MODIFIED DCT BASED SPEECH ENHANCEMENT IN VEHICULAR ENVIRONMENTS

S.PRASANNA VENKATESH, NITIN NARAYAN, KSAILESH BHARATHWAAJ,
M.P.ACTLIN JEEVA, P.VIJAYALAKSHMI

1,2,3,4,5 SSN College of Engineering, Chennai, India

Abstract- Speech enhancement aims to improve speech quality by using various algorithms. The objective of speech enhancement is improvement in intelligibility and or overall perceptual quality of speech signal using various signal processing techniques. Enhancing of speech corrupted by noise, or noise reduction, is an important field of speech enhancement, and used in many applications such as mobile phones, Voice Over Internet Protocol (VoIP), teleconferencing systems, speech recognition systems, and hearing aids. Our aim is to reduce the background noise (such as car, train noise, etc.) in applications as mentioned above. In our approach, we make use of Discrete Cosine Transform (DCT) in which we exploit the “energy compaction” property of DCT, which is the ability to pack the energy of the data sequence into as few frequency coefficients as possible. This method is expected to produce speech with greater intelligibility and lower noise levels and the effectiveness will be validated using qualitative and quantitative measures.

Keywords- Discrete Cosine Transform; Spectral Subtraction; Energy Compaction; Itakura-Saito Distance; Ip Spectrum

I. INTRODUCTION

The process of speech enhancement includes the method of noise removal in speech processing systems operating in the noisy environments. The presence of noise deprecates the efficiency of speech recognition systems and hence speech enhancement is most commonly incorporated as a preprocessing technique in these systems. Speech enhancement also works on improving the overall intelligibility and/or perceptual quality of speech signals especially those which undergo a transmission through a noisy transmission media. Other fields of applications include hands-free input systems like interactive GPS in cars, voice activated security systems etc.

Various noise removal algorithms prefer working in the transform domain where the noise removal is relatively easier. One such algorithm is the spectral subtraction which employs the Discrete Fourier Transform (DFT) in its process. In simple terms, DFT converts a list of samples of a signal into a list of coefficients ordered by their frequencies. The traditional method of incorporating DFT in spectral subtraction holds a few considerable downsides. Though the technique removes a certain amount of noise, the intelligibility obtained is not the best which has been asserted with examples in this paper.

DCT is similar to DFT except in the fact that it utilizes only real numbers. It expresses the samples of a signal as the sum of cosine functions. The use of cosine rather than sine is an important property since it promotes compression and thereby results in energy compaction... The implementation of DCT directly in the spectral subtraction algorithm proves to be better than the traditional DFT method, but the presence of high frequency noise components in the signal degrades the quality of the signal.

This paper focuses on a proposition to best this particular downside of the DCT implementation by applying its energy compaction property. Noisy speech corpus (NOIZEUS) database is used for the analysis and evaluation of the different spectral subtraction algorithms. This paper focuses on speech corrupted by car noise and train noise. The proposed preprocessing technique has been evinced in this paper through qualitative and quantitative analysis techniques.

The rest of the paper is organized as follows: Section II discusses the conventional spectral subtraction algorithm. The proposed DCT based spectral subtraction and modified DCT based enhancement algorithm is described in Section III and IV. Observations from various spectral subtraction algorithms and their performance evaluation are discussed in Section V and VI. Conclusions and future directions of the proposed work are given in Section VII.

II. SPECTRAL SUBTRACTION

One of the most popular methods of reducing the effect of additive background noise is Spectral Subtraction. In this method, the power spectrum of noise is estimated during silence regions and subtracted from the power spectrum of all the frames resulting in the power spectrum of the speech.

The basic principle of spectral subtraction appears in the literature in different forms. The above cited methods assume that the spectrum of a signal affected by uncorrelated noise is equal to the sum of the signal spectrum and the noise spectrum. This assumption is true only in the statistical sense. Since speech is a quasi-stationary signal, its properties can be assumed to be constant over a short duration of time(say...
Modified DCT Based Speech Enhancement In Vehicular Environments

25ms). The method initially used computed the power spectrum of each windowed segment of speech and subtracted from it an estimate of the noise power spectrum. Noise was estimated during periods of "silence". The original phase of the input signal is used for reconstruction.

\[
P_s(\omega) = \begin{cases} P_s(\omega), & \text{if } \hat{P}_n(\omega) > 0 \\ 0, & \text{otherwise} \end{cases} \tag{1}
\]

where \( P_s(\omega) \) is the spectrum of the input noise-corrupted speech, and \( P_n(\omega) \) is the estimate of the noise spectrum. The enhanced speech signal is recovered from both the estimated spectrum and the original phase by an inverse Fourier transform,

\[
s(t) = \mathcal{F}^{-1} \left[ e^{j \theta(\omega)} \sqrt{P_s(\omega)} \right] \tag{2}
\]

where \( \theta(\omega) \) is the phase of the noise-corrupted speech signal. The negative values in the modified signal spectrum are set to zero as shown in (1). However, this traditional spectrum subtraction method produces a musical noise. An alternative proposed to combat this problem of musical noise consists of subtracting an overestimate of the noise spectrum, and restricting the resultant spectrum from falling below a preset minimum level (spectral floor). The modified method consists of changing the algorithm in (1) to the following:

\[
P_s(\omega) = P_s(\omega) - \alpha P_n(\omega)
\]

\[
\hat{P}_n(\omega) = \begin{cases} \hat{P}_n(\omega), & \text{if } \hat{P}_n(\omega) > \beta P_n(\omega) \\ \beta P_n(\omega), & \text{otherwise} \end{cases} \tag{3}
\]

with \( \alpha \geq 1 \), and \( 0 < \beta \ll 1 \) where \( \alpha \) is the subtraction factor and \( \beta \) is the spectral floor parameter. The modified method is with \( \alpha > 1 \). The above method reduces the musical noise but with a reduced intelligibility of speech. In this paper, an alternative DCT implementation is proposed instead of the existing implementations using DFT to obtain an improved quality of speech.

### III. PROPOSED SOLUTION

The spectral subtraction method that we have proposed uses Discrete Cosine Transform (DCT) instead of Discrete Fourier Transform (DFT) with a modification in the parameters. A Discrete Cosine Transform expresses a sequence of data as a sum of cosine functions. The use of cosine rather than sine functions is important for energy compression, since it turns out that only less number of cosine functions is needed to approximate a signal. The DCT (Type II) and its inverse for the real sequence \( x(n) \) is given in eq.(4) and eq.(5)

\[
Y(k) = w(k)\sqrt{2/N} \sum_{n=0}^{N-1} \cos(\pi(2n+1)k/2N) x(n) \tag{4}
\]

where \( k=0,1,2 \ldots \ N-1 \)

\[
w(k) = \begin{cases} \sqrt{1/2}, & k=0 \text{ or } N \\ 1, & 1 \leq k \leq N-1 \end{cases}
\]

Here, \( N \) is the length of the sequence. The proposed solution uses DCT instead of DFT to compute the spectra of both noise-corrupted speech and noise signals. The method of spectral subtraction using DFT proposed employs various parameters such as spectral subtraction factor, gain factor, spectral floor factor and voice activity detection factor. These parameters were tweaked to give optimum result for spectral subtraction using DCT. In the proposed method, the absolute value of Discrete Cosine Transform (DCT) of the noise estimate is subtracted from the Discrete Cosine Transform (DCT) of the noise-corrupted speech signal. The algorithm uses the first five frames to estimate the power spectral density of noise and updates the estimate periodically with the help of voice activity detector. The phase information of the original noise corrupted signal is multiplied with the subtracted value to recover back the reconstructed signal.

\[
x(n) = \sqrt{2/N} \sum_{k=0}^{N-1} w(k)Y(k)\cos(\pi(2n+1)k/2N), \tag{5}
\]

where \( n = 0,1,2 \ldots N-1 \),

\[
w(k) = \begin{cases} \sqrt{1/2}, & k=0 \text{ or } N \\ 1, & 1 \leq k \leq N-1 \end{cases}
\]

### IV. ENHANCEMENT OF SPEECH USING MODIFIED DCT

The method of using DCT in spectral subtraction proposed earlier results in an improved intelligibility of speech. However, the recovered speech signal has an annoying high frequency noise in the background. This noise can be removed to a large extent by exploiting the energy compaction property of DCT, which is the ability to pack the energy of the spatial sequence into as few frequency coefficients as possible. This is because speech energy is concentrated into a few coefficients predominantly while the noise energy is distributed throughout the spectrum. This property is useful for noise removal. By eliminating the coefficients associated with high
frequencies, this high frequency noise can be reduced. In order to eliminate the undesired coefficients, an energy threshold is fixed. By trial and error, the energy threshold was fixed to be 55 percent of the total energy of the noise-corrupted signal. The number of coefficients is then increased from one until the energy of each frame exceeds this threshold. The remaining coefficients are set to zero to maintain a constant frame length of 25ms.

Spectral subtraction using DFT results in the loss of intelligibility in the recovered speech. In this case, the phrase “broke the man’s fall” is not intelligible when spectral subtraction is applied using DFT. However, when spectral subtraction is applied using DCT, the reconstructed speech is highly intelligible, but it has an annoying high frequency noise in the background. This noise is reduced when the modified DCT algorithm is applied. This is because the speech energy is concentrated into a few coefficients predominantly while the noise energy is distributed throughout the spectrum. The removal of high frequency coefficients by fixing the threshold removes the noise components while maintaining the intelligibility.

VI. PERFORMANCE ANALYSIS

Performance of enhancement algorithms are evaluated using quantitative measures such as Itakura-Saito distance, cosh distance for spectral distortion and using qualitative measure, mean opinion score (MOS) test for quality and intelligibility. The performance evaluation is carried out using NOIZEUS database [3,4]. The database has 30 IEEE sentences and eight different real world noises taken from the AURORA database. The noises available in the NOIZEUS database include idle car noise, train noise, inside flight noise, etc. The noise signals are down sampled from 44.1 kHz to 25 kHz. Quantitative spectral distortion measures are calculated for 20 sentences, from the NOIZEUS database, recovered using DFT, DCT, and modified DCT based spectral subtraction techniques.

V. OBSERVATIONS

Fig.3 a shows the plot of the clean speech “The soft cushion broke the man’s fall”. Fig.3 b shows the speech corrupted with train noise. Fig.3 c shows the signal recovered using spectral subtraction using DFT. Fig.3 d shows the signal reconstructed using spectral subtraction using DCT. Fig.3 e shows the signal recovered using spectral subtraction using modified DCT algorithm proposed.

Itakura-Saito distance is computed to show the perceptual difference between original LP spectrum and an estimation of that spectrum.

$$D_{\lambda} (\omega) = \left( \frac{P(\omega)}{P(\omega)} \right) \cdot \left( 1 + \frac{P(\omega)}{P(\omega)} \right) \cdot \left( \frac{1}{2\pi} \right) \int_{-\pi}^{\pi} \log \left( \frac{P(\omega)}{P(\omega)} \right) d\omega$$

(7)
modified DCT based speech enhancement in vehicular environments. The order of linear prediction coefficients is 20. Fig. 4 shows the LP spectra of the original speech signal, speech signal recovered using spectral subtraction using DFT, DCT, and modified DCT algorithm respectively. In the LP spectrum of signal recovered using spectral subtraction using DFT, spurious peaks are present (refer Fig. 4 b) whereas the LP spectrum of signal recovered by spectral subtraction using DCT, high frequency components are observed (refer Fig. 4 c).

These high frequency components are responsible for the annoying noise in the background.

These high frequency components are reduced considerably in the LP spectrum of signal recovered by spectral subtraction using modified DCT algorithm (refer Fig. 4 d). On an average, it is observed that the Itakura-Saito distance is less for spectra of speech reconstructed using modified DCT than that of the other two techniques for about 50 percent of the frames. This value is calculated by considering both the voiced and unvoiced frames.

The enhancement algorithms are also evaluated for spectral distortion using the symmetric measure, the Cosh distance measure. Here also it is observed that the Cosh distance is less for spectra of speech recovered using modified DCT for about half of the total frames. This value could be enhanced further by considering only the cosh distances of voiced frames.

From the analysis of two spectral distortion measures, namely Itakura-Saito and Cosh distance, it is clear that the spectral information is recovered in a better manner in the modified DCT algorithm when compared to the other spectral subtraction techniques.

**QUALITATIVE ANALYSIS:**

A Mean Opinion Score (MOS) test is used to assess the quality and intelligibility of the enhanced speech signals. MOS test is a subjective listening test in which, the participants were asked to score the intelligibility and perceptual quality of the signal enhanced using various spectral subtraction techniques.

Ten participants were asked to rate the enhanced speech on a scale of 0 to 5, where 5 being the most intelligible and 0 being the least intelligible. Ten sentences were corrupted with three types of noises (Car noise at 60mph, Train noise-1, Train noise-2) and the speech signal is recovered from them using DFT, DCT and modified DCT based spectral subtraction techniques. A total of 90 sentences (3 noises and 3 techniques for 10 sentences) were rated by the participants. The score obtained from each participant is averaged out and the Mean Opinion Score is displayed in Table I. Participants rated the speech enhanced using modified DCT algorithm to be significantly better in terms of intelligibility and perceptual quality than the speech recovered using other spectral subtraction techniques.

![Fig.4 LP spectra of (a)clean speech (b) speech recovered using DFT (c) speech recovered using DCT (d) speech recovered using modified DCT](image-url)
As observed from Table I, there is an improvement of 0.4 in the MOS score of modified DCT when compared to DCT and an improvement of about 1.7 when compared to DFT.

<table>
<thead>
<tr>
<th>Noise</th>
<th>Mean Opinion Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DFT</td>
</tr>
<tr>
<td>Car Noise 60mph</td>
<td>1.99</td>
</tr>
<tr>
<td>Train noise 1</td>
<td>1.84</td>
</tr>
<tr>
<td>Train noise 2</td>
<td>1.8</td>
</tr>
<tr>
<td>Average MOS value</td>
<td>1.8766</td>
</tr>
</tbody>
</table>

CONCLUSION

In conclusion, the proposed method of spectral subtraction using modified DCT algorithm has been empirically proved to produce speech of better intelligibility and lower noise levels. In spectral subtraction using DFT, some of the syllables were lost upon recovery. However, these syllables were recovered successfully in spectral subtraction using DCT. But, annoying high frequency components were present. These noise components were removed in the proposed method wherein the energy compaction property of DCT is exploited. Quantitative measures such as Itakura-Saito distance and qualitative measures such as MOS were used to justify the efficacy of the proposed method.

The analysis in this paper was restricted to car noise and train noise. This analysis could be extended to other noise sources in the future. In order to compute DCT, segments of fixed size were used.

Instead of choosing segments with fixed size, segmentation could be performed based on the sound units (or phonemes). This method of segmenting the speech file based on the sound unit before applying the modified DCT algorithm might improve the intelligibility even further.

REFERENCES