SPEECH ENHANCEMENT USING SGJMAP ALGORITHM FOR MOBILE APPLICIATION

V.Shruthi
PG scholar, Department of Electronics and Communication
Easwari Engineering College
Chennai, India

R.Senthamizhselvi
Associate professor, Department of Electronics and Communication
Easwari Engineering College
Chennai, India

G.R.Suresh
Professor, Department of Electronics and Communication
Rajalakshmi Institute of Technology
Chennai, India

Abstract— **The aim of the speech enhancement is to improve the intelligibility and quality of the speech. By recording the speech signal in the noisy environment, clean speech signals are degraded. Speech enhancement reduces the noise without distorting the original (clean) signal. The proposed method super Gaussian Joint Maximum a Posterior (SGJMAP) is based on gain function. In this concept, the amount of noise suppression and speech distortion is controlled in real-time based on the level of hearing comfort perceived in noisy real world acoustic environment. Objective quality and intelligibility measurement show the effectiveness of the proposed method in comparison to benchmark techniques considered in this paper.**

Keywords- Super Gaussian joint Maximum a Posteriori(SGJMAP).

I.INTRODUCTION

Across the world, 360 million people suffer from hearing loss. Statistics reported by National Institute on Deafness and other Communication Disorders (NIDCD) show that in United States, 15% of American adults (37million) aged 18 and over report some kind of hearing loss. Researchers in academia and industry are developing viable solutions for hearing impaired in the form of Hearing Aids (HA) and Cochlear Implants (CI).

Speech Enhancement (SE) is a key component in the HA pipeline. Existing HA devices do not carry the computational power to handle complex but indispensable signal processing algorithms [1-3]. Recently, HA manufacturers are using an external microphone in the form of a pen or a necklace to capture speech with higher Signal to Noise Ratio (SNR) and wireless transmit to HA [4]. The problem with these existing auxiliary devices is that they are too expensive and are not portable. One strong contending auxiliary device is our personal smart phone that has the capability to capture the noisy speech data using its microphone, perform complex computations and wirelessly transmit the data to the HA device. Recently, extensively used smart phones such as

Apple iPhone and other Android smart phones, are coming up with new HA features such as Live Listen by Apple [5], and many 3rd party HA applications to enhance the overall quality and intelligibility of the speech perceived by hearing impaired. Most of these HA applications on the smart phone use single microphone, to avoid audio Input/output latencies

The most challenging task in a single microphone SE is to suppress the background noise without distorting the clean speech. Traditional methods like Spectral Subtraction [6] introduce musical noise due to half-wave rectification problem [7], which is prominent at lower SNRs. This problem is solved by estimating the clean speech magnitude spectrum by minimizing a statistical error criterion, proposed by Ephraim and Malah [8, 9]. In [10], a computationally efficient alternative is proposed for SE methods in [8, 9]. In this new method, speech is estimated by applying the joint maximum a posteriori (JMAP) estimation rule. In [13], super-Gaussian extension of the JMAP (SGJMAP) is proposed which is shown to outperform algorithms proposed in [8-10]. Super-Gaussian statistical model of the clean speech and noise spectral components (especially Babble) attains a lower mean squared error compared to Gaussian model. The challenge with existing single microphone SE techniques for HA applications is that the amount of noise suppression cannot be controlled in real-time. Therefore, the amount of speech distortion cannot be restrained below tolerable level. Recent developments include SE based on deep neural networks (DNN) [11, 12], which requires rigorous training data. Although these methods yield supreme noise suppression, the preservation of Spectra-temporal characteristics of speech, the quality and natural attributes remains as a prime challenge. Hence, these methods are not suitable for HA applications, where the hearing impaired prefers to hear speech that sounds natural, like a normal hearing.

In this paper, we present a parameter called "compensation" factor in the optimization of the SGJMAP cost function to estimate the magnitude spectrum of the clean speech. The proposed gain is a function of the compensation parameter that is designed to vary in real time, which allows the smart phone user to control the amount of noise suppression and distortion of the voice. The developed method is computationally economic and requires no training. The change in the exchange parameter influences the performance of the SE in conditions of reverberation and changing noise conditions. The objective and subjective evaluations of the proposed method are carried out to evaluate the effectiveness of the method with respect to the reference techniques considered and to discuss the general usability of the developed algorithm.

II. METHODOLOGY USED

A. SGJMAP Based Speech Enhancement

In the SGJMAP method, a super Gaussian speech model is used by considering non-Gaussianity property in spectral domain noise reduction framework [14, 15] and by knowing that speech spectral coefficients have a super-Gaussian distribution. Spectral amplitude estimator using super Gaussian speech model allows the probability density function (PDF) of the speech spectral amplitude to be approximated by the function of two parameters μ and v. These two parameters can be adjusted to fit the underlying PDF to the real distribution of the speech magnitude. Considering the additive mixture model for noisy speech y(n), with clean speech s(n) and noise w(n).

$$\mathbf{Y(n)= S(n) + W(n)} \quad \textbf{(1)}$$

The noisy kth Discrete Fourier Transform (DFT) coefficient of (n) for frame λ is given by,

$$\mathbf{Y_k(\lambda) = S_k(\lambda) + W_k(\lambda)} \quad \textbf{(2)}$$

where S and W are the clean speech and noise DFT coefficients respectively. In polar coordinates, (2) can be written as,

$$\mathbf{R_K(\lambda)}e^{j\theta}\mathbf{Y_k(\lambda) = A_k(\lambda)}e^{j\theta}\mathbf{S_k(\lambda) + B_k(\lambda)}e^{j\theta}\mathbf{W_K(\lambda)} \quad \textbf{(3)}$$

Where $Rk(\lambda)$, $Ak(\lambda)$, $Bk(\lambda)$ are magnitude spectrums of noisy speech, clean speech and noise respectively. (λ), $\theta Sk(\lambda)$, $\theta Wk(\lambda)$ are the phase spectrums of noisy speech, clean speech and noise respectively. The goal of any SE technique is to estimate clean speech magnitude spectrum (λ) and its phase spectrum $\theta S(\lambda)$.

Figure 1 shows the block diagram of the proposed method. In (8), the gain of SGJMAP is a function of four parameters (v,$\hat{\xi}k$,$\hat{\gamma}k$). The accuracy of $\hat{\xi}k$,k depends on the VAD and the SE gain function of the previous frames. The values of v and μ can be set empirically to achieve good noise reduction without distorting the speech, as discussed in [16]. However, the optimal values of these parameters in real world rapidly fluctuate with changing acoustical and environmental conditions, owing to the fact that the gain is designed by assuming super-Gaussian PDF for speech only in ideal acoustic conditions. In the presence of reverberation and noise (especially babble), the real PDF of speech received at the microphone changes. Therefore, having fixed μ and v is not feasible to give robust noise reduction in dynamic conditions.

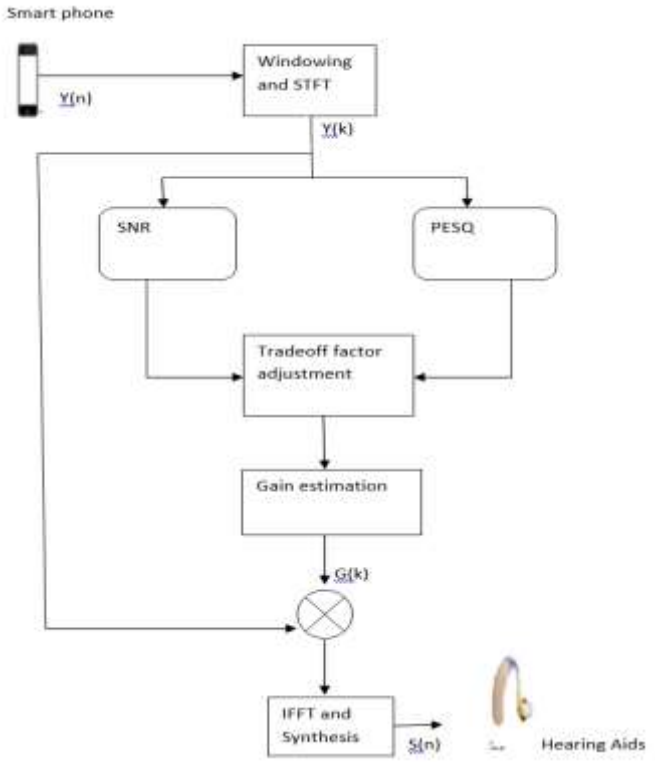


Figure1: Block diagram of SGJMAP

In order to compensate for these inaccuracies in the model, we introduce a "trade-off" parameter β into the cost function optimization for optimal clean speech magnitude estimation. Taking natural logarithm of (4), and differentiating with respect to Ak gives,

$$\mathbf{G_K}=\left[\left(\frac{1}{2\beta}\right)-\left(\frac{\mu}{\sqrt[4]{\gamma\hat{}k\xi\hat{}k}}\right)+\sqrt{\left(\frac{\mu}{\sqrt[4]{\gamma\hat{}k\xi\hat{}k}}-\frac{1}{2\beta}\right)+\frac{v}{2\gamma\hat{}\beta 2}}\right] \quad \textbf{(4)}$$

The final clean speech spectrum estimate is

$$\mathbf{\hat{S}_k = G_k Y_k} \quad \textbf{(5)}$$

The time domain sequence $\hat{s}(n)$ is obtained by taking Inverse Fast Fourier Transform (IFFT) of $\hat{S}k$. At very low values of β and v, the gain Gk becomes less dependent on $\hat{\xi}k$, which minimizes speech distortion while compromising on noise suppression. This makes the algorithm robust to inaccuracies in the estimation of $\hat{\xi}k$. In most of the statistical model based SE algorithms, the accuracy of clean speech magnitude spectrum directly depends on how accurately $\hat{\xi}k$ is estimated. However, inaccurate $\hat{\xi}k$ results in distortion of speech and introduces musical noise in the background. The proposed method circumvents this problem by allowing the user to select lower β. At higher values of β, the overall gain Gk decreases yielding good noise suppression, but ends up attenuating speech as well. Although, higher values of β is not useful when there is speech of interest, but it is useful in conditions when the user is exposed to loud noisy environment with no speech of interest. At $\beta\approx1$, the proposed method reduces to SGJMAP. Setting appropriate intermediate values for β yields noise suppression with considerable speech distortion.

III. RESULTS AND DISSCUSSION

A. Noise database

NOIZEUS is a noisy speech corpus recorded at the Center for Robust Speech Systems, Department of Electrical Engineering, University of Texas and Dallas. The noisy database contains 30 IEEE sentences produced by three male and three female speakers (five sentences /speaker), and was corrupted by eight different real-world noises at different SNRs. Thirty sentences from the IEEE sentence database were recorded in a sound proof booth using Tucker Davis Technologies (TDT) recording equipment.

The sentences were originally sampled at 25 KHz and down sampled to 8 KHz. To simulate the receiving frequency characteristics of telephone handsets, the speech and noise signals were filtered by the modified Intermediate Reference System (IRS) filters used in ITU-T P.862 for evaluation of the PESQ measure. Noise signals were taken from the AURORA database and included the following recordings from different places: babble (crowd of people), car, exhibition hall, restaurant, street, airport, train, station, and train.

B .Experimentation

There are no algorithms that are developed to our knowledge that provide similar functionality of achieving the balance between noise suppression and speech distortion in real time without any pre or post filtering. We therefore fix the values of few parameters and evaluate the performance of the proposed method by comparing with JMAP [10] and SGJMAP [13] method, as our two-benchmark single microphone SE techniques that have shown promising results. Also, the developed method is an improved extension of these two methods. The experimental evaluations are performed for 3 different noise types: CAR, babble and train noise.

For objective evaluation, all the files are sampled at 16 kHz and 20 ms frames with 50% overlap are considered. As objective evaluation criteria, we choose the perceptual evaluation of speech quality (PESQ) for speech quality measurement and short time objective intelligibility (STOI) to measure speech intelligibility. PESQ ranges between 0.5 and 4.5, with 4.5 being high perceptual quality. Higher the score of STOI better is the speech intelligibility. Figure 3 shows the plots of PESQ and STOI versus SNR for the 3 noise types. The best values of μ and v were empirically determined over large dataset as they largely control the statistical properties of the noisy signal.

Although objective measures give useful evaluation results during the development phase of our method, they give very little information about the usability of our application by the end user. We performed Mean Opinion Score (MOS) tests on 15 expert normal hearing subjects who were presented with noisy speech and enhanced speech using the proposed, JMAP and SGJMAP methods at SNR levels of -5 dB, 0 dB and 5 dB. The key contribution of this paper is in providing the user the ability to customize the parameters for their listening preference. Before starting the actual tests, the subjects were instructed to set β, μ and v for each noise type as per their preference. One key observation was, the preferred values of β, μ and v varied across subjects. This supports our claim that the developed application is user customizable. Therefore, for each audio file the subjects were instructed to score in the range 1 to 5 with 5 being excellent speech quality and 1 being bad speech quality. The detailed description of scoring procedure is in [20]. Subjective test results in Figure 3 illustrate the effectiveness of the proposed method in reducing the background musical noise, simultaneously preserving the quality and intelligibility of the speech. We also conducted a field test of our application in real world noisy conditions, which change dynamically. Varying the β, μ and v in real-time provides tremendous flexibility for the end user to control the perceived speech.

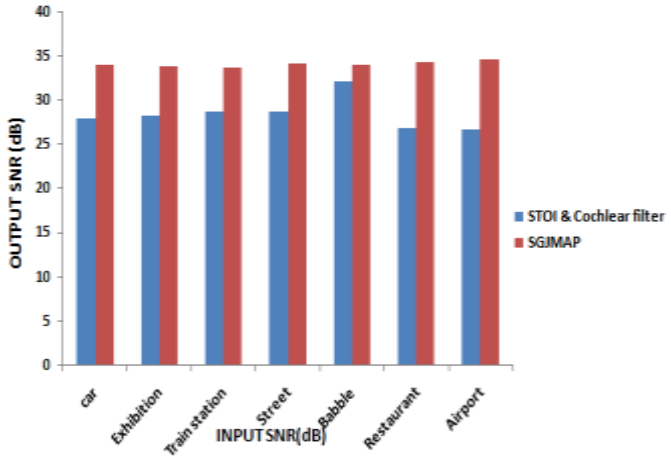
C. SNR estimation

The signal-to-noise ratio (SNR) is one of the oldest and most used objective measures. It is mathematically simple to calculate, but requires distorted and non-distorted (clean) speech samples. Where, x (n) is a clean speech, x (n) a distorted speech and N the number of samples. This classic definition of SNR is not well correlated with the quality of speech for a wide range of distortions. Therefore, there are several variations of classic SNR that show a much higher correlation with subjective quality. It has been observed that classic SNR is not well correlated with voice quality because although the voice is not a stationary signal, SNR averages the relationship in the whole signal. The energy of the speech fluctuates over time, so the parts where the speech energy is large and the relatively inaudible noise should not be washed from other parts where the speech energy is small and the noise can be heard from the speech . Therefore, the SNR was calculated in short squares and then calculated as an average. This measure is called segmental SNR and can be defined as Where L is the frame length (number of samples), and M the number of frames in the signal $(N = ML)$. The frame length is normally set between 15 and 20ms. Since, the logarithm of the ratio is calculated before averaging, the frames with an exceptionally large ratio is somewhat weighed less, while frames with low ratio is weighed somewhat higher. It can be observed that this matches the perceptual quality well, i.e., frames with large speech and no audible noise does not dominate the overall perceptual quality, but the existence of noisy frames stands out and will drive the over all quality lower.

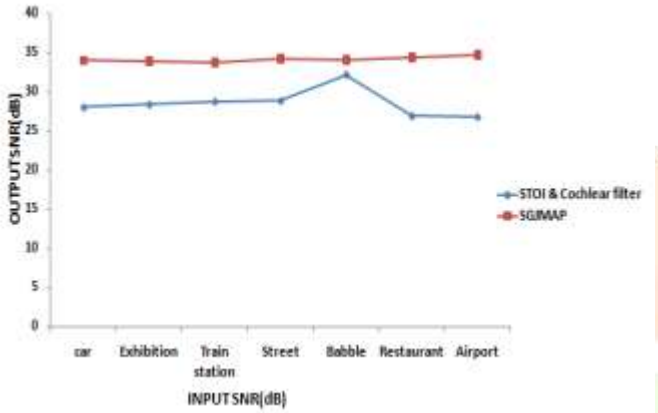
However, if the speech sample contains excessive silence, the overall *SNR*seg values will decrease significantly since silent frames generally show large negative *SNR*seg values. In this case, silent portions should be excluded from the averaging using speech activity detectors. In the same manner, exclusion of frames with excessively large or small values from averaging generally results in *SNR*seg values that agree well with the subjective quality. A typical value for the upper and the lower ratio limit is 35 and −10 dB. These ranges are also used for *SNR*seg calculation throughout this book. Another variation to the SNR is the frequency-weighed SNR (fwSNRseg). This is essentially a weighted SNRseg within a frequency band proportional to the critical band. The fwSNRseg can be defined as follows

where $W(j,m)$ is the weight on the j^{th} sub band in the m^{th} frame, K is the number of sub bands, $X(j,m)$ is the spectrum magnitude of the j^{th} sub band in the m^{th} frame, and $\hat{}X(j,m)$ its distorted spectrum magnitude.

SNR COMPARISION



2(a)



2(b)

Figure(2a) &2(b):Output SNR comparison for SGJMAP STOI &Cochlear filter

From the fig:2(a) &2(b) shows the SNR comparison in dB for 0dB various noise .It can be seen from the figure that compared to STOI and cochlear filter methods ,SGJMAP showed improved performance for various noise types and at various input SNR levels .

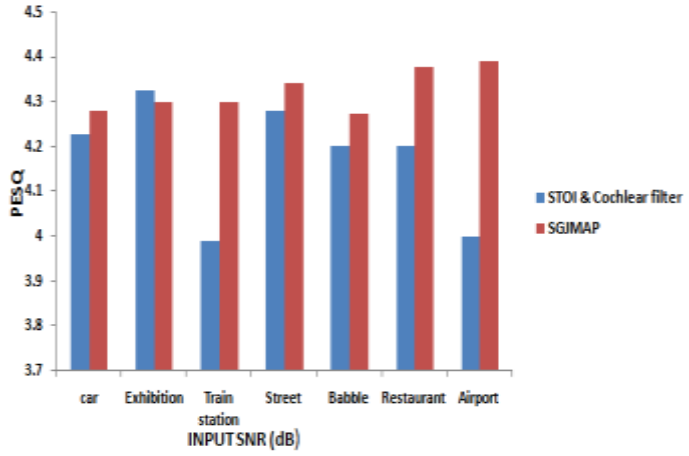
Table 1
Output signal to noise ratio result at different input SNR levels

NOISE	METHOD	SNR (0 dB)	SNR (5 dB)	SNR (10dB)	SNR (15dB)
Car	STOI & Cochlear filter	24.2915	25.6691	27.9897	28.0092
	SGJMAP	29.8037	30.3684	31.9239	34.0078
Exhibition	STOI & Cochlear filter	23.5172	25.9813	27.2866	28.3547
	SGJMAP	29.8585	30.4088	32.1703	33.9224
Train station	STOI & Cochlear filter	25.2273	27.0873	27.8157	28.7331
	SGJMAP	29.7602	30.3898	31.8978	33.7542
Street	STOI & Cochlear filter	23.3205	25.6672	26.7805	28.8370
	SGJMAP	29.7587	30.5153	32.2914	34.2872
Babble	STOI & Cochlear filter	27.8752	30.0132	31.7752	32.1848
	SGJMAP	29.9093	30.6454	31.7050	34.0410
Restaurant	STOI & Cochlear filter	24.1617	26.0033	26.3161	26.8870
	SGJMAP	31.2588	31.7622	32.8272	34.3626
Airport	STOI & Cochlear filter	23.9806	25.6491	26.4531	26.7620
	SGJMAP	**31.2437**	**31.8165**	**32.8501**	**34.7086**

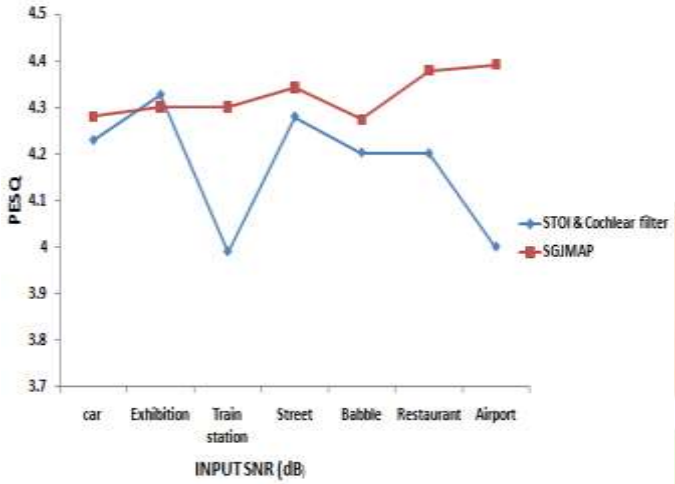
D. PESQ estimation

Perceptual assessment of speech quality (PESQ) is an international standard for estimating the average opinion score (MOS) of both the clean signal and its degraded signal. It has been developed from a number of previous MOS estimation attempts and is considered one of the most sophisticated and accurate estimation methods available today. PESQ has been officiallystandardized by the International Telecommunication Union Telecommunication Standardization Sector (ITU-T) as standard P.862 in February 2001PESQ uses a perceptive model to hide the degraded input and speech in an internal representation. The degraded entry is aligned over time with the original signal to compensate for the delay that may be associated with degradation. The difference in the internal representations of the two signals is used by the cognitive model to estimate the MOS. The PESQ values obtained using the cochlear filter and the STOI method and the same methods used separately are compared and the values are tabulated. PESQ scores were expressed using the mean auditory quality objective score scale (MOS LQO) and range from 1 (worst quality) to 5 (best quality)

.PESQ COMPARISON



3(a)



3(b)

Figure3(a)3(b):Output PESQ Comparison for SGJMAP, STOI &Cochlear filter

Table 2
Result of objective measure(PESQ) with different input SNR(0dB,10dB,15dB)

NOISE	METHOD	PESQ with (0dB)	PESQ with (5dB)	PESQ with (10dB)	PESQ with (15dB)
Car	STOI & Cochlear filter	3.8982	3.9403	4.2004	4.2302
	SGJMAP	4.2440	4.2404	4.2708	4.2799
Exhibition	STOI & Cochlear filter	3.5962	3.9366	4.0116	4.3276
	SGJMAP	4.2101	4.2490	4.1638	4.3004
Train station	STOI & Cochlear filter	3.8572	3.9275	3.9523	3.9906
	SGJMAP	4.2049	4.2570	4.2674	4.3007
Street	STOI & Cochlear filter	3.6684	3.7244	3.7860	4.2807
	SGJMAP	4.3672	4.2861	4.3077	4.3432
Babble	STOI & Cochlear filter	3.8968	3.9450	4.0855	4.2033
	SGJMAP	4.1878	4.2005	4.2049	4.2743
Restaurant	STOI & Cochlear filter	3.6321	3.9130	3.9964	4.2017
	SGJMAP	4.2047	4.2084	4.3082	4.3803
Airport	STOI & Cochlear filter	3.8038	3.8234	3.9176	4.0010
	SGJMAP	**4.1990**	**4.2263**	**4.2946**	**4.3921**

IV. CONCLUSION

The proposed speech enhancement SGJMAP method has been introduced to overcome the tradeoff between amount of noise suppression and speech distortion. The proposed algorithm used in smart phone device, which works as an assistive device for HA. Varying the tradeoff enables the smart phone user to control the amount of noise suppression and speech distortion. The SGJMAP algorithm shows the better results in terms of evaluation parameters such as SNR (dB) and PESQ (Out of 4.5) at various noise levels than the existing algorithms. The maximum SNR obtained as 34.7086 dB and maximum PESQ of 4.3921 have been achieved by the proposed method.

REFERENCES

[1] Y-T. Kuo, T-J. Lin, W-H Chang, Y-T Li, C-W Liu and S-T Young, "Complexity-effective auditory compensation for digital hearing aids," *IEEE Int. Symp on Circuits ad Systems (ISCAS)*, May 2008.

[2] T. J. Klasen, T. V Bogaert den, M. Moonen, J. Wouters, "Binaural Noise Reduction algorithms for hearing aids that preserve interaural time delay cues," *IEEE Trans. Signal Process*, vol. 55, pp. 1579-1585, April 2007.

[3] C. K. A. Reddy, Y. Hao, I. Panahi, "Two microphones spectral-coherence based speech enhancement for hearing aids using smartphone as an assistive device," *IEEE Int. Conf. on Eng. In Medicne and Biology soc.*, Oct 2016.

[4] B. Edwards, "The future of Hearing Aid technology," *Journal List, Trends Amplif*, v.11(1): 31-45, Mar 2007.

[5] https://support.apple.com/en-us/HT203990

[6] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoustic, Speech and Signal Process*, vol. 27, pp. 113-120, Apr 1979.

[7] M. Berouti, M. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," *Proc of IEEE Conf. on Acoustic SpeechSignal Processing*, pp. 208-211, Washington D.C, 1979.

[8] Y. Ephraim and D.Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[9] Y. Ephraim and D.Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.
[10] P. J. Wolfe and S. J. Godsill, "Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 10, pp. 1043–1051, 2003, special issue: Digital Audio for Multimedia CommunicationsT.
[11] Y. Xu, J. Du, L-R. Dai, C-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Proc. Letters*, pp. 65-68, Nov 2013.
[12] F. Weninger, J. R. Hershey, J. L. Roux,B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," *IEEE Global Conf. on Signal and Inf Processing*, Dec 2014.
[13] Lotter, P. Vary, "Speech Enhancement by MAP Spectral Amplitude Estimation using a super-gaussian speech model," *EURASIP Journal on Applied Sig. Process*, pp. 1110-1126, 2005.
[14] R. Martin, "Speech enhancement using MMSE short time spectral estimation with gamma distributed speech priors," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '02)*, vol. 1, pp. 253–256, Orlando, Fla, USA, May 2002.
[15] R. Martin and C. Breithaupt, "Speech enhancement in the DFT domain using Laplacian speech priors," in *Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC'03)*, pp. 87–90, Kyoto, Japan, September 2003.

[16] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*., vol. 6, no. 1, pp. 1–3, 1999.
[17] P. Vary, "Noise suppression by spectral magnitude estimation— mechanisms and theoretical limits," *Signal Processing*, vol. 8, no. 4, pp. 387–400, 1985.
[18] A. W. Rix, J. G. Beerends, M. P Hollier, A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ) – a new method for speech quality assessment of telephone networks and codecs," *IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, 2, pp. 749-752., May 2001.
[19] C. H Taal, R. C. Hendricks, R. Heusdens, R. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE trans. Audio, Speech, Lang. Process*. 19(7), pp. 2125-2136., Feb 2011.
[20] ITU-T Rec. P.830, "Subjective performance assessment of telephone- band and wideband digital codecs," 1996.