Spectral-Warping Based Noise-Robust Enhanced Children ASR System

Puneet Bawa
Chitkara University
Virender Kadyan (ervirenderkadyan@gmail.com)
University of Petroleum and Energy Studies https://orcid.org/0000-0001-8708-9738
Vaibhav Kumar
Chitkara University
Ghanshyam Raghuwanshi
Manipal University - Jaipur Campus

Research Article

Keywords: Kalman Filtering, Wiener Filtering, Feature Extraction, Spectral Warping, Children Speech Recognition

Posted Date: December 28th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-976955/v1

Spectral-warping based noise-robust enhanced children ASR system

Puneet Bawa[1], Virender Kadyan[2], Vaibhav Kumar[3], Ghanshyam Raghuwanshi[4]
[1,3] Centre of Excellence for Speech and Multimodal Laboratory, Chitkara University Institute of Engineering & Technology, Chitkara University, Punjab, India.
[2,3]Speech and Language Research Centre, School of Computer Science, University of Petroleum & Energy Studies (UPES), Energy Acres, Bidholi, Dehradun- 248007, Uttarakhand, India.
Department of Computer and Communication Engineering, Manipal University Jaipur, Jaipur
[1] puneet.bawa@chitkara.edu.in, [2] vkadyan@ddn.upes.ac.in, [3] ervaibhavkumar@outlook.com, [4]ghanshyam.raghuwanshi@jaipur.manipal.edu

Spectral-warping based noise-robust enhanced children ASR system

Abstract

In real-life applications, noise originating from different sound sources modifies the characteristics of an input signal which affects the development of an enhanced ASR system. This contamination degrades the quality and comprehension of speech variables while impacting the performance of human-machine communication systems. This paper aims to minimise noise challenges by using a robust feature extraction methodology through introduction of an optimised filtering technique. Initially, the evaluations for enhancing input signals are constructed by using state transformation matrix and minimising a mean square error based upon the linear time variance techniques of Kalman and Adaptive Wiener Filtering. Consequently, Mel-frequency cepstral coefficients (MFCC), Linear Predictive Cepstral Coefficient (LPCC), RelAtive SpecTrAl-Perceptual Linear Prediction (RASTA-PLP) and Gammatone Frequency cepstral coefficient (GFCC) based feature extraction methods have been synthesised with their comparable efficiency in order to derive the adequate characteristics of a signal. It also handle the large-scale training complexities lies among the training and testing dataset. Consequently, the acoustic mismatch and linguistic complexity of large-scale variations lies within small set of speakers have been handle by utilising the Vocal Tract Length Normalization (VTLN) based warping of the test utterances. Furthermore, the spectral warping approach has been used by time reversing the samples inside a frame and passing them into the filter network corresponding to each frame. Finally, the overall Relative Improvement (RI) of 16.13% on 5-way perturbed spectral warped based noise augmented dataset through Wiener Filtering in comparison to other systems respectively.

Keywords: Kalman Filtering, Wiener Filtering, Feature Extraction, Spectral Warping, Children Speech Recognition;

1. Introduction

Speech signals in real-time systems usually determined through stress, tiredness, environmental aggravations, and with incited fluctuations in pronunciation of the speaker. However, ASR frameworks are generally being designed by utilizing the ideal conditions where just clean speech corpus is considered for sufficient training of the speech-based models (Barker et al. 2001). Subsequently, this outcomes act as a simple recognizable proof for the spoken words and at the same time streamlined the conversion of spoken words into deciphered string. It is much needed for building of enhanced human-machine interface system. In general, real-time audio signals are affected by noise which results into disruptive and undesirable information (Kim et al. 1999). Consequently, the major challenge is the optimisation of ASR technology through reduction of disparities among real-time and ideal environmental conditions. Thereby, different strategies of system induction with background noise (noise augmentation) (Ko et al. 2017; Pervaiz et al. 2020); its tuning based on its real-time applications (Gopalakrishna et al. 2012; Lang et al. 1996), and enhancement of noisy signal before its conversion to text (Boashash and Mesbah 2004; Sivasankaran et al. 2015; Zhang et al. 2017), have been adopted by various researchers. Additionally, other experimentations are performed by utilization of numerous noise-cancellation methods which

are included by most commonly Adaptive Filtering methodology (Espy-Wilson et al. 1996) and different enhancements techniques of Neuro-Fuzzy filters (Esposito et al. 2000), Kalman Filtering (Goh et al. 1999). Apart, these methods employed filtering techniques which can be viewed as an aspect of enhanced quality. In like manner, the noise corrupted vocal signing clip is cleaned with tuned Kalman filter (Das et al. 2016). Thereby, Kalman filtering technique is highly preferred by considering the presence of non-linear noise such that instantaneous state in a linear dynamic system is injected by noise at lower SNRs (Sorqvist et al. 1997). On the other hand, adaptive filters are generally derived from Wiener Filtering (Abd El-Fattah et al. 2014) where Least Mean Square Error based algorithm helped in reduction of the impact of linear noise. This technique is able to smoothen the step factor in time domain and also employed Sigmoid function for controlling the direction.

The information present in real-time signal is too cumbersome to deal with the development of acceptable classification, recognition and verifications frameworks (Kim et al. 1999). This part can be accomplished by expulsion of undesirable information before extraction of the significant features in speech recognition and identification systems. In like way, this front-end process of feature extraction helped in transformation of processed speech signal. A compact yet logical representation is more discriminative as well as reliable than the actual signal. Nonetheless, in present ASR frameworks, various feature extraction procedures yielded into a multidimensional feature vectors. They are being utilised for portraying the dependable information of an input speech signal (Mesgarani et al. 2017). Hence, the different scope of options for such parametric representation of signal is performed using LPC (Gupta and Gupta 2016), MFCC (Bassan and Kadyan 2019) for recognition measures. MFCC has been broadly utilized and mainstream front-end method which is used for ASR frameworks. It tried to process the most relevant portion of a signal. In many case, these signal being propagated in noisy or mismatched conditions (Zhao and Wang 2013). Apart, PLP (Hermansky 1990) has been introduced as a way of distorting a spectra with a goal of minimizing the inter-speaker variations lies due to acoustically mismatched conditions. However, RASTA-PLP (Hermansky et al. 1991) strategy utilised a band-pass filter to energy component present in each frequency sub-band. It has smoothen the short-term noise-based variations alongside it handled the inter-speaker variations. Subsequently, many advanced noise robust feature extraction methodologies - zero crossing peak amplitude (Scarr 1968), average localized synchrony detection (Ali et al. 2004), BFCC, GPLP (Gulzar et al. 2014) and gammatone frequency cepstral coefficient (GFCC) (Zhao and Wang 2013) have been experimented. Recently, the utilization of Gammatone filters for accurate modelling of the critical bands is performed. Rather than utilising the triangular filters it out-performed conventional strategies of feature extraction in the field of recognition.

Nowadays, individuals generally feel more comfortable in recognizing the words being articulated in their respective native languages as compared to obscure foreign languages. Since, the improvement of ASR system in local language is totally reliant upon the adequate availability of the labelled data and phonetic transcriptions. In this manner, the majority of resource rich dialects based spoken dialogue frameworks are commercially accessible whereas a very few concentrations have been made towards the context for usage of native languages including Punjabi, Mizo, Bodo (Singh et al. 2019; Kaur et al. 2020). These under-resourced languages lack the web presence, availability of linguistic expertise and mainly the lack of resources which required text corpora and pronunciation rich lexicon (Besacier et al.

2014). Therefore, to overcome the challenge of data scarcity, various aspects of limited data for both acoustic & language models (Novotney et al. 2009), multi-lingual knowledge transfer (Ma et al. 2017) and construction of adequate pronunciation lexicon (Robinson et al. 1995) have been experimented. Another challenge is sufficient advancement of children ASR system where intelligent speech innovations: YouTube Kids, Amazon Alexa, and computer-aided language learning has been currently crucial in the process of classroom learning (Valente et al. 2012). Since, the acoustic and linguistic patterns in case of children speech signals are very unique which indulge speaking rate, vocal tract length when contrasted to an adult speech signal (Subramanian et al. 2019). Additionally, the accessibility of limited children speech datasets even in the context of native language prompts obstruction in development of efficient children speech recognition systems. Moreover, different procedures of data augmentation (Ko et al. 2015), have been utilized by researchers with a goal for inciting the artificial data. It tried to build it essential for performance improvement of data hungry deep learning approaches.

In this paper, two filtering techniques: Kalman and Wiener Filtering have been employed with an effort of reducing unwanted information in an input children speech signal. Since, limited resource Punjabi children ASR system is constructed on own developed speech corpus. Therefore, robustness to ASR system is provided through combination of original speech corpus with synthetic dataset which contains more ideal SNR ratio (Koopmans et al. 2018). Motivated by this, following efforts have been made for improving the performance accuracy of children speech recognition framework through:

- Classic approach of in-domain noise augmentation has been applied by inducing four distinctive type of noises – self-recorded classroom, cafeteria, white and pink noise at varying SNRs while keeping the class labels fixed.
- Comparative analysis between filtering procedures of Kalman and Wiener has been made by utilizing the feature extraction technique of MFCC, LPCC, RASTA-PLP and GFCC. So, an enhanced signal being generated using both the filters are mixed together with original dataset such that de-noising is operable for audio being recorded in both clean and noisy conditions.
- Indulging tonal characteristics using normalised VTLN methodology for filtered signal with an end goal of eliminating the existing inter-speaker variations.

The rest of the paper is structured as follows: Section 2 includes a literature analysis for building noise-robust ASR system. In addition, the theoretical context for filtering and feature extraction techniques is portrayed in Section 3. Section 4 and Section 5 give descriptions of the experimental configurations and the proposed system architecture. In section 6, discussions on the efficiency of different systems in varying environment conditions is outlined with conclusion made in Section 7.

2. Related Work

Gong et al. (1995) analysed the effects of noise in automated speech recognition systems. They revealed integral part of time and frequency associations in recognition systems. Further exploited task-specific a priori awareness of speech and noise are presented which showed the significance of high SNR values. Earlier, Lim and Oppenheim (1979) explored speech degradation by additional background noise and analysed various techniques proposed for enhancing speech and bandwidth compression. The experimentations resulted into

adequate compression while it retained the required information of original audio signal. Further, Boll (1979) calculated spectral noise bias during non-speech activity and supressed the stationary signal by subtracting the calculated spectral noise from it. Further it applied secondary procedures to attenuate the residual noise left after subtraction. The researchers performed perception test with DRT database of 192 words. It indulge noise and found comparable results on intelligibility and quality of signal. With advancement in time, Ephraim and Malah (1984) capitalized on importance of short-time spectral amplitude (STSA). For construction of an enhanced signal, the minimum mean square error (MMSE) STSA estimator are combined with complex exponential noisy phase. They performed similar investigation of MMSE STSA and Wiener STSA and found that MMSE STSA resulted into fundamentally less blunder and inclination when SNR is low. They also tried to improve speech signal for revising vocal tract resonance disorders. The parameters are thought of in such a way it adapt F1 and F2 formant frequencies of an input speech signal (Goncharoff et al. (1988)). However, Etter and Moschytz (1994) underlined the idea of noise-adaptive spectral magnitude expansion with an effort of adapting the crossover point of the spectral magnitude. The expansion are performed in each frequency channel which is based on the noise level. Nowadays, researchers have gone beyond just enhancing the audio signal. Various studies have been conducted on impact analysis of enhancement of speech recognition systems. (Umesh et al. (1996)) Frequency warping function has also been proposed which is derived from scale-transform based acoustic features to effectively separate vowels. The results showed clear distinction between different formant frequencies scale which lies differently among speakers. Also, frequency warping is explored in the field of automatic speech recognition by sampling it with warping function. An audio signal with high energy regions remained sampled with more than low density regions because it is believed that high energy regions carries more linguistic information (Paliwal et al. (2009)). Likewise, Sameti et al. (1998) intentionally corrupted the signal with white, simulated helicopter, and multi-talker (cocktail party) noise. The HMM based MMSE speech enhancement system had been consistently superior in performance to the spectral subtraction-based system in handling noise in non-stationarity. In (Saldanha 2016), Harmonic Regeneration Noise Reduction (HRNR) and adaptive Wiener filter with Two Step Noise Reduction (TSNR) methods are used to enhance the noisy speech signal. They also augmented the speech signal with fan noise and processed it with adaptive wiener filtering. The output is plotted on MATLAB which showed improvement in SNR of an audio signal. Lee et al. (2014) proposed a phase-dependent priori signal-to-noise ratio (SNR) estimator in log-mel spectral domain. It utilized both size and phase information, where the decision-directed (DD) approach is used to determine a priori SNR from a noisy speech. Lately, Haque and Bhattacharyya (2019) investigated a portion of it by separating procedures which are depend on direct and nonlinear methodologies. These procedures incorporate diverse versatile by separating it from dependent calculation like LMS, NLMS and RLS.

However, Gurugubelli and Vuppala (2019) proposed another component persuaded from the human hear-able discernment and high time-recurrence. As a piece of SFF procedure execution, audio signals are gone through a single pole. It is complex bandpass filter bank which tried to get high-goal time-recurrence circulation. Then, at the same time, the circulation is upgraded by utilizing a bunch of hear-able perceptual administrators. Similarly, Narayana and Kopparapu (2009) have experimented a study on the effect of additive Gaussian noise with improvement in performance of commonly used MFCC feature

extraction technique. They experimented an estimation error while tuning the parameters of MFCC during Gaussian noise. However, the vast majority of work conducted for building noise robust ASR framework utilised linear predictive coding (LPC) for speech signal modelling. However, Nair et al (2016) have experimented the shortcomings of Kalman filters with LPC and concluded the superiority of MFCC over LPC. They outlined the dependence of refinement of parameters on the choice of R and Q parameters. It resulted into easier modulation of MFCC parameters for smaller amount of noise. Further, Zhao and Wang (2013) analysed the boosted performance of novel speaker feature extraction technique. They indicated that non-linear corrections account mainly for variations in noise and have been adequately handled by implementation of a different time-frequency representation. In addition, they also experimented the boosted robustness of MFCC features in presence of noise. Consequently, Zhao et al. (2011) experimented on Gammatone filter based feature extraction method which can be extended in audio security system. The experiment concluded with an efficient and fair extraction of feature vectors. It resulted into satisfactory classification performance using SVM. Furthermore, Sárosi et al. (2011) carried out experiments for comparative study of novel front-end techniques in six languages: English, Italian, German, Spanish, French and Hungarian. They concluded with the presence of a substantial difference in implementations of MFCC and significant improvements is obtained in PNCC variants. It lies with separate bandwidths and differentiated SNR levels. Kadyan et al. (2021) also investigated adult-child mismatch using Punjabi corpus while formulating ASR system and used vocal tract length normalization to showcase with better output.

3. Theoretical Background

3.1 Filtering Techniques

Adequate noise reduction in input speech relies on the output of linear time-varying filter. It is being induced by intermittent pulses or presence of noise. However, closer observation with noise-reduction methodologies revealed speech modelling as a n^{th} order auto-regressive process. Thus, the present sample $x(k,t)$ is explicitly reliant on linear combination of previous $x(k-1,t)$ sample with random noise at varying SNRs. In other words, the representation is an all-pole FIR filter with input as an Additive White Gaussian Noise given by equation(1):

$$x(k) = -\sum_{i=1}^{p} a_i x(k-1) + u(k) \quad (1)$$

where $u(k)$ corresponds to a zero-mean Gaussian noise (process noise) and a_i refers to the linear prediction coefficients (LPCs) evaluated using auto-corelation function (ACF) as in equation(2):

$$A = R^{-1} * r \ (2)$$

3.1.1 Kalman Filter

Kalman filtering is utilized by a progression of estimations which is saw over the long haul, containing commotion (irregular varieties) and different mistakes. It delivered evaluations of obscure factors which is more exact than those dependent on a solitary estimation alone. The application of Kalman filter to autoregressive models are detailed in equation (3) and is first

performed by (Paliwal and Basu 1987). They represented it through state vector representation as state-space model. It is used as a dependent on the state transition matrix. It's coefficients is calculated from additive noisy signal. In addition, the internal use of state-space model makes Kalman Filtering able to handle dynamic models with varying parameters. The application of Kalman filter is employed with G as $n * 1$ matrix input along with corrupted noise $g(k)$ at the given kth instance by:

$$Y(k,t) = Y(k-1,t) + G * g(k) \quad (3)$$

where $Y(k,t)$ corresponds to $(n * 1)$ state matrix with $(n * n)$ state transition matrix. The LPC for the noisy signal is computed through equation (2).

3.1.2 Wiener filter

The Wiener filter is a common technique of filtering. It is employed for noisy signal and is used in many signal enhancement procedures. It is also used to measure an approximation of desired signal by performing linear time-invariant filtering of an observed noisy signal. This resulted into minimisation of mean square error between assessed estimated random process and the target process. The Wiener filter is also utilized to filter out the noise from the corrupted signal. It is used to provide an estimate of the underlying signal of interest. The frequency domain solution to this optimization problem gives the filter function as illustrated in below equation(4):

$$H(\omega) = \frac{P_s(\omega)}{P_s(\omega) + P_v(\omega)} \quad (4)$$

where $P_s(\omega)$ and $v(\omega)$ are the power spectral densities of a clean and noise signals, respectively with an assumption of uncorrelation between both the signals. Thus, signal-to-noise ratio (SNR) can be computed as in equation(5):

$$SNR = \frac{P_s(\omega)}{P_v(\omega)} \quad (5)$$

Finally, wiener filter equation can be interpreted as in equation(6):

$$H(\omega) = \left[1 + \frac{1}{SNR}\right]^{-1} \quad (6)$$

3.2 Feature Extraction

The feature vectors corresponding to an input speech signal plays a vital role in extraction of unique information. It is possible through segregation of a speaker from others by reduction of magnitude of a signal. It is devoid by causing any damage to the power of speech signal. As a result, the processing of features in degraded environmental conditions largely influenced the performance of an ASR framework. Two most widely used methods have been used to evaluate their effect in both the noisy and clean environments. Thousands of coefficients for specific signals are extracted. Apart, only hundreds of randomly selected signals are used as input features for further study (Kadyan et al. 2017). However, the techniques differ in pre-processing phases, pre-emphasis where usually a signal is passed into first-order finite impulse response (FIR) filter. This process is succeeded by a method of

partitioning speech signal into frames. It is beneficial in removal of acoustic interface existing at both starting and ending parts of an input speech signal.

3.2.1 Mel-Frequency Cepstral Coefficients (MFCC)

MFCC is a representation of short-term power spectrum. It is defined as a real cepstrum of a windowed short-time signal. The derivation for the same is performed using fast Fourier transform of a speech signal. MFCC makes the use of non-linear frequency scale. It is possible through approximation of behaviour of an auditory system (Davis and Mermelstein 1980). Further, the modification in the magnitude spectrum is demonstrated to Mel spectrum. It involved the broad variety of frequencies in FFT spectrum. Thus, the pitch value corresponding to each tone with frequency $f(k,t)$ is measured in $Hertz(Hz)$ is represented on Mel Scale as in equation(7):

$$Mel\big(f(k,t)\big) = 2595 * log_{10}\left(1 + \frac{f(k,t)}{700}\right) \quad (7)$$

3.2.2 Gammatone frequency cepstral coefficients (GFCC)

GFCC is based on an auditory model of peripheral. Although a gamma filter bank is used in it which decomposes the input voice to a temporal frequency representation by a auditory model. The GFCC is calculated using a bank of gammatone filters (Zhao and Wang 2013). Consequently, down-sampling of filter bank are responded along with the time dimension. It is possible by decomposing the input speech signal into T-F (Time Frequency) domain. In addition, an equivalent rectangular bandwidth (ERB) represented the bandwidth corresponding to a resulting filter as in equation(8):

$$b_m = b * ERB(f_{cm}) \quad (8)$$

where f_{cm} related to the central frequency where corresponded to m^{th} Gammatone filter. The magnitudes of the down-sampled responses are loudness-compressed which are employed using a cubic root operation as in equation(9):

$$G_m[i] = \left|g_{downsampled}[i,m]\right|^{\frac{1}{3}};\ i = 0 \ldots N-1, m = 0 \ldots M-1 \quad (9)$$

where N refers to the numbers of filters and M represents the number of time frames obtained after the down sampling operation.

3.2.3 Linear Predictive Cepstral Coefficient (LPCC)

Since spectral characteristics are generated directly from spectra, they essentially represent phonetic information. The contribution of all frequency components of a voice signal is equally emphasised by employing the LPCC features which are generated from spectra. It tried to employ the energy values of linearly organised filter banks. Cepstrum can be extracted from a voice stream using linear prediction analysis. The essential premise of linear predictive analysis is calculated by n^{th} speech sample. It may be predicted using a linear combination of the preceding j samples, as detailed in the below equation(10):

$$s(n,t) = \sum_{i=1}^{j} a_j * s(n-j) \quad (10)$$

Over such a speech analysis frames, $a_1, a_2, a_3 \ldots \ldots a_j$ are presumed to be constants. Thereby, the speech samples are predicted using these coefficients where an error is the discrepancy between real and anticipated speech samples which are evaluated by equation(11):

$$e(n,t) = s(n,t) - \sum_{i=1}^{j} a_j * s(n-j) \quad (11)$$

3.2.4 Relative spectral-perceptual linear prediction (RASTA-PLP)

In original PLP technique, a specific band-pass filter is applied to each frequency sub-band to smooth out short-term noise fluctuations and eliminate any consistent offset in the voice channel. The most important processes in RASTA-PLP are calculating the critical-band power spectrum as in PLP, transforming spectral amplitude through a compressing static nonlinear transformation, filtering the time trajectory of each transformed spectral component by the band pass filter using the equation (12) given below:

$$H(p,t) = (0.1) * \left(\frac{2 + p^{-1} - p^{-3} - 2p^{-4}}{p^{-4}(1 - 0.98 * p^{-1})} \right) \quad (12)$$

3.3 Spectral Warping

Spectral warping is a change of the time domain by a signal which effectively distorts the frequency content of the original signal. The matrix of transformation for such kind of augmentation is broken down into three stages. The first one is DFT which tried to convert time signal into frequency field. The second step is basically a matrix of interpolation which helped in getting the desired new frequency samples. The frequency warping is efficiently provided to determine the signal content. The spectral warping procedure corresponds to the z-transformation of the input signal at the uniform sample points (N_s) of the unit circle and the inverse DFT of the output as a part of final step. The spectral warp is achieved by treating non-uniform z-transform samples as evenly as possible, by applying the opposite Fourier transformation represented in matrix $g[n,t]$ as in equation(13).

$$g[n,t] = \frac{1}{N_s} \sum_{i=0}^{N_s-1} G[i] W^{-in} = \frac{1}{N_s} \sum_{i=0}^{N_s-1} F(e^{j\omega_i}) W^{-in}, where \ W_i^k = e^{-\frac{i2\pi k}{N_s}} \quad (13)$$

4. Experimental Setup

4.1 Original Dataset

For training and testing of our proposed system, the speech data was recorded by utilizing mono channel in clean acoustic conditions. It employed a sampling frequency of 16 kHz. The speech data incorporated 2159 utterances with total number of 20 male speakers and 19 female speakers with absolute duration of speech spanning to 4.15 hours. The data was orchestrated into training and testing sets as shown in Table 1 to such an extent that the convincing presentation for developing the noise-robust Punjabi children ASR system is accomplished.

Table 1. Speech Data employed for training and testing set

Data Characteristics	Training Set	Testing Set
No of Utterances	1575	584
Gender	13 Female and 15 Male	6 Female and 5 Male
Recording Environment	Open and Closed Environment	Open and Closed Environment
No of Hours	3hrs 24 minutes	51 minutes

4.2 Noisy Dataset

The scalable and adaptable noisy database using clean dataset is synthetically created while setting the ideal length and sampling frequency. It indulged the presence of significant information in an input speech signal. Similarly, the noise clips including self-recorded classroom, self-recorded cafeteria, white Gaussian and pink were chosen and injected into clean speech corpus. These clips were cautiously hand-picked ensuring the quality of the recordings and can generally be scaled for accommodating new noise types and desired SNR levels going from $15dB\ to\ 5dB$ with a step size of $-5dB$. For test dataset, the noise clips were embedded as irregular SNRs value which is similar to training set. However same category of noise as in training set was employed for additional experimentations.

5. System Overview

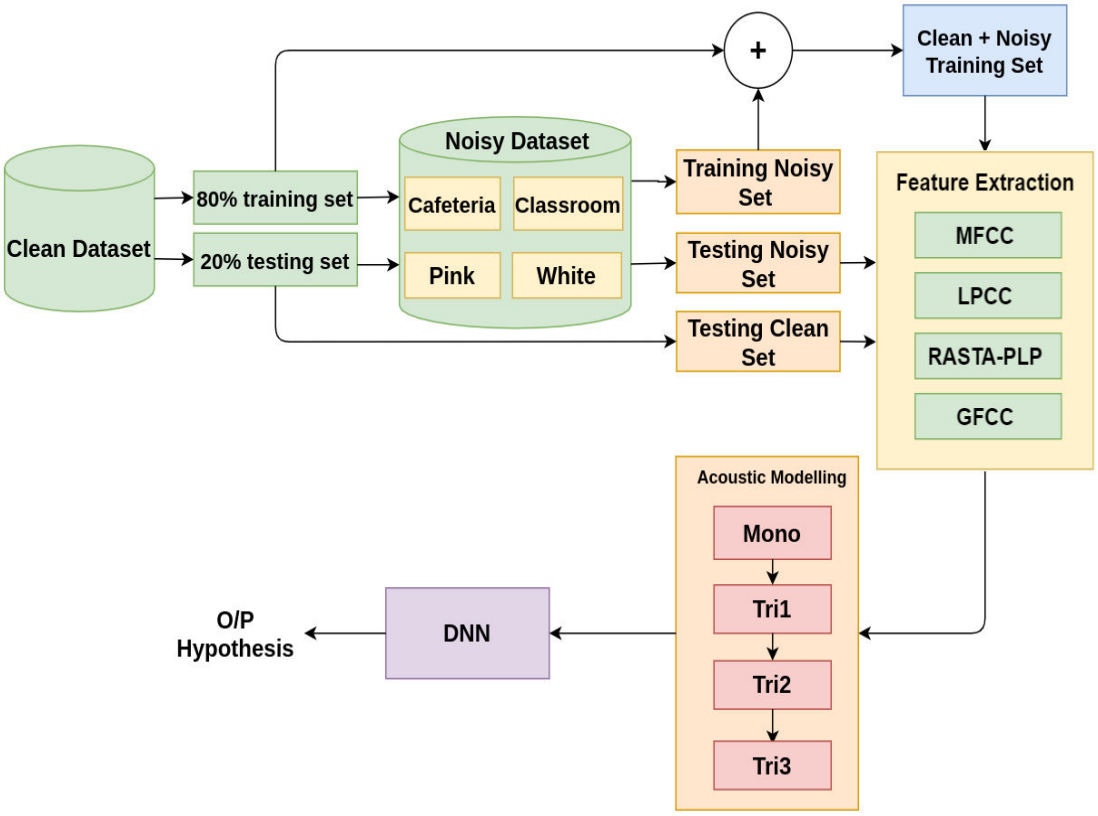


Figure 1. Block diagram summarising the steps involved for baseline system through involvement of MFCC and GFCC feature extraction techniques.

The noise augmented dataset has been demonstrated in the block diagram as detailed in Figure 1. Initially, the system is presented (a) clean input speech signal and (b) noisy signal

which conferred mismatched conditions between training and testing set. The two datasets are trained and tested by extraction of viable features. It is possible through four feature extraction techniques- MFCC, LPCC, RASTA-PLP and GFCC. For MFCC depiction, 13 coefficients (consisting of 12 finalized coefficients where the 13th coefficient is the energy parameter corresponding to each frame) are extracted for the frame length of 25ms and frame shift of 10ms based on equation(7). Consequently, LPCCs are utilised in this context to collect emotion-specific information expressed via vocal tract characteristics. The voice signal is subjected to a 10th order LP analysis in order to get 13 LPCCs for each speech frame of 25ms and with a 10ms frame shift. Furthermore, 12 lower order coefficients corresponding to the noise robust feature approach of GFCC are extracted over hamming window with 25ms frame length and 10ms frame shift based on the filter-bank utilised as in equation(8).

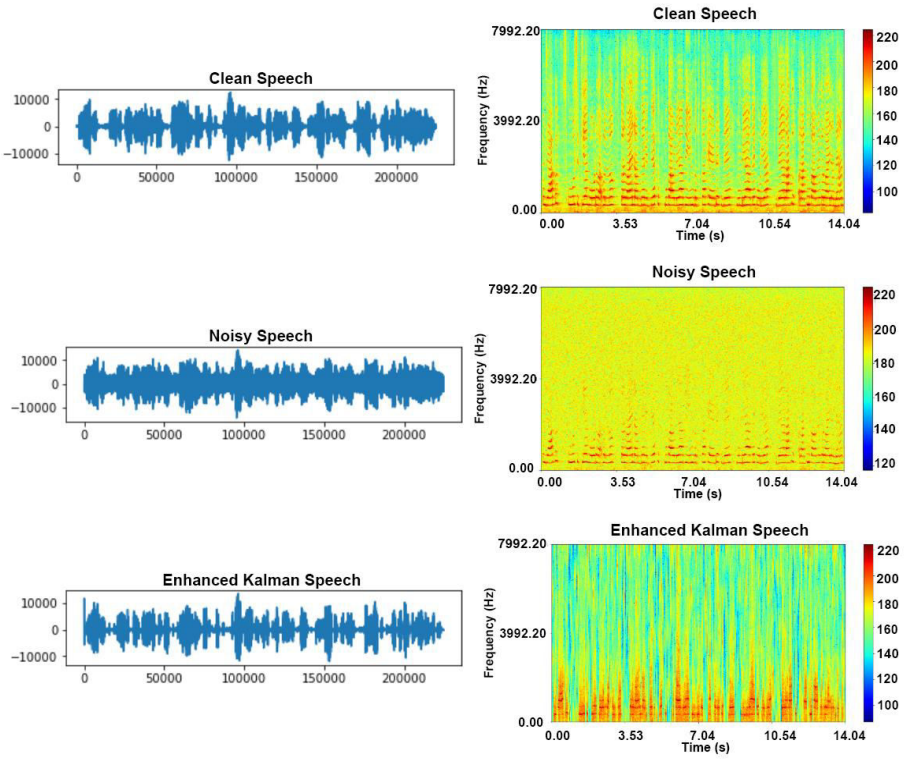


Figure 2(a). The plots illustrating the significance of clean, noisy signals along with Kalman filter based speech signal.

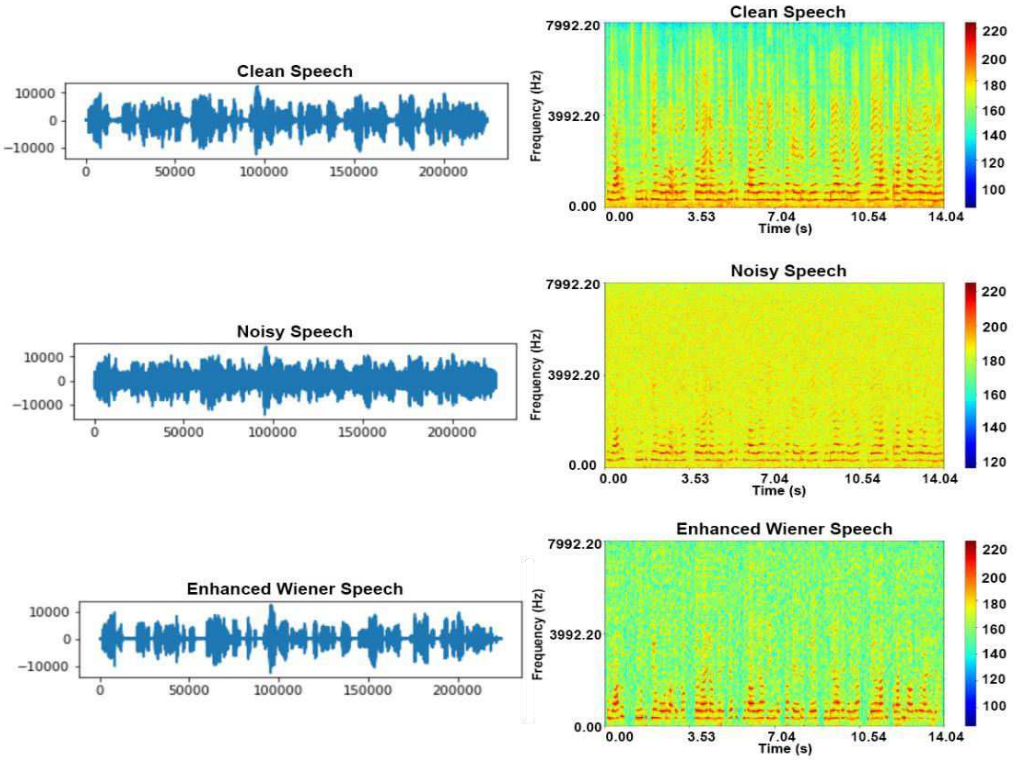


Figure 2(b). The plots illustrating the significance of clean, noisy signals along with Wiener filter based speech signal.

Further, the developed baseline system is decoded against the noisy test set for the real-time evaluation and the accuracy of the system is evaluated at varying SNRs. In this way, the mechanism for the application of two filtering techniques including Kalman Filtering in Algorithm 1 and Wiener Filtering is applied on the testing dataset as described in Figure 2(a) and Figure 2(b) respectively. For real-time performance evaluation of ASR system, both MFCC and GFCC front-end feature extraction techniques are employed. Furthermore, the normalisation process with CMVN is applied on extracted vector and later dividing it by the standard deviation by:

$$y(t,i) = \frac{y(t,i) - \mu(t,i)}{\sigma(y(t,i)} \quad (10)$$

Algorithm 1: Speech enhancement using Kalman Filtering Technique

Step 1: Initialise the frame size as 30ms and 10ms as window length

Step 2: Propagation Step

Step 2.1: Predict the next state

$$\boldsymbol{d_k = A * d_{k-1} + B * u_k}$$

where $\boldsymbol{d_k}$ represents desired signal at time $\boldsymbol{k}$, $\boldsymbol{A}$ is the state-transition model, $\boldsymbol{B}$ is the control input model, $\boldsymbol{u_k}$ is the original signal at time $\boldsymbol{k}$.

Step 2.2: Predict the error covariance ahead

$$\boldsymbol{P_k = A * P_{k-1} * A^T + Q}$$

where $\boldsymbol{P_k}$ is the error covariance matrix and $\boldsymbol{Q}$ being covariance of original signal.

Step 3: Measurement Update (Correction)

Step 3.1: Compute the Kalman gain

$$\boldsymbol{K_k = P_k * H^T * (H * P_k * H^T + R)^{-1}}$$

where $\boldsymbol{K_k}$ is the Kalman gain, $\boldsymbol{H}$ is the observation model and $\boldsymbol{R}$ is the covariance of noisy signal.

Step 3.2: Update the projected state via.

$$d_k = d_k + K_k * (x_k - H * d_k)$$

where x_k is prior state estimate.

Step 3.3: Update the predicted error covariance

$$P_k = (1 - K_k * H) * P_k$$

Step 4: Reiterate the process, using outputs k as input for $k + 1$ with d_k holding the predicted value, the desired result of the Kalman filter.

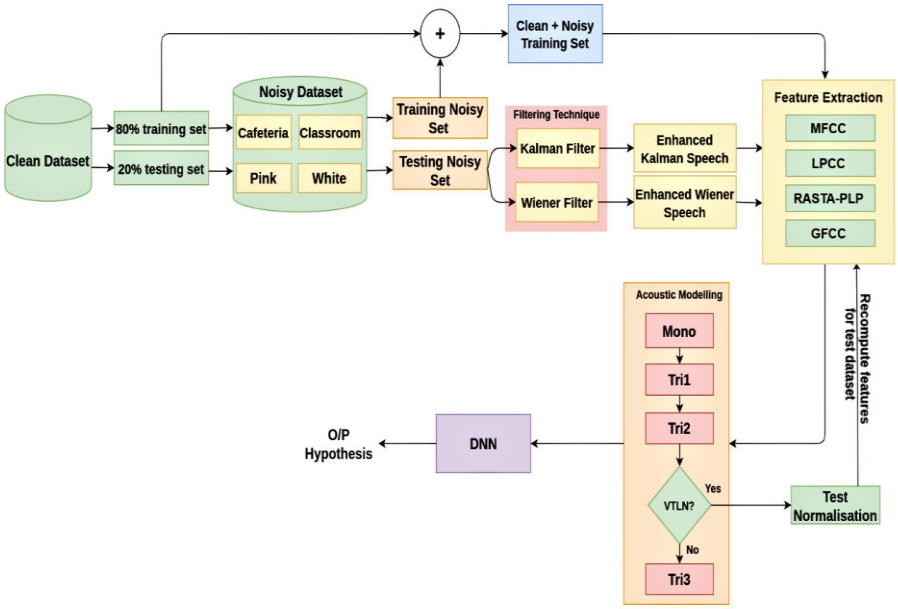


Figure 3. Block diagram summarising the steps involved for external integration of filtering and feature extraction techniques with vocal tract length normalisation approach

In comparison, acoustic models utilised both the linguistic knowledge of the original and the augmented dataset which are further trained on context modelling techniques of – monophonic training (mono), delta (tri1) and delta-delta (tri2). It is based upon triphones training. Consequently, the existing speaker variations are reduced by embedding VTLN warping function after evaluation of delta-delta (tri2). This process of normalisation used piecewise linear function (Zhang et al. 2004). It tried to help in mapping of the corresponding frequency on a large scale after computation of central segment. In addition, the test normalised VTLN is employed such that only test dataset based enhanced Kalman dataset and enhanced wiener dataset is normalised. It is relied on the best warping factor which is evaluated in the form of transformation matrix. The proposed system of aforementioned methodologies is detailed in Algorithm 2 such that the features are re-computed for the test datasets after triphone (tri2) modelling. Furthermore, the parameters needs to be reduced with an objective for boosting the feasibility of the system. It helped in efficient predictive analysis of word sequence. Similar to tri2, LDA(tri3) based triphone modelling helped in reducing the triphones into smaller amount of acoustically distinct units. Thus, this process employed LDA which act as a part of original feature space. It helped in presentation of information by reduction of dimensions from 117 to 40. It is possible through diagonal application of MLLT on the lower-dimension feature vectors. Finally, the systems for direct processing of (a) raw input speech signal (b) noisy signal (c) enhanced filtered signal is being trained on DNN-HMM based hybrid acoustic modelling. It utilised tangent hyperbolic activation function of Kaldi Toolkit (Povey et al. 2011). Finally, the efficiency of the enhanced ASR system which utilised the filtering technique is determined using two main parameters – WER and RI.

Algorithm 2: Step by step process for data augmentation through spectral warping using Wiener Filter on GFCC technique.

Step 1: Initialise child dataset as $\boldsymbol{child_{data}}$

Step 2: Split the $\boldsymbol{child_{data}}$ into 80%-20% ratio with initialisation of training dataset as:

$$\boldsymbol{training_{set}} = (\mathbf{80\%})\ \boldsymbol{of\ child_{data}}$$

,and testing dataset as:

$$\boldsymbol{testing_{set}} = (\mathbf{20\%})\ \boldsymbol{of\ child_{data}}$$

Step 3: Perform noise augmentation

Step 3.1: Initialise SNR values as

$$\boldsymbol{snrTrain_{val}} = [-\mathbf{5}:\mathbf{15}] \text{ and}$$

$$\boldsymbol{snrTest_{val}} = [\mathbf{5}:\mathbf{15}]$$

Step 3.2: Inject data into $\boldsymbol{training_{set}}$ as:

$$\boldsymbol{noisyTraining_{set}} = \boldsymbol{randn}(\boldsymbol{snrTrain_{val}}) * \boldsymbol{randn}(\boldsymbol{len}(\boldsymbol{training_{set}})$$

Step 3.3: Inject data into $\boldsymbol{testing_{set}}$ as:

$$\boldsymbol{noisyTesting_{set}} = \boldsymbol{randn}(\boldsymbol{snrTest_{val}}) * \boldsymbol{randn}(\boldsymbol{len}(\boldsymbol{testing_{set}})$$

Step 4: Pool clean $\boldsymbol{training_{set}}$ and $\boldsymbol{noisyTraining_{set}}$ to perform train noise augmentation as:

$$\boldsymbol{finalTraining_{set}} = \boldsymbol{training_{set}} + \boldsymbol{noisyTraining_{set}}$$

Step 5: Enhance the $\boldsymbol{testing_{set}}$ using Wiener filtering technique

Step 5.1: Apply the filter function as:

$$\boldsymbol{finalTesting_{set}} = \boldsymbol{wiener}(\boldsymbol{noisyTesting_{set}})$$ *//using equation* $(\mathbf{4})$

Step 5.2: Apply the equation $(\mathbf{6})$ on each set of the audio

Step 6: Select the phone based 3-gram language model, $\boldsymbol{n = 3}$ used for decoding process

Step 7: Augment $\boldsymbol{finalTraining_{set}}$ by employing spectral warping

Step 7.1: Initialize the process of Spectral Warping as

$$\boldsymbol{warp_{fac}} = [-\mathbf{0.15}, -\mathbf{0.10}, -\mathbf{0.05}, \mathbf{0.05}, \mathbf{0.10}, \mathbf{0.15}]$$

Step 7.2: Apply the warping factor on $\boldsymbol{finalTraining_{set}}$ as:

$$\boldsymbol{finalTraining_{warp_{set}}} = \boldsymbol{spec\_warp}(\boldsymbol{finalTraining_{set}}[\boldsymbol{warp\_fac}\ ])$$//using equation $(\mathbf{13})$

Step 8: Extract GFCC features as:

$$\boldsymbol{gfcc}(\boldsymbol{finalTraining_{warp_{set}}}), \boldsymbol{gfcc}(\boldsymbol{finalTesting_{set}})$$ *//using equation* $(\mathbf{8})$ *and* $(\mathbf{9})$

Step 9: Perform CMVN training on the top of extracted GFCC features:

$$\boldsymbol{cmvn}(\boldsymbol{finalTraining_{warp_{set}}}), \boldsymbol{cmvn}(\boldsymbol{finalTesting_{Set}})$$ *//using equation*$(\mathbf{10})$

Step 10: Perform mono-phone training (mono), tri-phone training (tri1, tri2, tri3) and align them for further computational training

Step 11: Reduce inter-speaker variations

Step 11.1: For alignment of tri2, compute vocal tract length normalisation on each test dataset samples by selection of optimal warp value lies between $\mathbf{0.8}$ and $\mathbf{1.2}$

Step 11.2: Go back to **Step 7** and recompute the features for normalised test data

Step 12: Train the model using DNN based hybrid architecture.

Step 13: Obtain the best result on DNN, otherwise go to **Step 7** for finding the best warp factor.

---

6. Results and Discussions

6.1 Performance evaluation of ASR system on both clean and noisy test conditions

For the first set of experiments, the baseline system is being experimented using four front-end feature extraction techniques- MFCC, LPCC, RASTA-PLP and GFCC on clean child audio signals in both training and testing dataset. The MFCC feature extraction technique achieved 15.43% WER which performed better than LPCC with 16.02%, RASTA-PLP with 15.46% and GFCC with 15.61% under clean conditions as depicted in Table 2. However, in real-world conditions, the test data is a mixture of clean and distorted audio signals with distinct SNRs and different forms of noise. So, the test dataset has been expanded with four types of noises including self-recorded classroom, self-recorded cafeteria, white and pink noise at varying SNR's ranging from 0-15dB. These sets are evaluated on clean training set in baseline system. It is employed on entire above mentioned feature extraction techniques of MFCC, LPCC, RASTA-PLP and GFCC with an effort of experimenting the performance

analysis of system under degraded environmental conditions. Under contrast, RASTA-PLP has been almost comparable to MFCC and GFCC in clean and noisy environments respectively, taking into account both the settings for the test datasets. Likewise, GFCC have led to greater performance than MFCC in non-ideal (noisy) conditions with a RI of 7.11% as shown in Table 2.

Table 2. WER (%) obtained on varying test conditions using baseline MFCC and GFCC feature extraction approaches.

Training Set	Testing Set	DNN (%)			
		MFCC	LPCC	RASTA-PLP	GFCC
Clean Child	Clean Child	15.43	16.02	15.46	15.57
	Noisy Child	20.84	21.06	19.43	19.36

6.2 Performance evaluation of the system at varying SNR values

Although quantitative findings are helpful in analysing speech enhancement algorithms, where diverse training conditions plays a key role. Additionally, test audio signals when evaluated onto trained models have been found to be extremely subjective to inter-speaker variations and affected the efficiency of ASR system even under varying environmental conditions. Therefore, an attempt in the following set of experiments have been made to balance the real-life recording conditions which are captured with a microphone. It comprised of noise-reducing filters. However, the noise test data set has been improved by employing the Kalman and Wiener filtering techniques. Instead of using random noise, the effectiveness for utilisation of filtering techniques are tested with noise induction at particular SNR values such as -5dB, 0dB, 5dB, 10dB, and 15dB.

6.2.1 Performance evaluation of filtering techniques on clean training set

In this set of experiment, the effectiveness of the system under clean training conditions is assessed to further analyse the filtering techniques on noisy dataset. It is being injected at specific SNR values which tried to reproduce enhanced signal. The enhanced signal is further evaluated on feature vectors of MFCC, LPCC, RASTA-PLP and GFCC on clean training conditions. LPCC is recognised to be a loss compression method, meaning data are lost on lengthy ranges. This means both LPCC as well as GFCC are unable to use trained data for filtering. Likewise, in such a scenario RASTA-PLP could not surpass MFCC feature extraction, which leaded to degrade performance of the system at higher greater SNR values.

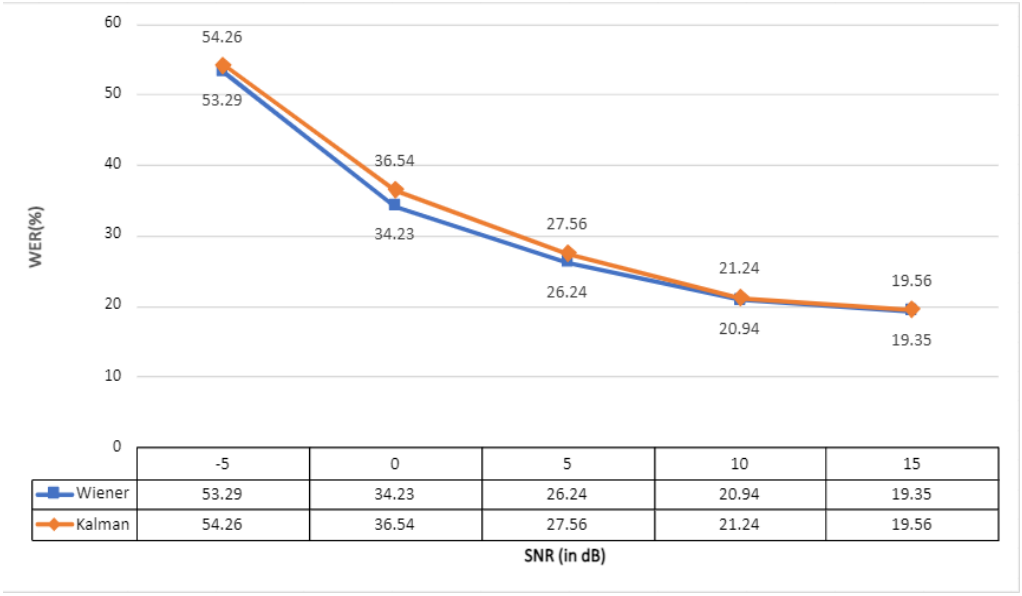


Figure 4(a). WER(%) obtained on Kalman and Wiener Filtering utilising MFCC feature extraction technique on clean training dataset

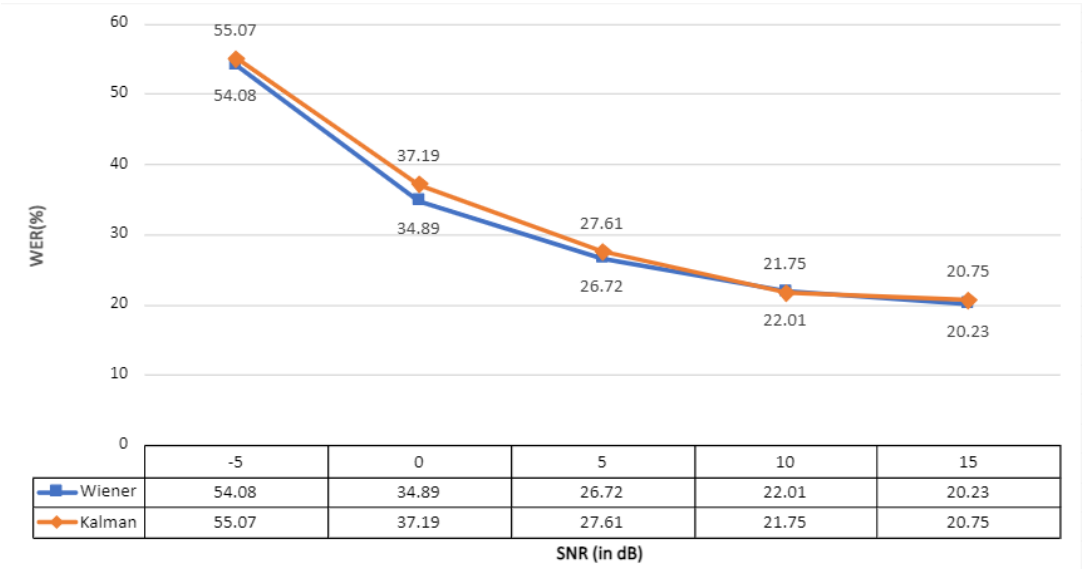


Figure 4(b). WER(%) obtained on Kalman and Wiener Filtering utilising GFCC feature extraction techniques on clean training dataset

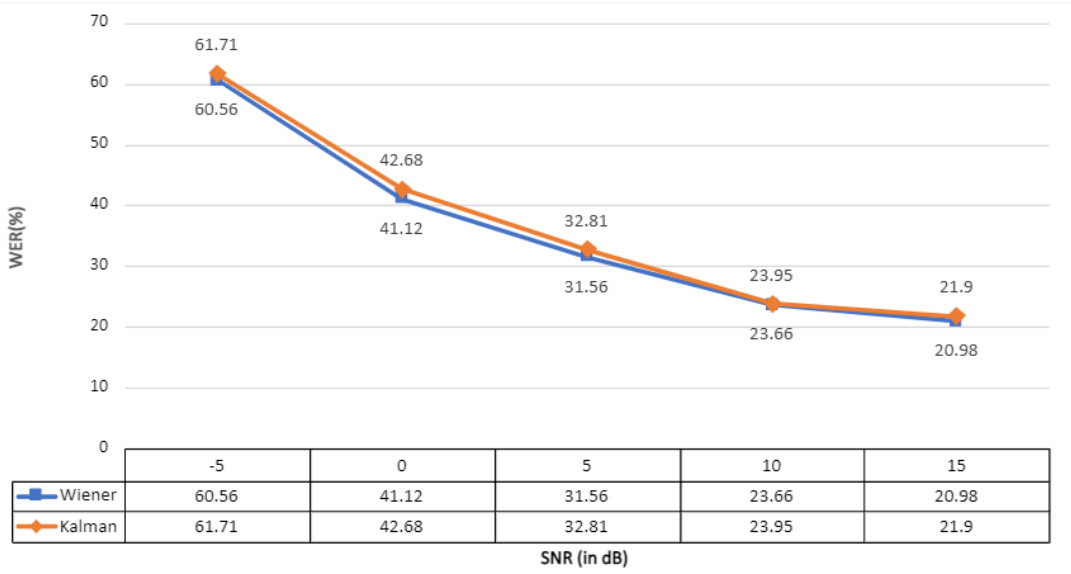


Figure 4(c). WER(%) obtained on Kalman and Wiener Filtering utilising LPCC feature extraction techniques on clean training dataset

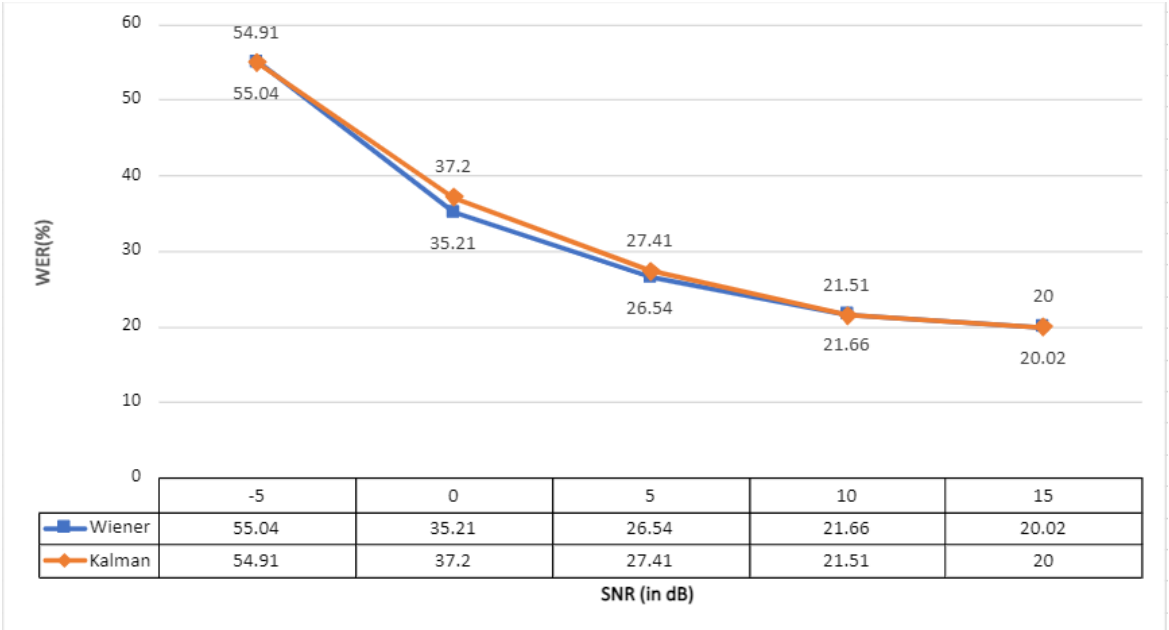


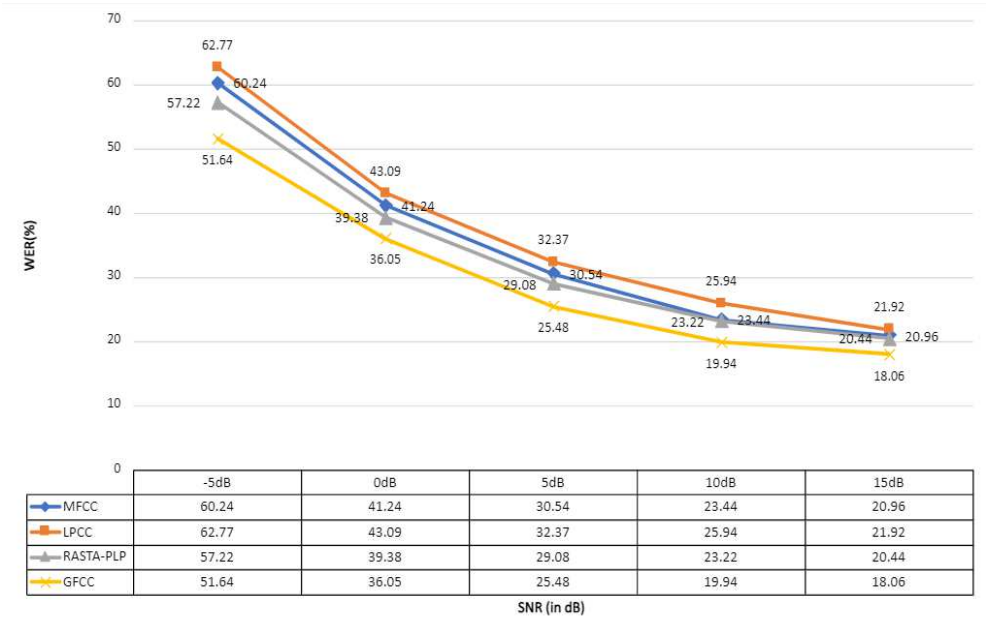
Figure 4(d). WER(%) obtained on Kalman and Wiener Filtering utilising RASTA-PLP feature extraction techniques on clean training dataset.

Furthermore, somewhat degraded performance of system is visible even at higher SNR values of 10dB, 15dB with worst output at lower SNRs –5dB and 0dB with respect to both Kalman and Weiner Filtering techniques as shown in Figure 4(a), Figure 4(b), Figure 4(c), and Figure 4(d). Moreover, the Wiener filtering approach has outperformed Kalman Filter approach with average Relative Improvement of 3.08%, 2.95%, 1.49%, and 2.51% with MFCC, LPCC, RASTA-PLP and GFCC feature extraction techniques respectively.

6.2.2 Performance evaluation of feature extraction techniques with external filtering techniques on noise augmented training dataset

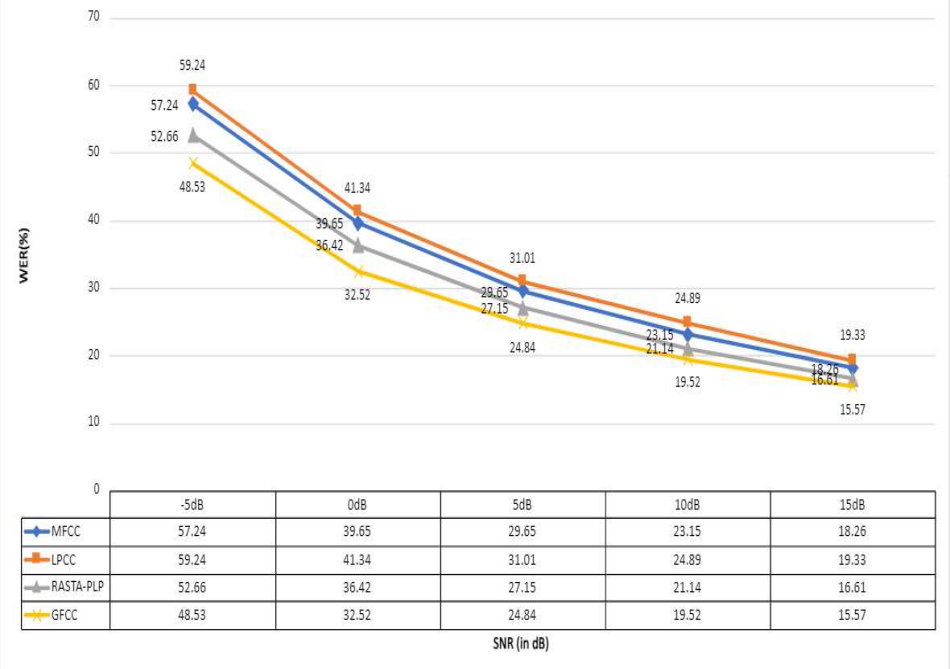
In this set of experiment, the noise clips are randomly injected into clean dataset with a SNR values ranging from –5dB to 15dB and combined into the clean dataset. This addition of noise during the evaluation on both enhanced Kalman speech set and enhanced Wiener

speech set have been created a regularisation effect which leads to an increased robustness of the model. The experiments have been proven to show the large-scale average Relative Improvement of 7.67% and 19.53% corresponding to enhanced Kalman child and enhanced Wiener child dataset using MFCC, LPCC, RASTA-PLP and GFCC feature extraction technique as detailed in Figure 5(a) and Figure 5(b). However, MFCC feature vectors for noise augmented training set has leads to degraded performance of the system with increased WER at every value of SNR for both enhanced Kalman child and enhanced wiener child dataset. In addition, noise reduction approach is utilised with both the filtering techniques that have been consequently failed in the adaptation of intelligibility factors at lower SNR values of –5dB and 0dB, which is similar to clean conditions. It resulted into decreased performance of the system.



	-5dB	0dB	5dB	10dB	15dB
MFCC	60.24	41.24	30.54	23.44	20.96
LPCC	62.77	43.09	32.37	25.94	21.92
RASTA-PLP	57.22	39.38	29.08	23.22	20.44
GFCC	51.64	36.05	25.48	19.94	18.06

Figure 5(a). WER (%) obtained on various feature extraction techniques utilising Kalman Filtering on noise augmented training dataset.



	-5dB	0dB	5dB	10dB	15dB
MFCC	57.24	39.65	29.65	23.15	18.26
LPCC	59.24	41.34	31.01	24.89	19.33
RASTA-PLP	52.66	36.42	27.15	21.14	16.61
GFCC	48.53	32.52	24.84	19.52	15.57

Figure 5(b). WER (%) obtained on various feature extraction techniques utilising Wiener Filtering on noise augmented training dataset.

6.3 Performance Evaluation on train-augmented datasets using vocal tract length normalisation

Table 3. WER (%) obtained on varying filtering techniques employing test normalisation with MFCC and GFCC feature extraction approaches.

Training Set	Testing Set	DNN (%)			
		MFCC	**MFCC+VTLN**	**GFCC**	**GFCC+VTLN**
Clean Child + Noisy Child	Enhanced Kalman Child	19.59	18.95	15.34	14.78
	Enhanced Wiener Child	18.66	17.84	14.98	14.21

It should be noticed from the findings made in earlier experimentations that both filtering strategies alone did not improve the robustness of the system at lower SNR values of –5dB and 0dB. Therefore, test set has been first augmented with random SNR values which ranging from 5dB to 15dB. Further, both Kalman and Wiener filtering techniques have been employed on testing signal for generation of enhanced speech signals. The enhanced signals corresponding to both the filtering techniques are evaluated using both feature extraction technique of MFCC and GFCC on noise-enhanced training set at SNR values ranging from –5dB to 15dB. From above findings, it is evident that MFCC is more resilient in clean conditions and GFCC in noisy situations, despite the employment of sufficient filtering techniques, as contrasted to LPCC and RASTA-PLP feature extraction methods. The result obtained showed that it achieved better performance with GFCC compared to MFCC feature extraction technique with a relative improvement of 24.55% and 30.65% using Kalman and Wiener filtering technique as shown in Table 3. Further, the vocal length in children is shorter which resulted into the need of implication of VTLN on the test dataset. This leads to reduction in variation existing among the training and testing dataset. The test normalisation are followed by implementation of Wiener Filtering technique. It resulted into RI of 4.39% with MFCC and 5.14% with GFCC feature extraction methods. The overall RI of 4.21% and 7.91% in case of Kalman and Wiener Filtering technique which are employed with GFCC feature extraction is obtained in comparison to the baseline system.

6.4 Performance evaluation on perturbed noise augmented dataset using spectral warping

The improved performance of the system is obtained by employing GFCC+VTLN along with enhanced Wiener Child which is quite observable. In like manner, the frequency domain features or transfer function of a test device are often relevance in both analogue and mixed signal devices. In region our interest, an emphasis is typically laid on a specific area of frequency spectrum rather than entire spectrum. So the technique of spectral warping is applied by time reversing the samples inside a frame and feed them into filter network, the spectral warping method is implemented as per each frame. The outputs from each of the first order filter stages are provided as a sample for each twisted signal. A warp factor of varying of values ranged from -0.1 to 0.1 along with a step size of 0.0025 is detailed in Figure 6. In

this way, the best value of the warp factor has been found to be -0.075 such that the effective improvement is obtained with real-time system. It is possible by employing GFCC+VTLN along with external enhanced wiener filtering technique. It is reported with a RI of 2.53% in contrast to GFCC+VTLN based system.

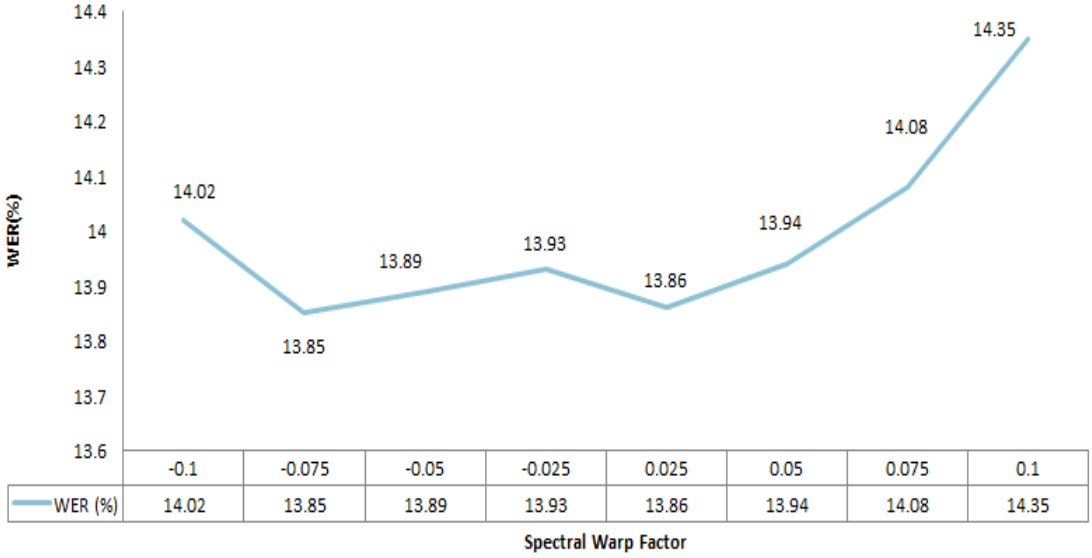


Figure 6. WER (%) obtained on varying spectral warping factor for noise augmented dataset trained on Enhanced Wiener Children dataset.

Finally, an experiment is conducted with a goal of reducing data scarcity using DNN-HMM based hybrid technique. It employed artificially generated noise enhanced training dataset and afterwards pooling it with 3, 4, 5-way based perturbation. Furthermore, an alignments is performed with speed perturbed data. It is rebuilt using DNN-HMM system due to a change in the duration of the signal. However, in low-resource language, we found that a very little improvement is obtained as illustrated in Table 4. It is probably because of the data that had previously been supplemented with simulated reverberation. The overall RI of 16.13% on 5-way perturbed spectral warped noise augmented dataset is achieved in case of Wiener filtering technique on GFCC+VTLN approach in comparison to the baseline system.

Table 4. WER (%) obtained on perturbed trained dataset trained on Enhanced Wiener Children dataset employing GFCC and VTLN

Training Set	Augmentation Type	Warp Factor	DNN (%)
Clean Child + Noisy Child	None	-0.075	13.85
	3-way		13.06
	4-way		13.02
	5-way		12.94

6.5 Comparative analysis of noise-robust system with earlier proposed approaches

The rapidly expanding area of automatic speech recognition is confronted with a number of challenges, including vocabulary size, style of speech, speaker mode, and, most all, environmental resilience. Deep learning has become a strong science in the process of speech recognition, based on sophisticated network architectures and large model parameters, with broad and far-reaching implications. Most of the researchers focused on fully resource datasets but lacks in low-resource speech recognition systems. In past researchers have been addressing these challenges in order to improve the performance of ASR systems. In the application domains, automatic speech recognition is being investigated to a considerable

extent where an effort of developing noise-robust ASR system has been made as shown in Table 5. Apart, issue of data scarcity and noise conditions disturbance are overcome in this proposed research using efficient spectral warping method and noise filtering is performed with proposed hybrid approach in comparison to other state of the art work.

Table 5. Comparative Analysis of noise-robust system with earlier proposed approaches

Authors	**Feature Extraction**	**De-noising Techniques**	**Summary**
Bawa et al. 2021	MFCC;GFCC	-	The researcher worked upon Punjabi children dataset under mismatched conditions by employing DNN-HMM based hybrid architecture. An overall RI of 30.94% was proposed on gender-pooled noisy dataset.
Kumar et al. 2021	MFCC	AutoSSR approach	The researchers created Punjabi dataset of around 200 mins with 87.10% sentence-level accuracy and word-level accuracy of 94.19%
Tian et al. 2018	MFCC	Running spectrum filtering	The researcher explored children speech recognition under noisy circumstances. With proposed filtering technique system was able to reach 96.6% accuracy for Japanese dataset.
Qian et al. 2016	MFCC	VTLN along with DNN-HMM adaptation	The researchers used CMU kids' corpus for experimenting with multiple types of noise synthetically injected into clean corpus and attained a relative improvement of 10% on the corpus.
Upadhyaya et al. 2017	MFCC, PLP	-	The researchers used 1000 continuous sentence from Hindi language and attained 16.09% WER after experimented it with MFCC and PLP feature extraction along with 2-gram to 4-gram language modelling techniques.
Proposed Approach	MFCC, GFCC, RASTA-PLP, GFCC	Wiener and Kalman Filtering	In contrast to the baseline system, the Wiener Filtering Technique using GFCC+VTLN technique yielded an overall RI of 16.13% on a 5-way perturbed spectral warped noise augmented dataset.

7. Conclusion

The research for acoustic and linguistic constructs in children speech is addressed in this study via comparative analysis of filtering strategies. It demonstrated with four front end feature extraction methods of MFCC, LPCC, RASTA-PLP and GFCC. These methods have been further evaluated by use of test-based normalisation technique through VTLN for reduction of inter-speaker differences and excitation source characteristics on scarce resources under deteriorated environmental conditions. Two type of corpus, including the clean children dataset and noise-enhanced children dataset using DNN-HMM classifier, have been shown to increase the robustness of the ASR system under non-ideal environments. The obtained findings using synthetic training data were shown to be more beneficial on children speech corpus with an overall improvement of 24.55% (Kalman Filtering) and 30.65% (Wiener Filtering) under noisy conditions. The overall experimental analysis demonstrated the effectiveness of the proposed spectral warped noise augmented system utilising Wiener filter alongside the use of GFCC feature extraction. An overall relative improvement of

16.13% compared to the baseline system is achieved. In future, the comprehensive study can be expanded with the implementation of these filtering methods in speech-based systems, including speaker verification & authentication, gender & emotion classification, using various augmentative methodologies and more broadly, an out-domain speech augmentation approach.

Conflict of Interest: The authors declare that they have no conflict of interest.

References

Abd El-Fattah, M. A., Dessouky, M. I., Abbas, A. M., Diab, S. M., El-Rabaie, E. S. M., Al-Nuaimy, W., & Abd El-Samie, F. E. (2014). Speech enhancement with an adaptive Wiener filter. International Journal of Speech Technology, 17(1), 53-64. https://doi.org/10.1007/s10772-013-9205-5

Ali, A. A., Van Der Speigel, J., & Mueller, P. (2000). Auditory-based speech processing based on the average localized synchrony detection. In 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100) (Vol. 3, pp. 1623-1626). IEEE. https://doi.org/10.1109/ICASSP.2000.862016

Barker, J., Cooke, M., & Green, P. (2001). Robust ASR based on clean speech models: An evaluation of missing data techniques for connected digit recognition in noise. In Seventh European Conference on Speech Communication and Technology.

Bassan, N., & Kadyan, V. (2019). An Experimental Study of Continuous Automatic Speech Recognition System Using MFCC with Reference to Punjabi Language. In Recent Findings in Intelligent Computing Techniques (pp. 267-275). Springer, Singapore. https://doi.org/10.1007/978-981-10-8639-7_28

Besacier, L., Barnard, E., Karpov, A., & Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. Speech Communication, 56, 85-100. https://doi.org/10.1016/j.specom.2013.07.008

Boashash, B., & Mesbah, M. (2004). Signal enhancement by time-frequency peak filtering. IEEE Transactions on signal processing, 52(4), 929-937. https://doi.org/10.1109/TSP.2004.823510

Boll, S. (1979). Suppression of acoustic noise in speech using spectral subtraction. IEEE Transactions on acoustics, speech, and signal processing, 27(2), 113-120. https://doi.org/10.1109/TASSP.1979.1163209

Das, O., Goswami, B., & Ghosh, R. (2016). Application of the tuned kalman filter in speech enhancement. In 2016 IEEE first international conference on control, measurement and instrumentation (CMI) (pp. 62-66). IEEE. https://doi.org/10.1109/CMI.2016.7413711

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE transactions on acoustics, speech, and signal processing, 28(4), 357-366. https://doi.org/10.1109/TASSP.1980.1163420

Ephraim, Y., & Malah, D. (1984). Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Transactions on acoustics, speech, and signal processing, 32(6), 1109-1121. https://doi.org/10.1109/TASSP.1984.1164453

Esposito, A., Ezin, E. C., & Reyes-Garcia, C. A. (2000). Designing a fast neuro-fuzzy system for speech noise cancellation. In Mexican International Conference on Artificial Intelligence (pp. 482-492). Springer, Berlin, Heidelberg. https://doi.org/10.1007/10720076_44

Espy-Wilson, C. Y., Chari, V. R., & Huang, C. B. (1996). Enhancement of alaryngeal speech by adaptive filtering. In Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96 (Vol. 2, pp. 764-767). IEEE. https://doi.org/10.1109/ICSLP.1996.607475

Etter, W., & Moschytz, G. S. (1994). Noise reduction by noise-adaptive spectral magnitude expansion. Journal of the Audio Engineering Society, 42(5), 341-349. http://www.aes.org/e-lib/browse.cfm?elib=6947

Goh, Z., Tan, K. C., & Tan, B. T. G. (1999). Kalman-filtering speech enhancement method based on a voiced-unvoiced speech model. IEEE Transactions on speech and audio processing, 7(5), 510-524. https://doi.org/10.1109/89.784103

Gong, Y. (1995). Speech recognition in noisy environments: A survey. Speech communication, 16(3), 261-291. https://doi.org/10.1016/0167-6393(94)00059-J

Gopalakrishna, V., Kehtarnavaz, N., Mirzahasanloo, T. S., & Loizou, P. C. (2012). Real-time automatic tuning of noise suppression algorithms for cochlear implant applications. IEEE Transactions on Biomedical Engineering, 59(6), 1691-1700. https://doi.org/10.1109/TBME.2012.2191968

Gulzar, T., Singh, A., & Sharma, S. (2014). Comparative analysis of LPCC, MFCC and BFCC for the recognition of Hindi words using artificial neural networks. International Journal of Computer Applications, 101(12), 22-27.

Gupta, H., & Gupta, D. (2016). LPC and LPCC method of feature extraction in Speech Recognition System. In 2016 6th International Conference-Cloud System and Big Data Engineering (Confluence) (pp. 498-502). IEEE. https://doi.org/10.1109/CONFLUENCE.2016.7508171

Gurugubelli, K., & Vuppala, A. K. (2019). Perceptually enhanced single frequency filtering for dysarthric speech detection and intelligibility assessment. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6410-6414). IEEE. https://doi.org/10.1109/ICASSP.2019.8683314

Haque, M., & Bhattacharyya, K. (2019). A study on different linear and non-linear filtering techniques of speech and speech recognition. ADBU Journal of Engineering Technology, 8.

Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. the Journal of the Acoustical Society of America, 87(4), 1738-1752. https://doi.org/10.1121/1.399423

Hermansky, H., Morgan, N., Bayya, A., & Kohn, P. (1991). RASTA-PLP speech analysis. In Proc. IEEE Int'l Conf. Acoustics, Speech and Signal Processing (Vol. 1, pp. 121-124).

Kadyan, V., Mantri, A., & Aggarwal, R. K. (2017). A heterogeneous speech feature vectors generation approach with hybrid hmm classifiers. International Journal of Speech Technology, 20(4), 761-769. https://doi.org/10.1007/s10772-017-9446-9

Kaur, J., Singh, A., & Kadyan, V. (2020). Automatic Speech Recognition System for Tonal Languages: State-of-the-Art Survey. Archives of Computational Methods in Engineering, 1-30. https://doi.org/10.1007/s11831-020-09414-4

Kim, D. S., Lee, S. Y., & Kil, R. M. (1999). Auditory processing of speech signals for robust speech recognition in real-world noisy environments. IEEE Transactions on speech and audio processing, 7(1), 55-69. https://doi.org/10.1109/89.736331

Ko, T., Peddinti, V., Povey, D., & Khudanpur, S. (2015). Audio augmentation for speech recognition. In Sixteenth Annual Conference of the International Speech Communication Association.

Ko, T., Peddinti, V., Povey, D., Seltzer, M. L., & Khudanpur, S. (2017). A study on data augmentation of reverberant speech for robust speech recognition. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5220-5224). IEEE. https://doi.org/10.1109/ICASSP.2017.7953152

Koopmans, W. J., Goverts, S. T., & Smits, C. (2018). Speech recognition abilities in normal-hearing children 4 to 12 years of age in stationary and interrupted noise. Ear and Hearing, 39(6), 1091-1103. https://doi.org/10.1097/AUD.0000000000000569

Lang, M., Guo, H., Odegard, J. E., Burrus, C. S., & Wells, R. O. (1996). Noise reduction using an undecimated discrete wavelet transform. IEEE Signal Processing Letters, 3(1), 10-12. https://doi.org/10.1109/97.475823

Lee, Y. K., Park, J. G., Lee, Y. K., & Kwon, O. W. (2014). Speech Enhancement Using Phase-Dependent A Priori SNR Estimator in Log-Mel Spectral Domain. ETRI Journal, 36(5), 721-729. https://doi.org/10.1109/ICCE.2011.5722657

Lim, J. S., & Oppenheim, A. V. (1979). Enhancement and bandwidth compression of noisy speech. Proceedings of the IEEE, 67(12), 1586-1604. https://doi.org/10.1109/PROC.1979.11540

Ma, J. Z., Keith, F., Ng, T., Siu, M. H., & Kimball, O. (2017). Improving Deliverable Speech-to-Text Systems with Multilingual Knowledge Transfer. In INTERSPEECH (pp. 127-131).

Mesgarani, N., Slaney, M., & Shamma, S. A. (2006). Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations. IEEE Transactions on Audio, Speech, and Language Processing, 14(3), 920-930. https://doi.org/10.1109/TSA.2005.858055

Nair, A. P., Krishnan, S., & Saquib, Z. (2016). Mfcc based noise reduction in asr using kalman filtering. In 2016 Conference on Advances in Signal Processing (CASP) (pp. 474-478). IEEE. https://doi.org/10.1109/CASP.2016.7746218

Narayana, M. L., & Kopparapu, S. K. (2009). Effect of noise-in-speech on mfcc parameters. In Proceedings of the 9th WSEAS international conference on signal, speech and image processing, and 9th WSEAS international conference on Multimedia, internet & video technologies (pp. 39-43).

Novotney, S., Schwartz, R., & Ma, J. (2009). Unsupervised acoustic and language model training with small amounts of labelled data. In 2009 IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 4297-4300). IEEE. https://doi.org/10.1109/ICASSP.2009.4960579

Paliwal, K., & Basu, A. (1987). A speech enhancement method based on Kalman filtering. In ICASSP'87. IEEE International Conference on Acoustics, Speech, and Signal Processing (Vol. 12, pp. 177-180). IEEE. https://doi.org/10.1109/ICASSP.1987.1169756

Pervaiz, A., Hussain, F., Israr, H., Tahir, M. A., Raja, F. R., Baloch, N. K., ... & Zikria, Y. B. (2020). Incorporating Noise Robustness in Speech Command Recognition by Noise Augmentation of Training Data. Sensors, 20(8), 2326. https://doi.org/10.3390/s20082326

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., ... & Silovsky, J. (2011). The Kaldi speech recognition toolkit. In IEEE 2011 workshop on automatic speech recognition and understanding (No. CONF). IEEE Signal Processing Society.

Robinson, T., Fransen, J., Pye, D., Foote, J., & Renals, S. (1995). WSJCAMO: a British English speech corpus for large vocabulary continuous speech recognition. In 1995 International Conference on Acoustics, Speech, and Signal Processing (Vol. 1, pp. 81-84). IEEE. https://doi.org/10.1109/ICASSP.1995.479278

Saldanha, J. (2016). Speech Enhancement Using Filtering Techniques. 3rd National Conference on Emerging Trends in Electronics and Communication (NCETEC-16)

Sameti, H., Sheikhzadeh, H., Deng, L., & Brennan, R. L. (1998). HMM-based strategies for enhancement of speech signals embedded in nonstationary noise. IEEE Transactions on Speech and Audio processing, 6(5), 445-455. https://doi.org/10.1109/89.709670

Scarr, R. (1968). Zero crossings as a means of obtaining spectral information in speech analysis. IEEE Transactions on Audio and Electroacoustics, 16(2), 247-255. https://doi.org/10.1109/TAU.1968.1161984

Singh, A., Kadyan, V., Kumar, M., & Bassan, N. (2019). ASRoIL: a comprehensive survey for automatic speech recognition of Indian languages. Artificial Intelligence Review, 1-32. https://doi.org/10.1007/s10462-019-09775-8

Sivasankaran, S., Nugraha, A. A., Vincent, E., Morales-Cordovilla, J. A., Dalmia, S., Illina, I., & Liutkus, A. (2015). Robust ASR using neural network based speech enhancement and feature simulation. In 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU) (pp. 482-489). IEEE. https://doi.org/10.1109/ASRU.2015.7404834

Sorqvist, P., Handel, P., & Ottersten, B. (1997). Kalman filtering for low distortion speech enhancement in mobile communication. In 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing (Vol. 2, pp. 1219-1222). IEEE. https://doi.org/10.1109/ICASSP.1997.596164

Subramanian, A. S., Wang, X., Watanabe, S., Taniguchi, T., Tran, D., & Fujita, Y. (2019). An investigation of end-to-end multichannel speech recognition for reverberant and mismatch conditions. arXiv preprint arXiv:1904.09049.

Sárosi, G., Mozsáry, M., Mihajlik, P., & Fegyó, T. (2011). Comparison of feature extraction methods for speech recognition in noise-free and in traffic noise environment. In 2011 6th Conference on Speech Technology and Human-Computer Dialogue (SpeD) (pp. 1-8). IEEE. https://doi.org/10.1109/SPED.2011.5940729

Valente, D. L., Plevinsky, H. M., Franco, J. M., Heinrichs-Graham, E. C., & Lewis, D. E. (2012). Experimental investigation of the effects of the acoustical conditions in a simulated classroom on speech recognition and learning in children. The Journal of the Acoustical Society of America, 131(1), 232-246. https://doi.org/10.1121/1.3662059

Zhang, X., Wang, Z. Q., & Wang, D. (2017). A speech enhancement algorithm by iterating single-and multi-microphone processing and its application to robust ASR. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 276-280). IEEE. https://doi.org/10.1109/ICASSP.2017.7952161

Zhao, X., & Wang, D. (2013). Analyzing noise robustness of MFCC and GFCC features in speaker identification. In 2013 IEEE international conference on acoustics, speech and signal processing (pp. 7204-7208). IEEE. https://doi.org/10.1109/ICASSP.2013.6639061

Zhao, Y., Wang, H., & Cui, R. (2011). An approach to sound feature extraction method based on gammatone filter. In Advances in Multimedia, Software Engineering and Computing Vol. 2 (pp. 371-376). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-25986-9_57

Goncharoff, V., & Chandran, S. (1988, January). Adaptive speech modification by spectral warping. In ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing (pp. 343-344). IEEE Computer Society.

Umesh, S., Cohen, L., Marinovic, N., & Nelson, D. (1996). Frequency-warping in speech. In Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96 (Vol. 1, pp. 414-417). IEEE.

Paliwal, K., Shannon, B., Lyons, J., & Wójcicki, K. (2009). Speech-signal-based frequency warping. IEEE signal processing letters, 16(4), 319-322.

Kadyan, V., Shanawazuddin, S., & Singh, A. (2021). Developing children's speech recognition system for low resource Punjabi language. Applied Acoustics, 178, 108002.

Bawa, P., & Kadyan, V. (2021). Noise robust in-domain children speech enhancement for automatic Punjabi recognition system under mismatched conditions. Applied Acoustics, 175, 107810.

Kumar, Y., Singh, N., Kumar, M., & Singh, A. (2021). AutoSSR: an efficient approach for automatic spontaneous speech recognition model for the Punjabi Language. Soft Computing, 25(2), 1617-1630.

Tian, Y., Tang, J., Jiang, X., Tsutsui, H., & Miyanaga, Y. (2018). Accuracy on Children's Speech Recognition under Noisy Circumstances. In 2018 18th International Symposium on Communications and Information Technologies (ISCIT) (pp. 101-104). IEEE.

Qian, M., McLoughlin, I., Quo, W., & Dai, L. (2016). Mismatched training data enhancement for automatic recognition of children's speech using DNN-HMM. In 2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP) (pp. 1-5). IEEE.

Upadhyaya, P., Farooq, O., Abidi, M. R., & Varshney, Y. V. (2017). Continuous Hindi speech recognition model based on Kaldi ASR toolkit. In 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET) (pp. 786-789). IEEE.