Speech Enhancement Based on Spectral Estimation from Higher-lag Autocorrelation

Benjamin J. Shannon, Kuldip K. Paliwal and Climent Nadeu[†]

School of Engineering, Griffith University
Brisbane, QLD 4111, Australia
Ben.Shannon@student.griffith.edu.au, K.Paliwal@griffith.edu.au, climent@talp.upc.es

Abstract

In this paper, we propose a unique approach to enhance speech signals that have been corrupted by non-stationary noises. This approach is not based on a spectral subtraction algorithm, but on an algorithm that separates the speech signal and noise signal contributions in the autocorrelation domain. We call this technique the AR-HASE speech enhancement algorithm.

In this initial study, we evaluate the performance of the new algorithm using the average PESQ score computed from 10 male utterances and 10 female utterances taken from the TIMIT database as a measure of speech quality. We test the algorithm using one broadband stationary noise and two non-stationary noises. We will show that the AR-HASE enhancement algorithm produces near transparent quality for clean speech, gives poor enhancement performance for broadband stationary noises, and gives significantly enhanced quality for the two non-stationary noises.

Index Terms: speech enhancement, autocorrelation, impulsive noise.

1. Introduction

Many of the state-of-the-art speech enhancement algorithms use the analysis-modification-synthesis framework [1] in their operation. In this framework, the corrupted speech signal is broken up into short-time segments, which are transformed to the frequency domain where only the spectral magnitude is modified. The speech signal is then reconstructed with an inverse short-time Fourier transform followed by an overlap-add operation. This structure is used by the popular spectral subtraction algorithm, originally proposed by Boll [2] in 1979, and also by techniques related to Wiener filtering, such as Ephraim-Malah's method [3] and all its more recent variants.

These spectral enhancement algorithms require an estimate of the noise spectrum, which can be obtained from non-speech segments indicated by a voice activity detector or, alternatively, with a minimum statistics approach [4], i.e. by tracking spectral minima in each frequency band. In consequence, they are effective only when the noise signals are stationary or at least do not show rapidly varying statistical characteristics. The worst type of noise for these systems is when the noise signal is typically coincident with the speech signal, and absent at other times. This situation, for example, could arise with an impulsive noise. In this case, most of the non-speech frames could be completely devoid of impulsive noise, but the speech frames could contain a large amount of this noise. To handle these situations, noise reduction techniques that operate intra-frame (within the current frame) are required; these techniques cannot use the noise power spectrum estimate from other non-speech frames.

In previous work, we have proposed a noise robust spectral estimation technique for short-time speech signals that operates intra-frame. This method uses the periodic correlation property of short-time speech signals and the autocorrelation domain to perform noise reduction. It is well known that the pitch period of human speech is typically constrained to values between 2 ms and 12 ms. This means that in the autocorrelation domain, we will have large magnitude coefficients at these periods. This property, conversely, is generally not true for noise signals. By computing a spectral estimate using only the higher-lag autocorrelation coefficients, we have a way of separating the speech and noise signal without having to estimate the noise signal directly. We call this method, Higher-lag Autocorrelation Spectral Estimation (HASE) [5] [6].

The HASE method was motivated by the large volume of previous work on noise robust Automatic Speech Recognition ASR feature extraction based on autocorrelation domain processing [7] [8] [9] [10]. This method has been successfully applied to the noise robust ASR problem, particularly where the noise signal had rapidly changing characteristics. The goal of ASR feature extraction is to produce features that have a low dimensionality, are insensitive to speaker and environmental changes and are effective in discriminating the linguistic units. These goals have little in common with the goals of speech enhancement.

In this paper, we investigate the HASE algorithm for speech enhancement. We show that this algorithm has some inherent limitations for enhancement applications. We propose to overcome these limitations by using an Auto-Regressive (AR) model of high order. We refer to this extended HASE algorithm as the AR-HASE algorithm. It is our aim in this work to explore the potential of this technique for the enhancement of speech signals corrupted by both stationary and non-stationary disturbances.

2. Speech Enhancement using Higher-lag Autocorrelation Spectral Estimation

A brief description of the previously proposed Higher-lag Autocorrelation Spectral Estimation (HASE) technique proceeds as follows. The short-time speech segment (approx. 32 ms) is first

[†] This work was performed while Climent Nadeu was on leave from the Signal Theory and Communications Department, Universitat Politecnica de Catalunya, 08034 Barcelona, Spain.

windowed using a Hamming window. Following this, a biased estimate of the autocorrelation sequence is made. Once the autocorrelation sequence is computed, the higher-lag range (2 ms to 32 ms) of one-side of the autocorrelation sequence is windowed using a high dynamic range window function. The Double Dynamic Range (DDR) window function design method [5] is used to compute this window. The magnitude spectrum of the windowed higher-lag autocorrelation sequence is then computed as an estimate of the short-time power spectral density.

The speech enhancement framework that we first used to evaluate the performance of the HASE algorithm for speech enhancement is shown in Fig.1. Here, we have taken the typical spectral subtraction algorithm and modified it. We have substituted the enhanced short-time power spectrum estimate in the spectral subtraction framework with the power spectral estimate computed using the HASE algorithm.

As mentioned previously, the spectral subtraction algorithm requires an estimate of the noise power spectrum. In the proposed framework, this estimate is not required. The speech signal enhancement is performed based on prior knowledge of the autocorrelation sequences of typical speech and noise signals. Speech signals (particularly voiced) have autocorrelation sequences with large magnitude coefficients at higher-lag values. This property is not typically observed in noise signals. Therefore, by using only the higher-lag portion of the autocorrelation sequence to compute a spectral estimate, the noise contribution is reduced.

The first problem we encountered in applying the HASE algorithm in this framework is the Fourier phase spectrum and the HASE magnitude spectrum are not well matched. To achieve good results in the synthesis stage, the pitch harmonic features in the phase spectrum and magnitude spectrum need to match well. This problem is demonstrated in the analysis shown in Fig.2. This figure shows the Fourier power spectrum of a 32 ms frame containing an /iy/ sound (plot (a) dashed line) and the group delay sequence computed from the Fourier phase spectrum (plot (b)). Wherever a pitch harmonic is present in the Fourier power spectrum, the corresponding group delay sequence shows a near constant value. However, the low power regions between the pitch harmonics give spurious values in the group delay sequence. Figure 2(a) also shows the HASE spectral estimate (solid line) for the same frame. Due to the extra windowing steps in the HASE algorithm, the bandwidths of the pitch harmonics are larger than in the direct case. This means that in the synthesis stage, relatively high magnitude spectral coefficients are matched with spurious phase coefficients. This is the cause of noticeable distortion in the output speech.

There are several ways to reduce the problem of pitch harmonic bandwidth mismatch between the magnitude and phase spectrum. The approach we have chosen for this study is to increase the number of samples used in estimating the magnitude spectrum. For example, in the case of the HASE algorithm, a 32 ms frame is processed. This allows a one-sided biased autocorrelation sequence to be computed that has a lag range of up to 32 ms. To reduce the bandwidth of the pitch harmonics, we need a one-sided autocorrelation sequence with a lag range greater than 32 ms. Extension of this sequence can be achieved with the aid of Auto-Regressive (AR) modelling. Rather than computing the biased estimate of the autocorrelation sequence using the FFT algorithm, we propose to compute it as the inverse Fourier transform of a high order AR power spectral estimate, thus extending the non-zero lag range beyond 32 ms. This approach also provides a further degree of freedom. By manipulating the order of the AR model, we can tune the performance of the enhancement algorithm. A brief evaluation of the proposed HASE and AR-HASE based enhancement algorithm are now presented.

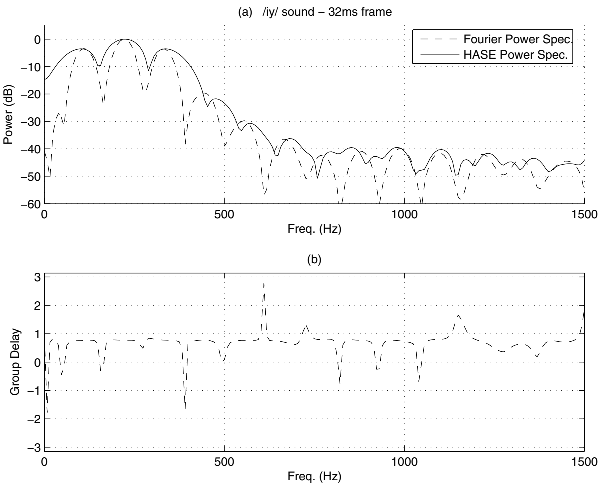


Figure 2: Comparison of Fourier power spectrum and the HASE power spectrum. (a) Power spectrum estimate of a 32 ms frame containing an /iy/ sound using Fourier transform (dashed line) and the HASE algorithm (solid line). (b) Group delay sequence computed from the Fourier phase spectrum.

3. Experimental Evaluation

In this section, we evaluate the performance of both the HASE and the AR-HASE algorithm. We first explore the performance of the HASE algorithm in clean conditions to determine how significant the pitch pulse bandwidth mismatch problem discussed in section 2 is. We then go on and test the enhancement potential of the AR-HASE algorithm using three types of noise. One of the noises is a stationary type noise and the other two are non-stationary.

To evaluate the performance of the proposed speech enhancement algorithms, we took 20 speech files from the TIMIT database and down-sampled them to a sampling frequency of 8 kHz. The 20 utterances came from 10 different male and 10 different female speakers. Using these 20 samples, the average PESQ [11] score was computed as a measure of performance. PESQ stands for ”Perceptual Evaluation of Speech Quality”. This algorithm was designed to provide a way to estimate the subjective quality of speech. The output from the algorithm is an estimate of the Mean Opinion Score (MOS), which is a number between 1 and 5. The meanings assigned to the scores in relation to the speech quality are: 1-Bad 2-Poor 3-Fair 4-Good 5-Excellent.

The three noise samples used in the evaluation are theoretically ideal for the HASE algorithm. That is, for an analysis frame size of 32 ms, the theoretical autocorrelation sequence has high magnitude coefficients for time lags between 0 and 2 ms and zero value coefficients for time lags greater than 2 ms. These three noises are white Gaussian noise, repeating impulse noise and repeating chirp noise.

The three noises were created using the following steps. The artificial white noise was obtained using a Gaussian random number generator. To create the artificial impulsive noise, we first be-

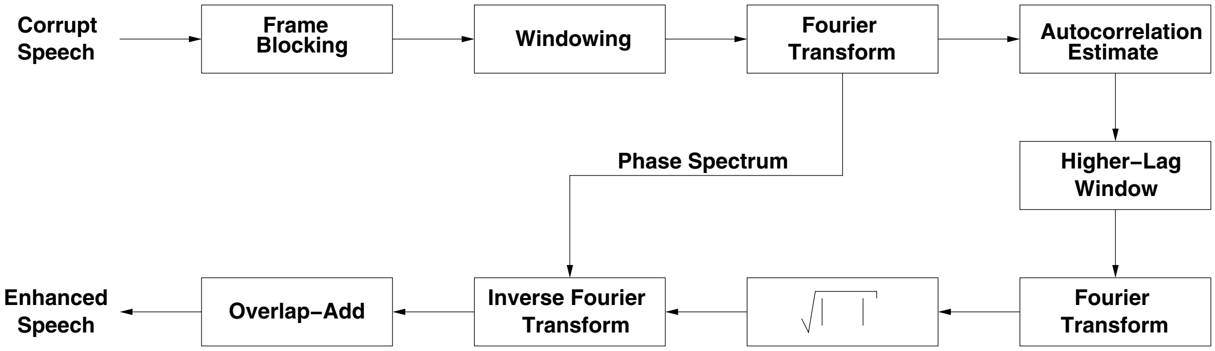


Figure 1: Block diagram of the proposed AR-HASE based speech enhancement algorithm.

Model Order	Male	Female	Combined
32	2.73	2.50	2.61
64	3.29	3.23	3.26
96	3.25	3.85	**3.55**
128	3.13	3.83	3.48
160	3.11	3.79	3.45
192	3.06	3.72	3.39
224	2.97	3.65	3.31
255	2.89	3.59	3.24

Table 1: Mean PESQ scores of AR-HASE algorithm with different AR model orders tested on clean speech.

gan with a 32 ms block of zeros. To this block, we added a unit pulse of 2 ms duration. The starting position of the 2 ms pulse was randomly selected between 0 and 30 ms using a uniform random number generator. We then concatenated this block with another 32 ms block that contained only zeros. These two steps were then repeated, but this time the sign of the 2 ms pulse was reversed to maintain zero mean. These four steps were then repeated continuously to get a sufficiently long sequence of the impulsive noise. Thus, for this noise, the separation between successive pulses randomly varies between 32 to 92 ms. Finally, the artificial chirp noise was created by defining one period of the chirp as a sinusoidal signal whose frequency changes linearly from 0 kHz to 4 kHz (half of the sampling frequency) over a period of 32 ms. This period was then repeated to give a sequence of sufficient length.

3.1. HASE enhancement

Using the HASE algorithm in the proposed modified analysis - modification - synthesis speech enhancement framework gave a mean PESQ score of 2.85 for clean speech. This is considered a low score for clean speech. As expected, distortion was also noted during listening.

3.2. AR-HASE enhancement

The first evaluation of the AR-HASE algorithm is performed on clean speech for different AR model orders. A high model order is expected to give better performance; therefore, we start at a model order of 32 and increase it by 32 until all the frame data is used in the AR modelling. These results are shown in Table 1.

Since a model order of 96 gave the best performance in clean conditions, this model order is used in the enhancement evaluation. The results comparing the AR(96)-HASE enhanced speech with unenhanced speech is given in Fig.3.

4. Discussion

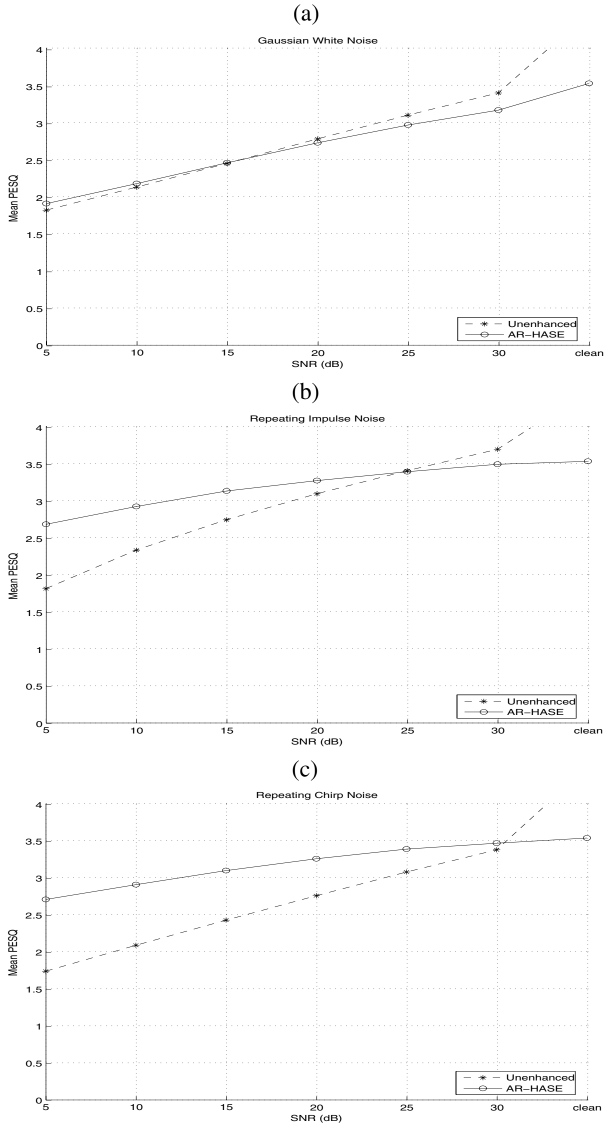
When we apply the HASE algorithm to clean speech utterances and listen to the HASE enhanced utterances, the speech is easily understood, but it sounds like the speakers pitch has been distorted. We get an average PESQ score of 2.85 which is approximately equivalent in speech quality to speech corrupted with white Gaussian noise at a global SNR of 20 dB. Thus, the HASE enhancement algorithm reduces the speech quality significantly for clean speech signals. Therefore, we have disregarded this algorithm.

When we apply the AR-HASE algorithm to clean speech signals and investigate its performance as a function of AR model order, the peak in speech quality occurs at a model order of 96. To compute the AR model of order 96, autocorrelation lags up to 12 ms are used. This is sufficient to cover the pitch period of most human speakers. For example, if we take a voiced speech frame from a speaker that has a pitch of 100 Hz, then compute a Fourier spectrum from 0 to 4 kHz, we expect to see 40 peaks. To make an AR spectrum match well with each of the 40 peaks in the Fourier spectrum, we would require a minimum of 80 poles in the AR model. Therefore, an order of 96 makes intuitive sense.

The AR-HASE algorithm is nearly transparent for clean speech. Where there is noticeable distortion, it sounds more like a reverberant distortion than an additive background distortion. The average PESQ score for clean speech was 3.55. This was equivalent to a speech quality of speech corrupted with white Gaussian noise at a global SNR $>$30 dB.

The enhancement properties of the AR-HASE algorithm were dependent on the corrupting additive noise. For the broadband white Gaussian noise, no enhancement in quality was achieved using a model order of 96. We attribute the poor performance for this case to the estimate of the short-time autocorrelation sequence. Over a short analysis frame, the autocorrelation estimate of a white broadband noise is far from the asymptotic estimate. In informal testing, it was found that by using a very low model order (12-24), the white Gaussian noise could be eliminated from the speech, but this was at the expense of significant speech distortion.

The AR-HASE algorithm worked very well for the non-stationary noises. For the impulsive noise and repeating chirp noise at 5 dB SNR, the average PESQ scores were 0.87 and 0.97


(a)

(b)

(c)

Figure 3: PESQ performance of AR-HASE speech enhancement compared to unenhanced speech. (a) White Gaussian noise. (b) Repeating impulse noise. (c) Repeating chirp noise.

higher than the unenhanced scores respectively. This is equivalent to a listener's opinion of the speech quality moving from poor to fair.

Since the AR-HASE algorithm gave good enhancement performance for the non-stationary noises and poor performance for the broadband stationary noise, it could be possible to get better performance by combining this algorithm with an existing enhancement algorithm such as spectral subtraction. For this type of approach, the contributions from both algorithms may be complimentary. That is, if we use the spectral subtraction and the AR-HASE algorithm in cascade, the spectral subtraction algorithm could remove the stationary noise, and the following AR-HASE algorithm could reduce any residual non-stationary noise.

5. Conclusion

In this paper, we have proposed a new approach to the enhancement of speech signals that have been corrupted by non-stationary, additive and uncorrelated noise signals. This approach was not based on a spectral subtraction algorithm, but on an algorithm that separates the speech signal and noise signal contributions in the autocorrelation domain. This technique was called the AR-HASE algorithm.

The AR-HASE algorithm was first tested on clean speech signals. It was shown that after choosing an appropriate AR model order, near transparent quality could be achieved for clean speech. The algorithm was then tested on three types of noise signals using the average PESQ score as a speech quality measure.

For broadband stationary noise, little enhancement of the speech quality was gained using the AR-HASE algorithm. For the other two noises tested, repeating chirp and impulsive noise, a large improvement in speech quality was measured.

6. References

[1] J. B Allen, "Short Term Spectral Analysis, Synthesis, and Modification by Discrete Fourier Transform," *IEEE Trans. ASSP*, vol. 25, no. 3, pp. 235–238, 1977.

[2] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *Acoustics, Speech and Signal Processing*, vol. ASSP-27, no. 2, pp. 113–120, 1979.

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

[4] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. SAP*, vol. 9, no. 5, pp. 504–512, 2001.

[5] B. J. Shannon and K. K. Paliwal, "MFCC Computation from Magnitude Spectrum of Higher Lag Autocorrelation Coefficients for Robust Speech Recognition," in *Proc. ICSLP*, 2004.

[6] B. J. Shannon and K. K. Paliwal, "Spectral Estimation using Higher-lag Autocorrelation Coefficients with Applications to Speech Recognition," in *Proc. ISSPA*, 2005, pp. 599–602.

[7] J. A. Cadzow, "Spectral Estimation: An overdetermined rational model equation approach," in *Proc. IEEE*, 1982, vol. 70, pp. 907–939.

[8] Y. T. Chan and R. P. Langford, "Spectral Estimation via the High-Order Yule-Walker Equations," *IEEE Trans. on ASSP*, , no. 5, pp. 689–698, 1982.

[9] D. Mansour and B. H. Juang, "The Short-Time Modified Coherence Representation and Noisy Speech Recognition," *IEEE Transactions on ASSP*, vol. 37, no. 6, pp. 795–804, 1989.

[10] J. Hernando and C. Nadeu, "Linear Prediction of the One-Sided Autocorrelation Sequence for Noisy Speech Recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 1, pp. 80–84, 1997.

[11] A. W Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, 2001, vol. 2, pp. 749–752.