Enhanced modulation spectral subtraction for IOVT speech recognition application

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 share our emotions, thoughts by speaking with each other. If we consider an automatic machine, voice control is the most convenient way for us than carrying a remote controller. Automatic speech recognition system(ASRS) works by breaking down the audio of a speech recording into individual sounds, analyzing each sound, using algorithms to find the most probable word fit in that language, transcribing those sounds into text and use that text as a command. But here comes a drawback due to noisy environment. We cannot deliver a clean voice to a machine since speech is degraded by background noise signals. This degraded speech reduces the speech recognition rate. The purpose of this proposed method is the enhancement of noisy speech signals and its effects on emotion recognition. This method can be applied as pre-processing stage to smart Internet of Vehicle Things (IOVT). The quality of enhanced speech is evaluated by subjective and objective evaluation parameters such as, PESQ, SNRLoss, and overall signal quality. 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 %. For traffic and train noise SNR seg. Improvement is 44.99 % and 39.69 % respectively at 0 dB input SNR is reported.

Keywords—IOVT internet of vehicle Things, Enhanced Modulation Spectral Subtraction (EMSS)

1 INTRODUCTION

Recently there is a huge demand of preprocessing stage in smart automatic vehicles. Many speech enhancement systems may degrade speech recognition performance of emotions due to background noise.

Figure 1 shows generalized system applications for secure IoVT .

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.

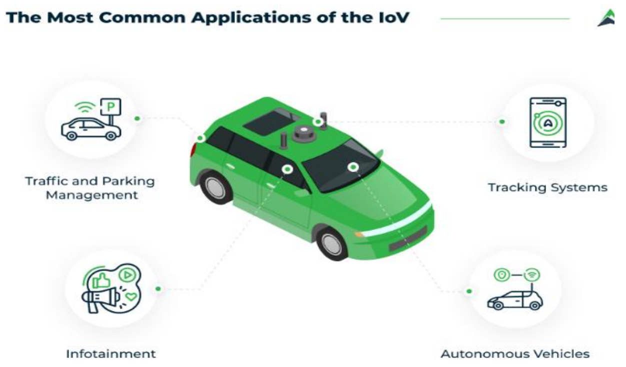
The Most Common Applications of the IoV
Traffic and Parking Management
Tracking Systems
Infotainment
Autonomous Vehicles

Fig 1: Generalized IOVT system

To investigate the effect of background noise (such as airport, car, restaurant, railway station etc.) on a typical speech emotion recognition system (such as anger, happiness, fear, sadness etc.) using proposed Enhanced Modulation Spectral Subtraction (EMSS) method as a pre-processing stage. In order to evaluate the potential performance of proposed approach, objective evaluation have been performed.

In this study we investigated the speech emotion recognition problem under various real-time noise conditions by considering modulation domain processing as a preprocessing stage. To investigate speech emotion recognition performance of proposed EMSS enhancement method applied, as preprocessing stages, 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 degraded by different noise type such as airport, car, train and traffic at different input SNR to construct noisy emotion speech stimuli.

AMS framework processes the degraded signal in the frequency domain using Fourier analysis. For spectral analysis, many speech processing techniques employ AMS framework. In order to achieve this some speech enhancement methods some method employ short time Fourier transform (STFT) [1, 3, 4]. Here in this thesis, the later approach of STFT spectrum which is composed of short time magnitude spectrum and short time phase spectrum is investigated. The modification on this magnitude spectrum is applied to enhance degraded speech. Hence, we have to built the phase spectrum the magnitude before the stage of synthesis. After the overlap-add stage that rebuilt stimuli generated are selected for the listening tests that are subjective test and objective tests to check out enhanced speech quality. To analyze it, we require a particular framework in order to attain modifications in short time spectral domain. We will consider an AMS framework established by Allen Rabiner, 1977 Grifin Lim 1984. In order to apply Fourier transform, it is compulsory that the input signal be in infinite in length and stationary in nature. This is contradictory to both requirements as speech is non-stationary and infinite in length. The speech signal conveys information thus it cant be stationary. That is why for more obvious reasons, it is impractical to be infinite. Therefore to make Fourier transform practically, we need to use short-time analysis. The generalized AMS framework in figure 1 decomposes the speech signal into short time frames. Since speech can be considered as quasi-stationary, it can be analyzed frame wise using short-time Fourier Transform.

2 EMSS METHOD

2.2 *AMS method*

AMS method [40] is an efficient method for signal enhancement. AMS uses following steps.

First, framing of the input speech signal with suitable window function and Second, STFT of windowed frames with some frame shift. Third, inverse Fourier Transform and fourth retrieving signal by overlap and add (OLA) method. Let's consider additive noise scenario as in Eq.

$$x(n) = s(n) + N(n) \tag{1}$$

Where x(n) is noisy speech, s(n) is clean speech and N(n) is background noise. In this the discrete time index.

As due to non-stationary nature of speech the AMS framework, processing of speech is done over a short frame duration applying short-Time Fourier Transform. Now the STFT of noise m corrupted speech in equ 2 x(n) is

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

Where M is acoustic frame duration in samples, *l* is an acoustic frame number 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 acoustic and modulation domains. This Hamming window is found to be efficient over other window function. In modulation domain processing the AMS framework is repeated after acoustic domain processing. The speech signal spectral subtraction is done in modulation domain [2] speech signal with the speech enhancement technique [1, 2, 3] as shown in Figure 3. Now apply STFT to Equ 2, as which gives following

$$X(n,k) = S(n,k) + N(n,k) \tag{3}$$

Where X(n,k) is noisy speech, s(n,k) is clean speech and N(n,k) is background noise. The fourier transforms representation of X(n,k) is combination of acoustic magnitude spectrum, acoustic phase spectrum as shown in Eq. 4.

$$x(n,k) = |x(n,k)|ej < x(n,k) \tag{4}$$

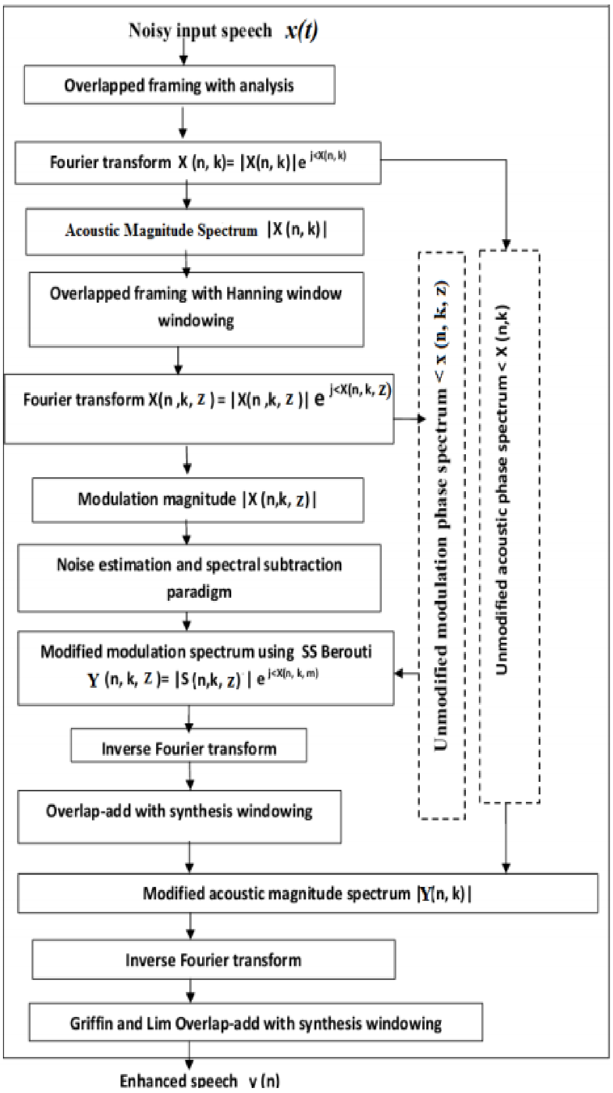
Noisy input speech x(t)
Overlapped framing with analysis
Fourier transform X (n, k)= |X(n, k)|e j<X(n, k)
Acoustic Magnitude Spectrum |X (n, k)|
Overlapped framing with Hanning window windowing
Fourier transform X(n ,k, Z) = |X(n ,k, Z)| e j<X(n, k, Z)
Modulation magnitude |X (n,k, z)|
Noise estimation and spectral subtraction paradigm
Modified modulation spectrum using SS Berouti Y (n, k, Z)= |S (n,k, z) | e j<X(n, k, m)
Inverse Fourier transform
Overlap-add with synthesis windowing
Modified acoustic magnitude spectrum |Y(n, k)|
Inverse Fourier transform
Griffin and Lim Overlap-add with synthesis windowing
Enhanced speech y (n)
Unmodified modulation phase spectrum < x (n, k, z)
Unmodified acoustic phase spectrum < X (n,k)

Fig. 2: Flow chart of a proposed EMSS, AMS-based speech enhancement method

2.3 Traditional Spectral Subtraction

Traditionally the spectral subtraction by S S boll method I done by subtracting short time spectral amplitude of the estimated noise from background noise. This subtraction yields negative spikes magnitudes spectra. To remove this noise flooring B a shown in Equ 5 is applied as a function of the over-subtraction factor. The modified spectrum is given by the Eq. 4

$$\hat{S}(n,k) = X(n,k)^{\gamma} - \alpha N(n,k)^{\gamma} \qquad (4)$$

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. Noise floor B is as follow

$$B = \beta \mid N(n,k,m) \mid^{\gamma} \qquad (5)$$

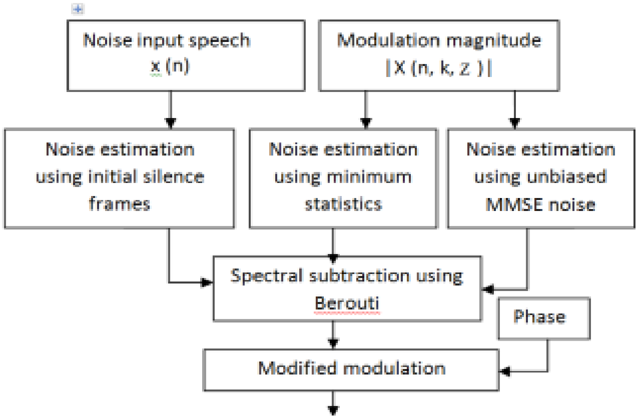
Noise input speech x (n)
Modulation magnitude |X (n, k, z)|
Noise estimation using initial silence frames
Noise estimation using minimum statistics
Noise estimation using unbiased MMSE noise
Spectral subtraction using Berouti
Phase
Modified modulation

Fig. 3: Noise estimation and spectral subtraction Paradigm

Fig. 3 shows noise estimation and subtraction paradigm.

The modulation spectrum X(n,k,z) is derived from traditional Allen and Rebiners 1977 AMS based acoustic spectrum elaborated in Section 2.2 . It is computed using every frequency bin achieved during acoustic spectrum transform by STFT. The frame by frame each frequency component derived in the acoustic processing by repeating AMS framework along time. The modulation spectrum X(n,k,z) is

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

Where n, k is number of discrete acoustic frame and index of discrete acoustic frequency respectively. z is known as an index of the discrete modulation frequency. The modulation frame duration L is in terms of acoustic frame. The w(n) is modified Hamming analysis window function. In our study the modified Hamming window with optimal frame duration of 128 ms and frame shift of 16 ms is applied for second AMS framework that is modulation domain.

3 MODIFICATION

Most important step in spectral subtraction for enhancement of speech is appropriate estimation noise. We examine 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. In the fiirst, estimation of noise using initial silence frame is done and in the second, minimum statistic noise estimation approach is used. The first approach employs a voice activity detection(VAD) algorithm to renew the noise during pause between the utterances and nonspeech frames. Hence, there is greater computational load. In the proposed EMSS method, it is observed that during frame shift and atlarge frame duration, no appreciable effect of noise renewing is found during the modulation domain processing in experimental evaluation. Therefore, to reduce the computational load on the conventional ModSpecSub [2] method, we deter the use of the VAD [7] algorithm to update noise and apply minimum statistic noise estimation perspective in the modulation domain.

3.1 Modulation domain spectral enhancment subtraction:

Following Eq. 7 computes modulation domain spectra

$$\hat{S}(n,k,m) = \begin{cases} (X_R(n,k,m)\,|^{\gamma} - \alpha\,|\,N(n,k,m)\,|^{\gamma})^{1/\gamma}, & \\ if X_R(n,k,m)\,|^{\gamma} - \alpha\,|\,N(n,k,m)\,|^{\gamma} & \geq B \\ \beta\,|\,N(n,k,m)\,|^{\gamma}\ otherwise & \end{cases}$$

(8)

Where clean speech signal estimates is S(n, k, z).

3.2 Database

The Modulation domain processing in different aspects of 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 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 SNR seg., PESQ

3.3 Result Analysis

The over-subtraction factor α is manipulate the amount of subtraction of noise estimate from the noisy speech signal. Table 1 shows the confusion matrix for car noise.

TABLE1 Confusion matrix results for different methods in car noise Over-subtraction and is traditionally can be used between 0-6. $^{\gamma}$=1 it is Magnitude spectral subtraction and when $^{\gamma}$=2 it I power spectral subtraction. In minimum statistics method [12, 13 of noise estimation case α, this should be between 0 and 3. The enhanced output results were obtained at α= 1. The second noise estimation method unbiased MMSE noise estimator, yields enhanced objective scores between 0-1 for α For unbiased MMSE noise estimator It has been observed that α= 0.1 yields enhanced objective scores, but for α= 1, objective scores decays. In our study over-subtraction factor α is 0:1≤α≤3. For implementation and result analysis we used α= 1, β= 0:0001 and power spectral subtraction domain. The observation study shows that spectral subtraction gives enhanced objective scores at $^{\gamma}$= 2, α= 1. Here we meet best scores by proposed EMSS i.e.

Table -1: Speech Recognition scores: **car noise**

Car Noise							
		Recognised (%)					
		Type of Stimuli	Neutral	Anger	Joy	Sad	Fear
TESTED	**Neutral**	Noisy	16.5	0	33	50.5	0
		EMSS	**30.5**	16.5	26.5	21.5	5
		Traditional Spectral (S S Boll)	12.5	18.5	9.5	51.5	8
		Paliwal's ModSpecsub	8.25	12.5	79.25	0	0
	Anger	Noisy	0	18.8	28.5	14.7	8
		EMSS	0	93.5	6.5	0	0
		Traditional Spectral (S S Boll)	12.5	9.5	18.5	51.5	8
		Paliwal's ModSpecsub	0	52.5	47.5	0	0
	Joy	Noisy	11.5	38.5	41.5	8.5	0
		EMSS	0	19.5	81.5	0	0
		Traditional Spectral (S S Boll)	12.5	51.5	18.5	8	9.5
		Paliwal's ModSpecsub	0	23.5	76.5	0	0
	Sad	Noisy	5.5	0	44.5	9.5	40.5
		EMSS	19.5	0	0	46.5	34.25
		Paliwal's ModSpecsub	0	8.5	68.25	23.25	0
	Fear	Noisy	0	0	52.2	32.5	15.25
		EMSS	4.5	0	0	42.5	53
		Traditional Spectral (S S Boll)	8	9.5	18.5	12.5	51.5
		Paliwal's ModSpecsub	6.2 5	0	0	46.25	47.5

about 52 % improvement than Paliwals ModSpecSub and noisy speech stimuli. For airport noise SNR seg. Improvement is 55.14 %. For car noise SNR seg. is improved by 65.82 %. For train and traffic noise SNR seg. enhancement is 39.69 % and 40.50 % respectively at 0 dB input SNR.

4 CONCLUSION

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. For the speech emotion type anger and happy (with different noise type and input SNR) on structured by treatment type of the proposed scheme, as compared with the traditional ModSpecSub method. 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 %. For traffic and train noise SNR seg. Improvement is 44.99 % and 39.69 % respectively at 0 dB input SNR is reported.

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.