The Second International Conference "Problems of Cybernetics and Informatics"
September 10-12, 2008, Baku, Azerbaijan. Section #3 "Modeling and Identification"
www.pci2008.science.az/3/43.pdf

EVALUATION OF DERIVATIVE TIME-DELAY MODELING FOR ROBUST PITCH DETECTION IN VERY HIGH AND NONSTATIONARY NOISE

Irina Gorodnitsky

University of California, San Diego, USA, *igorodni@ucsd.edu*

Analysis of acoustic signals in noisy and nonstationary environments remains very challenging. Over the last decade, interest in development of new, physically motivated methods to address this problem has grown. In particular, it has been suggested that the physical phenomena involved in speech production lend themselves naturally to nonlinear modeling approaches. The most successful approach in this domain appears to be nonlinear prediction, of which nonlinear oscillators (NOs) [1] is a special case. The basic premise of this 'classical NO' approach is to estimate adaptively a function (model) *f*(.) that predicts a future data point from the time-delay embedded state-space vector **x** [1]:

$$x(n+1) = f(\mathbf{x}(n)).$$

The prediction function f(.) can be estimated using a number of methods, including function approximations using a truncated Volterra series with quadratic filters, lookup tables, and trainable mapping functions, e.g. a radial basis functions (RBFs)-based neural network. Among these, the most successful prediction (synthesis) behavior has been reported with RBF networks [1]. However, even the 'state-of-the-art' implementation of this method described in [1], designed to optimize NO stability, has shown poor performance for overall speech. Stable re-synthesis was reported for only 18% of the vowels (pure extended vowels, not vowels extracted from natural speech) [1]. To enable re-synthesis of the more difficult vowels, which were found to have 'more complicated' structure in state-space, e.g. /a/, /e/, and /i/, a low-pass filter had to be employed to remove structure in these phonemes that was due to the influence of the vocal tract on a glottis source signal. Unfortunately, this step can also remove important speaker-related characteristics. With it, a 56% success rate for stable re-synthesis of vowels was obtained [1], which is still cannot be considered as sufficient for practical use. Besides the apparent issue with the fixed-point stability, two other problems are encountered with the classical NO approach [1]. The first is the large parameterization space needed for successful prediction. Models involving 200 or more parameters are common. The second problem is related to the first, namely the lack of any relationship between the model parameters and the physical properties of speech. This lack of physical connection undermines the motivation that precipitated development of physically plausible models for speech in the first place.

A different nonlinear method has been proposed in the context of pitch detection in [2]. This method does not attempt to model data but solely to identify repetitive cycles in data by utilizing a variant of the nearest neighbors procedure. While a number of claims were made in [2], including superior robustness, they were not substantiated. Considerable doubts remain about the method's efficiency, robustness, and how well the presumption of constant pitch for an entire phoneme would work for speech overall, beyond the one vowel demonstrated in [2].

The current paper presents a novel nonlinear oscillator method for signal analysis whose properties have the potential to remedy many of the drawbacks listed above. The method models the dynamics (first derivative) of the data rather than data themselves with the parameter space being spanned by time-delayed vectors of data. To clarify, the scalar two-dimensional linear model being investigated in this paper, for example, has the form

$$\dot{\mathbf{x}} = \alpha_1 \mathbf{x}(t-\tau_1) + \alpha_2 \mathbf{x}(t-\tau_2), \quad (1)$$

where $\dot{\mathbf{x}}$ indicates a derivative of a current data vector of some length and $\mathbf{x}_{\tau_i} = \mathbf{x}(t-\tau_i)$ indicate time-delayed vectors of data $\mathbf{x}$ where τ_i are the parameters we want to estimate. Several unrelated sources support this approach. Inclusion of time-differentiated spectral features (derived from cepstral coefficients) has been shown to improve noise robustness in speech and speaker recognition [3]. More generally, studies of time-delay differential models

found them to be highly robust to noise [4]. Additionally, even the scalar time-delay differential models in [4] containing as few as one time delay were observed to provide a unique low dimensional projection of the data dynamics with no restriction placed on the dimensionality of the dynamical system that could be described. A third key observation in [4] was that such models revealed nonlocal correlations in data.

A formal result showing the dimensionality reducing property of this representation was derived recently in [5]. The result in [5] showed that output of up to an infinite-dimensional system of uncoupled linear oscillators (damped and undamped) can be described by one dimensional scalar linear delay differential models. This property opens up the possibility of modeling the so called 'complicated structures', including the vowels studied in [1] and unvoiced speech. Furthermore, the superior robustness property of this representation, observed in [4,5], raises the possibility that it could be practical for analysis of very noisy data.

Motivated by these prospects, the current paper presents an exploratory study of the derivative-based NO in the context of speech processing in noise. Being an early investigation, the paper considers only one aspect of this approach - the robust estimation of pitch. It is shown here that the simple 2-delay linear model in (1) does improve robustness of pitch estimation in very high, nonstationary noise conditions. In the subsequent presentation, the time-delay derivative model will be referred to as here the Interval Domain (ID) model or ID representation.

Evaluation Framework

It was shown in [5] that in the case of a one-delay ID model $\dot{\mathbf{x}} = \alpha\, \mathbf{x}(t-\tau)$, there is a direct relationship between the smallest optimal delay and the fundamental frequency (F0) of a periodic time-series:

$$\tau = \frac{f_s}{4F0},$$

where f_s denotes the sampling rate. The delay τ in the one-delay model can be solved for in the Least Squares (LS) sense and the frequency then calculated directly as the reciprocal of τ. However, this ID model was found to provide somewhat inferior performance in noise compared to the two-delay model ID in (1) which is studied here. The relationship between F0 and the parameters of model (1) can be shown to be [6]

$$F0 = f_s/(\tau_2 - \tau_1 + 2). \qquad (2)$$

Pitch is closely related to the F0 of the voiced part of speech. Hence, we can use Eq. (2) to assess the feasibility of estimating pitch from the parameters of the NO model (1). A total of 42 utterances from the DARPA TIMIT database of sentences [7] were evaluated and the pitch estimates using Eq. (2) were compared to those obtained with the widely used ESPS [8] pitch determination algorithm. Noisy speech was generated by adding scaled noise samples to the noise free signal using three levels of signal to noise ratio (SNR) ranging from –10 dB to 0 dB. The comparison was made for three distinct types of noise: 1) white, stationary noise; 2) babble noise, which is a colored noise with a spectrum similar that of the speech signal; and 3) noise measured inside a moving M109 tank, which is deterministic low-frequency noise with power that dominates the power of speech in the critical F0 range. The last two noise types were selected in order to present challenging test environments. White Gaussian noise was generated using the MATLAB 'randn' command. The babble and M109 tank noise samples were downloaded from the Signal Processing Information Base (SPIB) [9]. Babble noise is background speech, in this sample generated by 100 people speaking simultaneously. As such, it is nonstationary and its spectrum overlaps almost completely with that of our signal of interest. In many automatic recognition systems babble is considered the hardest noise to deal with. The M109 noise, on the other hand, has power concentrated in the frequencies below 200Hz, where it dominates the power of the speech signal. Since F0 often takes values in the range below 200Hz and since the time-delay equation models the deterministic structure of data, the tank noise is expected to present a worst-case scenario for NO models.

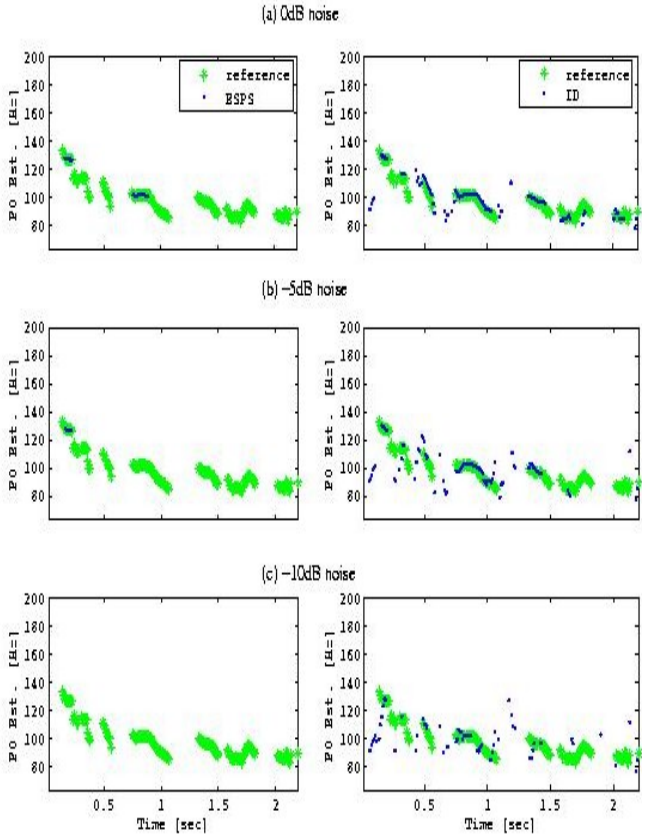

Fig. 1. ID and ESPS pitch estimates in white background noise.

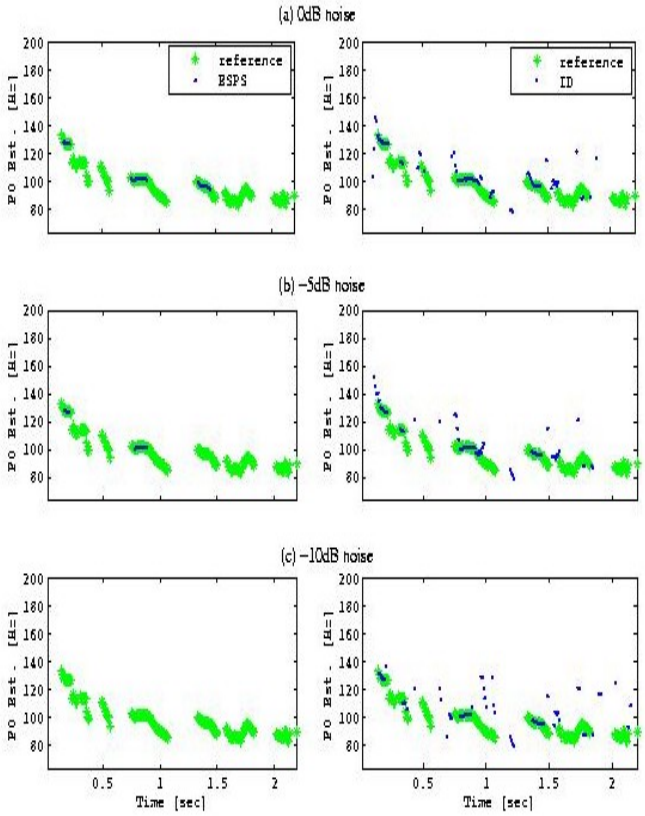

Fig. 2. ID and ESPS pitch estimates in babble noise.

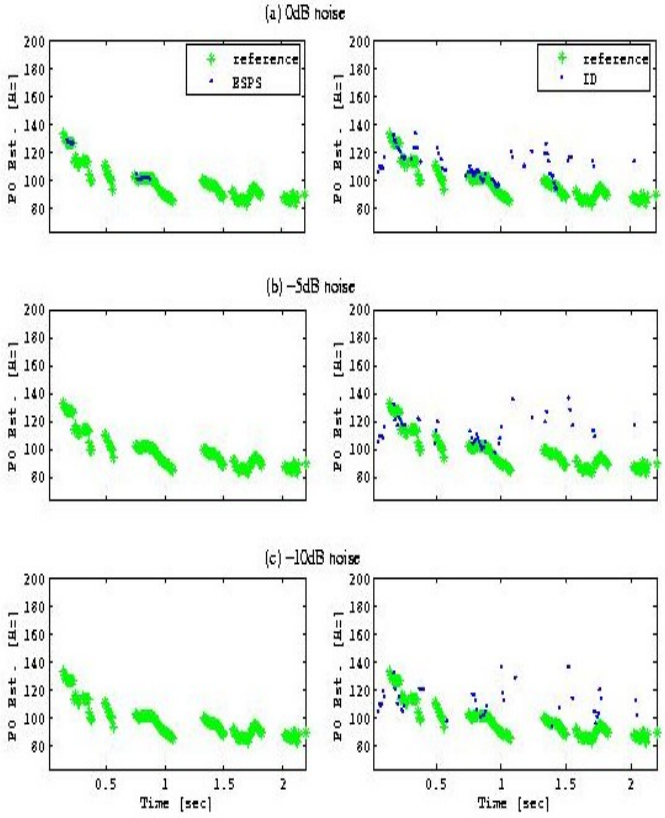

Fig. 3. ID and ESPS pitch estimates in M109 tank background noise.

All computations, with the exception of ESPS, were done in MATLAB. ESPS estimates were obtained from the Tcl/Tk SNACK library audio analysis module for Linux platforms. Speech segments (30 msec frames with 10 msec frame shifts (the framelength in SNACK)) were modeled using Eq. 1 and the ESPS method. The parameters of model (1) were estimated using the FOCUSS algorithm [9] for sparse optimization, which was extended to use multiple measurement vectors. The 16000Hz sampling rate of the original recordings was used throughout. Model (1) was run using a 50-400Hz limit on pitch while a 60-400Hz limit was used for ESPS. Setting the lower bound to 60Hz instead of 50Hz for ESPS was necessary to avoid occurrence of pitch halving errors in a part of the utterance.

The results were consistent across the three tested speakers, one female, two male, and speaker gender appeared to be irrelevant to the observed performance. In the interest of compactness, only one sentence SX42 ('Biblical scholars argue history'), spoken by male speaker MTAB0, is chosen for the presentation. Figures 1-3 show pitch estimated for the segments that are deemed voiced by the ID and ESPS methods for each of the noise conditions. Both method make a voiced/unvoiced decision for the individual analysis frames based on the continuity of the obtained F0 estimates. The pitch is set to zero for the frames that are deemed unvoiced. The difference in robustness of the two methods is evident for all noise types even at 0dB, since ESPS's ability to make voiced/unvoiced decisions is significantly degraded while the ID estimates are only marginally affected, particularly in the case of babble and white noise. The performance is somewhat worse in the case of M109 noise, but the pitch estimates could still be used for applications such as voice activity detection.

The drop in performance between 0dB and -5dB SNR is gradual for the ID method. On the other hand, at –5 dB SNR, ESPS in unable to detect any voiced speech except for one short segment in the babble noise example. Since babble is composed of multiple waveforms that individually are weak relative to the signal of interest, it appears to present the most benign case of the three noise types for both methods. At -10dB we observe degradation in ID performance for all noise cases. However, short continuous contours coincident with pitch are still identified by ID in a couple of the sections in the cases of white and babble noise and could be exploited for voiced activity detection even at the –10dB SNR levels.

References

1. G. Kubin, C.Lainscsek, E. Rank. Identification of nonlinear oscillator models for speech analysis and synthesis. Nonlinear Speech Modeling and Apps. 3445, (2005), pp. 74–113.
2. D.E. Terez. Robust pitch determination using nonlinear state-space embedding. In: Proc. IEEE Int. Conf. Acoustics, Speech, and Sig. Processing. Vol. I.(2002) pp. 345–48.
3. C. Yang, F. Soong, T. Lee. Static and dynamic spectral features: Their noise robustness and optimal weights for asr. IEEE Trans. Audio, Sp. and Lan. Proc. 15(3), (2007) pp. 1087–97.
4. M. J. Bunner, T. Meyer, A. Kittel, and J. Parisi, Recovery of the time-evolution equation of time-delay systems from time series. Phys. Rev. E 56, (1997), pp. 5083–9.
5. I. Gorodnitsky, Dynamical theory formalism for robust modeling of damped, undamped, and nonlinear oscillatory signals. In: Proc. IEEE Int. Conf. on Acoustic, Speech and Sig. Processing. Vol. III. (2007), pp. 725–28.
6. I. Gorodnitsky, A nonlinear oscillator with physically meaningful parameters for robust pitch detection in noise. Accepted to J. Accoust. Soc. Am., to appear in (2009).
7. J.S. Garofolo, L.F. Lamel, W. M. Fisher, J.G. Fiscus, D.S., Pallett, N.L. Dahlgren,. Darpa timit acoustic-phonetic continuous speech corpus. NTIS \# PB91-100354. (1993).
8. B.G. Secrest, G.R. Doddington. An integrated pitch tracking algorithm for speech systems. In: Proc. IEEE Int. Conf. Acoustics, Speech, and Sig. Proc., (1983) pp. 1352–55.
9. SPIB, 1995. Noise data: http://spib.rice.edu/spib/select_noise.html. Last accessed 6/12/08.
10. I. Gorodnitsky, B. Rao. Sparse signal reconstruction from limited data using focuss: A recursive weighted minimum norm algorithm. IEEE Trans. Sig. Process: 45 (3), (1997), pp. 600–16.