PERFORMANCE OF NONLINEAR SPEECH ENHANCEMENT USING PHASE SPACE RECONSTRUCTION

Michael T. Johnson , Andrew C. Lindgren, Richard J. Povinelli, Xiaolong Yuan

Department of Electrical and Computer Engineering
Marquette University, Milwaukee, WI
{mike.johnson, andrew.lindgren, richard.povinelli, xiaolong.yuan}@mu.edu

ABSTRACT

This paper presents the implementation of two nonlinear noise reduction methods applied to speech enhancement. The methods are based on embedding the noisy signal in a high-dimensional reconstructed phase space and applying singular value decomposition to project the signal into a lower dimension. The advantages of these nonlinear methods include that they do not require explicit models of noise spectra and do not have the typical "musical tone" side effects associated with traditional linear speech enhancement methods. The proposed nonlinear methods are compared with traditional speech enhancement techniques, including spectral subtraction, Wiener filtering, and Ephraim-Malah filtering, on example speech utterances with additive white noise for a variety of SNR levels. The results show that the local nonlinear noise reduction method outperforms Wiener filtering and spectral subtraction but not Ephraim-Malah filtering, as had been suggested by previous studies.

1. INTRODUCTION

Speech enhancement methods endeavor to separate and remove contaminating noise from the speech signal of interest. Noise reduction techniques are crucial for human intelligibility and speech technologies such as speech recognition and speaker identification [1]. The conventional techniques used in speech enhancement typically rely upon models of the spectral characteristics of both noise and speech in order to perform the separation and filtering.

As an alternative to these traditional techniques and to conventional frequency domain speech processing theory, interest has emerged into studying speech as a nonlinear, dynamical system [2, 3]. Nonlinear time series methods perform analysis and processing in a reconstructed phase space, a time-domain vector space whose dimensions are time-lagged versions of the original time series [4]. The reconstructed phase space is therefore simply a plot of the time-lagged signal vectors, a parametric graph of the time series in which geometric structures of the underlying signal, called attractors or trajectories, appear. Reconstructed phase spaces have been shown to be topologically equivalent to the original system, if the embedding dimension is large enough [5]. This implies that the full dynamics of the system are accessible in this space, and for that reason, a phase space reconstruction potentially contains more information than a spectral representation [4, 6].

A noise free signal has a well-defined attractor structure that evolves and unfolds in a finite dimension. Truly random noise, however, is time independent and therefore spreads out into an infinite dimensional phase space without structure. An example of a reconstructed phase space for a typical vowel is shown in Figure 1.

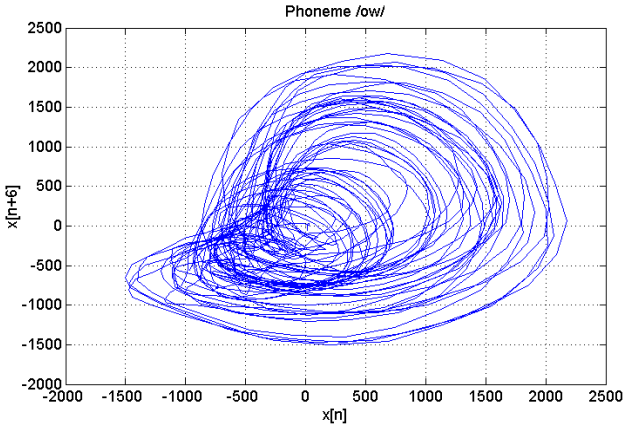

Figure 1: Phase space plot of phoneme /ow/ (τ =6, m = 2)

Several different nonlinear noise reduction techniques exist that utilize a phase space for signal and noise separation [7]. These methods have been successfully applied to known deterministic chaotic systems as well as to experimental time series data [6-10]. The two approaches that are used in this paper are global and local projection methods that have previously been demonstrated on speech, with results superior to Ephraim Malah filtering over a small set of isolated phonemes [9, 10]. We extend this to comparison against several traditional methods and enhancement of full sentences. The techniques are particularly advantageous because they do not require estimation of either noise or speech spectra, which is a requirement of their linear counterparts, and do not have the "musical tone" side effects that typically arise in frequency domain methods as a result of spectral estimation errors.

This material is based on work supported by the National Science Foundation under Grant No. IIS-0113508 and the Department of Education GAANN Fellowship.

2. DESCRIPTION OF METHODS

The dynamical systems techniques presented here embed the noisy signal in a very high-dimensional phase space, where the underlying signal will have a characteristic attractor structure but the additive noise will be distributed throughout the space. The data is then projected to a lower dimension, the final embedding dimension, where the true dynamics of the system reside [6].

2.1. Parameters

The time-lag used in a reconstructed phase space is typically guided by empirical analysis of key measures such as mutual information and autocorrelation [4, 6]. Too small of a time lag compresses the attractor, and too large of a time lag causes it expand out. The first zero of the autocorrelation function and the first minimum of the automutual information curve both give indication of which time lags may be desirable. Since the true embedding dimension of the time series is also unknown, it must also be determined empirically. The false nearest neighbors (FNN) algorithm is used to establish the dimension [6]. Algorithms for both of these tasks are available in [11]. Histograms of these metrics across a sample of speech phonemes are shown in Figure 1.

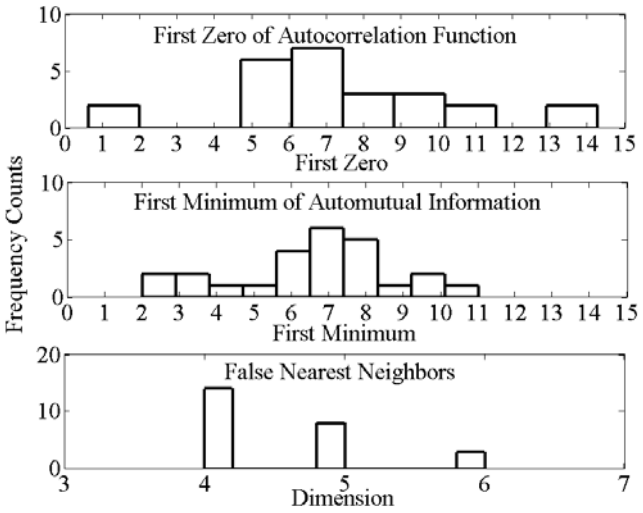

Figure 2: Autocorrelation, mutual information, and FNN

Based on these plots, a time delay of six and a final embedding dimension of five were selected for these experiments. The original embedding dimension before projection was chosen as ten.

2.2. Global nonlinear noise reduction (GNNR)

The global nonlinear projection scheme [12] is based on a decomposition of the high-dimensional data matrix of the embedded signal asserts. The projection is then accomplished using the dominant components of that decomposition. There is a relationship between this model and enhancement algorithms based on spectral sub-space decomposition [13], in the sense that for $\tau = 1$ a reconstructed phase space is a standard data matrix and the decomposition and transformation matrices are identical to those of sub-space methods.

The observed noisy time series x is zero-meaned and embedded in the phase space by creating vectors in $\mathbb{R}^m$.

$$\vec{x} = [x_i, x_{i-\tau}, \ldots, x_{i-(m-1)\tau}] \tag{1}$$

where m is the embedding dimension, chosen significantly larger than the dimension of the attractor, τ is the time lag, and i is the time index. These vectors are compiled into a trajectory matrix,

$$\mathbf{X} = \begin{bmatrix} x_0 & x_1 & \cdots & x_{n-(m-1)\tau} \\ x_\tau & x_{1+\tau} & \cdots & x_{n-(m-1)\tau+\tau} \\ \vdots & \vdots & \ddots & \vdots \\ x_{(m-1)\tau} & x_{1+(m-1)\tau} & \cdots & x_n \end{bmatrix} \tag{2}$$

and a singular value decomposition (SVD) is performed using

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T \tag{3}$$

Decomposing the trajectory matrix through an SVD can be thought of as approximating the attractor by a hyper-ellipsoid [6]. The singular values of **S** represent the magnitude of the principle axes, and the right singular vectors **V** represent the directions of these axes. If the attractor is noise free, then some or even most of the singular values will be near zero.

The signal and the noise is separated and filtered by putting the largest singular values from **S** into a new diagonal matrix $\mathbf{S}_1$ and projecting the data matrix via

$$\bar{\mathbf{X}} = \mathbf{U}\mathbf{S}_1\mathbf{V}^T \tag{4}$$

The number of singular values to be used is chosen appropriately to filter the subspace that contains the noise from the subspace that contains the signal of interest, which is the true dimension of the attractor. $\bar{\mathbf{X}}$ is the filtered version of the original trajectory matrix.

To apply this method to speech enhancement, the speech utterance is first divided into frames of equal length, with fifty percent overlap between subsequent frames to reduce edge effects. The signal in each frame is then embedded in a phase space, and its trajectory matrix compiled (τ, the time lag as well as the embedding dimension, m, are input parameters to the algorithm). A SVD is performed on the trajectory matrix as described above, and an enhanced trajectory matrix is created. The rows of the trajectory matrix are time aligned and averaged to generate a one-dimensional time series. The final enhanced signal is generated using a straightforward overlap-and-add technique.

2.3. Local nonlinear noise reduction (LNNR)

The local projection method [6-10] employs a similar strategy to that of the global method, except that the procedure is done locally in neighborhood regions along the attractor. Each region is analyzed using an SVD and projected as with the global method. Formulas (2)-(4) remain unchanged, except that they are applied to each point individually, using a smaller data matrix. This method has the significant advantage that it tracks the attractor within the phase space, capturing the underlying dynamical behavior of the system, and performs enhancement by modeling that dynamical behavior within each region.

For application to speech enhancement, the speech utterance is again divided into frames with a fifty percent overlap and embedded in a phase space. The neighborhood regions in the phase space are found (the number of points in a neighborhood is an input parameter), after which the vectors in each neighborhood are zero-meaned and the local SVD computed. The data points are projected onto the dominant directions using equation (4) above. As before, the rows of the trajectory matrix are time aligned and averaged to generate a one-dimensional time series, and the enhanced signal is generated using an overlap-and-add technique. In the local method, the algorithm is iterated several times for convergence. The core algorithm to perform the noise reduction is available in [11]. Neighborhoods of fifteen data points were used, and the algorithm was iterated ten times.

2.4. Comparative baseline methods

2.4.1 Spectral subtraction

Spectral subtraction [1] is a classic technique used in speech enhancement, and is the algorithm most often used for comparison purposes. The method estimates the noise power spectra from silent frames in the signal, and then subtracts the noise spectrum from the individual speech frame spectra. Reconstruction is done through a simple inverse DFT, using the enhanced magnitude spectrum and the original phase components. The implementation used here incorporates spectral flooring and uses fifty percent overlapping frames multiplied by a Hanning window to reduce edge effects.

2.4.2 Wiener filtering

Iterative Wiener filtering [1] constructs an optimal linear filter using estimates of both the underlying speech and underlying noise spectra. The noise spectrum is estimated from silence frames as in spectral subtraction, while the speech spectrum in each frame is estimated iteratively, beginning with the noisy signal spectrum and using the Wiener filter output to get an improved estimate. The version used here is unconstrained iterative Wiener filtering with all-pole modeling. Ten iterations were performed on each frame for convergence. Reconstruction is again done through the overlap-add technique.

2.4.3. Ephraim-Malah filtering

Ephraim-Malah filtering [14] is based on a maximum likelihood short time spectral amplitude estimator, modeling speech and noise spectral components as statistically independent Gaussian random variables. Although more complex to implement than spectral subtraction or Wiener filtering, it has several theoretical advantages, including intrinsically varying the degree of enhancement as a function of signal-to-noise ratio.

3. RESULTS

Ten example sentences were taken from the TIMIT data set [15], contaminated with additive white Gaussian noise, and enhanced using the methods described above.

The signal-to-noise ratio of the contaminated speech was varied from –10 to +10 dB.

3.1. Performance criteria

The performance measures used were signal-to-noise ratio

$$SNR = 10 \log_{10} \left(\frac{\sum S_{orig}^2}{\sum \left(S_{orig} - S_{enh} \right)^2} \right) \tag{5}$$

and the segmental signal-to-noise ratio

$$SSNR = \frac{10}{M} \sum \log_{10} \left(\frac{\sum S_{orig}^2}{\sum \left(S_{orig} - S_{enh} \right)^2} \right) \tag{6}$$

The classic SNR is not as meaningful as SSNR perceptually, because it is sensitive to outliers and fluctuations, but it is a commonly presented metric for enhancement and thus included. Thresholds of –10 and 35 dB were placed on the SSNR per frame, to match human perception more accurately [1].

3.3. Results on additive white Gaussian noise

Resulting SNR and SSNR numbers for additive white Gaussian noise, averaged across the ten example sentences, are shown in Figures 3 and 4. The amount of enhancement is more at lower initial SNRs, as is typical of nearly all speech enhancement methods.

The LNNR method outperforms the GNNR method, indicating that the localization of the projection regions is an important part of using the phase space representation for enhancement. Compared to the baseline methods, the LNNR method outperforms traditional spectral subtraction and Wiener filtering, but is not as effective as Ephraim-Malah filtering.

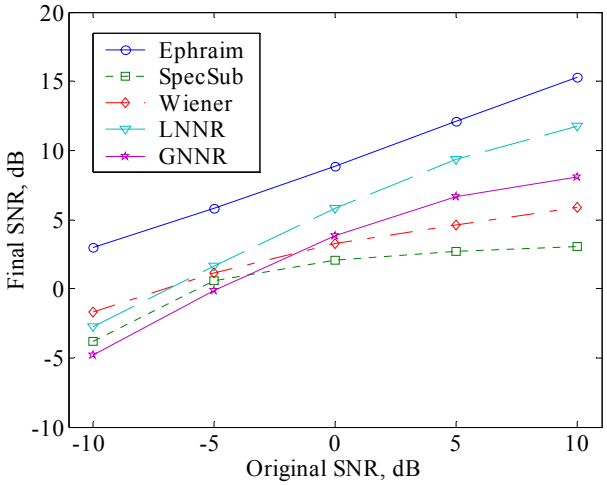

Figure 3: SNR results for additive white Gaussian noise

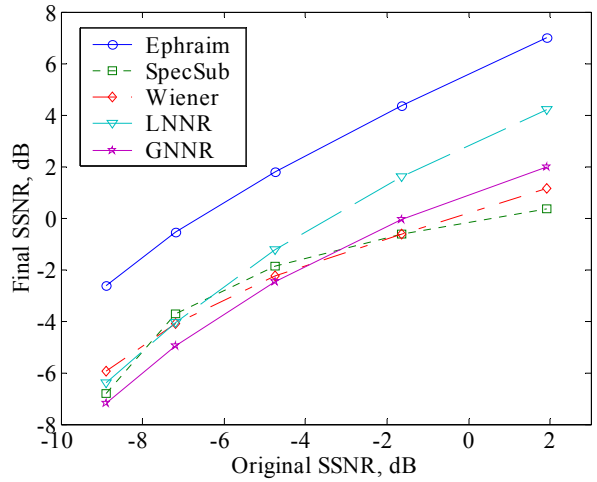

Figure 4: SSNR results for additive white Gaussian noise

4. CONCLUSIONS AND DISCUSSION

The nonlinear noise reduction schemes, and in particular the localized projection method, are clearly effective for enhancement. These approaches have the advantage of not requiring explicit spectral models for speech and noise, but also have the equivalent limitation that they cannot therefore take advantage of the information about the noise characteristics that is present during silence intervals. Since the errors produced are not due to spectral misestimation, common side effects such as the presence of extraneous tone patterns do not occur.

One note of interest is that the results do not support the claims of some previous work [9, 10] that the LNNR method is superior to Ephraim-Malah filtering. This may be due to the fact that the comparisons in that work were carried out on isolated extended-length phonemes. Despite this finding, the method shows good performance superior to spectral subtraction and Wiener filtering, and given the potential benefits of nonlinear models further work seems warranted. Future research will focus on developing methods for incorporating explicit noise and signal models into the LNNR method, and performing more thorough evaluations of the methods, including perceptual studies.

5. REFERENCES

[1] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*, vol. IEEE Press, Second ed. New York, 2000.

[2] M. Banbrook, S. McLaughlin, and I. Mann, "Speech characterization and synthesis by nonlinear methods," *IEEE Transactions on Speech and Audio Processing*, vol. 7, pp. 1 -17, 1999.

[3] A. Kumar and S. K. Mullick, "Nonlinear Dynamical Analysis of Speech," *Journal of the Acoustical Society of America*, vol. 100, pp. 615-629, 1996.

[4] H. D. I. Abarbanel, *Analysis of observed chaotic data*. New York: Springer, 1996.

[5] T. Sauer, J. A. Yorke, and M. Casdagli, "Embedology," *Journal of Statistical Physics*, vol. 65, pp. 579-616, 1991.

[6] H. Kantz and T. Schreiber, *Nonlinear time series analysis*. Cambridge: Cambridge University Press, 1997.

[7] E. Kostelich and T. Schreiber, "Noise Reduction in Chaotic Time Series: A Survey of Common Methods," *Physical Review E*, vol. 48, pp. 1752-1763, 1993.

[8] H. Kantz, T. Schreiber, I. Hoffmann, T. Burug, G. Pfister, L. Flepp, J. Simonet, R. Badii, and E. Brun, "Nonlinear Noise Reduction: A Case Study on Experimental Data," *Physical Review E*, vol. 48, 1993.

[9] R. Hegger, H. Kantz, and L. Matassini, "Noise reduction for human speech signals by local projections in embedding spaces," *IEEE Transactions on Circuits and Systems*, vol. 48, pp. 1454-1461, 2001.

[10] R. Hegger, H. Kantz, and L. Matassini, "Denoising human speech signals using chaoslike features," *Physical Review Letters*, vol. 84, pp. 3197-3200, 2000.

[11] R. Hegger, H. Kantz, and T. Schreiber, "Practical implementation of nonlinear time series methods: The TISEAN package," *Chaos*, vol. 9, 1999.

[12] D. S. Broomhead and G. King, "Extracting Qualitative Dynamics from Experimental Data," *Physica D*, pp. 217-236, 1986.

[13] Y. Ephraim and H. L. V. Trees, "A Signal Subspace Approach for Speech Enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 251-266, 1995.

[14] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Log-spectral Amplitude Estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-33, pp. 443-445, 1985.

[15] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "TIMIT Acoustic-Phonetic Continuous Speech Corpus," : Linguistic Data Consortium, 1993.