Sinusoidal Modeling and Modification of Unvoiced Speech[1]

(paper SA-386)

Michael W. Macon, *Member, IEEE* [2]
(503) 690-4067 (voice)   (503) 690-1406 (fax)
macon@ee.ogi.edu

Mark A. Clements, *Senior Member, IEEE*
(404) 894-4584 (voice)   (404) 894-8363 (fax)
clements@ece.gatech.edu

Center for Signal and Image Processing
School of Electrical and Computer Engineering
The Georgia Institute of Technology
Atlanta, GA 30332-0250

Abstract

Although sinusoidal models have been shown to be useful for time-scale and pitch modification of voiced speech, objectionable artifacts often arise when such models are applied to unvoiced speech. This correspondence presents a sinusoidal model-based speech modification algorithm that preserves the natural character of unvoiced speech sounds after pitch and time-scale modification, eliminating commonly-encountered artifacts. This advance is accomplished via a perceptually-motivated modulation of the sinusoidal component phases that mitigates artifacts in the reconstructed signal after time-scale and pitch modification.

EDICS number - SA 1.5.3   Other enhancements (Speech modification)

[1] This work was supported in part by Sanders Inc., a Lockheed-Martin Company

[2] M. W. Macon is currently with the Oregon Graduate Institute of Science and Technology, P.O. Box 91000, Portland, OR 97291

I Introduction

Sinusoidal models have been shown to be useful for signal transformations such as pitch and time-scale modification of speech and music signals [1, 2]. The validity of such representations for modeling quasi-stationary harmonic signals (e.g., voiced speech) has been well-documented. For convenience, however, unvoiced speech is typically represented by the same model within speech modification algorithms. This approach can impart an undesirable and often-cited "tonal" character to the noise-like unvoiced signal, especially during time scale expansion and other transformations [1].

Some researchers have proposed harmonic/stochastic decompositions of the signal for speech coding [3, 4] or modification [5, 6]. Most of these are based on representing the periodic portion of the signal by a sinusoidal model and then modeling the residual signal as the output of a time-varying filter excited by white noise. Although these types of decompositions can mitigate some artifacts, it is more desirable to handle harmonic and stochastic elements of the signal within a single, unified framework.

The algorithm presented here is an extension of the *Analysis-by-Synthesis/Overlap-Add* (ABS/OLA) sinusoidal model [2, 7]. In this extension, unvoiced or noise-like segments of the signal are represented by sinusoidal components, but the phases of these sinusoids are modulated to eliminate the tonal artifact in the signal after modification and preserve its noise-like character. A perceptual motivation for the algorithm is given, followed by a frequency-domain interpretation of its resulting effect on the signal and the results of a subjective evaluation of the method.

II The ABS/OLA Sinusoidal Model

In the ABS/OLA model, the input signal $s[n]$ is represented by a sum of overlapped short-time signal frames $s_k[n]$,

$$s[n] = \sum_{k=0}^{K-1} w[n - kN_s]s_k[n] \tag{1}$$

where K is the number of synthesis frames, N_s is the frame length, $w[n]$ is a window function that is nonzero over the interval $[-N_s, N_s]$, and $s_k[n]$ represents the kth frame of the synthesized signal. Each frame $s_k[n]$ is represented as the sum of a small number of sinusoidal components, given by

$$s_k[n] = \sum_{l=0}^{L-1} A_l^k \cos(\omega_l^k n + \phi_l^k) \tag{2}$$

where L is the number of sinusoidal components in the frame, and A_l^k, ω_l^k, and ϕ_l^k are the kth frame sinusoidal amplitudes, frequencies, and phases, respectively. An iterative analysis-by-synthesis procedure is performed to find the optimal component amplitudes, frequencies, and phases for each frame, based on a mean-squared error criterion. The frequencies ω_l^k are not constrained to be harmonically related, but components are organized into a "quasiharmonic" structure with one sinusoid associated with each harmonic of the fundamental.

Overlap-add synthesis is performed by a computationally-efficient procedure that uses the inverse fast Fourier transform to compute each frame $s_k[n]$, rather than sets of oscillator functions as in [1]. Time-scale modification is performed by expanding or contracting each frame $s_k[n]$, while adjusting component frequencies and phases to preserve pitch pulse shape. Pitch modification is performed by altering the component frequencies, phases, and amplitudes such that the fundamental frequency is modified while the speech formant structure and general waveform shape characteristics are maintained [7].

III Phase Randomization

Perceptual motivation

Empirically, it has been found that the ABS/OLA model is capable of faithfully reproducing both voiced and unvoiced sounds when a frame update of 10 milliseconds or less is used in synthesis. However, when time-scale expansion and/or pitch raising operations are performed, the unvoiced segments take on the above-mentioned "tonal" character.

This role of time-scale expansion in causing this artifact can be explained in terms of current theories of pitch perception. One theory suggests that the brain assigns the perceived pitch of a tone complex based on the intervals between peaks in the fine time structure of the signal at various points on the basilar membrane, integrated over a time interval on the order of several milliseconds [8]. Thus, any arbitrary set of sinusoidal components with constant amplitude and frequency will produce regular patterns at various places across the basilar membrane, and the brain will recognize prominent periodicities in these patterns. When the sinusoidal components remain stationary for a duration significantly large with respect to the integration time of this human pitch detection mechanism, the resynthesized speech signal begins to take on a tonal character.

It has also been observed that this tonal artifact is exacerbated by pitch-raising modification. In [9], McAulay and Quatieri justify the use of the sinusoidal representation for unvoiced speech by an argument based on the Karhunen-Loève expansion for noise-like signals. They conclude that this representation for unvoiced speech is valid when the sinusoidal components are spaced "closely enough" together that the ensemble power spectral density is relatively smooth across frequency. When the fundamental frequency of the sinusoidal components is raised in a given frame, the components become more widely spaced in frequency, leaving a spectral shape that possesses more distinct spectral lines, as shown in Figure V. Thus, the model for the noise becomes less mathematically representative of the signal characteristics. This effect tends to worsen the perceived tonal noise artifact.

In a previous study of vowel perception under various acoustic manipulations [10], it was found that randomizing the phases of a sinusoidal model of *voiced* sounds resulted in an aperiodic, noise-like signal. Similarly, it has been noted [11] that the aperiodicity of *unvoiced* sounds can be preserved under time-scale and pitch modification by randomly modulating the phase of the components of the sinusoidal model. The nominal frequency of each component is kept roughly the same, but the time structure of combined sets of these components along the basilar membrane no longer exhibits the periodicities originally detectable by the listener [10].

The above arguments suggest that applying a random phase modulation of the sinusoidal components can maintain perception of randomness in the modified signal by (*i*) insuring that long-term periodicities in the time waveform are disrupted over the course of the synthesis frame, and (*ii*) maintaining the smoothness of the original signal spectrum.

Overlap-add phase randomization

This phase randomization approach can be implemented within the context of an *overlap-add* model by subdividing each synthesis frame and randomizing the phase offsets between components prior to synthesis of each subframe. Referring to Equation (2), each N_s–sample frame can be divided into several smaller subframes of length N_{sub}, as shown in Figure 2. It is possible to resynthesize a signal *identical to the original* synthesis frame by

$$s_k[n] = \sum_{m=-\infty}^{\infty} w_s[n - mN_{sub}] \sum_{l=0}^{L-1} A_l \cos(\omega_l n + \phi_{l,m}) \tag{3}$$

where $w_s[n]$ is a window function that is nonzero over $[-N_{sub}, N_{sub}]$ (the frame k notation has been suppressed). Equations (2) and (3) are made equal by letting $\phi_{l,m} = \phi_l^k$ for all m, where ϕ_l^k is the original phase estimate for the frame in Equation (2).

Alternatively, the phase offsets between sinusoidal components in each subframe can be *varied* by adding a random offset to each phase term:

$$\phi_{l,m} = \phi_l^k + V_l\, \psi_{l,m} \tag{4}$$

where $\psi_{l,m}$ is a uniform random variable over some subinterval of $[-\pi, \pi]$ and $V_l \in [0, 1]$. Thus, when $V_l = 0$ for all l, the frame $s_k[n]$ will be resynthesized in its original form, but when $V_l = 1$, the phase offsets will be completely random from subframe to subframe. This suggests the possibility of using a "soft-decision" weighting of $V_l \in [0,1]$ to produce varying degrees of phase randomization.

Although the previous equations have been presented as time-domain summations of cosines, $s_k[n]$ can be computed much more efficiently using a sequence of N_s/N_{sub} IFFT's and an overlap-add procedure analogous to that used in the original model [2]. In practice, the number of subframes is usually made proportional to the time-scale expansion factor.

Frequency-domain interpretation

Interpreting the above algorithm in the frequency domain provides several interesting insights into the behavior of the algorithm. Specifically, the effect on each component can be described as a modulation that increases the effective bandwidth of each component, smoothing the signal spectrum.

Rewriting the subframe overlap-add equation (3) in terms of complex signals and substituting Equation (4) produces

$$s_k[n] = \Re e \left\{ \sum_{m=-\infty}^{\infty} w_s[n - mN_{sub}] \sum_{l=0}^{L-1} A_l e^{j(\omega_l n + \phi_l + V_l \psi_{l,m})} \right\}. \tag{5}$$

This equation can be rewritten to incorporate a function $b_l[n]$ that modulates the lth sinusoidal signal component,

$$s_k[n] = \Re e \left\{ \sum_{l=0}^{L-1} b_l[n] A_l e^{j(\omega_l n + \phi_l)} \right\} \tag{6}$$

where

$$b_l[n] = \sum_{m=-\infty}^{\infty} w_s[n - mN_{sub}] e^{jV_l \psi_{l,m}}.$$

The function $b_l[n]$ has the Fourier transform

$$B_l(e^{j\omega}) = W_s(e^{j\omega}) \sum_{m=-\infty}^{\infty} e^{-j(m\omega N_{sub} - V_l \psi_{l,m})}. \tag{7}$$

where $W_s(e^{j\omega})$ is the Fourier transform of the subframe synthesis window.

If V_l is set to 0 for all l, then the summation in Equation (7) will produce a pulse train whose pulses coincide with the nulls of $W_s(e^{j\omega})$ for $\omega \neq 0$, resulting in $B_l(e^{j\omega}) = \delta(\omega)$. In contrast, if $V_l = 1$ for all l, it can be shown that the summation will result (on average) in a flat spectrum across all frequencies, and $B_l(e^{j\omega})$ will, on average, assume the shape of the window transform $W_s(e^{j\omega})$. Thus, as V_l is gradually varied between 0 and 1, the basis function $B_l(e^{j\omega})$ will transition from a spectral line to a stochastic function with the frequency-domain shape of $W_s(e^{j\omega})$, as shown in Figure 3. The maximum bandwidth of $B_l(e^{j\omega})$ can be varied by varying the subframe length N_{sub}, since this will alter the mainlobe width of the window transform.

The increase in bandwidth of each sinusoidal component results in a smoothing of the resynthesized signal spectrum. This is demonstrated in the bottom and middle panels of Figure V, where the modified signal spectrum is shown with and without the phase randomization algorithm applied, respectively.

The use of a modulating function such as $b_l[n]$ to preserve randomness in the sinusoidal representation is reminiscent of ideas in [4], where "narrowband basis functions" were used to represent unvoiced speech in a speech coding application. In contrast, here a straightforward extension of the overlap-add synthesis procedure provides for a computationally efficient synthesis of these modulated components, avoiding the explicit generation and filtering of long random sequences, as in [4].

Voicing measure

As mentioned above, the model lends itself well to a mapping of the amount of aperiodicity in the input signal to the parameter V_l in Equation (4). This parameter can also be varied across *frequency* in the synthesis of signals that contain both a voiced and unvoiced component. Several approaches to estimating the "degree of voicing" are mentioned in the sinusoidal modeling and speech coding literature. In [12], for example, the signal-to-noise ratio between a set of harmonic components and the original speech spectrum is mapped to the degree of voicing, with the implication that a harmonic model will fit the spectrum better in voiced speech. A similar notion is used in a frequency-dependent voicing decision in [3]. The synthesis method developed in this paper can be coupled with any of these analysis methods to implement frequency-dependent voicing decisions and to provide smooth transition from voiced to unvoiced states.

IV Subjective comparison

To confirm the the appropriateness of the phase randomization approach, a subjective comparison test was conducted using 25 volunteer subjects. Of these 25 subjects, two were experienced in subjective speech quality assessments, and 23 were naïve listeners. The subjects were asked to compare pairs of utterances presented via headphones, where each pair consisted of one utterance synthesized using the phase randomization algorithm and one synthesized using ABS/OLA without this extension. The order of the sentence pairs and of the elements within each pair was selected randomly for each subject. Subjects were instructed to select utterance "A" or "B" according to preference "in terms of overall sound quality," and were allowed to replay the stimuli as many times as desired.

The speech material used as input to the algorithm consisted of eight short phrases selected to represent an equal number of male and female voices and to contain several unvoiced phonemes. The voicing analysis method used was similar to that suggested in [12], in which all frequencies above a cutoff are declared "unvoiced," and this cutoff frequency is varied according to voicing characteristics. Four test conditions were applied to each of the eight sentences. Time-scale modifications by factors of 2.0, 3.0, and 4.0 (slower speech) were applied with no pitch modification, and time-scale modification by a factor of 3.0 was also applied in combination with a pitch modification by a factor of 1.5 (higher pitch).

The results of the four test conditions described were as follows:

test	modification factors	% preferring phase rand
A	$\beta = 1.0,\ \rho = 2.0$	81.0
B	$\beta = 1.0,\ \rho = 3.0$	79.0
C	$\beta = 1.0,\ \rho = 4.0$	73.5
D	$\beta = 1.5,\ \rho = 3.0$	72.5

The factors β and ρ correspond to pitch modification and time-scale modification factors, respectively. Each value given represents a percentage of responses preferring the phase randomization method over the standard modification method, averaged over the eight utterances and 25 subjects. Based on this number of trials, the test results show a preference for the phase randomization method that is statistically significant ($p < 0.001$) in all cases.

Although it should be expected that the algorithm would provide greater improvement of speech quality in more drastic modifications, this was not observed in the response percentages for Tests B, C, and D. One explanation of this effect is as follows: the subjects were instructed only to compare "overall sound quality" and not any specific aspect of the speech signals. Since most of the subjects participating in the test were not experienced in critical listening tests, they tended to judge *both* exemplars as more "unnatural" than unmodified speech for drastic modifications of time scale or pitch. Because of this, the response percentages tended to gravitate slightly towards a result more consistent with guessing rather than definite preference of one or another method. This hypothesis was confirmed by interviews with subjects after the experiment. It is also interesting to note that the two subjects who had previous critical listening experience chose the phase randomization method in 100% of the tested cases.

V Summary

In this correspondence, an extension to the ABS/OLA sinusoidal speech modeling and modification algorithm has been presented, along with a perceptual motivation for this algorithm and an analysis of its frequency-domain effects on the signal. This refined model enables the application of sinusoidal time-scale and pitch modification algorithms to unvoiced and noise-like signal segments as well as voiced speech and eliminates the problem of unnatural "tonal" artifacts that often arise in modification of unvoiced speech.

References

[1] T. F. Quatieri and R. J. McAulay, "Shape invariant time-scale and pitch modification of speech," *IEEE Transactions on Signal Processing*, vol. 40, pp. 497–510, March 1992.

[2] E. B. George and M. J. T. Smith, "An analysis-by-synthesis approach to sinusoidal modeling applied to the analysis and synthesis of musical tones," *Journal of the Audio Engineering Society*, vol. 40, pp. 497–516, June 1992.

[3] D. W. Griffin and J. S. Lim, "Multiband excitation vocoder," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, pp. 1223–1235, August 1988.

[4] J. S. Marques and A. J. Abrantes, "Hybrid harmonic coding of speech at low bit-rates," *Speech Communication*, vol. 14, pp. 231–247, June 1994.

[5] J. Laroche, Y. Stylianou, and E. Moulines, "HNS: Speech modification based on a harmonic + noise model," *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pp. II–550–553, April 1993.

[6] X. Serra and J. S. Smith, "Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition," *Computer Music Journal*, vol. 14, no. 4, pp. 12–23, 1990.

[7] E. B. George and M. J. T. Smith, "Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model," to appear in *IEEE Transactions on Speech and Audio Processing*, 1997.

[8] B. C. J. Moore, *An Introduction to the Psychology of Hearing.* Academic Press Limited, 3rd ed., 1989.

[9] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-34, pp. 744–754, August 1986.

[10] R. Carlson, B. Granström, and D. Klatt, "Vowel perception: the relative perceptual salience of selected acoustic manipulations," Tech. Rep. QPSR 3–4, Speech Transmission Laboratory, KTH, Stockholm, 1979.

[11] T. F. Quatieri and R. J. McAulay, "Phase coherence in speech reconstruction for enhancement and coding applications," *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pp. 207–210, April 1989.

[12] R. J. McAulay and T. F. Quatieri, "Low-rate speech coding based on the sinusoidal model," in *Advances in Speech Signal Processing* (S. Furui and M. M. Sondhi, eds.), New York: M. Dekker, 1992.

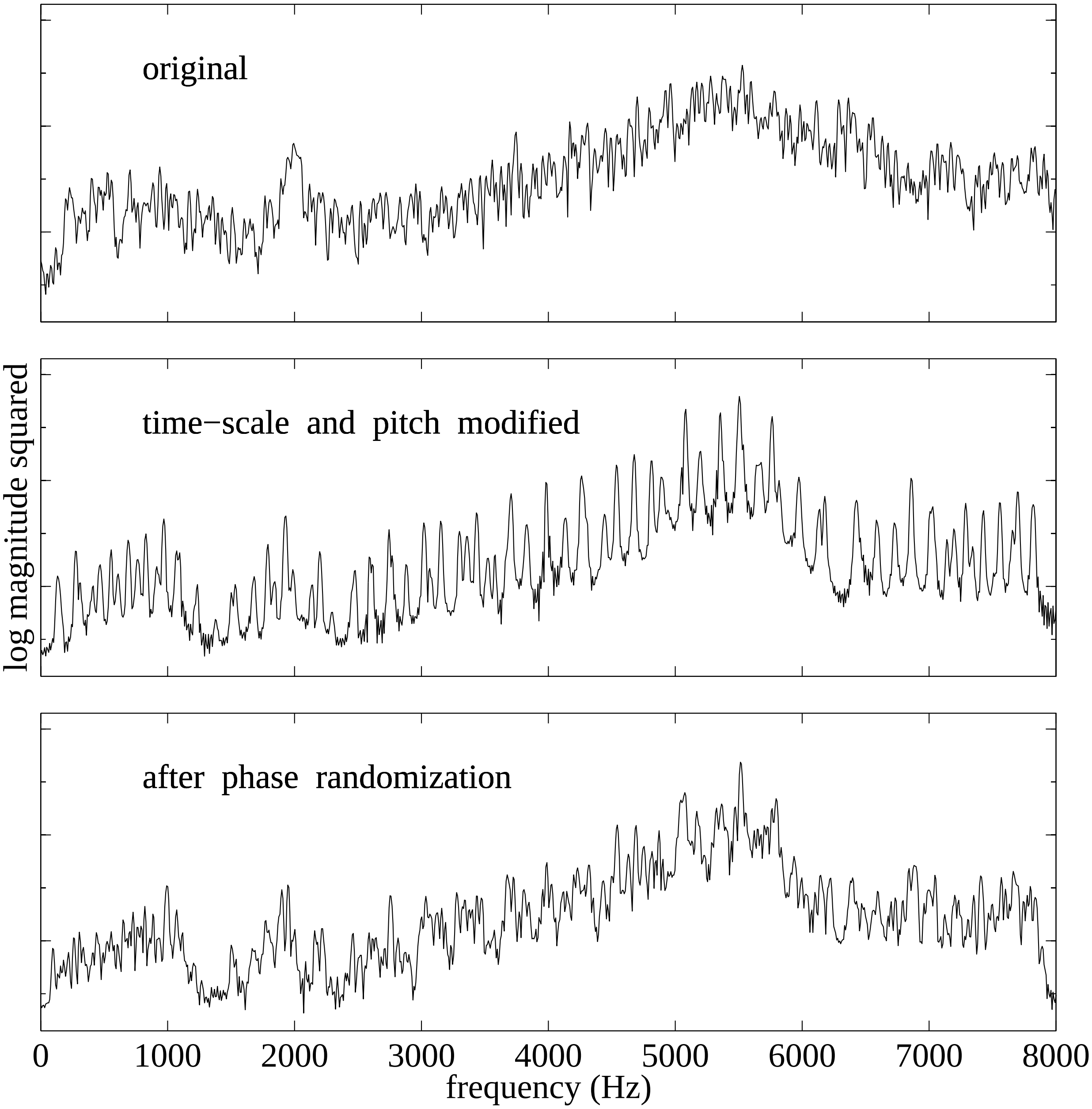
original
time−scale and pitch modified
after phase randomization
log magnitude squared
0
1000
2000
3000
4000
5000
6000
7000
8000
frequency (Hz)

Figure 1: Periodogram (50 ms rectangular window) of 80 ms unvoiced speech segment. *(top)* original signal; *(middle)* signal after time-scale expansion by a factor of 4 and pitch shift by a factor of 2; *(bottom)* resulting signal after modification using phase randomization algorithm. (frame length before modification = 10 ms, $N_{sub} = 5$

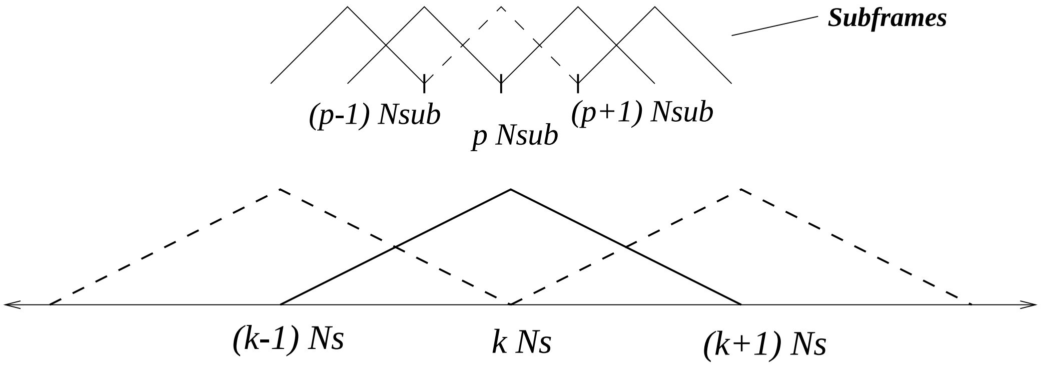
Subframes
(p-1) Nsub
p Nsub
(p+1) Nsub
(k-1) Ns
k Ns
(k+1) Ns

Figure 2: Subframe overlap-add synthesis.

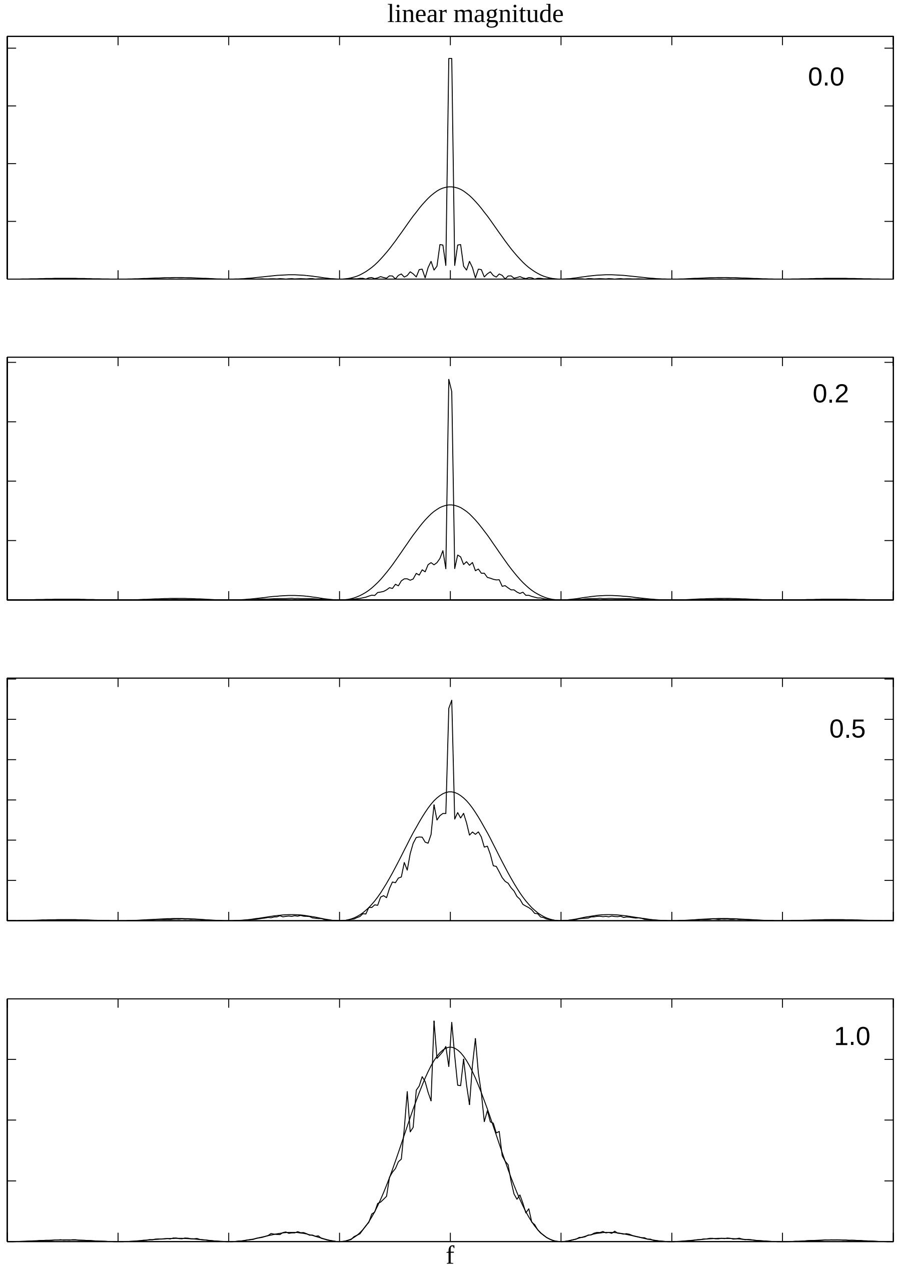
linear magnitude
0.0
0.2
0.5
1.0
f

Figure 3: Illustration of effect of phase randomization on frequency domain sinusoidal basis functions for $V_l = \{0.0, 0.2, 0.5, 1.0\}$ (from top to bottom) in Equation (4), averaged over 30 trials. Note that the bandwidth of the basis functions approaches the bandwidth of the window transform $W_s(e^{j\omega})$ as V_l approaches 1.0.