A New Technique for Compensation in Speech Enhancement Using Magnitude and Phase Spectrum

S Chandra Sekhar (M.tech)[1]

T. Vijay Kumar (Associate Professor and HOD) [2]

[1,2]Dr. k. v. subba reddy institute of technology, Dupadu, Kurnool, Andhra Pradesh, 518218, INDIA

ABSTRACT

Foundation commotion is an extreme issue in discourse related frameworks. Keeping in mind the end goal to take care of this issue, it is imperative to dispense with the clamor from the loud discourse, which is called discourse improvement. Run of the mill discourse improvement calculations just work on the brief timeframe size range, while keeping the brief timeframe stage range unaltered for blend. Or on the other hand just remunerate the stage range while keeping the greatness range unaltered. In this paper, we display a novel technique by changing both size and stage spectra to create an altered complex range. The trial of a target discourse quality measure PESQ, and spectrogram examination had demonstrated that the proposed technique can get better upgrade execution.

I.INTRODUCTION

Speech enhancement is a noise suppression technology It has important significance for solving the problem of noise disturbance. And it can improve the quality and intelligibility of voice communications. The purpose of speech enhancement is to restore the original signal from noisy observations corrupted by various noises [1]. Let us consider an additive noise mode

$$x(n) = s(n) + d(n) \quad (1)$$

Where$x(n), s(n), d(n)$ denote discrete-time signals of noisy speech, clean speech, and noise, respectively. The discrete short-time Fourier transform (DSTFT) of the corrupted speech signal $x(n)$ is given by

$$\mathrm{X(n,k)} = \sum_{\mathrm{m}=-\alpha}^{\alpha} \mathrm{x(m)\omega(n-m)e^{-j2\pi km/N}} \quad (2)$$

Where k denotes the k th discrete-frequency of N uniformly spaced frequencies and ω(n)is an analysis window function of short duration. By using DSTFT we can obtain Eq.(1) as.

$$\mathrm{X(n,k) = S(n,k) + D(n,k)} \quad (3)$$

Where$X(\mathrm{n,k})$, $S(\mathrm{n,k})$,and $D(\mathrm{n,k})$are the DSTFTs of noisy speech, clean speech, and noise, respectively. Each of them can be described in terms of the DSTFT magnitude spectrum and the DSTFT phase spectrum. For example, $S(\mathrm{n,k})$can be written in polar form as

$$S(\mathrm{n,k}) = |S(\mathrm{n,k})|\mathrm{e}^{-\mathrm{j}\angle S(\mathrm{n,k})} \quad (4)$$

Where $S(\mathrm{n,k})$ is the magnitude spectrum, and $\angle S(\mathrm{n,k})$ is the phase spectrum.

Most of the existing speech enhancement algorithms only change the magnitude spectrum of the noisy speech. The modified magnitude then recombined with the unchanged phase spectrum to produce a modified complex spectrum, which is the estimated clean speech spectrum. These algorithms are called magnitude spectrum based methods. Boll proposed the method of spectral subtraction (SSUB) in 1979. Its basic principle is to subtract the magnitude spectrum of the noise from the noisy speech magnitude spectrum, and obtain the estimate of the clean signal magnitude spectrum, but the phase spectrum is unchanged [2].

The MMSE estimator, which is presented by Ephraim and Malah in 1984. Its main idea is to minimize the mean-squared error (MSE) between the clean and estimated (magnitude or power) spectra [3]. Wiener filter [4] was proposed by Wiener. Hansen and Jensen first presented the Wiener method in the single channel case enhancement [5].

Doclo and Moonen further extended the Wiener method in the multichannel case [6]. Ephraim and Van Trees proposed the linear predictive factors to estimate the pure speech signal [7].

The reason for ignoring the phase impact is that the phase spectrum has been found to have less perceptual effect at significantly higher signal to noise ratio (SNR) levels[8].But recently, it is found that the phase spectrum may be useful in speech processing applications[9].

Kamil Wójcicki et.al. proposed the speech enhancement method of phase spectrum compensation (PSC) in 2008[10][11]. This paper proposes a new method by changing both magnitude spectrum and phase spectrum to produce a modified complex spectrum. The proposed method obtains better performance in terms of an objective speech quality

The challenge to the researchers is to create software that works with imperfect, but historically invaluable, audio tracks. Commercial voice recognition software designed for personal computer users works well if the speaker talks clearly in English, or another common West European language. The personal accounts have been given in many different East European languages. Further challenges are presented by strong accents and the emotion shown in many recordings.

The researchers do not aim to design a system that can transcribe the recordings word-for-word, but a searchable database that links different testimonies to key events and places. "We want to build a speech recognition system that is good enough to recognise some of words," Although the largest strides in the development of voice recognition technology have occurred in the past two decades, this technology really began with Alexander Graham Bell's inventions in the 1870s. By discovering how to convert air pressure waves (sound) into electrical impulses, he began the process of uncovering the scientific and mathematical basis of understanding speech.

In the 1950s, Bell Laboratories developed the first effective speech recognizer for numbers. In the 1970s, the ARPA Speech Understanding Research project developed the technology further - in particular by recognizing that the objective of automatic speech recognition is the understanding of speech not merely the recognition of words.

By the 1980s, two distinct types of commercial products were available. The first offered speaker-independent recognition of small vocabularies. It was most useful for telephone transaction processing. The second, offered by Kurzweil Applied Intelligence, Dragon Systems, and IBM, focused on the development of large-vocabulary voice recognition systems so that text documents could be created by voice dictation.

Over the past two decades, voice recognition technology has developed to the point of real-time, continuous speech systems that augment command, security, and content creation tasks with exceptionally high accuracy.

1.2 History of Speech Recognition

Designing a machine that mimics human behavior, particularly the capability of speaking naturally and responding properly to spoken language, has intrigued engineers and scientists for centuries. Since the 1930s, when Homer Dudley of Bell Laboratories proposed a system model for speech analysis and synthesis, the problem of automatic speech recognition has been approached progressively, from a simple machine that responds to a small set of sounds to a sophisticated system that responds to fluently spoken natural language and takes into account the varying statistics of the language in which the speech is produced. Based on major advances in statistical modeling of speech in the 1980s, automatic speech recognition systems today find widespread application in tasks that require a human-machine interface, such as automatic call processing in the telephone network and query-based information systems that do things like provide updated travel information, stock price quotations, weather reports, etc. In this article, we review some major highlights in the research and development of automatic speech recognition during the last few decades so as to provide a technological perspective and an appreciation of the fundamental progress that has been made in this important area of information and communication technology.

1.2.1 Early automatic speech recognizers

Early attempts to design systems for automatic speech recognition were mostly guided by the theory of acoustic-phonetics, which describes the phonetic elements of speech (the basic sounds of the language) and tries to explain how they are acoustically realized in a spoken utterance. These elements include the phonemes and the corresponding place and manner of articulation used to produce the sound in various phonetic contexts. For example, in order to produce a steady vowel sound, the vocal cords need to vibrate (to excite the vocal tract), and the air that propagates through the vocal tract results in sound with natural modes of resonance similar to what occurs in an acoustic tube. These natural modes of resonance, called the formants or formant frequencies, are manifested as major regions of energy concentration in the speech power spectrum. In 1952, Davis, Biddulph, and Balashek of Bell Laboratories built a system for isolated digit recognition for a single speaker [1], using the formant frequencies measured (or estimated) during vowel regions of each digit.

1.2.2 Technology drivers since the 1970's

In the late 1960's, Atal and Itakura independently formulated the fundamental concepts of Linear Predictive Coding (LPC, which greatly simplified the estimation of the vocal tract response from speech waveforms. By the mid 1970's, the basic ideas of applying fundamental pattern recognition technology to speech recognition, based on LPC methods, were proposed by Itakura, Rabiner and Levinson and others.

Other systems developed under DARPA's SUR program included CMU's Hearsay (-II) and BBN's HWIM, Neither Hearsay-II nor HWIM (Hear What I Mean) met the DARPA program's performance goal at its conclusion in 1976. However, the approach proposed by Hearsay-II of using parallel asynchronous processes that simulate the component knowledge sources in a speech system was a pioneering concept.

II.PROPOSED METHOD

This method is based on modified magnitude and phase spectrum compensation. The block diagram is shown in Fig. 1

The magnitude estimation of clean speech is [3]

$$\hat{A}_k = \frac{\sqrt{\pi}}{2}\frac{\sqrt{V_k}}{\gamma_k} exp\left(-\frac{V_k}{2}\right)\left[(1+V_k)I_o\left(\frac{V_k}{2}\right) + V_k I_1\left(\frac{V_k}{2}\right)\right] Y_k \quad (5)$$

I_o (•) And I_1(•) denotes modified Bessel functions of zero and first order respectively,

Then V_k is defined as

$$V_k = \frac{\xi_k}{1+\xi_k}\gamma_k \quad (6)$$

Where ξ_k and γ_k are defined as

$$\xi_k = \frac{\lambda_x(k)}{\lambda_d(k)}\gamma_k = \frac{{Y_k}^2}{\lambda_d(k)} \quad (7)$$

The phase spectrum compensation function obtained from [10]

$$\Lambda(n\,,k) = \lambda\psi(k)\left|\widehat{D}(n\,,k)\right| \qquad (8)$$

λ = real-valued empirically determined constant. Where λ=3.74

$\psi(k)$=anti symmetry function

$$\psi(k) = \begin{cases} 1 & if\ 0 < k/N < 0.5 \\ -1 & if\ 0.5 < k/N < 1 \\ \quad 0, & otherwise \end{cases} \qquad (9)$$

Non-conjugate vectors of DSTFT values are given to the zero weighting (i.e k=0 value and possible singleton k=N/2 for even). The next step is offsetting the complex spectrum of the noisy speech by the additive real-valued phase spectrum compensation function Λ(K).

$$X_\Lambda(n\,,k) = X(n\,,k) + \Lambda(n\,,k) \quad (10)$$

The compensated phase spectrum is obtained by

$$\angle X_\Lambda(n\,,k) = ARG[X_\Lambda(n\,,k)] \quad (11)$$

ARG= complex angle function.

From above we got the magnitude estimation and compensated phase spectrum so then we can get the modified complex spectrum by

rlap-add, we obtain the enhanced time-
$\hat{s}(n)$

$$\hat{S}(n\,,k) = \hat{A}_k e^{j\angle X_A(n\,,k)} \qquad (12)$$

The last IDSTFT is used t frequency domain $\hat{S}(n\,,k)$ to the time-domain representation. The result may be complex; in PSC method the imaginary component is

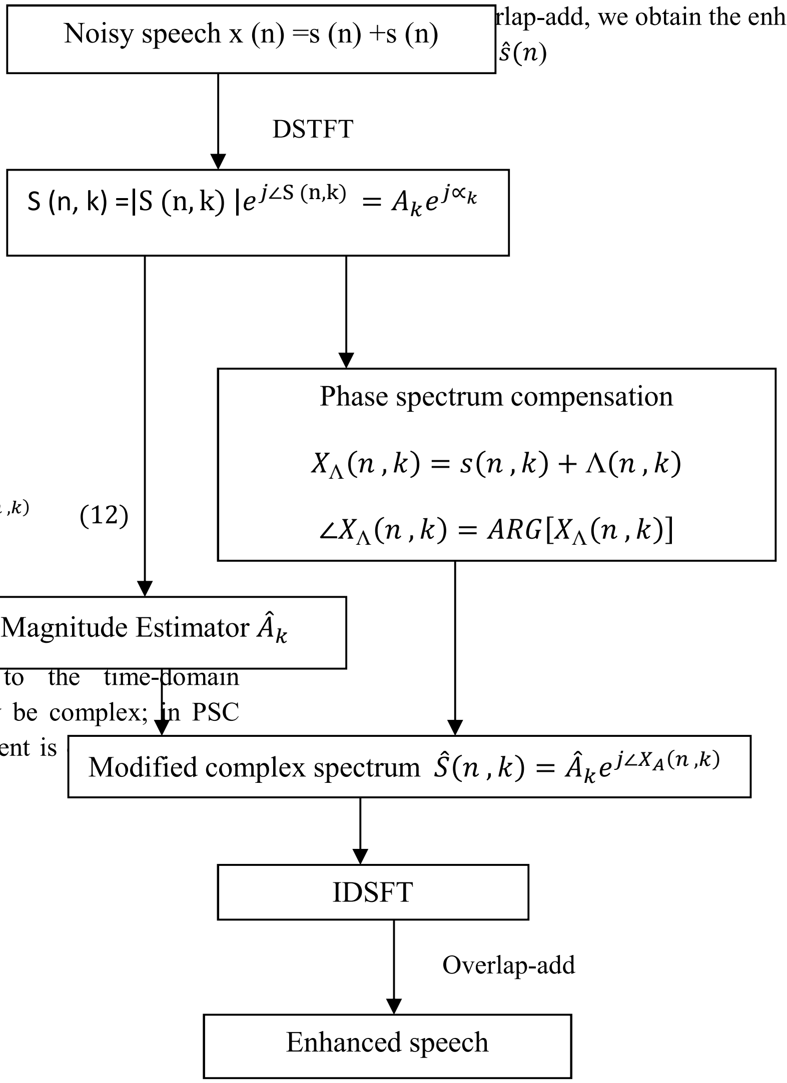
Noisy speech x (n) =s (n) +s (n)
DSTFT
$S(n,k) = |S(n,k)|e^{j\angle S(n,k)} = A_k e^{j\propto_k}$
Phase spectrum compensation
$X_\Lambda(n\,,k) = s(n\,,k) + \Lambda(n\,,k)$
$\angle X_\Lambda(n\,,k) = ARG[X_\Lambda(n\,,k)]$
Magnitude Estimator $\hat{A}_k$
Modified complex spectrum $\hat{S}(n\,,k) = \hat{A}_k e^{j\angle X_A(n\,,k)}$
IDSFT
Overlap-add
Enhanced speech

Fig: 1. Block diagram of proposed speech enhancement method

III. EXPERIMENTAL RESULTS

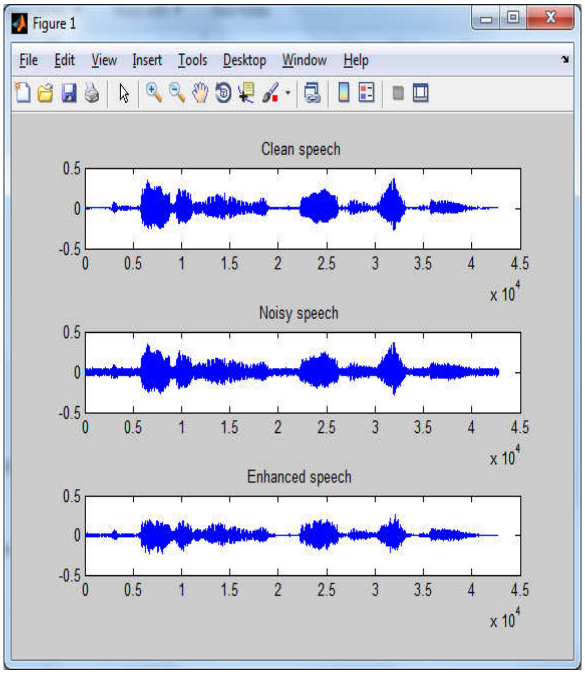

Fig: 2.Plot for **sp10 audio files** (a) Clean Speech (b) Noisy Speech (c) Enhanced Speech.

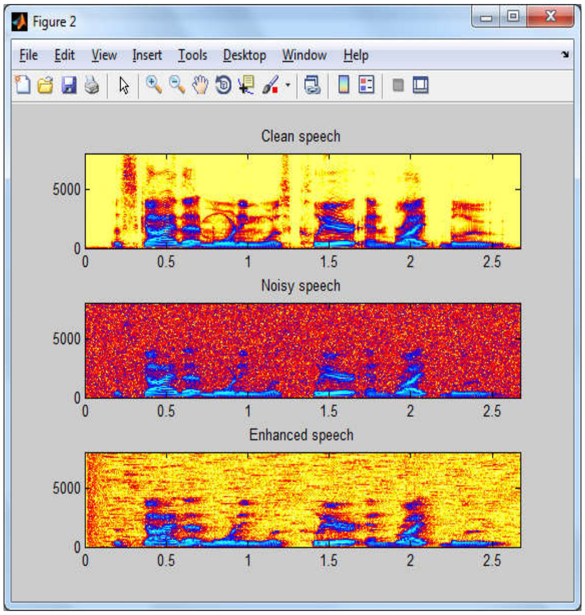

Fig: 3.Plot for **sp10_white_sn10 audio files** (a) Clean Speech (b) Noisy Speech (c) Enhanced Speech.

IV.CONCLUSION

Here, we have proposed another strategy for discourse upgrade which depends on altered greatness and remunerated stage range. Trial consequences of the target discourse quality measure PESQ, and spectrogram investigation had demonstrated that the proposed strategy accomplish preferred discourse quality over the customary discourse improvement strategies. The technique can be utilized as a part of the frameworks which need to wipe out the foundation commotions, for example, discourse acknowledgment, discourse correspondence, and so on, and it can additionally enhance the discourse quality and comprehensibility.

[1] P. Loizou, Speech Enhancement: Theory and Practice. Boca Raton,FL: CRC, 2007.

[2] BOLL S F. Suppression of acoustic noise in speech using spectral subtraction[J]. IEEE Trans. Acoustics, Speech, Signal Processing, 1979, 27(2):113-120.

[3] Ephraim Y, Malah D. Speech enhancement using a minimum mean square error short time spectral amplitude estimator. IEEE Transactions on Acoustics, Speech, Signal Processing, 1984, 32(6): 1109-1121

[4] N. Wiener, The Extrapolation, Interpolation, and Smoothing of Stationary Time Series With Engineering Applications. New York:Wiley, 1949.

[5] P. C. Hansen and S. H. Jensen, “FIR filter representations of Reduced rank noise reduction,” IEEE Trans. Signal Process., vol. 46, no.6, pp.1737--1741, Jun. 1998.

[6] S. Doclo and M. Moonen, “On the output SNR of the speech-distortion weighted multichannel Wiener filter,” IEEE Signal Process. Lett., vol.12, no. 12, pp. 809--811, Dec. 2005.

[7] Y. Ephraim and H. V. Trees, “A signal subspace approach for speech enhancement,” IEEE Trans. Speech Audio Process., vol. 3, no. 4, pp.251–266, Jul. 1995.

[8] D.L. Wang and J.S. Lim, “The unimportance of phase in speech enhancements”, IEEE Trans. Acoust., Speech and Signal Process., Vol.30, pp. 679-681, Aug. 1982.

[9] K. Paliwal, L Alsteris, "Usefulness of phase in speech processing",Proc. IPSJ Spoken Language Processing Workshop, Gifu, Japan, pp. 1-6, 2003.