NONLINEAR STATISTICAL MODELING OF SPEECH

S. Srinivasan[1], T. Ma[1], D. May[1], G. Lazarou[2] and J. Picone[1]
[1] Department of Electrical and Computer Eng., Mississippi State University, MS, USA
[2] New York City Transit Authority, New York, New York, USA
{ss754, tm334, dom5}@ece.msstate.edu, {georgios.lazarou, joseph.picone}@gmail.com

Abstract

Contemporary approaches to speech and speaker recognition decompose the problem into four components: feature extraction, acoustic modeling, language modeling and search. Statistical signal processing is an integral part of each of these components, and Bayes Rule is used to merge these components into a single optimal choice. Acoustic models typically use hidden Markov models based on Gaussian mixture models for state output probabilities. This popular approach suffers from an inherent assumption of linearity in speech signal dynamics. Language models often employ a variety of maximum entropy techniques, but can employ many of the same statistical techniques used for acoustic models.

In this paper, we focus on introducing nonlinear statistical models to the feature extraction and acoustic modeling problems as a first step towards speech and speaker recognition systems based on notions of chaos and strange attractors. Our goal in this work is to improve the generalization and robustness properties of a speech recognition sysetm. Three nonlinear invariants [1] are proposed for feature extraction: Lyapunov exponents, correlation fractal dimension, and correlation entropy. We demonstrate an 11% relative improvement on speech recorded under noise-free conditions, but show a comparable degradation occurs for mismatched training conditions on noisy speech. We conjecture that the degradation is due to difficulties in estimating invariants reliably from noisy data.

To circumvent these problems, we introduce two dynamic models to the acoustic modeling problem: (1) a linear dynamic model (LDM) [2] that uses a state space-like formulation to explicitly model the evolution of hidden states using an autoregressive process, and (2) a data-dependent mixture of autoregressive (MixAR) models [3]. Results show that LDM and MixAR models can achieve comparable performance with HMM systems while using significantly fewer parameters. The LDM model provided a reduction of the speech recognition error rate by 4.9% relative on clean data and 6.5% relative on noisy data. The MixAR model provided a reduction in the speaker recongition error rate of 6.4% relative on noisy evaluation data. Currently we are devolping Bayesian parameter estimation and discriminative training algorithms for these new models to improve noise robustness.

References:

[1] D. May, *Nonlinear Dynamic Invariants For Continuous Speech Recognition*, M.S. Thesis, Department of Elect. and Comp. Eng., Mississippi State University, May 2008.

[2] J. Frankel and S. King, "Speech Recognition Using Linear Dynamic Models," *IEEE Transactions on Speech and Audio Processing*, vol. 15, no. 1, pp. 246–256, Jan. 2007.

[3] Wong, C. S. and Li, W. K., "On a Mixture Autoregressive Model," *Journal of the Royal Statistical Society*, vol. 62, no. 1, pp. 95-115, Feb. 2000.

Keywords: nonlinear statistical models, chaos, machine learning, speech recognition