Speech Recognition Enhancement by Psychoacoustic Modeled Noise Suppression

Yiu-Pong LAI, Man-Chun HUI, Chi-Wah KOK and Man-Hung SIU
Department of Electrical and Electronic Engineering, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, HONG KONG

Abstract—This paper proposed a spectral subtraction based speech enhancement algorithm that improves computer based speech recognition. Speech recognition can not be improved by traditional spectral subtraction techniques because of the associated artifacts, such as musical noise. This paper proposes a tonal noise suppression algorithm which can be used for accurate psychoacoustic model parameter estimation. Simulation results showed that the spectral subtraction based speech enhancement algorithm with psychoacoustic masking produced by the estimated parameters produces high quality clean speech, which can be used as a preprocessing tool to enhance the speech recognition rate. Simulation results using Aurora 3 database are presented.

I. INTRODUCTION

Speech recognition systems work reasonably well in laboratory conditions, but their performance deteriorates drastically when deployed in practical situations where the speech is corrupted by additive noise distortion. One way to improve the performance of speech recognition systems is to enhance the speech signal (and remove noise) prior to its recognition. Various speech enhancement techniques have been applied as a preprocessor for recognizing speech in the presence of noise and were found to improve the recognition performance significantly, especially at low SNR.

Speech enhancement is a special case of time-varying signal estimation that aims at making human voice clear and easier to understand for both human and computer listeners. The speech enhancement algorithm finds an optimal estimate preferred by a human listener. The speech signals are estimated by modeling the speech production or the human auditory system. In comparison, the noise spectrum is relatively easier to estimate than that of the speech signal, because the noise component is relatively stationary.

In this paper, we focus on single channel speech enhancement. This is the most difficult task, since the noise and the speech are in the same channel. In this case, noise is usually estimated during speech pauses. Since the speech signal is assumed to be corrupted by additive noise, therefore, clean speech can be obtained by spectral subtraction technique [1] with the estimated noise spectrum. Through advance estimation techniques, clean speech can be generated with all noise components being removed. Unfortunately, spectral subtraction introduces “musical noise” into the enhanced speech which can be very annoying and unnatural. This perceptually annoying noise is composed of tones at random frequencies and has an increased variance that affects the recognition performance.

Various techniques have been developed to reduce the artifacts associated with spectral subtraction. In this paper, it is proposed to incorporate a human hearing model that is widely used in audio coding [4]. Instead of attempting to remove all noise components from the speech signal, the proposed algorithm attempts to attenuate the noise below the audible threshold. Thus reduces the amount of modification to the spectral magnitude, and hence the amount of artifacts introduced into the cleaned speech signal. This allows one to find the best tradeoff between the amount of noise reduction, the speech distortion and the level of residual noise in a perceptual sense.

Auditory model is used in [2] to adjust the parameters of a non-auditory noise suppression procedure. Haulick [3] uses the auditory masking threshold to identify and then suppress musical noise. Thiemann [4] directly constructed the spectral subtraction levels from a high-resolution psychoacoustic model originally developed for the evaluation of audio quality. High quality clean speech can be produced, however, the algorithm does not work well on noisy speech obtained from environments with tonal noise nature, such as background speech environments, static noise, etc. In case of inaccurately calculated masking parameters, spectral subtraction associated artifacts will be enhanced and results in annoying music noise. To obtain clear speech, overestimated noise components are used in spectral subtraction [4], and thus leads to musical noise too.

To remedy this problem, this paper proposed a tonal noise suppression scheme that is shown to work well with the auditory masking based spectral subtraction speech enhancement algorithm. Simulation results have shown that high quality clean speech can be obtained with inaudible spectral subtraction artifacts. A high percentage of improvement is obtained from the speech recognition results of the cleaned speech when compared to the original noise corrupted speech over a wide signal to noise ratio range. The proposed speech enhancement algorithm is discussed in Section II, where the tonal noise component is suppressed by tonal noise detector. The noise spectrum of the tonal noise suppressed speech is estimated in Section IIA with the psychoacoustic masking model in [7]. The estimated masking threshold is applied to the spectral subtraction algorithm to produce the clean speech. Simulation results are presented in Section III and the paper is concluded in Section IV.

II. PROPOSED ALGORITHM

The speech signal sampled at $f = 8000Hz$, and grouped into subframes of 16ms, or 128 samples. A processing frame is formed by two adjacent subframes and is sample-by-sample multiplied to the raised-cosine window $h(n) = 0.54 - 0.46 \cos \left( \frac{2 \pi n}{N - 1} \right).$ (1)
where $N = 256$ is the frame size. Notice that the processed frame can be perfectly reconstructed to the original speech signal through an overlap add process. The high-resolution spectral response $X(k)$ of the windowed signal is computed using a 1024 point DFT. The magnitude response $|X(k)|$ is preprocessed to suppress tonal noise, while the phase $\angle X(k)$ is reserved for the reconstruction of the noise suppressed signal. A perceptually modeled spectral subtraction algorithm is then applied to the tonal noise suppressed signal $X(k)$ to generate the magnitude response of the clean speech $S(k)$. The clean speech is obtained by the IDFT of the signal $S(k) \cdot \angle X(k)$. Figure 1 shows the detail block diagram of the proposed algorithm.

A. Tonal Noise Suppression and Spectral Masking Estimation

We followed the tonal analysis method described in MPEG1 audio coder to detect tonal component from the power spectrum $P(k,p)$ of the high resolution spectrum $|X(k,p)|$ of the $p$-th speech signal.
The power spectrum $P(k, p)$ is normalized by a reference level 96dB as
\[ P(k, p) = P(k, p) - \max(P(k, p)) + 96. \]

Tonal signal (both speech and noise) are detected by first locating the peaks of $P(k, p)$. The located spectral peaks are known as tonal components if and only if
\[ X(k, p) \text{ is tonal} \iff \begin{cases} P(k) - P(k + s) > t_s, \\ P(k) - P(k - s) > t_s. \end{cases} \]

To determines the tonal component to be speech or noise, the algorithm relies on the relative stationarity of the noise signal when compared to that of the speech signal. A long time estimator for tonal noise is applied which employs a counter to monitor the tonal components. Furthermore, in order to combat for the chaotic nature of the tonal components in real world applications, the spectrum is divided into 40 bands such that each bands contains two frequency bins. One counter is assigned to each band, which is increased by one when a tonal is detected in that band. Otherwise the counter is decreased by one. When the counter values exceed a chosen threshold, the frequency bin is corrupted by tonal noise and are suppressed by replacing $|X(k, p)|$ with the geometric mean of the spectral components around $X(k, p)$.
\[ |\hat{X}(k, p)| = \left( \prod_{i=1}^{\tau} |X(k + i, p)| \right)^{1/\tau}. \]

For all other frequency, $|\hat{X}(k, p)| = |X(k, p)|$.

### B. Voice Activity Detector

The noise spectrum $\hat{W}(k, p)$ is estimated from the tonal noise suppressed signal $|\hat{X}(k, p)|$. The noise estimation is taken from the speech pauses which are identified using a voice activity detector given by
\[ V = \frac{1}{N} \sum_{k=1}^{N} \frac{|\hat{X}(k, p)|^2}{|W(k, p - 1)|^2} - \log \frac{|\hat{X}(k, p)|^2}{|W(k, p - 1)|^2} - 1, \]
where $W(k, p - 1)$ is the estimated noise power spectrum in the $p - 1$th frame. The $p$-th frame is determined to be speech or noisy by
\[ V > 0.6 \text{ speech,} \]
\[ V < 0.6 \text{ noise.} \]

Since the spectrum of the noise signal is assumed to be a short-time stationary process, therefore, the noise power spectrum is updated from the current and previous estimates according to
\[ |\hat{W}(k, p)|^2 = \lambda |\hat{W}(k, p - 1)|^2 + (1 - \lambda)|\hat{X}(k, p)|^2, \]
where $\lambda$ is the forgetting factor, and is chosen to be 0.7 in our simulation, which leads to an average of 16 frames (256 ms) to be included in each estimation.

### C. Spectral Subtraction

The spectral subtraction process can be reformulated as a spectral smoothing process with gain factor $G(k, p)$ for the $k$-th frequency bin in the $p$-th frame. The clean speech is produced by
\[ \hat{S}(k, p) = |\hat{X}(k, p)|G(k, p) - \hat{X}(k, p), \]
where $G(k, p)$ is the Wiener filter of the tonal suppressed signal
\[ G(k, p) = \frac{\hat{S}(k, p)}{|\hat{X}(k, p)|}, \]
and $\hat{S}(k, p)$ is the estimated clean speech obtained by spectral subtraction of the previous noise estimation $\hat{W}(k, p)$ as
\[ |\hat{S}(k, p)|^2 = \max(|\hat{X}(k, p)|^2 - |W(k, p)|^2, 0). \]
speech and that of the cleaned speech. If we zoom-in the voice inactive region, we can observe that a smooth transition to zero is observed in the clean speech when compared to that obtained by [4] where time domain Gibbs phenomenon are observed which contribute to the musical noise artifacts. Showing in Figure 3(a) and (b) are the spectral responses of the speech signal before and after speech enhancement using the proposed algorithm, which clearly showed that the noisy formant signals are not completely removed. Instead they are suppressed to a level lower than the audible threshold obtained from the psychoacoustic model as shown in Figure 3(c), where the dotted lines are the audible threshold resulted from the tonal masking effects of the psychoacoustic model. Notice that by suppressing the tonal noise and other noise components to a level smaller than the audible threshold value, the algorithm efficiently cleans the noisy speech and at the same time reduces the amount of artifacts induced into the clean speech. Although the noisy formant-like tones are not completely removed, reducing the noisy formant-like tones to a level lower than the audible threshold would help to increase the correct recognition rate because it is the relative power of individual spectral peaks that are important to the performance of the recognition system. As a result, the psychoacoustic masking effectively “removed” the noisy formant signal from the speech system. To confirm our conjecture, three experiments were performed.

Three speech recognition experiments, the well-matched (WM), Mid-mismatch (MM) and High-mismatch (HM), were conducted according to different level of mismatch between the training and testing conditions [8]. Both noisy speech and the enhanced speech were parameterized by the Mel-frequency Cepstral Coefficient (MFCC) [9] based feature vector in the front-end process of HTK 3.1 [10]. Similar to other speech recognition experiments, the feature vector consists of 12 static coefficients and log frame energy as well as their first and second derivatives. Besides using HTK, two ETSI front-end processing tools (W1007 and W1008) [11], [12] were used to perform the same experiments as references. The training and testing procedure for all the experiments were based on the script provided with the database. Noted that the first 250ms of the cleaned speech signal were removed before feeding into the speech recognition system. This is because the first few frames of the speech signal were used to estimate the noise spectrum, as a result, they do not benefit from the speech enhancement. Similarly, the first 250 ms of the noisy speech were removed before feeding into the speech recognition system for fair comparison.

The performance of the recognition experiments are tabulated in Table I. The second to the fourth columns of the table represent the recognition accuracy of different evaluation conditions, and the last two columns represent the average accuracy and the weighted average accuracy according to the evaluation in ICSLP 2002 [13] respectively. Observed from the first two rows, the MFCC on the enhanced speech can improve the recognition performance where the mismatch between the training and testing speech has been reduced by more than 20% relative word error rate showing. In the third and fourth rows are the recognition results that use Cepstral Mean Subtraction (CMS). The CMS is commonly used to remove convolutive noise to improve speech recognition rate. Observed from the second row of both tables showed that the recognition rate of enhanced speech has out-performed that obtained by CMS processed speech results in the third row of both tales. As a result, we can conclude that the effect of tonal noise, such as background speech etc, will affect the performance of the speech recognition system a lot more than that of convolutive noise. Furthermore, it can be observed from fourth row of both table which showed that the recognition rate of CMS processed enhanced speech signal are higher than that of all the previous cases. As a result, we can conclude that the speech enhancement algorithm proposed in this paper collaborate well with CMS algorithm which results in a clean speech that is free from additive and convolutive noises.

Compared with other front-end processing tools, it can be observed that the speech recognition results obtained by the proposed speech enhancement algorithm is compatible to that of using standard feature processing tool (W1007) [11] provided by ETSI which is listed in the fifth row in both tables. However, when compared with the advanced DSR feature processing tool (W1008) [12], the speech recognition performance obtained by proposed speech enhancement algorithm is shown to inferior. This is because a different speech enhancement model is considered in W1008 which is aimed at MFCC based speech recognition system, while the proposed system uses general auditory model. As a result, with the special target considered by W1008, better speech recognition performance is observed. However, it should be noted that there are two drawback concerning about W1008. First, since it is not constructed according to the human auditory system, as a result, it is not as robust as the proposed speech enhancement system when considered as a front-end processing tool for a general class of speech recognition system which may not be MFCC based. Second, the cleaned speech obtained from W1008 may not be easier to understand by human and computer listeners. As a result, it will be difficult to establish the noisy environment conditions where the W1008 algorithm breaks down.

In addition to the speech recognition rate comparison using different front-end processing tools, we have also investigated the robustness of the speech recognizer with that obtained with the speech end-points given a prior. The comparison results are shown in Table II, where more than 20% relative word error rate reduction are observed using the proposed speech enhancement process. This indicates that the proposed algorithm does not only removes the noise from speech, but also reduces the mismatch problem between the speech and noisy environment and that in cleaned speech.

Note that although the speech recognition performance of the proposed speech enhancement technique is not as good as that of W1008, the computational complexity of the proposed algorithm is a lot lower than that of W1008. As a result, we argued that the proposed algorithm is useful in constrained power applications. In the above experiments, although the performance of our approach is not as well as the ETSI W1008 feature extraction processing, which is well designed for speech recognition, the noise suppression algorithm shows significant improvement on the performance of speech recognition.

| TABLE I Recognition Performance of Aurora 3 Danish Subset |
|---------------|---------|---------|---------|-------|-------|
| feature vector | WM      | MM      | HM      | avg   | w.avg |
| HTK MFCC       | 84.29   | 57.20   | 32.48   | 57.99 | 61.86 |
| HTK MFCC (enhanced) | 88.13   | 66.53   | 48.00   | 63.89 | 66.69 |
| HTK MFCC (CMS) | 87.50   | 66.38   | 54.72   | 62.87 | 66.91 |
| HTK MFCC (CMS) (enhanced) | 87.39   | 67.66   | 50.63   | 68.56 | 71.29 |
| ETSI (W1007)   | 87.39   | 64.09   | 47.06   | 66.38 | 69.36 |
| ETSI (W1008)   | 91.57   | 74.58   | 74.53   | 80.26 | 81.36 |

IV. Conclusion

A speech enhancement algorithm that works well in a wide range of SNR is presented. The proposed algorithm makes use with spectral subtraction technique to suppress noise in speech signal. Psychoacoustic model is applied to estimate a noise spectral mask, which when applied to spectral subtraction will mask the noise to become inaudible. We further investigated the incorporation of a
TABLE II
Recognition Performance of Aurora 3 Danish Subset with end-point noise Conditions

<table>
<thead>
<tr>
<th>feature vector</th>
<th>WM</th>
<th>MM</th>
<th>HM</th>
<th>avg</th>
<th>w.avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>HTK MFCC</td>
<td>86.21</td>
<td>61.72</td>
<td>32.13</td>
<td>60.02</td>
<td>64.12</td>
</tr>
<tr>
<td>HTK MFCC (enhanced)</td>
<td>89.56</td>
<td>69.63</td>
<td>53.64</td>
<td>70.94</td>
<td>73.61</td>
</tr>
<tr>
<td>HTK MFCC (CMS)</td>
<td>88.65</td>
<td>67.51</td>
<td>31.39</td>
<td>62.52</td>
<td>66.94</td>
</tr>
<tr>
<td>HTK MFCC (CMS) (enhanced)</td>
<td>89.24</td>
<td>70.06</td>
<td>52.47</td>
<td>70.59</td>
<td>73.33</td>
</tr>
<tr>
<td>ETSI (WI007)</td>
<td>87.28</td>
<td>67.32</td>
<td>39.37</td>
<td>64.66</td>
<td>68.32</td>
</tr>
<tr>
<td>ETSI (WI008)</td>
<td>93.37</td>
<td>81.49</td>
<td>79.59</td>
<td>84.82</td>
<td>85.77</td>
</tr>
</tbody>
</table>

Tonal noise removal components into the speech enhancement system which will provide a more accurate estimation of the psychoacoustic model and thus achieve better noise suppression results. The speech enhancement system is than applied as front-end processing tool for the speech recognition system. The proposed speech enhancement system provides clean speech that are free from musical noise and time-domain Gibbs phenomenon, which are shown to be a major source of noisy signals that affect the speech recognition rate. Simulation results showed that a consistent 20% improvement can be obtained by using the proposed speech enhancement algorithm when compared to traditional system. Improvement are also observed when applying the proposed system with CMS. Finally, because of the good performance of the system and very low computational complexity of the proposed system (when compared to that of WI008), therefore, we can concluded that the proposed speech enhancement system is a very attractive speech recognition system front-end processing tool.

REFERENCES
[12] ETSI standard doc. “Speech processing, Transmussion and Quality aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms” ETSI ES 202 050 v1.1.1, October 2002.

Fig. 1. The block diagram of a psychoacoustic modelled noise suppression speech enhancement system with tonal noise suppression preprocessing.

Fig. 2. Time domain waveform of a noisy speech, and (b) cleaned speech.

Fig. 3. Spectral response of (a) a noisy speech frame, (b) cleaned speech, and (c) with psychoacoustic mask plotted in dotted line.