International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-2 Issue-5, June 2013

Speech Enhancement Using DCT

S.C.Shekokar, M. B. Mali

Abstract- The speech enhancement problem comprises of various problems characterized by the type of noise source, the nature of interaction between speech and noise, the number of sensor signals (microphone outputs) available for enhancement and the nature of the speech application. Noise reduction remains a demanding problem due to wide variety of background noise types (car noise, babble noise, cockpit noise, train noise, subway noise, etc.) and the difficulty in estimating their statistics. The connection of noise and clean signal is usually classified as additive/multiplicative/convolution. The additive model very often dominates in real-world applications.

Keywords–Speech enhancement, Speech processing

I. INTRODUCTION

Whenever speech is recorded by a microphone, unwanted noise is also recorded. This noise depends on the environment and can range from anything such as computer fan noise, car engine noise to factory floor noise. The goal of any speech enhancement system is to suppress or completely remove the unwanted noise while maintaining the quality and/or intelligibility of the speech. The Fig.1 shows basic overview of speech system. It consists of speech and noise, adding this we will get noisy speech.

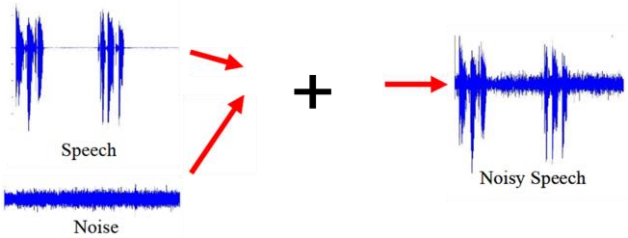

Fig. 1 Basic overview of Additive System [6].

Fig.2 shows speech enhancement algorithm, after performing enhancement algorithm, it will produce clean speech.

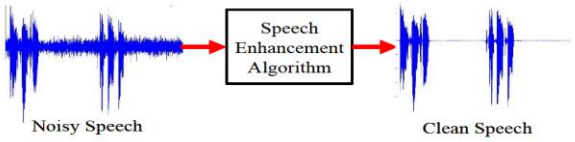

Fig. 2 Basic overview of speech enhancement system [6].

Speech enhancement can be performed in the time domain as well as in the frequency domain. Time domain filters include Finite Impulse Response (FIR) and Infinite Impulse Response (IIR) filters, Linear Predictive Coefficients (LPC) and Hidden Markov Model (HMM) etc.

Manuscript published on 30 June 2013.
* Correspondence Author (s)
S.C.Shekokar, Sinhgad College of Engineering,Vadgaon(Bk), Off Sinhgad Road Pune, India
Prof. M. B. Mali, Department of Electronics and Telecommunication, University of Pune, India

In transform domain techniques first transformation is performed on noisy speech before filtering then the inverse transformation take to give original speech. In Transform domain filters are those which calculate the transform coefficients first followed by the enhancement process. At the end, the inverse transform must be applied to obtain the enhanced speech. The main advantage of noise filtering process is relative ease of distinguishing and removing noise from the speech. Many speech enhancement algorithms prefer operating in the transform domain since the speech energy is not present in all the transform coefficients and it is therefore easier to filter off the noise especially for the noise-only coefficients. Different types of transforms may require different type's analysis methods.
The use of speech processing systems for voices communication and recognition tasks is becoming more and more common because of increasing power and falling cost of digital signal processors and the availability of cheap memory chips. One outstanding example of a voice communication product is cellular radio telephony system. Numerous examples of voice recognition products include hands-free input system for voice dialing, voice activated security systems, and etc. speaking is arguably a more natural way of communicating with machine than typing. Due to its accurate recognition, it is more efficient and faster. As the presence of noise significantly degrade the performance of speech coders and voice recognition systems, it is therefore imperative to incorporate speech enhancement as preprocessing step in these systems.
Noise can be defined as unwanted signal and there are many forms of noise. One of the most common sources of noise is background noise which is always present in different degrees in any location apart from a soundproof room. Operating a hands-free mobile phone in a car can be affected by at least 3 types of background noises, namely wind, road as well as engine noise. Other examples of noisy speech inputs are found in pay phones in noisy environments such as food courts and bus terminals, voice communication systems in cockpits, cellular phones in machine rooms, etc. A second source of noise is channel noise which affects both digital and analogue transmissions and therefore degrades the resulting speech at the receiver end. Different types of noise require different noise models and their own unique set of solutions. The scope of the speech enhancement explored in this project is focused on suppression of background noise.

II. METHODOLOGY

The new speech enhancement technique is introduced an adaptive time-shift analysis speech (ATSA) enhancement system. This proposed technique work on the pitch period of speech. The proposed technique used in no. of applications.

1) Speech Enhancement System with Pitch Synchronous Analysis: This method of pitch synchronous analysis shifts the analysis window by the pitch period to obtain the speech segments and it will theoretically produce constant DCT coefficients for stationary signals. This is valid for only voiced speech signal, for unvoiced/silence part of speech signal is lags. Despite this, significant improvements can still be obtained as voiced speech is more dominant than unvoiced speech.

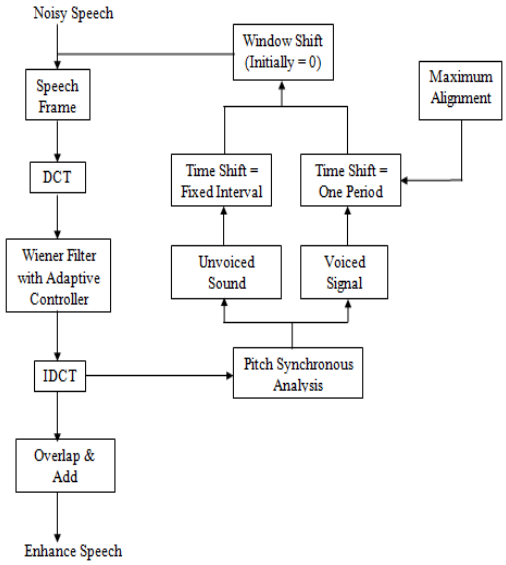

Fig.3. Block diagram of the proposed adaptive time-shift analysis speech (ATSA) enhancement system.

The structure of this proposed adaptive time-shift analysis speech enhancement system [5] is shown in Figure 3. By using noise reduction technique voiced/unvoiced decision is made from the initial speech frame. If it contains voiced signal, the time-shift will be changed to one pitch period or else the time-shift will fall back to the original fixed value. In this manner, the analysis window shift adapts to the underlying speech properties and it is no longer fixed.

2) *Windowing Function:* A window function can be applied when a signal is observed for a finite duration for truncating the signal, in signal processing. Rectangular window is the simplest window function which creates the well-known problem, spectral leakage effect. Spectral leakage effect means, if there are two sinusoids waves are present with same frequencies, then leakage are present with both waves interface with each other. If their frequencies are different, then leakage interferes are present of those having less amplitude than other. The main reason of using rectangular window is, it gives strong side-lobes in frequency domain where the first side-lobe is only around 13 dB lower than the main lobe. Similar to Fourier transform, DCT has the same problem with the rectangular window. The rectangular window also has some disadvantages such as discontinuities at the endpoints or maximum scalloping loss for frequency component that is exactly in the middle of two FFT coefficients.

Rectangular window also have some advantages.

1. Due to narrower main-lobe it is able to resolve comparable strength signals.
2. By using rectangular window there is no discontinuity problem at the end point in DCT as compared to DFT. Though DCT is based on an even symmetrical extension during the transform of a finite signal.

Though there are other windows such as Hann window one of the very popular [1] and trapezoidal window [2] use in speech analysis, this paper is focusing on rectangular window though it has some unsatisfactory aspects but some interesting properties.

3) *Weiner Filter with Adaptive controller:* The Wiener filter solves the signal estimation problem for stationary signals. The filter is optimal in the sense of the MMSE (Minimum Mean Square Error). It depends on estimation of the a-priori SNR which can be calculated as

Let the y=noisy speech, s=clean speech and n=noise signal, and their respective DCT representations are $Y_{m.k}$, $S_{m.k}$ and $N_{m.k}$ where m is the time frame index and k is the frequency index. Then the a-priori SNR, ξ, can be expressed as follow:

$$\tilde{\xi}_{m,k} = \alpha \frac{|\tilde{S}_{m-1,k}|^2}{\lambda_N} + (1-\alpha) \max\left[\frac{|Y_{m,k}|^2}{\lambda_N} - 1, 0\right] \quad (1)$$

Where $\tilde{S}_{m-1,k}$ = estimated clean speech in the previous frame, max is the maximum function

λ_N = noise variance which equals to the expectation of the power magnitude of the noise signal, $E[|Nm,k|^2]$.

In above λ_N is assumed to be known as noise signal is a wide sense stationary random process and can be computed during the silence period.

In Equation (1), the parameter α is used to set a proportion of contributions from the previous frames to the current estimate. In Fourier transform domain, the value of α is normally set to 0.98. The same value of α is also commonly used in DCT speech enhancement schemes [1], [4]. DCT coefficients may require a different value of α or even an adaptive one.

For estimation of adaptive controller Minimun Mean Square Error (MMSE) criteria is used along with decision directed approach. It leads to improved version of Weiner filter.

Recall the decision-directed approach in Equation (1), the a-priori SNR can be expressed as:

$$\hat{\xi}_{m,k} = \alpha_{m,k}\tilde{\xi}m\text{-}1,k + (1-\alpha_{m,k}) \max\left(\gamma_{m,k} - 1, 0\right) \quad (2)$$

where $\alpha_{m,k}$ is an adaptive version of α, $\tilde{\xi}_{m-1,k} = |\tilde{S}_{m-1,k}|^2 / \lambda_N$ and $\gamma_{m,k} = |Ym,k|^2 / \lambda_N$

The error between estimated a-priori SNR $\hat{\xi}_{m,k}$ and then real one $\xi_{m,k}$ is

$$J_\alpha = E\left\{\left(\hat{\xi}_{m,k} - \xi_{m,k}\right)^2\right\} \quad (3)$$

4) *Pitch Synchronisation:* Ideally all algorithms work well only in clean situations therefore the above noise reduction filtering is performed first. As in this pitch synchronisation pitch period should extracted first.

The Wiener filtered speech $\hat{S}_{m,k}$ can be given by:

$$\hat{S}_{m,k} = \frac{\hat{\xi}_{m,k}}{\hat{\xi}_{m,k}+1} + Y_{m,k} \quad (4)$$

where the estimated a-priori SNR $\hat{\xi}_{m,k}$ is obtained by using equation (2).

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

Following this noise reduction filtering, the enhanced speech after inverse DCT, $\hat{s}(n)$, is utilized for pitch detection to obtain a more accurate estimation.

Out various algorithms for detecting the pitch period the time domain autocorrelation method [3] is quite a common for solving this problem, since it is simple and robust for some noise corruption conditions. It is selected in this for extracting the pitch period to be used for the time-shift. The autocorrelation function of the resulting signal $\hat{s}(n)$ can be defined as:

$$R(n) = \sum_{m=0}^{N-m-1} \hat{s}(m)\,\hat{s}(n+m) \qquad (5)$$

Since the fundamental frequency in spoken English language is range bound between 80Hz to 500Hz. The frequency of a peal is defined it voiced or unvoiced. A distinct peak is defined to be greater than 0.5 times of $R(0)$. If no distinct peak found, it means that it is likely to be a silence or unvoiced frame. For voiced frames, the pitch period is extracted and used as the analysis window shift. In this method, the window length needs to be at least twice as long as the longest pitch period of the observed speech signals. The final enhanced speech is obtained by overlap & adds process. It is different from the original process due to the adaptive window shifting. A convenient solution is to produce a weighting function which records all the windows frame by frame and calculates the net weighting function. The weighting function can be calculated from the current and the previous frames and hence can be performed in real time. Thereafter, the enhanced speech has to be normalized by the weighting function. The pitch synchronous analysis can be further improved by using maximum alignment technique. In this technique the speech analysis window starts from the short-term maximum amplitude of the speech signals and the time shift equals to one period. For more accurate pitch period compared with noisy speech Wiener filtered speech is used along with maximum alignment technique. Several impulses with period equal to the pitch period of the current voiced frame are generated to calculate the cross-correlation with the Wiener filtered speech. Let $\hat{s}(n)$ is the Wiener filtered speech sequence and $im(n)$ is the impulse sequence then the discrete cross-correlation of these two real signals is given by:

$$R_{\hat{s}im}(n) = \sum_{m=-\infty}^{\infty} \hat{s}(m)im(n+m) \qquad (6)$$

Where the impulse sequence $im(n)$ can be expressed as

$$im(n) = \sum_{m=-N}^{N} \delta(n - mT) \qquad (7)$$

where $\delta(n)$ is the delta function which equals to 1 only when $n = 0$, else $\delta(n) = 0$.

T is the pitch period and

N could be ∞ or a fixed value which indicates the impulse train is infinite or finite respectively. In this paper,The number of impulses are empirically fixed to 5 on $N = 2$.

After words, the position of the maximum amplitude of the current speech frame will be identified by tracking the maximum of the cross-correlation values.

III. RESULTS

Ten different segments of speeches, half females and half males, are randomly selected from the TIMIT database. They are resample at 8 kHz and corrupted by four additive noise types including white noise, fan noise, car noise and F16 aircraft noise. The total speech duration of all these test speech segments is 313.998s including the silence period.

Table 1 Segmental SNR

Noise Type	SNR (dB)	ΔSegSNR Result
White	0	5
	5	4.9
	10	4.3
	15	2.96
	20	2.7
Car	0	2.85
	5	2.1
	10	1.96
	15	3.24
	20	3.99
F 16 Aircraft	-10	6.54
	-5	5
	0	3.97
	5	5.2
	10	2.67

The proposed ATSA technique is evaluated segmental SNR (SegSNR) measure. SegSNR has been widely used to qualify the enhanced speech.

IV. CONCLUSION

In ATSA speech enhancement method the pitch period is extracted first after that the voiced/ unvoiced signal decision is taken. A variety of algorithms are proposed in the past to detect the pitch period. In this thesis the time domain autocorrelation method is used, since it is simple and robust for some noise corruption conditions. It is selected for extracting the pitch period to be used for the time-shift. The wiener filter is used for filtering the noise present in the signal. This method produces good quality enhanced speech.

REFERENCES

[1] I. Y. Soon, S. N. Koh, and C. K. Yeo, "Noisy speech enhancement using discrete cosine transform," *Speech Communication*, vol. 24, pp. 249–257, 1998.

[2] S. Ou, X. Zhao, and J. Dong, "Combining DCT and Adaptive KLT for Noisy Speech Enhancement," in *Wireless Communications, Networking and Mobile Computing, 2007. WiCom 2007. International Conference on*, 2007, pp. 2857– 2860.

[3] L. Rabiner, M. Cheng, A. Rosenberg, and C. McGonegal, "A comparative performance study of several pitch detection algorithms," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 24, no. 5, pp. 399–418, 1976.

[4] H. Ding and I. Y. Soon, "An adaptive time-shift analysis for DCT based speech enhancement," in *Proceedings ICICS*, 2009, pp. 1–4

[5] Huijun Ding, Ing Yann Soon and Chai Kiat Yeo, "A DCT-based speech enhancement system with pitch synchronous analysis*", IEEE Transactions on Audio, Speech and Language Processing* 2011.

[6] Barry Commin, "Signal Subspace Speech Enhancement with Adaptive Noise Estimation" *Department of Electronic and Computer Engineering, National University of Ireland, Galway*, Sept. 2005.