{2+} & Φ_{2-}, then Φ_{3+} & Φ_{3-}, etc… Also, a two-hidden layer network architecture was chosen with few neurons in the first hidden layer, in order to minimise the computational expense.[2] A consequence of this is that the first hidden layer is effectively an encoding function, whilst the remaining layers perform the mapping. Therefore it can be considered that the encoding layer is dealing with noise and producing reduced redundancy data, which is then mapped as in the previous 2-input networks.

As a first step in the transition, the network architecture was changed and experiments were carried out using noise-free training data (as before), with the following results:

- Using two hidden layers (in the described *encoder layer – mapping layers* architecture) produced much better performance and used less neurons overall, where there were the additional phase shift inputs.
- Increasing the number of neurons in the encoder layer from 2 to 3 improved the performance substantially (an order of magnitude); further increases, however, yielded minimal improvement that did not justify the additional computational cost.
- Testing these architectures with noisy phase-shift data produced high and erratic error values. Tests with 80dB and 40dB SNR produced errors that were usually 10-100 times and 1000-10000 times higher than noise-free tests, respectively. This performance varied considerably between different training runs of the same network.

Next, training continued with noisy training data:

- Training data was generated with simulated AWGN at signal-to-noise ratios of 80dB, 60dB, 40dB and 20dB. The same (*delf*, *dela*) values were used as before.
- The introduction of noise, even at an SNR of 80dB caused a dramatic improvement in performance and consistency between training runs, when tested with noisy data (at SNRs of

[2] The number of multiplications required, for applying the network weights, is the total number of neuron-to-neuron connections. Separating the inputs from the mapping layer with a very small layer dramatically reduces the computational load.

80dB and 40dB), with virtually no degradation in the noise-free test case; see figure 4.

- Alternative training data sets were created using combinations of noise levels and multiple data at each (*delf*, *dela*) pair value including, in some cases, the ideal noise-free training data set. These gave no benefit at all for a considerable increase in training time and so they were abandoned.
- Figure 4 shows that increasing the training data noise improves the overall resilience to noise, but at the cost of worse performance in the low-noise and no-noise tests. The best noise level for training will depend largely on the desired response for the implementation.
- For the noise-free training, there were performance benefits in increasing the number of phase shift inputs. However, with the noisy data sets, the SNR of the training data proved more significant. In the tests conducted, six phase shift inputs appears to be the optimum trade-off between accuracy and computational load (pre- and post-training).

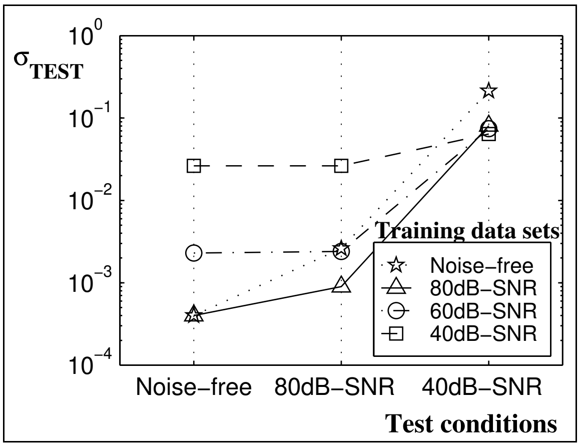

Figure 4. Error resilience for various levels of noise in training data. (Examples of typical performance)

It was initially expected that using noisy training data would improve noise resilience, because the network would be trained to give greater significance to phase shifts closer to the mainlobe maximum, where the FFT magnitude is greater (and therefore suffers less). In practice the result was as expected, but not for the reason proposed.

Up to a distance of about 6 padded bins from the mainlobe maximum, there is actually very little relative change in magnitude, particularly for the cases of greater nonstationarity, where the mainlobe is wider. Therefore, instead of the phase measure-ments closer to the maximum being of greater quality (and therefore more important to the network), all the phase inputs were roughly of equal merit. Nevertheless, the basis of the expectation appears to have been correct. Without noise, the neural network had no incentive to adjust the *relative* weights of the encoder layer to toward any particular goal, since it was able to achieve a good quality mapping with any combination of 2 or more inputs. Therefore, the relative contribution between inputs was arbitrary (leading to erratic results when tested with noisy data). Conversely, in the noisy case, every input was subject to noise that reduced its individual reliability, so for overall good performance the network tended towards relying on a roughly equal contribution between all the inputs. This conclusion is based upon observations of the network weights after training, and is to be verified with further experiments.

4 Conclusions and Future Work

This paper has established that feedforward neural networks are suitable for implementing the PDA mapping function to a high degree of accuracy. Furthermore, with the aid of noisy training data and additional phase shift inputs, a network architecture has been found that provides noise resilience, without requiring a large increase in network complexity or computational load.

There are both short-term and long-term goals for extending the research into phase distortion analysis. In the short term, additional NN estimators can be designed to provide correction to the frequency, amplitude and phase estimates, which are also degraded by nonstationarities. Also, experiments can be carried out to specifically train the neural networks to be robust to the type of additive interference caused by neighbouring sinusoids. In the longer term, the method might be extendable to second and higher order modulation.

5 References

[1] Masri, P., Bateman, A. 1995. “Identification of nonstationary audio signals using the FFT, with application to analysis-based synthesis of sound”. Proc. IEE Colloquium on Audio Engineering. pp. 11.1-6.

[2] McAulay, R.J., Quatieri, T.F. 1986. “Speech analysis/synthesis based on a sinusoidal representation”. Proc. IEEE Transactions on Acoustics, Speech and Signal Processing, Vol.34:4, pp. 744-754.

[3] Serra, X., Smith, J.O. 1989. “Spectral Modeling Synthesis”. Proc. International Computer Music Conference (ICMC), pp. 281-284.