International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, Number 18 (2015) pp 39550-39557

Intelligibility Of Speech Using Short Time Fourier Transform Phase Spectrum

Prabhavathi C. N

Research Scholar, Dept. of ECE Jain University Bangalore, India
Email:prabhacngowda@gmail.com

Dr. K. M Ravikumar

Research Guide Prof. and Head, Dept. Of ECE SJCIT, Chickballapur, India
Email :kmravikumar@rediffmail.com

Abstract
In speech area it is known fact that the short-time phase spectrum plays very short role in human perception tasks and also in automatic speech recognition systems. The usefulness of information obtained from phase is explored in human speech perception as well as in automatic speech recognition. By conducting many human perception experiments, it is shown that the short-time phase spectrum contributes to speech intelligibility as much as the magnitude spectrum. The short-time Fourier transform (STFT) of a speech signal has two components: the magnitude spectrum and the phase spectrum. In this paper, the importance of short-time magnitude and phase spectra for speech perception is evaluated. It is shown in this paper that even for small window durations, the phase spectrum can contribute to speech intelligibility as much as the magnitude spectrum if the analysis–modification–synthesis parameters are properly selected.

Keywords: Short-time Fourier Transform, Phase spectrum, magnitude spectrum, Speech perception.

I. INTRODUCTION

Speech is a continuously varying acoustic pressure wave. The evolution of speech has made it the primary form of communication between humans. There occur conditions where we measure the speech signal which is usually in the time domain and then transform the speech signal to another form to enhance our ability to communicate. A very early example of this is the transduction by telephone handset of the continuously varying speech signals to a continuously varying electric voltage signal. The resulting signal can be transmitted and processed electrically with analog circuitry and transduced back by the receiving handset to a speech pressure signal.

Short Time Fourier Transform (STFT): In **Short Time Fourier Transform,** the signal is divided into small enough segments, where these segments (portions) of the signal can be assumed to be stationary.

Kannada Vowels and Consonants:
In this paper the experiment is conducted for the clean speech signal. The input is the recording in vowel-consonant-vowel format in linguistic language. The linguistic language considered is kannada language. The detailed description of vowels and consonants in kannada language are explained in the following table. The kannada language vowels and consonants and their equivalent English vowels and consonants are explained in the Table 1. There are 14 swaralagu (vowels) in Kannada.

TABLE 1: KANNADA VOWELS AND THEIR ENGLISH EQUIVALENT

Swaragalu (vowels in kannada)	Equivalent vowel (English)	Swaragalu (vowels in kannada)	Equivalent vowel(English)
ಅ	A	ಋ	Rū
ಆ	Ā	ಎ	Ye
ಇ	I	ಏ	Yē
ಈ	Ī	ಐ	Ai
ಉ	Ou	ಒ	O
ಊ	Oū	ಓ	Ō
ಋ	Ru	ಔ	Au

Yogavaahakas
There are 2 yogavaahakas- Ansuswara and visarga. The Table 2 shows the ansuswara and visargas with their english equivalent.

TABLE 2: KANNADA VOWELS AND THEIR ENGLISH EQUIVALENT

Anuswara	Equivalent (English)	Anuswara	Equivalent (English)
ಅಂ	Aom	ಅಃ	(ahā)

Consonant Letter:
Two categories of consonant letters (vyanjana) are defined in Kannada: the structured consonants and the unstructured consonants. The structured consonants are classified according to where the tongue touches the palate of the mouth and are classified accordingly into five structured groups. The structured consonants and the unstructured consonants with voiceless aspirate and nasal are explained in Table 3.

Structured Consonants:

TABLE 3: STRUCTURED CONSONANTS IN KANNADA

Voiceless aspirate	Voiceless aspirate	Voiced aspirate	Voiced aspirate	Nasal
ಕ (ka)	ಖ (kha)	ಗ (ga)	ಘ (gha)	ಙ (nga)
ಚ (cha)	ಛ (chha)	ಜ (ja)	ಝ (jha)	ಞ (ña)
ಟ (ṭ a)	ಠ (ṭ ha)	ಡ (ḍ a)	ಢ (ḍ ha)	ಣ (ṇ a)
ತ (ta)	ಥ (tha)	ದ (da)	ಧ (dha)	ನ (na)
ಪ (pa)	ಫ (pha)	ಬ (ba)	ಭ (bha)	ಮ (ma)

Unstructured Consonants
The unstructured consonants and their equivalent in English is as shown in Table 4. The unstructured consonants are not classified into five categories as in structured category.

TABLE 4: UNSTRUCTURED CONSONANTS IN KANNADA

ಯ (ya)	ಲ(la)	ಷ (ṣ a)	ಳ (ḷ a)
ರ (ra)	ವ (va)	ಸ (sa)	ೞ (l̲)
ಱ (r̲ a)	ಶ (śa)	ಹ (ha)	

Speech Enhancement:
The main aim of speech enhancement is to improve the quality of speech by using various algorithms. Improvement in various intelligibility and the overall perceptual quality of degraded speech signal using audio signal processing techniques is the major objective of speech enhancement.
The most important field in the speech enhancement is the enhancing of speech degraded by noise, or noise reduction. This enhanced speech is used for many applications such as speech recognition, mobiles, VoIP, teleconferencing subsystems and hearing aids.

II. LITERATURE SURVEY

The short-time Fourier transform of a speech signal has two parts the magnitude and the phase spectrum.
The improvement of one or more aspects of speech perception is important in many situations like environments with interfering background noises of streets, offices, schools etc. This is also important in speech identification systems, hearing aids and many more. Researchers and engineers have evolved to the wide number of methods to solve this problem. Still because of complexities, this area of research poses a considerable challenge.
A few studies have been reported in the literature which discuss whether the phase spectrum provides any information which can contribute to intelligibility for human speech recognition Schroeder [14], and Oppenheim and Lim [8] performed some informal perception experiments concluding that the phase spectrum is important for intelligibility when the window duration of the short-time Fourier transform (STFT) is large (Tw > 1s), while it seems to convey negligible intelligibility at small window durations(20–40 ms). Liu et al. (1997) have investigated the intelligibility of phase spectra through a more formal human speech perception study. Fig 1 shows the performance of magnitude only and phase only stimuli with respect to the window size.

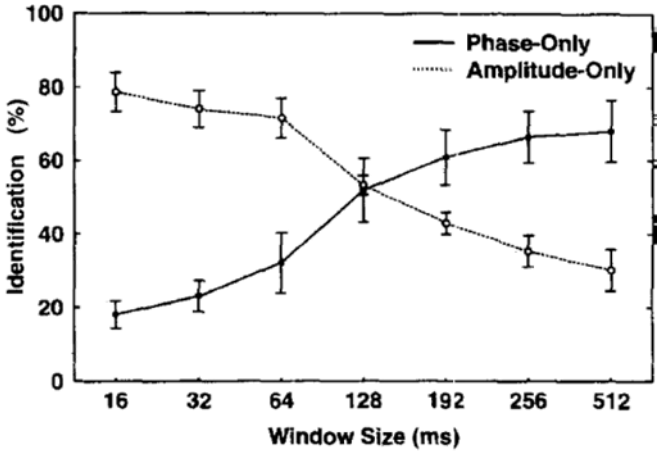

Fig. 1 Performance of magnitude only and phase only stimuli as a function of window size

Based on the literature survey, it is noticed that the most of the work is centered on the magnitude spectrum of speech intelligibility. It was assumed that phase does not contain any information or it contains very less information. As it is found that phase also contains some information which can contribute for speech intelligibility.

III. STFT MODIFICATION SYSTEM

The generalized modification system and the detailed modification systems are shown in the Fig. 2 and Fig. 3 respectively. The implementation details are also explained in this section.
The detailed analysis modification and synthesis technique is explained in the figure. The modification refers to making the magnitude unity. The phase only content is retained and the signal is reconstructed.

Generalized Modification System:

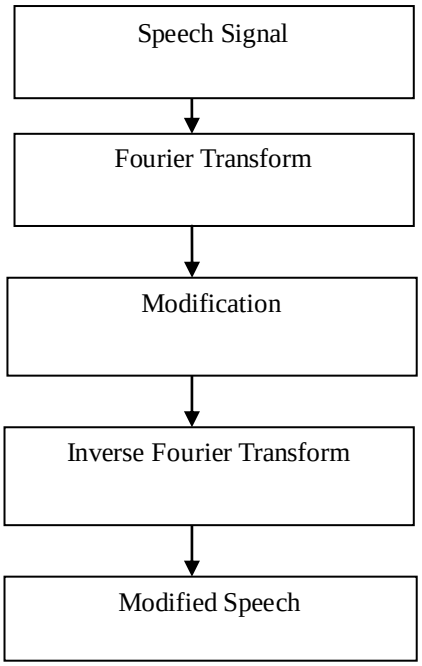

Fig. 2: Generalized Modification System

Implementation Details

1. The speech signal is a Real signal. Hence, the FFT of this signal obeys the conjugate symmetry property.
2. The speech signal is divided into smaller segments of 20ms duration.
3. The overlap of 50% is used in framing.
4. The Hamming Window is optimized to minimize the maximum (nearest) side lobes. The different windows are used in this paper.
5. 'N' point FFT is determined for each frame. Also the magnitude and phase of each FFT point is obtained.
6. Magnitude and phase are separated.
7. The magnitude in magnitude is done to obtain the signal with phase alone.
8. The Inverse Fourier Transform is applied tothe signal. of the signal is done
9. The reconstruction of signal is done in order to get the signal back to time domain.
10. Finally the output is analyzed.

The Experiments are conducted as follows:
The experiment 1 is carried out for identification of vowel consonant vowel for linguistic language. Here the experiments is carried for magnitude only and phase only for different windows and results are tabulated. The experiment 2 is carried out for noisy signals for speech enhancement as similar to experiment 1.

Detailed Modification System

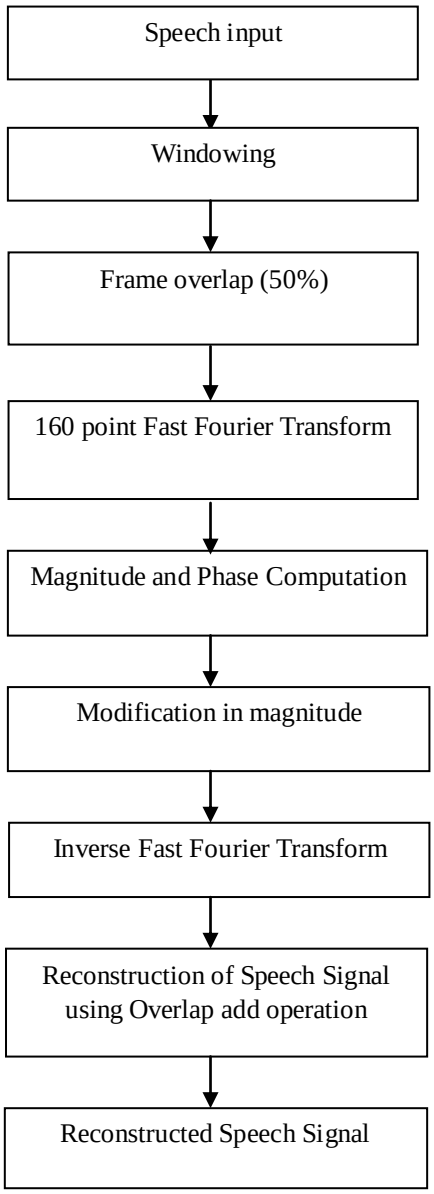

Fig 3. Detailed Modification System

IV. STFT ANALYSIS

Though speech is a non-stationary signal, it is generally assumed to be quasi-stationary and, therefore can be processed through a short-time Fourier Transform analysis. The term 'short-time' implies a finite-time window over which the properties of speech may be assumed stationary[11].

$$S(f, t) = \int_{-\infty}^{\infty} s(\tau)\, w(t-\tau) e^{-j2\pi ft} d\tau \quad (1)$$

where w(t) is a window function of duration Tw.
The STFT of a speech signal s(t) is given in Eq.(1) In speech processing, the Hamming window function is typically used and its width Tw is normally 20–40ms. S(f, t) can be decomposed as given in Eq.(2)

$$S(f, t) = |S(f,t)| e^{j\psi(f,t)} \quad (2)$$

where |S(f, t)| is the short-time magnitude spectrum and $\psi(f,t)$=angle(S(f, t)) is the short-time phase spectrum.

The signal s(t) is completely characterized by its short-time magnitude and phase spectra. The aim is to determine the contribution that the phase and magnitude spectra provide towards speech intelligibility. Accordingly, stimuli are created either from phase or magnitude spectrum. In order to construct, for example, an utterance with only phase spectra, the signal is processed through the STFT analysis using the Eq. (1) and the magnitude spectrum is made unity in the modified STFT $\hat{S}(f, t)$ is given in Eq.(3)

$$\hat{S}(f, t) = e^{j\psi(f,t)} \quad (3)$$

This modified STFT is then used to synthesize the signal $\hat{S}(t)$ using the overlap-add method.

The synthesized signal $\hat{S}(t)$ contains all of the information about the short-time phase spectra contained in the original signal s(t), but no information about the short-time magnitude spectra is obtained. This procedure is referred as the STFT phase-only synthesis. The modified STFT is computed as given in Eq.(4)

$$\hat{S}(f, t) = |S(f,t)|e^{j\phi} \quad (4)$$

Where 'ϕ' is a random variable uniformly distributed between 0 and 2π. It may also seem plausible to set to zero for all values of f and t.

V. EVALUATION METHODS

The following two methods are considered for evaluating the performance of STFT.

1. **Subjective Listening Test:**

Subjective listening test is one of the evaluation method. The degree of hearing of normal hearing people ranges from 0dB to 25dB and that of hearing impaired is above 30dB. In this test the subjects like normal hearing people [S1, S2, ….S10] and hearing impaired people [S1, S2, … S5] are made to listen the noisy and enhanced speech and asked to give the rating according to their hearing capability. The rating is from 0 to 5.

TABLE 5. MEAN OPINION RATINGS AND REMARKS

Opinion score	Remarks
1	More noisy speech
2	Average
3	Good
4	Better
5	Almost clean speech

The rating criteria are considered based on the quality of speech and subjects are asked to give the appropriate mean opinion score according to the speech quality. The rating criteria is given in the following table 5.

2. **Spectrogram Analysis:**

A spectrogram, is a visual representation of the spectrum of frequencies in a sound or other signal as they vary with time or some other variable. Also called spectral waterfalls, voiceprints, or voice grams.

V. RESULTS

The input to the clean speech signals are in the vowel-consonant-vowel format in linguistic language is considered in this paper. The speech database is as shown in the Table 6.

TABLE 6: INPUT SPEECH DATABASE

Input	Input	Input	Input
ಅಬ ಅ	ಅಗ ಅ	ಅನ ಅ	ಅಡ ಅ
ಅದ ಅ	ಅಕ ಅ	ಅಪ ಅ	ಅವ ಅ
ಅಹ ಅ	ಅಮ ಅ	ಅಸ ಅ	ಅಜ ಅ
ಅಧ ಅ	ಅಶ ಅ	ಅಠ ಅ	ಅಥ ಅ

Waveform representation of ಅಮ ಅ :

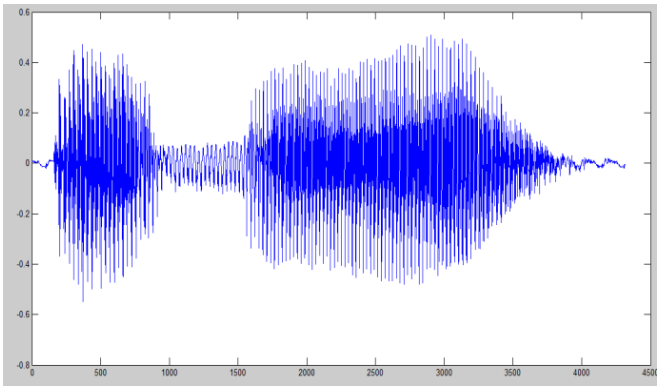

Fig 4. Time domain plot for input 'ಅಮ ಅ' signal

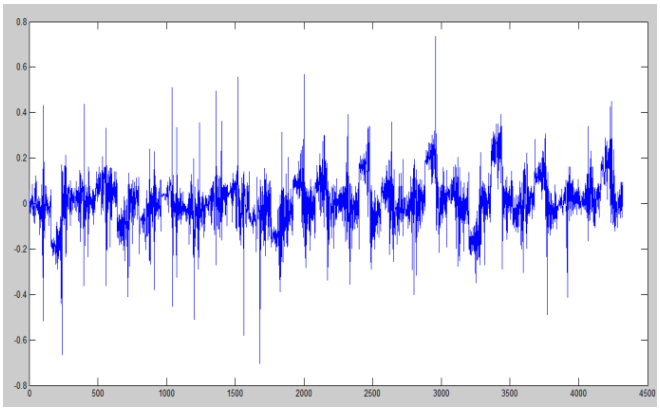

Fig 5. Reconstructed phase only plot using hamming window

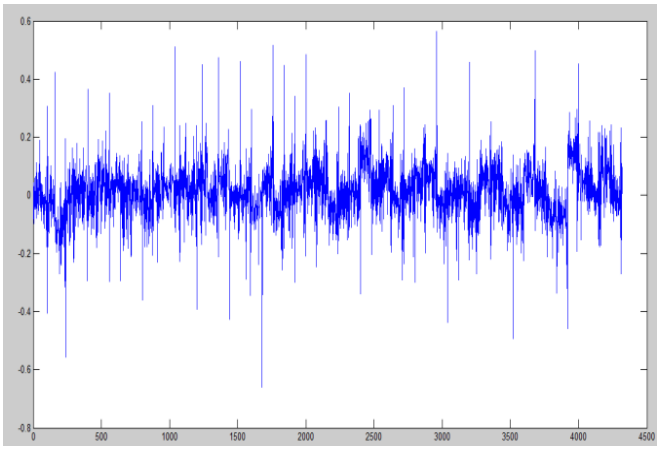

Fig 6. Reconstructed phase only plot using rectangular window

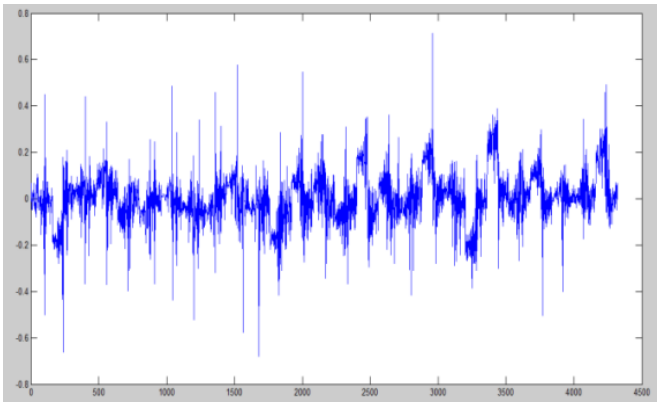

Fig 7. Reconstructed phase only plot using rectangular window

Results for vowel consonant vowel using subjective method:

The magnitude only and phase only stimuli were considered for smaller (20ms) duration and larger (1024ms) durations. The ratings given by the participants are as given in the table 7.

Table 7: Mean opinion scores of the participants for larger window durations of 1024ms

Stimuli Type	Hamming Window	Rectangular Window
Original	5	5
Magnitude only	4	4.5
Phase only	1	1

The magnitude only and phase only stimuli were sonsidered for smaller (20ms) duration and larger (1024ms) durations. The ratings given by the participants are as shown in the table 8.

Table 8: Mean opinion scores of the participants for smaller window durations

Stimuli Type	Hamming Window (20ms)	Rectangular Window (20ms)
Original	5	5
Magnitude only	3	3
Phase only	5	5

Hence from table 7 and 8 it is observed that for larger duration magnitude only gives better identification and for smaller duration phase only contributes for better identification of vowel consonant vowel.

Experiment 2: RESULTS FOR NOISY INPUT:

Table 9.Rating for subjective test analysis for different windows.

Participants	Different windows			
Listeners	***Hamming***	***Rectangular***	***Hanning***	***Bartlett***
S1	3	2	1.5	1.5
S2	2.5	2	1.5	1.5
S3	3.5	2.5	2.5	2.5
S4	3.5	2.5	3	1
S5	3.5	2.5	3	1
S6	3	3.4	2.2	2

The mean opinion score of the listeners for noisy input speech signal is shown in table 10.

The average mean for different windows is calculated in table 5 and a bar graph is drawn in order to analyse the response of the test. The bar graph is shown in fi.g 3.

TABLE 10. Mean of the ratings.

	Hamming	Rectangular	Hanning	Bartlet
Mean	3.16	2.48	2.03	1.6

From the subjective test and the bar graph it is concluded that the hamming window gives the best response when compared with different windows(hanning, bertlet, rectangular windows).

Bar graph repsentation of various windows with their mean is as shown in Fig 8.

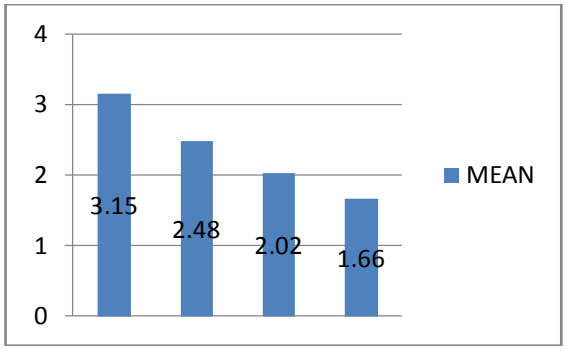

Fig 8. Bar graph resulting the performance of different window functions

Spectrogram Analysis:

Spectrogram analysis for experiment 1 (Clean Speech signal):

The spectrogram of the input clean speech signal is shown in fig 9(a). The phase only reconstrcuted spectrograms for hamming, hanning and rectangular windows are as shown in the figures from fig 9(b) – fig 9(d).

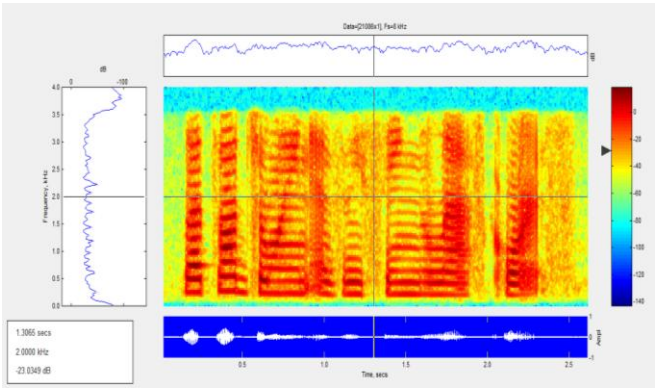

Fig 9(a). Spectrum of the Input signal

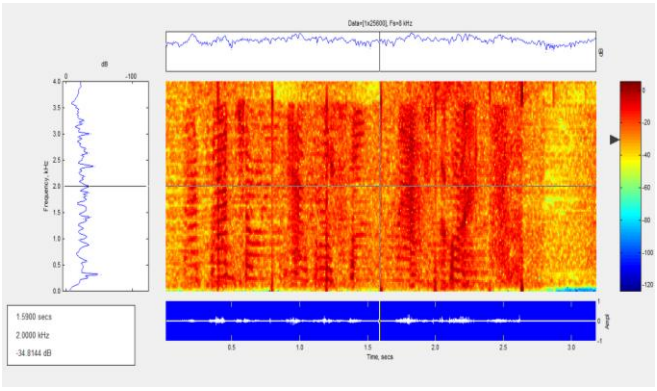

Fig 9(b). Reconstructed phase only Spectrum using hamming window

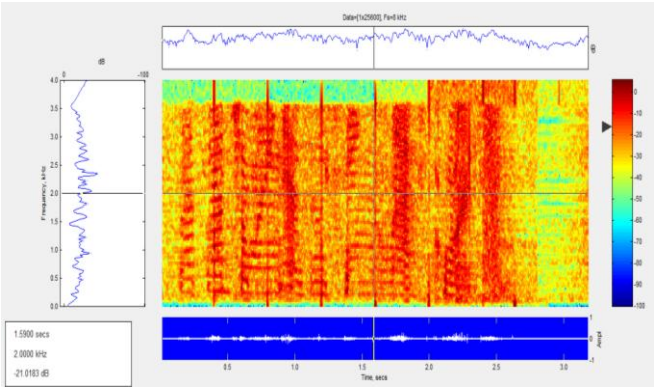

Fig 9(c). Reconstructed phase only spectrum using rectangular window

The spectrogram of the input noisy speech signal is shown in fig 10(a). The phase only reconstrcuted spectrograms for hamming, hanning and rectangular windows are as shown in the figures from fig 10(b) – fig 10(d).

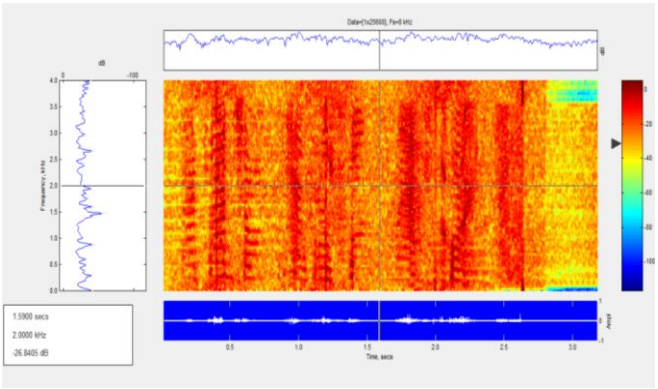

Fig 9(d). Reconstructed phase only spectrum using Hanning window

Spectrogram Analysis for Noisy speech signal input:

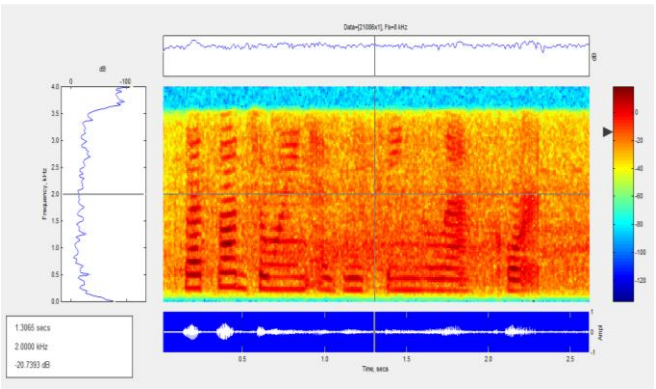

Fig 10(a). Spectrogram of the noisy speech signal

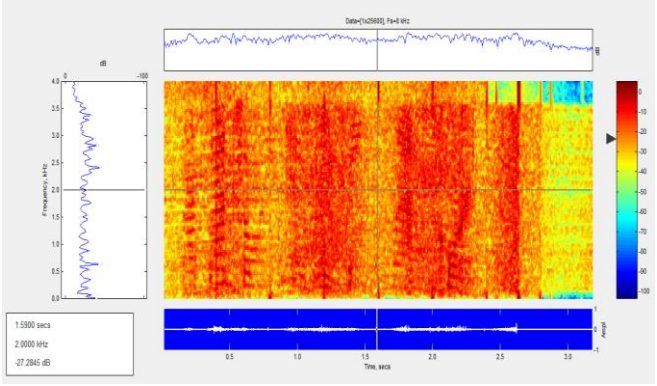

Fig 10(b). Reconstructed phase only spectrum using Hamming window

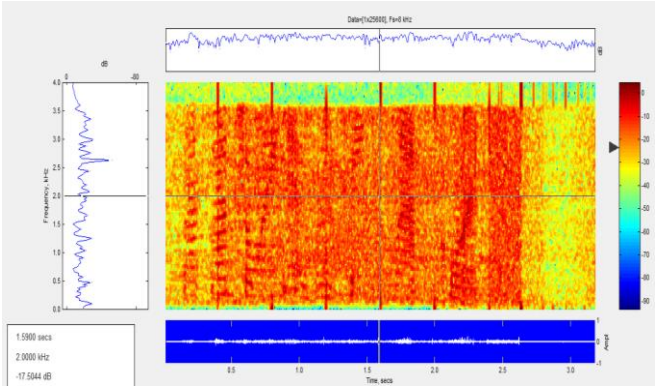

Fig 10(c). Reconstructed phase only spectrum using Rectangular window

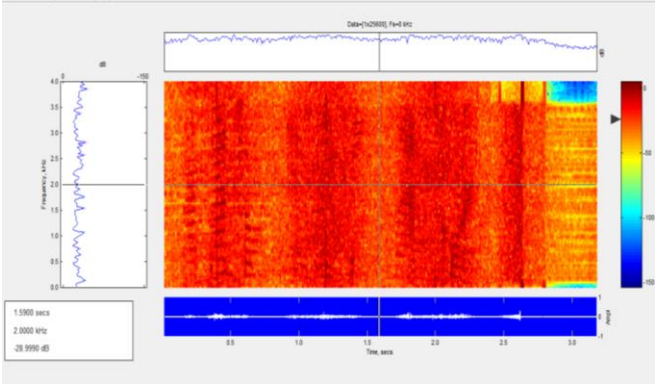

Fig 10(d). Reconstructed phase only spectrum using Hanning window

VI. CONCLUSION

A detailed analysis of short time Fourier transform using phase only by keeping magnitude constant for various windows were analysed with smaller window duration and larger window duration. And also analysed for magnitude only for different windows for both smaller and larger window duration.It is observed that for larger window magnitude gives better identification of consonants and for smaller window duration phase has better identification consonants.

Also Hamming window gives better performance in enhancing the quality of the noisy speech signal using phase only of STFT.

It may be possible to improve the intelligibility of the reconstructed stimuli if we make some assumptions about the speech signal such as if we assume speech to be a minimum phase signal, the phase spectra and magnitude spectra are related through Hilbert transform. This makes it possible to reconstruct the phase spectrum of a speech frame given its magnitude spectrum or to reconstruct the magnitude given its phase spectrum.

This makes it possible to reconstruct the phase spectrum of a speech frame given its magnitude spectrum or to reconstruct the magnitude given its phase spectrum. The stimuli were created from recordings made in clean conditions and also in noisy conditions.

REFERENCES

[1] Clark, J.E. & Mannell R.H. "Some comparative characteristics of uniform and auditorily scaled channel synthesis", *Proc. SST-88*, 282-287 1988.

[2] Fant, G. *Acoustic Theory of Speech Production*, (Mouton: The Hague, second printing 1970)

[3] Fant, G. "Analysis and synthesis of speech processes", in Malmberg, B. (ed.) *Manual of Phonetics* (North Holland: Amsterdam)1968

[4] Flanagan, J.L. & Golden, R.M. "Phase vocoder", *Bell Sys. Tech. J.*, 1493-1509 1966.

[5] Gold, B"Experiment with speechlike phase in a spectrally flattened pitchexcited "1964

[6] channel vocoder", *J. Acoust. Soc. Am.* 36, 1892-1894.

[7] J.B. Allen and L.R. Rabiner, "A unified approach to short-time Fourier analysis and synthesis" Proc. IEEE, Vol. 65, No. 11, pp. 1558-1564, 1977.

[8] D.W. Griffin and J.S. Lim, "Signal estimation from modified short-time Fourier transform", IEEE Trans. Acoust., Speech and Signal Processing, Vol. ASSP-32, pp. 236-243, 1984.

[9] L. Liu, J. He and G. Palm, "Effects of phase on the perception of intervocalic stop consonants", Speech Communication, Vol. 22, pp. 403- 417, 1997.

[10] A.V. Oppenheim and J.S. Lim, "The importance of phase in signals" Proc. IEEE, Vol. 69, pp. 529-541, 1981.

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

[12] M.R. Portnoff "Short-time Fourier analysis of sampled speech" IEEE Trans. Acoust., Speech and Signal Processing, Vol. ASSP-29, pp. 364-373, 1981

[13] L.R. Rabiner and R.W. Schafer, *Discrete-time speech signal processing, principles and practice*, Prentice Hall, Englewood Cliffs, NJ, 1978.

[14] M.R. Schroeder, "Models of hearing", Proc. IEEE, Vol. 63, pp. 1332- 1350, 1975.

[15] Atlas, L., Li, Q., Thompson, J., 2004. "Homomorphic modulation spectra. In: Proc. IEEE

Internat. Conf. Acoust. Speech Signal Process". (ICASSP), Vol. 2, Montreal, Quebec, Canada, pp. 761–764.

[16] Atlas, L., Vinton, M., "Modulation frequency and efficient audio coding". In: Proc. SPIE Internat. Soc. Opt. Eng., Vol. 4474, pp. 1–8 2001.

[17] Drullman, R., Festen, J., Plomp, R. "Effect of reducing slow temporal modulations on speech reception". J. Acoust. Soc. Amer. 95 (5), 2670–2680.1994

[18] Falk, T., Stadler, S., Kleijn, W.B., Chan, W.-Y., "Noise suppression based on extending a speech-dominated modulation band". In: Proc. ISCA Conf. Internat. Speech Comm. Assoc. (INTERSPEECH), Antwerp, Belgium, pp. 970–97 2007.

[19] Falk, T.H., Chan, W.-Y.,. A non-intrusive qualitymeasure of dereverberated speech. In: Proc. Internat. Workshop Acoustic. Echo Noise Control.2008

[20] Falk, T.H., Chan, W.-Y., Modulation spectral features for robust far-field speaker identification. IEEE Trans. Audio Speech Lang. Process. 18 (1), 90–100 2010.