GARCH COEFFICIENTS AS FEATURE FOR SPEECH RECOGNITION IN PERSIAN ISOLATED DIGIT

Mohamad Abdolahi, Hamidreza Amindavar

Amirkabir University of Technology, Department of Electrical Engineering,
424 Hafez, Tehran, Iran
abdolahi@cic.aut.ac.ir, hamidami@cic.aut.ac.ir

ABSTRACT

This paper describes a new technique that gives high performance based on GARCH (Generalized Autoregressive Conditional Heteroskedastic) time series modeling incorporating past variances to predict future variances. This is particularly suitable since no transformation on the speech signal is performed, rather we have a new statistical feature extraction, moreover, the speech signals are among non stationary processes whose variance are heteroskedastic; e.g., time varying. Therefore, we provide a new parametric speech modeling using GARCH coefficients. The features resulting from GARCH modeling are used for recognition of isolated digits 1 to 10 in Persian language. The results show a significant improvement in the recognition accuracy compared to results based on Mel-frequency cepstrum coefficients (MFCC).

1. INTRODUCTION

The structure of many successful systems for speech recognition typically consists of a signal preprocessing feature extraction followed by a pattern classifier. Automatic extraction of useful information from speech has been a subject of active research for many decades. The Mel-frequency cepstral feature extraction methods that are currently used in many speech recognition systems are motivated by the properties of the human auditory system and speech perception. However, despite their general acceptance as the standard features in clean speech recognition, the cepstral features are widely acknowledged not to cope well with the noisy speech. In order to improve the robustness of front-ends with respect to noise and distortion some alternative features have been proposed [5, 6, 7]. On the other hand, the MFCC have proven to be one of the effective set of features for speech recognition. They are computed by applying a Mel-scaled filter bank either to the short-term fast Fourier transform magnitude spectrum or to the short-term linear prediction coefficient-based spectrum to obtain a perceptually meaningful smoothed spectrum. Despite the empirical superiority of MFCC's over many other types of signal processing techniques, there are no theoretical reasons why the linear transformation associated with the discrete cosine transformation performed on Mel-filter bank log channel energies could construct an optimal transformation since it is fixed a priori and it is also independent of HMM states and of the speech classes. For this reason, a search for a new statistical model of speech has led to the so called optimum-transformed HMM [8] based on minimum classification error criterion. However, there is a transformation involved namely the Mel-warped discrete Fourier transform DFT in addition to the training of HMM using the gradient descent method. The state-dependent transformation on the Mel-warped DFT, together with the HMM parameters, is automatically trained using the gradient descent method, resulting in a minimization of a measure of an overall empirical error count. Mostly the statistical models are either segment or frame based strategies, however, features are extracted by a deterministic method at frame level. Many statistical models such as stochastic segment model SSM [3] or the simpler parametric trajectory model [4] that models the time variation, are operating at the segment level. However, at the frame duration, for the frequency domain transformation we must resort to the stationarity assumption. In the frequency domain approach, a plausible speech recognition can be attained, but, at the expense of a large number of features; therefore, we are faced with a high feature space dimension. For analysis simplicity the increase in the size of the feature space forces us to assume the independency among features, this shortcoming is accumulated as the observation vector dependency problems grow larger. Moreover, with the assumption on the independency of features; i.e, a diagonal covariance matrix assumption, and the conditional independency of feature vector (such as in HMM), we are apt to lose more information of speech signal behavior. In addition to being sensitive to additive and conventional noise, a critical issue in recognition using MFCC features, MFCC is known as a deterministic mapping from the signal space to the feature space and therefore the whole role of the statistical modeling has to be done by the recognizer. Since the dimension on the feature space is large, we need some assumptions such as independency between feature components; in order to have a diagonal covariance matrix, and we need the conditional independency between feature vectors in HMM in order to limit the computational complexity. Furthermore, with a large number of features, the search space is large as well to find the optimum model in the training phase, and this could result in an ill conditioned covariance matrix for a small database and/or a large model parameter set. GARCH for parametric time series modeling, with the simultaneous capability of time varying temporal variance modeling , can be used instead of MFCC for speech recognition. GARCH, which stands for generalized autoregressive conditional heteroskedasticity model provides a leverage on the assumption of finite variance for all stochastic processes. Generally speaking, we consider heteroskedasticity as time-varying variance. Conditional implies a dependence on the observation of the immediate past, and autoregressive describes a feedback mechanism that incorporates past observation into the present. GARCH then is a mechanism that includes past variance in the description of the future variance. The pioneering research by [10], and then [9, 11, 12] has shown that a time-varying variance instead of a constant is more useful in modeling non stationary phenomena such as

economic series. GARCH models account for heavy tailed probability distributions as excess kurtosis. Speech signals are known to resemble a non stationary process, therefore, a successful candidate for a GARCH model. However low order GARCH, i.e. GARCH(1,1) which is used in this paper, couldn't model speech as a whole signal by a few parameters, and high order GARCH is too computationally expensive for this purpose; also we couldn't segment speech sequence by GARCH. In this paper, we use GARCH speech modeling to obtain some features in the time domain for speech recognition.

The paper is organized as follows: In section 2 we discuss the GARCH model. In section 3 the GARCH coefficients are used to extract the GARCH features for speech recognition and validate the performance over Farsi digits in clean and noisy environments, and some concluding remarks are provided at the end.

2. GARCH MODELING

Bollerslev [9] developed GARCH as a generalization of Engle's [10] original autoregressive conditional heteroskedasticity (ARCH) volatility modeling technique. Bollerslev designed GARCH to offer a more parsimonious model that lessens the computational burden. Next, we briefly discuss GARCH modeling. Let's consider the time series X_t defined by

$$X_t = E\{X_t \,|\Psi_{t-1}\} + \varepsilon_t, \tag{1}$$

where Ψ_{t-1} denotes all information about X_t until time $t-1$ and ε_t is a residual error, and furthermore, assume that $E\{X_t \,|\Psi_{t-1}\}$=0 , and if

$$\varepsilon_t = \sqrt{h_t}\, z_t, \quad z_t \sim \mathcal{N}(0,1) \tag{2}$$

$$h_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{q} \beta_i h_{t-j}, \tag{3}$$

then X_t is a generalized autoregressive conditionally heteroscedastic (GARCH) process of order (p, q), z_t is an IID sequence of random variables [9] here assumed Gaussian, where $\mathcal{N}(\cdot)$ stands for the Gaussian distribution of mean zero and variance 1, and α_i and β_j are non-negative constants with the convention that $\alpha_i > 0$, and $\beta_j > 0$. The model in (1)-(2) describes a relationship between the variance at time t and the past variances. According to this model small changes follow small changes and large changes follow large. This describes the volatility clustering and the capability of heavy tailed distribution modeling of GARCH process. For GARCH(1,1) , we have:

$$h_t = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 h_{t-1}, \tag{4}$$

and in order to have finite unconditional variance, the following constraint is imposed on the GARCH coefficients

$$\sum_i \alpha_i + \sum_j \beta_j < 1. \tag{5}$$

In IGARCH [13] this inequality is replaced with equality and allows modeling time series whose variance is not even finite. We use GARCH(1,1) to model speech signals. Next we adopt the maximum likelihood principal to estimate the parameters of the GARCH model. GARCH model parameters can be estimated using maximum likelihood estimation (MLE). We assume that $\{X_1, \cdots, X_T\}$ are generated through some mechanism modelled by GARCH(1,1), then the likelihood function is formulated as

$$\mathcal{L}(\alpha_0, \alpha_1, \beta_1) = f_{X_2,\cdots,X_T|X_1,h_1}(X_2, \cdots, X_T|X_1, \mathbf{h}) \tag{6}$$

$$= \prod_{j=2}^{T} \frac{1}{\sqrt{2\pi h_j}} \exp\left(-\frac{x_j^2}{2h_j}\right), \tag{7}$$

$$h_j = \alpha_0 + \alpha_1 X_{t-1}^2 + \beta_1 h_{t-1},$$

where h_j are obtained recursively. By taking the logarithm and neglecting the constant term, we obtain the log likelihood function as

$$\ell(\alpha_0, \alpha_1, \beta_1|\mathbf{X}, \mathbf{h}) = -0.5 \sum_{j=2}^{T} \log h_j + x_j^2/h_j^2, \tag{8}$$

where $\mathbf{X} = (x_1, \cdots, x_T)^T$, and $\mathbf{h} = (h_1, \cdots, h_T)^T$, superscript T denotes transposition, with the aid of a constrained nonlinear optimizations technique subject to the constraint in (5), the GARCH coefficients of the model are obtained, these coefficients are to serve as the new speech features. Next, we examine the performance of GARCH modeling in speech recognition.

Mixture	2	4	8
MFCC Viterbi	90.6%	95.0%	96.33%
MFCC Baum-Welch	93.5%	97.0%	97.33%

Table 1: Recognition performance using MFCC.

3. IMPLEMENTATION RESULT

We use GARCH modeling to represent the speech signal, and then include these coefficients into MFCC features for increased performance of the speech recognizer in the isolated digit classification in Persian, where HMM is used for classification with 5 states, constant for all classes, with 2, 4, and 8 Gaussian mixtures. The MFCC features are used with the Viterbi and Baum-Welch HMM training method to compare the gain of the model training according to the expectation maximization (EM) method in Baum-Welch with the K-means algorithm in Viterbi. The other models include MFCC plus GARCH features and only the GARCH coefficients are trained and tested with the Viterbi method. The speech database included pronunciation of digits 1 to 10 in Persian language, 30 utterances per speaker per digit, 6 male speakers, in 3 different days for any speaker, 1200 samples for training and 600 samples for tests are used. The benchmark features were 25 MFCC coefficients (12 MFCC +12 ΔMFCC + 1 log energy), and the GARCH features include 3 coefficients, the unconditional variance α_0 , ARCH and GARCH coefficients α_1, β_1. In order to decrease the feature extraction time the iterations are restricted, without any loss of performance in recognition, throughout our simulations we use 2 iterations. The three GARCH coefficients are extracted from any frame with a duration of about 15 msec. HMM training with a Gaussian mixture according to the Viterbi and Baum-Welch methods are applied and 0.001 is selected as the minimum variance. The results in Table 1 as a benchmark demonstrate the improvement achieved by using the Baum-Welch algorithm and the EM algorithm versus the Viterbi method using the clustering method. However, the improvements are not staggering given the computational burden and the complexity of EM algorithm. The results

Mixture	2	4	8
MFCC+GARCH	99.50%	99.83%	100%
GARCH(3 coefficients)	99.17%	99.83%	100%

Table 2: Recognition performance using MFCC and GARCH, Viterbi training throughout.

Mixture	2	4	8
MFCC(25 coefficients)	82.17%	86.33%	85%
GARCH(3 coefficients)	76.00%	95.67%	99.83%

Table 3: Recognition performance in noisy environment.

in Table 2 demonstrate that GARCH coefficients improve recognition performance once added to the MFCC features. Noting that the second row of Table 2 is obtained using 25 MFCC features and 3 GARCH coefficients, and comparing the second row of Tables 1 and 2, in spite of a weak HMM training (Viterbi) and a small number of mixtures, by dispensing with the MFCC coefficients all together in the third row of Table 2 but keeping the GARCH coefficients, the simulations achieve similar performance results. The second and third row of Table 2 claims that the advantage gained by including 25 MFCC coefficients is not overwhelming, as it was stated in the introduction the theoretical basis for MFCC is not established yet. Tables 1 and 2 are in a noise free environment. Next, we demonstrate the performance for noisy speech.

When Gaussian noise is present, MFCC and GARCH performances are compared under the condition that signal to noise ratio (SNR) is set to -3 dB, including for both the training and test data. As in previous scenarios, the models in the two cases train 25 MFCC and 3 GARCH coefficient features. The results of recognition in Table 3 demonstrate that the recognition by only 3 GARCH coefficients is more robust and consistent than 25 MFCC features, and as the number of mixtures increases the performance of GARCH based feature recognition method consistently increases. In fact, the additive white noise has a constant power over all frequency bands and infects all of the 25 MFCC features, but the effect of white noise on the GARCH model is analyzed as follows. According to (2), and (3) for white noise only α_0 is nonzero and α_1 and β_1 are zero. This conclusion can be easily drawn because of the unconditional variance nature of the white noise, and no conditional variance aspects are present in the additive white noise. However, the addition of white noise is not additive in the GARCH model coefficients, but only shifts the statistical coefficients that are proportional to variance, and its effect according to the empirical results is not as critical as that of white noise on cepstral coefficients as white noise can be on the cepstral coefficients. As an example, let's consider

$$Y = X + N, \tag{9}$$

where X is the clean speech, N is a zero mean white noise of variance δ. As $\delta \rightarrow \infty$ the GARCH(1,1) parameters denoted by $(\alpha_0, \alpha_1, \beta_1)$ approach $(\delta, 0, 0)$, therefore, $\alpha_0 \rightarrow \infty$, and (α_1, β_1) $\rightarrow (0, 0)$ monotonically; i.e., the parameters of the GARCH model are shifting proportionally. In a noisy environment, this important result is manifested in the feature space. The clusters created by $(\alpha_0, \alpha_1, \beta_1)$ are only shifted in the feature space proportionally, the noise only shifts different clusters and does not cause them to interfere with each other. For one of the utterances, this analysis is verified by Figure 1 where the addition of noise only shifts the unconditional variance GARCH coefficient; α_0, with respect to the noiseless speech unconditional GARCH coefficient, and in Figure 2 where the addition of noise again only shifts the conditional variance GARCH coefficient; β_1, and it approaches zero under the noisy environment. But the characteristics of this coefficient sequence under either noisy or clean speech are preserved. According to Table 3, we have a decrease in GARCH model performance with 2 mixtures where the problem is attributed to the clusterings of small number of features; 3 features per frame in Viterbi training that exploits the K-means algorithm per segment. On the other hand, usually, HMM training in addition to Baum-Welch training is used in practical applications where the number of mixtures is a lot more than 2.

4. CONCLUSION

In this paper, we discussed the feasibility of GARCH modeling for speech signals. GARCH speech modeling can capture the time varying variance nature of speech signals effectively. The recognition results illustrate that GARCH(p, q) coefficients can replace the MFCC features for speech recognition applications successfully. This is especially noteworthy since $p + q$ coefficients of GARCH can replace 25 MFCC features where the dimensions of the feature space are significantly reduced and subsequently the volume of search space to find the optimal model in training phase is reduced by a factor of 3/25. Furthermore, the size of covariance matrix is reduced; hence, it is well posed for feature vector. Thus using GARCH modeling the ill conditioning problem of the covariance matrix for small database is less critical. By using GARCH modeling, not only the model variation at any frame duration is captured explicitly, but also the theoretical optimizations that model variations at segment level, (that were laid dormant because they were too complexity expensive for usage), are now more at hand since the feature space is reduced.

Acknowledgment

Authors appreciate the efforts by Mehdi Karimi during the preparation process for this paper.

5. REFERENCES

[1] F. Norden, T. Eriksson, "Time evolution in LPC spectrum coding,"*IEEE Trans. on Speech and Audio Processing*, Vol. 12 , NO. 3 , pp. 290-301, May 2004.

[2] J. Chen, Y. A. Huang, Q. Li, K. K. Paliwal,"Recognition of noisy speech using dynamic spectral subband centroids," *IEEE Signal Processing Letters*, Vol. 11, NO. 2, pp. 258-261, Feb. 2004.

[3] M. Ostendorf, V. Digalakis, O. A. Kimball, "From HMM's to segment models: A unified view of stochastic modeling for speech recognition," *IEEE Trans.* Vol.4, NO. 5, pp. 360-378, Sept. 1996.

[4] L. Deng, M. Aksmanovic, Xiaodong Sun, C. F. J. Wu, "Speech Recognition using Markov Models with polynomial Regression Function and Non stationary states," *IEEE Trans. on Speech and Audio Processing*, Vol. 2 , NO. 4,pp. 507-520, Oct. 1994.

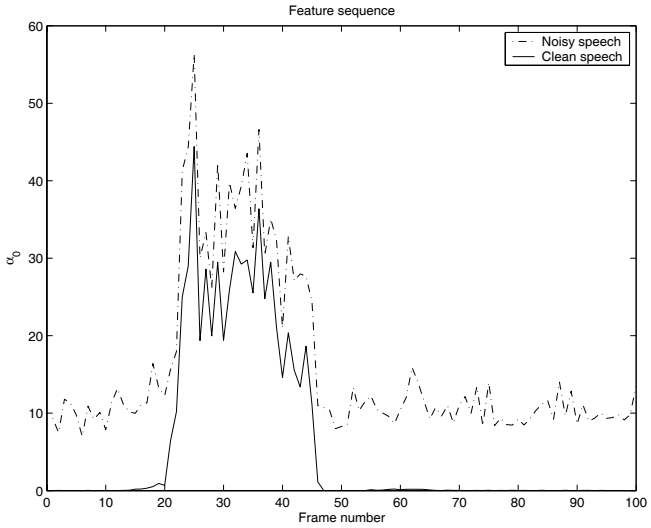

Figure 1: The effect of noise on the unconditional variance feature, SNR=-3dB.

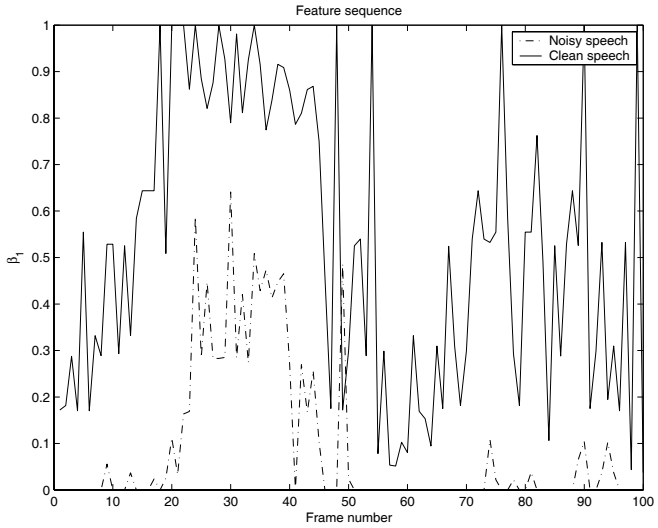

Figure 2: The effect of noise on the conditional variance feature, SNR=-3dB.

[5] J. W. Pitton, K. Wang, B. H. Juang, "Timefrequency analysis and auditory modeling for automatic recognition of speech," *Proceedings of IEEE*, Vol. 84 , NO. 9, pp. 1199-1214, Sept. 1996.

[6] A. Potamianos, P. Maragos, "Time-frequency distributions for automatic speech recognition," *IEEE Trans. Speech Audio Processing*, Vol. 9 , NO. 3 , pp. 196–200, March 2001.

[7] O. Ghitza, "Auditory models and human performance in tasks related to speech coding and speech recognition," *IEEE Trans. Speech Audio Processing*, Vol. 2, NO. 1, pp. 115–132, Jan. 1994.

[8] R. Chengalvarayan, L. Deng,"HMM-based speech recognition using statedependent, discriminatively derived transforms on Melwarped DFT features," *IEEE Trans. on Speech and Audio Processing*, Vol. 5, NO. 3, May 1997.

[9] T. Bullerslev, "Generalized autoregressive conditional heteroskedasticity," *J. Econometrica*, Vol. 31, pp. 307-327, 1986.

[10] R. F. Engle, "Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation," *J. Econometrica*, Vol. 50, NO. 4, pp. 987-1007, 1982.

[11] Y. Cheung, L. Xu ,"Dual multivariate auto-regressive modeling in state space for temporal signal seperation," *IEEE Trans. System, Man, and Cybernetics*, NO. 10, pp. 1–13, 2003.

[12] W. K. Li, S. Ling, H. Wong , "Estimation for partially nonstationary multivariate autoregressive models with conditional heteroskedasticity," *J. Biometrica*, Vol. 88, NO. 2, 2001.

[13] P. Reinhard Hansen and A. Lunde, "Does anything beat a GARCH(1,l)? a comparison based on test for superior predictive ability," *IEEE Int'l Conf. on Computational Intelligence for Financial Engineering*, March 2003.