MARKOV-SWITCHING GARCH MODEL AND APPLICATION TO SPEECH ENHANCEMENT IN SUBBANDS

Ari Abramson and Israel Cohen

Department of Electrical Engineering, Technion - Israel Institute of Technology
Technion City, Haifa 32000, Israel
{aari@tx,icohen@ee}.technion.ac.il

ABSTRACT

In this paper, we introduce a Markov-switching generalized autoregressive conditional heteroscedasticity (GARCH) model in the short-time Fourier transform (STFT) domain. A GARCH model is utilized with Markov switching regimes, where the parameters are assumed to be frequency variant. The model parameters are evaluated in each frequency subband and a special state (regime) is defined for the case where speech coefficients are absent or below a threshold level. The problem of speech enhancement under speech presence uncertainty is addressed and it is shown a soft voice activity detector may be inherently incorporated within the algorithm. Experimental results demonstrate the potential of our proposed model to improve noise reduction while retaining weak components of the speech signal.

1. INTRODUCTION

Statistical modeling of speech signals in the short-time Fourier transform (STFT) domain is of much interest in many speech enhancement applications. The Gaussian model [1] enables to derive useful estimators for the speech expansion coefficients such as the minimum mean-square error (MMSE) of the short-term spectral amplitude (STSA), as well as MMSE of the log-spectral amplitude (LSA) [1, 2]. Recently, a generalized autoregressive conditional heteroscedasticity (GARCH) model has been introduced for statistically modeling speech signals in the STFT domain [3]. However, the proposed model assumes that the parameters are both time and frequency invariant and it also requires an independent detector for speech activity in the time-frequency domain. A Markov-switching time-frequency GARCH (MSTF-GARCH) model has been proposed in [4] for modeling nonstationary signals in the time-frequency domain. Accordingly, the parameters are allowed to change in time according to the state of a hidden Markov chain (*e.g.*, switching between speech phonemes), but the parameters are still frequency-invariant. The model is estimated using training signals based on maximum likelihood (ML) approach and a recursive algorithm has been derived for conditional variance estimation and signal reconstruction from noisy observations. However, not only that different phonemes may result in different GARCH parameters, speech signals are generally characterized by different both volatility and energy levels in various frequency bands. Therefore, different parameters may better represent different frequency subbands.

In this paper, we modify the MSTF-GARCH model by assuming different Markov chains in distinct subbands with identical state transition probabilities. The GARCH parameters are state dependent and frequency variant. We define an additional state for the case where speech coefficients are absent (or below a certain threshold level) and introduce parameter estimation method which is computationally more efficient than the traditional ML approach. Furthermore, the probability of the speech absence state can be used as a soft voice activity detector which is naturally generated in the reconstruction algorithm. Experimental results demonstrate improved noise reduction performance while preserving weak components of the speech signal.

Section 1 introduces the statistical model. In Section 2, we show how the model parameters can be estimated and in Section 3, we derive the speech enhancement algorithm based on the proposed model. Finally, in Section 4 we evaluate the performance of the proposed algorithm.

2. MODEL FORMULATION

Let $\{X_{tk} \,|\, t = 0, 1, ...T-1, k = 0, 1, ..., K-1\}$ denote the coefficients of a speech signal in a STFT domain, where t is the time frame index and k is the frequency-bin index. Let $\{v_{tk}\}$ be iid complex Gaussian random variables with zero-mean and unit variance, let κ_n denote the nth frequency subband with $n \in \{1, 2, ..., N\}$ and $N < K$. An $(m+1)$-state hidden Markov chain is assumed for each frequency subband, denoted by $S_t(\kappa_n)$, with a realization $s_t(\kappa_n) \in \{0, 1, ..., m\}$ and state trans-

This research was supported by the Israel Science Foundation (grant no. 1085/05).

ition probabilities which are independent of the subband index. Let $\mathcal{I}^t$ denote all available information up to time t, *i.e.*, $\{X_{\tau k} \,|\, \tau = 0, 1, ..., t,\ k = 0, 1, ..., K-1\}$ and the regimes (states) path. Given the active state $S_t(\kappa_n) = s_t(\kappa_n)$, the *one-frame-ahead conditional variance* of the spectral coefficient X_{tk}, $k \in \kappa_n$ is defined by $\lambda_{tk|t-1,s_t} \triangleq E\left\{|X_{tk}|^2 \,|\, \mathcal{I}^{t-1}, s_t\right\}$, with $s_t = s_t(\kappa_n)$. The speech spectral coefficients are assumed to follow an MSTF-GARCH process of order $(1,1)$ [4]:

$$X_{tk} = \sqrt{\lambda_{tk|t-1,s_t}}\, v_{tk}\,, \quad k \in \kappa_n \tag{1}$$

$$\lambda_{tk|t-1,s_t} = \lambda_{\min,n,s_t} + \alpha_{n,s_t} |X_{t-1,k}|^2 + \beta_{n,s_t}\left(\lambda_{t-1,k|t-2,s_{t-1}} - \lambda_{\min,n,s_{t-1}}\right)\,, \tag{2}$$

where $\lambda_{\min,n,s_t} > 0$ and $\alpha_{n,s_t}, \beta_{n,s_t} \geq 0$ are sufficient constrains for the positivity of the one-frame-ahead conditional variance, given that the initial conditions satisfy $\lambda_{0k|-1,s_0} \geq \lambda_{\min,n,s_0}$ for all $k \in \kappa_n$ and $s_0 = 0, 1, ..., m$. Note that the model formulation in [4] is slightly different. We assume that the parameters are frequency dependent while each $\lambda_{\min,n,s_t}$ defines the minimum value of the conditional variance in subband κ_n under $S_t(\kappa_n) = s_t$. Let $a_{s_{t-1},s_t} \triangleq p(S_t = s_t \,|\, S_{t-1} = s_{t-1})$, let π_s denotes the stationary probability of state s and let Ψ be an $(m+1)\times(m+1)$ matrix with elements

$$\psi_{s+1,\tilde{s}+1} = \frac{\pi_{\tilde{s}}}{\pi_s} a_{\tilde{s},s}\left(\alpha_{n,s} + \beta_{n,s}\right)\,, \quad s, \tilde{s} = 0, 1, ..., m\,. \tag{3}$$

Then, a necessary and sufficient condition for asymptotic wide-sense stationarity of the model defined in (1) and (2) is $\rho(\Psi) < 1$, where $\rho(\cdot)$ denotes spectral radius [5]. This condition is also necessary to ensure a finite second order moment for the process.

The unconditional expectation of the state-dependent one-frame-ahead conditional variance follows

$$\begin{aligned} E\left\{\lambda_{tk|t-1,s_t}\right\} = {} & \lambda_{\min,n,s_t} + \alpha_{n,s_t} E\left\{|X_{t-1,k}|^2 \,|\, s_t\right\} \\ & + \beta_{n,s_t} E\left\{\lambda_{t-1,k|t-2,S_{t-1}} \,|\, s_t\right\} \\ & - \beta_{n,s_t} E\left\{\lambda_{\min,n,S_{t-1}} \,|\, s_t\right\} \end{aligned} \tag{4}$$

with

$$\begin{aligned} E\left\{\lambda_{\min,n,S_{t-1}} \,|\, s_t\right\} &= \sum_{s_{t-1}} p(s_{t-1} \,|\, s_t)\, \lambda_{\min,n,s_{t-1}} \\ &= \sum_{s_{t-1}} \frac{\pi_{s_{t-1}}}{\pi_{s_t}} a_{s_{t-1},s_t} \lambda_{\min,n,s_{t-1}}\,. \end{aligned} \tag{5}$$

Therefore, the stationary variance of the process is given by (see [5])

$$\lim_{t\to\infty} E\left\{|X_{tk}|^2\right\} = \boldsymbol{\pi}\left(I_{m+1} - \Psi\right)^{-1} \tilde{\boldsymbol{\lambda}}_{\min,n}\,, \tag{6}$$

where $\boldsymbol{\pi}$ is a row vector of the stationary probabilities, I_{m+1} is the identity matrix of order $m+1$,

$$\tilde{\boldsymbol{\lambda}}_{\min,n} \triangleq \left[\tilde{\lambda}_{\min,n,0}, \tilde{\lambda}_{\min,n,1}, ..., \tilde{\lambda}_{\min,n,m}\right]^T \tag{7}$$

and

$$\tilde{\lambda}_{\min,n,s} \triangleq \lambda_{\min,n,s} - \frac{\beta_{n,s}}{\pi_s} \sum_{\tilde{s}} \pi_{\tilde{s}} a_{\tilde{s},s} \lambda_{\min,n,\tilde{s}}\,. \tag{8}$$

3. MODEL ESTIMATION

The estimation of a GARCH model with Markov regimes is generally obtained from a training set using ML approach [6, 7]. However, the maximization of the likelihood function is numerically unstable for multi-regime processes and only a local maxima can be generally obtained. Assuming an $(m+1)$-state Markov chain with GARCH of order $(1,1)$ in each regime, the maximization process generates $(m+1)^2$ variables for the transition probabilities and additional $3\times(m+1)$ variables for the GARCH parameters in each regime. Speech signals in the STFT domain demonstrate different levels of magnitudes in different subbands and the coefficients are generally sparse. Therefore, we limit the conditional variances in each subband within a dynamic range of η_g dB and define a special state for speech absence hypothesis. Let $\zeta_g \triangleq \max_{t,k} |X_{tk}|^2$ and $\zeta_n \triangleq \max_{t,k\in\kappa_n} |X_{tk}|^2$ denote the global maximum energy and the local maximum energy of the coefficients (in subband κ_n), respectively. Then, for the speech absence state (namely, $s_t = 0$), we set

$$\lambda_{\min,n,0} = 10^{\log_{10}\zeta_g - \eta_g/10}\,,\ \alpha_{n,0} = \beta_{n,0} = 0\,. \tag{9}$$

Under speech presence, a local dynamic range of η_ℓ dB ($\eta_\ell < \eta_g$) is assumed for the conditional variances. Furthermore, the parameters $\lambda_{\min,n,s}$, $s > 0$ are chosen to enable tracking any transients between different levels of magnitudes results in switching the active state. Without loss of generality, we sort the states according to the minimum variance level such that

$$\lambda_{\min,n,1} = \max\left\{\lambda_{\min,n,0}, 10^{\log_{10}\zeta_n - \eta_\ell/10}\right\}\,, \tag{10}$$

and for $s = 2, ..., m$, $\lambda_{\min,n,s}$ are log-spaced between $\lambda_{\min,n,1}$ and ζ_n. Each state practically represents different floor level for the spectral coefficients' variance. The parameters $\alpha_{n,s}, \beta_{n,s}$ for $s > 0$ set the volatility level of the conditional variance and they are chosen as follows. Assuming an immutable state s, the stationary variance follows

$$\lambda_{\infty,n,s} \triangleq \lim_{t\to\infty,\, k\in\kappa_n} \lambda_{tk|t-1,s} = \lambda_{\min,n,s} \frac{1-\beta_{n,s}}{1-\alpha_{n,s}-\beta_{n,s}} \tag{11}$$

provided that $\alpha_{n,s} + \beta_{n,s} < 1$. Since different states are related to different dynamic ranges in ascending order, we constrain $\lambda_{\infty,n,s} \leq \lambda_{\min,n,s+1}$ and therefore

$$\frac{1-\beta_{n,s}}{1-\alpha_{n,s}-\beta_{n,s}} \leq \frac{\lambda_{\min,n,s+1}}{\lambda_{\min,n,s}} \,. \tag{12}$$

The autoregressive parameters, $\beta_{n,s}$, are chosen experimentally while the moving average parameters, $\alpha_{n,s}$, are chosen to satisfy equality in (12). Although the clean signal is assumed to be available for the model estimation, it is only the high energy values that are needed in each subband. These values can be practically estimated from the noisy coefficients using the spectral subtraction approach. The state transition probabilities can be estimated from test signals such that each active state is determined by the energy level of the subband.

4. SPECTRAL ENHANCEMENT OF NOISY SPEECH

Let D_{tk} denote the spectral coefficients of a noise signal which is uncorrelated with the speech signal and assume that $D_{tk} \sim \mathcal{CN}\left(0, \sigma_{tk}^2\right)$. Let $Y_{tk} = X_{tk} + D_{tk}$ be the noisy observations and let $\mathcal{Y}^t \triangleq \{Y_{\tau k} \,|\, \tau = 0, 1, ..., t\,, k = 0, 1, ..., K-1\}$ denote the set of the observed coefficients up to time t. The noise variance σ_{tk}^2 is assumed to be known and it can be practically estimated using the improved minima controlled recursive averaging approach [8]. Reconstruction of the one-frame-ahead conditional variances of the speech coefficients is carried out recursively for each state by

$$\begin{aligned}\hat{\lambda}_{tk|t-1,s_t} &= \lambda_{\min,n,s_t} + \alpha_{n,s_t} E\left\{|X_{t-1,k}|^2 \,|\, \mathcal{Y}^{t-1}, s_t\right\} \\ &\quad + \beta_{n,s_t} E\left\{\lambda_{t-1,k|t-2,S_{t-1}} \,|\, \mathcal{Y}^{t-1}, s_t\right\} \\ &\quad - \beta_{n,s_t} E\left\{\lambda_{\min,n,S_{t-1}} \,|\, \mathcal{Y}^{t-1}, s_t\right\}\,, \end{aligned} \tag{13}$$

where

$$\begin{aligned} &E\left\{|X_{t-1,k}|^2 \,|\, \mathcal{Y}^{t-1}, s_t\right\} \\ &= \sum_{s_{t-1}} p\left(s_{t-1} \,|\, s_t, \mathcal{Y}^{t-1}\right) E\left\{|X_{t-1,k}|^2 \,|\, \mathcal{Y}^{t-1}, s_{t-1}\right\} \\ &\triangleq \sum_{s_{t-1}} p\left(s_{t-1} \,|\, s_t, \mathcal{Y}^{t-1}\right) \hat{\lambda}_{t-1,k|t-1,s_{t-1}}\,, \end{aligned} \tag{14}$$

$$\begin{aligned} &E\left\{\lambda_{t-1,k|t-2,S_{t-1}} \,|\, \mathcal{Y}^{t-1}, s_t\right\} \\ &\quad \simeq \sum_{s_{t-1}} p\left(s_{t-1} \,|\, s_t, \mathcal{Y}^{t-1}\right) \hat{\lambda}_{t-1,k|t-2,s_{t-1}} \end{aligned} \tag{15}$$

and

$$\begin{aligned} &E\left\{\lambda_{\min,n,S_{t-1}} \,|\, \mathcal{Y}^{t-1}, s_t\right\} \\ &\quad = \sum_{s_{t-1}} p\left(s_{t-1} \,|\, s_t, \mathcal{Y}^{t-1}\right) \lambda_{\min,n,s_{t-1}}\,. \end{aligned} \tag{16}$$

A detailed algorithm for the conditional variance restoration is described in [4].

Having an estimate for the speech coefficient's second order moment under each state, $\hat{\lambda}_{tk|t,s_t}$, estimates of the speech coefficients are obtained by minimizing the mean-square error of the log-spectral amplitude (LSA). Let

$$\hat{\xi}_{tk,s_t} \triangleq \frac{\hat{\lambda}_{tk|t,s_t}}{\sigma_{tk}^2}\,, \quad \hat{\vartheta}_{tk,s_t} \triangleq \frac{\hat{\xi}_{tk,s_t}}{1+\hat{\xi}_{tk,s_t}} \cdot \frac{|Y_{tk}|^2}{\sigma_{tk}^2}\,. \tag{17}$$

Then, the LSA estimation of the speech coefficients is given by

$$\hat{X}_{tk} = Y_{tk} \prod_{s_t} G\left(\hat{\xi}_{tk,s_t}, \hat{\vartheta}_{tk,s_t}\right)^{p\left(s_t \,|\, \mathcal{Y}^t\right)}\,, \tag{18}$$

where

$$G\left(\xi, \vartheta\right) = \frac{\xi}{1+\xi} \exp\left(\frac{1}{2}\int_{\vartheta}^{\infty} \frac{e^{-t}}{t} dt\right) \tag{19}$$

is the LSA gain function [2] and the state probabilities, $p\left(s_t \,|\, \mathcal{Y}^t\right)$, are evaluated according to [4].

5. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, we demonstrate the application of the proposed model to speech enhancement and to speech presence probability estimation.

The enhancement evaluation includes two objective quality measures; segmental SNR and log-spectral distortion (LSD). The speech signals used in our evaluation are taken from the TIMIT database. The signals are sampled in 16 kHz, degraded by a nonstationary factory noise and transformed into the STFT domain using half overlapping Hamming windows of 32 msec length. Twenty subbands are considered with global and local dynamic ranges of $\eta_g = 50$ dB and $\eta_\ell = 20$ dB, and four-state Markov chains (*i.e.*, $m = 3$) for each subband. The autoregressive parameters used in our simulations are $\beta_{n,s} = 0.8$ for all n and $s > 0$. In each subband, the state persistence probability is 0.8 and $a_{s,\tilde{s}}$ are equally chosen for all $s \neq \tilde{s}$. Figure 1 demonstrates the spectrograms and waveforms of a clean signal, noisy signal with SNR of 5 dB, and the enhanced signal obtained by the proposed algorithm. It shows that the background noise is highly attenuated while weak speech components are retained, even while noise transients occur. Furthermore, the segmental SNR and the LSD are improved. A subjective study of speech spectrograms and informal listening tests confirm that the quality of the enhanced speech is improved by using frequency-dependent parameters which are derived from the different energy levels.

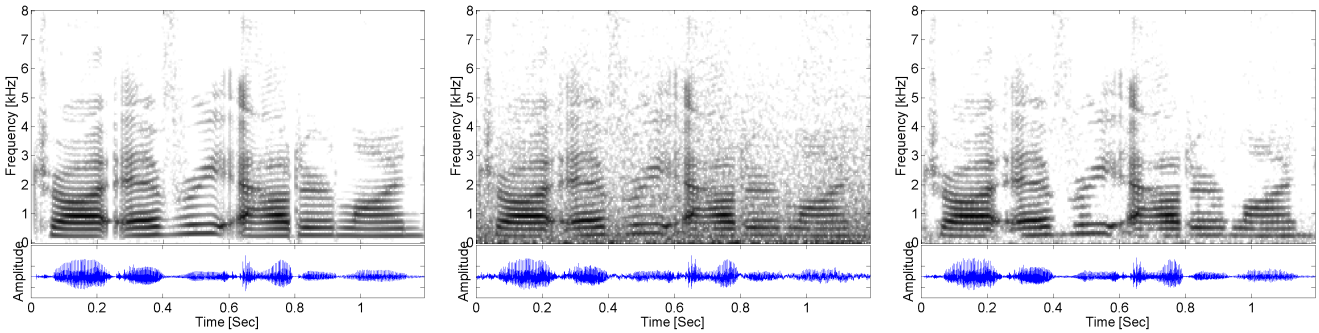

Figure 1: Speech spectograms and waveforms. (a) Clean signal: "Try any other line."; (b) speech corrupted by factory noise with 5 dB SNR (LSD= 6.68 dB, SegSNR= 0.05 dB); (c) speech reconstructed by using 4-state model (LSD= 3.14 dB, SegSNR= 6.76 dB).

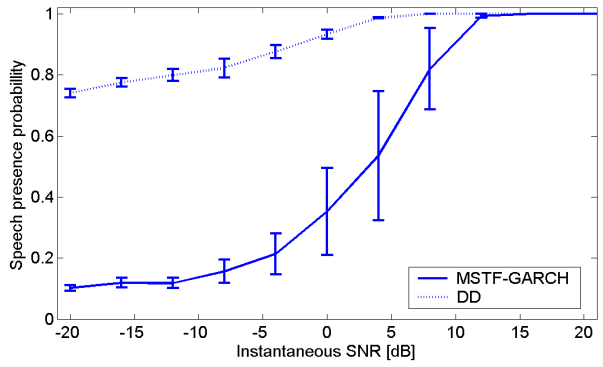

Figure 2: Conditional speech presence probability obtained by the proposed algorithm and by the decision-directed based algorithm.

The conditional speech presence probability results from the enhancement algorithm is compared with the statistical model-based voice activity detector (with *hangover*) of Sohn *et al.* [9] when applied to subbands. The later evaluates the conditional likelihood $\mathcal{L}_t \triangleq p\left(\mathcal{Y}^t \mid S_t \neq 0\right) / p\left(\mathcal{Y}^t \mid S_t = 0\right)$ by utilizing the decision-directed approach for the a priori SNR estimation (assuming only two states). The conditional speech presence probability is obtained by $p\left(S_t \neq 0 \mid \mathcal{Y}^t\right) = \mu\mathcal{L}_t / \left(1 + \mu\mathcal{L}_t\right)$, where $\mu \triangleq p\left(S_t \neq 0\right) / p\left(S_t = 0\right)$ is the *a priori* probabilities ratio. Figure 2 demonstrates the speech presence probabilities achieved when both algorithms are applied to a speech signal corrupted by a white Gaussian noise with SNR of 15 dB. The *instantaneous SNR* is defined as the ratio between the norms of the clean signal and the noise signal in each subband. It can be seen that the speech presence probability, derived from our proposed algorithm, results in a higher dynamic range for the probabilities and in much lower values for low energy coefficients. Furthermore, the probabilities ascribed to each instantaneous SNR are with higher variance resulting from the Markovian nature of the model.

6. REFERENCES

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. ASSP-32, no. 6, pp. 1109–1121, December 1984.

[2] ——, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 33, no. 2, pp. 443–445, April 1985.

[3] I. Cohen, "Speech spectral modeling and enhancement based on autoregressive conditional heteroscedasticity models," *Signal Processing*, vol. 68, no. 4, pp. 698–709, Apr. 2006.

[4] A. Abramson and I. Cohen, "Recursive supervised estimation of a Markov-switching GARCH process in the short-time Fourier transform domain," *submitted to IEEE Trans. on Signal Processing*.

[5] ——, "Asymptotic stationarity of Markov-switching time-frequency GARCH processes," in *Proc. of 30th IEEE Intern. Conf. On Acoustics, Speech, and Signal Processing, ICASSP-06.*, May 2006, pp. III 452–445.

[6] J. D. Hamilton, *Time Series Analysis*. Princeton University Press, 1994.

[7] J. D. Hamilton and R. Susmel, "Autoregressive conditional heteroskedasticity and changes in regime," *Journal of Econometrics*, vol. 64, pp. 307–333, July 1994.

[8] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Processing*, vol. 11, no. 5, pp. 466–475, September 2003.

[9] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, 1999.