Volume : 1 | Issue : 4 | April 2012 ISSN - 2250-1991

Research Paper Science

Voice Activation Detection Algorithm for Estimating the Noise From Human Speech Signal

* Kanu Patel ** Sameena Zafar

* PG Student, Patel College of Science and Technology, Bhopal

** Assistant Professor, Patel College of Science and Technology, Bhopal

ABSTRACT

Speech enhancement seeks to eliminate noise in a variety of environments, the most prominent of which are telecommunications applications. After over thirty years of research throughout the world, no perfect solution exists to this problem. The objective of our work is to implement a novel speech enhancement algorithm, which offers superior noise reduction over current methods. All speech enhancement systems suffer from distortion or residual noise due to imperfect noise removal. Some variations are more promising than others. One such method is signal subspace speech enhancement. However, this algorithm can only update the noise estimate when speech is absent, and suffers degradation in performance in many different noise types. The system designed in this thesis takes the subspace method as its basis and develops a robust and accurate noise estimation algorithm that can update the noise estimate throughout the signal, not just in speech absence.

Keywords : Speech Enhancement; Noise reduction; Speech estimation

1. Introduction

Whenever a microphone records speech, unwanted noise is record. This noise depends on the environment and can range from anything such as computer fan noise, car engine noise to factory floor noise. Figure 1 shows how noise introduces in speech using a microphone. 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. This has been an ongoing area of research was proposed in 1979 by Boll in [1].

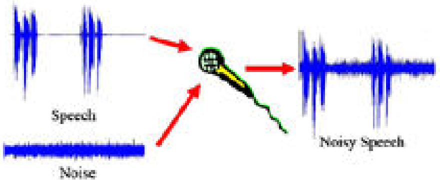

Figure: 1 Basic diagram for noise introduce in speech via microphone

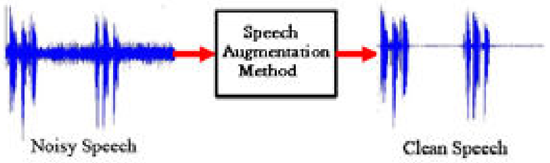

Figure: 2 Basic blocks to removed/reduced noise

Much progress has made in the development of single- microphone noise reduction for hearing aid and speech communication. Figure 2 shows the basic method for removing the noise from noisy speech signal. Noise is everywhere and in most applications that are related to audio and speech, such as human-machine interfaces, hands-free communications, voice over IP (VoIP), hearing aids, teleconferencing / telepresence / telecollaboration systems, and so many others, the signal of interest (usually speech) that is picked up by a an audio source is generally contaminated by noise. In many speech related systems like mobile communication in an adverse environment, the desired signal is not available directly; rather it is mostly contaminate with some interference sources of noise. These background noise signals degrade the quality and intelligibility of the original speech, resulting in a severe drop in the performance of the applications. The degradation of the speech signal due to the background noise is a severe problem in speech related systems and therefore should eliminate through speech enhancement algorithms. The majority of these algorithms have found to improve listening comfort and speech quality. Past intelligibility studies conducted in the late 1970s found no intelligibility improvement with the spectral subtraction algorithm. Noise-reduction algorithms implemented in wearable hearing aids revealed no significant intelligibility benefit, although they have found to improve speech quality and ease of listening in hearing-impaired. Some of the noise-reduction algorithms proposed for hearing aids rely on modulation spectrum filtering, others rely on which in some cases, might be more damaging than the background noise itself.

2. Brief Review

The basic fundamental spectral and noise information indicates in this section.

2.1. Spectral Subtraction

The earliest and most commonly used method of speech enhancement is magnitude spectral subtraction, [1]. Figure 3 shows the spectral subtraction method. Speech and noise are assumed to be additive and un correlated, therefore, if an estimate of the noise spectrum can be found for a particular frame of a noisy speech signal, then an estimate of the clean speech signal can be calculated by subtracting it from the noisy signal as described by the flowing expression:

$$\hat{X}(\omega) = Y(\omega) - \widehat{W}(\omega) \quad (1)$$

Where, $\hat{X}(\omega)\hat{X}(\omega)$is the estimate of the clean frequency spectrum for a given frame, $Y(\omega)Y(\omega)$ is the noisy spectrum for that frame, and $\widehat{W}(\omega)\widehat{W}(\omega)$ is the noise spectrum estimate. An estimate of the clean speech has recovered by applying the inverse DFT to $\hat{X}(\omega)\hat{X}(\omega)$, to give $\hat{X}(t)$ $\hat{X}(t)$. Since the human ear is relatively insensitive to phase, the phase angle of the noisy signal can be use when reconstructing the speech.

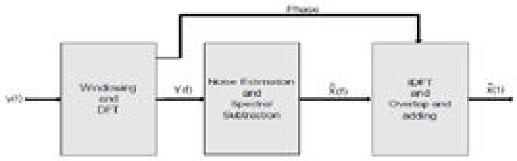
Phase
y(t)
Windowing and DFT
Y(f)
Noise Estimation and Spectral Subtraction
$\hat{X}(f)$
IDFT and Overlap and adding
$\hat{x}(t)$

Figure: 3 Spectral subtraction system overview

This is quite a computationally simple system; however, noise estimation is a non-trivial problem, still without an optimal solution after almost thirty years of research. There are many different methods of noise estimation.

2.2. Drawbacks of Spectral Subtraction
Spectral Subtraction may be straightforward to implement, and although reducing the noise significantly, it has some severe drawbacks. It is clear that the effectiveness of spectral subtraction is heavily dependent on accurate noise estimation, which is a difficult task to achieve in most conditions. When the noise estimate is less than perfect, two major problems occur, musical noise and distortion.

2.3. Musical Noise
Musical noise occurs when random short sinusoids, which are tone-like in sound, is created due to flaws in the noise estimate, making the noise removal imperfect, [2].

$$e = \hat{X}(\omega) - X(\omega) \quad (2)$$

e is the error, the difference between the clean estimate $\hat{X}(\omega)\hat{X}(\omega)$ and the actual clean signal $X(\omega)X(\omega)$.

Musical noise artifacts are randomly distributed over time and frequency because some, but not all, of the frequency components are removed from the noisy signal. Perceptually they are very annoying to the listener, due to their randomness and unnatural quality. Studies show than many listeners find musical noise more disturbing than the original noisy signal. Since the ultimate goal of speech enhancement is to provide good quality speech for human listeners, this is a severe flaw.

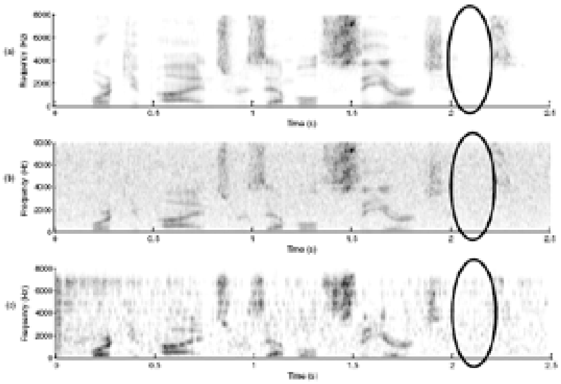

Figure: 4 Spectrograms of (a) clean, (b) noisy and (c) enhanced signals, respectively.

In figure 4 it can be seen that while spectral subtraction enhances the speech, it is at the cost of random, perceptually annoying, musical noise. Comparing the sections circled in figure 4 (a), (b) and (c) it is evident that the enhanced signal (c) contains random frequencies that are not present in the clean signal (a). This is an example of musical noise.

3. Noise Estimation

There are several distinct approaches to finding a good noise estimate, some of which are more useful in different situations.

3.1. Two-Channel Method

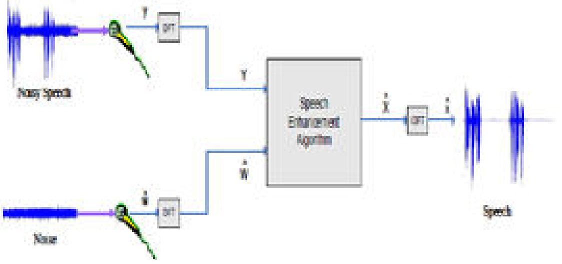

Figure: 5 Two-Channel Speech Enhancement overview

A two-channel method, also known as adaptive noise cancellation [3,4], involves the use of two recording devices as shown in figure 5. One microphone is used to record the speech plus noise while a second microphone is placed in a position where it can pick up noise only. The input from the latter microphone is use to calculate the spectrum of the noise in the noisy signal. This only works in a stationary environment, where the noise source is known, for example where the noise from a car's engine might interfere with mobile phone conversation. A microphone attached to a hands-free mobile phone docking station could be place appropriately to pick-up car engine noise, but not speech.

In practice, it is problematic to find a suitable location for the second microphone, due to the difficulty of placing the microphone in a location where speech cannot hear, but noise can. The level of input noise detected by the second microphone may be different from the actual noise picked up by the mobile phone's microphone due to their relative distances from the noise source.

3.2. Voice Activity Detection
Voice activity detectors aim to calculate which frames of a noisy signal contain speech and which do not. Noise statistics are estimated and updated every time a frame is judge not to contain speech, i.e. noise-only frames.

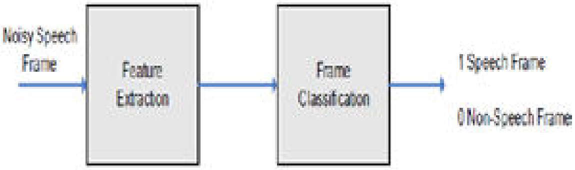

Figure: 6 Voice Activity Detection overview

One example of a VAD involves calculating the number of times the signal amplitude crosses the x-axis (i.e. the amplitude is zero) in a frame. This is called the Zero Crossing Rate (ZCR). Non-speech frames have a lower average ZCR than noisy speech frames, since they contain less signal information. Therefore, if the ZCR for a given frame is below a certain threshold value, δ, it is determined to be a noise-only frame. Otherwise, it is determined that the frame contains speech as well as noise.

Rabiner's VAD algorithm [7] uses the ZCR along with the short-term energy of the noisy signal to determine the presence (or absence) of speech in each frame. If the noisy signal energy in frame mm,$Y(m)\,Y(m)$, rises above the average estimated noise energy $\widehat{W}(m)\widehat{W}(m)$, then it is likely that frame mm contains speech plus noise. Otherwise, it is a noise-only frame.

$$D(m) = \begin{cases} 1 & ZCR(m) > \delta \; and \; Y(m) < \hat{W} \\ 0 & else \end{cases} \quad (3)$$

Where $D(m)D(m)$is the VAD output, with 1 representing a speech frame and 0 representing a noise-only frame.

However, there are difficulties common to all VADs [9]. Firstly, there is the situation where the noisy signal contains mostly speech with very few speech pauses, meaning the noise updates are rare. In that, time the noise may have varied sufficiently to make the estimate inaccurate. This produces errors in the enhanced speech, such as musical noise and distortion, as discussed earlier. Even theoretical VADs, which perfectly decide between noise and speech frames, can produce poor results if the speech pauses are too infrequent or if the noise changes too rapidly.

Secondly, most VADs have difficulty distinguishing correctly between noise and speech at low SNRs, these results in the estimated noise spectrum containing speech components, become incorrectly attenuated, resulting in loss of speech information or distortion.

4. Results

Here, we get the result by using a method "Computationally Efficient Speech Enhancement by Spectral Minima Tracking in Sub bands". Doblinger, [8] was amongst the first to implement a noise estimation algorithm that did not rely on voice activity detection. This technique based on tracking the minima of the noise power in each subband.

The algorithm summarized as follows:

· The noisy signal power in the Kth sub band at frame mm is denoted by $Y_K(m)Y_K(m)$

· The noisy signal power estimate is smoothed using

$$\hat{Y}_K(m) = \alpha\, \hat{Y}_K(m-1) + (1-\alpha)Y_K(m) \quad (4)$$

Where $\hat{Y}_K(m)\hat{Y}_K(m)$ the average is noisy signal power, and α is the smoothing parameter.

· The noise power estimate in frame m is calculated by:

$$\hat{W}_k(m) = \begin{cases} \gamma\hat{W}_k(m-1) + \dfrac{1-\gamma}{1-\beta}\left(\hat{Y}_k(m) - \beta\hat{Y}_k(m-1)\right) & \hat{W}_k(m-1) < \hat{Y}_k(m) \\ \hat{W}_k(m-1) & \text{else} \end{cases}$$

Where $\hat{W}(k)\hat{W}(k)$represent the noise power estimate.

The parameters a, b and γ were experimentally chosen as 0.7 ≤ a ≤ 0.9, β = 0.96 and the noise smoothing parameter γ was chosen very close to one, γ = 0.998.

The major drawback in Doblinger's algorithm is evident in figure 7. The algorithm cannot distinguish between an increase in noise power and an increase in speech power, resulting in an overestimation of noise power during speech frames. For example in the region between frame 180 and frame 230 above the noise power estimate, $\widehat{W}(m)$ $\hat{W}(k)\hat{W}(k)$has followed the speech power. Hence, distortion occurs in the estimated clean signal, due to this noise over-estimate.

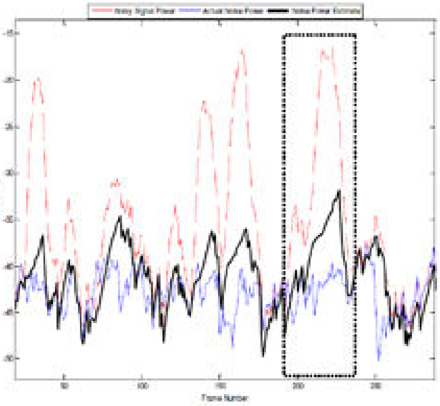

Figure: 7 Noise estimation in subband 17 using Doblinger's algorithm

REFERENCES

[1] Li, N. and Loizou, P. (2008). "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," Journal of Acoustical Society of America, 123(3), 1673-1682 [2] M. E. Hamid, K. Ogawa, and T. Fukabayashi, "Improved Single-channel Noise Reduction Method of Speech by Blind Source Separation", Acoust. Sci. & Tech., Japan, 28(3):153-164, 2007. [3] C. He, and G. Zweig, "Adaptive two-band spectral subtraction with multi-window spectral estimation", ICASSP, vol. 2, pp. 793-796, 1999. [4] S. C. Liu, "An approach to time-varying spectral analysis", J. EM. Div. ASCE 98, 245-253, 1973. [5] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shin, Q. Zheng, N. C. Yen, C. C. Tung, and H. H. Liu, "The Empirical Mode Decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis", Proceeding Royal Society London A, vol. 454, pp. 903-995, 1998. [6] R. Martin, "Spectral Subtraction Based on Minimum Statistics", Proc. EUSIPCO, pp. 1182-1185, 1994. [7] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shin, Q. Zheng, N. C. Yen, C. C. Tung, and H. H. Liu, "The Empirical Mode Decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis", Proceeding Royal Society London A, vol. 454, pp. 903-995, 1998. [8] Z. Xiaojie, L. Xueyao, Z. Rabu, "Speech Enhancement Based on Hilbert-Huang Transform Theory", in First International Multi-Symposiums on Computer and Computational Sciences, pp. 208-213, 2006. [9] P. Flandrin, P. Goncalves and G. Rilling, "Detrending and Denoising with Empirical Mode Decompositions", In Proc., EUSIPCO, pp.1581-1584, 2004.