ANNUAL REPORT FOR AWARD # 0414450

Joseph Picone ; *Mississippi State Univ*
Nonlinear Statistical Modeling of Speech

Participant Individuals:
Senior personnel(s) : Georgios Y Lazarou
Graduate student(s) : Sanjay Patil; Sridhar Raghavan; Saurabh Prasad; Madhulika Pannuri; Sundararajan Srinivasan
Undergraduate student(s) : Daniel May; Ryan Irwin; Wesley Holland

Participants' Detail

Partner Organizations:
Department of Defense: Collaborative Research

DoD is a partial sponsor of this work. I am also currently working on a sabbatical with the group that sponsored the work. We have a close working relationship with them and transfer software from MS State to DoD on an as-needed basis. We try to address DoD needs that arise that are relevent to the project. Two examples of such work are confidence measures and lattice rescoring. We delivered some Perl tools to DoD for these computations, and also released these as part of our public domain software.

Activities and findings:

Findings:

Convergence of Kalman filtering, particle filtering, and other such iterative algorithms is very sensitive to prior knowledge. This is typically why these techniques don't work well for robust speech processing applications.

Our initial SVM-based speaker recognition system provides a small improvement over the GMM baseline, a finding that is consistent with what others in the community have found.

Training and Development:

Several of our students have improved their technical writing skills through publications and documentation related to this project.

All students are learning a strict software engineering process we use in our lab, and have increased their knowledge considerably.

All graduate students have received special training through a

graduate level course in Natural Language Processing that we were able to offer in conjunction with this project.

Outreach Activities:

We have hosted an open house for a local high school that specializes in mathematics and sciences: The Mississippi School for Mathematics and Science. It is one of the best math and science schools in the country and attracts the best students from all over Mississippi.

Journal Publications:

Book(s) of other one-time publications(s):

J. Picone, "Risk Minimization Approaches in Speech Recognition" , bibl. G. Camps-Valls, J.L. Rojo-µlvarez, and M. Mart¡nez-Ram¢n, "Kernel Methods in bioengineering, communications, and signal processing," The Idea Group, Inc., Hershey Pennsylvania, USA, (2006). *Book* Submitted of Collection: Idea Group Inc., USA, "Kernel Methods in bioengineering, communications, and signal processing"

Other Specific Products:

Teaching aids

A tutorial on particle filtering.

This work is disseminated via a URL:

http://www.cavs.msstate.edu/hse/ies/whats_new/2005_06/

Teaching aids

We have developed a number of tutorials on key core technologies for this project.

These tutorials are available at:

http://www.cavs.msstate.edu/hse/ies/projects/nsf_nonlinear/doc/

Software (or netware)

We have augmented our pattern recognition applet to include three time series analysis techniques: linear prediction, Kalman filtering, and Particle filtering.

The applet is available at:

http://www.cavs.msstate.edu/hse/ies/projects/speech/software/demonstrations/applets/util/pattern_recognition/current/index.html

Software (or netware)

We have developed baseline implementations of several techniques for

estimating nonlinearities in a signal. These are provided in MATLAB and used as reference implementations for our C++ code.

This software is located at:

http://www.cavs.msstate.edu/hse/ies/projects/nsf_nonlinear/downloads/software/matlab/

Internet Dissemination:

http://www.cavs.msstate.edu/project/nsf_nonlinear

We always maintain a web site for every project we execute. This web site includes software, data, publications, etc., related to the project.

Contributions:

Contributions within Discipline:

We have replicated previously published work on particle filtering, Kalman filtering, and Lyapunov exponent estimation. We have applied these techniques to more comprehensive speech databases as a first step in characterizing their performance on a large-scale application.

Contributions to Other Disciplines:

We have made software available as part of our public domain system, and also released a number of tutorials on our web site.

Contributions to Education and Human Resources:

We have introduced two undergraduate students to speech research. Both are showing great promise and plan to pursue graduate studies.

A third undergraduate will enter graduate school in Spring'2006 and continue working on this project.

Contributions to Resources for Research and Education:

A project web site has been developed and maintained.

Special Requirements for Annual Project Report:

Categories for which nothing is reported:
Participants: Other Collaborators
Research and Education Activities
Products: Journal Publications
Contributions Beyond Science and Engineering
Special Reporting Requirements
Animal, Human Subjects, Biohazards

Submit Return

View Activities PDF File

We welcome comments on this system

09/30/04 — 09/30/05: RESEARCH AND EDUCATIONAL ACTIVITIES

The overall goal of this IIS project was to create a new class of speaker recognition and verification systems based on nonlinear statistical models. Compelling evidence has been presented in recent years demonstrating that nonlinear statistical models can approach the performance of context-dependent phonetic models with almost a two order of magnitude reduction in complexity. Nonlinear dynamics provide a framework in which we can develop parsimonious statistical models that overcome many of the limitations of current hidden Markov model based techniques. We specifically propose the following:

- extend the traditional supervised-learning HMM paradigm to support a chaotic acoustic model that incorporates a **nonlinear statistical model of observation vectors;**
- evaluate the impact of this model on **text independent speaker verification** applications;
- outline extensions of these nonlinear statistical models to the **acoustic and language modeling** components of the conversational speech recognition problem.

Major accomplishments in the first year of this project include a book chapter contribution that provides an overview of the impact kernel theory and other nonlinear techniques have had on all aspects of the speech recognition problem, augmentation of our Java-based pattern recognition applet with a new time series analysis approach based on particle filtering, and release of a wide variety of core technologies to estimate nonlinear parameters of a time series. These accomplishments are described in more detail in the following sections.

A. Nonlinear Time Series Estimation

Particle filtering is a sequential Monte-Carlo simulation for nonlinear and non-Gaussian state space modeling [1],[2]. It is an attempt to integrate the Bayesian models and state-space representations that is based on te principles of a Parzen windowing approach to estimating statistics from data. Particle filtering has enjoyed some success in speech enhancement applications. Particle filtering is also closely related to Kalman filtering.

In the first year of this project, we implemented three forms of filtering: Particle filtering, Kalman filtering, and the unscented Kalman filtering. Within the generic framework of sequential importance sampling (SIS) algorithm a specific variant of particle filtering algorithm – sampling and importance Resampling (SIR) was implemented [3],[4]. The speech state space model is assumed to be Gaussian and the state space parameters are estimated using the modified Yule-Walker equations. A comparative analysis of the results from Kalman filtering approach and particle filtering approach is underway.

Lyapunov exponents are an indicator of the sensitive dependence to initial conditions for a dynamical system. Lyapunov exponents indicate the rate of divergence or convergence and thus the stability of the dynamical system. Lyapunov spectrum is also an indicator of the dimensions of the dynamical system.

For a time series data, Lyapunov spectrum estimation was implemented using the singular value decomposition based time-delay embedding [5]-[7]. The method reconstructs the phase space for the unknown dimensions. Lyapunov exponents can represent the chaotic behavior of the dynamical system. Experiments were carried on the simulated Lorenz time data series to check the effect of change in embedding dimension, window size, and evolution steps. The results are in agreement with the expected results.

A speech signal [8] demonstrates a low dimensional chaotic behavior. The Lyapunov spectrum of a chaotic time series was determined by first reconstructing a pseudo-phase space from the time series and then using this phase-space to estimate the local dynamics around the attractor. Phase space

reconstruction was achieved using both the time delay embedding method and the singular value decomposition (SVD) method. Both these methods operate by constructing a multi-dimensional space from a one-dimensional time series. To extract the local dynamics at a point around the attractor, the tangent map was calculated by examining how the neighboring points evolve with time. By performing a QR-decomposition on the tangent map at each point, the local Lyapunov exponents were evaluated from the diagonal elements of the matrix. The global Lyapunov exponents were then obtained by averaging the local exponents over all the points. We are investigating the dimensionality of the feature stream associated with a speech recognition system.

B. Kalman Filtering

Kalman filtering is a popular tool in the research community for estimating non-stationary processes when it is possible to model the system dynamics by linear behavior and Gaussian statistics [11]. Previously, Kalman filters have been applied for speech enhancement applications when the corrupting process was additive white Gaussian noise [12]. We now have a working implementation of the Kalman filter and plan to use it in our feature extraction front-end for robust speech recognition. It is hoped that use of the Kalman filter as a non-stationary process estimator will enable us to recover clean features from noisy observations and will increase noise robustness of our baseline recognition system.

In a speech recognition setup, the normally assumed inter-frame independence need not be necessarily observed. A state-space representation allows us to capture inter-frame correlations by modeling dependencies between successive observed feature vectors. Hence it is anticipated that the recursive filtering provided by Kalman filters will lead to better estimates of the clean signal since the algorithm uses all past observations to give the filtered output (intuitively similar to Auto Regressive filtering). Further, since Kalman filters are capable of modeling time-varying system behavior (the general formulation of Kalman filters allows the system matrices in the state variable model to be functions of time), we feel that they should potentially be useful in filtering applications when the dynamic model of the system is changing with time.

Kalman filters are optimal linear filters when the statistics of the state-space are modeled by Gaussian pdfs. However, when the system dynamics exhibit nonlinear behavior or non-Gaussian statistics, the recursive Kalman filtering process must be modified to accommodate for the nonlinear behavior. Previously, the extended Kalman filter (EKF) has been employed [11] in extending the recursive Kalman filtering algorithm to nonlinear systems by linearizing the nonlinear state space model and then propagating Gaussian statistics through this linearized model. A significant concern with this approach is that violation of the local linearity assumption will lead to unstable filters.

As an alternative, unscented Kalman filters use properties of the unscented transform to extend the idea of recursive Kalman filtering to nonlinear systems without linearizing the state-space model. The unscented transform[13] is a method for accurately determining the statistics of a random variable which undergoes a nonlinear transformation. This is a novel technique which can capture the first two moments of data in the transformed space by a deterministically chosen set of points in the domain space. We can use this idea to propagate Gaussian statistics through a nonlinear state-space model[14] and use the standard algorithmic formulation of Kalman filters to obtain the filtered estimates. We believe that this implementation will give us flexibility in our system modeling for recursive filtering, and we need not have to restrict our pre-processing setup to a linear (or linearized) model.

C. Baseline Speaker Recognition and Verification Systems

We have regenerated our baseline results for text independent speaker verification using the NIST 2001 dataset and a new version of our public domain speech recognition software that will be released in

October 2005. We ported this techology to the Mississippi State University supercomputing cluster and developed a program that simplifies job submission and tracking on the cluster. This program has been integrated into our public domain ASR software and will be part of our upcoming release. This program simplifies running several parallel experiments hence increases the productivity by eliminating the task of creating separate scripts for every job. This is especially useful for running experiments on resource intensive tasks such as support vector machines (SVMs), relevance vector machines (RVMs) and other nonlinear modeling techniques that are currently being developed in the lab.

We also designed the experimental setup for SVM and RVM-based speaker verification system. Since we pioneered the use of SVMs for speech recognition in a previous NSF project, SVMs have become extremely popular in speaker recognition. However, the community has not yet understood and accepted our RVM technology. We have tested our implementation of SVMs using the ISIP Foundation Clases (IFCs) on a speaker verification task and observed an improvement over the baseline GMM based system that is consistent with the literature. We made changes to the SVM and RVM code base in the production system in order to build a powerful discriminative training system that could be used for speech recognition and speaker recognition. We refined a utility that will learn the relevance vectors from data [15], and are in the process of testing this new technology on the same NIST 2001 data set.

Once we have these baselines in order, we will proceed in two directions. First, we need to update the data sets to the latest NIST evaluation sets (2005). Second, we can then insert our nonlinear modeling technology because this software is already integrated into the ISIP IFCs.

We also developed a utility to compute word-posteriors from word-graphs[16]. This utility was written in Perl and can be downloaded from our website[17]. This utility is useful for annotating one-best output with word-posteriors that could be used as confidence scores and hence help in reducing word error rates. This is tool could be used on our word-graphs or HTK format word graphs – including those generated by Bolt Beranek and Newman (BBN). This utility can also perform lattice word error rate computation and confusion network generation. The ultimate goal for building such a utility is to reduce the word error rates. It can be used on lattices built from different types of acoustic models, and will be a useful tool for generating publishable baselines using our nonlinear techniques.

D. Research Experience for Undergraduates (REU)

There were two components to our REU project this year: (1) adding nonlinear statistical modeling techniques to our Java-based Pattern Recognition applet, and (2) enhancing our ability to convert and transform between different grammar formats. Both projects made significant progress over Summer'2005, and will contribute software to our planned software release in October 2005.

Our Java-based applet is an excellent mechansim for teaching the fundamentals of pattern recognition to undergraduates and entry-level graduate students. It allows users to create data sets and to process them through a variety of popular patttern recognition techniques, and to compare the results both in terms of error rate and decision surfaces. The visualization capabilities are particularly important as they help students develop intuition about how these algorithms behave on difficult data sets. The preloaded data sets available in the applet allow users to easily replicate many classic problems. Example output from the applet is shown in Figure 1. The applet now has three time series analysis techniques: linear prediction, Kalman filtering, and Particle filtering. In addition to supporting education and training, the applet has been used to understand how these algorithms behave on special data sets as we construct the baseline implementations. Our typical development cycle starts in Matlab, proceeds to the Java-based applet, and then concludes with a C++ IFC-based implementation.

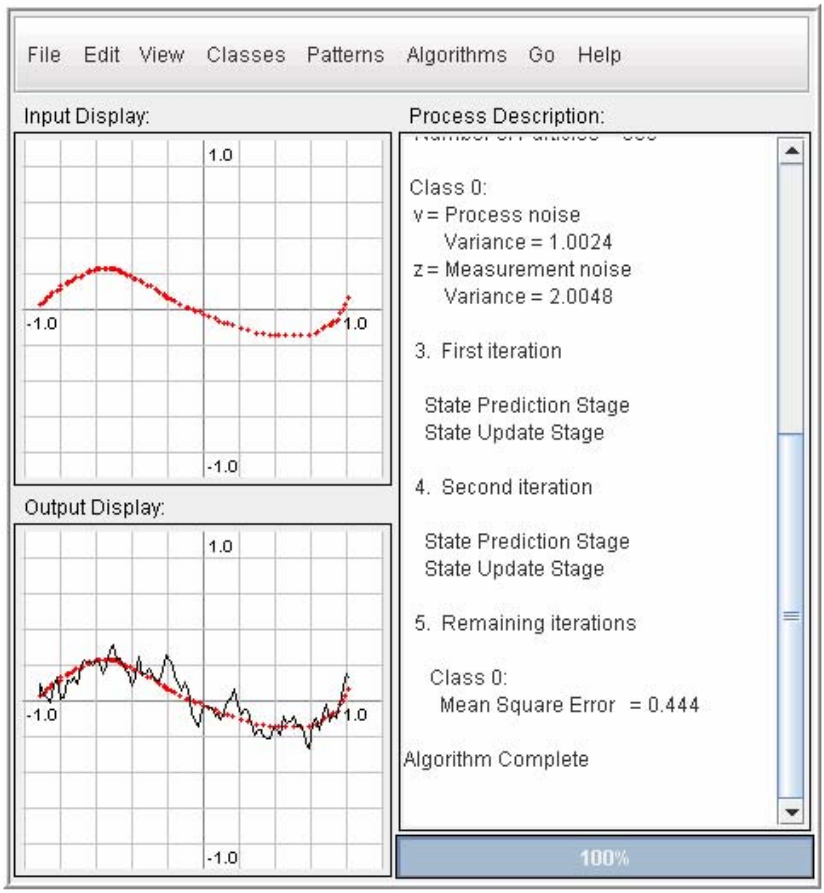

Figure 1. A screenshot of the new particle filtering component in our Java-based Pattern Recognition applet. This contribution was developed as part of our REU extension to this project.

The second component of our REU program involves grammar transformations. In the past year, there have been tremendous advances in our grammar transformation capabilities. Most prominent is the addition of XML Speech Recognition Grammar Specification (SRGS) [18] to the formats supported by our system. Such support requires the transparent conversion of a Chomsky Type-2 context-free grammar to a Type-3 regular grammar; a nontrivial task. This conversion allows our software to process XML-SRGS language models with a finite-state automaton instead of the more complex push-down automaton normally required to recognize Type-2 grammars. Also implemented was the corresponding backwards conversion from Type-3 grammar to Type-2 grammar.

This addition complements our previously implemented support for another grammar format, the JSpeech Grammar Format (JSGF). We now have the ability to convert between these two industry standards via our internal format, ISIP Hierarchical DiGraph (IHD). Theoretically, this is a conversion from a context-free grammar specification to a different context-free grammar specification through an intermediary regular grammar. Although this intermediary grammar lacks the descriptive power of a context-free grammar, our conversion process ensures preservation of grammar meaning during this state through

redundancy. To verify the equivalence of these grammar specifications, we ran a series of parallel speech training and recognition experiments in both XML-SRGS and JSGF [19]; we concluded that the differences in grammar specification did not affect the accuracy of speech recognition.

With the escalating deployment of speech applications and the widespread adoption of XML as a data encapsulation format, it is important to understand these transformations. Although theoretical investigations into these techniques have been conducted, little is known concerning the practical details of implementation; we plan to rectify this situation [20]. Also, by providing these conversion utilities in the public domain, we hope to increase comprehension of said transformations by the community and pave the way for future web applications.

E. Other Issues

The first year of this project started slowly due to a combination of unforeseen events. Several high quality Ph.D. candidates that were recruited for this project failed to obtain visas in Fall'2004 and Spring'2005. Hence, the project was understaffed through much of 2004 and early 2005. This problem was rectified in Fall'2005, as we had our best recruiting year in a long time. The project is now fully staffed with four quality Ph.D. students and one M.S. student, and is rapidly making progress.

Also, the PIs decision to take an sabbatical (IPA) with DoD, one of the sponsors of the first year of this work, also impacted our project. We have established suitable management infrastructure to minimize the impact of this transition, and the PI has been able to contribute to the project remotely.

We expect the second year of the project to result in two significant outcomes: generation of comprehensive speaker recognition and verification benchmarks and publication in first-tier journals and conferences.

F. References

[1] P.M. Djuric, J.H. Kotecha, J. Zhang, Y. Huang, T. Ghirmai, M. Bugallo, and J. Miguez, "Particle Filtering," IEEE Magazine on Signal Processing, vol 20, no 5, pp. 19-38, September 2003.

[2] N. Arulampalam, S. Maskell, N. Gordan, and T. Clapp, "Tutorial On Particle Filters For Online Nonlinear/ Non-Gaussian Bayesian Tracking," IEEE Transactions on Signal Processing, vol. 50, no. 2, pp. 174-188, February 2002.

[3] R. van der Merve, N. de Freitas, A. Doucet, and E. Wan, "The Unscented Particle Filter," Technical Report CUED/F-INFENG/TR 380, Cambridge University Engineering Department, Cambridge University, U.K., August 2000.

[4] M.W. Andrews, "Learning and Inference in Nonlinear State-Space Models," Gatsby Unit for Computational Neuroscience, University College, London, U.K., December 2004.

[5] M. Banbrook, G. Ushaw, and S. McLaughlin, "Lyapunov exponents from a time series: a noise robust extraction algorithm," submitted to IEEE transactions on Signal Processing, July 1995.

[6] M. Banbrook, S. McLaughlin, and I. Mann, "Speech Characterization and synthesis by nonlinear methods," IEEE Transactions on Speech and Audio Processing, vol. 7, no. 1, pp. 1-17, January 1999.

[7] P. Bryant and R. Brown, “Lyapunov exponents from observed time series,” Physical Review Letters, vol. 65, no. 13, pp. 1523-1526, September 1990.

[8] R. Povinelli, M. Johnson, A. Lindgren, and J. Ye, “Time Series Classification using Gaussian Mixture Models of Reconstructed Phase Spaces,” IEEE Transactions on Knowledge and Data Engineering, vol. 16, no. 6, pp. 770-783, June 2004.

[9] M. Banbrook, G. Ushaw, and S. McLaughlin, “Lyapunov exponents from a time series: a noise robust extraction algorithm,” submitted to IEEE transactions on Signal Processing, July 1995.

[10] P. Bryant and R. Brown, “Lyapunov exponents from observed time series,” Physical Review Letters, vol. 65, no. 13, pp. 1523-1526, September 1990.

[11] S. Haykin, *Adaptive Filter Theory*, Fourth edition (September 14, 2001), Prentice-Hall Engineering.

[12] M. Gabrea, “Robust Adaptive Kalman Filtering-Based Speech Enhancement Algorithm”, *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Montreal, Canada, May 17-19, 2004.

[13] S. Julier and J. Uhlmann, “A New Extension of the Kalman Filter to Nonlinear Systems,” *SPIE AeroSense Symposium*, Orlando, FL, SPIE, April 21–24, 1997.

[14] E.A. Wan and R. van der Merwe, “The Unscented Kalman Filter for Nonlinear Estimation,” *Proceedings of Symposium 2000 on Adaptive Systems for Signal Processing, Communication and Control (AS-SPCC)*, IEEE, Lake Louise, Alberta, Canada, Oct, 2000.

[15] J. Hamaker, “*Sparse Bayesian Methods for Continuous Speech Recognition"*, Ph.D. Dissertation Proposal, Department of Electrical and Computer Engineering, Mississippi State University, March 2002.

[16] S. Raghavan and J. Picone, *"Confidence Measures Using Word Posteriors and Word Graphs," http://www.cavs.msstate.edu/hse/ies/publications/seminars/msstate/2005/confidence/*, IES Spring'05 Seminar Series, Intelligent Electronic Systems, Center for Advanced Vehicular Systems, Mississippi State University, Mississippi State, Mississippi, USA, January 2005.

[17] S. Raghavan and J. Picone, “Lattice Tools,” *http://www.cavs.msstate.edu/hse/ies/projects/speech/software/legacy/lattice_tools/*, Intelligent Electronic Systems, Center for Advanced Vehicular Systems, Mississippi State University, Mississippi State, Mississippi, USA, January 2005.

[18] Voice Browser Working Group, “Speech Recognition Grammar Specification Version 1.0,” *http://www.w3.org/TR/speech-grammar/*, World Wide Web Consortium, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA, 2004

[19] Voice Browser Working Group, “JSpeech Grammar Format”, *http://www.w3.org/TR/jsgf/*, World Wide Web Consortium, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA, 2000

[20] D. May, “Automatic Speech Recognition: Monthly Tutorial”, *http://www.cavs.msstate.edu/hse/ies/projects/speech/software/tutorials/monthly/2004/2004_08/*, Institute for Signal and

Information Processing, Mississippi State University, Mississippi State, Mississippi, USA, August 2004.