Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 50 (2008) 476–486
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Voice activity detection based on adjustable linear prediction and GARCH models

Hiroko Kato Solvang [a,b,*], Kentaro Ishizuka [c], Masakiyo Fujimoto [c]

[a] Department of Genetics, Institute for Cancer Research, Norwegian Radium Hospital, Rikshospitalet University Hospital, Montebello, 0310 Oslo, Norway
[b] Department of Biostatistics, Institute of Basic Medical Science, University of Oslo, P.O. Box 1122, Blindern, 0317 Oslo, Norway
[c] NTT Communication Science Laboratories, NTT Corporation, 2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan

Received 9 November 2006; received in revised form 29 January 2008; accepted 14 February 2008

Abstract

We propose a method for voice activity detection (VAD) that employs a class of the Autoregressive–Generalized Autoregressive Conditional Heteroskedasticity (AR-GARCH) model. As regards correlated speech signals, we represent the AR part of the AR-GARCH model with a state-space to obtain the appropriate linear prediction error series. By applying the GARCH model to the residual, we estimate the conditional variance sequences corresponding to the voice activity parts. To detect voice activity, we establish an appropriate threshold for the conditional variance sequences. To confirm the performance of our proposed VAD method, we conduct experiments using speech signals with real background noise (signal-to-noise ratios (SNRs) of 10, 5 and 0 dB) of an airport and a street. Furthermore, using receiver operating characteristics curves and equal error rates, we compare our results with those of previous standardized VAD algorithms (ITU-T G.729B, ETSI ES 202 050, and ETSI EN 301 708) as well as recently developed methods (VAD with long-term spectral divergence, likelihood ratio tests, and higher-order statistics for VAD). In terms of the signals with background noise at an SNR of 0 dB, the experimental results show a significant performance improvement compared with standardized VAD algorithms and more than 10% improvement compared with recently developed VAD methods.

Keywords: Voice activity detection; AR-GARCH model; State-space representation; Kalman filter; Linear prediction

1. Introduction

Voice activity detection (VAD) is a technique for detecting the period within an input signal that includes conversational speech, and it plays a crucial role in speech signal processing techniques. VAD in the presence of noise is particularly important for tasks such as speech enhancement utilizing information from a period that includes only noise (Le Bouquin-Jeannès and Faucon, 1995), speech coding that allows variable bit-rates (Srinivasan and Gersho, 1993), and automatic speech recognition (Junqua et al., 1994). Since these techniques require information about the period of the speech segment in the observed signals or assume that the information is given a priori, their practical performance is greatly affected by the VAD performance. This drawback has made it necessary to develop more robust VAD in real environments (Karray and Martin, 2003).

In general, VAD consists of two parts: an 'acoustic feature extraction' part and a 'decision mechanism' part. Although both parts influence VAD performance, this paper focuses on the acoustic features.

The short-term signal energy and zero-crossing rate have been widely used as simple acoustic features for VAD (Rabiner and Sambur, 1975). Although these features are indeed effective under high signal-to-noise ratio (SNR) conditions, they are easily contaminated by environmental noise. Therefore, various kinds of robust

* Corresponding author. Address: Department of Genetics, Institute for Cancer Research, Norwegian Radium Hospital, Rikshospitalet University Hospital, Montebello, 0310 Oslo, Norway. Tel.: +47 22 93 44 18; fax: +47 22 93 44 40.
E-mail address: hsolvang@rr-research.no (H.K. Solvang).

0167-6393/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2008.02.003

acoustic features for VAD have been proposed to reflect the inherent characteristics of speech signals. These features include an auto-correlation function (Basu, 2003; Kristjansson et al., 2005), harmonicity (Shen et al., 1998), pitch (Tucker, 1992), power in the band-limited region (Marzinzik and Kollmeier, 2002; ITU-T Recommendation G.729 Annex B, 1996; ETSI ES 202 050, 2005), mel-frequency cepstral coefficients (Kristjansson et al., 2005), delta line spectral frequencies (LSF) (ITU-T Recommendation G.729 Annex B, 1996), whole spectra (ETSI ES 202 050, 2005), minimum mean squared error estimated spectra (Sohn et al., 1999), long-term spectral envelope (Ramirez et al., 2004), and nonlinear fluctuations in the speech cycle period (Ishizuka and Kato, 2006). On the other hand, VAD methods based on higher-order statistics have been proposed to utilize the statistics of speech signals (Nemer et al., 2001; Li et al., 2005). In this paper, we adopt the latter approach and focus on inhomogeneous variance over time.

Fig. 1 shows the waveforms and histograms of speech signals for one utterance. The signals were obtained by adding street noise to the speech at SNRs of 10, 5 and 0 dB. As in the figures for clean speech (without noise), the histograms mostly show leptokurtic and heavy-tailed distributions. Also, as the SNRs become lower, the noisy speech histograms begin to show a mesokurtic and medium-tailed distribution. Although the signals in Fig. 1 are for one utterance, we have confirmed the same configuration for histograms of a speech signal including 1000 utterances (total duration of about 74 min). It is important to consider the histograms since those of the sample data reveal the statistical characteristics of the distributions from which the variables are generated. Therefore, we consider that it is reasonable to assume that conventional linear prediction approaches based on stationarity and Gaussianity are appropriate for eliminating the background noise part and a non-Gaussian distribution may be appropriate for representing the distribution of target speech signals. The above is the modeling strategy that we consider for our proposed VAD method.

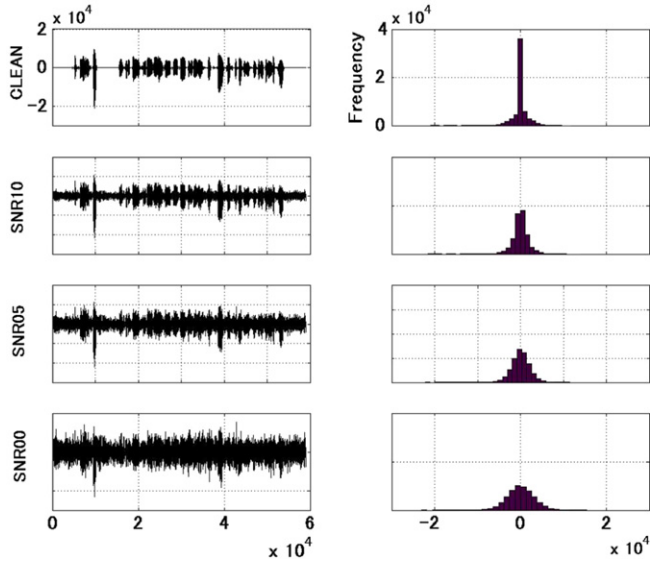

Fig. 1. Speech signals (left); vertical axis: amplitude of speech, horizontal axis: time [s]. Their histograms (right); vertical axis: frequency, horizontal axis: amplitude of speech.

In order to consider modeling the target speech signals with the background noise, we show the upper panel in Fig. 2, which is a daily return series from the finance domain. The sudden changes of the variance in the figure are called 'volatility' in the finance. Financial analysts have attempted more suitable time series modeling for estimating this volatility. In fact, speech signals with high SNR background noise resemble the daily return and we can confirm that the histogram (the lower panel in Fig. 2) of the daily return series exhibits heavier tails than a Gaussian distribution and is very similar to the histograms of clean speech signals shown in Fig. 1. For such data characterized by heavy-tailed distributions, a generalized autoregressive conditional heteroskedasticity (GARCH) model (Bollerslev, 1986) has been widely used in time series analysis. The GARCH model is a natural generalization of the ARCH (autoregressive conditional heteroskedasticity) process (Engle, 1982), and it allows lagged conditional variances to take account of the long memory process. GARCH and ARCH models are the most basic models as regards financial time series data. Recently, this model has also been used to model speech characteristics in the time–frequency domain for speech enhancement (Cohen, 2005; Abramson and Cohen, 2006) and also as feature parameters for robust automatic speech recognition (Abdolahi and Amindavar, 2005). Employing a similar approach to ours, Tahmasbi and Rezaei (2007) proposed a soft voice activity detection method using the GARCH model for pre-processing the multiple observations of a likelihood ratio test (Ramírez and Segura, 2005); however, their method applies the GARCH model to observed signals without strict consideration of the correlation to the

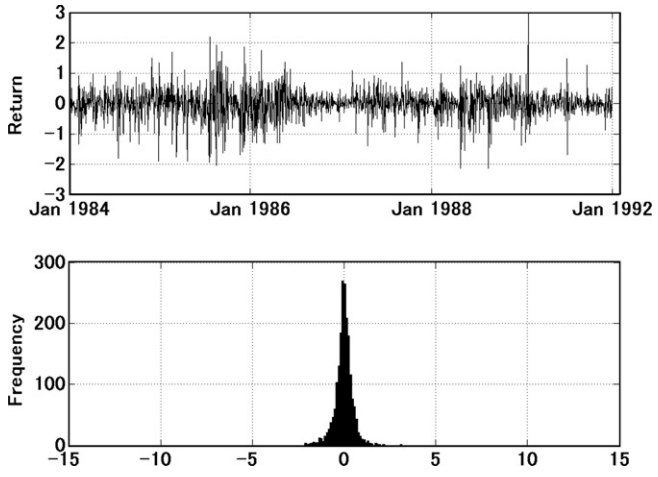

Fig. 2. Time series data of Deutschmark/British pound daily returns from January 1984 to January 1992 (upper); the vertical axis: rate of return, horizontal axis: days. Histogram of the daily returns (lower); vertical axis: frequency, horizontal axis: rate of return.