# USC–SIPI REPORT #196

## Cumulant-Based Adaptive Analysis of Speech Signals

by

**Mithat C. Dogan and Jerry M. Mendel**

**January 1992**

Signal and Image Processing Institute
**UNIVERSITY OF SOUTHERN CALIFORNIA**
Department of Electrical Engineering-Systems
3740 McClintock Avenue, Room 400
Los Angeles, CA 90089-2564 U.S.A.

# Abstract

This report describes a speech processing method consisting of an adaptive predictor, a voicing decision (V/UV), and a pitch period estimator. The focus of this report is on robust detection of speech state and estimation of pitch period. This is accomplished by observing the behavior of an adaptive predictor which processes the speech signal. Higher-order- statistical analysis is proposed for discrimination of speech states. Comparing the energy of the original speech signal with that of the prediction-error residual yields the decision method. Both covariance and cumulant-based prediction methods are investigated and the latter is shown to be a more robust way of making (V/UV) decision. Pitch estimation is accomplished by using correlation-based approaches that operate on the energy estimate of the cumulant-based prediction residual rather than the original speech signal. Pitch estimation by our method yields better performance than currently existing batch procedures.

# Contents

# List of Figures

# Chapter 1

# Introduction

Voiced/Unvoiced (V/UV) decision is an important problem in speech processing. Almost all speech coding, recognition and speaker identification systems require this information for an effective processing of speech data. In addition, low-delay speech processing systems require this decision be provided in real-time. In [1] some commonly employed features are described, and a subset of them are used to train an artificial neural network to perform V/UV decision.

In frame-based analysis of speech signals, feature extraction is performed on the current block of data, and a decision is given at the end of the period. For this reason, frame-based methods are incapable of tracking rapid changes in signal characteristics. Transitions of the state of speech within a frame period affect the decisions resulting from a frame-based analyzer. In general, this mixed state of speech within a period can not be identified and incorrect decisions will be made. This will degrade the performance of the overall speech processing system. In addition, frame-based analysis introduces delay, which may not be tolerable in low-delay systems.

Severe non-stationarity observed in speech signals and low-delay requirements of the contemporary speech processing systems motivate the use of adaptive algorithms for feature extraction in place of their batch counterparts. In general, adaptive processing techniques are designed to minimize some least-squares error criterion. Their use is motivated by the assumption that the processes are Gaussian and the performance analysis is tractable with this assumption [2]; however, this approach ignores the non-Gaussian nature of the underlying signal.

Adaptive prediction of the incoming signal and continuous monitoring of prediction error power makes detecting changes in the spectral characteristics of the process possible. We may consider such a change as an *event*. After an event, an adaptive unit will require a period to adjust itself for the new configuration. During this learning period, prediction error power will temporarily increase. This observation was used in [3], to detect abrupt changes in the autoregressive (AR) parameters of a linear process. If a lattice form is used rather than a finite impulse response (FIR) filter, reflection coefficients will be available for monitoring purposes. In addition, adaptive lattice filters exhibit better learning characteristics than their FIR counterparts. This may improve the ability to localize the event when prediction error power is monitored.

In this report, we shall investigate the application of adaptive prediction methods to detect V/UV transitions in speech signals. Hence, events of interest will be V/UV or UV/V transitions. Our approach will take the speech production model into account and utilize higher than second-order statistics of speech signals.

# Chapter 2

# Higher-Order-Statistics

During the last decade, there has been an intense research activity on higher-order statistics, also known as cumulants. Their blindness to Gaussian noise and ability to reveal phase information are major reasons for this activity. For a tutorial on this important topic, we refer the reader to [4]. In this chapter, we shall briefly summarize the properties of cumulants.

Let $\{x_1, x_2, \ldots, x_n\}$ be a set of n real random variables, and $\{v_1, v_2, \ldots, v_n\}$ be a set of n real and deterministic variables. Then, the $n^{th}$ order cumulant of $\{x_1, x_2, \ldots x_n\}$ is defined as the coefficient of $v_1, v_2, \ldots, v_n$ in the Taylor series expansion around the origin of the cumulant-generating function defined by,

$$K(v_1, v_2, \ldots, v_n) = ln\, E\{\, exp(\, j\,(v_1 x_1 + v_2 x_2 + \ldots + v_n x_n)\} \tag{2.1}$$

Based on the above definition, cumulants of zero-mean random variables can be expressed in the form,

$$C(x_1, x_2) = E\{x_1 x_2\}$$

$$C(x_1, x_2, x_3) = E\{x_1 x_2 x_3\} \tag{2.2}$$

$$C(x_1, x_2, x_3, x_4) = E\{x_1 x_2 x_3 x_4\} - E\{x_1 x_2\} E\{x_3 x_4\} -$$

$$E\{x_1 x_3\} E\{x_2 x_4\} - E\{x_1 x_4\} E\{x_2 x_3\}$$

From the above expressions for cumulants, one can derive the following properties:

- If $z_1, z_2, \ldots, z_n$ are Gaussian random variables, with $n > 2$, we have

$$C(z_1, z_2, \ldots, z_n) = 0 \tag{2.3}$$

- If we have two sets of independent random variables, $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$, then,

$$C(x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n) = C(x_1, x_2, \ldots, x_n) + C(y_1, y_2, \ldots, y_n) \tag{2.4}$$

- Based on (2.3) and (2.4) it is easy to verify that cumulants suppress Gaussian noise, i.e., if $z_1, z_2, \ldots, z_n$ are Gaussian random variables independent of $x_1, x_2, \ldots, x_n$ and $n > 2$, we have

$$C(x_1 + z_1, x_2 + z_2, \ldots, x_n + z_n) = C(x_1, x_2, \ldots, x_n) \tag{2.5}$$

In our work on speech processing we employed fourth-order cumulants since the third-order cumulants are blind to sources with symmetric probability density function. Although it is hard to propose an explicit density function for speech signals, a symmetric probability density function appears to be a reasonable one.

Linear-prediction is a popular method to analyze speech signals. This method assumes an autoregressive (AR) model for speech production

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) \; + u(n) \tag{2.6}$$

where $s(n)$ is the speech signal and $u(n)$ is the excitation sequence.

In [4] it is shown that the AR parameters ( i.e., the $a_k$'s ) satisfy the following equations,

$$\sum_{k=0}^{p} a_k \, C(s(n), s(n+k_0), s(n+m-k), s(n), \ldots, s(n)) \; = \; 0 \tag{2.7}$$

where $m > 0$, $a_0 = 1$ and $k_0$ is a parameter whose selection is addressed in [4]. Concatenating (2.7), for $m = 1, 2, \ldots, p+M$, where $M \geq 0$, we obtain the so-called cumulant-based normal equations,

$$\mathbf{C}(k_0)\mathbf{a} = \mathbf{0} \tag{2.8}$$

where $\mathbf{C}(k_0)$ is a Toeplitz matrix and $\mathbf{a} = col(1, a_1, a_2, \ldots, a_p)$. The AR parameters can be obtained by solving (2.8) if the matrix $\mathbf{C}(k_0)$ is full-rank. Note that, if the excitation sequence $u(n)$ is Gaussian and we employ higher than second-order cumulants, the identification problem becomes ill-posed. This observation is the starting point of our research on the application of cumulants to

speech processing. In the following chapter, we shall describe the speech production model in more detail.

# Chapter 3

# Speech Production Model

The state of speech signal belongs to three categories: voiced, unvoiced and silence. Silent periods can be detected easily by monitoring zero crossing rate and energy of the received signals [5]. For this reason, we shall concentrate on voiced/unvoiced classification of speech.

Unvoiced sounds are generated by forming a constriction at some point in the vocal tract and forcing air through the constriction at a high velocity to produce turbulence. This creates a broad spectrum noise source to excite the vocal tract. The energy concentration is shifted to the high-frequency end of the spectrum for unvoiced sounds, but the spectrum is relatively flat when compared with that of voiced speech. Due to large number of random effects involved in the production of unvoiced speech, Gaussian noise is a valid candidate as the excitation source. This assumption is validated by Wells [6]. In his work, the bispectrum is used to make V/UV decision. It has been found that bispectrum of English fricatives tend to zero, but for vowels the situation is just the opposite. A typical unvoiced segment of speech is shown in Figure 3.1a.

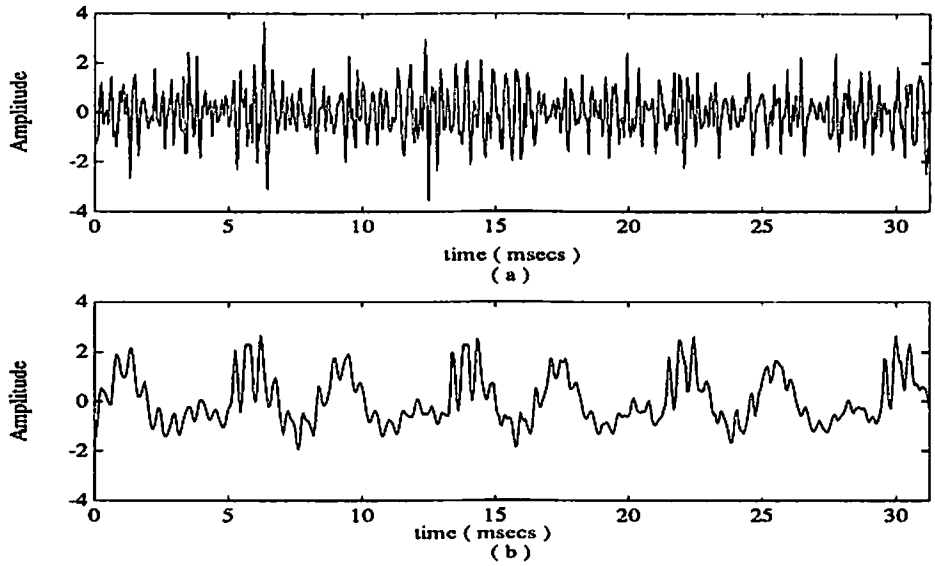


Figure 3.1: Typical speech signals: (a) Unvoiced speech, (b) Voiced speech.

Voiced sounds are produced by forcing air through the glottis with the tension of the vocal cords adjusted so that they vibrate in a relaxation oscillation, thereby producing quasi-periodic pulses of air which excite the vocal tract. This excitation is clearly non-Gaussian. The energy concentration is in the low-frequency side of the spectrum in the form of a fundamental component and its harmonics. In addition, voiced sounds have more energy than unvoiced sounds. A typical voiced speech segment is shown in Figure 3.1b.

For voiced sounds, the vocal tract can be modelled as an all-pole linear system. The same model also holds for unvoiced sounds but the AR order is less. Correlation between adjacent samples is high for voiced sounds. On the other hand, unvoiced speech resembles white noise since its spectrum is relatively flat, yielding small correlation between adjacent samples. Correlation sequences for voiced and unvoiced cases are illustrated in Figure 3.2.

The differences in the excitation and correlation properties for these two cases can be used to

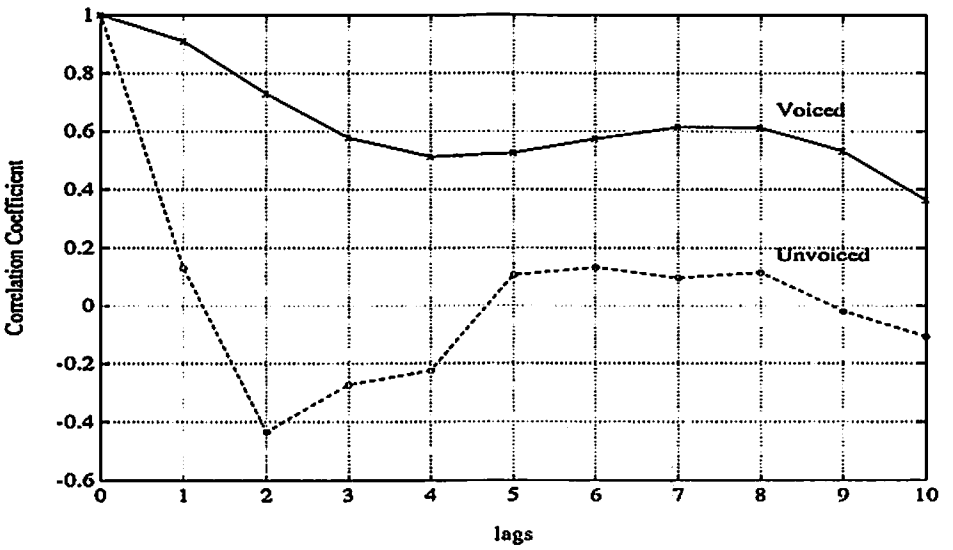


Figure 3.2: Adjacent sample correlation of speech signals

discriminate between them; however, with second-order statistics we can only use the correlation properties but can not utilize the information about the excitation model. This motivates the use of higher-order cumulants of speech signals.

# Chapter 4

# Our Approach

In the previous chapter, we mentioned the distinctions between voiced and unvoiced sounds: correlation among adjacent samples and excitation models. In this chapter, we shall investigate methods that fully utilize this information.

Linear prediction (LP) methods are employed to accomplish our goal; however, we shall not use batch-type methods for reasons outlined in the Introduction. Linear prediction can be based on second-or higher-order statistics, however the former is usually employed. Linear prediction is essentially identifying the inverse of a linear system driven by white noise; hence, it can be considered as a system identification problem. The system under consideration can be approximated by an AR model, so an FIR prediction filter will whiten the spectrum of the incoming signal. We shall investigate the differences between cumulant-and covariance-based adaptive prediction methods in this chapter.

## 4.1 Second-order statistics based adaptive filtering

Correlation-based adaptive prediction filters tend to minimize the prediction error power at the output of the filter. Since correlation among adjacent samples is high for voiced signals, we can remove a large proportion of energy from the original speech signal using prediction. On the other hand, in the case of unvoiced sounds, LP will not be that successful due to small correlation among samples. Therefore, a comparison of the input signal power with the power in the prediction residual will reveal the state of the speech signal.

Lattice prediction filters enable monitoring the variation of prediction error power with model order due to their specific structure. Autoregressive model-order-selection can be performed by selecting the tap which results in minimum prediction-error power. This leads to another discrimination between voiced and unvoiced sounds , since this order will be relatively lower for the unvoiced case.

## 4.2 Fourth-order statistics based adaptive filtering

In this section, we shall investigate the behavior of a fourth-order cumulant- based adaptive filter. An adaptive algorithm for estimating the parameters of nonstationary AR processes, excited by non-Gaussian signals is proposed in [7], and some modifications are suggested in [8]. We used the method of [7], which is in the software package $Hi - Spec^{TM}$ (trademark of United Signals and Systems, Inc.) [9]. The ideas for the covariance-based filter directly apply to this case with one important exception: the cumulant-based adaptive filter provides the solution to the cumulant-based normal equations, and this solution is not *the* one that minimizes the prediction- error power;

however, one may argue that if the speech production system can be identified accurately, then the prediction error should be close to the minimum possible value.

With higher-order statistics, we have the diversity of using the excitation information: for voiced sounds, the excitation is non-Gaussian; hence, the speech production mechanism can be identified by cumulant-based AR equations. On the other hand, for unvoiced sounds the excitation is Gaussian, *making the identification problem ill-posed*[1]. *The cumulant-based adaptive filter will not be able to identify the system and, since there is no associated output-power minimization criterion, prediction-error power may arbitrarily increase.* In this case, a cumulant-based filter may even amplify the speech signal making the power reduction by prediction comparison more clear than when using a covariance-based method.

To validate our ideas about covariance and cumulant-based adaptive prediction of speech signals, we performed some experiments using data from the TIMIT speech recognition database. The results verify our claims and are provided in the next chapter.

[1] A cumulant-based filter provides the solution of cumulant-based normal equations in an adaptive fashion; however, this set of equations becomes trivial when the input to be analyzed is a Gaussian linear process, because higher than second-order cumulants of Gaussian processes are zero.

# Chapter 5

# Experiments

## 5.1 Experiment I: Prediction performance of adaptive processors

We start our experiments by investigating the prediction performance of correlation-and cumulant-based linear predictors in voiced speech case. An indication of performance is the energy of prediction-error residual at the output of the filter. For this purpose, we selected a voiced speech segment from the TIMIT database and performed adaptive filtering based on both correlation and cumulants. We expected that the correlation-based filter would yield better performance, since it is designed to minimize prediction-error power. The original speech signal is scaled so that estimate of its variance is unity. The results of this experiment are shown in Figure 5.1. Energy values reported in this figure represent the estimate of the variance of the signal averaged over the data window. Interestingly enough, the cumulant-based filter performed better than its covariance counterpart,

although the latter is designed to minimize the power of the prediction residual. We repeated this experiment with other speech segments and in all of the cases, cumulant-based filter outperformed covariance-based filter.

In voiced speech, a conventional system identification approach for estimating the AR parameters, using a least-squares fit procedure, suffers due to the nature of the excitation sequence. It is known that, for voiced speech, the source is definitely non-Gaussian ; it is quasi-periodic in nature with spiky excitations. The impulsive nature of the excitation in voiced speech is exploited in [10], by making a Bernoulli-Gaussian assumption to develop a multipulse coding scheme. In [11] , a robust linear prediction algorithm is proposed which takes into account the non-Gaussian nature of source excitation for voiced speech by assuming the excitation is from a mixture distribution, such that a large portion of the excitation sequence is from a normal distribution with small variance while a small portion comes from an unknown distribution of higher variance. Such a distribution is called *heavy-tailed Gaussian.* Based on the above mixture model, a linear prediction algorithm is devised which employs robust statistical procedures (developed in [12]) that operate in a batch mode. Although satisfactory performance is observed, the method can not track the transitions in the input data. This points out a very important fact : conventional linear prediction can be unsatisfactory due to incorrect modelling of the excitation. Of course, this carries over to the adaptive domain, i.e., a correlation-based adaptive algorithm may not be able to yield the best possible fit in the presence of outliers in the data. On the other hand, a non-Gaussian excitation is required by higher-order-statistics-based identification algorithms. A cumulant-based adaptive filter is able to reduce the power in the signal by effective prediction, although it is not based on a

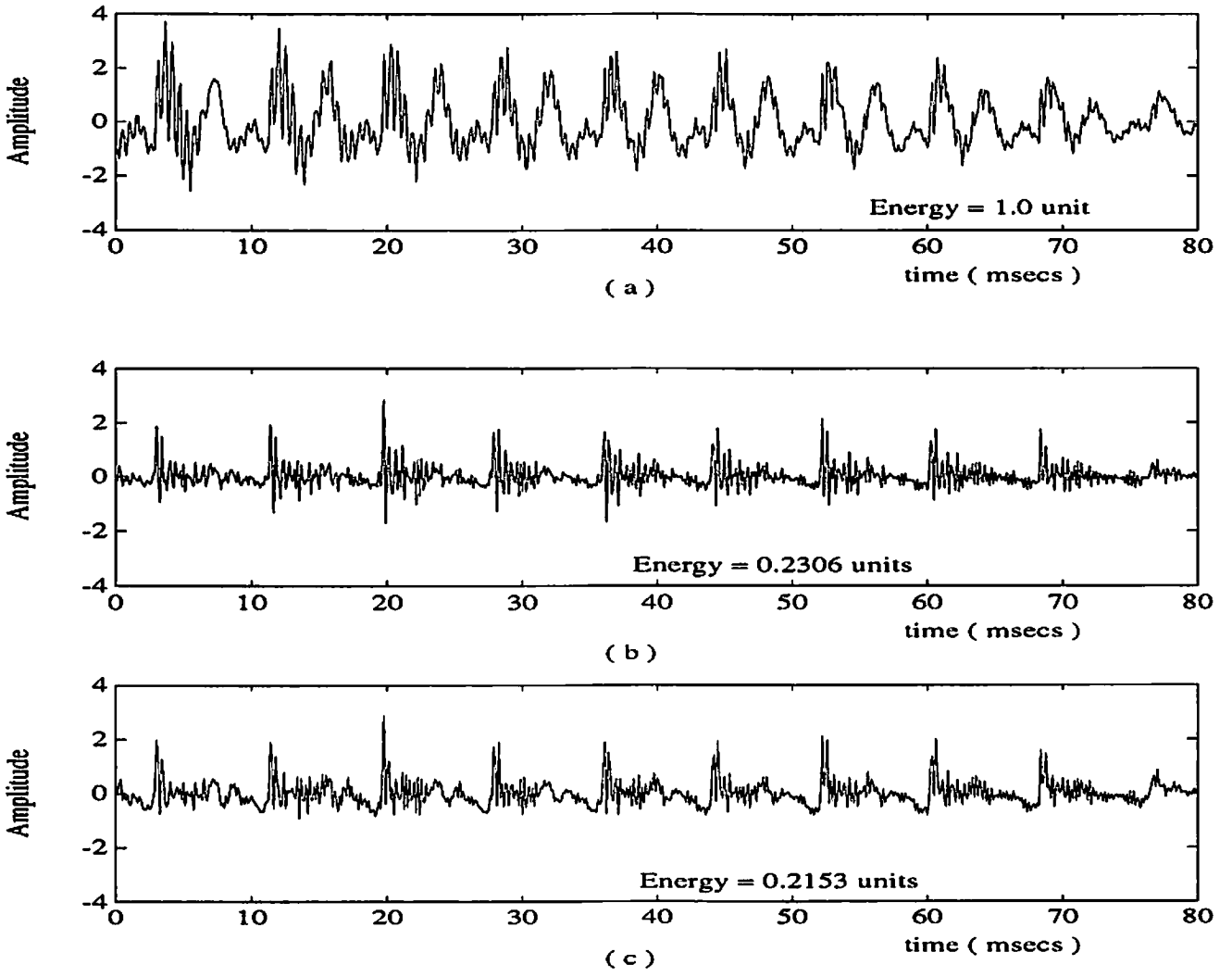


Figure 5.1: Energy comparisons. (a) Original speech signal; (b) prediction residual from covariance-based filter; and (c) prediction residual from cumulant-based filter.

criterion for minimizing the power of prediction residual. Power reduction may be even more than that provided by a covariance-based filter due to the just described outlier problem.

## 5.2 Experiment II: Behavior for voiced and unvoiced speech

To analyze the behavior of adaptive predictors in voiced and unvoiced speech states, we selected a 250 msec period of speech segment in which there are two transitions: voiced (0-75 msec), unvoiced (75-190 msec) and again voiced (190-250 msec). This signal is shown in Figure 5.2.

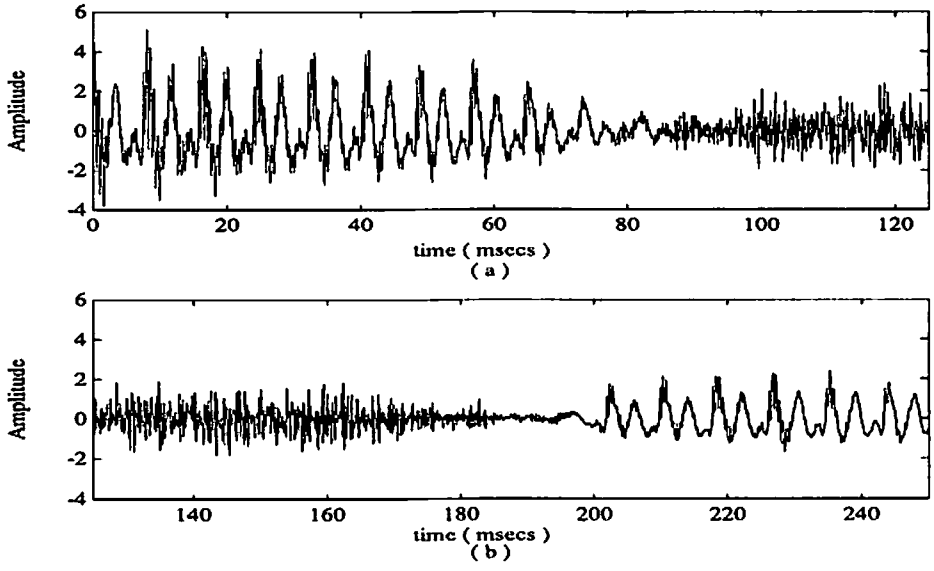


Figure 5.2: Speech signal to be used in the experiment: (a) first 125 msecs, (b) last 125 msecs.

We used an order ten predictor for adaptive filtering of the speech waveform. Figure 5.3 shows the prediction-error from a covariance-based filter. Observe that an adaptive filter based on a power minimization criterion will turn off during the unvoiced period; hence, this segment passes undistorted through the filter. The reason for this (as explained previously) is the small adjacent-

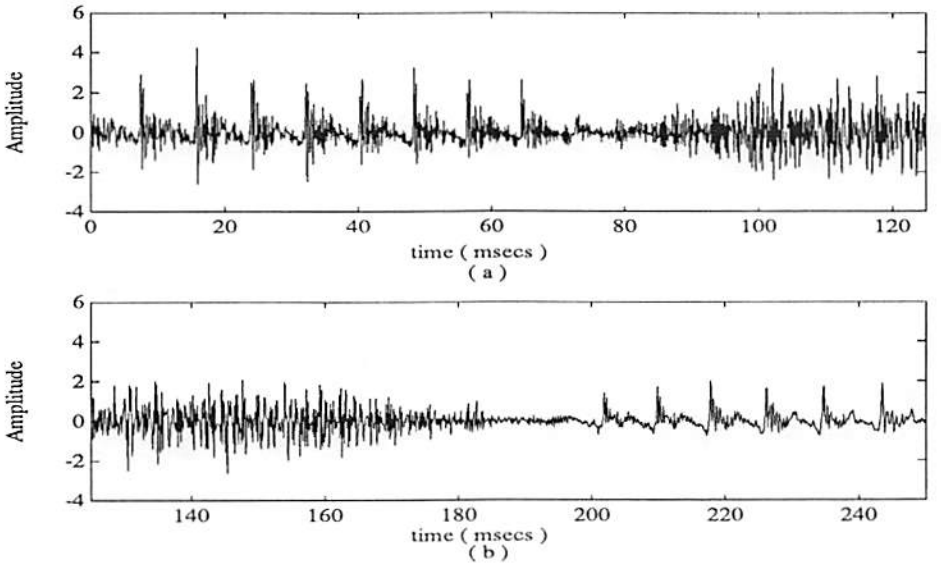


Figure 5.3: Prediction residual from covariance-based adaptive filter: (a) first 125 msecs, (b) last 125 msecs.

sample correlation for unvoiced sounds which makes the process unpredictable. To minimize the output power,the filter turns off; however, during voiced segments deconvolution is successful. We observe a quasi-periodic pulse train for the prediction residual, which is in accordance with the excitation model for voiced speech production.

Figure 5.4 depicts the cumulant-based filter residual. During voiced periods, successful deconvolution is possible since the excitation is non-Gaussian, and again a quasi-periodic pulse train is observed at the output of the filter. Now, however, the filter amplifies the speech signal during the unvoiced segment. As explained before, during this mode of operation, the system identification task is ill-posed, and, since this filter has no power minimization criterion, the power of the prediction residual becomes higher than the unvoiced speech signal.

To make better comparisons concerning the energy of the original speech and prediction resid-

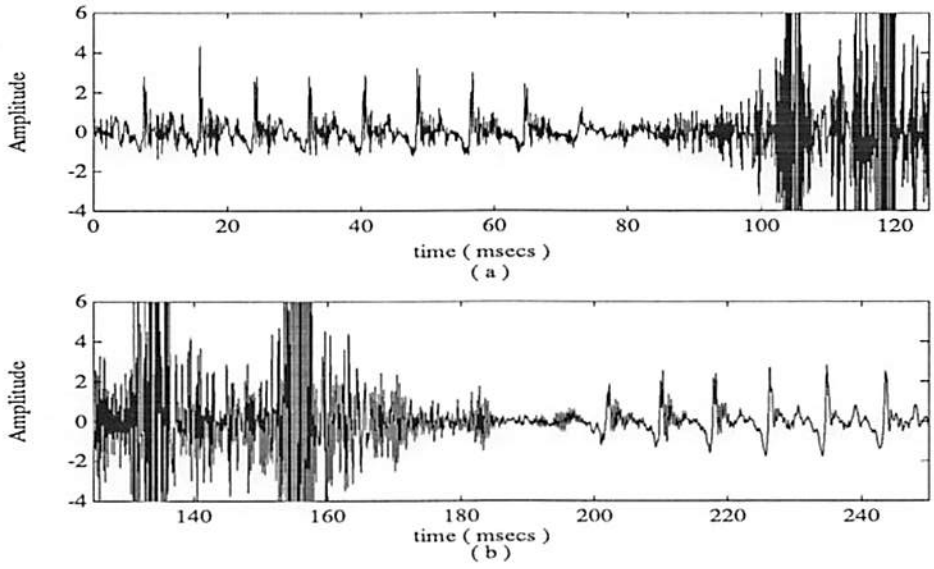


Figure 5.4: Prediction residual from cumulant-based adaptive filter: (a) first 125 msecs, (b) last 125 msecs.

uals, obtained via the two different filters, we illustrate the energy estimates in Figure 5.5. Energy is estimated by first squaring the signal and then performing low-pass filtering using a 15 point Hamming window. Figure 5.5 shows that, by comparing the prediction-residual power and the original-signal power, it is possible to make reliable V/UV decisions. With the cumulant-based method, even better results are obtained, because it amplifies the input data during unvoiced periods.

## 5.3 Experiment III: Pitch Prediction

The observations from the previous experiment validate our earlier statements; however, using a predictor may bring additional advantages as well. One important by-product is pitch period estimation. Pitch period is the time difference between the quasi-periodic excitation pulses during

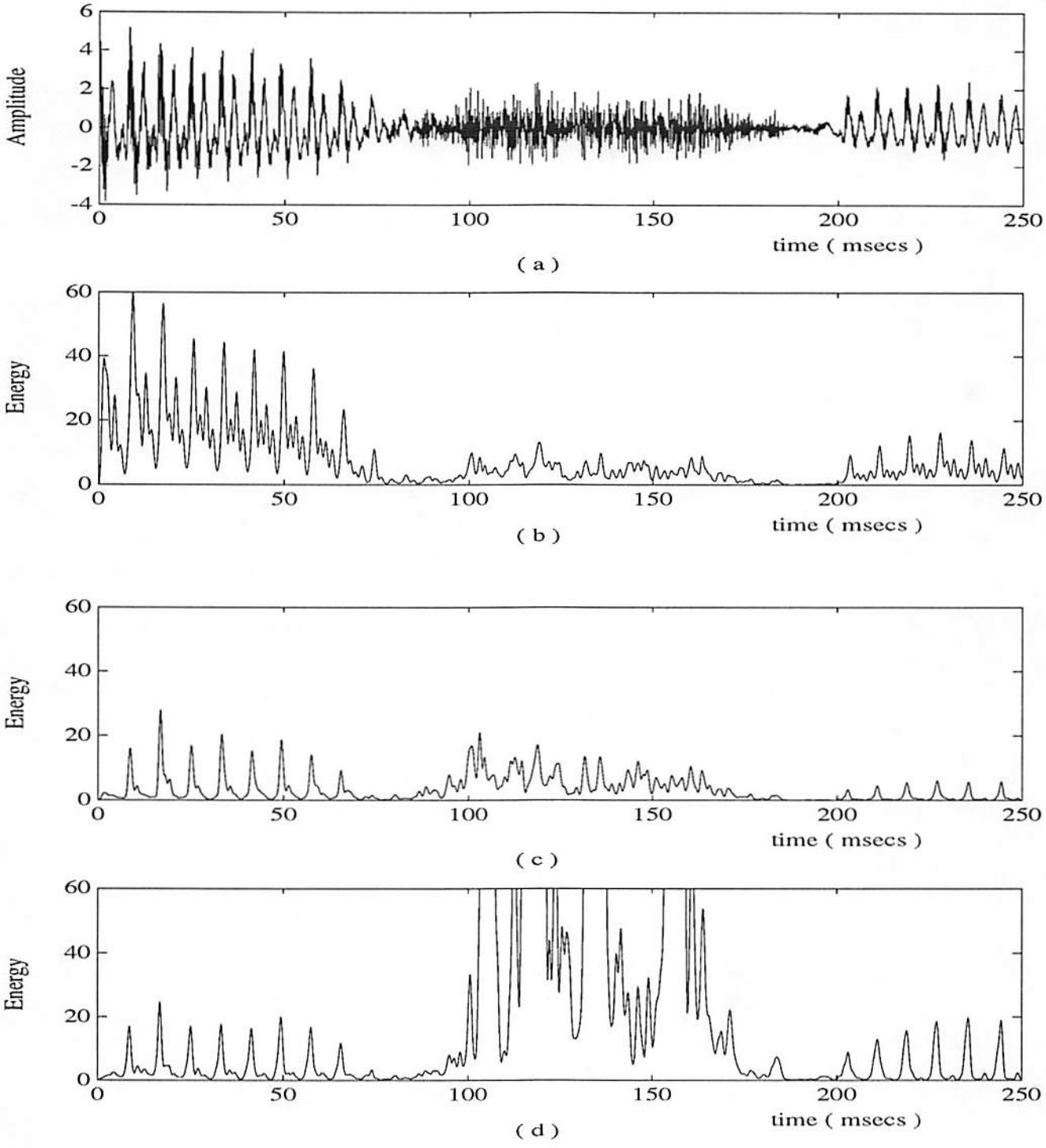


Figure 5.5: Energy estimates. (a) Original speech signal; (b) energy estimate of original speech signal; (c) energy estimate of prediction-error residual from covariance-based filter, (d) energy estimate of prediction-error residual from cumulant-based filter.

voiced speech. After the V/UV detection step, better pitch estimation is possible by operating on the energy estimate of prediction-residual rather than on the original speech signal. From Figure 5.5, we observe that the peaks in the energy estimate sequence are spaced by a pitch period during voiced periods and they are sharper than the ones in the original speech signal due to combined filtering and squaring operations. Consequently, we may apply the correlation-based approach described in [13] to the energy estimate sequence, for a reliable, simple but robust calculation of pitch period. In [13], pitch estimation is accomplished as follows: low-pass filtered speech signal is quantized to three levels; -1,0,1 and the correlation sequence of this quantized signal is obtained. Covariance calculation is simple with the quantized sequence, since it can be performed only by addition. Finally, a peak-picking method estimates the pitch period. Peak-search is performed on the possible range of values that pitch-period can take, which is called the admissible pitch range. We applied this method to the energy estimate of prediction-residual from the cumulant-based predictor that processes the speech segment in Figure 5.2. Since the energy estimates are non-negative, they are quantized to two levels. The original signal and pitch estimates are given in Figure 5.6.

The decisions and estimates agree with the signal characteristics. Results from the correlation-based filter are also accurate for this speech segment; however, the accuracy of the correlation-based method depends more on the threshold employed in comparing the power of prediction residual to that of the input, than in the cumulant-based counterpart, since the latter amplifies unvoiced speech. Therefore, we can observe degradation in the correlation-based case since it is sensitive to the value of the threshold.

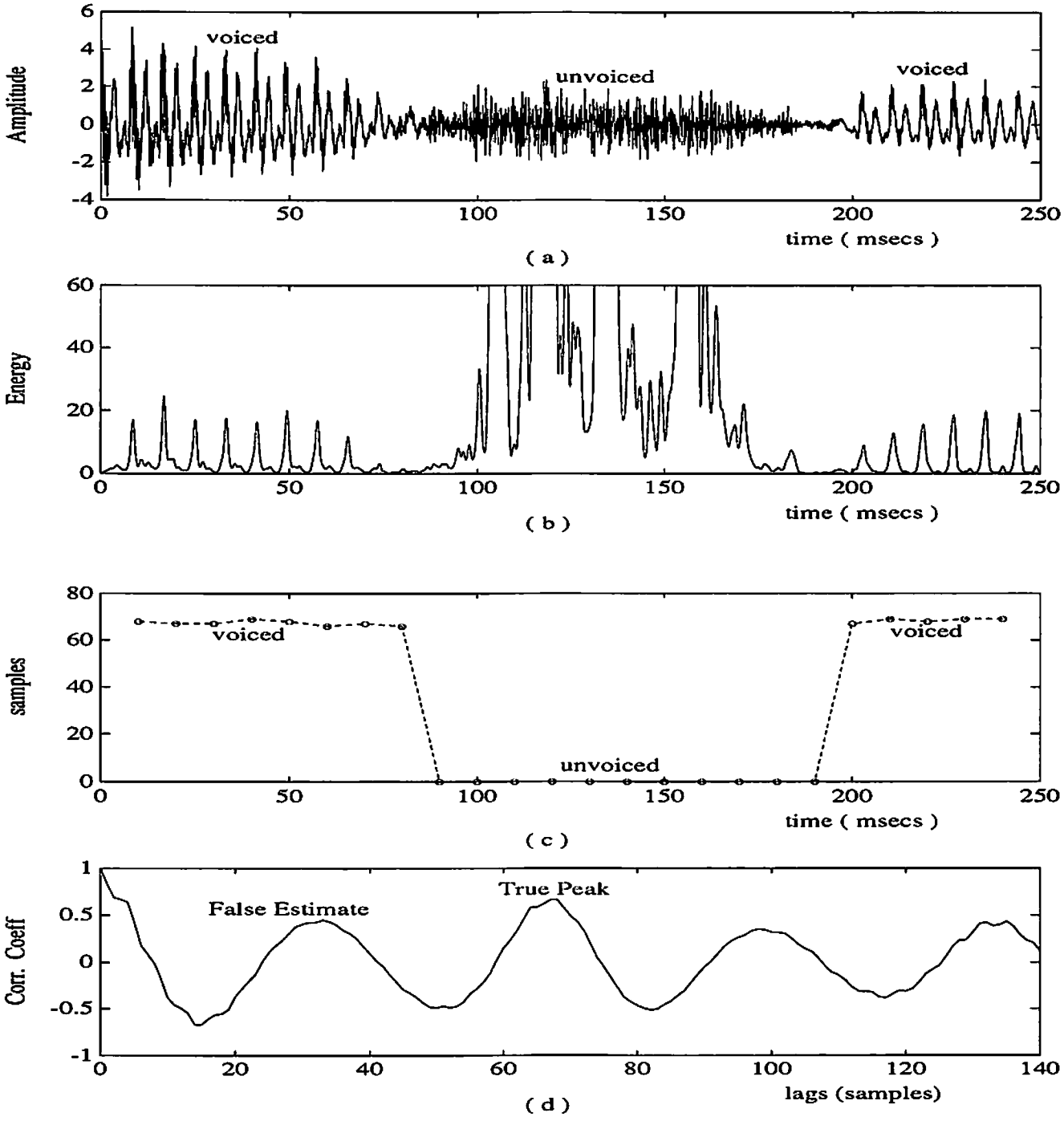


Figure 5.6: Pitch-period estimation experiment. (a) Original speech signal; (b) energy estimate of prediction-error residual from cumulant-based filter; (c) pitch contour obtained by processing energy-estimate sequence using the method in [13]; and (d) autocorrelation sequence of the second voiced speech segment processed by the method in [14], leading to a gross error.

The second voiced speech segment in Figure 5.2 is an example of the situation when harmonics are stronger than the fundamental frequency component. In general, correlation-based approaches operating directly on the speech signal fail when this event is present. To demonstrate this, we implemented the method described in [14]. In [14] pitch estimation is accomplished by calculating the correlation sequence of the low-pass filtered speech signal, and employing a peak-picking algorithm on the correlation sequence. Peak-searching is done on the admissible pitch range. For reliability purposes, the algorithm also investigates the possibility of pitch errors, by checking for peaks at one-half, one-third, one-fourth, one-fifth, and one-sixth of the first estimate of the pitch period, if they are in the admissible pitch range. If a peak at these locations is larger in amplitude than half of that of the current estimate, the pitch estimate is changed to the location of this peak. In our experiment, the pitch detector of [14] locates the major peak at lag 68; however, its decision rule identifies another peak around lag 34 which is in the admissible pitch range. Since the amplitude of the peak at lag 34 is larger than half of that of the major peak, the final pitch estimate is chosen to be half of the correct value, which is a gross error.

# Chapter 6

# Conclusions

In this work, we showed that it is possible to track transitions in the state of speech using adaptive linear prediction. Both covariance and cumulant-based methods are investigated, and greater contrast between V/UV cases is demonstrated by the latter method because cumulants can use the difference in the excitation model of the two speech states.

Pitch-period estimation is also possible by linear prediction. Rather than operating on the original signal, we prefer to employ the prediction-error residual available from an adaptive filter. Cumulant-based approach operating on the power estimate of the residual process is shown to be a practical way of pitch estimation.

We investigated the prediction performance of adaptive predictors based on correlation and cumulants and found that cumulant-based prediction can outperform correlation-based prediction, although the latter is designed to minimize the power of the prediction residual. We conjectured that outliers in the excitation model of voiced speech result in this phenomena. Better predic-

tion performance obtained via cumulants is worth investigating analytically; however, this is not tractable with real or synthesized speech since there are many parameters involved. Simpler cases, such as a single sinusoid in Gaussian noise can be analyzed to evaluate the performance of cumulant and covariance-based adaptive-line-enhancers.

# Chapter 7

# Acknowledgement

The authors wish to acknowledge support of this research from Rockwell Science Center.

# Bibliography


[1] A. Bendiksen and K. Steiglitz, " Neural Nets for Voiced/Unvoiced Speech Classification," *in Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp.521–524, 1990.

[2] A. Benveniste, M. Metivier and P. Priouret, *Adaptive Algorithms and Stochastic Approximations*, Springer-Verlag, 1990.

[3] A. Johanson, G. Ahlbom and L.H. Zetterberg, "Event detection using recursively updated lattice filters,"*in Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp.632–635, 1985.

[4] J.M. Mendel, "Tutorial on higher-order statistics (spectra) in signal processing and system theory: theoretical results and some applications," *in Proc. IEEE*, vol.79, no.3, pp.278–305, March 1991.

[5] L. Rabiner and R. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, 1978.

[6] B. Wells, "Voiced/Unvoiced decision based on the bispectrum," *in Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing* , pp. 1589–1592, 1985.

[7] A. Swami and J.M. Mendel, " Adaptive system identification using cumulants," *in Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp.2248–2251, 1988.

[8] J.R. Fonollosa, J. Vidal and E. Masgrau , "Adaptive system identification based on higher order statistics," *in Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp.3437–3440, 1991.

[9] $Hi-Spec^{TM}$: Software package for signal processing with higher-order-spectra, United Signals and Systems, Culver City, California, 1991.

[10] K.Y. Lee, B.G. Lee, I. Song and S. Ann, "On Bernoulli-Gaussian modelling of speech excitation source," *in Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp.217–220, 1990.

[11] C. Lee, "Robust linear prediction of speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-36 , no.5, pp.642–650, May 1988.

[12] P. Huber, *Robust Statistical Procedures* , CBMS - NSF Regional Conference Series in Applied Mathematics, 1977.

[13] Dubnowski, R.W. Schafer and L.R. Rabiner , " Real-time digital hardware pitch detector," *IEEE Trans. Acoust. , Speech , Signal Process.* , vol. ASSP-24, no.1, pp.2–8, February 1976.

[14] D.A. Krubsack and R.J. Niederjohn, " An autocorrelation pitch detector and voicing decision with confidence measures developed for noise corrupted speech," *IEEE Trans. Acoust., Speech, Signal Process.* , vol. ASSP-39, no.2, pp. 319-329, February 1991.