SPEECH ENHANCEMENT USING ARCH MODEL

Aviva Atkins and Israel Cohen

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

ABSTRACT

In this paper, we investigate the use of the *autoregressive conditional heteroscedasticity* (ARCH) model as a replacement to the decision-directed method in the log-spectral amplitude estimator for speech enhancement. We employ three sound quality measures: speech distortion, noise reduction and musical noise, and explain the effect the ARCH model parameters have on these measures. We demonstrate and compare the use of the decision-directed and ARCH estimators and show that the ARCH model achieves better results than the decision-directed for some of these measures, while compromising between the speech distortion and noise reduction.

Index Terms— Speech Enhancement, Time-frequency analysis, Musical noise, Noise reduction, ARCH

1. INTRODUCTION

Spectral domain estimators for speech enhancement, such as the minimum mean-square error (MMSE) estimator [1], the short-time spectral amplitude (STSA) estimator [2], and the log-spectral amplitude (LSA) estimator [3], require an estimate of the a-priori signal-to-noise ratio (SNR). One of the most commonly-used approaches for this estimate is the decision-directed method [2]. Due to non-linearity of the processing methods in these algorithms, artificial noise distortion may be introduced at the output, causing what is known as musical noise. Cappé [4] has shown that using the decision-directed estimator it is possible to reduce this artifact, though at the expense of some distortion to the estimated speech signal and a higher level of residual background noise, which masks the musical noise.

Recently, it was proposed in [5] to use *generalized ARCH* (GARCH) for statistically modeling the speech signals in the time-frequency domain. The GARCH model is extensively used in financial applications where it is necessary to model time varying volatility while taking into account the time-series heavy tailed behavior and volatility clustering. As the short-time Fourier transform (STFT) expansion coefficients exhibit such heavy-tailed behavior and "volatility clustering" (in the sense that large magnitudes tend to follow large magnitudes and small magnitudes tend to follow small magnitudes, while the phase is uncorrelated), the GARCH is a reasonable model.

In this paper, we investigate the use of a simplified case of the GARCH model, the ARCH model, as part of a spectral domain noise reduction algorithm. We define three measures representing different components of the sound quality: speech distortion, noise reduction, and musical noise, and we explain the effect that the ARCH model parameters have on these measures. We present a similar evaluation of the decision-directed estimator, with a comparison to the ARCH estimator.

The paper is organized as follows. In Section 2, we review the signal estimation problem, along with the decision-directed and ARCH estimators. In Section 3, we present the performance measures employed in our evaluation of the algorithms. Experimental results and discussion are presented in Section 4. Finally, Section 5 concludes the paper and offers future research directions.

2. SIGNAL ESTIMATION

We consider an observed speech signal $y(n) = x(n) + d(n)$ composed of a clean speech signal $x(n)$ which is corrupted by uncorrelated additive noise $d(n)$, where n is a discrete time index. In the time-frequency domain, applying the STFT, the observed signal is:

$$Y_\ell(k) = X_\ell(k) + D_\ell(k) \tag{1}$$

where k is the frequency bin index $(k = 0, 1, \ldots, N-1)$, and ℓ is the time frame index $(\ell = 0, 1, \ldots, L)$. $X_\ell(k)$ and $D_\ell(k)$ are the respective STFT of $x(n)$ and $d(n)$. Our objective is to find an estimate $\hat{X}_\ell(k)$ for the STFT of the clean speech signal.

Given an error function between the STFT of the clean signal and its estimate $e\left(X_\ell(k), \hat{X}_\ell(k)\right)$, the estimated sig-

This Research was supported by Qualcomm Research Fund and MAFAAT - Israel Ministry of Defense.

978-1-5090-2007-2/16/$31.00 ©2016 IEEE

nal is derived from:

$$\hat{X}_\ell(k) = \arg\min_{\hat{X}} E\left\{ e\left(X_\ell(k), \hat{X}(k)\right) |Y_0(k), \ldots, Y_{\ell'}(k) \right\}. \tag{2}$$

We shall consider the causal case in which $\ell' \leq \ell$, and the LSA error function

$$e_{\mathrm{LSA}}\left(X_\ell(k), \hat{X}_\ell(k)\right) = \left(\log|X_\ell(k)| - \log\left|\hat{X}_\ell(k)\right|\right)^2. \tag{3}$$

Using the statistical signal model in [6], an estimate $\hat{X}_\ell(k)$ can be found independently for each frequency bin index k. Hence, for simplicity, the frequency bin index k will be omitted from this point forward. Also, similarly to [6], we assume knowledge of the noise probability density distribution, which in practice can be estimated using the *Minima Controlled Recursive Averaging* method [7]. The estimate for X_ℓ is obtained by applying a spectral gain function to each noisy spectral component of the observed signal:

$$\hat{X}_\ell = G\left(\xi_{\ell|\ell'}, \gamma_\ell\right) \cdot Y_\ell, \tag{4}$$

where the *a-priori* and *a-posteriori* SNRs are defined, respectively, by:

$$\xi_{\ell|\ell'} \triangleq \frac{\lambda_{\ell|\ell'}}{\sigma_\ell^2}, \quad \gamma_\ell \triangleq \frac{|Y_\ell|^2}{\sigma_\ell^2}. \tag{5}$$

$\sigma_\ell^2 \triangleq E\left\{|D_\ell|^2\right\}$ denotes the short-term spectrum of the noise, and $\lambda_{\ell|\ell'} \triangleq E\left\{|X_\ell|^2 |Y_0(k), \ldots, Y_{\ell'}\right\}$ denotes the short-term spectrum of the speech signal.

For the LSA error function in (3), the gain function is [3]

$$G_{\mathrm{LSA}}\left(\xi_{\ell|\ell'}, \gamma_\ell\right) = \frac{\xi_{\ell|\ell'}}{\xi_{\ell|\ell'}+1} \exp\left(0.5 \int_{v_\ell}^{\infty} \frac{e^{-t}}{t} dt\right), \tag{6}$$

where v_ℓ is defined by $v_\ell \triangleq \frac{\gamma_\ell \xi_{\ell|\ell'}}{\xi_{\ell|\ell'}+1}$. Hence, the problem in this case reduces to finding an estimator for the a-priori SNR.

2.1. Decision-directed estimator

The widely used decision-directed (DD) estimator of Ephraim and Malah [2] is given by:

$$\hat{\xi}_{\ell|\ell} = \max\left\{ \alpha \frac{\left|\hat{X}_{\ell-1}\right|^2}{\sigma_\ell^2} + (1-\alpha) P(\gamma_\ell - 1), \xi_{\min} \right\}, \tag{7}$$

where $P(x) = x$ if $x \geq 0$ and $P(x) = 0$ otherwise, and α is a smoothing parameter, $\alpha \in [0, 1)$. As Cappé clearly explained [4], there is a trade-off in the choice of the parameter α. To reduce the musical noise, it is necessary to choose α close to 1; however, the closer α gets to 1, the higher the distortion introduced into the signal is. A typical value for α that has been found to provide a good compromise is 0.98. The parameter $\xi_{\min}$ is the noise floor, essentially controlling the noise reduction and the perceptual masking of the residual musical noise. Applying a lower limit of $\xi_{\min}$ is an additional option for this algorithm which is typically set to -15 dB.

2.2. ARCH model

Instead of the DD estimator, we use a two step estimator [6], composed of a propagation step and an update step to recursively update the estimate of the conditional a-priori SNR.

Suppose we are given an estimate of the *one-frame-ahead a-priori SNR* $\hat{\xi}_{\ell|\ell-1}$ and a new noisy spectral component Y_ℓ, then the estimate $\hat{\xi}_{\ell|\ell}$ can be updated by computing the conditional a-priori SNR:

$$\hat{\xi}_{\ell|\ell} = E\left\{ \frac{|X_\ell|^2}{\sigma_\ell^2} \middle| \hat{\xi}_{\ell|\ell-1}, Y_\ell \right\}. \tag{8}$$

The gain function

$$G_{\mathrm{SP}}\left(\xi_{\ell|\ell'}, \gamma_\ell\right) = \sqrt{\frac{\xi_{\ell|\ell'}}{\xi_{\ell|\ell'}+1}\left(\frac{1}{\gamma_\ell} + \frac{\xi_{\ell|\ell'}}{\xi_{\ell|\ell'}+1}\right)} \tag{9}$$

minimizes the expected spectral power distortion [8], yielding:

$$\hat{\xi}_{\ell|\ell} = G_{\mathrm{SP}}^2\left(\hat{\xi}_{\ell|\ell-1}, \gamma_\ell\right) \cdot \frac{|Y_\ell|^2}{\sigma_\ell^2} = G_{\mathrm{SP}}^2\left(\hat{\xi}_{\ell|\ell-1}, \gamma_\ell\right) \cdot \gamma_\ell. \tag{10}$$

Using (9) and (10), we can write

$$\hat{\xi}_{\ell|\ell} = \alpha_\ell \hat{\xi}_{\ell|\ell-1} + (1-\alpha_\ell)(\gamma_\ell - 1), \tag{11}$$

where

$$\alpha_\ell = 1 - \left(\frac{\hat{\xi}_{\ell|\ell-1}}{\hat{\xi}_{\ell|\ell-1}+1}\right)^2, \quad \alpha_\ell \in [0, 1]. \tag{12}$$

Computation of the update step requires an estimate of $\hat{\xi}_{\ell|\ell-1}$. According to the $\mathrm{GARCH}(p, q)$ model presented in [5],

$$\hat{\xi}_{\ell|\ell-1} = \kappa + \sum_{i=1}^{q} \mu_i \hat{\xi}_{\ell-i|\ell-i} + \sum_{j=1}^{p} \delta_j \hat{\xi}_{\ell-j|\ell-j-1}, \tag{13}$$

where the values of the parameters are constrained by

$$\kappa > 0,\ \mu_i \geq 0,\ \delta_j \geq 0, \quad i = 1, \ldots, q, j = 1, \ldots, p$$
$$\sum_{i=1}^{q} \mu_i + \sum_{j=1}^{p} \delta_j < 1.$$

Using the special case $\mathrm{GARCH}(0, 1)$, also known as $\mathrm{ARCH}(1)$, we get the propagation step:

$$\hat{\xi}_{\ell|\ell-1} = \kappa + \mu \hat{\xi}_{\ell-1|\ell-1}, \quad \kappa > 0,\ 0 \leq \mu < 1. \tag{14}$$

Since the a-priori SNRs need to be equal to $\xi_{\min}$ when speech is absent, we obtain from (14) a condition on κ, $\kappa = (1-\mu)\,\xi_{\min}$, implying

$$\hat{\xi}_{\ell|\ell-1} = (1-\mu)\,\xi_{\min} + \mu\hat{\xi}_{\ell-1|\ell-1}. \tag{15}$$

The effect that $\xi_{\min}$ and μ have on the processed signal is investigated in Section 4.

Note that from (7) and (11) we can observe that an a-priori SNR estimator, which is based on a GARCH model, has a similar form of the decision-directed estimator but with a *time-varying frequency-dependent* weighting factor α_ℓ.

3. PERFORMANCE MEASURES

We employ three performance measures commonly used for the quality assessments of a speech enhancement algorithm. First is the speech distortion, which is used to assess the quality of the estimated speech component. The second measure is the noise reduction, and the third measure is the artificial distortion of the noise, i.e, the musical noise.

3.1. Distortion and noise reduction ratio

Combining (1) and (4) we obtain:

$$\hat{X}_\ell = G\left(\xi_{\ell|\ell'}, \gamma_\ell\right) X_\ell + G\left(\xi_{\ell|\ell'}, \gamma_\ell\right) D_\ell, \tag{16}$$

where the first element is the filtered desired signal, and the second element is the residual noise. With the same gain applied to both the signal and the noise, the trade-off between the distortion and the noise reduction when speech is present is clear. From this expression we derive the distortion measure

$$\mathrm{J_X} \triangleq E\left\{\left(\log|X_\ell| - \log\left|G\left(\xi_{\ell|\ell'}, \gamma_\ell\right) X_\ell\right|\right)^2\right\}, \tag{17}$$

which is evaluated only in time-frequency bins containing the speech signal. From (16) we also derive the noise reduction ratio (NRR), which is evaluated across all time-frequency bins,

$$\mathrm{NRR} \triangleq \frac{E\left\{|D_\ell|^2\right\}}{E\left\{\left|G\left(\xi_{\ell|\ell'}, \gamma_\ell\right) D_\ell\right|^2\right\}}. \tag{18}$$

3.2. Musical noise via higher order statistics

The attenuated noise, as a result of the processing, will be composed of isolated spectral components, also known as *tonal* components, or musical noise. These tonal components can be quantified, as they are related to the weight of the tail of the noise components' probability density function (pdf). Hence the kurtosis, defined as $\mathrm{kurtosis} = \mu_4/\mu_2^2$, where μ_m is the mth order moment of the signal, can be used to evaluate the percentage of components which are tonal. However, as the original noise could also contain some tonal components, and we are interested in the amount of tonal components caused by the processing, we use the ratio of the kurtosis before and after the processing. Hence, we define the third measure, the log of the kurtosis ratio (LKR) as:

$$\mathrm{LKR} \triangleq \log_{10}\left(\frac{\mathrm{kurtosis_{proc}}}{\mathrm{kurtosis_{org}}}\right), \tag{19}$$

which is evaluated on noise only frames. $\mathrm{kurtosis_{org}}$ is the kurtosis of the input noise and $\mathrm{kurtosis_{proc}}$ is the kurtosis of the processed noise. The LKR increases as the musical noise increases [9], and the absence of musical noise corresponds to LKR of zero and below.

Analytical calculation of the kurtosis ratio requires the use of a specific noise reduction method or assumptions about the statistical spectral components [9, 10]. Here, we use the sample kurtosis [11]:

$$\mathrm{kurtosis} = \frac{1}{L}\sum_{\ell=0}^{L}\left[\frac{\frac{1}{N}\sum_{k=0}^{N-1}\left(\left|D_\ell(k)\right|^2 - \overline{\left|D_\ell(k)\right|^2}\right)^4}{\left(\frac{1}{N}\sum_{k=0}^{N-1}\left(\left|D_\ell(k)\right|^2 - \overline{\left|D_\ell(k)\right|^2}\right)^2\right)^2}\right] \tag{20}$$

where $\overline{\left|D_\ell(k)\right|^2} = \frac{1}{N}\sum_{k=0}^{N-1}\left|D_\ell(k)\right|^2$. The sample kurtosis is calculated for the input signal and the processed signal separately and then plugged into the LKR (19).

4. EXPERIMENTAL RESULTS AND DISCUSSION

The evaluation of the signal distortion was done on speech signals taken from the TIMIT database, including 20 different utterances from 20 different speakers, half male and half female. The signals are sampled at 16kHz and degraded by white Gaussian noise with SNR in the range $[0, 20]$ dB. The noisy signals are transformed to the time-frequency domain using STFT, with 75% overlapping Hamming analysis windows of 32 ms length. To calculate the distortion (17) the time-frequency bins containing speech were defined as $\mathcal{H}_1 = \{\ell, k \mid 20\log_{10}|X_\ell(k)| > \epsilon\}$, where $\epsilon = \max_{\ell,k}\{20\log_{10}|X_\ell(k)|\} - 50$, confining the dynamic range of the log-spectrum to 50 dB.

The evaluation of the musical noise (represented by the LKR) was done separately on a complex white Gaussian noise in the time-frequency domain, to emulate performance in noise only frames.

Figure 1(a)–(c) clearly demonstrates Cappé's explanation of the influence of the parameters α and $\xi_{\min}$ on the distortion of the signal, the musical noise, and the residual noise level for the decision-directed estimator. The distortion increases as α increases, while the musical noise decreases and is in fact almost eliminated for high enough α. By taking a higher noise floor $\xi_{\min}$ there is more residual noise, and the value of α required to reduce or eliminate the musical noise decreases.

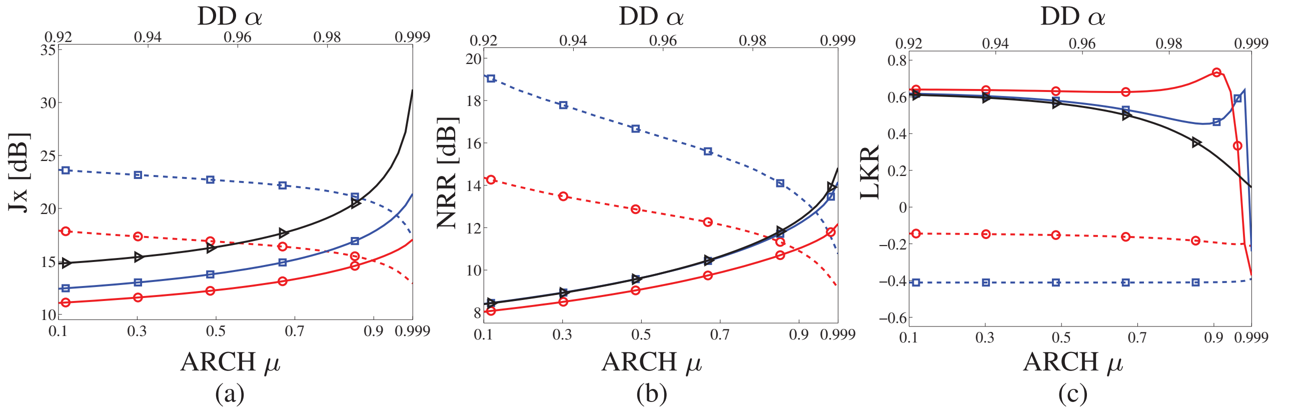

Fig. 1: Comparison of DD (solid lines) and ARCH (dashed lines) estimators for 5 dB SNR: (a) Distortion, (b) NRR and, (c) LKR, with varying α (upper axis) and μ (lower axis) respectively per estimator, and $\xi_{\min}$ of -20 dB (square), -15 dB (circle), and for DD method only $\xi_{\min} = 0$ (triangle).

It is interesting to note that for α close to 1 and without a noise floor limit $\xi_{\min}$ (i.e. $\xi_{\min} = 0$), the musical noise monotonically decreases as α increases, but with $\xi_{\min} > 0$ there is actually a slight increase in the amount of musical noise before it drops. When we add a noise floor limit, intervals in which the a-priori SNR was lower than this limit are replaced by that value, thus distorting the spectral noise components pdf.

From Figure 1 we can observe that for the ARCH estimator, increasing the value of μ actually decreases the distortion. This can be easily understood from (11) and (15). As μ increases, the one-frame-ahead a-priori SNR $\hat{\xi}_{\ell|\ell-1}$ becomes more dependent on the previous frame's a-priori SNR $\hat{\xi}_{\ell|\ell-1} \approx \hat{\xi}_{\ell-1|\ell-1}$. Hence, when a signal component appears abruptly ($\gamma_\ell \gg 1$), after a one frame delay, α_ℓ will *decrease* approximately according to the previous frame's a-priori SNR, and the a-priori SNR will follow the a-posteriori SNR of the current frame $\hat{\xi}_{\ell|\ell} \approx \gamma_\ell$ (whereas for the DD after a frame delay it is approximately $\gamma_{\ell-1}$). It is important to note that when the signal component *disappears*, the conditional a-priori SNR immediately drops and there is no frame delay, as opposed to the DD which has one frame delay. This strong dependence on the one-frame-ahead a-priori SNR when μ increases also causes the NRR to decrease, as the a-priori SNR does not drop to $\xi_{\min}$ immediately after frames that contain speech. Clearly, the lower we take the noise floor $\xi_{\min}$, the more noise reduction we get.

While the distortion and NRR strongly depend on the value of μ, the musical noise (LKR) mainly depends on the noise floor $\xi_{\min}$. Lower $\xi_{\min}$ means higher α_ℓ, resulting in a smoother a-priori SNR around $\xi_{\min}$, thus reducing the musical noise. This is in contrast to the DD estimator where for low $\xi_{\min}$ less variation is being masked, resulting in more musical noise and less attenuation, until α is large enough to smooth out the variations.

For the DD estimator we have to compromise between the amount of distortion and amount of musical noise, while for the ARCH estimator, the musical noise can be eliminated by choosing an appropriate value of $\xi_{\min}$. However, for the ARCH estimator we need to compromise between the amount of distortion and the amount of residual noise. In comparison to the typical DD estimator values of $\alpha = 0.98$ and $\xi_{\min} = -15$ dB, where we still have musical noise, we can see that by using the ARCH model we can eliminate the musical noise almost completely if we choose the same $\xi_{\min} = -15$ dB. If we choose, for example, $\mu = 0.98$ we also get less distortion; however, we get slightly less noise reduction than the DD estimator. Maintaining the same $\xi_{\min} = -15$ dB, if α for the DD estimator is increased so that we perceive no musical noise, the speech distortion and noise reduction also increase. Using the ARCH model, we can either use the same values of parameters as before, obtaining less distortion than the DD estimator but also less noise reduction, or we can decrease μ to get higher noise reduction at the expense of higher speech distortion. In both cases we can perceptually avoid musical noise. On the other hand, if we can tolerate a slightly higher distortion than the DD estimator, we can choose smaller μ so that more noise reduction can be obtained compared to DD estimator (with no impact on the musical noise).

5. CONCLUSION

We have demonstrated and compared the use of the DD and ARCH estimators for spectral speech enhancement. We investigated the influence of the ARCH parameters on the different measures of the sound quality of the processed signal, and demonstrated that by using the ARCH model it is possible to achieve better results than the DD method for some of those measures, specifically the musical noise, while compromising between the speech distortion and noise reduction. It would be interesting to expand the model to a full GARCH(p,q) model and conduct a similar analysis.

6. REFERENCES

[1] P. C. Loizou, *Speech Enhancement Theory and Practice*, CRC Press, Taylor & Francis Group FL, 2007.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust. Speech Signal Process*, vol. ASSP-32 (6), pp. 1109–1121, 1984.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error log-spectral amplitude estimator," *IEEE Trans. Acoust. Speech Signal Process*, vol. ASSP-33 (2), pp. 443–445, 1985.

[4] O. Cappé, "Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor," *IEEE Trans. Acoust. Speech Signal Process*, vol. 2 (2), pp. 345–349, 1994.

[5] I. Cohen, "Modeling speech signals in the time-frequency domain using GARCH," *Signal Processing*, vol. 84 (12), pp. 2453–2459, 2004.

[6] I. Cohen, "Relaxed statistical model for speech enhancement and a priori SNR estimation," *IEEE Trans. Speech and Audio Processing*, vol. 13 (5), pp. 870–881, 2005.

[7] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech and Audio Processing*, vol. 11 (5), pp. 466–475, 2003.

[8] P. J. Wolfe and S. J. Godsill, "Efficient alternatives to the ephraim and malah suppression rule for audio signal enhancement," *EURASIP Journal on Applied Signal Processing*, vol. 2003:10, pp. 1043–1051, 2003.

[9] Y. Uemura, Y. Takahashi, H. Saruwatari, K. Shikano, and K. Kondo, "Automatic optimization scheme of spectral subtraction based on musical noise assessment via higher-order statistics," *in Proc. International Workshop on Acoustic Echo and Noise Control*, 2008.

[10] S. Kanehara, H. Saruwatari, K. Shikano R. Miyazaki, and K. Kondo, "Theoretical analysis of musical noise generation in noise reduction methods with decision-directed a priori SNR estimator," *in Proc. International Workshop on Acoustic Signal Enhancement*, pp. 1–4, 2012.

[11] H. Yu and T. Fingscheidt, "Black box measurement of musical tones produced by noise reduction systems," *in Proc. IEEE International Conference on Acoustics Speech and Signal Processing*, pp. 4573–4576, 2012.