CCIT Report #478 April 2004

Speech Spectral Modeling and Enhancement Based on Autoregressive Conditional Heteroscedasticity Model

Israel Cohen

Abstract

In this paper, we introduce a novel approach for statistically modeling speech signals in the short-time Fourier transform (STFT) domain. The proposed model is based on autoregressive conditional heteroscedasticity (ARCH) modeling, which is widely-used for modeling the volatility of financial time-series such as exchange rates and stock returns. Generalized ARCH models account for excess kurtosis (*i.e.*, heavy-tailed distribution) and volatility clustering, two important characteristics of financial time-series. Speech signals in the STFT domain exhibit both “volatility clustering” and heavy tail behavior, and thus are well suited for such modeling. We define the conditional “volatility” of the STFT expansion coefficients, and propose to model the one-frame-ahead conditional variance of the expansion coefficients as a generalized ARCH process. Taking into account speech presence uncertainty, we derive recursive estimators for the variances and magnitudes of the STFT expansion coefficients. Experimental results show that the proposed model and speech enhancement algorithm yield a higher segmental signal-to-noise ratio, lower log-spectral distortion, and better Perceptual Evaluation of Speech Quality scores (PESQ, ITU-T P.862) than those obtained by using the Gaussian statistical model and the decision-directed estimation approach of Ephraim and Malah.

I. Introduction

Statistical modeling of speech signals in the short-time Fourier transform (STFT) domain has recently received much attention, but is still a puzzling problem. Ephraim and Malah [1] proposed to model the individual STFT expansion coefficients of the speech signal as zero-mean

The author is with the Department of Electrical Engineering, Technion - Israel Institute of Technology, Technion City, Haifa 32000, Israel (email: icohen@ee.technion.ac.il; tel.: +972-4-8294731; fax: +972-4-8295757).