Acoustic Echo Suppression Using Speech Uncertainty In Modulation Domain

Muhammed PV Shifas, E.P. Jayakumar, P.S. Sathidevi .

Abstract—Quality degradation of near-end speech in mobile communication or hands free devices is mainly due to acoustic echoes and background noises. The received far-end speech gets reflected from the obstacles present in the surroundings creating acoustic echo. All other disturbances from the near-end environment are considered as background noises. A novel acoustic echo suppression scheme using speech uncertainty in modulation domain (MD) is proposed in this paper. State of the art acoustic echo suppression systems are based on either time domain or frequency domain analysis. In recent times, the modulation domain analysis is popularly used in speech processing, as it captures the human perceptual properties. Modulation domain provides the temporal variation of the acoustic magnitude spectra which acts as an information bearing signal. In this paper, a new method is developed and implemented to model the echo path and estimate the echo in modulation domain. Echo cancellation is done effectively by manipulating the modulation spectrum and employing speech uncertainty. In this method, the microphone input is modelled as a binary hypothesis process and the gain function is modified accordingly. The proposed method shows better performance as compared to other competitive methods for acoustic echo suppression with no audible degradation in the near-end speech.

Index Terms—Speech enhancement, acoustic echo suppression, modulation domain analysis, speech uncertainty.

I. INTRODUCTION

THE quality of the speech in mobile or hands free devices gets deteriorated because of the effects of acoustic echoes and background noises. The reflection of incoming far-end signal from the near-end surroundings will disturb the input to the microphone input. This reflected signal is known as acoustic echo. Since it is due to the reflection from surroundings, the intensity of acoustic echo would obviously depend on the operating surroundings of the devices. All other disturbances are considered as background noises. Some of the commonly used methods for noise elimination are spectral subtraction and Minimum Mean Square Error (MMSE) estimator [1], [2].

Acoustic echo can be imagined as a high intensity babble noise generated by the same speaker, and hence it is very difficult to tackle it through conventional speech enhancement techniques. Since the acoustic echo produced by the reflections of incoming signal from the surroundings and the incoming signal from the far-end side is accessible from the loud speaker, we can create a copy of echo components by

estimating the echo path impulse response effectively. Hence, in almost all acoustic echo analysis techniques, the basic challenge is to effectively estimate the acoustic echo path response and then determine the echo components.

Recently, Faller et.al [3] had proposed a new frequency domain modelling of echo path response. It is modelled as a system in frequency domain which introduces some delay and spectral modification to the incoming far-end speech. The delay parameter introduced from the echo path is estimated by observing the peak of cross correlation spectrum between the microphone and loud speaker signals. They have shown that the echo path gain could be calculated through minimizing the mean square error [4]. This frequency domain analysis of echo effects gave better performance than the normal time domain counterpart. But, as in any frequency domain analysis, this frequency domain echo cancellation setup also introduces some random spikes in the final echo suppressed signal, which we call as musical noise effects.

To mitigate the musical noise, we have recently proposed a new modulation domain analysis for acoustic echo suppression [5], [6]. This is through modelling the echo path effects in modulation domain rather than in time domain or frequency domain. Experimental results show that this modulation domain technique is better than the other existing echo suppression systems. In this work, we have modified the modulation domain gain filter for echo suppression by incorporating the near-end speech uncertainty. It is done by modelling the microphone input as a binary hypothesis process considering presence and absence of the relevant near-end speech signal. The gain filter should give maximum suppression (zero gain) to the microphone signal during the echo only situation and the suppression should be minimum during the occurrence of relevant speech. The detailed theory is given in Section. IV.

In order to make a comparison of the performance of our method with that of the state of art technique, we have implemented the system proposed by Park.et.al [7] also and then tested experimentally. The performance comparison is done using subjective as well as objective measures.

The proposed modulation domain technique for acoustic echo suppression is detailed in Section. II and the modified system employing the speech uncertainty is discussed in Section. IV. In Section. V, performance analysis of the modified system and comparison with the existing methods are included. Paper is concluded in Section. VI

Muhammed PV Shifas is with the Speech Signal Processing Laboratory, Department of Computer Science, University of Crete, Greece (e-mail: shifaspv@csd.uoc.gr).

E.P. Jayakumar and P.S. Sathidevi are with the Department of Electronics and Communication Engineering, National Institute of Technology Calicut, Kozhikode, Kerala, India

II. ACOUSTIC ECHO SUPPRESSION IN MODULATION DOMAIN

As stated in literature [3], the practical echo path response has only few high amplitude early reflection coefficients and low amplitude tails extending to infinity. This high intensity early reflection coefficients bring in some delay in the received far-end signal and the effects of tail can be captured with under-estimation of echo. With this modelling, the acoustic echo path can be treated as a system which introduces some delay and spectra modifications to the received far-end signal. Hence, the problem of echo path estimation has now become estimation of these two parameters.

Here, instead of modelling the echo path effects in frequency domain, we have moved to modulation domain since it captures the temporal variation of the acoustic magnitude spectrum, where the linguistic information is gathered. Therefore, echo path is modelled as a system that introduces some delay and spectral modification to the modulation spectra of the incoming far-end signal. The delay introduced by the echo path is estimated using the cross correlation based delay estimation algorithm, where the peak of the correlation spectrum of the microphone and loud speaker signal gives the sample delay introduced by the acoustic echo path. Since echo path response changes continuously due to the changes in the surroundings, the estimation of those parameters needs to be done continuously.

Mathematical modelling of echo path effects in modulation domain is formulated as follows. As we know, the microphone input will be having two signal components namely acoustic echo and the relevant near-end speech. Let us take $y(n)$ as the microphone signal, which has the echo component $d(n)$ and the near-end speech component $s(n)$. As we know that the echo is created because of the reflection of incoming far-end signal $x(n)$ on the obstacles in the echo path $g(n)$, the echo components can be modelled as a convolution expression $d(n) = g(n) * x(n)$. Hence, the mathematical expression for microphone input will be

$$y(n) = g(n) * x(n) + s(n) \quad (1)$$

Now, as we discussed initially, the echo path can be modelled as a system which introduces some delay and spectral modification to the received far-end signal. These parameters of echo path are estimated independently. Let $x_d(n)$ is the delayed version of incoming far-end signal $x(n)$ after estimating the delay parameter introduced by the echo path and $G_c(k, f_{ind}, l)$ be the spectral modification introduced by the echo path in the modulation domain. Then the above expression for microphone input can be re-written in modulation domain as

$$Y(k, f_{ind}, l) = G_c(k, f_{ind}, l)X_d(k, f_{ind}, l) + S(k, f_{ind}, l) \quad (2)$$

where, $G_c(k, f_{ind}, l)$, $Y(k, f_{ind}, l)$, $X_d(k, f_{ind}, l)$ and $S(k, f_{ind}, l)$ are the modulation spectra of the respective signals. The symbols k indicates the acoustic frequency index, f_{ind} indicates the frame index and l indicates the modulation frequency index.

The estimation of these parameters is carried out independently. To estimate the sample delay caused by the echo path, initially the cross-correlation between the reflected wave (microphone input) and the original signal (received far-end signal) is calculated. The sample in which the peak of the correlation spectra occurs will be the delay introduced from the acoustic echo path. To estimate the second echo path parameter, which is the modulation domain spectral modification, the following mathematical analysis is carried out. If we can efficiently estimate the impulse response of the echo path, then the estimation of echo disturbance spectra will be a simple task. After this, we can effectively suppress these echo disturbances through the spectral modification as in the conventional noise suppression algorithms. To get the estimate of spectral modification, multiply Eq.2 with $X_d^*(k, f_{ind}, l)$. Taking expectation on both sides and as the expectation is a linear operator, the expression becomes

$$\begin{aligned} E\{Y(k, f_{ind}, l)X_d^*(k, f_{ind}, l)\} = \\ E\{G_c(k, f_{ind}, l)X_d(k, f_{ind}, l)X_d^*(k, f_{ind}, l)\} + \\ E\{S(k, f_{ind}, l)X_d^*(k, f_{ind}, l)\} \end{aligned} \quad (3)$$

Since the near-end and far-end speakers are different, the corresponding speech signals $S(k, f_{ind}, l)$ and $X_d(k, f_{ind}, l)$ are linearly independent. Then, the expectation of product becomes product of individual expectations.

$$\begin{aligned} E\{Y(k, f_{ind}, l)X_d^*(k, f_{ind}, l)\} = \\ E\{G_c(k, f_{ind}, l)X_d(k, f_{ind}, l)X_d^*(k, f_{ind}, l)\} + \\ E\{S(k, f_{ind}, l)\}E\{X_d^*(k, f_{ind}, l)\} \end{aligned} \quad (4)$$

In general, the speech signal is assumed to be a Gaussian process with zero mean and finite variance, i.e., $E\{S(k, f_{ind}, l)\} = 0$. Hence, the second term on the right hand side of the Eq.4 becomes zero and the expression becomes

$$\begin{aligned} E\{Y(k, f_{ind}, l)X_d^*(k, f_{ind}, l)\} = \\ E\{G_c(k, f_{ind}, l)X_d(k, f_{ind}, l)X_d^*(k, f_{ind}, l)\} \end{aligned} \quad (5)$$

Since the spectral gain of echo path $G_c(k, f_{ind}, l)$ is a deterministic unknown, we can take it out from the expectation as

$$\begin{aligned} E\{Y(k, f_{ind}, l)X_d^*(k, f_{ind}, l)\} = \\ G_c(k, f_{ind}, l)E\{X_d(k, f_{ind}, l)X_d^*(k, f_{ind}, l)\} \end{aligned} \quad (6)$$

Then, the final estimate of modulation domain coloration effect filter $G_c(k, f_{ind}, l)$ will be a least square estimator as given in Eq.7

$$G_c(k, f_{ind}, l) = \frac{E\{Y(k, f_{ind}, l)X_d^*(k, f_{ind}, l)\}}{E\{X_d(k, f_{ind}, l)X_d^*(k, f_{ind}, l)\}} \quad (7)$$

In practice, the response of echo path changes continuously. Hence the estimate of $G_c(k, f_{ind}, l)$ is done recursively using the first order recursive relations with recursion parameter α ,

$$\begin{aligned} E\{X_d^*(k, f_{ind}, l)Y(k, f_{ind}, l)\} = \\ \alpha E\{X_d^*(k, f_{ind} - 1, l)Y(k, f_{ind} - 1, l)\} + \\ (1 - \alpha)X_d^*(k, f_{ind}, l)Y(k, f_{ind}, l) \end{aligned} \quad (8)$$

$$E\{X_d^*(k, f_{ind}, l)X_d(k, f_{ind}, l)\} = \alpha E\{X_d^*(k, f_{ind} - 1, l)X_d(k, f_{ind} - 1, l)\} + (1 - \alpha)|X_d^*(k, f_{ind}, l)X_d(k, f_{ind}, l)| \quad (9)$$

The signal retrieved from the microphone at any time instant can be categorized into two sets, namely single talk and double talk situations. The single talk happens when the microphone signal has the echo components only and in the double talk case, the microphone input will have the relevant speech signal along with the echo components.

The above mentioned least square estimator for echo path tracking will diverge from the actual estimate in the double talk situation. This will lead to the quality degradation of near-end speech transmission. To overcome this problem, we have employed a Double Talk Detector (DTD) in modulation domain. It will identify the double talk situation and pause the adaptation for a while.

III. MODULATION DOMAIN DOUBLE TALK DETECTOR

The main challenge in any acoustic echo cancellation setup is the estimation of echo path response. As discussed in Eq.7, the echo path response is estimated using the correlation measurements between the incoming signal and far-end signal which causes the formation of the echo signal. Accuracy of echo path estimation depends on how effectively one can estimate the correlation coefficients. In some time instance, the microphone input may have speech components in the near-end environment mixed with the echoes, which makes the estimation of correlation coefficients very difficult. Hence, we need a setup to detect the echo only time instance on the microphone input and hence estimate the relevant correlation coefficients. This detection is achieved through a DTD.

A number of double talk detection algorithms are available in time domain and frequency domain literature [8], [9], [10]. All these algorithms are based on the correlation measurements between the signals. As the proposed system is operating in modulation domain, correlation based DTD algorithm in this domain is introduced and is described below.

Let $y(t)$ be the microphone input signal, $x(t)$ be the incoming far-end signal, $d(t)$ be the echo estimate and $e(t)$ be the echo cancelled signal. Let $Y(k, f_{ind}, l)$, $X(k, f_{ind}, l)$, $D(k, f_{ind}, l)$ and $E(k, f_{ind}, l)$ be the corresponding modulation spectra. Then, the power spectral densities between these signals are calculated using the below expressions.

$$\begin{aligned} C_{ye}(k, f_{ind}, l) &= \alpha C_{ye}(k, f_{ind} - 1, l) + (1 - \alpha)|Y(k, f_{ind}, l)E(k, f_{ind}, l)'| \\ C_{yd}(k, f_{ind}, l) &= \alpha C_{yd}(k, f_{ind} - 1, l) + (1 - \alpha)|Y(k, f_{ind}, l)D(k, f_{ind}, l)'| \\ C_{yy}(k, f_{ind}, l) &= \alpha C_{yy}(k, f_{ind} - 1, l) + (1 - \alpha)|Y(k, f_{ind}, l)Y(k, f_{ind}, l)'| \\ C_{ee}(k, f_{ind}, l) &= \alpha C_{ee}(k, f_{ind} - 1, l) + (1 - \alpha)|E(k, f_{ind}, l)E(k, f_{ind}, l)'| \\ C_{dd}(k, f_{ind}, l) &= \alpha C_{dd}(k, f_{ind} - 1, l) + (1 - \alpha)|D(k, f_{ind}, l)D(k, f_{ind}, l)'| \end{aligned} \quad (10)$$

Here, The cross spectra $C_{yd}(k, f_{ind}, l)$ and $C_{ye}(k, f_{ind}, l)$ are the cross power spectrum of the echo signal and echo cancelled signal with the microphone signal and $C_{yy}(k, f_{ind}, l)$, $C_{ee}(k, f_{ind}, l)$ and $C_{dd}(k, f_{ind}, l)$ represent the auto power spectra of the corresponding signals indicated with the prefix. Let α be the recursion parameter, which is closer to unity.

After estimating the above parameters, the normalized cross-correlation is calculated as

$$R_{ye}(k, f_{ind}, l) = \frac{C_{ye}(k, f_{ind}, l)}{\sqrt{C_{yy}(k, f_{ind}, l) * C_{ee}(k, f_{ind}, l)}} \quad (11)$$

$$R_{yd}(k, f_{ind}, l) = \frac{C_{yd}(k, f_{ind}, l)}{\sqrt{C_{yy}(k, f_{ind}, l) * C_{dd}(k, f_{ind}, l)}} \quad (12)$$

Then, the Double Talk Detection (DTD) algorithm sets the flag to indicate the presence of double talk situation through the algorithm specified below. If the Flag is set to ON, it indicates the presence of relevant near-end speech in the microphone input and hence the filter adaptation should pause a while till the flag becomes OFF.

Algorithm 1 Proposed Modulation Domain DTD

```

if  $DTD_{begin} = ON$  then
     $T1 \leftarrow 0.8$ 
     $T2 \leftarrow 0.2$ 
    for  $f_{ind} = 1$  to  $end$  do
        if  $R_{ye}(k, f_{ind}, l) > T1$  then
            if  $R_{yd}(k, f_{ind}, l) < T2$  then
                 $Flag \leftarrow OFF$ 
            else
                 $Flag \leftarrow ON$ 
            else
                 $Flag \leftarrow OFF$ 

```

After estimating the echo path response $G_c(k, f_{ind}, l)$ employing the above mentioned double talk detector, the echo estimate in the modulation domain can be obtained using the estimated coloration filter gain.

$$\hat{Y}(k, f_{ind}, l) = G_c(k, f_{ind}, l)|X_d(k, f_{ind}, l)| \quad (13)$$

The modulation domain Weiner gain filter expression for the suppression of the estimated disturbance spectra is similar to the normal noise suppression filter with estimated echo spectra being the disturbance spectra instead of noise spectra. The expression for the gain is given by

$$G_w(k, f_{ind}, l) = \left[\frac{\max(|Y(k, f_{ind}, l)|^2 - \beta|\hat{Y}(k, f_{ind}, l)|^2, 0)}{|Y(k, f_{ind}, l)|^2} \right]^{\frac{1}{2}} \quad (14)$$

where the parameter β represents the echo estimation efficiency. The value of β is chosen to be greater than one, when echo is under estimated and less than one, when overestimated. As we are considering only the later reflections from the path while modelling the echo path, it is an under estimation case.

IV. MODIFICATION BY EMPLOYING SPEECH UNCERTAINTY

The basic objective of any acoustic echo cancellation method is to suppress the echo components at the microphone input completely. Instead of using the simple Wiener gain filter to suppress the echo components, it will be better if we adapt the gain based on the presence and absence of the relevant near-end speech. The idea is that the system should give maximum attenuation to the microphone signal when the retrieved signal has only echo components and the attenuation should be minimum under the near-end speech only situation.

To make the gain adaptive as discussed above, we have modelled the microphone signal as a binary hypothesis process. The first hypothesis ($H0$) indicates the case where microphone has echo components ($E(k, f_{ind}, l)$) only whereas the second hypothesis ($H1$) indicates the echo plus near-end speech ($S(k, f_{ind}, l)$) situation.

$$H0 : Y(k, f_{ind}, l) = E(k, f_{ind}, l) \quad (15)$$

$$H1 : Y(k, f_{ind}, l) = E(k, f_{ind}, l) + S(k, f_{ind}, l) \quad (16)$$

Under the assumption that the signals $S(k, f_{ind}, l)$ and $E(k, f_{ind}, l)$ are zero mean complex Gaussian process with variance $\sigma_s(k, f_{ind}, l)$ and $\sigma_e(k, f_{ind}, l)$. Then the corresponding distribution becomes

$$p(Y(k, f_{ind}, l)/H0) = \frac{1}{\pi(\sigma_e(k, f_{ind}, l) + \sigma_n(k, f_{ind}, l))} \exp\left(-\frac{|Y(k, f_{ind}, l)|^2}{\sigma_e(k, f_{ind}, l) + \sigma_n(k, f_{ind}, l)}\right) \quad (17)$$

$$p(Y(k, f_{ind}, l)/H1) = \frac{1}{\pi(\sigma_e(k, f_{ind}, l) + \sigma_n(k, f_{ind}, l) + \sigma_s(k, f_{ind}, l))} \exp\left(-\frac{|Y(k, f_{ind}, l)|^2}{\sigma_e(k, f_{ind}, l) + \sigma_n(k, f_{ind}, l) + \sigma_s(k, f_{ind}, l)}\right) \quad (18)$$

Then the probability of hypothesis one to happen can be computed using the Bayesian theorem as,

$$p(H0/Y(k, f_{ind}, l)) = \frac{p(Y(k, f_{ind}, l)/H0)p(H0)}{p(Y(k, f_{ind}, l)/H0)p(H0) + p(Y(k, f_{ind}, l)/H1)p(H1)} = \frac{1}{qf(k, f_{ind}, l)} \quad (19)$$

where $q = p(H1)/p(H0)$ with $p(H1) = 1 - p(H0)$ and the function $f(k, f_{ind}, l)$ will be

$$f(k, f_{ind}, l) = \frac{p(Y(k, f_{ind}, l)/H1)}{p(Y(k, f_{ind}, l)/H0)} = \frac{1}{1 + \zeta(k, f_{ind}, l)} \exp\left(\frac{\gamma(k, f_{ind}, l)\zeta(k, f_{ind}, l)}{1 + \zeta(k, f_{ind}, l)}\right) \quad (20)$$

The parameters $\gamma(k, f_{ind}, l)$ is the *a posteriori* signal to disturbance ratio (SDR) and $\zeta(k, f_{ind}, l)$ is the *a priori* SDR

Fig. 1: Block diagram of the proposed Acoustic Echo Suppression system

which is estimated using the decision direct approach [2] with parameter $\alpha=0.85$ as

$$\hat{\zeta}(k, f_{ind}, l) = \alpha \frac{|\hat{S}(k, f_{ind} - 1, l)|^2}{\lambda(k, f_{ind}, l)} + (1-\alpha)P[\gamma(k, f_{ind}, l) - 1] \quad (21)$$

where the function $P[x]$ is defined as

$$P[x] = \begin{cases} 0 & \text{if } x < 0, \\ x & \text{otherwise} \end{cases}$$

The parameter $\gamma(k, f_{ind}, l)$ can be calculated through the expression

$$\gamma(k, f_{ind}, l) = \frac{|Y(k, f_{ind}, l)|^2}{\lambda(k, f_{ind}, l)} \quad (22)$$

After estimating the hypothesis probability, the final combined gain filter expression considering the speech uncertainty becomes

$$G(k, f_{ind}, l) = G_w(k, f_{ind}, l)(1 - p(H0/Y(k, f_{ind}, l))) \quad (23)$$

Then, the echo suppressed signal can be obtained by filtering the microphone input using the estimated filter gain.

$$E(k, f_{ind}, l) = G(k, f_{ind}, l)|Y(k, f_{ind}, l)| \quad (24)$$

This is the expression of the echo suppressed signal in modulation domain and the corresponding time domain signal can be obtained by taking the inverse transformation. The block diagram of the proposed Acoustic Echo Suppression system employing speech uncertainty is shown in Fig.1

V. EXPERIMENTAL ANALYSIS

For analysis purpose, the male and female speech segments with a sampling frequency of $8kHz$ are taken from the NOIZEUS data base [11]. Overlapping speech frames of size 32ms with an overlapping factor 0.25 are used for processing. To make the spectral continuity in the final processed signal, the signal is windowed using a hamming window before taking the 256 point Short Time Fourier Transform (STFT). The modulation spectra are obtained by taking the windowed STFT of the acoustic domain spectral amplitudes choosing a frame length equal to 30ms in the time domain and with 10ms overlapping [12].

The performance of the proposed system is compared with that of the existing frequency domain echo suppression technique proposed by Park et.al [7]. To evaluate the performance

TABLE I: Mean Opinion Score (MOS)

Noises	SNR (dB)	Park et.al	MD	Proposed
White	10	2.52	2.55	2.65
	20	2.75	2.70	2.79
	30	3.10	3.10	3.25
Babble	10	3.11	3.25	3.45
	20	3.65	3.60	3.90
	30	3.85	3.85	3.95
Factory	10	2.30	2.33	2.75
	20	3.25	3.20	3.55
	30	3.40	3.66	3.90
Cockpit	10	2.64	3.00	3.12
	20	3.13	3.44	3.52
	30	3.65	3.80	3.95

improvement by introducing the speech uncertainty gain term, the comparison is done with our previous work on acoustic echo suppression system in modulation domain with simple MMSE gain [5]. The robustness of the proposed system under noisy conditions is tested using a number of noisy speech samples with different SNR (5dB to 20dB) values which are generated by mixing the noises from NOISEX data base [13] to the clean speech segments. The echo path is modelled as a system to fit with a room with dimensions $5 \times 4 \times 3m^3$ and reflection coefficients 0.6 [14]. The echo is generated by filtering the incoming signal with the generated filter response with the above specifications. For analysis purpose, the echo power at the microphone input is corrected so as to keep the mean power 3.5dB less than the near-end speech [15].

The objective quality evaluation of the proposed system is performed through the evaluation of the Echo Return Loss Enhancement (ERLE) and Speech Attenuation (SA). The ERLE defines the system robustness towards the echo suppression and SA defines the deterioration happened to the near-end speech while performing the echo suppression.

$$ERLE(n) = 10 * \log \frac{E\{\hat{s}(n)^2\}}{E\{y(n)^2\}} \quad (25)$$

$$SA(n) = 10 * \log \frac{P_s}{P} \quad (26)$$

where the signals $\hat{s}(n)$ represents the echo cancelled signal and $y(n)$ represents the microphone input, P_s is the mean power of the clean near-end speech signal and P is the mean power of the processed near-end speech components at the output of the echo canceller.

ERLE values are obtained for 12 speech files and the mean value is shown in the Fig.4 over a range of SNR values. For the visual inspection, the speech signals are plotted in time and frequency domain in Fig.2 and 3 respectively. To determine the effectiveness of echo suppression, the speech attenuation (SA) factor of the near-end speech is found out. Subjective quality measurement is done using the mean opinion score (MOS). MOS is obtained for the 12 speech samples and the mean MOS are tabulated in Table I for a range of SNR values.

Fig. 2: Output waveforms from different algorithms. (a) microphone input (echo+noise+near-end speech). (b) output obtained by Park et al. (c) output obtained by the modulation domain technique. (d) output obtained by proposed method.

Fig. 3: Speech Spectrogram from different algorithms. (a) microphone input (echo+noise+near-end speech). (b) output obtained by Park et al. (c) output obtained by the modulation domain technique. (d) output obtained by proposed method.

Fig. 4: Echo Return Loss Enhancement (ERLE) plot.

Fig. 5: Speech Attenuation (SA) plot.

A. Results and Discussions

The robustness of the proposed system employing the speech uncertainty is analysed and compared with that of the competent system in frequency domain. The advantage of incorporating the speech uncertainty over the simple MMSE modulation domain echo canceller is also illustrated. Performance comparison is also done with our previous work on echo cancellation in modulation domain. The speech segment was created in such a way that the initial echo only duration is followed by a double talk situation (echo+near-end). From these plots, it is clearly observed that the proposed system in modulation domain completely attenuates the echo components under the single talk case, where the frequency domain technique fails. The spectral plot of the proposed system shows the suppression of random spectral spikes, which were not suppressed using the frequency domain methods.

In order to perform a quantitative analysis, the Echo Return

Loss Enhancement (ERLE) of the different speech segments are calculated and averaged to get mean ERLE. ERLE conveys the robustness of a system towards the echo suppression. Here, we have plotted the mean ERLE values under different SNR conditions. This is to get a better understanding on the variation in echo estimation and suppression under different noise conditions. From Fig.4, it is evident that the proposed system in modulation domain attenuates the echo components effectively under various SNR conditions and the system performs much better than the conventional systems.

The speech attenuation (SA) measures the degradation happened to the relevant near-end speech signal while performing the acoustic echo suppression. A robust system should give less attenuation to the near-end speech. Speech attenuation (SA) plots in Fig.5 clearly indicate that the proposed system outperforms the conventional frequency domain method. But, when comparing with the basic modulation domain MMSE method, the proposed system gives slightly more attenuation in the high SNR range. Finally, the Mean Opinion Scores (MOS) of the various speech segments have been collected from different listeners, and are tabulated in Table I. This also indicates the performance improvement of the proposed system over the existing methods under different noisy conditions.

VI. CONCLUSION

A novel robust acoustic echo suppression system in modulation domain by employing the speech uncertainty is developed in this paper. The gain adaptation and modification are achieved through modelling the microphone input as a binary hypothesis problem corresponding to the presence and absence of near-end speech components. Performance evaluation using speech samples from the NOIZEUS data base reveals that the proposed framework performs convincingly better than other echo suppression methods in terms of Echo Return Loss Enhancement (ERLE) and perceptual quality (MOS) of the reconstructed near-end speech signal.

REFERENCES

- [1] R. Martin, "Spectral subtraction based on minimum statistics," *power*, vol. 6, p. 8, 1994.
- [2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 32, no. 6, pp. 1109–1121, 1984.
- [3] C. Faller and C. Tournery, "Robust acoustic echo control using a simple echo path model," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 5. IEEE, 2006, pp. V–V.
- [4] —, "Estimating the delay and coloration effect of the acoustic echo path for low complexity echo suppression," in *Proc. Intl. Works. on Acoust. Echo and Noise Control (IWAENC)*, no. LCAV-CONF-2005-030, 2005.
- [5] P. M. Shifas, E. Jayakumar, and P. Sathidevi, "Robust acoustic echo suppression in modulation domain," in *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*. Springer, 2018, pp. 527–537.
- [6] E. Jayakumar, P. M. Shifas, and P. Sathidevi, "Integrated acoustic echo and noise suppression in modulation domain," *International Journal of Speech Technology*, vol. 19, no. 3, pp. 611–621, 2016.
- [7] Y.-S. Park and J.-H. Chang, "Frequency domain acoustic echo suppression based on soft decision," *Signal Processing Letters, IEEE*, vol. 16, no. 1, pp. 53–56, 2009.
- [8] T. Gänslér and J. Benesty, "A frequency-domain double-talk detector based on a normalized cross-correlation vector," *Signal Processing*, vol. 81, no. 8, pp. 1783–1787, 2001.

- [9] V. Das, A. Kar, and M. Chandra, "A new cross correlation based double talk detection algorithm for nonlinear acoustic echo cancellation," in *TENCON 2014-2014 IEEE Region 10 Conference*. IEEE, 2014, pp. 1–6.
- [10] J. Benesty, D. R. Morgan, and J. H. Cho, "A new class of doubletalk detectors based on cross-correlation," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 2, pp. 168–172, 2000.
- [11] Y. Hu and P. C. Loizou, "Evaluation of objective measures for speech enhancement," in *Interspeech*, 2006.
- [12] K. Paliwal, B. Schwerin, and K. Wójcicki, "Speech enhancement using a minimum mean-square error short-time spectral modulation magnitude estimator," *Speech Communication*, vol. 54, no. 2, pp. 282–305, 2012.
- [13] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.
- [14] S. McGovern, "<http://sgm-audio.com/research/rir/rir.html>," 2003.
- [15] S. Gustafsson, R. Martin, and P. Vary, "Combined acoustic echo control and noise reduction for hands-free telephony," *Signal Processing*, vol. 64, no. 1, pp. 21–32, 1998.