Impulsive noise removal from Speech Signals using Rank Order Mean Method

Arjun Kumar[1],Akash kumar[1], Saurabh Kumar Shrivastava[1], Dr. R. K. Singh[2]
[1]M.Tech Student Kamla Nehru Institute of Technology, Sultanpur
[2]Professor, Electronics Engg. Department, Kamla Nehru Institute of Technology, Sultanpur

Abstract—*This paper presents a noise cancellation technique to remove impulsive noises that commonly corrupt speech signals using Rank Order Mean is proposed. The rank order differentiation is applied to input signal to estimate the time occurrence of impulsive noise. Then rank order mean is used for replacing the noisy samples to get the noise free signal. The above described technique shows improvement in terms of Signal to Noise Ratio (SNR) and Perceptual Evaluation of Speech Quality (PESQ) with respect to the existing techniques.*

Index Terms— wavelets, transient removal, impulsive noise removal, speech enhancement, Rank order mean.

Introduction

The most common problem in speech processing is the effect of interference noise in speech signals. Interference noise masks the speech signal and reduces its intelligibility. Interference noise can come from acoustical sources such as ventilation equipment, traffic, crowds and commonly, reverberation and echoes. It can also arise electronically from thermal noise, tape hiss or distortion products. If the sound system has unusually large peaks in its frequency response, the speech signal can even end up masking itself.

One relationship between the strength of the speech signal and the masking sound is called the signal-to-noise ratio, expressed in decibels. Ideally, the S/N ratio is greater than 0dB, indicating that the speech is louder than the noise. Just how much louder the speech needs to be in order to be understood varies with, among other things, the type and spectral content of the masking noise.

The most uniformly effective mask is broadband noise. Although, narrow-band noise is less effective at masking speech than broadband noise, the degree of masking varies with frequency.

High-frequency noise masks only the consonants, and its effectiveness as a mask decreases as the noise gets louder. But low-frequency noise is a much more effective mask when the noise is louder than the speech signal, and at high sound pressure levels it masks both vowels and consonants.

Regularity of speech and impulse noise

The regularity of a signal is defined as the number of continuous derivatives that the signal possesses. It can be estimated from the Holder exponent, also known as the Lipschitz exponent [7], which is defined as follows: If we assume that a signal f(t) can be approximated locally at υ by a polynomial of the form

$$f(t) = C_0 + C_1(t - v) + \ldots + Cn(t - v)^n + C|t - v|^\alpha$$
$$= p_n(t - v) + C|t - v|^\alpha \ldots\ldots\ldots\ldots\ldots (1)$$

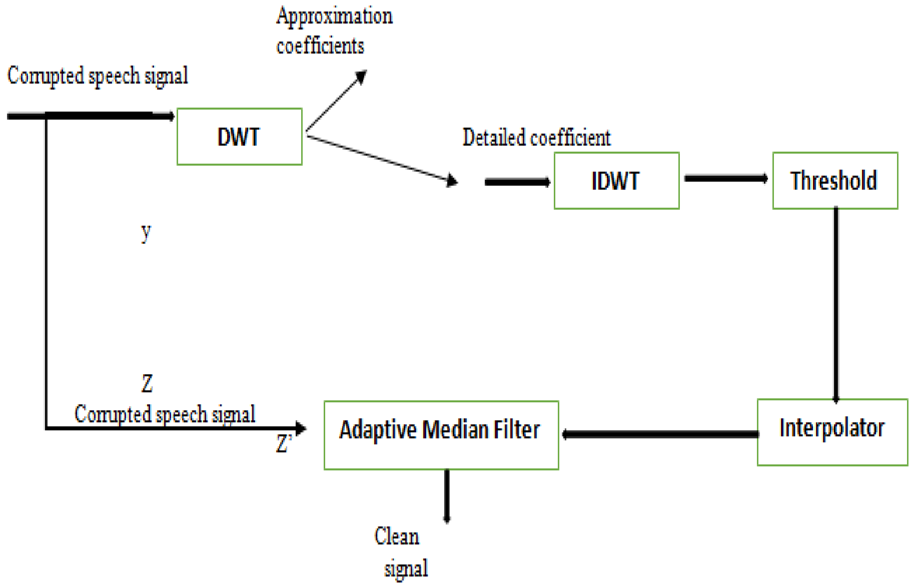
where p_n is a polynomial of order n and C is a coefficient, then the term associated with Lipschitz exponent, α, is the part of the signal that does not fit into the n + 1 approximation [6]. The local regularity of a function at υ can be characterize by α such that

$$/f(t) - p_n(t - v) \mid \leq C|t - v|^\alpha \ldots\ldots\ldots\ldots (2)$$

Where a higher value of indicates a better regularity or smoother function. In [7], it is shown that a bounded function *f(x)* is uniformly Lipschitz over an interval *R* if it satisfies

$$\int_{-\infty}^{+\infty} |F(w)|(1 + |w|^\alpha)dw < +\infty \ldots\ldots\ldots (3)$$

where F(ω) is the Fourier transform of *f(ω)*. The condition gives the global regularity for the whole real line and implies that F(ω) has a decay faster than 1/w A speech signal can be considered to be broadly made up of vowels, consonants and silence portions. The vowel portion is generated by vibrations in the vocal chords which are then low pass filtered by the vocal tract. As such, vowels are usually periodic with an upper cutoff frequency that does not exceed 5 kHz. The consonants, on the other hand, are generated by constriction in the mouth; they are usually random with a spectrum that can extend up to 20 kHz. If we use condition [3], it is clear that a vowel due to its lower bandwidth will have larger positive Lipschitz value than a consonant which may even be negative. The silence portion of speech is essentially background noise that is random with a small or negative Lipschitz value. Consequently, by knowing the random or periodic nature of different parts of speech and their corresponding regularity, it is possible to make a better decision when removing impulse noise from speech. Another important characteristics of speech, is the slow time varying nature of the temporal and spectral envelop in comparison to an impulse; this slow-time varying nature is because speech is generated by the movements of muscles in the mouth and vocal tract, which is a relatively slow process. A higher value of indicates a better regularity or a smoother function. To detect an impulse, we need a transform that ignores the polynomial part in [1]. A wavelet transforms$\psi(x)$ with n vanishing moments is able to ignore a polynomial up to order n.



Such a wavelet satisfies the condition

$$\int_{-\infty}^{+\infty} t^n \psi(t)dt = 0 \text{ .... } (4)$$

and using it to transform [2] we get the inequality

$$|Wf(s,t)| \leq Cs^{\alpha} \text{ .... } (5)$$

Detailed coefficient

where W*f(s, t)* is the wavelet transform of *f(t)* at scale s [7]. To estimate the local regularity of a signal we set the inequality in (5) to equality and take the logarithm on both sides:

$$\log|Wf(t,s)| = \log(c) + \alpha \log s \text{ ...... } (6)$$

The Lipschitz value is simply the slope of the decay of the wavelet coefficients across scales,
given by

$$= \frac{\Delta \log |Wf(t,s)|}{\Delta \log s} \text{......... } (7)$$

An impulse is characterized by a sudden change in the signal or a sudden shift in the signal mean value. Large magnitude coefficients, termed modulus maxima, will be present at time points where

the impulses have occurred. Impulses are distinguishable from noise by the presence of modulus maxima at all of the scale levels; noise, on the other hand, produces modulus maxima only at finer scales. In [7], Mallet and Hwang used a method for detecting singularities by analyzing the evolution of the wavelet modulus maxima across scales, for a continuous wavelet transform; the decay of the maxima line was used to determine the regularity at a given point. However, in practical application a dyadic discrete wavelet transform is preferredover the continuous wavelet transform due to its lower computational effort. Therefore, if a discrete wavelet transform is usedthe number of wavelet coefficients will reduce by half for the next increase in scale. In such a case, rather than extracting the maxima line it is computationally more efficient to simply estimate the decay of the wavelet coefficients for each time point in the smallest scale.Points where large changes occur in the signal, such as an impulse, would have large coefficients at all the different scales, thus having little decay across scales. Noise, on the other hand, would have large coefficients only at finer scales and would, therefore, show faster decay towards coarser scales. Likewise, the consonants in speech being similar to high-pass filtered white noise would also show more decay towards coarser scales; however, vowels, with frequency spectrum that does not exceed 5 kHz, are characterized by small coefficients in the smallest scale that increases as the scale increases.

Removal of impulse noise from speech

Since our objective is the removal of impulse noise from speech, it is not critical that impulses of lower magnitudes that are perceptually inaudible be removed. It is important, however, that impulses of larger magnitudes be suppressed below the just-noticeable level difference (JNLD) to make them inaudible. The JNLD is not a fixed value and varies with the nature of the signal and sound pressure level (SPL). For example, for white noise the JNLD is around 0.7 Db for SPL between 40 and 100 dB, while for a 1 kHz tone the JNLD decreases from 1 to 0.2 dB as the SPL increases from 40 to 100 Db [8]. Consequently, an impulse would be perceptually inaudible formost SPL levels if it is suppressed below 0.2 dB above the speech signal level.

As discussed in Section 2, the temporal and spectral envelop of speech is slow time-varying in comparison to an impulse. This property is used to detect and suppress the wavelet coefficients that correspond to an impulse. Therefore, what is needed is a dynamic threshold for each wavelet level that varies in proportion to the smooth envelop of the absolute wavelet coefficients values, but, at the same time, not affected by impulse noise. That is, for scale s and sample n, such a dynamic threshold, τ(s, n), can be defined as

$$\tau(s,n) = k_s . Env[|Wf(n,s)|] \ldots\ldots\ (8)$$

where operator Env[·] is the envelop of the signal that is unaffected by impulse noise and ks is a factor that is determined empirically for each level on the basis of the JNLD and the nature of the impulse noise. A median filter is known to possess the property where step function type signals are preserved while at the same time robust to impulse noise [9]. As such, the operator Env[·] in [8] can be replaced by a median filter of length N = 2K + 1 so that τ(s, n) becomes

$$[\tau(s,n) = ks.MED[|Wf(n-K,s)| ,\ldots..,|Wf(n,s)|,\ldots\ldots.|Wf(n+K,s)| \ldots\ldots\ldots\ (9)$$

The length of the median filter needs be adjusted so that it is sufficiently long in comparison to an impulse but short in comparison to a vowel or consonant. A wavelet coefficient would be considered to be that of an impulse if it greaterthan (s, n); to suppress the impulse the coefficient is attenuated to a new coefficient, dw*f(n, s),*given by

$$\widehat{Wf}(n,s) = \begin{cases} Wf(n,s) & \text{if } |Wf(n,s)| < \tau(s,n), \\ \tau(s,n)\frac{Wf(n,s)}{|Wf(n,s)|} & \text{otherwise.} \end{cases} \ldots\ldots\ldots\ (10)$$

For impulses that occur in a consonant or in the non-voice portion of speech, suppression of the wavelet coefficients at coarser scales is as important as in the finer scales; this is because an impulse has a Lipschitz exponent that is usually greater than a consonant or background noise and, as such, its coefficients will not decay as fast at coarser scales. And if these coefficients at coarser scales are not suppressed adequately, they will be audible as low frequency thuds. However, for an impulse that occurs in the middle of a vowel, suppression of the corresponding coefficients at coarser scales is not as critical as in the finer scales. This is because the Lipschitz exponent of a vowel is much greater than that of an impulse, and at coarser scales the contribution to the coefficients comes mainly from the vowel. This also implies that the vowel will usually mask out the low-frequency portion of an impulse. At larger scales, the use of [8] and [9] to detect the coefficients that correspond to an impulse become less effective if a discrete wavelet transform is used since the number of sample points decreases by half for the next increase in wavelet scale, thereby reducing the time resolution by half; furthermore, at larger scales an impulse will have much lesser contributions on a wavelet coefficient since other portions of the signal within the wavelet support length will also contribute to the coefficient. Therefore, for wavelet coefficients at larger scales it becomes more effective if the attenuation is done on the basis of the impulses detected in the smaller scales.

If I (s) is the average decay of an impulse and I (s0) = 1, where s0 is the finest scale, then the attenuated wavelet coefficient, *dwf(sc, n)*, for the coarser scale, *sc*, is given by

$$\widehat{Wf}(s_c, n) = \begin{cases} 0 & \text{if } K < 0, \\ K & \text{otherwise.} \end{cases}$$

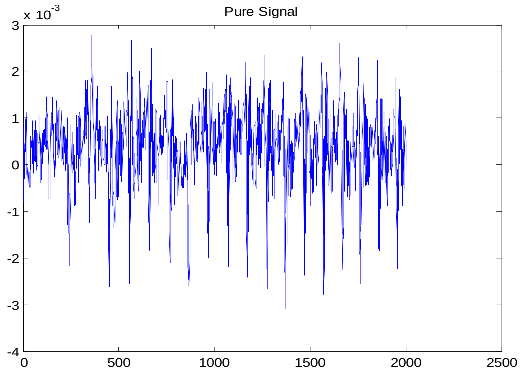
…………… (11)

|Wf(sf , n)| is the absolute wavelet coefficient of the detected impulse in the finer scale sf , and kc is a constant that is empirically determined depending on the type of impulse noise and the JNLD at that scale.

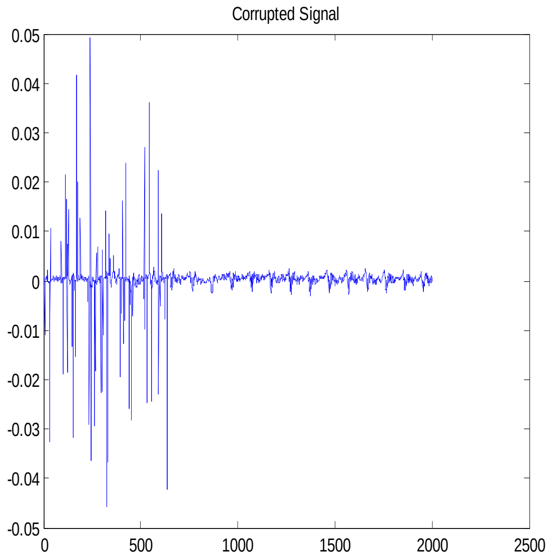
Result and discussion

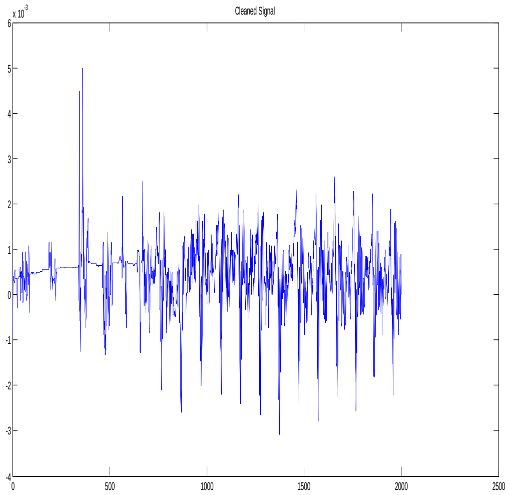
In this thesis work we addresses the problem of reducing the impulsive noise in speech signal using compressive sensing approach. The results are compared against well-known speech enhancement methods, spectral subtraction, Total variation denoising.

The impulse noise corrupted speech signal and the enhanced speech signal (the output of the noise cancellation system) are given as input to the classification system. The speech recognition system gives 12.3 % accuracy for noisy signal where as 92.3 % accuracy for the enhanced signal.



Objective and subjective quality evaluation are performed for the four speech enhancement scheme. Results show that the signal processed by the compressive sensing based method outperforms the other methods.matrix and DCT bases. It gives better results compared to the traditional methods like total variation and median filtering techniques. The proposed algorithm is evaluated using different objective and subjective tests like LLR, SNR Seg, PESQ etc. The result shows that the quality of speech has been improved durably. Also the output of the enhanced speech signal is shows 92.3 % accuracy in automatic speech recognition of Malayalam digits whereas the noisy speech signal shows only 12.3 % accuracy. This showsthe importance of impulsive noise removal in speech processing algorithms as a pre-processing step.





Conclusion and scope of future work

In this work, we have introduced Signal Dependent Rank Order Mean method for noise cancellation. A new method for removing impulse noise from speech in the wavelet transform domain has been described. The method utilized the multi-resolution property of the wavelet transform, which provides finer time resolution at high frequencies, to effectively identify and remove the impulse noise. To discriminate the impulse from speechit uses the low time-varying nature of speech relative to an impulse, and the difference in regularity between an impulse and various parts of speech. On the basis of these differences, an algorithm was developed to identify and suppress wavelet coefficients that correspond to impulse noise. Experiment results have shown that the new method is able to significantly reduce impulse noise without degrading the quality of speech signal or introducing any artifacts.

References

[1]P. Esquef, M. Karjalainen, and V. Valimaki, "Detection of clicks in audio signals using warped linear prediction," 14th International Conference on Digital Signal Processing 2002, vol. 2, pp. 1085-1088, July 2002.

[2] C. Chandra, M. S. Moore, S. K. Mitra, "An efficient method for the removal of impulse noise from speech and audio signals," Proceedings of ISCAS 1998, vol. 4, pp. 206-208, Jun 1998.

[3] Z. Liu, A. Subramanya, Z. Zhang, J. Droppo, and A. Acero, "Leakage model and teeth clack removal for air- and bone-conductive integrated microphones," Proceedings of ICASSP 2005, vol. 1, pp. 1093-1096, Mar 2005.

[4] S. V. Vaseghi and R. Frayling-Cork, "Restoration of old gramophone recordings," Journal of Audio Eng. Soc., vol. 40, No. 10, pp. 791-801, 1992.

[5] S. MallatandW. L. Hwang, "Singularity detection and processing with wavelets," IEEE Trans. Inform. Theory, vol. 38, no. 2, pp. 617-643, 1992.

[6] L. Yin, R. Yang, M. Gabbouj, and Y. Neuvo, "Weighted median filters: a tutorial," IEEE Trans. Circuits Syst, vol. 43, no. 3, pp. 157-192, 1996.

[7] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoust., Speech, Signal Processing, vol. 27, pp. 113-120, Apr. 1979.

[8] Y. Ephraim and D. Mallah, "Speech enhancement using a minimal mean-square error short-time spectral estimator," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-32, no. 6, pp. 1109-1121, Dec. 1984.

[9] S. Mallat, A Wavelet tour of signal processing, Second Edition, Academic Press 1998.

[10] E. Zwicker and H. Fastl, Psychoacoustics - Facts and Models, Second Edition, Springer 1999

.