Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 54 (2015) 818 – 828

Procedia Computer Science

Eleventh International Multi-Conference on Information Processing-2015 (IMCIP-2015)

Perceptual Subspace Speech Enhancement with Variance Normalization

Sudeep Surendran* and T. Kishore Kumar

National Institute of Technology Warangal, Warangal 506 004, Telangana, India

Abstract

In this paper a perceptual subspace speech enhancement method using masking property of human auditory system with variance normalization is presented. The masking property of the human auditory system is used while deciding the gain parameters for the algorithm. Spectral Domain Constrained estimator was employed in determining the filter coefficients and colored noise was handled by replacing the noise variance by Rayleigh quotient. Variance normalization is further done to remove the spikes in the values so as to avoid abrupt increase or decrease in power of the output samples making the output more intelligible. The objective measures SNR_{Loss} and SNR_{LESC} were chosen for performance evaluation based on their efficiency in determining the intelligibility of the output. The results show an improved performance of the proposed method over some of the existing speech enhancement methods in terms of intelligibility.

Peer-review under responsibility of organizing committee of the Eleventh International Multi-Conference on Information Processing-2015 (IMCIP-2015)

Keywords: Intelligibility; Perceptual features; Signal subspace approach; Speech enhancement; Variance normalization.

1. Introduction

Speech enhancement is the improvement in the quality and intelligibility of noise corrupted speech signals by using various techniques for removing noise from it. Speech enhancement is commonly used as a pre-processing block in a lot of applications like automatic speech recognizer and other communication systems. The performance of speech communication systems can be improved using various speech enhancement techniques like spectral subtraction, adaptive wiener filtering, model based methods etc. Spectral subtraction has been widely used for enhancing speech because of its simplicity and ease of implementation in single channel systems but it suffers from the production of musical noise after enhancement and is one of its major drawbacks. The use of perceptual features in speech enhancement has been the latest trend in the field. Masking property of human auditory system makes the noise of a particular band of frequency inaudible to the listener if it falls below the masking threshold of that particular frequency. One of the earliest works utilizing perceptual features was done by Johnston[1].

Signal distortion is considered as another important issue in speech enhancement. There has always been an effort to develop speech enhancement techniques that give good compromise between residual noise and signal distortion

*Corresponding author. Tel.: +91-833-190-3362.
E-mail address: sudipsuren@nitw.ac.in

Peer-review under responsibility of organizing committee of the Eleventh International Multi-Conference on Information Processing-2015 (IMCIP-2015)
doi:10.1016/j.procs.2015.06.096

of the output signal. Signal subspace approach (SSA),[2–4,6] have shown to give a better compromise between the two compared to the other existing techniques. The use of perceptual features in subspace method by Jabloun and Champagne[12] had shown a reduction in residual noise compared to the conventional signal subspace methods. Variance normalization was used in spectral subtraction speech enhancement algorithm by Maganti and Matassoni[5] 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.

This paper proposes a speech enhancement algorithm which reduces speech distortion and increases speech intelligibility. For the purpose, signal subspace method utilizing the perceptual features and variance normalization is employed. The use of perceptual features reduces signal distortion and variance normalization reduces abrupt changes in the output making it more intelligible.

The rest of the paper is arranged as follows. In section 2, the subspace method of speech enhancement is explained. Section 3 explains the perceptual subspace method of speech enhancement. Section 4 shows how variance normalization is done. Section 5 mentions the steps involved in the proposed method. Section 6 explains about the performance evaluation. Section 7 gives the results and section 8 provides the conclusion.

2. Signal subspace method of speech enhancement

Signal subspace approach of speech enhancement used by Ephraim and Van trees[2] employed Eigen Value Decomposition (EVD) and used Karhunen-Loeve Transform (KLT) to project clean speech into signal + noise subspace called the signal subspace and removed noise which falls in the orthogonal noise subspace. Further, the elements of the signal subspace were processed separately to remove any elements of noise from it using a diagonal gain matrix based on the uncorrelated nature of the coefficients in the subspace. The gain matrix elements were decided based on estimators like Time domain Constrained (TDC) or Spectral Domain Constrained (SDC) estimators. Then inverse KLT was applied to get the enhanced output speech signal which gave a better compromise between the signal distortion and residual noise compared to the other exiting speech enhancement methods.

One of the important features of SSA is dimensionality reduction which is achieved by reducing the rank of the noise corrupted data matrix by forcing it back to that of the uncorrupted signal. Further, with proper tuning of the parameters like window size, rank of the matrix etc, SSA offers a better compromise between signal distortion and residual noise level over other speech enhancement methods.

The noisy signal (y) composed of the clean speech signal (s) and the additive noise (w) can be represented as in (1)

$$y = s + w \tag{1}$$

The covariance matrix R_x of the noisy speech can be represented as

$$R_\chi = R_s + R_w \tag{2}$$

where R_s and R_w are the covariance matrices of clean speech and noise respectively with R_x assumed to have a higher rank than $R_s \cdot R_x$ and R_s are toeplitz matrices, the nature of which was well studied by Gray[7] The EVD of R_x and R_s is given by (3) and (4) respectively

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

$$R_s = U_P \Lambda_s U_P^H \tag{4}$$

where Λ and Λ_s given by (5) and (6) represent the diagonal matrices of the Eigen values respectively of R_x and R_s

$$\Lambda = \text{diag}(\lambda_1, \lambda_2 \ldots\ldots \lambda_Q) \tag{5}$$

$$\Lambda_s = \text{diag}(\lambda_{s,1r} \lambda_{s,2} \ldots\ldots \lambda_{s,P}) \tag{6}$$

The dimension of U is Q and that of U_P is P such that $Q > P$. Also, R_w is given by:

$$R_w = \sigma^2 I \tag{7}$$

where σ^2 is the variance of noise and I represents identity matrix. Thus from (4) and (7), R_x can be represented as

$$R_x = U(\Lambda_s + \sigma^2 I)U^H \tag{8}$$

Here, $U = [U_P U_{Q-P}]$ where, $U_P = [u_1 u_2, \ldots\ldots u_P]$ spans the signal subspace and $U_{Q-P} = [u_{P+1}, u_{P+2}, \ldots u_Q]$ spans the noise subspace where u_i represents the Eigen vector corresponding to the Eigen value λ_i.

A linear filter H is designed so as to separate the signal subspace from the noise subspace. Thus,

$$\hat{s} = Hx \tag{9}$$

The residual error is

$$r = \hat{s} - s = Hx - Is \tag{10}$$

Then,

$$r = Hs - Hw - Is = (H - I)s + Hw = r_s + r_w \tag{11}$$

where r_s represents the signal distortion given by (12) and r_w represents the residual noise given by (13)

$$r_s = (H - I)s \tag{12}$$

$$r_w = Hw \tag{13}$$

2.1 Estimation of H

Various optimization criteria can be employed for calculating H, which includes the following:

1. Least squares estimator (LSE)
   LSE minimizes the squared fitting error between the observation vector x and a linear low order model.
2. Linear Minimum mean square error estimator (LMMSE)
   In LMMSE, the residual error energy is minimized to get optimum value for H.
3. Time Domain Constrained (TDC) estimator.
   TDC minimizes signal distortion subject to keeping the residual noise energy within a limit.
4. Spectral Domain Constrained Estimator (SDC)
   In SDC, H 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 given by (14). In this paper SDC is considered

$$\min_H E\{\|r_s\|^2\} \quad \text{subject to} \begin{cases} E\{|u_i^H r_w|^2\} \le \alpha_i \sigma^2 & \text{for} \quad 1 \le i \le P \\ E\{|u_i^H r_w|^2\} = 0 & \text{for} \quad P < i \le Q \end{cases} \tag{14}$$

where α_i, is a set of non-negative constants[2]. The solution of matrix H is given by (15)

$$H = U_P G U_P^H \tag{15}$$

where U_P^H is called the Karhunen-Loeve Transform (KLT) and G is the gain matrix given by (16)

$$G = \text{diag (gain value corresponding to each } u_i) = g_i = e^{-\upsilon\sigma^2/\lambda_{s,i}} \quad \text{for} \quad i = 1, 2, \ldots P \tag{16}$$

where $\boldsymbol{v}$ is a control parameter.

2.2 Considering colored noise

Some of the methods employed for handling the case of colored noise[2,3,8–11] in subspace method are:

1. Pre-whitening and De-whitening.
   In this approach proposed[2], KLT is applied on the input speech by assuming that the noise is white. To deal with colored noise, the noise is converted to white by multiplying the entire signal by the whitening matrix $R_W^{-1/2}$ and later on dewhitening the enhanced signal by multiplying it with the de-whitening matrix $R_W^{-1/2}$.
2. Using a common diagonalization matrix.
   A common diagonalization matrix $R_w^{-1}R_s$, for both noise and speech, was proposed by Yi Hu and Loizou[8]. The EVD of $R_w^{-1}R_s$ gives
   $$R_w^{-1}R_s = U_C \Lambda_C U_C^T \tag{17}$$
   such that
   $$U_C^T R_s U_C = \Lambda_s \tag{18}$$
   $$U_C^T R_w U_C = I \tag{19}$$
   where, U_C and Λ_C are the Eigen vector and Eigen *value* matrices of $R_w^{-1}R_s \cdot U_C^T$ represents the transpose of U_C.
3. Rayleigh quotient method
   Rayleigh quotient method used by Jabloun *et al.*[12] replaced the variance of the noise by Rayleigh quotient. This method is found to shape the noise better than the other existing methods and reduces the computational load as well and hence is employed in this paper. The noise variance, σ^2, is taken as the noise energy in the direction of the i^{th} Eigenvector, which is the Rayleigh Quotient ξ associated with the i^{th} Eigen vector u_i of $\hat{R}_s$ and R_w given by
   $$\xi_i = u_i^T R_w u_i \tag{20}$$
   which in matrix notation gives
   $$\sigma^2 = \xi = \frac{1}{K} U^T \Phi_w \tag{21}$$
   where Φ_w is the power spectral density estimate of noise.

3. Perceptual Subspace Method

Perceptual features have been used in speech enhancement to reduce the signal distortion and improve intelligibility[12–17]. Initial work in the area was done by Johnston[1] for coding of audio signals. Use of masking property as the perceptual feature is based on the fact that within a critical band of frequency, one sound having a greater magnitude masks the other with lesser magnitude. In the bark scale, one bark covers a critical band and hence Eigen domain to Bark scale conversion has to be performed for efficiently including the masking property in subspace method. For this conversion, Eigen to frequency domain conversion followed by frequency to bark domain conversion is done.

For the conversion of Eigen to frequency domain the following equations are used

$$\Phi_B = \frac{1}{Q} \sum_{i=1}^{P} \lambda_i v_i \tag{22}$$

where Φ_B is Blackman-Tukey Spectrum Estimator and v_i is given by (23) for a K point DFT

$$v_i(k) = \left| V_i \left(\frac{2\pi k}{K} \right) \right|^2 \quad \text{for} \quad k = 0, 1, 2 \ldots K \tag{23}$$

where $V_i(\omega)$ is given by (24)

$$V_i(\omega) = \sum_{q=1}^{Q-1} u_i(q) e^{-j\omega q} \tag{24}$$

In matrix form (22) can be represented by (25)

$$\Phi_B = \frac{1}{Q} v\lambda \tag{25}$$

Φ_B is then used to calculate the masking threshold Φ_{lhr} as given by Jabloun *et al.*[12] which is described below.

3.1 Steps in the calculation of masking threshold

First, frequency (f) to bark scale (z) conversion is done using (26)

$$z(f) = 13 \arctan(0.0076f) + 35 \arctan\left[\left(\frac{f}{7500}\right)^2\right] \tag{26}$$

The masking threshold at i barks due to the masking component located at j barks of tonal and nontonal component are given by (27) and (28) respectively

$$T_{tm}(j,i) = X_{tm}(j) + O_{tm}(j) + SF(j,i) \tag{27}$$

$$T_{nm}(j,i) = X_{nm}(j) + O_{nm}(j) + SF(j,i) \tag{28}$$

Masking components below the masking threshold are discarded, reducing distortion.

$X_{tm}(j)$ is the sound pressure level in dB of the masking component with critical band index j which is given by (29)

$$X\left(z\left(\frac{F_s k}{K}\right)\right) = \Phi_B(k) \tag{29}$$

$O_{tm}(j)$ and $O_{nm}(j)$ are the threshold offsets given by (30) and (31) respectively.

$$O_{tm}(j) = -1.525 - 0.275j - 4.5\,dB \tag{30}$$

$$O_{nm}(j) = -1.525 - 0.175j - 0.5\,dB \tag{31}$$

$SF(j,i)$ is the spreading function given by

$$SF(j,i) = \begin{cases} (17(dz+1) - 0.4X(j) - 6\,dB & -3 \le dz < -1 \\ (0.4X(j)+6)dz\,dB & -1 \le dz < 0 \\ -17\,dz\,dB & 0 \le dz < 1 \\ -(dz-1)(17 - 0.15X(j)) - 17\,dB & 1 \le dz < 8 \end{cases} \tag{32}$$

Φ_{thr} is calculated from (27) and (28).

The perceptual features are then converted to Eigen domain using

$$\theta = \frac{1}{K} V^T \Phi_{thr} \tag{33}$$

Modified gain matrix G is thus obtained as

$$G = e^{-\upsilon\sigma^2/\max(\lambda,\theta)} \tag{34}$$

The KLT using perceptual feature as mentioned in the above subsection is represented by PKLT.

4. Variance Normalization

To smoothen the output of the enhancement algorithm Ching Ta Lu[17], employed an optimal smoothing factor, adapted by the variation of signal to spectral deviation ratio (SSDR) in successive frames. To reduce the effect of any present tones which are caused by increased variance at random frequencies, Maganti *et al.*[5] performed variance normalization across the critical bands for spectral subtraction speech enhancement algorithm. This variance normalization[5] is used in this paper to smoothen the output. The variance is computed as

$$v(m) = \frac{1}{K-1}\sum_{i=1}^{K}(v_i(m) - \hat{v}(m))^2 \tag{35}$$

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

$$w(m) = \frac{v(m)}{\max\{v(m)\}} \tag{36}$$

$w(m)$ gives the normalized values which are then multiplied with the energies to obtain a smoother output

$$\hat{Y}_k(m) = Y_k(m)w(m) \tag{37}$$

where $Y_k(m)$ is the energy of the m^{th} frame and $\hat{Y}_k(m)$ is the normalized energy of the m^{th} frame.

5. Steps in the Proposed Algorithm

The following steps are involved in the proposed speech enhancement method called perceptual KLT with variance normalization represented by PKLTV.

1. Calculation of noise covariance matrix R_w and noisy speech covariance matrix R_x by considering that the first
2. Estimation of the speech covariance matrix R_s using equation (2).
3. Performing the Eigen value decomposition of R_x and R_s to get U, U_P, Λ and Λ_s using equation (3) and (4) after determining the order P from the number of Λs that are greater than 0.
4. Calculation of the power spectral density using (25).
5. Computation of the auditory masking threshold using (26)–(32).
6. Variance normalization is done using (35)–(37).
7. Conversion of the perceptual features to Eigen domain using (33).
8. Calculation of gain matrix G using (34) with a very low value for v and $\sigma^2 = \xi$.
9. Estimation of the enhanced speech signal $\hat{s}$ by multiplying the noisy speech subframes with H as in (9).

6. Performance Evaluation

The purpose of the proposed algorithm is to improve the intelligibility of the output speech. Hence the performance evaluation was done using the parameters which gave a clear inference about the intelligibility. Work done by Hu and Loizou18 gives a comparison of different objective measures used for performance evaluation. In this paper, the objective parameters provided by Maa and Loizou[19] namely SNR_{LOSS} and SNR_{LESC} were used which gives better measure of intelligibility compared to the other objective measures. The lesser the values of these parameters, greater is the performance of the algorithm in terms of intelligibility.

6.1 SNR_{LOSS}

$\mathrm{SNR}_{\mathrm{LOSS}}$ in the band j and frame m is defined as

$$L(j,m) = \mathrm{SNR}_x(j,m) - \mathrm{SNR}_{\hat{x}}(j,m) \tag{38}$$

where $\mathrm{SNR}_x(j,m)$ and $\mathrm{SNR}_{\hat{x}}(j,m)$ are the SNRs of the j^{th} frequency band of the m^{th} frame of the input signal and the enhanced signal respectively.

The limited value of $L(j,m)$ used due to the dependence of $L(j,m)$ on input SNR values is given by

$$\hat{L}(j,m) = \min(\max(L(j,m), -\mathrm{SNR}_{\mathrm{Lim}}), \mathrm{SNR}_{\mathrm{Lim}}) \tag{39}$$

where $[-\mathrm{SNR}_{\mathrm{Lim}}, -\mathrm{SNR}_{\mathrm{Lim}}]$ is the restricted SNR range.

L is mapped to the range [0, 1] using the equation

$$\mathrm{SNR}_{\mathrm{LOSS}}(j,m) = \begin{cases} -\frac{C_-}{\mathrm{SNR}_{\mathrm{Lim}}}\hat{L}(j,m) & \text{if } \hat{L}(j,m) < 0 \\ \frac{C_+}{\mathrm{SNR}_{\mathrm{Lim}}} & \text{if } \hat{L}(j,m) \geq 0 \end{cases} \tag{40}$$

C_- and C_+ are the parameters controlling the slopes of the mapping function. The average $\mathrm{SNR}_{\mathrm{LOSS}}$ is given by

$$\overline{\mathrm{SNR}_{\mathrm{LOSS}}} = \frac{1}{M}\sum_{m=0}^{M-1} f\ \mathrm{SNR}_{\mathrm{LOSS}}(m) \tag{41}$$

where

$$f\ \mathrm{SNR}_{\mathrm{LOSS}}(m) = \frac{\sum_{j=1}^{k} W(j)\mathrm{SNR}_{\mathrm{LOSS}}(i,m)}{\sum_{j=1}^{k} W(j)} \tag{42}$$

where $W(j)$ is the weight used.

6.2 SNR_{LESC}

The excitation spectral correlation (ESC) measure at frame m is computed as follows:

$$r^2(m) = \frac{\left(\sum_{k=1}^{K} X(k,m)\hat{X}(k,m)\right)^2}{\sum_{k=1}^{K} X^2(k,m)\hat{X}^2(k,m)} \tag{43}$$

K is the number of bands ($K = 25$ in our study).

$$\mathrm{ESC} = \frac{1}{M}\sum_{m=1}^{M} r^2(m) \tag{44}$$

Combining the $\mathrm{SNR}_{\mathrm{LOSS}}$ and ESC measures gives the parameter $\mathrm{SNR}_{\mathrm{LESC}}$ as in

$$\mathrm{SNR}_{\mathrm{LESC}}(m) = (1 - r^2(m)) f\ \mathrm{SNR}_{\mathrm{LOSS}}(m) \tag{45}$$

The speech segments were divided into three level regions namely the high level, mid-level and low level regions and ESC was calculated separately for these regions as described by Loizou *et al.*[19] The ESC measures were used to get three different $\mathrm{SNR}_{\mathrm{LESC}}$ values denoted as $\mathrm{SNR}_{\mathrm{LESC}}$ High, $\mathrm{SNR}_{\mathrm{LESC}}$ Mid and $\mathrm{SNR}_{\mathrm{LESC}}$ Low obtained for the high, mid and low level segments respectively. $\mathrm{SNR}_{\mathrm{LESC}}$ Mid gives an approximation of the entire $\mathrm{SNR}_{\mathrm{LESC}}$ values and hence in this paper, $\mathrm{SNR}_{\mathrm{LESC}}$ Mid values are calculated and tabulated for different algorithms.

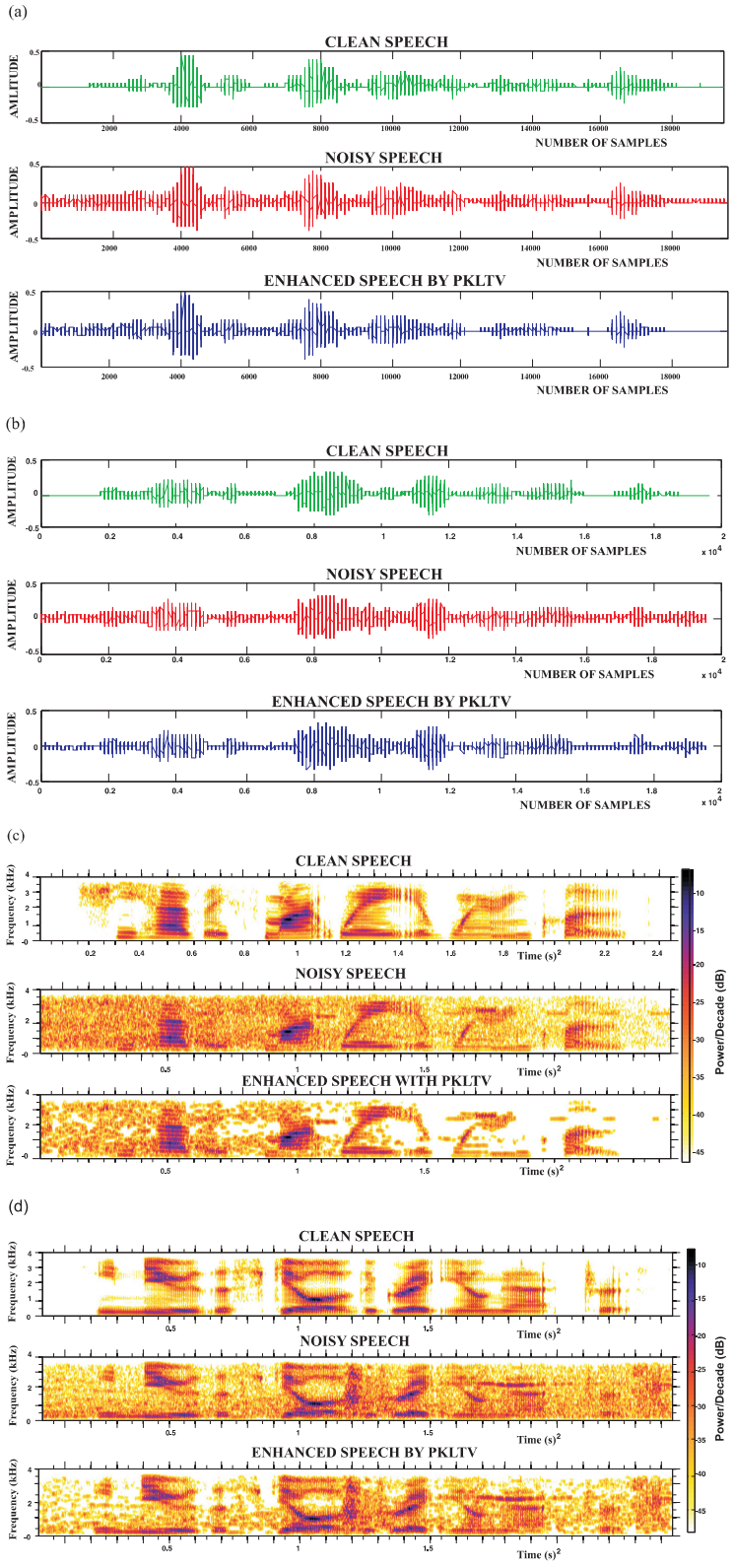

Fig. 1. (a) Waveforms for a female speaker; (b) Waveforms for a male speaker; (c) Spectrograms for the female speaker; (d) Spectrograms for the male speaker.

7. Results

The speech samples used for evaluation of the proposed algorithm were taken from the NOIZEUS database[20]. Clean Female and male speech samples and their corresponding corrupted samples by various noises, sampled at the rate of 8000 kz with 16 bits/sample, were taken for evaluating the performance of the proposed algorithm. Seven different noise environments considered were airport, babble, car, exhibition, restaurant, station and street for 0 dB, 5 dB, 10 dB and 15 dB SNRs. A hanning window of length 256 was used for the frames in the analysis and the overlap between frames is 50%. A subframe window of length 128 is used in the calculation of H as mentioned in step 9 of section 5. The different algorithms compared with the proposed algorithm are spectral subtraction, KLT, non-KLT and PKLT represented respectively by specsub, klt, nonklt and pklt in the analysis. The proposed algorithm (PKLTV) is represented by pkltv.

The results are shown in terms of:

1. *Waveforms and spectrograms*
2. *Comparative tables for* $\mathrm{SNR_{LOSS}}$ *and* $\mathrm{SNR_{LESC}Mid}$ *values*.

7.1 Waveforms and spectrograms

Figure 1(a) and 1(b) shows the waveforms corresponding to the clean speech, speech corrupted by 5dB street noise and the enhanced speech by PKLTV for a female speaker and a male speaker respectively. Figure 1(c) and 1(d) shows the respective spectrograms for Figures 1(a) and 1(c) respectively.

Table 1. $\mathrm{SNR_{LOSS}}$ of various algorithms used.

INPUT		$\mathrm{SNR_{LOSS}}$									
SNR	TYPE OF	FEMALE					MALE				
(dB)	NOISE	specsub	kltevd	nonklt	pklt	pkltv	specsub	kltevd	nonklt	pklt	pkltv
0	airport	0.8455	0.8404	0.8645	0.8449	0.7920	0.0068	0.8604	0.8645	0.8661	0.8296
0	babble	0.8788	0.8719	0.8760	0.8779	0.8390	0.8640	0.8439	0.8447	0.8407	0.8651
0	car	0.8879	0.8759	0.8796	0.8781	0.8462	0.9067	0.8830	0.8868	0.8766	0.8413
0	exhibition	0.8741	0.8466	0.8486	0.8502	0.8668	0.8720	0.8465	0.8497	0.8455	0.8526
0	restaurant	0.8700	0.9192	0.9227	0.9184	0.8533	0.8919	0.9094	0.9155	0.9092	0.8361
0	station	0.8806	0.9123	0.9175	0.9139	0.8399	0.9137	0.9002	0.9063	0.9032	0.8325
0	street	0.9183	0.9734	0.9741	0.9662	0.8933	0.9214	0.9380	0.9382	0.9308	0.8600
5	airport	0.8039	0.8044	0.8106	0.8177	0.7638	0.8558	0.7933	0.7994	0.8028	0.7852
5	babble	0.8338	0.8430	0.8479	0.8453	0.8338	0.8354	0.8015	0.8063	0.8076	0.8354
5	car	0.8430	0.8195	0.8193	0.8217	0.8045	0.8622	0.8017	0.8075	0.8035	0.7810
5	exhibition	0.8206	0.8061	0.8096	0.8098	0.7868	0.8436	0.7840	0.7871	0.7907	0.7937
5	restaurant	0.7505	0.6978	0.7016	0.7096	0.6924	0.8115	0.7995	0.8062	0.8093	0.7668
5	station	0.8523	0.8177	0.8177	0.8232	0.8083	0.8609	0.8058	0.8084	0.8106	0.8039
5	street	0.8940	0.9104	0.9093	0.9015	0.8235	0.8693	0.8250	0.8308	0.8194	0.7726
10	airport	0.7865	0.7063	0.7145	0.7370	0.7021	0.7625	0.7068	0.7125	0.7305	0.7061
10	babble	0.7504	0.6969	0.7016	0.7159	0.6998	0.7740	0.7108	0.7141	0.7253	0.7032
10	car	0.8092	0.7461	0.7535	0.7611	0.6973	0.8193	0.7334	0.7354	0.7484	0.7075
10	exhibition	0.7921	0.7778	0.7819	0.7800	0.7122	0.7507	0.6963	0.6988	0.7121	0.6821
10	restaurant	0.7736	0.7564	0.7618	0.7709	0.7001	0.7393	0.7263	0.7343	0.7450	0.6834
10	station	0.7274	0.6981	0.7075	0.7257	0.6523	0.8117	0.7701	0.7726	0.7716	0.7228
10	street	0.7830	0.7548	0.7568	0.7623	0.6910	0.7742	0.7021	0.7125	0.7260	0.6911
15	airport	0.6939	0.6124	0.6206	0.6492	0.5937	0.7187	0.6501	0.6580	0.6850	0.6133
15	babble	0.6978	0.6201	0.6271	0.6455	0.5779	0.7041	0.6275	0.6314	0.6484	0.6285
15	car	0.7156	0.6353	0.6401	0.670	0.6153	0.7389	0.6478	0.6507	0.6750	0.6277
15	exhibition	0.6935	0.6182	0.6219	0.6394	0.6270	0.7204	0.6495	0.6541	0.6725	0.6459
15	restaurant	0.6729	0.6262	0.6367	0.6647	0.5615	0.6951	0.6261	0.6313	0.6548	0.6035
15	station	0.7667	0.7163	0.7193	0.9381	0.6313	0.6791	0.6214	0.6294	0.8869	0.5786
15	street	0.7773	0.8002	0.7996	0.7801	0.6657	0.7368	0.7023	0.7037	0.7054	0.6501

Table 2. SNR_{LESC}Mid values of various algorithms used.

INPUT SNR (dB)	TYPE OF NOISE	SNR_{LESC}Mid FEMALE specsub	kltevd	nonklt	pklt	pkltv	MALE specsub	kltevd	nonklt	pklt	pkltv
0	airport	0.4484	0.4538	0.4817	0.4674	0.4314	0.4794	0.4806	0.4817	0.4932	0.4964
0	babble	0.5613	0.5640	0.5687	0.5853	0.5209	0.4180	0.3803	0.3786	0.3861	0.4335
0	car	0.4769	0.5410	0.5430	0.5454	0.5800	0.4025	0.3475	0.3486	0.3545	0.4145
0	exhibition	0.6432	0.6040	0.6053	0.6134	0.6099	0.3203	0.3241	0.3252	0.3128	0.3557
0	restaurant	0.5146	0.6358	0.6375	0.6687	0.5530	0.4490	0.4558	0.4573	0.4680	0.4182
0	station	0.5001	0.5246	0.5260	0.5190	0.4964	0.3694	0.3673	0.3680	0.3704	0.3957
0	street	0.4517	0.6048	0.6057	0.6579	0.5581	0.2553	0.3863	0.3857	0.3925	0.3360
5	airport	0.3442	0.3239	0.3252	0.3366	0.3500	0.3796	0.3030	0.3137	0.2943	0.2530
5	babble	0.3600	0.4020	0.4029	0.3929	0.3600	0.2087	0.2006	0.2011	0.2050	0.2087
5	car	0.2689	0.2714	0.2720	0.2718	0.3278	0.1541	0.1493	0.1515	0.1505	0.1824
5	exhibition	0.2304	0.2194	0.2196	0.2203	0.2570	0.1357	0.1178	0.1171	0.1228	0.1691
5	restaurant	0.2771	0.2307	0.2313	0.2300	0.2393	0.2083	0.2257	0.2287	0.2164	0.2109
5	station	0.3255	0.2934	0.2928	0.3020	0.3060	0.2479	0.2016	0.2039	0.2000	0.2126
5	street	0.3675	0.4133	0.4130	0.4311	0.2809	0.3388	0.2957	0.3097	0.2832	0.2086
10	airport	0.2655	0.1950	0.1968	0.1920	0.1803	0.1010	0.0734	0.0734	0.0804	0.0861
10	babble	0.1738	0.1809	0.1805	0.1788	0.1884	0.1147	0.0749	0.0759	0.0802	0.0984
10	car	0.1314	0.1160	0.1201	0.1155	0.1314	0.0602	0.0492	0.0491	0.0494	0.0810
10	exhibition	0.1210	0.0955	0.0957	0.0945	0.0981	0.0443	0.0326	0.0328	0.0336	0.0370
10	restaurant	0.2390	0.2191	0.2222	0.2209	0.2074	0.1103	0.1056	0.1065	0.1090	0.1053
10	station	0.1472	0.1601	0.1623	0.1630	0.1576	0.0595	0.0503	0.0501	0.0508	0.0589
10	street	0.1788	0.1840	0.1856	0.1753	0.1494	0.0712	0.0476	0.0501	0.0518	0.0514
15	airport	0.0830	0.0572	0.0559	0.0591	0.0616	0.0491	0.0331	0.0336	0.0354	0.0383
15	babble	0.0496	0.0403	0.0409	0.0393	0.0385	0.0358	0.0255	0.0249	0.0265	0.0303
15	car	0.0531	0.0430	0.0429	0.0446	0.0519	0.0453	0.0205	0.0205	0.0216	0.0244
15	exhibition	0.0531	0.0328	0.0323	0.0342	0.0409	0.0151	0.0126	0.0128	0.0132	0.0136
15	restaurant	0.1036	0.0901	0.0912	0.0919	0.0804	0.0356	0.0195	0.0196	0.0202	0.0210
15	station	0.0932	0.0806	0.0804	0.6533	0.0486	0.0290	0.0228	0.0237	0.4409	0.0192
15	street	0.0422	0.0499	0.0495	0.0396	0.0247	0.0229	0.0209	0.0210	0.0205	0.0235

7.2 Comparative tables for SNR_{LOSS} and SNR_{LESC}Mid

Table 1 gives the SNR_{LOSS} values of different speech enhancement algorithms under various noisy conditions for both female and male speakers.

Table 2 gives the SEGLESC-Mid values of different speech enhancement algorithms under various noisy conditions for both female and male speakers.

From the waveforms and spectrograms figure, it can be observed that the proposed algorithm removes a lot of noise from the noisy speech to give an enhanced speech.

The Tables 1 and 2 show that the proposed algorithm has low SNR_{LOSS} and SNR_{LESC}Mid values in a majority of conditions which indicate that it outperforms some of the existing speech enhancement algorithms in terms of intelligibility measure which is the main aim of the proposed algorithm. The results obtained by the objective measures are supported by informal listening tests.

8. Conclusion

A perceptual subspace speech enhancement approach with variance normalization is proposed. In this paper, Eigen Value Decomposition was used for the purpose. Perceptual features were included in the subspace method by determining the auditory masking threshold at all frequency bands and changing the gain function according to it. Conversion from Eigen domain to bark domain and back was done for incorporating the perceptual feature in the enhancement algorithm. Finally variance normalization was done to the output. Use of perceptual features clearly added to the performance of the speech enhancement system in terms of SNR_{LOSS} and SNR_{LESC} measure. The use of variance normalization improved the intelligibility over using the perceptual features alone, as is evident from its

lower SNR loss and SNR_{LESC} values. It provided a smooth output in terms of intelligibility since it normalized abrupt changes in the output values. Informal subjective tests support the objective measures.

References

[1] James D. Johnston, Transform Coding of Audio Signals using Perceptual Noise Criteria, *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 2, February (1988).

[2] Y. Ephraim and H. L. Van Trees, A Signal Subspace Approach for Speech Enhancement, *IEEE Transaction on Speech and Audio Processing*, vol. 3, no. 4, pp. 251–266, July (1995).

[3] U. Mittal and N. Phamdo, Signal/noise KLT based Approach for Enhancing Speech Degraded by Colored Noise, *IEEE Transactions Speech Audio Processing*, vol. 8, pp. 159–167, March (2000).

[4] Chang Huai You, Susanto Rahardja and Soo Ngee Koh, Audible Noise Reduction in Eigen Domain for Speech Enhancement, *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 6, August (2007).

[5] Hari Krishna Maganti and Marco Matassoni, A Perceptual Masking Approach for Noise Robust Speech Recognition, *EURASIP Journal on Audio, Speech, and Music Processing*, (2012).

[6] Rezayee and S. Gazor, An Adaptive KLT Approach for Speech Enhancement, *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 2, pp. 87–95, February (2001).

[7] Robert Molten Gray, On the Asymptotic Eigenvalue Distribution of Toeplitz Matrices, *IEEE Transactions on Information Theory*, vol. IT-18, no. 6, November (1972).

[8] Yi Hu and Philipos C. Loizou, A Subspace Approach for Enhancing Speech Corrupted by Colored Noise, *IEEE Signal Processing Letters*, vol. 9, no. 7, July (2002).

[9] Y. Hu and P. C. Loizou, A Generalized Subspace Approach for Enhancing Speech Corrupted by Colored Noise, *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 4, pp. 334–341, July (2003).

[10] Hanoch Lev-Ari and Yariv Ephraim, Extension of the Signal Subspace Speech Enhancement Approach to Colored Noise, *IEEE Signal Processing Letters*, vol. 10, no. 4, April (2003).

[11] Junfeng Sun, Jie Zhang and Michael Small, Extension of the Local Subspace Method to Enhancement of Speech with Colored Noise, *Signal Processing*, vol. 88, pp. 1881–1888, (2008).

[12] Firas Jabloun and Benoît Champagne, Incorporating the Human Hearing Properties in the, Signal Subspace Approach for Speech Enhancement, *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, November (2003).

[13] Nathalie Virag, Single Channel Speech Enhancement based on Masking Properties of the Human Auditory System, *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 2, March (1999).

[14] Adam Borowicz and Alexander Petrovsky, Signal Subspace Approach for Psychoacoustically Motivated Speech Enhancement, *Speech Communication*, vol. 53, pp. 210–219, (2011).

[15] Adda Saadoune, Abderrahmane Amroucheb and Sid-Ahmed Selouani, Perceptual Subspace Speech Enhancement using Variance of the Reconstruction Error, *Digital Signal Processing*, vol. 24, pp. 187–196, (2014).

[16] Ching-Ta Lu and Hsiao-Chuan Wang, Enhancement of Single Channel Speech based on Masking Property and Wavelet Transform, *Speech Communication*, vol. 41, pp. 409–427, (2003).

[17] Ching-Ta Lu, Reduction of Musical Residual Noise for Speech Enhancement using Masking Properties and Optimal Smoothing, *Pattern Recognition Letters*, vol. 28, pp. 1300–1306, (2007).

[18] Yi Hu and Philipos C. Loizou, Evaluation of Objective Quality Measures for Speech Enhancement, *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, January (2008)

[19] Jianfen Maa, Philipos C. Loizou, SNR Loss: A New Objective Measure for Predicting the Intelligibility of Noise-suppressed Speech, *Speech Communication*, vol. 53, pp. 340–354, (2011).

[20] Y. Hu and P. Loizou, Subjective Evaluation and Comparison of Speech Enhancement Algorithms, *Speech Communication*, vol. 49, pp. 588–601, (2007).