International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 10 | Oct 2022 www.irjet.net p-ISSN: 2395-0072

# Enhanced modulation spectral subtraction incorporating various real time noise environment

**Nikita G Bangar [1], Dr. S. N. Holambe [2]**

*[1] Department of computer science and Engg, TPCT's COE Osmanabad, Osmanabad, India*
*[2] Professor, Department of computer science and Engg, TPCT's COE Osmanabad*

---

***Abstract*** — We humans may have different communication medium such as text or nonverbal communication but speech is the only active and powerful way of communication. Speech is the result of speech signals. These speech signals are pressure variations travelling through air. These variations in pressure are known as sound waves. Speech intelligibility can be degraded due to multiple factors, such as noisy environments, technical difficulties or biological conditions. It is possible to reduce the background noise, but at the expense of introducing speech distortion, which in turn may impair speech quality and intelligibility. Hence, the main challenge in designing effective speech enhancement algorithms is to suppress noise without introducing any perceptible distortion in the signal.

**Keywords—Enhanced Modulation Spectral Subtraction (EMSS)**

## 1 INTRODUCTION

In the process of speech enhancement, it is very important to acquaint with the speech output , the speech signal, and a lot of acoustic features of speech perception used by individuals. While doing so, we must preserve the properties of speech, need to have high quality and intelligibility of speech. This requires knowledge of Electronic Engineering, Biomedical, and Computer engineering.

In speech communication, intelligibility is a measure of how comprehensible speech is in given conditions. Speech signal can be evaluated depending on many important attributes such as quality.

Intelligibility is not equivalent and relation between these two attributes is not under-stood. As in case of quality assessment, intelligibility is not subjective since it has been readily measures, measure by counting correctly recognized speech contents. The method of Subjective intelligibility measurement cannot be easy for reproduced. These are also time consuming as well as costly.

Hence many studies have been put forth on objective intelligibility assessment measures in the literature staring from pioneer work by Frentch and Steinberg et. al. at Bell laboratories. The objective intelligibility parameters are considerably faster to apply. They are also less expensive as compared to subjective evaluation. It has the plus point that their results can be easily and accurately reproducible.

So to study the concurrent effect of real time noises like airport, car, restaurant, railway station etc. on proposed speech enhancement method is very essential. Also the effects of this enhancement on intelligibility of enhanced speech need to be explored. There for in this study we explore the effect of speech enhancement on speech intelligibility. Similarly in concerned with the field of artificial intelligence applications such as smart vehicle, robotic the intelligibility of speech play a vital role on a typical speech emotion recognition system (such as anger, happiness, fear, sadness etc.) In order to evaluate the potential performance of this study, objective evaluation has been performed.

## 2 EMSS METHOD

### 2.2 *Framework for enahancement*

Analysis modification and synthesis (AMS) is frequently applied in speech enhancement for signal enhancement. AMS method van be explained as follows.

1. Input signal breaking in to small with suitable window function. 2. Short Time Fourier Transform of each windowed frames along with appropriate some frame shift. 3. Then inverse Fourier Transform, 4. Synthesizing the original ignal by overlap and add (OLA) technique. The generalized speech model can be represented as follow in Equ. 1 with additive noise N(n) is

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

In this speech model x(n) is real time unpure speech, s(n) is idle speech and N(n) is additive noise.

Speech is non-stationary nature therefore analysis, speech is done over a short frame duration. Over the short frame duration the speech can be considered as station. Now applying short-Time Fourier Transform on each single frame. The STFT of noise corrupted speech in equ 2 is

$$X(n,k) = \sum_{l=-\infty}^{\infty} x(n)w(n-l) \times e^{\frac{-j\times 2\times \pi \times k\times l}{M}} \quad (2)$$

Where M is acoustic frame duration in samples, acoustic frame number is *l* and index of discrete acoustic frequency represented by k. In our method we applied modified W(n) Hamming window as an analysis window function for both domains acoustic and modulation. This Hamming window is proved to be efficient over numerous window functions.

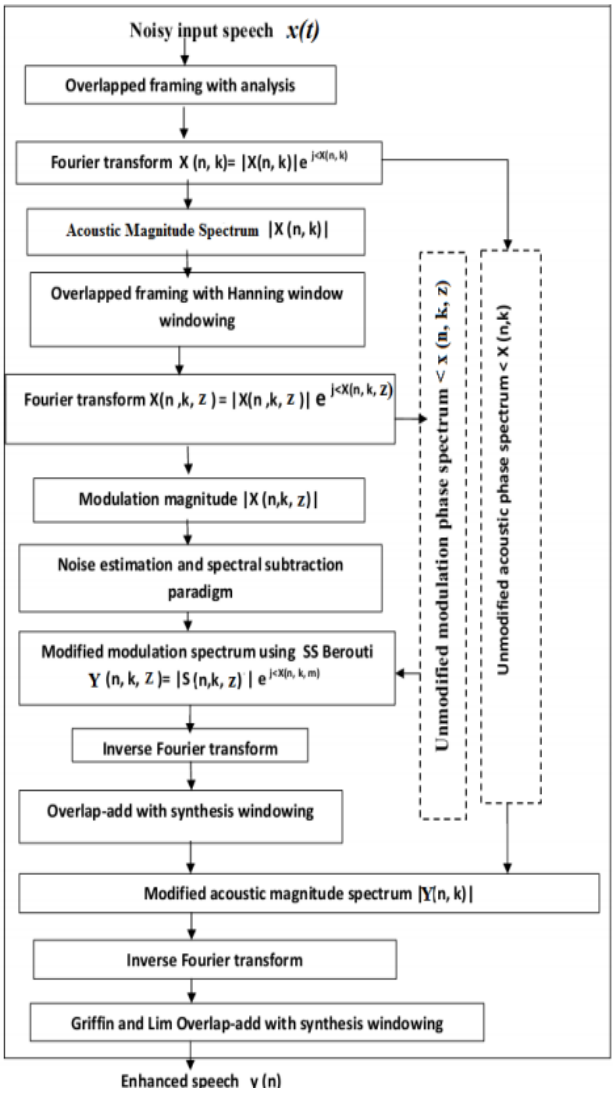


Fig. 1: Enahced modulation pectral subtraction AMS-based approach

## 2 Analysis

### 2.1 Effect of Clean Phase Spectrum in Modulation Domain

A speech signal on the application of STFT produce two parts: ie. magnitude and phase spectrum. Generally we to do not process the phase information since most of the features are computed complely from the spectrum of short-time magnitude. Thi is beacaue of following reaon: (i) experimental evaluation of some well-known subjective listening scores have demonstrated that the short-time phase spectrum (for tiny window 20 - 40 ms durations) provides a negligible precentage of intelligibility in spectral analysis, and (ii) difficulty from a signal processing viewpoint in using the short-time phase spectrum directly, due to phase-wrapping and other problems. Contrary to previous studies, over small window durations of 20-40 ms, results in indicate that the short-time phase spectrum will make speech intelligibility prominant. It is also studied that use of noisy phase in place of the clean signal phase is safe, provided the spectral SNR is high enough. In the analysis section described so far, we have considered modification on magnitude spectrum in modulation domain processing while leaving noisy phase spectrum as it is. It will be interesting to see how useful the clean phase estimate is in modulation domain processing.

$$\hat{S}(n,k) = X(n,k)^{\gamma} - \alpha N(n,k)^{\gamma} \quad (3)$$

In Equ 4 when $^{\gamma}$=1 it is Magnitude spectral subtraction and when $^{\gamma}$=2 it I power spectral subtraction. α is known a spectral subtraction factor.

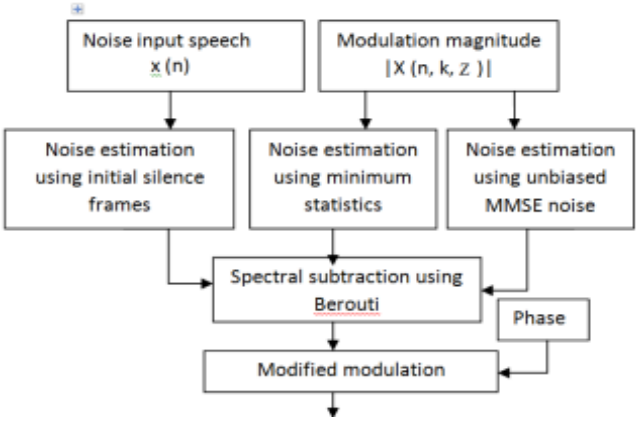


Fig. 2: Noise estimation and spectral subtraction paradigm

Fig. 3 shows noise estimation and subtraction paradigm.

processing by repeating AMS framework along time. The modulation spectrum X(n,k,z) is

$$X(n,k,z) = \sum_{l=-\infty}^{\infty} x(n)\, w(n-l) \times e^{\frac{-j\times 2\times \pi \times k\times l}{N}} ( \quad (6)$$

Where n, k is number of discrete acoustic frame.

### 2.2 Speech Corrupted by Babble Noise

In this section we evaluate the effect of several noise estimation methods on the proposed method. To reduce the computational load, optimal noise estimates for speech enhancement is computed. In modulation domain spectral subtraction, extensive experimental evaluation based on different noise estimation

methods are done. The babble noise, for instance, affect the low frequencies more than high frequency,

where most of the speech energy resides, are affected more than the high frequencies.

Therefore it is imperative to study effect of modulation

TABLE I. SHOWS SPEECH TRANSMISSION OBJECTIVE INTELLIGIBILITY (STOI) SCORES FOR CAR NOISE ENVIRONMENT PC_SP25_SN0_AMS67_G2_A_2, WHERE CAR BACKGROUND NOISE SPEECH SMPLE SP25 IN POWER SPECTRAL SUBTRACTION (G=2) AND ALPHA = 2

| | Input SNR (dB) | | | |
|---|---|---|---|---|
| **Noisy s** | **0dB** | **5dB** | **10dB** | **15dB** |
| | STOI = 0.6113 loss = 0.8908 | STOI = 0.7392 loss = 0.8599 | STOI = 0.852 loss = 0.8294 | STOI = 0.9216 loss = 0.8189 |

domain spectral subtraction in babble noise, and also to prevent destructive subtraction of the speech while moving most of the residual noise. In the proposed EMSS method, it is observed that during frame shift and at large frame duration, no appreciable effect of noise renewing is found during the modulation domain processing in experimental evaluation.

### 2.3 Database

We have used two standard speech database, NOIZEUS and EMOCAP. NOIZEUS consist of 30 sentences with male female speaker which is freely available and EMOCAP is paid dataset and is not freely accessible.

TABLE II. THE MODULATION DOMAIN PROCESSING IN DIFFERENT ASPECTS OF OVERALL IGNAL QUALITY SCORES FOR PROPOSED SPEECH ENHANCEMENT TECHNIQUE AT DIERENT INPUT SNR

| | Input SNR (dB) | | | |
|---|---|---|---|---|
| **Noisy s** | **0dB** | **5dB** | **10dB** | **15dB** |
| Speech sample | Csig=2.599266 Cbak=1.953793 Covl=2.161618 | Csig=3.147339 Cbak=2.344388 Covl=2.632825 | Csig=3.613310 Cbak=2.665438 Covl=3.015106 | Csig=3.835048 Cbak=2.87291 Covl=3.220404 |

noise estimation is evaluated by the application of NOIZEUS speech corpus database. The speech emotion stimuli such as anger, happy, fear and neutral are taken from speech emotion database IMMOCAP.

The clean speech emotion stimuli are the degreed by different noise type such as airport, car, train and traffic emotion speech stimuli. We evaluate performance result of proposed EMSS method in terms of objective evaluation parameters such as SNR seg., PESQ

### 2.4 Result Analysis

Table 1 shows pc_sp25_sn0_ams67_g2_a_2, where car background noise speech sample sp25 in power spectral subtraction (g=2) and alpha = 2.

TABLE III. INPUT SNR (DB) FOR NOIZEUS UTTERANCES (SP25) FOR SPEECH CORRUPTED CAR NOISE

| | Input SNR (dB) | | | |
|---|---|---|---|---|
| **Noisy s** | **0dB** | **5dB** | **10dB** | **15dB** |
| | LLR=1.048523 SNRseg=-2.4246 WSS=64.453471 PESQ=1.93246 | LLR=0.812026 SNRseg=-0.0347 WSS=53.152166 PESQ=2.269127 | LLR=0.580774 SNRseg=2.20952 WSS=45.073462 PESQ=2.52667 | LLR=0.516235 SNRseg=3.563 WSS=38.07301 PESQ=2.6797 |

Table 2 shows pc_sp25_sn0_ams67_g2_a_2, where car background noise speech smple sp25 in power spectral subtraction (g=2) and alpha = 2 for objective parameter LLR, Segmental SNR (SNRseg), weighted spectral slope (WSS) and perceptual evaluation of speech quality (PESQ) scores.

Table III shows LLR, PESQ, WSS scores for car noise environment where car background noise speech in power spectral subtraction (g=2) and alpha = 2

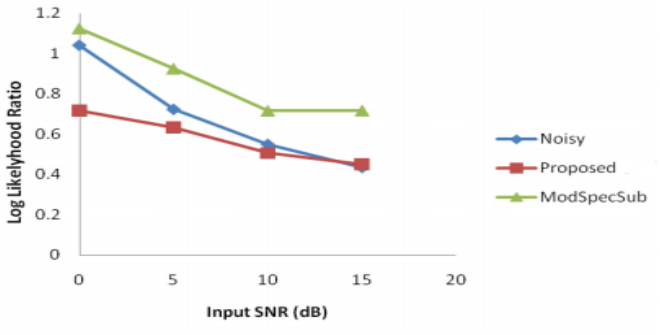


Figure 3: LLR scores improvement for babble noise.

Figure 3 shows the LLR improvement compared with Paliwals ModSpecSub in case of speech stimuli corrupted by babble noise. Lower LLR values indicate closeness between the spectral magnitudes of the clean and enhanced speech signals. From results in Figure

4 it is confirmed that the proposed method gives consistent lower values of LLR.

## 3 CONCLUSION

So to study the concurrent effect of real time noises like airport, car, restaurant, railway station etc. on proposed speech enhancement method is very essential. Also the effects of this enhancement on intelligibility of enhanced speech need to be explored. There for in this study we explore the effect of speech enhancement on speech intelligibility. Similarly in concerned with the field of artificial intelligence applications such as smart vehicle, robotic the intelligibility of speech play a vital role on a typical speech emotion recognition system (such as anger, happiness, fear, sadness etc.)

To investigate speech emotion recognition performance of proposed EMSS enhancement method applied, as pre-processing stages in IOVT to speech recognition systems different speech emotion and noise type are employed. The speech emotion stimuli such as anger, happy, fear and neutral are taken from speech emotion database IMMOCAP. The clean speech emotion stimuli are the degreed by different noise type such as airport, car, train and traffic at different input SNR to construct noisy emotion speech stimuli. We evaluate performance result of proposed EMSS method in terms of objective evaluation parameters such as LLR, SNR seg., PESQ, SNR loss. Here we meet best scores by proposed EMSS i.e. about 50 % improvement than ModSpecSub and noisy speech stimuli. For airport noise SNR seg. improvement is 55.14 %. For car noise SNR seg. is improved by 60.97 %.

## REFERENCES

[1] Sunil Kamath and Philipos Loizou. A multi-band spectral subtraction method for enhancing speech corrupted by colored noise. In ICASSP, volume 4, pages 44164{44164. Citeseer, 2002.

[2] Kuldip Paliwal, Kamil Wojcicki, and Belinda Schwerin. Single-channel speech en-hancement using spectral subtraction in the short-time modulation domain. Speech communication, 52(5):450{475, 2010.

[3] Rainer Martin. Bias compensation methods for minimum statistics noise power spec-tral density estimation. Signal Processing, 86(6):1215{1229, 2006.

[4] Yariv Ephraim and David Malah. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Transactions on acous-tics, speech, and signal processing, 32(6):1109{1121, 1984.

[5] P Loizou. Noizeus: A noisy speech corpus for evaluation of speech enhancement algorithms.Speech Commun, 49:588{601, 2017

[6] Philipos C Loizou.Speech enhancement: theory and practice. CRC press, 2007.

[7] Nathalie Virag. Single channel speech enhancement based on masking propertiesof the human auditory system. IEEE Transactions on speech and audio processing, 7(2):126{137, 1999

[8] Rainer Martin. Noise power spectral density estimation based on optimal smoothing and minimum statistics.IEEE Transactions on speech and audio processing, 9(5):504{512, 2001.

[9] Yi Hu and Philipos C Loizou. Evaluation of objective quality measures for speech en-hancement.IEEE Transactions on audio, speech, and language processing, 16(1):229{238, 2008.

[10] PC Loizou. Subjective evaluation and comparison of speech enhancement algorithmsSpeech Commun, 49:588{601, 2007

[11] Pavan D Paikrao, Sanjay L. Nalbalwar, 'Analysis Modification synthesis based Opti-mized Modulation Spectral Subtraction for speech enhancement',International jour-nal of Circuits, Systems and Signal Processing, Vol . 11, pg 343-352,2017.