Signal Subspace Speech Enhancement with Oblique Projection and Normalization

Sudeep SURENDRAN, T. Kishore KUMAR

Dept. of E.C.E, National Institute of Technology, Warangal, Telangana, India

{sudipsuren, kishoret}@nitw.ac.in

Submitted January 20, 2017 / Accepted April 7, 2017

Abstract. *In this paper, a subspace speech enhancement method handling colored noise using oblique projection is proposed. Perceptual features and variance normalization are used to reduce residual noise and improve speech intelligibility of the output. Initially, additive noise is removed from the noisy speech by removing the orthogonal noise subspace from the noisy speech subspace to obtain the speech subspace. Then, the oblique projection of the noise subspace on the speech subspace along the additive noise subspace is used to determine the colored noise that remains. The enhanced clean speech signal is estimated using Spectral Domain Constrained Estimator, incorporating the masking property of the auditory system and the variance of the colored noise. To avoid the occurrence of any abrupt spikes in the output, variance normalization is performed by adaptively changing the control parameter of the gain matrix. The spectrogram, objective measures and subjective intelligibility test show superior performance of the proposed method over other existing speech enhancement methods.*

Keywords

Speech enhancement, signal subspace approach, masking property, oblique projection, variance normalization

1. Introduction

Speech enhancement aims at improving the quality and/or intelligibility of speech signals corrupted by noise using various signal processing techniques. It is imperative in a lot of applications like automatic speech recognition, wireless/wired communication systems etc., as a pre-processing block. Speech enhancement techniques can be broadly classified into supervised and unsupervised methods. Supervised methods like Hidden Markov model (HMM) based methods [1] and Gaussian Mixture Models(GMM) [2] consider a model for both the speech and noise signals whose parameters are estimated from the training samples of that signal and achieves noise reduction by defining an interaction model. The performance of the supervised approaches is limited by the prior information available and fails to deliver the desired performance in non-stationary noise environments. In Unsupervised methods, which assume a statistical model for the speech and noise signals, noise reduction is achieved by estimating clean speech from noisy observations without any prior information. Spectral subtraction [3], Wiener filtering [4], Kalman filtering [5], short-time spectral amplitude (STSA) estimators [6], Signal Subspace Approach (SSA) etc. belong to this category. Adaptive filters are also used for speech enhancement as used in [7] for echo cancellation. Due to the simplicity and ease of implementation in single channel systems, spectral subtraction and Wiener filtering have been widely used for enhancing speech. But these and most of the other existing methods have major drawbacks of generating unpleasant residual noise called musical noise after enhancement.

To reduce the amount of musical noise, smoothing methods like the decision directed approach [6] or Wiener filtering based on apriori SNR estimation [4] are often utilized. Most algorithms often produce high signal distortion in an attempt to reduce residual noise. There has always been an effort to develop speech enhancement techniques that give good compromise between residual noise and signal distortion of the output signal. SSA has shown to give a better compromise between the two, compared to the other existing techniques. Hilkhuysen et al. [8] has shown that the intelligibility of subspace enhancement [9] output is better compared to that of spectral subtraction [10] and minimum mean squared error spectral subtraction [11] methods.

To deal with the case of colored noise instead of approximating the noise covariance matrix with an identity matrix, Rezayee and Gazor [12] developed a Time-Domain Constraint (TDC) estimator using a diagonal matrix for coloured-noise power spectrum which resulted in a sub-optimal estimator. Hu and Loizou [9] presented a TDC estimator based on the joint diagonalization of the covariance matrices of the clean signal and the noise process. Lev-Ari and Ephraim [13] extended their original speech enhancement subspace approach to coloured-noise processes in the time and spectral domains using whitening of the input noise. Pre-whitening and De-whitening [14], use of a common diagonalization matrix [15], use of Rayleigh quotient method [16] etc. are also employed for handling the case. In Rayleigh quotient method, the noise variance, σ^2, is taken as the noise energy in the direction of the i^{th} eigenvector, which is the Rayleigh

DOI: 10.13164/re.2017.1161

Quotient associated with the i^{th} eigenvector of the speech covariance matrix and noise covariance matrix. Oblique projection is used in this paper to handle the case of colored noise. It has been used in certain signal processing applications [17]. Oblique projection allows decomposing a matrix into two non-orthogonal components which makes it suitable for filtering out colored noise from noisy speech.

The use of perceptual features in subspace method by Jabloun and Champagne [16] reduced the residual noise compared to the conventional signal subspace methods. Frequency masking property of human auditory system was used to decide the gain parameters for filtering the noisy speech. This property makes the noise of a particular band of frequency inaudible to the listener if it falls below the masking threshold of that particular frequency, making it possible to hide inaudible noise signals in the enhanced speech signal. This reduces the amount of filtering required, which in effect reduces the residual noise and signal distortion.

After enhancement, even after reducing musical noise to an extent, the output signal may suffer from abrupt changes or spikes which are audible and compromise intelligibility performance. Normalizations could be performed on the output signal to reduce such spikes and distortions. Lu [18], employed an optimal smoothing factor, adapted by the variation of signal to spectral deviation ratio in successive frames. Variance normalization was used in spectral subtraction speech enhancement algorithm by Maganti and Matassoni [19] across the critical bands to smoothen the output signal, removing the spikes in the output which reduced the effect of increased variance at random frequencies.

The paper proposes a speech enhancement algorithm using SSA employing Eigen Value Decomposition (EVD). Additive noise is removed by subtracting the orthogonal noise on to the signal subspace. Colored noise is handled by a novel technique using oblique projection. Perceptual features (simultaneous and temporal masking) are incorporated in Spectral Domain Constrained (SDC) estimator of clean speech signal, reducing signal distortion and residual noise. Further, the filtering level is adaptively varied according to the normalized variance, to avoid the occurrence of abrupt changes in the output making it more intelligible.

This paper is structured as follows. Sec. 2 explains the signal subspace approach and Sec. 3 explains the proposed speech enhancement method. The results in terms of spectrograms, objective tests and subjective test are presented in Sec. 4 and the conclusion is provided in Sec. 5.

2. Signal Subspace Approach

Subspace algorithms decompose the vector space of the noisy signal into noise subspace having noise signals and speech subspace having speech signals. They are rooted on linear algebra theory and are based on the principle that the clean signal is mostly confined to a subspace of the noisy Euclidean space. Orthogonal matrix factorization techniques such as Singular Value Decomposition (SVD) and EVD are employed for the decomposition. Speech enhancement is obtained by estimating the clean speech signal by removing the noise subspace and filtering noise elements from the speech subspace.

SSA provides dimensionality reduction and a better compromise between signal distortion and residual noise over other speech enhancement methods. Tufts et al. [20] presented a least squares estimator (LS) method for retrieving the signal component from a noisy data set by projecting the noisy signal onto the signal subspace using SVD. Dendrinos et al. [21] first utilized signal subspace techniques to enhance speech who proposed the use of SVD on a data matrix containing time-domain amplitude values. Karhunen-Loeve Transform (KLT) was used by Ephraim and Van Trees [14] for speech enhancement in which filtering was performed using a diagonal gain matrix based on the uncorrelated nature of the coefficients in the subspace. The additive noise was assumed to be zero-mean, white, and uncorrelated with the speech signal. The gain matrix elements were decided based on TDC or SDC estimators. The former attempted to spectrally shape the residual noise while the latter constrained residual noise energy. Variance of the reconstruction error was included in perceptual KLT in order to optimize the subspace decomposition mode in [25].

Noisy speech vector $x = [x_0, x_1, ...x_{N-1}]^{\mathrm{T}}$ composed of clean speech signal (s) and noise (w) assumed to be uncorrelated with (s), can be represented as in (1). Their corresponding covariance matrices follow (2).

$$x = s + w, \tag{1}$$

$$\boldsymbol{R_x} = \boldsymbol{R_s} + \boldsymbol{R_w}, \tag{2}$$

$\boldsymbol{R_x}$ is assumed to have a higher rank than $\boldsymbol{R_s}$. $\boldsymbol{R_x}$ and $\boldsymbol{R_s}$ are Toeplitz matrices, the EVDs of which are given below.

$$\boldsymbol{R_x} = \boldsymbol{U\Lambda U}^{\mathrm{H}}, \tag{3}$$

$$\boldsymbol{R_s} = \boldsymbol{U_P}\boldsymbol{\Lambda}_{\mathrm{s}}\boldsymbol{U_P}^{\mathrm{H}}. \tag{4}$$

Here, $\boldsymbol{U} = [\boldsymbol{U_P U_{Q-P}}]$ where, $\boldsymbol{U_P} = [u_1, u_2, \ldots\ldots u_P]$ spans the signal subspace and $\boldsymbol{U_{Q-P}} = [u_{P+1}, u_{P+2}, \ldots u_Q]$ spans the noise subspace, where u_i represents the eigenvector corresponding to the eigenvalue λ_i. The dimension of $\boldsymbol{U}$ is $\boldsymbol{Q}$ and that of $\boldsymbol{U_P}$ is P such that $Q > P$. $\boldsymbol{\Lambda}$ and $\boldsymbol{\Lambda}_{\mathrm{s}}$ are the diagonal eigenvalue matrices of $\boldsymbol{R_x}$ and $\boldsymbol{R_s}$ respectively. Also, $\boldsymbol{R_w}$ is given by (5),

$$\boldsymbol{R_w} = \sigma^2\boldsymbol{I} \tag{5}$$

where σ^2 is the variance of noise and $\boldsymbol{I}$ represents identity matrix. Thus,

$$\boldsymbol{R_x} = \boldsymbol{U}(\boldsymbol{\Lambda}_{\mathrm{s}} + \sigma^2\boldsymbol{I})\boldsymbol{U}^{\mathrm{H}}. \tag{6}$$

A linear filter $\boldsymbol{H}$ is designed so as to separate the signal subspace from the noise subspace to get an estimate of the clean speech signal $\widehat{s}$ as shown in (7).

$$\widehat{s} = \boldsymbol{H}x. \tag{7}$$

Also, the signal distortion (r_{s}) and the residual noise (r_w) are given by (8) and (9) respectively.

$$r_{\mathrm{s}} = (\boldsymbol{H} - \boldsymbol{I})s, \tag{8}$$

$$r_w = \boldsymbol{H}w. \tag{9}$$

2.1 Signal Estimator

Estimators like Maximum Likelihood, Least Squares, Minimum Variance, TDC etc., could be used to estimate the clean speech from the noisy speech. An estimator of interest, employed in this paper, is SDC estimator which is the solution for the optimization problem which minimizes the signal distortion subject to keeping every spectral component of the residual noise in signal subspace below a threshold as provided in (10),

$$\min_{\boldsymbol{H}} E\{\|r_{\mathrm{s}}\|^2\}$$
$$\text{subject to} \begin{cases} E\{|u_i^{\mathrm{H}} r_w|^2\} \le \alpha_i \sigma^2, & \text{for } 1 \le i \le P, \\ E\{|u_i^{\mathrm{H}} r_w|^2\} = 0, & \text{for } P < i \le Q \end{cases} \tag{10}$$

where α_i, is a set of non-negative constants [14]. The solution of matrix $\boldsymbol{H}$ is given by (11), where $\boldsymbol{U_P}^{\mathrm{H}}$ is the KLT and G is the gain matrix given by (12) with control parameter ν.

$$\boldsymbol{H} = \boldsymbol{U_P} \boldsymbol{G} \boldsymbol{U_P}^{\mathrm{H}}, \tag{11}$$

$$\begin{aligned} \boldsymbol{G} &= \text{diag (gain values corresponding to each } u_i), \\ &= g_i = \mathrm{e}^{(-\nu\sigma^2/\lambda_{s,i})}, \quad \text{for } i = 1, 2, \ldots\ldots P. \end{aligned} \tag{12}$$

3. Proposed PKLT-OBL Method

A perceptual subspace speech enhancement method using oblique projection with variance normalization is proposed (PKLT-OBL). A combination of simultaneous and temporal masking properties of human auditory system in subspace approach of speech enhancement together with variance normalization to reduce residual noise and oblique projection to handle the case of colored noise is considered. Variance normalization is used to adaptively vary the control parameter in the gain of the filter matrix according to the variance of the power spectral density (PSD) of the noisy speech signal in the speech subspace to fine tune the filtering to further reduce the occurrence of residual noise in the output.

The flow diagram of the proposed (PKLT-OBL) method is provided in Fig. 1. Initially, $\boldsymbol{R_x}$ is computed from the noisy speech sample and $\boldsymbol{R_w}$ is obtained during the silence periods of the noisy speech. Additive noise is removed by calculating the orthogonal projection of the eigenvectors of noise subspace on the noisy speech covariance matrix eigenvectors and then subtracting the reconstructed covariance matrix from the noisy speech covariance matrix to obtain speech covariance matrix. To handle the case of colored noise, oblique projection of the noisy speech covariance matrix eigenvectors on the speech covariance matrix eigenvectors is determined. Covariance matrix of the colored noise is constructed from the oblique projection, which is used to compute its PSD. Finally, the noisy speech is filtered to obtain the enhanced speech signal using filter matrix obtained using a modified SDC estimator.

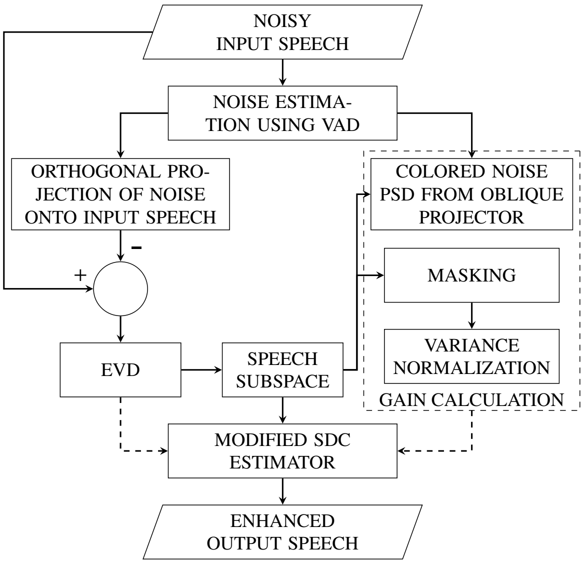

Fig. 1. Flow diagram of the proposed PKLT-OBL algorithm.

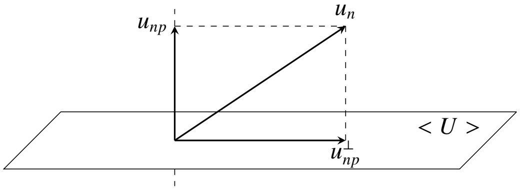

Fig. 2. Orthogonal decomposition.

3.1 Additive Noise Removal

Additive noise is removed by using the orthogonal projection of the noise subspace on the noisy speech. In Fig. 2, orthogonal projection is depicted, where a noise eigenvector $\boldsymbol{u}_{\mathrm{n}}$ is decomposed into two orthogonal projections such that $\boldsymbol{u}_{\mathrm{np}}$ is along <$\boldsymbol{U}$> and $\boldsymbol{u}_{\mathrm{np}}{}^{\perp}$ is perpendicular to <$\boldsymbol{U}$>.

In matrix form, the orthogonal projection $\boldsymbol{U}_{\mathrm{np}}$ of the eigenvector matrix $\boldsymbol{U}_{\mathrm{n}}$ of $\boldsymbol{R_n}$ along $<\boldsymbol{U}>$ is given by (13).

$$\boldsymbol{U}_{\mathrm{np}} = \boldsymbol{P_U} \boldsymbol{U}_{\mathrm{n}} \tag{13}$$

where $\boldsymbol{P_U}$, given by (14), is the orthogonal projection operator whose range is <$\boldsymbol{U}$>.

$$\boldsymbol{P_U} = \boldsymbol{U}(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U})^{-1}\boldsymbol{U}^{\mathrm{T}}. \tag{14}$$

The orthogonal projection operator with range $< \boldsymbol{H}^{\perp} >$ is given by (15)

$$\boldsymbol{P_U}^{\perp} = \boldsymbol{I_U} - \boldsymbol{P_U} \tag{15}$$

where $\boldsymbol{I_U}$ is the identity matrix with dimension same as that of $\boldsymbol{U}$. The autocorrelation matrix of the orthogonal noise with <$\boldsymbol{U}_{\mathrm{np}}$> is given by (16).

$$\boldsymbol{R}_{\mathrm{np}} = \boldsymbol{U}_{\mathrm{np}} \boldsymbol{\Lambda}_{\mathrm{n}} \boldsymbol{U}_{\mathrm{np}}{}^{\mathrm{T}} \tag{16}$$

Subtracting the orthogonal additive noise component from the noisy speech gives the estimate of the speech subspace having only the remaining correlated colored noise:

$$\boldsymbol{R_s} = \boldsymbol{R_x} - \boldsymbol{R}_{\text{np}}. \tag{17}$$

3.2 Colored Noise Removal

In the proposed method, to handle the colored noise, oblique projection of the noise subspace on the speech subspace parallel to the subspace $<\boldsymbol{U}_{\text{np}}>$ is considered. Behrens and Scharf [22] have discussed the applications of oblique projections in signal processing in detail.

The oblique projection matrix $\boldsymbol{E}_{\text{HS}}$ with range $<\boldsymbol{H}>$ and null space $<\boldsymbol{S}>$ is given by (18).

$$\boldsymbol{E}_{\text{HS}} = \boldsymbol{H}(\boldsymbol{H}^{\text{H}}\boldsymbol{P}_{\text{S}}{}^{\perp}\boldsymbol{H})^{-1}\boldsymbol{H}^{\text{H}}\boldsymbol{P}_{\text{S}}{}^{\perp} \tag{18}$$

where $\boldsymbol{P}_{\text{S}}{}^{\perp}$ is the orthogonal projection perpendicular to $<\boldsymbol{S}>$. To complete the null space, $\boldsymbol{A}$ is defined to span $<\boldsymbol{HS}^{\perp}>$ such that $\boldsymbol{E}_{\text{HS}}\boldsymbol{A} = 0$. Similarly, $\boldsymbol{E}_{\text{SH}}$ represents the oblique projection matrix with range $<\boldsymbol{S}>$ and null space $<\boldsymbol{H}>$. Oblique projection is represented in Fig. 3.

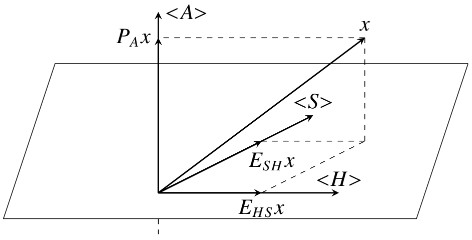

Fig. 3. Oblique projection.

Modifying (18) gives the oblique projection matrix $\boldsymbol{E}_{\boldsymbol{U}_{\text{s}}\boldsymbol{U}_{\text{np}}}$ with range $<\boldsymbol{U}_{\text{s}}>$ and null space $<\boldsymbol{U}_{\text{np}}>$ having orthogonal projection operator $\boldsymbol{P}_{\boldsymbol{U}_{\text{np}}}{}^{\perp}$ whose range is $<\boldsymbol{U}_{\text{np}}{}^{\perp}>$,

$$\boldsymbol{E}_{\boldsymbol{U}_{\text{s}}\boldsymbol{U}_{\text{np}}} = \boldsymbol{U}_{\text{s}}(\boldsymbol{U}_{\text{s}}{}^{\text{H}}\boldsymbol{P}_{\boldsymbol{U}_{\text{np}}}{}^{\perp}\boldsymbol{U}_{\text{s}})^{-1}\boldsymbol{U}_{\text{s}}{}^{\text{H}}\boldsymbol{P}_{\boldsymbol{U}_{\text{np}}}{}^{\perp}. \tag{19}$$

The projection of the noisy speech on the speech subspace along the noise subspace is obtained by (20).

$$\boldsymbol{U}_{\text{nc}} = \boldsymbol{E}_{\boldsymbol{U}_{\text{s}}\boldsymbol{U}_{\text{np}}}\boldsymbol{U}_{\text{n}} \tag{20}$$

This would provide non additive, colored noise present in the speech subspace with autocorrelation matrix as:

$$\boldsymbol{R}_{\text{nob}} = \boldsymbol{U}_{\text{nc}}\boldsymbol{D}_{\text{n}}\boldsymbol{U}_{\text{nc}}{}^{\text{H}}. \tag{21}$$

The PSD of colored noise ($\boldsymbol{\Phi}_{\text{c}}$) can be obtained as the Fourier transform of the autocorrelation obtained from $\boldsymbol{R}_{\text{nob}}$. Variance of the colored noise obtained similar to the variance of noise obtained in [16], is provided in (22) with $\boldsymbol{V}$ being the square of the K point fft of the eigenvalue matrix of $\boldsymbol{R_s}$

$$\sigma_{\text{c}}{}^2 = \boldsymbol{V}^{\text{T}}\boldsymbol{\Phi}_{\text{c}}/K. \tag{22}$$

3.3 Combined Masking Threshold

In this paper, a combination (T_{MC}) of simultaneous and temporal masking is considered, as shown in (23),

$$T_{\text{MC}} = \max(T_{\text{MS}}, T_{\text{MT}}) \tag{23}$$

where T_{MS} is the simultaneous masking threshold [16] and T_{MT} is the temporal masking threshold [23]. T_{MS} at i barks due to the masking component located at j barks is given as

$$T_{\text{MS}}(j,i) = X(j) + O(j) + \text{SF}(j,i) \tag{24}$$

where, $X(j)$ is the sound pressure level in dB of the masking component with critical band index j, $O(j)$ is the threshold offset and $\text{SF}(j,i)$ is the spreading function.

T_{MT} in the m^{th} band with t ms time difference between the masker and the maskee is given by (25)

$$T_{\text{MT}} = a(b - \log_{10} t)(L_{\text{m}} - c) \tag{25}$$

where L_{m} is the masker level in dB obtained by taking the average power of all samples in a particular critical band

$$L_{\text{m}}(i) = 10 \log \frac{1}{N} \sum_{k=1}^{N} s^2(k) \tag{26}$$

where $s^2(k)$ is the power of the samples and N is the number of samples in the particular critical band. The values for the parameters a, b, and c are provided in [23].

Considering the maximum value among the two thresholds would provide the actual masking effect experienced.

3.4 Varinace Normalization

To reduce the effect of any present tones which are caused by increased variance at random frequencies, Maganti and Matassoni [19] performed variance normalization across the critical bands for spectral subtraction speech enhancement algorithm. The variance is computed as in (27)

$$v(m) = \frac{1}{K-1} \sum_{i=1}^{K} (v_i(m) - \widehat{v}(m))^2 \tag{27}$$

where K is the number of bands, m is the frame index, $\widehat{v}$ is the mean and v_i is the i^{th} element. The peaks of noise present in the enhanced speech are suppressed by normalizing them with respect to the maximum value across the bands as in (28),

$$w(m) = \frac{v(m)}{\max\{v(m)\}}. \tag{28}$$

In the previous works on signal subspace speech enhancement, control parameter ν, in the calculation of gain (12) was mostly kept constant and no much attention was given to make it change adaptively in accordance with the noise levels in the signal to be enhanced. In [14], ν was chosen to be 2 and in [16], it was selected as 2 for RQSS and PWSS and 0.8 for PSS. The proposed work provides a novel method to adaptively change the control parameter to reflect the noise content and filter the speech signal accordingly to produce improvement in speech enhancement. It is made to adaptively vary according to the normalized variance of the PSD of the previous frame as given in (29), to smoothen the output and make it more intelligible.

$$v = 1/(w(m))^2, \qquad (29)$$

v made inversely proportional to the square of the normalized variance ensures that it modifies the gain effectively avoiding abrupt variations and providing a smooth output.

4. Results

Clean narrowband (NB)and wideband (WB) speech samples for the evaluation of the proposed algorithm were taken from the TSP database [26] consisting of over 1400 utterances spoken by 12 male and 12 female speakers of the phonetically-balanced harvard sentences [27] with relatively low word-context predictability. For NB, speech samples were sampled at 8 kHz and IRS filtered and for WB, they were sampled at 16 kHz and filtered by 7 kHz low pass filter. Noisy speech samples were created by adding six real-world noises (airport, babble, car, restaurant, station and street) at different SNRs (0 dB, 5 dB and 10 dB) to the clean speech samples using the standard FaNT tool [28]. A hanning window of length 256 was used for the frames in the analysis and the overlap between the frames was taken to be 50 %. The different algorithms compared with the proposed PKLT-OBL algorithm represented by $pklt_{\mathrm{obl}}$ are Wiener filtering [4], spectral subtraction [3], KLT [15], Non-KLT [9], PKLT [16] and PKLT-VRE [25], represented respectively by $wiener$, $specsub$, $kltevd$, $nonklt$, $pklt$ and $pklt_{\mathrm{vre}}$ in the analysis. KLT uses KL Transform for enhancement, Non-KLT is based on the simultaneous diagonalization of the clean speech and noise covariance matrices and PKLT uses simultaneous masking property together with KL Transform. PKLT-VRE includes Variance of the Reconstruction Error (VRE) criterion in PKLT.

The results are shown in terms of spectrograms, objective parameters, namely, d_{WSS}, C_{OVRL}, STOI and d_{LLR} values and subjective intelligibility test.

4.1 Spectrograms

Figure 4 shows the spectrograms corresponding to the clean speech, speech corrupted by 0 dB airport noise and the enhanced output speech by the proposed PKLT-OBL algorithm for NB and WB speech.

From the spectrograms, it can be observed that the proposed algorithm retains most of the speech signals reducing signal distortion and removes noise at the same time.

4.2 d_{WSS}

The weighted spectral slope distance (d_{WSS}) measure has a high correlation with subjective quality ratings and is computed as in (30), where $S_x(j,m)$ and $\bar{S}_{\hat{x}}(j,m)$ denote the spectral slopes of the clean and enhanced signals respectively of the j^{th} band, m^{th} frame and $W(j,m)$ represents the weight [29].

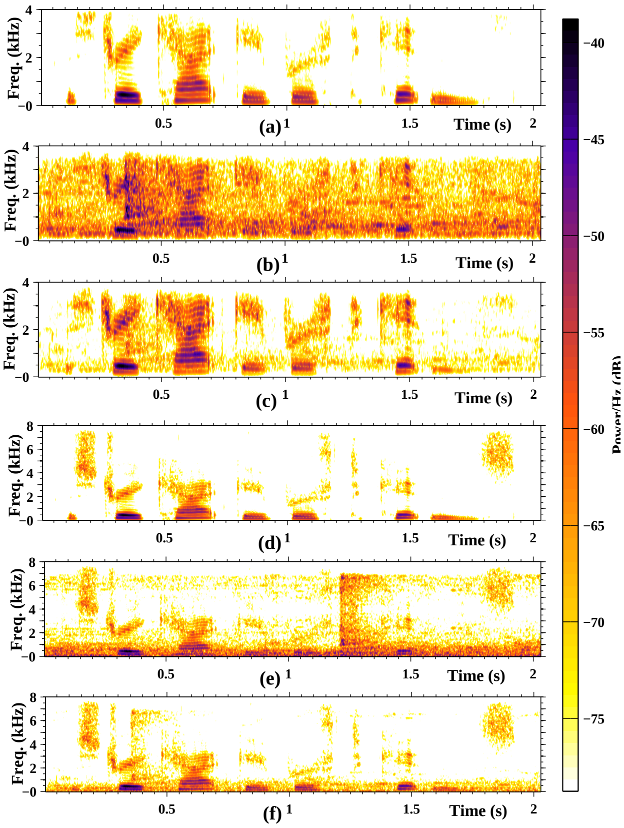

Fig. 4. Spectrograms of (a) NB clean speech, (b) NB noisy speech, (c) NB enhanced speech by PKLT-OBL, (d) WB clean speech, (e) WB noisy speech, (f) WB enhanced speech by PKLT-OBL.

$$d_{\mathrm{WSS}} = \frac{1}{M}\sum_{m=0}^{M-1}\frac{\sum_{j=1}^{25} W(j,m)(S_x(j,m)-\bar{S}_{\hat{x}}(j,m))^2}{\sum_{j=1}^{25} W(j,m)}. \qquad (30)$$

d_{WSS} values shown in Tab. 1 give the lowest values for the proposed PKLT-OBL algorithm in every case. For instance, for NB speech with 0 dB car noise and WB speech with 5 dB airport noise, the d_{WSS} values for PKLT-OBL are the lowest with 81 and 57 respectively.

4.3 Composite Objective Measure

The objective composite measure C_X [29] comprising of C_{SIG}, C_{BAK} and C_{OVRL} were shown to have high correlation to the ITU-T P.835 standard speech quality measures for systems that include noise suppression algorithm. C_X calculated from the basic objective measures (O_k) using multi-variable adaptive regression splines (MARS) technique with γ_k being the MARS coefficient is obtained as in (31). C_{OVRL} is shown to give a high correlation to the overall speech quality and so is provided in Fig. 5 (a) and (b) for NB and WB speech respectively.

Input	Type	d_{WSS} values													
SNR	Of	wiener		specsub		kltevd		nonklt		pklt		pklt$_{vre}$		pklt$_{obl}$	
[dB]	Noise	NB	WB	NB	WB	NB	WB	NB	WB	NB	WB	NB	WB	NB	WB
0	airport	130	90	107	100	121	112	121	112	128	151	117	122	87	89
0	babble	134	108	110	82	121	85	121	85	127	85	113	77	85	55
0	car	123	110	90	105	117	112	117	112	127	113	112	105	81	105
0	restaurant	129	127	102	97	119	105	119	105	127	103	116	97	79	73
0	station	122	113	88	89	109	100	109	100	118	99	106	91	77	73
0	street	125	115	92	96	118	105	118	105	126	104	112	99	75	84
5	airport	102	64	83	60	95	102	96	101	101	132	91	108	66	57
5	babble	108	93	88	68	100	70	100	70	105	69	94	64	67	44
5	car	95	94	74	89	91	97	91	97	98	99	86	92	65	84
5	restaurant	97	108	80	83	87	86	87	85	92	84	84	79	63	58
5	station	99	96	72	76	90	84	90	84	98	83	87	77	63	59
5	street	101	101	70	84	92	87	92	87	97	88	88	84	58	69
10	airport	80	78	70	80	73	87	73	86	76	108	69	91	50	65
10	babble	81	77	66	56	72	57	72	56	75	56	69	52	50	34
10	car	73	80	57	73	71	80	71	80	76	83	68	78	49	66
10	restaurant	78	89	64	68	70	67	70	67	73	66	67	63	48	45
10	station	80	81	58	64	70	67	70	66	76	66	68	62	51	45
10	street	77	84	57	70	71	71	71	70	74	72	67	68	46	55

Tab. 1. d_{WSS} values.

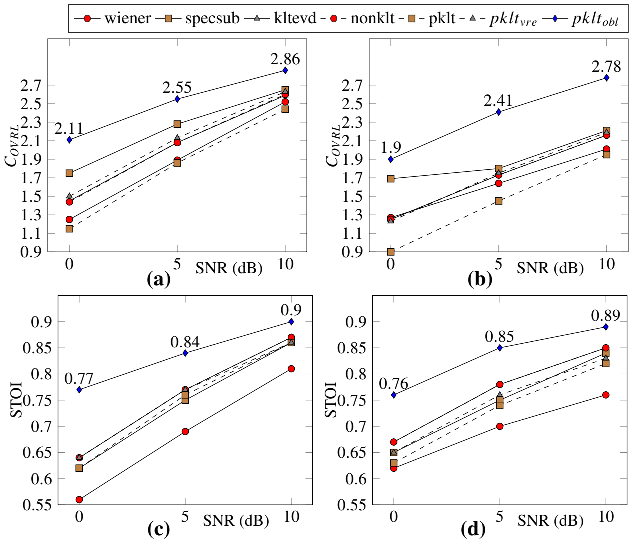

Fig. 5. C_{OVRL} and STOI values.

$$C_X = \gamma_0 + \sum_{k=1}^{5} \gamma_k O_k \tag{31}$$

It can be observed that the proposed PKLT-OBL algorithm has the highest C_{OVRL} value compared to the others, for both NB and WB speech. For instance, the C_{OVRL} of PKLT-OBL for 0 dB NB and WB speech are 2.11 and 1.9 respectively, which are the highest.

4.4 STOI

Short-time objective intelligibility measure (STOI) is based on mean cross-correlations between processed and reference signals across time-frequency cells [30]. The STOI values for NB and WB speech are shown in Fig. 5 (c) and (d) respectively.

The proposed PKLT-OBL algorithm has the highest STOI value compared to the others for NB and WB speech. For instance, the STOI of PKLT-OBL for 5 dB NB and WB speech are the highest with 0.84 and 0.85 respectively.

4.5 d_{LLR}

Log likelihood ratio distance (d_{LLR}) is an LPC based objective measure. The d_{LLR} measure is defined as in (32)

$$d_{\mathrm{LLR}}(\alpha_p, \alpha_c) = \log\left(\frac{\alpha_p \boldsymbol{R_c} \alpha_p^{\mathrm{T}}}{\alpha_c \boldsymbol{R_c} \alpha_c^{\mathrm{T}}}\right) \tag{32}$$

where α_c and α_p are the LPC vectors of the original and the enhanced speech signal frames respectively, and $\boldsymbol{R_c}$ is the auto-correlation matrix of the original speech signal. A lower d_{LLR} indicates a better speech quality.

Table 2 shows that the proposed PKLT-OBL algorithm has the lowest d_{LLR} values in all the cases. For instance, for NB speech with 0 dB babble noise and WB speech with 5 dB car noise, the d_{LLR} values are 0.90 and 0.27 which are the lowest.

4.6 Subjective Intelligibility Test

The subjective intelligibility test (SIT)in anechoic conditions was performed as in [31]. The test set consisted of the outputs of all the seven algorithms used for the analysis for both NB and WB speech, with all the 18 noisy conditions (three different noise levels for each of the six noise types) for 30 sentences from the database. In total, 1080 sets of speech samples, each having the outputs of seven enhancement algorithms were considered for SIT. 34 normal listeners in the

Input	Type	d_{LLR} values													
SNR	Of	wiener		specsub		kltevd		nonklt		pklt		pklt$_{vre}$		pklt$_{obl}$	
[dB]	Noise	NB	WB	NB	WB	NB	WB	NB	WB	NB	WB	NB	WB	NB	WB
0	airport	2.11	0.94	1.26	0.31	1.13	0.38	1.17	0.41	1.36	0.67	1.27	0.41	1.00	0.30
0	babble	1.87	1.35	1.24	0.51	1.04	0.47	1.08	0.48	1.25	0.58	1.21	0.43	0.90	0.23
0	car	2.55	1.27	1.63	0.42	1.28	0.41	1.29	0.45	1.51	0.56	1.44	0.41	1.05	0.33
0	restaurant	1.84	1.61	1.35	0.76	1.10	0.47	1.14	0.49	1.33	0.56	1.25	0.46	0.89	0.40
0	station	3.17	1.64	2.72	0.98	2.47	0.68	2.50	0.70	2.82	0.86	2.45	0.67	2.22	0.54
0	street	2.11	1.45	1.44	0.83	1.19	0.54	1.24	0.56	1.38	0.80	1.38	0.58	1.04	0.39
5	airport	1.67	0.89	1.09	0.26	0.91	0.35	0.95	0.38	1.09	0.58	1.01	0.38	0.72	0.15
5	babble	1.56	1.15	1.09	0.36	0.86	0.30	0.89	0.30	1.05	0.40	1.00	0.27	0.74	0.11
5	car	1.84	1.23	1.26	0.30	0.94	0.35	0.95	0.38	1.14	0.47	1.06	0.34	0.82	0.27
5	restaurant	1.41	1.28	1.08	0.41	0.79	0.36	0.82	0.37	0.97	0.43	0.90	0.33	0.70	0.19
5	station	2.53	1.36	2.26	0.59	2.14	0.48	2.16	0.50	2.43	0.65	2.08	0.50	1.92	0.30
5	street	1.92	1.42	1.33	0.62	1.02	0.42	1.04	0.44	1.21	0.58	1.15	0.42	0.82	0.41
10	airport	1.12	0.91	0.99	0.23	0.68	0.32	0.70	0.36	0.84	0.47	0.78	0.31	0.63	0.12
10	babble	1.19	0.95	0.99	0.23	0.69	0.22	0.70	0.22	0.84	0.29	0.79	0.20	0.64	0.05
10	car	1.36	1.14	1.03	0.25	0.75	0.30	0.76	0.33	0.91	0.40	0.86	0.29	0.68	0.17
10	restaurant	1.05	1.03	0.97	0.27	0.65	0.27	0.67	0.27	0.81	0.35	0.77	0.25	0.61	0.10
10	station	2.08	1.10	1.88	0.37	1.73	0.36	1.75	0.37	2.00	0.46	1.76	0.34	1.66	0.18
10	street	1.34	1.16	1.05	0.37	0.84	0.34	0.86	0.35	0.94	0.43	0.86	0.31	0.65	0.16

Tab. 2. d_{LLR} values.

Algorithm	Intelligibility [%]					
	NB			WB		
	0 dB	5 dB	10 dB	0 dB	5 dB	10 dB
wiener	60.2	75.0	84.1	60.9	80.7	91.2
specsub	58.1	77.4	82.0	61.0	79.1	88.6
kltevd	76.2	82.1	83.8	82.0	83.2	86.3
nonklt	75.0	80.1	85.4	79.2	82.0	90.2
pklt	72.3	79.0	88.1	78.3	82.0	92.1
$pklt_{vre}$	76.0	82.1	89.2	80.1	82.5	90.0
$pklt_{obl}$	83.2	90.1	92.0	89.3	91.8	96.4

Tab. 3. Subjective intelligibility test.

age group of 18-25 (mean 21.4 years) with normal hearing were asked to identify keywords from the sentences played to them. Each listener was provided with 32 sets (16 NB and 16 WB) of speech samples for the test. The average percentage of correctly identified words for each noise level, separately for NB and WB speech is shown in Tab. 3.

As can be seen in Tab. 3, the proposed PKLT-OBL algorithm gives the highest intelligibility percentage for all the noise levels for both NB and WB speech. For instance, PKLT-OBL gives 83.2 % for 0 dB NB speech and 91.8 % for 5 dB WB speech showing better intelligibility compared to the other enhancement methods used in the evaluation.

5. Conclusion

A perceptual subspace speech enhancement method employing EVD and handling colored noise using oblique projection is proposed. Orthogonal projectors removed the additive noise and oblique projection was used to determine the colored noise onto the speech subspace, with a null on the additive noise space. The variance of the colored noise calculated, was used in the SDC estimator of clean speech. Perceptual features were incorporated in it by changing the gain function according to the combined simultaneous and temporal auditory masking threshold. Use of perceptual features reduced the signal distortion of the output by avoiding unnecessary filtering of noise elements that are not perceptible, in effect reducing the filtering of desired speech elements. Variance normalization reduced the occurrence of residual noise by removing abrupt changes in the speech signal adaptively, through the control parameter of the gain of the filter.

The spectrogram clearly shows that the proposed $pklt_{\mathrm{obl}}$ algorithm removes most of the noise and retains most of the clean speech signal from the noisy speech. $pklt_{\mathrm{obl}}$ clearly outperforms the other state of the art speech enhancement methods and other signal subspace approach based methods for both NB and WB speech in terms of quality and intelligibility. Low values for the objective measures d_{WSS} and d_{LLR} and high value for C_{OVRL} show the superior quality and high values for STOI and subjective intelligibility test show the superior intelligibility of the output of $pklt_{\mathrm{obl}}$ over the other existing algorithms.

References

[1] VEISI, H., SAMETI, H. Speech enhancement using hidden Markov models in Mel-frequency domain. *Speech Communication*, 2013, vol. 55, no. 2, p. 205–220. DOI: 10.1016/j.specom.2012.08.005

[2] CHEHREHSA, S., MOIR, T. J. Speech enhancement using maximum a-posteriori and Gaussian mixture models for speech and noise periodogram estimation. *Computer Speech & Language*, 2016, vol. 36, p. 58–71. DOI: 10.1016/j.csl.2015.09.001

[3] BOLL, S. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1979, vol. 27, no. 2, p. 113–120. DOI: 10.1109/TASSP.1979.1163209

[4] SCALART, P., FILHO, J. V. Speech enhancement based on a priori signal to noise estimation. In *Proceedings of the IEEE International*

Conference on Acoustics, Speech, and Signal Processing. Atlanta, GA (USA), 1996, p. 629–632. DOI: 10.1109/ICASSP.1996.543199

[5] MELLAHI, T., HAMDI, R. LPC-based formant enhancement method in Kalman filtering for speech enhancement. *AEU - International Journal of Electronics and Communications*, 2015, vol. 69, no. 2, p. 545–554. DOI: 10.1016/j.aeue.2014.11.007

[6] EPHRAIM, Y., MALAH, D. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1984, vol. 32, no. 6, p. 1109–1121. DOI: 10.1109/TASSP.1984.1164453

[7] VERTELETSKAYA, E., SAKHNOV, K., SIMAK, B. Delay estimator and improved proportionate multi-delay adaptive filtering algorithm. *Radioengineering*, 2012, vol. 21, no. 1, p. 182–189. ISSN: 1805-9600

[8] HILKHUYSEN, G., GAUBITCH, N., BROOKES, M., et al. Effects of noise suppression on intelligibility: Dependency on signal-to-noise ratios. *The Journal of the Acoustical Society of America*, 2012, vol. 131, no. 1, p. 531–539. DOI: 10.1121/1.3665996

[9] HU, Y., LOIZOU, P. C. A generalized subspace approach for enhancing speech corrupted by colored noise. *IEEE Transactions on Speech and Audio Processing*, 2003, vol. 11, no. 4, p. 334–341. DOI: 10.1109/TSA.2003.814458

[10] BEROUTI, M., SCHWARTZ, R., MAKHOUL, J. Enhancement of speech corrupted by acoustic noise. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. 1979, p. 208–211. DOI: 10.1109/ICASSP.1979.1170788

[11] EPHRAIM, Y., MALAH, D. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1984, vol. 32, no. 6, p. 1109–1121. DOI: 10.1109/TASSP.1984.1164453

[12] REZAYEE, A., GAZOR, S. An adaptive KLT approach for speech enhancement. *IEEE Transactions on Speech and Audio Processing*, 2001, vol. 9, no. 2, p. 87–95. DOI: 10.1109/89.902276

[13] LEV-ARI, H., EPHRAIM, Y. Extension of the signal subspace speech enhancement approach to colored noise. *IEEE Signal Processing Letters*, 2003, vol. 10, no. 4, p. 104–106. DOI: 10.1109/LSP.2003.808544

[14] EPHRAIM, Y., VAN TREES, H. L. A signal subspace approach for speech enhancement. *IEEE Transactions on Speech and Audio Processing*, 1995, vol. 3, no. 4, p. 251–266. DOI: 10.1109/89.397090

[15] HU, Y., LOIZOU, P. C. A subspace approach for enhancing speech corrupted by colored noise. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. Orlando, FL (USA), 2002, p. 573–576. DOI: 10.1109/ICASSP.2002.5743782

[16] JABLOUN, F., CHAMPAGNE, B. Incorporating the human hearing properties in the signal subspace approach for speech enhancement. *IEEE Transactions on Speech and Audio Processing*, 2003, vol. 11, no. 6, p. 700–708. DOI: 10.1109/TSA.2003.818031

[17] HU, Y., ZHANG, X., ZHU, F., et al. Image recognition using iterative oblique projection. *Electronics Letters*, 2005, vol. 41, no. 20, p. 1109–1110. DOI: 10.1049/el:20051935

[18] LU, C. T. Reduction of musical residual noise for speech enhancement using masking properties and optimal smoothing. *Pattern Recognition Letters*, 2007, vol. 28, no. 11, p. 1300–1306. DOI: 10.1016/j.patrec.2007.03.001

[19] MAGANTI, H. K., MATASSONI, M. A perceptual masking approach for noise robust speech recognition. *EURASIP Journal on Audio, Speech, and Music Processing*, 2012, p. 1–9. DOI: 10.1186/1687-4722-2012-29

[20] TUFTS, D. W., KUMARESAN, R., KIRSTEINS, I. Data adaptive signal estimation by singular value decomposition of a data matrix. *Proceedings of the IEEE*, 1982, vol. 70, no. 6, p. 684–685. DOI: 10.1109/PROC.1982.12367

[21] DENDRINOS, M., BAKAMIDIS, S., CARAYANNIS, G. Speech enhancement from noise: A regenerative approach. *Speech Communication*, 1991, vol. 10, no. 1, p. 45–57. DOI: 10.1016/0167-6393(91)90027-Q

[22] BEHRENS, R. T., SCHARF, L. L. Signal processing applications of oblique projection operators. *IEEE Transactions on Signal Processing*, 1994, vol. 42, no. 6, p. 1413–1424. DOI: 10.1109/78.286957

[23] SINAGA, F., GUNAWAN, T. S., AMBIKAIRAJAH, E. Wavelet packet based audio coding using temporal masking. In *Proceedings of the Joint Fourth International Conference on Information, Communications and Signal Processing and the Fourth Pacific Rim Conference on Multimedia*. 2003, vol. 3, p. 1380–1383. DOI: 10.1109/ICICS.2003.1292691

[24] HU, Y., LOIZOU, P. C. Subjective comparison and evaluation of speech enhancement algorithms. *Speech Communication*, 2007, vol. 49, no. 7–8, p. 588–601. DOI: 10.1016/j.specom.2006.12.006

[25] SAADOUNE, A., AMROUCHE, A., SELOUANI, S. A. Perceptual subspace speech enhancement using variance of the reconstruction error. *Digital Signal Processing*, 2014, vol. 24, p. 187–196. DOI: 10.1016/j.dsp.2013.09.005

[26] KABAL, P. *TSP Speech Database*. 26 pages. [Online] Cited 2002-09-04. Available at: http://www-mmsp.ece.mcgill.ca/Documents/Downloads/TSPspeech/TSPspeech.pdf

[27] IEEE, IEEE Recommended Practice for Speech Quality Measurements. *IEEE No 297-1969*. June 1969, p. 1–24. DOI: 10.1109/IEEESTD.1969.7405210

[28] HIRSCH, H. G. *FaNT - Filtering and Noise Adding Tool*. 4 pages. [Online] Cited 2005-03-15. Available at: http://dnt.-kr.hsnr.de/download.html

[29] HU, Y., LOIZOU, P. C. Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, 2008, vol. 16, no. 1, p. 229–238. DOI: 10.1109/TASL.2007.911054

[30] TAAL, C. H., HENDRIKS, R. C., HEUSDENS, R., et al. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. Dallas, TX (USA), 2010, p. 4214–4217. DOI: 10.1109/ICASSP.2010.5495701

[31] COOKE, M., MAYO, C., BOTINHAO, C. V., et al. Evaluating the intelligibility benefit of speech modifications in known noise conditions. *Speech Communication*, 2013, vol. 55, no. 4, p. 572–585. DOI: 10.1016/j.specom.2013.01.001

About the Authors ...

Sudeep SURENDRAN was born in Kerala State, India, in 1987. He received his M. Tech. in Communication Engineering from The University of Calicut in 2013. He is currently a research scholar with the department of E.C.E, N.I.T. Warangal, India. His research interests are in the area of speech signal processing.

T. Kishore KUMAR was born in the state Andhra Pradesh, India, in 1971. He received his Ph.D. in Digital Signal Processing from JNTU Hyderabad in 2004. He is currently working as an associate professor and Head, in the department of E.C.E., N.I.T. Warangal, India. His research interests include Speech Signal Processing, Adaptive Signal processing etc.