Chapter 2

Speech Enhancement Techniques: State of Art

The speech signal degradations may be attributed to various factors; viz. disorders in production organs, different sensors (microphones) and their placement (hands free), acoustic non-speech and speech background, channel and reverberation effect and disorders in perception organs. Considerable research recently has examined ways to enhance speech, mostly related to speech distorted by background noise (occurring at the source or in transmission)-both wideband (and usually stationary) noise and (less often) narrowband noise, clicks, and other non-stationary interferences [1-7]. Most cases assume noise whose pertinent features change slowly (i.e., locally stationary over analysis frames of interest), so that it can be characterized in terms of mean and variance (i.e., second-order statistics), either during non-speech intervals (pauses) of the input signals or via a second microphone (called reference microphone) receiving little speech input [1].

In ideal scenario there should be no degradation in quality and/or intelligibility of original speech and/or human subjects have normal speech production and perception systems. In practical scenario there is degradation in quality and/or intelligibility and/or human subjects have impaired speech production and perception systems. So the goal of speech enhancement is to enhance quality and intelligibility. Except when inputs from multiple microphones are available (in some specially arranged cases), it has been very difficult for speech enhancement systems to improve intelligibility. Thus most speech enhancement methods raise quality, while minimizing any loss in intelligibility. As observed, certain aspects of speech are more perceptually important than others. The auditory system is more sensitive to the presence than absence of energy, and tends to ignore many aspects of phase. Thus speech enhancement algorithms often focus on accurate modeling of peaks in the speech amplitude spectrum, rather than on phase relationships or on energy at weaker frequencies. Voiced speech, with its high amplitude and concentration of energy at low frequency, is more perceptually important than unvoiced speech for preserving quality. Hence, speech enhancement usually emphasizes improving the periodic portions of speech. Good representation of spectral amplitudes at harmonic frequencies and especially in the first three formant regions is paramount for high speech quality. All enhancement algorithms introduce their own distortion and care to be taken to minimize distortion

Weaker, unvoiced energy is important for intelligibility, but obstruent are often the first to be lost in noise and the most difficult to recover. Some perceptual studies claim that such sounds are less important than strong voiced sounds (e.g., replacing the former by noise of

corresponding levels causes little decrease in intelligibility). In general, however, for good intelligibility, sections of speech (both voiced and unvoiced) undergoing spectral transitions (which correspond to vocal tract movements) are very important. Speech enhancement often attempts to take advantage of knowledge beyond simple estimates of SNR in different frequency bands. Some systems combine speech enhancement and automatic speech recognition (ASR), and adapt the speech enhancement methods to the estimated phonetic segments produced by the ASR component. Since ASR of noisy speech is often less reliable, simpler ASR of broad phonetic classes is more robust, yet allows improved speech enhancement [2].

2.1 Interferences and Suppression Techniques

Different types of interference may need different suppression techniques. Noise may be continuous, impulsive, or periodic, and its amplitude may vary across frequency (occupying broad or narrow spectral ranges); e.g., background or transmission noise is often continuous and broadband (sometimes modeled as "white noise"- uncorrelated time samples, with a flat spectrum). Other distortions may be abrupt and strong, but of very brief duration (e.g., radio, static, fading). Hum noise from machinery or from AC power lines may be continuous, but present only at a few frequencies. These noises are generally additive in nature. Most speech enhancement techniques are devised to handle the additive background noise. Noise which is not additive (e.g., multiplicative or convolutional) can be handled by applying a logarithmic transformation to the noisy signal, either in the time domain (for multiplicative noise) or in the frequency domain (for convolution noise), which converts the distortion to an additive one (allowing basic speech enhancement methods to be applied). Varieties of techniques are devised to handle convolutive distortion and reverberation.

Interfering speakers present a different problem for speech enhancement. When people hear several sound sources, they can often direct their attention to one specific source and perceptually exclude others. This "cocktail party effect" is facilitated by the stereo reception via a listener's two ears [3]. In binaural sound reception, the waves arriving at each ear are slightly different (e.g., in time delays and amplitudes); one can often localize the position of the source and attend to that source, suppressing perception of other sounds. How the brain suppresses such interference, however, is poorly understood. Monaural listening (e.g., via a telephone handset) has no directional cues, and the listener must rely on the desired sound source being stronger (or having major energy at different frequencies) than competing sources. When a desired source

can be monitored by several microphones, techniques can exploit the distance between microphones [3]. However, most practical speech enhancement applications involve monaural listening, with input from one microphone. Directional and head-mounted noise-cancelling microphones can often minimize the effects of echo and background noise. The speech of interfering speakers occupies the same overall frequency range as that of a desired speaker, but such voiced speech usually has fundamental (pitch) frequency F0 and harmonics at different frequencies. Thus some speech enhancement methods attempt to identify the strong frequencies either of the desired speaker or of the unwanted source, and to separate their spectral components to the extent that the components do not overlap. Interfering music has properties similar to speech, allowing the possibility of its suppression via similar methods (except that some musical chords have more than one F0, thus spreading energy to more frequencies than speech does). The multi speech separation (speaker separation) requires multiple microphone solution. The single microphone techniques are not sufficient for this type of interference. Very little literature is available and still this problem is not exactly solved for any general case.

2.2 Recent Trends - Speech Enhancement Techniques

The approach to speech enhancement varies considerably depending upon type of degradation. The speech enhancement techniques can be divided into two basic categories: (i) Single channel and (ii) Multiple channels (array processing) based on speech acquired from single microphone or multiple microphone sources respectively [3]. However, single channel (one microphone) signal is available for measurement or pick up in real environments and hence focus is here on single channel speech enhancement methods. Figure 2.1 shows the chart of the latest single channel speech enhancement methods for three different kinds of problems.

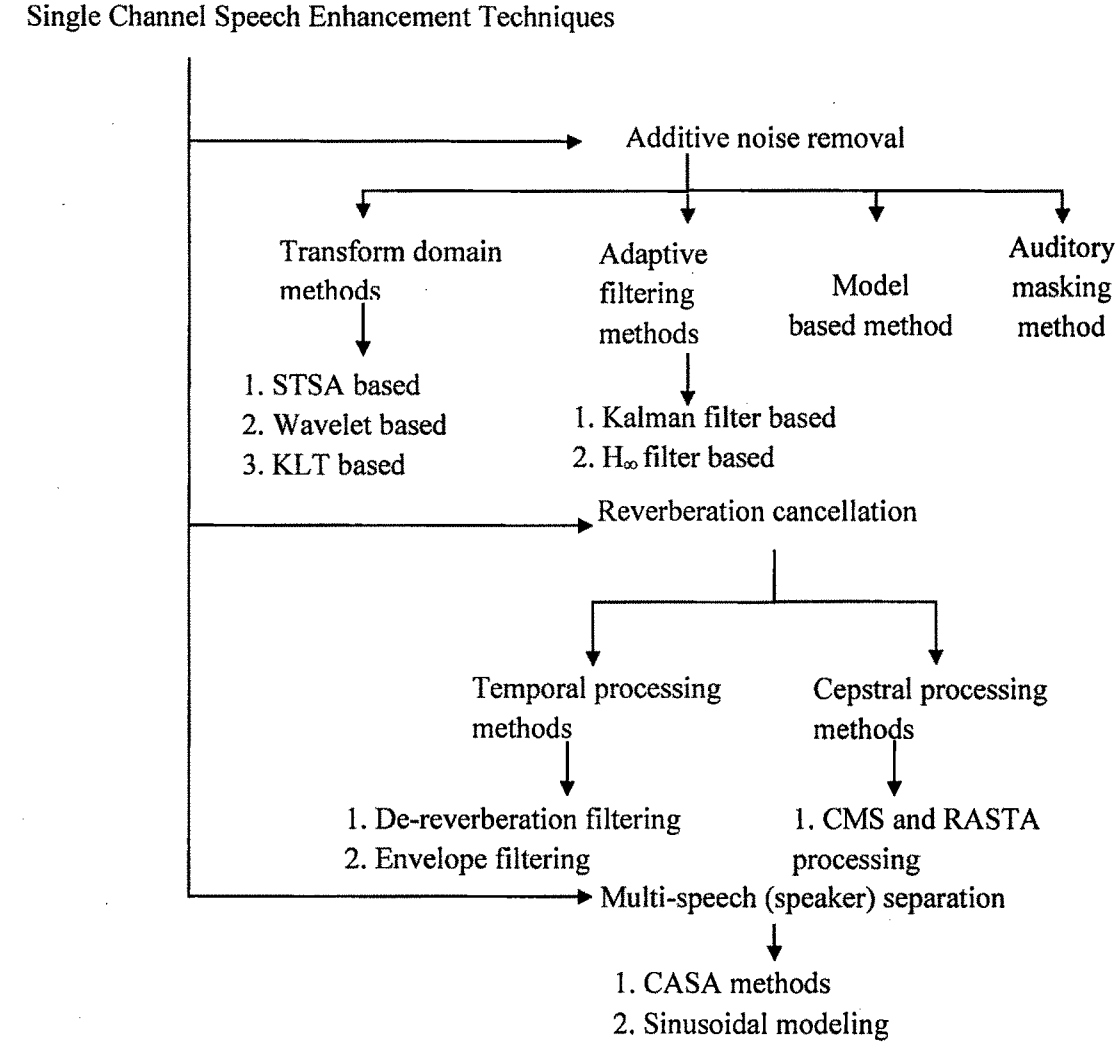


Fig. 2.1 A chart showing summary of existing speech enhancement methods

2.2.1 Additive Noise Removal

In most cases the background random noise is added with the desired speech signal and forms an additive mixture which is picked up by microphone. It can be stationary or non-stationary, white or colored and having no correlation with desired speech signal. Variety of methods suggested in literature so far to overcome this problem. The majority of them belong to following four categories.

2.2.1.1 Transform Domain Methods

The most commonly used methods are transform domain methods. They are most conventional methods. They transfer the time domain signal into other domain using different transforms and involve some kind of filtering to suppress noise and then inverse transform filtered signal into time domain. They follow the analysis-modify-synthesis approach. The transformation used is DFT, WT or KLT.

- **DFT based (STSA methods):** They are most popular as they have less computational complexity and easy implementation. They use short time DFT (STDFT) and have been intensively investigated; also known as spectral processing methods. They are based on the fact that human speech perception is not sensitive to spectral phase but the clean spectral amplitude must be properly extracted from the noisy speech to have acceptable quality speech at output and hence they are called short time spectral amplitude (STSA) based methods [5,7]. In practice power density of signal is used instead of amplitude. Methods of this category remove an estimate of noise from noisy signal using spectral subtraction (SS). The noise power spectrum estimation is obtained by averaging over multiple frames of a known noise segment; which can be detected using voice-activity detector (VAD) [4]. However the basic SS method suppresses noise but it has limitation in terms of an artefact called *musicality* [2]. This gives rise to distortion in enhanced speech. Several modifications in basic method are suggested by Boll and Berouti *et al.* [4] to reduce the musical noise. However this requires very careful parameter selections. The other modification in basic SS is using McAuly's maximum likelihood (ML) estimation [4] of output speech; which assumes noise with complex Gaussian distribution. In general all SS methods estimate *a posteriori* SNR. Also SS methods are suitable for stationary white noise only. The solutions to this are suggested using smoothing time varying filter called Wiener filter [4]. The combination of SS and Wiener filter is used in most real applications.

  The optimal Wiener filter for the noisy speech can be designed in frequency domain via the estimated ratio of the power spectrum of clean speech; called object power spectrum to that of noisy speech (*a priori* SNR). This spectrally varying attenuation accommodates coloured noise, and can be updated at any desired frame rate to handle non-stationary noise. A major problem with this approach is estimating background noise spectrum at every frame which is limited by the performance of VAD. This requires noise adaptation [4] in VAD for

every frame. However estimation of object power spectrum considering current frame only is non-realistic as well as time varying and non-stationary process. A solution to this is suggested by Ephrahim and Malah [4] known as decision direct (DD) method which estimates *a priori* SNR of current frame using *a posteriori* SNR of current frame, estimated noise for current frame and estimated clean speech in previous frame. So in practice a Wiener filter is combined with DD approach to give realistic system. The Wiener filter shows substantial reduction in musical noise artefacts compared to SS methods.

The realistic and optimal object power spectrum estimation without artefacts requires model based statistical methods. The stochastic estimation methods such as minimum mean square error (MMSE) and its variant MMSE log spectral amplitude (LSA) suggested by Ephrahim and Malah [4] are commonly used estimation methods. They are based on modelling spectral components of speech and noise processes as independent Gaussian variables. Almost all literature mentions that the performance of Wiener filter and MMSE LSA is outstanding in terms of both subjective and objective evaluations. The stochastic estimation method called MAP (maximum *a posteriori*) is very close in performance with MMSE LSA with simpler computations. All of these methods assume speech presence in the frequency bin under consideration; but it is not always true. These methods can be extended by incorporating a two state speech presence/absence model which leads to a soft decision based spectral estimation and further improves performance at the cost of computational complexity. Further improvements were observed by using Laplacian model for speech spectral coefficients rather than Gaussian model. The various kinds of noise adaptation strategies used like hard/soft/mixed decision also affect the performance. The soft decision based noise adaptation found satisfactory in removing musical artefact but at the cost of increased processing requirements.

A background noise suppression system developed by Motorola is included as a feature in IS-127, the TIA/EIA standard for the Enhanced Variable Rate Codec (EVRC) to be used in CDMA based telephone systems [8]. EVRC was modified to EVRC-B and later on replaced by Selectable Mode Vocoder (SMV) which retained the speech quality at the same time improved network capacity. Recently, however, SMV itself has been replaced by the new CDMA2000 4GV codecs. 4GV is the next generation 3GPP2 standards-based EVRC-B codec [9]. The EVRC based codec uses combination of STSA based approaches: multiband

spectral subtraction (MBSS) and minimum mean square error (MMSE) gain function estimator for background noise suppression as a pre-processor. The voice activity detector (VAD) used to decide speech/silence frame is embedded within the algorithm. Its quality has been proven good through commercial products. Nevertheless, the quality may not be sufficiently good for a wide range of SNRs, which were not given much attention when it was standardized. Another algorithm suggested by A.Sugiyama, M.Kato and M. Serizawa [3] uses modified MMSE-STSA approach based on weighted noise estimation. The subjective tests on this algorithm claim to give maximum difference in mean opinion score (MOS) of 0.35 to 0.40 compared to EVRC and hence its later version is equipped within 3G handsets. The modified STSA-MMSE algorithm based on weighted noise estimation is employed in millions of 3G handsets as the one and only commercially available 3GPP-endorsed noise suppressor [3]. But still there are open questions like how the parameters of statistical models can be estimated in a robust fashion and what can be meaningful optimization criteria for speech enhancement; which will require further research.

- **Wavelet based:** The DFT based methods use short time spectral measurements and hence are suffered by time-frequency resolution trade-offs. Wavelet based methods are developed which provides more flexibility in time-frequency representation of speech. The Wavelet de-noising algorithm is most commonly used and based on soft thresholding [7, 10] of the Wavelet coefficients. However uniform thresholding results in suppression of noise as well as unvoiced components of desired speech. So, Wavelet transform combined with smoothing filter like Wiener filter in Wavelet domain is suggested. Presently, a method is suggested in which the soft thresholding decision is taken based on statistical models. Unfortunately Wavelet based techniques are failed to achieve the great success and popularity in speech enhancement. The STFT and Wavelet based techniques are described in next chapter and simulation is presented in chapter 4.
- **KLT based:** The frequency domain methods are nowhere close to offering fully satisfactory solutions to their inherent problems: the musical noise artefact and the inevitable trade-off between signal distortion and the level of residual noise. The signal subspace approach (SSA) for speech enhancement has been originally introduced by Dendrinos *et al.* operates in eigen domain [11]. It uses the singular value decomposition (SVD) of a data matrix to remove the noise subspace and then reconstruct the desired speech signal from the remaining subspace.

This approach is modified by Ephrahim and Van Trees and proposes the use of Eigen value decomposition (EVD) of covariance matrix of input signal vector. This method consists in estimating a transform, namely the *Karhunen-Loeve transform (KLT)* [12], which will project the input signal vector into a subspace called the signal subspace hence readily eliminating the components in the orthogonal noise only subspace. The enhanced signal is reconstructed in time domain using inverse KLT. The SSA was found to outperform frequency domain methods but yet not received much attention and its use in practice is still scarce due to high computational load. However, with the sharp computation hardware available today this method can become a serious candidate to compete with the currently employed noise reduction methods.

2.2.1.2 Adaptive Filtering Methods

The adaptive filters which are mostly used in adaptive control applications can also be useful for speech enhancement. Mostly LMS and its variants are useful in multi microphone additive noise and echo cancellation problems. But for single channel speech enhancement Kalman and H_∞ adaptive filters are found suitable. They can also address the problem of colored noise removal as the noise is not always white is real environments. The transform domain methods degrade in such situations.

- **Kalman filter based**: In Wiener filtering approach the analysis has shown that the amount of noise attenuation is in general proportional to the amount of speech degradation. Kalman filtering [13] provides optimal time domain estimations and can be used instead of Wiener filtering at the cost of computational complexity and complicated implementation hardware. Literature suggests a large number of variants of basic Kalman filtering algorithm used in speech enhancement. It can be integrated with autoregressive (AR) speech models; but still the robust estimation of model parameters requires further research.
- **Robust-H_∞ filter based**: Recently H_∞ filtering [14] has been shown to overcome unrealistic assumptions of Wiener and Kalman filtering methods. Furthermore, both Wiener and Kalman estimators may not be sufficiently robust to the signal model errors. The estimation criterion in the H_∞ filter design is to minimize the worst possible effects of the modeling errors and additive noise on the signal estimation errors. Since the noise added to speech is not Gaussian in general, this filtering approach appears highly robust and more appropriate in practical speech enhancement. Furthermore, the H_∞ filtering algorithm is straightforward to

implement. Still this algorithm has not got enough attention in implementation for speech enhancement.

As a preliminary work the simulation and implementation of adaptive noise and echo cancellation is described in section 2.4.

2.2.1.3 Model Based Methods

The third method adopts a specific speech production model (e.g., from low-rate coding), and reconstructs a clean speech signal based on the model, using parameter estimates from the noisy speech [1, 7]. This method improves speech signals by parametric estimation and speech re-synthesis. Speech synthesizers generate noise-free speech from parametric representations of either a vocal tract model or previously analyzed speech. Most synthesizers employ separate representations for vocal tract shape and excitation information, coding the former with about 10 spectral parameters (modeling the equivalent of formant frequencies and bandwidths) and coding the latter with estimates of intensity and periodicity (e.g., F0). Standards methods (e.g., LPC) do not replicate the spectral envelope precisely, but usually preserve enough information to yield good output speech. Such synthesis suffers from the same mechanical quality as found in low-rate speech coding and from degraded parameter estimation (due to the noise), but can be free of direct noise interference, if the parameters model the original speech accurately. In general, re-synthesis is the least common of the speech enhancement techniques, due to the difficulty of estimating model parameters from distorted speech and due to the inherent flaws in most speech models. It nonetheless has application in certain cases like improving the speech of some handicapped speakers.

2.2.1.4 Auditory Masking Methods

Several perceptual based approaches are also investigated, where unwanted component of signal is masked by the presence of another component and taking advantage of simultaneous masking property of human auditory system. Instead of removing all noise from signal these methods attempt to attenuate the noise below the audible threshold. Virag [15] proposed the noise reduction algorithm based on this principle and shown that the auditory masking algorithm outperforms other noise suppression algorithms with respect to human perception; the algorithm was judged to reduce musical artifacts and give acceptable speech distortion. However, the disadvantage is the large computational load due to sub-band decomposition and additional DFT analyzer required for psychoacoustic modeling. The RelAtive SpecTral Amplitude processing

(RASTA) algorithm proposed by Hermnsky and Morgan [19] to enhance speech for automatic speech recognition in reverberant environment. This algorithm was later modified for additive noise removal. This algorithm is required further investigations and can be tested for real-time implementation. The RASTA algorithm and its simulation are described in chapter 5.

2.2.2 Reverberation Cancellation

The reverberation is a convolutive distortion that occurs to the speech while it is picked up by microphone. The speech signal is convolved with ambient or channel impulse response. The objective here is to recover the original speech without *a priori* information of channel or environment through which speech is collected or recorded. The acoustic echo can also be considered as one kind of reverberation effect. The *blind deconvolution* is the obvious remedy to the reverberation and acoustic echo which involves some kind of inverse filtering and equalization operation. They basically classified in two categories; however multistage algorithms [16, 17] which use combination of these methods are also proposed in literature.

2.2.2.1 Temporal Processing Methods

The temporal processing methods obtain the enhancement by processing the reverberant speech in time domain.

- **De-reverberation filtering**: Here the signal is passed through a filter having impulse response that is inverse of reverberation process. A blind estimation of filter is always difficult. Douglas *et al.* and Yagnanarayan *et al.* proposed inverse filter estimation based on LP residual and Gillespe *et al.* proposed same based on correlation shaping [7]. These methods were partially successful but failed in environment with long reverberation time because of assumption made about LP residue of speech signal that it is independent and identically distributed. A more robust filter can be obtained from the harmonic structure of reverberant speech signal called *harmonicity based de-reverberation filter (HERB)* [3]. It estimates the inverse or de-reverberation filter as the time average of a filter that transforms observed reverberant signals into the output of an adaptive harmonic filter. This achieves high quality de-reverberation, provided a sufficient number of observed signals (training data) are available. Several modifications still require making it useful in real practice like reduction in training data size, enhanced approximation to speech harmonicity etc. Further research in this direction is required.

- **Envelope filtering**: This method does not require obtaining impulse response of an environment. It is based on *modulation transfer function (MTF)* of speech [2]. It assumes that the temporal envelope of surrounding environment impulse response decays exponentially with time and the carrier signals of the impulse response and a speech signal can be modelled as mutually independent white noise functions. However, these assumptions are not accurate with regard to real speech and reverberation. So, this approach yet not achieved high quality de-reverberation.

2.2.2.2 Cepstral Processing Methods

The cepstral processing methods process the speech signal in cepstral domain. The homomorphic signal processing and cepstral mean subtraction (CMS) method proposed by Oppenheim et al.[2] achieved de-reverberation by removing cepstral components corresponding to the impulse response by applying low time lifter in cepstral domain. Also as an alternate the cepstral filtering can be done using a comb filter. It is successful for cancellation of simple echoes but have a limited performance for real environments. A generalization of CMS is RelAtive SpecTral Amplitude processing (RASTA) algorithm [19]. It uses a cepstral lifter to remove high and low modulation frequencies and not simply the DC component, as does CMS. It is also motivated by certain auditory principle that auditory system is particularly sensitive to signal change. Still there is a scope of research in proper implementation of this algorithm for speech enhancement. It can also be used to remove additive noise. The RASTA algorithm is described in detail in chapter 5.

2.2.3 Multi-speech (speaker) Separation

Here a low-level speaker may be sought in presence of a loud interfering speaker and the signal picked by single microphone containing additive mixture of both signals. Here speech of other speakers is degradation and speech of desired speaker to be enhanced. The problem of multi-speaker separation is the most difficult to handle and still the research done is limited in the context of problem solution [7, 18]. There are certain problems faced here like difficulty due to spectral similarity, pitch of different speakers may cross or overlap, number of talkers is not known, talker amplitude varies in an utterance etc. Very few approaches are proposed in literature for single microphone solution to this problem.

2.2.3.1 CASA Method

One approach called computational auditory scene analysis (CASA) which replicates the perceptual processes by which human listener segregate simultaneous sounds. It involves segregating speech of desired speaker in the presence of degradation, treat speech of desired speaker as stream of segments, approach to localize and select these streams and stitch them together in sequence to obtain speech of desired speaker. Most works had been carried out recently and it suffers from two deficiencies: First, it is not able to separate unvoiced segments and second, the vocal-tract related filter characteristics are not given importance compared to excitation signal. Also, evaluation is an important issue for CASA that requires further thought. However, it is still under research and adherence to the general principles of auditory processing is likely to give rise to CASA systems that make fewer assumptions and it will turn into superior performance in real acoustic environments [3, 18].

2.2.3.2 Sinusoidal Modeling

Another approach to the problem is to use sinusoidal modeling [2] of speech. Here the speech signal generated by two different simultaneous talkers can be represented by a sum of two sets of sine waves, each with time-varying amplitudes, frequencies and phases. The algorithm separates amplitudes, frequencies and phases for each speaker and re-synthesizes the signal for each speaker. Separation of the spectra of each speaker is done with the help of her/his pitch estimation. The performance depends on how best pitch of each speaker can be estimated and joint pitch estimation is the most difficult task in multi-speaker case.

The single channel techniques are not having enough power to solve this problem. However, the multi-microphone techniques like beam forming and blind source separation are far more superior and suitable for this problem. They exploit spatial information and additional reference for processing. This problem is ruled out here in the context of single channel solution.[1]

[1] A paper entitled "A Review on Single Channel Speech Enhancement Techniques for Wireless Communication Systems" is presented in National conference on Information Sciences (NCIS-2010) organized by MCIS, Manipal University, Manipal in April 2010.

2.3 A Case Study of Speech Enhancement Technique Using Adaptive Filtering Algorithms:

The initial preliminary research work carried out has focused on single channel speech enhancement techniques where no reference signal for noise is available. However, as a preliminary starting work the two microphone enhancement technique using adaptive algorithm called adaptive noise cancellation (ANC) is taken as a case study. It is simulated and implemented for additive noise reduction and echo cancellation purposes. When more than one microphone is available to furnish pertinent signals, speech degraded by many types of noise can be handled. Processed version of a second "reference" signal $u(n)$ (containing mostly or exclusively interference noise) is directly subtracted in time from the primary noisy speech signal $y(n)$. The block diagram is shown in figure 2.2.

While other speech enhancement filtering methods get good results with a dynamic filter that adapts over time to estimated changes in the distortion, such adaptation is essential in ANC. Since there will be a delay between the times the interference reaches different microphones and since the microphones may pick up different versions of the noise (e.g., the noise at the primary microphone may be subject to echoes and/or spectrally variable attenuation), a secondary signal must be filtered so that it closely resembles the noise present in the primary signal. In most adaptive system, the digital filter used is FIR because of simplicity and guaranteed stability. There are several ways to obtain the filter coefficients, of which the most attractive is the least-mean-squares (LMS) method via steepest descent [20], due to its simplicity and accuracy. More computationally expensive exact least-squares (LS) methods typically yield only marginal gains over the faster stochastic-gradient LMS method; the latter is also useful for enhancement of one-microphone speech degraded by additive noise [1].

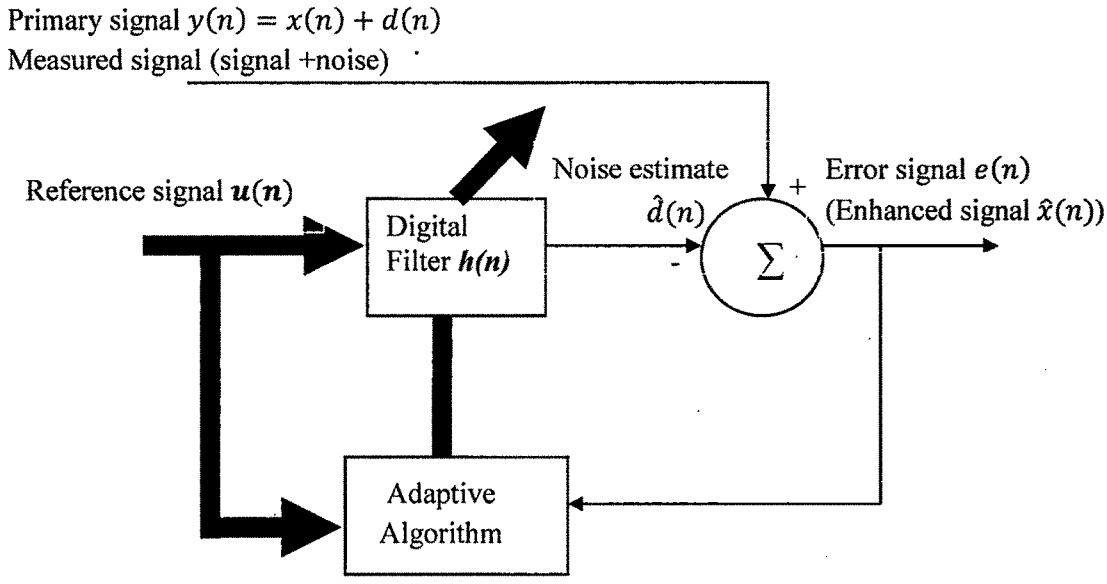


Fig. 2.2 Block diagram of adaptive noise cancellation (ANC) system

2.3.1 ANC Using NLMS Algorithm

Filter coefficients are chosen so that the energy in the difference or residual error signal $e(n)$ (i.e., the primary signal $y(n)$ minus a filtered version $\hat{d}(n)$ of the reference $u(n)$) is minimized. Thus one can select FIR filter weight co-efficients (or impulse response) $h(k)$ so that the energy in

$$e(n) = y(n) - \hat{d}(n) = y(n) - \sum_{k=1}^{L} h(k)u(n-k) \quad (2.1)$$

is minimized. Here $y(n)$ is a signal with noise to be processed and $\hat{d}(n)$ is a filtered version of reference signal $u(n)$. As long as the two microphone signals ($u(n)$ *and* $y(n)$) are uncorrelated, minimizing $e^2(n)$ (a "least mean squares" approach) over time should yield a filter that models the transformed reference, which can thus be subtracted from $y(n)$ to provide enhanced speech, which is actually the minimized residual $e(n)$. This provides the signal estimate or enhanced signal $\hat{x}(n)$. Correlation between $u(n)$*and* $y(n)$ is undesirable because then the $h(k)$ values are affected by speech and $\hat{d}(n)$ will partly contain speech rather than only transformed noise, and part of the desired speech will be suppressed. Solving Equation (2.1) can exploit LS or LPC methods, or simpler LMS techniques which do not require calculating correlation matrices or inverting them. The LMS approach uses steepest-gradient iteration [20] to get

$$h_i(n+1) = h_i(n) + \mu e(n)u(n-i) \quad (2.2)$$

Where μ is scalar parameter ($0< \mu < {}^{1}/_{MS_{max}}$). Here M is the tap size of the filter and S_{max} is power spectral density of reference input $u(n)$. It is called adaptation step size. A large value for μ speeds up convergence, but may lead to stability problems. A modified version with better stability is often used, called normalized LMS (NLMS) [20]:

$$h_i(n+1) = h_i(n) + \frac{\tilde{\mu}e(n)u(n-i)}{(\sum_{k=0}^{N-1} u^2(n-k))} = h_i(n) + \mu(n)e(n)u(n-i) \tag{2.3}$$

with control factor (step size) $0 < \tilde{\mu} < 2\frac{E[|u(n)|^2]D(n)}{E[|e(n)|^2]}$, where $E[|e(n)|^2]$ = error signal power, $E[|u(n)|^2]$ = input signal power and $D(n)$ = mean square deviation of filter weight co-efficients. It can be briefly described as follows:

- Initialization: If prior knowledge of the tap weight vector $\boldsymbol{h(n)}$ is available, use it to select an appropriate value for $\boldsymbol{h(0)}$. Otherwise, set $\boldsymbol{h(n)} = \boldsymbol{0}$.
- Data:
  - Given $\boldsymbol{u(n)}$ = M by 1 tap input vector at time $n = [u(n), u(n-1), \ldots\ldots, u(n-M-1)]^T$, $y(n)$ = noisy speech signal at time n.
  - To be computed: $\boldsymbol{h(n)}$ =estimate of tap-weight vector at time n
- Computation: $\hat{d}(n) = \boldsymbol{h(n)}^T \boldsymbol{u(n)}$,

  $$e(n) = y(n) - \hat{d}(n)$$

  $$\boldsymbol{h(n+1)} = \boldsymbol{h(n)} + \tilde{\mu}.e(n).\boldsymbol{u(n)}/||\boldsymbol{u(n)}||^2.$$

Figure 2.3 shows the flow chart to implement the algorithm.

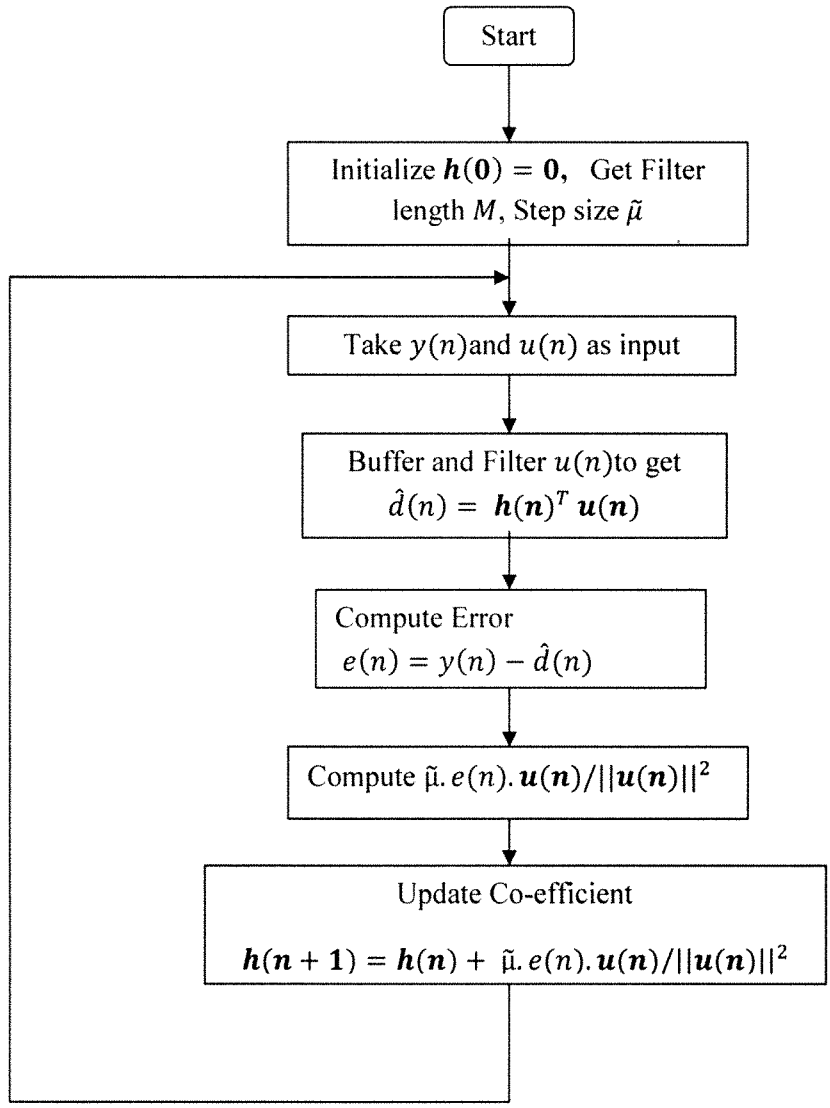
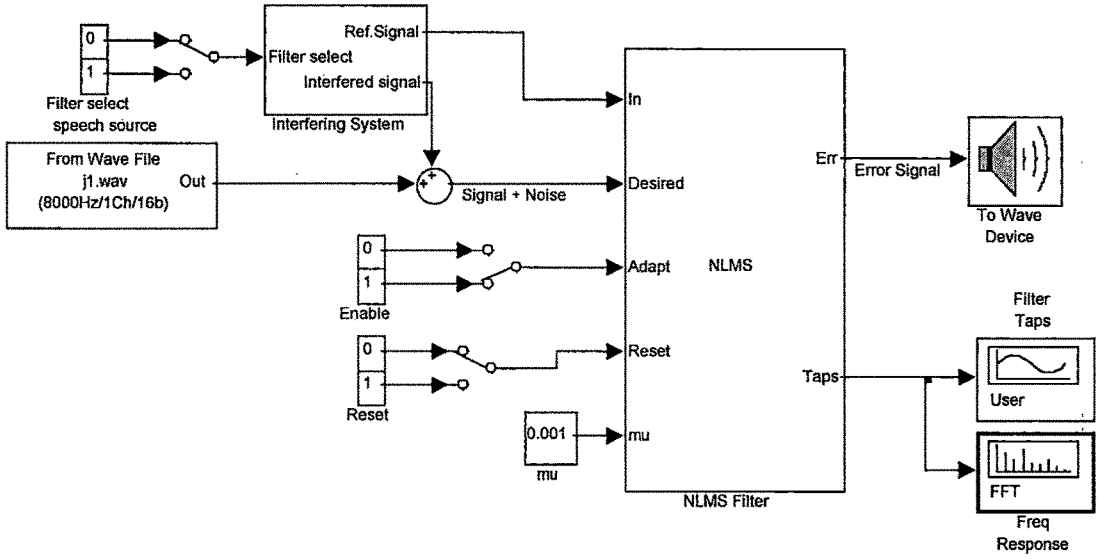


Fig. 2.3 Flow chart for implementation of ANC using NLMS algorithm

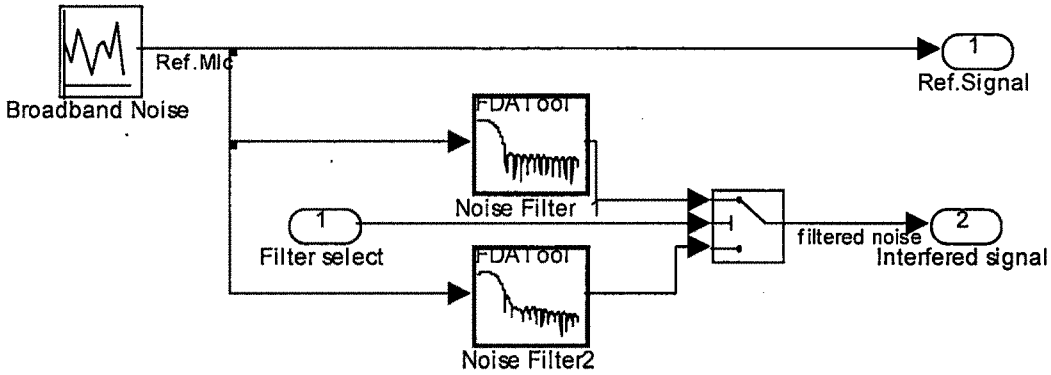
2.3.2 Practical Implementation of ANC with NLMS Algorithm

The NLMS algorithm is first implemented in SIMULINK. The SIMULINK model is prepared using the techniques like masking, subsystems, conditional subsystems, and in-built S functions etc. Here the *.wav file is used as a signal source. It is added with filtered random white noise. The parameters of filter can be selected to any suitable value using FDA Toolbox. Also the noise characteristics can be varied by selecting appropriate parameters for the Noise block in the model. The noise can be either filtered by low pass filter or by band pass filter. The

selection switch is provided in the model. The NLMS block accepts one input (on "In port") directly from reference source as white noise. The other input (on "Desired port") from added mixture of signal and filtered noise. The step size parameter (μ) can be set to any value from input port labeled (mu). Also, it has two control inputs, one to enable adaption and other to reset the filter weights to zero at any time and then allowing them to readapt. The output signal can be obtained from output port labeled "Error", which is actually clean output signal. This port is connected to speaker or headphone through PC sound card using the block "To Wave Device". To record the clean signal replace this block by "To Wave File" block and give the name of file in the parameter dialog box of that block. Also, the output port labeled "weights" can be used to see the updating of filter coefficients and variable frequency response by connecting suitable blocks at that port.



NLMS Adaptive Noise Cancellation using PC



Interfering System

Fig. 2.4 SIMULINK implementation of ANC using NLMS algorithm

Also, the same NLMS algorithm is implemented for real time application using the Real Time Workshop and Embedded Target for TI C600 Toolboxes. The hardware setup is shown in figure 2.5. Here the SIMULINK model is developed for C6713 DSP. Here the control signals like adaption enable, reset and noise filter selection is done by using switches on the DSK. The speech source signal can be applied to "Line In" or "Mic In" source of DSK depending on selected option in the block of ADC in model file. The noise source is again simulated here from the SIMULINK block using the same techniques as described above. Connecting speaker or

headphone at "Line Out" or "HP Out" port of DSK can obtain the output signal. To operate this model from beginning, it is required to follow the following procedure.

1. Connect a speech source to the 'line in' or 'mic in' jack of the DSK board.
2. Set the required parameters by choosing Simulation -> Configuration Parameters.
3. To generate code choose Tools -> Real-Time Workshop ->Build Model (or Ctrl-B).
4. After generating code, Real-Time Workshop connects to Code Composer Studio (CCS) and creates a new project. After compiling and linking the code, Real-Time Workshop downloads the COFF (Common Object File Format) file to the DSK and begins execution. At this time, if speakers (or headphone) are connected to the audio output jack of the DSK, one could hear the noisy signal.
6. Now, the system is ready to begin the adaptation algorithm. By Pressing down the user DIP switch (SW0) on the DSK, initiate the algorithm. One could hear the noise component of the signal slowly decrease in volume as the filter adapts.
7. To control the adaptive filter during execution, move the User DIP Switches as follows:

   Switch 0:

   'Off' — pause adaptation process, 'On' — start/resume NLMS adaptation process.

   Switch 1:

   'Off' — disable reset, 'On' — reset LMS adaptation process.

   Switch 2:

   'Off' — apply band pass noise model, 'On' — apply low-pass noise model.

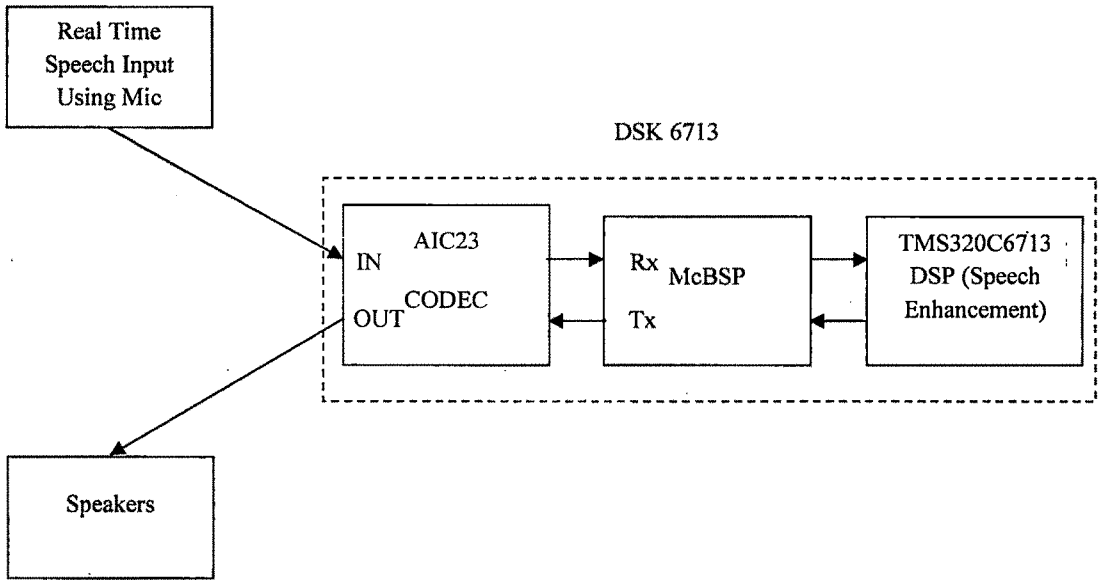
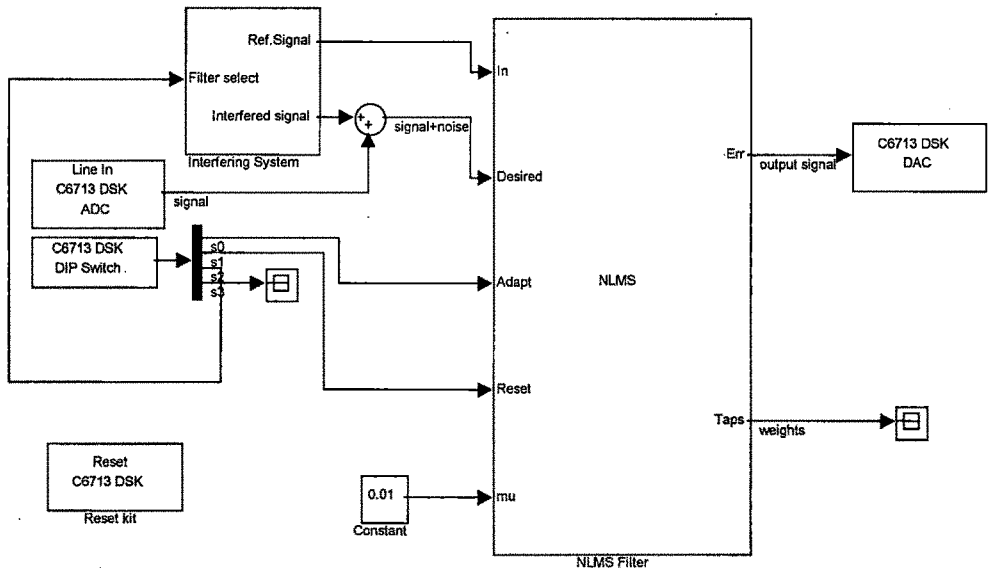


Fig. 2.5 Hardware setup for implementation on DSK 6713



Adaptive Noise Cancellation using NLMS & DSK6713

Fig. 2.6 DSK 6713 implementation of ANC using NLMS algorithm

2.3.3 Performance of NLMS algorithm for ANC

In order to design and implement any adaptive filter for a given application, it is required to determine the values of parameters such as the step size $\tilde{\mu}$, the filter length M, and the initial coefficient weight vector $\boldsymbol{h}(\mathbf{0})$. To properly select these parameters, it is required to understand important properties of adaptive algorithms [21] as summarize here.

1. Stability conditions

As seen in above sections, the adaptive filter uses FIR filter, which is inherently stable. However, the whole adaptive filter is not always stable. The stability depends on the algorithm that adjusts its coefficients. Different analysis and criteria shows that the step size $\tilde{\mu}$ must be within some range to satisfy the stability condition. In most practical cases for NLMS algorithm it should be between 0 and 1 according to optimization criterion. According to stability criterion, it should be between 0 and 2/M. The stability improves with the lower value of step size $\tilde{\mu}$, but it requires larger filter length M.

2. Convergence rate

In applications with slowly changing signal statistics, the performance function drifts in time. Adaptation is the process of tracking the signals and environments. Thus, speed of convergence is the most important considerations. Algorithm convergence is attained when the Mean Square Error (MSE) is reduced to minimum value. The analysis has shown that the average time needed for the algorithm to converge is inversely proportional to the step size $\tilde{\mu}$. But it is not recommended to use arbitrary large step sizes to speed up convergence because of the stability constraint.

3. Steady state performance

With a true gradient and under noise-free conditions, the adaptive algorithm converges to the minimum MSE and remains there because the gradient is zero at the optimum solution. But actually the NLMS algorithm not uses the true gradient but approximate estimate of it. This causes the coefficients to be updated randomly around the optimum values. This generates extra noise at the output in steady state. This is measured by a parameter called excess MSE and it is proportional to step size $\tilde{\mu}$, and filter length M. Thus, using a longer filter length not only requires higher cost, but also introduces more noise. To obtain a better steady state performance, a smaller value of $\tilde{\mu}$ is required, but results in slower convergence.

4. Finite precision effects

For adaptive filters, the dynamic range of the filter output is determined by the time-varying filter co-efficients, which are unknown at the design stage. Also, the feedback of $e(n)$ makes signal scaling (to avoid overflow) more complicated. The leaky NLMS algorithm can be used to reduce numerical errors accumulated in filter coefficients. This prevents overflow in a finite-precision implementation by providing a compromise between minimizing the MSE and constraining the values of the adaptive filter coefficients. In implementation with C6713 DSP double precision floating point arithmetic is used, which provides sufficient accuracy and hence there is no need to implement leaky NLMS here.

5. Computational complexity and filter order M

The NLMS algorithm requires $2*M+1$ additions and $2*M+1$ multiplications at any iteration n, where M is the tap length or filter order. So, the computation complexity depends on the order of filter and it must be carefully chosen. The order M of the filter is usually a function of the separation of the two sound sources as well as of any offset delay in synchronization between the two (or, equivalently, a function of the echo delay in telephony). In many cases, delays of 10-60 ms lead to fewer than 500 taps (at 8000 samples/sec), and NLMS algorithm is feasible on a single chip [4]. Unless the delay is directly estimated, M must be large enough to account for the maximum is possible delay, which may lead to as many as 1500 taps when the two microphones are separated by a few meters (or even exceeding 4000 taps in cases of acoustic echo cancellation in rooms). Such long filter responses can lead to convergence problems as well as to reverberation in the output speech [4]. The noise (echo) can be minimized by optimizing the step size ($\tilde{\mu}$ in Equation (2.3), which changes the filter coefficients each iteration), at the cost of increased settling time for the filter. For large delays, versions of ANC operating in the frequency domain may be more efficient [1], e.g., sub-band systems [20].

To test the NLMS algorithm for different values of parameters like step size and filter length and its effect on stability and convergence, the SIMULINK model as shown in figure 2.7 is used. Here the input reference is noise source and desired signal is only filtered noise. So in the steady state conditions the error signal must be zero and weights are adjusted so that it exactly adapts to the same filter that has filtered noise. Here the mean square error (MSE) signal and deviation of weight vector is measured.

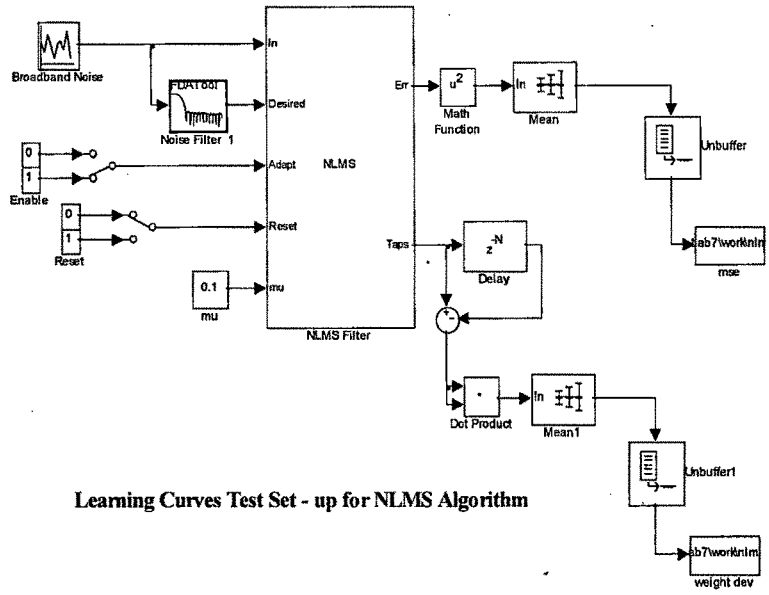


Fig. 2.7 SIMULINK model to obtain learning curves of ANC using NLMS algorithm

The results of the testing are given in figure 2.8 with different values of step size ($\tilde{\mu}$) and filter length (M). In this test, reference signal is Gaussian noise with zero mean and unity variance. The desired signal is given as filtering version of this signal, with 4000Hz Bandwidth. So, the actual speech signal applied is zero signal. In ideal situation the error signal output must be zero at all times, but due to stability and convergence properties of algorithm, it will not achieve ideal performance. The simulation time is set to 1 second and results are recorded. The graph of iterations (time) →mean square error (MSE) and frames (time)→ mean square deviation of 2nd norm of weight vector are obtained by test procedure and plotted in figure 2.8. They are termed as the "Learning Curves". Table 2.1 indicates the numerical values of parameters used for testing the algorithm.

Step Size ($\tilde{\mu}$))	0.001	0.01	0.1	1.0	1.5
Filter Length (M)	16	32	64	128	

Table 2.1 Parameter values for testing NLMS algorithm performance

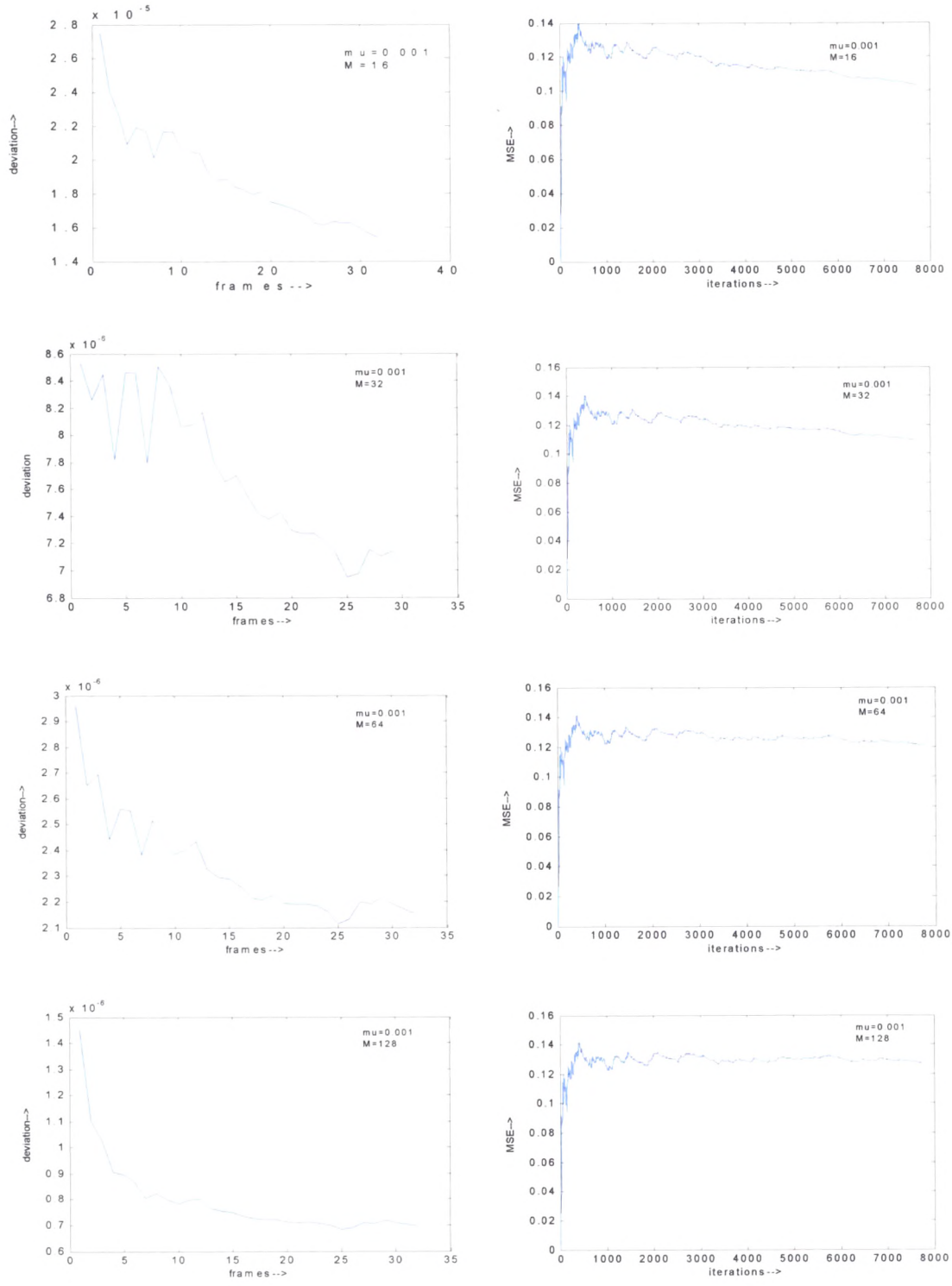


Fig. 2.8(a) Learning curves for NLMS algorithm with $\tilde{\mu} = 0.001$

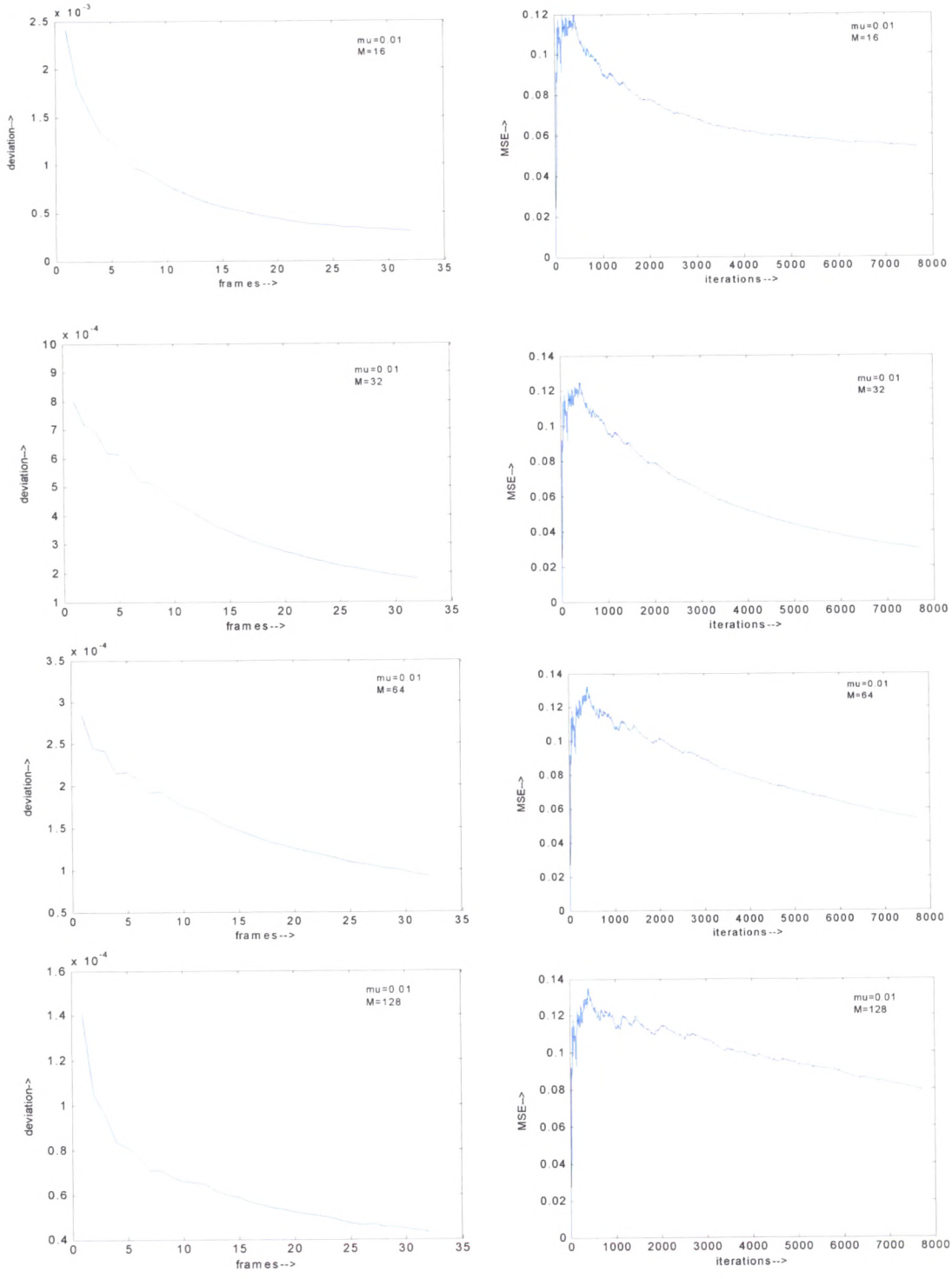


Fig. 2.8(b) Learning curves for NLMS algorithm with $\tilde{\mu} = 0.01$

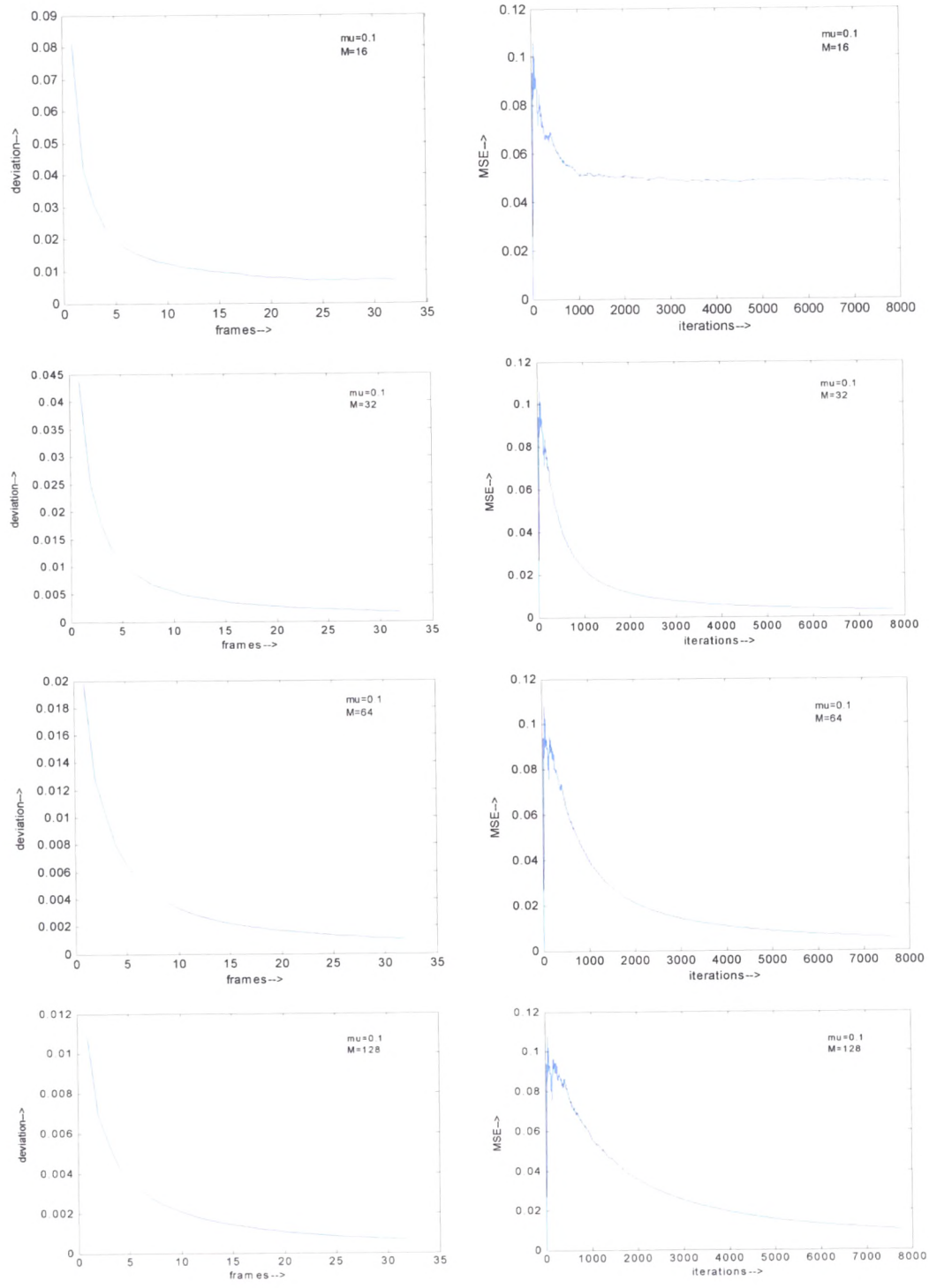


Fig. 2.8(c) Learning curves for NLMS algorithm with $\tilde{\mu} = 0.1$

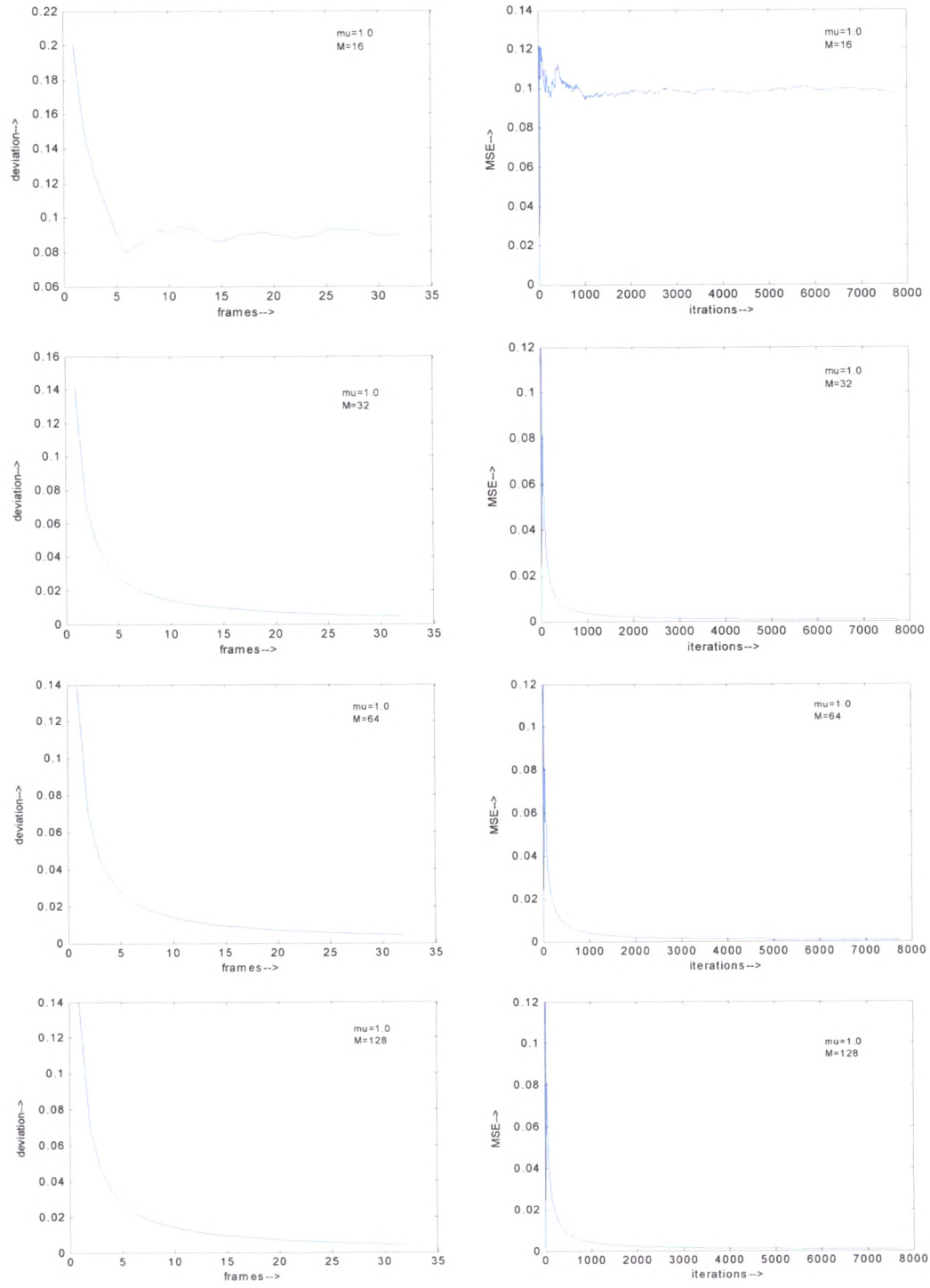


Fig. 2.8(d) Learning curves for NLMS algorithm with $\tilde{\mu} = 1.0$

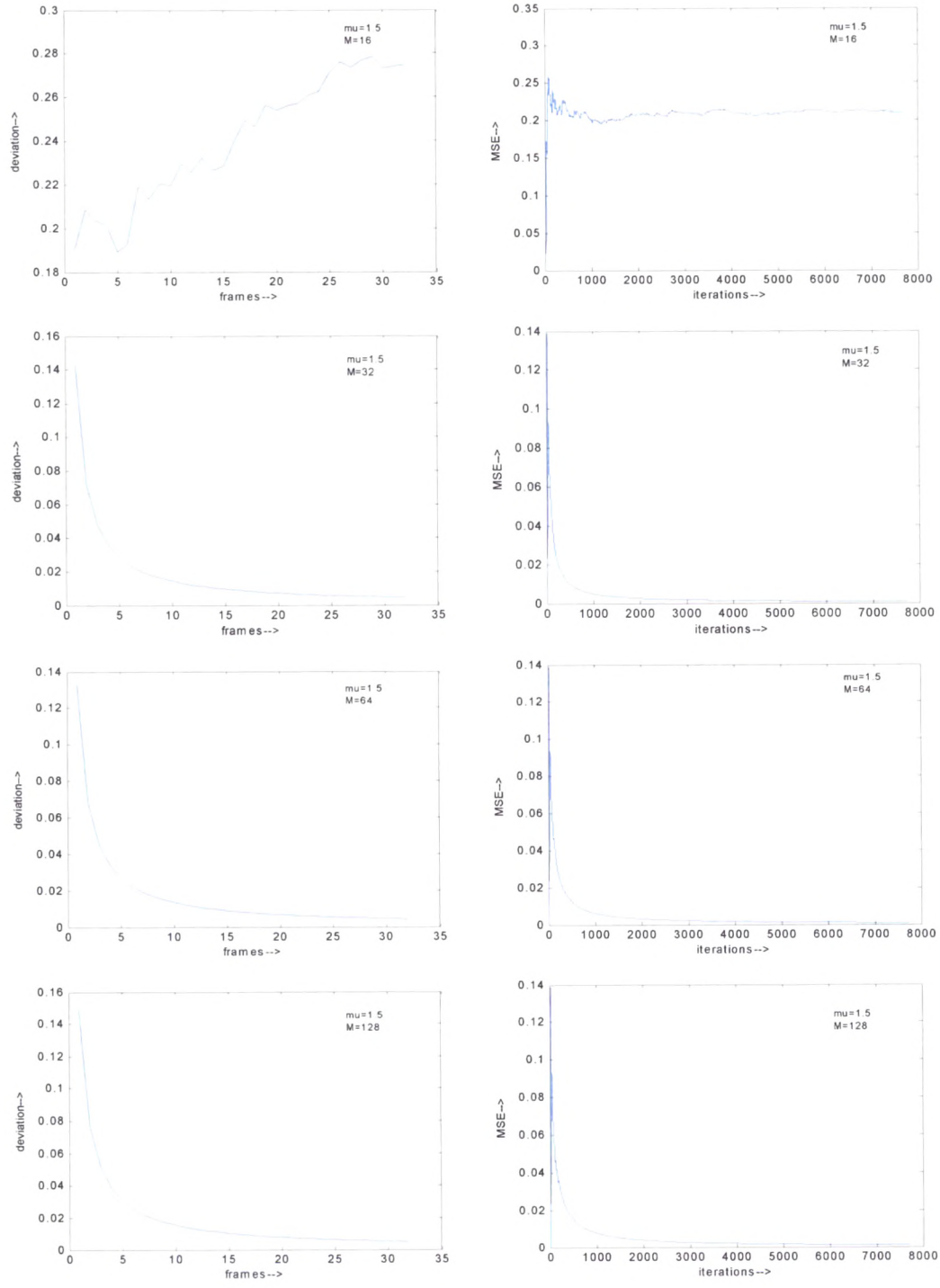


Fig. 2.8(e) Learning curves for NLMS algorithm with $\tilde{\mu} = 1.5$

From these curves it can be concluded that the convergence is faster if larger step size is selected and it is slower if step size is small. But it is not recommended to use very large step size, to account for stability. Also, for larger step size the deviation in weight coefficient vector is large, which introduces its own noise in output (excess MSE). Also, the upper bound on step size is inversely proportional to filter length. So, unnecessary lager filter length must be avoided. The increase in filter length can improve stability but it degrades the steady state performance by introducing excess MSE and more deviation of weight coefficients. Hence for given noise source, the step size of 0.01 to 0.1 and filter length from 32 to 64 is the optimized value. These values are used in all the programs implemented using this algorithm.

The ANC method relies on the microphones being sufficiently apart or on having an acoustic barrier between them. The ANC method is less successful when the secondary signal contains speech components from the primary source, or when there are several or distributed sources; its performance depends on locations of sound sources and microphones, reverberation, and filter length and updating. ANC does best when the microphones are separated enough so that no speech appears in secondary signal, but close enough so that the noise affecting the main signal is also strong in the secondary signal.

2.3.4 Echo Cancellation Using NLMS Algorithm

Echo in a telecommunication system is the delayed and distorted sound which is reflected back to the source. There are two types of echo encountered in telecommunications: acoustic echo, which results from the reflection of sound waves and acoustic coupling between the microphone and loudspeaker, and electrical (line) echo, generated at the two-to-four wire line conversion hybrid transformer due to imperfect impedance matching. Here the model is developed which is equally applicable to both the cases. Here $y(n)$ is a signal with echo or containing both desired speech $x(n)$ from the near end, plus undesired echo $d(n)$ from the far end and $u(n)$ is the near end receive input and $e(n)$ is the output [1, 4].

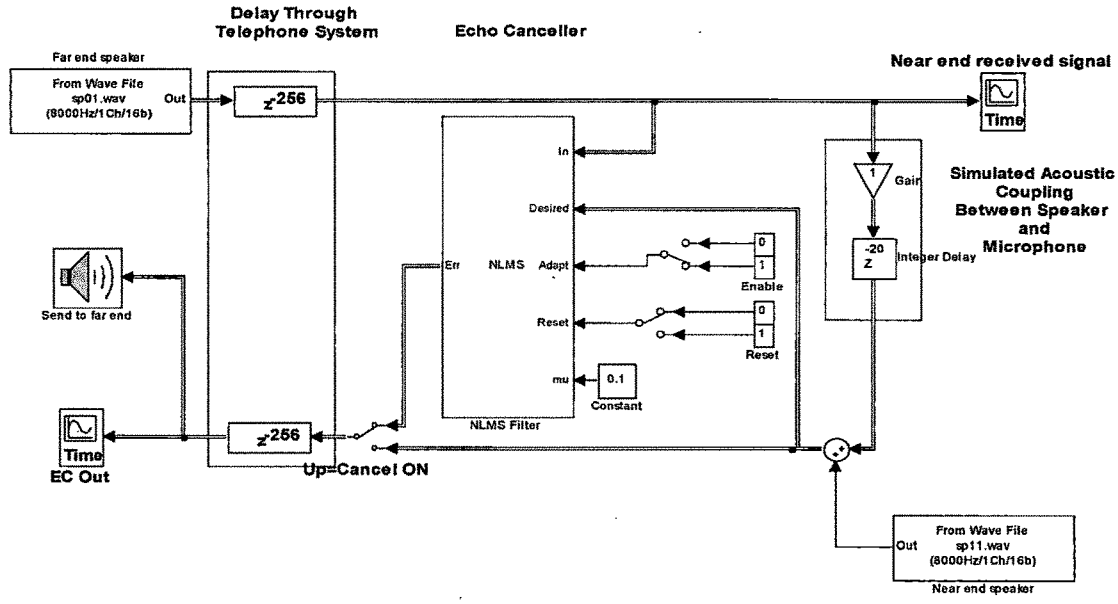


Fig. 2.9 Echo cancellation using NLMS algorithm - SIMULINK model

The SIMULINK implementation of echo cancellation using NLMS algorithm is shown in figure 2.9. The operation and arrangement of various blocks are very much similar to ANC. The same arrangement can be used for room reverberation cancellation. In practice, echo cancellers are applied on both ends to cancel the echoes in each direction.

2.4 Summary

In this chapter an exhaustive survey of various speech enhancement techniques useful for wireless communication systems has been described. Various techniques for all three kinds of major speech enhancement problems that arise in wireless communication are addressed. For noise removal problem it was stated that the DFT based approach is most common but most powerful. It estimates spectral amplitude of clean speech but no attempt is made to estimate the phase of the desired signal; rather the phase of the noisy signal is preserved. Further explanation is given in next chapter. For reverberation cancellation problem the single algorithm is not sufficient for all environments. Multistage algorithms must be used in some combination. Further scope for improvement is seen in RASTA processing. This can be used to handle both noise and reverberation cancellation. The details of RASTA processing are given in chapter 5. The proper investigation in this direction is suggested here. The problem of speaker separation is the most difficult to handle and still the research done is limited in the context of problem solution. The

adaptive algorithms like LMS and NLMS which are popular in adaptive control systems can also be used for speech enhancement. Their applications for additive noise removal and echo cancellation are described. The real time SIMULINK and DSK6713 implementation is also mentioned as a case study. However, the problem with this approach is the requirement of reference signal which can be obtained by placing the second microphone to pick up the background noise reference. This is not possible in every situation and hence the single channel solution is the prime requirement in communication systems.