Journal of Basic and Applied Engineering Research
p-ISSN: 2350-0077; e-ISSN: 2350-0255; Volume 2, Issue 22; October-December, 2015, pp. 1891-1895

http://www.krishisanskriti.org/Publication.html

Objective Quality and Intelligibility Evaluation for Speech Enhancement Algorithms

Kalpana Naruka[1] and O.P. Sahu[2]

[1,2]National Institute of Technology, Kurukshetra, Haryana, India
E-mail: [1]knaruka79@gmail.com, [2]ops_nitk@yahoo.co.in

Abstract—*This paper is about the objective assessment of some existing speech enhancement techniques including a new paradigm for enhancement of speech i.e. compressed sensing (CS).CS is remarkably advancing in speech processing applications because it provides sampling, compression and encryption simultaneously. Several objective measures are computed to assess the speech quality and intelligibility of speech enhanced by speech enhancement algorithms. In this paper, a total of 6 algorithms are evaluated encompassing Spectral Subtraction, Wiener filter, Log-MMSE, Kalman filter, Signal Subspace and CS. Performance of CS is evaluated by computing objective measures of reconstructed speech via Compressive Sampling Matching Pursuit (CoSaMP).Appreciable results are observed using CS in terms of perceptual evaluation of speech quality (PESQ) of objective quality score, SNRloss, LLR, SegSNR etc. in acceptable range.*

1. INTRODUCTION

Speech Enhancement aims at improving quality and intelligibility of speech signal. These two are the perceptual aspects of evaluating the goodness of speech. Quality is highly subjective because it refers to individual preference of listeners. One listener may find utterance good, at the same time another listener may find it bad or even very bad. Intelligibility tells the no. of correctly identified words by the listeners. The Quality and intelligibility of speech signal can be degraded by various types of noises and phenomena's such as reverberation, echoes, multipath propagation, spectral distortions, background noises etc. in communication link. Enhancement of speech can be done at any point in the link or at receiver itself to retrieve the original message with minimum loss.In a past few years, several speech enhancement algorithms have been proposed tosuppress noise in speech communication applications. It is still challenging to find such an algorithm that works in real time situations. A fair comparison is tough among algorithms because lack of common speech database and types of noises used in evaluating performance of algorithms.

Most of the speech enhancement algorithms are capable of improving only speech quality. They hardly improve speech intelligibility. The reason behind thisis not having a good estimate of the background noise spectrum which is required for implementation of most of the algorithms. In practice, background noise is of non-stationary type that makes impossible to track the spectrum accurately.

Subjective listening tests provide the most accurate results for evaluating speech because no objective measure yet matched to complex accuracy of human auditory system. The main disadvantage of these tests is requirement of healthy normal hearing and experienced listeners. Another minus is time consumption and expenses to conduct these tests so objective measures need to be developed. The aim is to obtain high correlation with subjective tests and to assess quality without need of original speech. Several objective measures have been proposed such as Itakura-Saito distortion, Articulation Index, Segmental SNR, LLR, WSS, PESQ, SNRloss etc. PESQ has 93.5% correlation with subjective listening test that is highest correlation attained by any objective measure [3]. In reality, objective measures require original speech to assess quality except PESQ.

In this paper, several objective and subjective measures are defined to assess quality and intelligibility of enhanced speech. Total 6 algorithms are evaluated encompassing Spectral Subtraction, MMSE, Wiener filter, Kalman filter, Signal Subspace and Compressed Sensing in terms of objective and subjective measures. Simulations results and Conclusions are summarized at end of paper.

2. SUBJECTIVE EVALUATION

Subjective listening test methodology is designed by ITU in recommendation ITU-T P.835 [15]. This methodology was designed to evaluate the speech quality along three dimensions: signal distortion (SIG), background distortion (BAK) and overall quality(OVRL). This evaluation removes the uncertainty of listeners in listening tests by increased readability in terms of rating given to the enhanced speech on a five point scale. The mean opinion score (MOS) for SIG, BAK and OVRL scales are described in table 1, 2 and 3 respectively taken from [1].

Table 1: Description of SIG scale

Score	Quality description
1	Very unnatural, very degraded
2	Fairly unnatural, fairly degraded
3	Somewhat natural, somewhat degraded
4	Fairly natural, little degradation
5	Very natural, no degradation

Table 2: Description of BAK scale

Score	Quality description
1	Very conspicuous, very intrusive
2	Fairly conspicuous, somewhat intrusive
3	Noticeable but not intrusive
4	Somewhat noticeable
5	Not noticeable

Table 3: Description of OVRL scale

Score	Quality description
1	Bad
2	Poor
3	Fair
4	Good
5	Excellent

3. OBJECTIVE EVALUATION

Several objective measures have been proposed to assess the quality and intelligibility. This paper reviews some existing objective measures: SegSNR, LLR, WSS, SNRloss, SNRESC, PESQ, composite measures.

3.1 Overall SNR

It is the most simple and common method because of directly comparing original and processed waveforms in time domain. In computation of this measure synchronization of original and processed speech signal is mandatory for good performance. It is computed as follows:

$$Overall\ SNR = 10\log_{10}\frac{\sum_{i=1}^{N} s^2(i)}{\sum_{i=1}^{N}(s(i)-x(i))^2} \quad (1)$$

where$s(i)$ and $x(i)$ are original and processed speech samples indexed by i and N is total no. of samples.

3.2 Segmental SNR (SegSNR)

It is frame based measure to assess speech quality. It is calculated by averaging of frame level SNR estimates in below manner [11]:

$$SegSNR = \frac{10}{N}\sum_{i=1}^{N}\log_{10}\left[\frac{\sum_{j=0}^{L-1} s^2(i,j)}{\sum_{j=0}^{L-1}[s(i,j)-x(i,j)]^2}\right] \quad (2)$$

where N is the no. of frames, L is frame length, $s(i,j)$ and $x(i,j)$ are the i^{th} frame of original and enhanced speech respectively. Difference between $s(i,j)$ and $x(i,j)$ is the i^{th} noise frame. The range of SegSNR is limited to [-10, 35] dB .

3.3 Log-likelihood Ratio (LLR)

It is a linear predictive coding (LPC) based measure showing similarity of spectral envelope and defined as [1]:

$$d_{LLR}(\vec{a}_x, \vec{a}_s) = \log\left(\frac{\vec{a}_x \mathcal{R}_s \vec{a}_x^T}{\vec{a}_s \mathcal{R}_s \vec{a}_s^T}\right) \quad (3)$$

where$\vec{a}_x$ and $\vec{a}_s$ are the LPC vectors of enhanced speech frame and original speech frame respectively. $\mathcal{R}_s$is the autocorrelation matrix of original speech signal. The range of LLR values are limited to [0,2].

3.4 Weighted Spectral Slope (WSS)

It calculates the weighted difference between the spectral slopes in each frequency band. This is done by measuring the difference between adjacent spectral magnitudes.This measure is evaluated as [1]:

$$d_{WSS} = \frac{1}{N}\sum_{i=0}^{N-1}\frac{\sum_{j=1}^{K} W(j,i)(S_s(j,i)-S_x(j,i))^2}{\sum_{j=1}^{K} W(j,i)} \quad (4)$$

Where N is no. of frames, K no. of bands and $W(j,i)$are the weights. $S_s(j,i)$and$S_x(j,i)$ are the spectral slopes for j^{th} frequency band at frame i of the clean and enhanced speech respectively.

3.5 Perceptual Evaluation of Speech Quality (PESQ)

PESQ uses the psychoacoustic model to predict the subjective quality of speech. With the help of model, original and degraded signals are drawn into an internal representation. This measure is defined in ITU-T recommendation P.862.ITU mapping function converts the raw PESQ score [0.5-4.5] onto the MOS-LQO (mean opinion score-listening quality objective) scale in range of [1= bad to 5= excellent].

The output mapping function used in PESQ is computed as a linear combination of average disturbance value D_{ind} and average asymmetrical value A_{ind} as defined in [1]:

$$PESQ = a_0 + a_1 D_{ind} + a_2 A_{ind} \quad (5)$$

Whereparameters a_0,a_1 and a_2 are chosen as different set to optimize PESQ measure in three rating scales signal distortion, background distortion and overall quality.

3.6 Composite Measures

Composite measures are calculated by combining some of above measures and compared with SIG, BAK and OVRL scores obtained by subjective listening tests.

$$Csig = 3.093 - 1.029 * LLR + 0.603 * PESQ - 0.009 * WSS \quad (6)$$

$$Cbak = 1.634 + 0.478 * PESQ - 0.007 * WSS + 0.063 * SegSNR \quad (7)$$

$$Covl = 1.594 + 0.805 * PESQ - 0.007 * WSS - 0.512 * LLR \quad (8)$$

High SIG score reflects lower signal distortion and higher BAK score reflects lower noise distortion. OVRL score is subjected to overall quality of processed speech.

3.7 SNRloss

Most of the objective measures discussed above mainly concerned with predicting quality of enhanced speech. They do not convey any information regarding the intelligibility of enhanced speech. SNRloss is the objective measure that can be used for predicting intelligibility of processed or enhanced speech.

The SNRloss is defined as follows [4]:

$$L(j,i) = SNR_s(j,i) - SNR_{\hat{s}}(j,i)$$
$$= 10\log_{10}\frac{S^2(j,i)}{\hat{S}^2(j,i)} \quad (9)$$

Where j and i designates band and frame respectively.$SNR_s(j,i)$and$SNR_{\hat{s}}(j,i)$ are the SNR's of original speech and enhanced speech in band j respectively. $S(j,i)$and$\hat{S}(j,i)$ are the spectrum of enhanced and original speech in j^{th} frequency band at i^{th} frame.

Clearly, when $S(j,i) = \hat{S}(j,i)$, SNRloss is zero. It means as the SNR level increases and tends to infinity, the estimated spectrum approaches the clean spectrum.

3.8 SNRLESC

It is the combination of two measures SNRloss and ESC (Excitation Spectra Correlation) defined as [4]:

$$SNRLESC(i) = \left(1 - r^2(i)\right).fSNRloss(i) \quad (10)$$

Here ESC measure at frame i is calculated first as follows:

$$r^2(i) = \frac{(\sum_{j=1}^{K} S(j,i).\hat{S}(j,i))^2}{\sum_{j=1}^{K} S^2(j,i).\sum_{j=1}^{K}\hat{S}^2(j,i)} \quad (11)$$

Where K is no. of bands and average ESC is computed by averaging $r^2(i)$ over all N frames.

$r^2(i)$issquared Pearson's correlation so limited to range in between [0,1]. A value of correlation coefficient close to 0 indicates that enhanced and original speech is uncorrelated and value close to 1 indicates opposite.

In case of $\hat{S}(j,i) = \alpha.S(j,i)$, when $\alpha > 1$ means $\hat{S}(j,i)$ is uniformly amplified and $\alpha < 1$ indicate uniform attenuation across all bands. Uniform distortions preserve the spectral shape which includes vowels consequently intelligibility remains preserved while in case of non-uniform distortions intelligibility suffers.

$fSNRloss(i)$is the average SNR loss across bands that is defined as:

$$fSNRloss(i) = \frac{\sum_{j=1}^{K} W(j).SNRloss(j,i)}{\sum_{j=1}^{K} W(j)} \quad (12)$$

Where $W(j)$ is the weight placed on j^{th} frequency band. The SNRLESC is limited to range in between [0,1] and this measure is obtained for three regions low, mid and high level segments of speech.

4. ALGORITHMS EVALUATED

A total of 6 algorithms are implemented i.e. Spectral subtraction [5], MMSE [8], Wiener filter [6], Kalman filter [9], Subspace method [10,11] andCompressed Sensing [13]. In implementation of Compressed Sensing (CS), Speech is reconstructed via Compressive Sampling Matching Pursuit [14] using no. of iteration (s=2). These algorithms are compared in terms of quality and intelligibility parameters in objective manner. In implementing these algorithms parameters used are same as in reference papers.

5. SIMULATION AND RESULTS

All algorithms are implemented in MATLAB. A common clean speech is taken from NOZEUS database [12], speech sample of male speaker of utterance 'A good book informs of what we ought to know' sampled at 25 kHz originally. The speech signal is resampled at 16 kHz and only a subset of first 19200 samples of speech is processed to reduce length of evaluations. Additive white Gaussian noise of SNR 10 dB is added to clean speech samples to make noisy speech signal. Speech signal is non-stationary in nature but assumed to be stationary in short segments. These short segments are called analysis frames and usually of 20-30msec. In our experiments, frame size is kept 20msec or length 320 samples and 50% overlapping is used. These frames are isolated, windowed and processed individually and at the end they are concatenated together to obtain the complete speech signal. The waveforms and spectrogram of reconstructed speech using Compressed Sensing is showed in Fig. 1 and Fig. 2 respectively. Objective parameters are calculated for these algorithms are listed in Table 4.

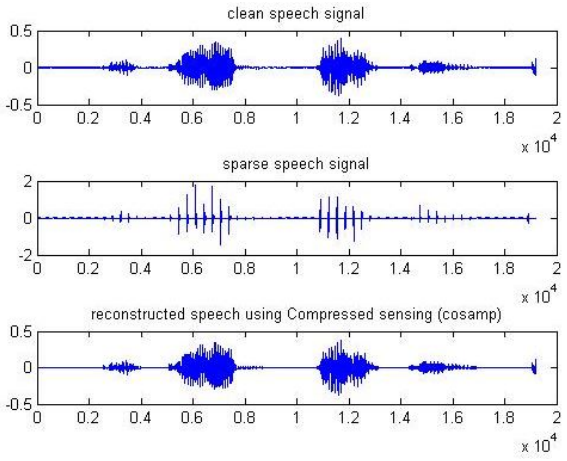

Fig. 1: Waveforms of clean speech, sparse representation in DCT and reconstructed speech using CoSaMP for s=2

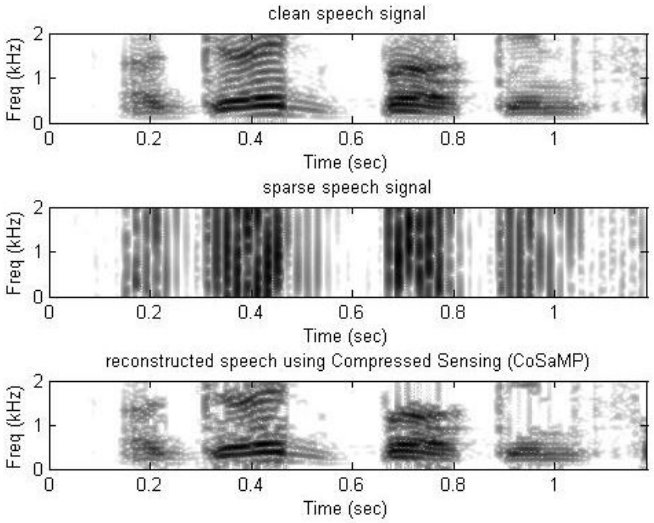

Fig. 2: Spectrogram of clean speech, sparse signal, reconstructed speech using CoSaMP for s=2

Table 4: Objective Evaluation of Speech Enhancement Algorithms for speech sample from NOIZEUS database

Techniques / Parameters	Spectral Subtraction	Wiener filter	Log-MMSE	Signal Subspace	Kalman filter	CSreconstructed speech
Overall SNR	14.5397	5.9424	3.6822	7.0163	0.4822	17.1773
SegSNR	1.5089	2.3854	-0.4291	-0.0251	5.3987	15.8903
LLR	5.5447	3.8226	5.3499	5.8384	1.5817	2.6549
WSS	50.3947	82.6106	51.2456	68.7264	1.6105	17.6584
PESQ MOS	2.0900	2.2130	1.7510	1.6220	2.9050	3.4590
MOS LQO	1.7060	1.8200	1.4580	1.3850	2.6830	3.4970
SIG	-1.8061	-0.2496	-1.8173	-2.5551	3.2026	2.2470
BAK	2.3751	2.2638	2.0854	1.9267	3.3515	4.1647
OVRL	0.0844	0.8399	-0.0941	-0.5705	3.1115	2.8750
SNRloss	0.7781	0.9956	0.9016	0.8825	0.7755	0.4477
SNRLESC_low	0.6631	0.6999	0.7422	0.7222	0.1956	0.0179
SNRLESC_mid	0.0510	0.0854	0.1409	0.0946	0.0526	0.0014
SNRLESC_high	0.0005	0.0015	0.0245	0.0178	0.1430	0.0012

6. CONCLUSIONS

Speech Enhancement is a tough task and still it is challenging after developing a number of algorithms. Tremendously growing speech processing applications require security of signals along with compressed versions of signals to save bandwidth. Compressed Sensing is a newly developed technique to acquire and reconstruct signal with fewer samples than traditional Nyquist sampling theorem. CS recovery algorithms are able to work in noisy environments in order to enhance speech. In this paper, 6 algorithms are evaluated; objective measures are computed for processed speech by these algorithms. A very low value of SNRloss and appreciable value of PESQ MOS is found to be in reconstructed speech using CS. All measures are found to be in acceptable range and comparable to other speech enhancement algorithms. Future direction is concentrated on developing Speech Enhancement algorithms based on CS, designing efficient measurement matrices and suitable sparse bases for representation of speech.

REFERENCES

[1] Yi Hu and P.C. Loizou, "Evaluation of Objective Quality Measures for Speech Enhancement", *IEEE Trans. On Audio, Speech and Language Proc.*, vol. 16, no. 1, Jan. 2008

[2] Yi Hu and P.C. Loizou, "Subjective Comparison of Speech Enhancement Algorithms", *in Proc. IEEE ICASSP*, 2006

[3] T.S. Gunawan, E.Ambikairajah, "Subjective Evaluation of Speech Enhancement Algorithms using ITU-T P.835 Standard", *in proc. Of IEEE*, 2006

[4] Ma.J. and P.C. Loizou, "SNR loss: A new objective measure for predicting intelligibility of noise- suppressed speech", *speech communication*, 53, 340-354, 2011

[5] Steven F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. On Acoust., Speech and Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr.1979

[6] C. Plapous, C.Marro, P. Scalart, "Improved Signal-to-NoiseRatio Estimation for Speech Enhancement", *IEEE Transc. On Audio, Speech and Language Proc.*, vol. 14, no. 6, Nov., 2006

[7] Ephraim Y., Malah D., "Speech Enhancement using a Mimumum Mean-Square Error Log-Spectral Amplitude Estimator", *IEEE Trans. On Acoustics, speech and signal proc.*, 33(2), 443-445, Apr, 1985

[8] Kalman Filter, *SIGGRAPH, Los Angeles, CA,* August 12-17, 2001

[9] . Ephraim, H.L. Van Trees, "A Signal Subspace Approach for Speech Enhancement", *IEEE Trans. On Speech and Audio processing,*Vol.3, no. 4, ; 251-266, July 1995

[10] W.G. Yan, G.Y.Xiang and Z.X. Qun, "A signal Subspace Speech Enhancement method for Various Noises", *TELKOMNIKA*, vol.11, no. 2, pp. 726-735, Feb, 2013

[11] *NOIZEUS*: a noisy speech copus for evaluation of speech enhancement algorithms

[12] F.F.Firouzeh, S. Ghorshi, S. Salsabili, "Compressed Sensing based Speech Enhancement" *IEEE Conference,* 2014

[13] D. Needell and J.A. Troop, "CoSaMP: iterative signal recovery from incomplete and accurate samples", *Applied and Computational Harmonic Analysis*, vol. 26, pp. 301-321, 2009

[14] ITU-T P.835, Subjective test methodology for evaluating speech communication systems that include noise suppression algorithms, *ITU-T Recommendation P.835*, 2003.