IJICTR International Journal of Information & Communication Technology Research

Volume 5- Number 2- Spring 2013 (41-54)

Speaker-Dependent Speech Enhancement Using Codebook-based Synthesis for Low SNR Applications

Roghayeh Doost
Amirkabir University of Technology
rdoost@aut.ac.ir

Abolghasem Sayadian
Amirkabir University of Technology
abs35@aut.ac.ir

Received: December19, 2011- Accepted: August 15, 2012

***Abstract*— In this paper, a speaker-dependent speech enhancement is performed by using the codebooks. For this purpose, making use of the STFT parameters, two codebooks are designed for speech and noise separately. In order to design the speech codebook, an adequate number of sentences of particular speakers are used. Utilizing an estimator based on a perceptually weighted distance function, we start searching within the codebooks to find the true indexes for each noisy frame. After finding the true indexes, we synthesize the enhanced speech by using the selected indexes of the speech codebook. As a modification, we suggest two methods to reduce the search time as follows: firstly, a new method for reduction of the codebook size is described. Secondly, by utilizing the relation between the spectral center-of-gravities of the speech, noise and noisy speech, the search area within the codebooks is effectively reduced. Simulation results show that the proposed method can enhance a noisy speech with low SNR. Moreover, since the proposed method is performed frame by frame and it does not use the previous frames of the noisy speech, therefore this method can enhance the noisy speech contaminated by a highly non-stationary noise.**

Keywords- Short time Fourier transform (STFT), speech enhancement, codebook.

I. NTRODUCTION

Speech enhancement is conventionally used as a preprocessing step in various areas of the speech processing [1]. Single channel methods are very applicable methods for speech enhancement. The capability of these methods depends on the SNR level of the noisy speech. For example, in the spectral subtraction [2] and wiener filter [3] methods, the voice activity detector (VAD) is required for silence detection. On the other hand, the performance of VAD is decreased when the SNR level of the noisy speech is decreased [4].

In adaptive spectral subtraction [5], adaptive wiener filter [6], subspace [7,8], and kalman filter [9] methods, it is required to employ the noise estimation approaches whereas the noise estimation approaches do not have good performance for low SNR levels [10].

In the HMM [11,12] and codebook based [1] methods, the linear predictive (LP) coefficients are used to compute the wiener filter. Then the noisy frame is passed through the wiener filter and consequently the enhanced speech is obtained. As a drawback of these methods, LP coefficients contain only the envelope data of the speech spectrum. They do not contain the more detailed data such as the pitch information. Therefore, for low SNR levels of the noisy speech, the enhanced speech does not have an appropriate perceptual quality. In other words, these methods are suitable for speech enhancement when the SNR level of the noisy speech is more than 0dB. [1,11,12].

In this paper, a novel method for enhancement of the noisy speech with negative SNR is proposed. We do not estimate the noise from silence or previous frames. In order to enhance the noisy speech, we do not also employ any filter. In the proposed method, the

enhanced speech is synthesized by using the speech codebook. This codebook contains the STFT parameters of the speech. For design of this codebook, an adequate number of sentences of particular speakers are used. It means that we propose a speaker-dependent speech enhancement [13]. Besides a noise codebook is designed for a particular noise. For each noisy frame, all combinations between the speech and noise code-vectors should be evaluated. Then, the selected speech code-vectors are used to synthesize the enhanced speech. Therefore, the quality of the enhanced speech depends on both the proper design of the codebooks and true selection of the speech code-vectors in the search process. Simulation results show that the proposed method can enhance the noisy speech with low SNR. Besides, since the proposed method is implemented frame by frame and it does not use the previous frames of the noisy speech, therefore this method can enhance the noisy speech contaminated by a highly non-stationary noise.

The application of the proposed method is the enhancement of the noisy speech of particular speakers. These speakers are modeled by using a codebook. Moreover, it is assumed that we know the type of the noise contaminating the speech [14]. So we use the particular codebook of that noise.

In section 2, a new algorithm for speech enhancement is proposed. In section 3, a new method for codebook design is presented. Then two new methods for reduction of the search time are proposed in section 4. Making use of MATLAB simulations, the proposed algorithm is verified in section 5 and finally the paper is concluded in section 6.

II. Speech Enhancement by Using the Codebook

Consider an additive noise model as follows where $y(n)$, $x(n)$, and $w(n)$ are the noisy speech, speech, and noise respectively. The parameter n denotes the time sample.

$$y(n) = x(n) + w(n) \tag{1}$$

The standard deviation of the noisy speech, speech, and noise are defined as follows: $\sigma_y = \sqrt{E\{y^2(n)\}}$, $\sigma_x = \sqrt{E\{x^2(n)\}}$, and $\sigma_w = \sqrt{E\{w^2(n)\}}$. In these relations $E\{.\}$ denotes the expectation function. We can normalize the speech and noise as shown below where $\hat{x}(n)$ and $\hat{w}(n)$ are the normalized values of $x(n)$ and $w(n)$, respectively. Thus, we have $E\{\hat{x}^2(n)\}=1$ and $E\{\hat{w}^2(n)\}=1$.

$$\hat{x}(n) = \frac{1}{\sigma_x} x(n) \tag{2}$$

$$\hat{w}(n) = \frac{1}{\sigma_w} w(n) \tag{3}$$

Making use of above relations, the noisy speech is rewritten as follows:

$$y(n) = \sigma_y \left(\frac{\sigma_x}{\sigma_y}\hat{x}(n) + \frac{\sigma_w}{\sigma_y}\hat{w}(n)\right) \tag{4}$$

Assuming $x(n)$ and $w(n)$ are independent and zero mean, we have:

$$\sigma_y^2 = \sigma_x^2 + \sigma_w^2 \tag{5}$$

Employing relation (5) and defining $\alpha = \frac{\sigma_x}{\sigma_y}$, relation (4) can be written as follows:

$$y(n) = \sigma_y (\alpha\hat{x}(n) + \sqrt{1-\alpha^2}\,\hat{w}(n)) \tag{6}$$

In this relation, $\sigma_y \alpha\hat{x}(n)$ implies to the speech. Since in the silence frames, there is not any speech signal and only the noise signal exists, therefore we have $\alpha = 0$. Correspondingly, $\alpha = 1$ implies that the frames are not contaminated with any noise.

In this paper, making use of a priori information contained in the speech and noise codebooks, a new method is proposed to enhance the noisy speech. The speech codebook is trained by using an adequate number of sentences of particular speakers. The noise codebook is also designed for a particular noise of the database. These codebooks contain STFT parameters of the speech and noise. We assume that the speech and noise codebooks have N_x and N_w vectors, respectively.

We model the observed noisy speech, *y(n)*, by using the codebooks entries. For this purpose we select STFT vectors of $\overline{X}_c^i$ and $\overline{W}_c^j$ ($i = 1....N_x$, $j = 1...N_w$) from the speech and noise codebooks, respectively. The complex vectors, $\overline{X}_c^i$ and $\overline{W}_c^j$, contain both the amplitude and phase of STFT parameters. The amplitude of these vectors is represented by X_c^i and W_c^j in which the subscript "*c*" denotes the codebook. Then, the time domain variables, x_c^i and w_c^j, are calculated as follows:

$$x_c^i(n) = IFFT(\overline{X}_c^i) \tag{7}$$

$$w_c^j(n) = IFFT(\overline{W}_c^j) \tag{8}$$

The standard deviation of x_c^i and w_c^j are defined as follows:

$$\sigma_{xc}^i = \sqrt{E\{x_c^{i\,2}(n)\}} \tag{9}$$

$$\sigma_{wc}^j = \sqrt{E\{w_c^{j\,2}(n)\}} \tag{10}$$

where we have [15, 16]:

$$E\{x_c^{i\,2}(n)\} = \frac{1}{FL}\sum_{n=1}^{FL} x_c^{i\,2}(n) \tag{11}$$

† The authors are with Amirkabir University of Technology, Tehran, Iran.

$$E\{w_c^{j\,2}(n)\} = \frac{1}{FL}\sum_{n=1}^{FL} w_c^{j\,2}(n) \qquad (12)$$

and FL denotes the frame length. According to relations (9) to (12), the standard deviation of x_c^i and w_c^j are obtained as follows:

$$\sigma_{xc}^i = \sqrt{\frac{1}{FL}\sum_{n=1}^{FL} x_c^{i\,2}(n)} \qquad (13)$$

$$\sigma_{wc}^j = \sqrt{\frac{1}{FL}\sum_{n=1}^{FL} w_c^{j\,2}(n)} \qquad (14)$$

Therefore, x_c^i and w_c^j are normalized as follows:

$$\hat{x}_c^i(n) = \frac{x_c^i(n)}{\sigma_{xc}^i} \qquad (15)$$

$$\hat{w}_c^j(n) = \frac{w_c^j(n)}{\sigma_{wc}^j} \qquad (16)$$

Substituting x_c^i by y in relations (9), (11), and (13), σ_y is calculated in each frame:

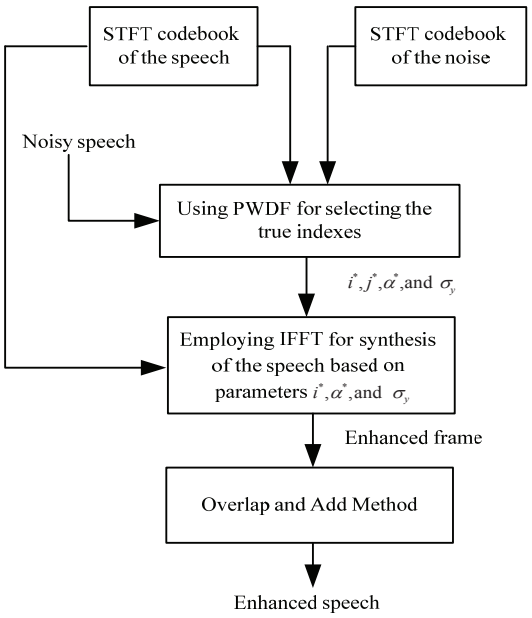

Figure 1. Block diagram of the proposed method for enhancement of the noisy speech with low (negative) SNR.

$$\sigma_y = \sqrt{\frac{1}{FL}\sum_{n=1}^{FL} y^2(n)} \qquad (17)$$

Finally, for $0 \le \alpha \le 1$ and all combinations of i and j, we determine the modeled noisy speech, $z_\alpha^{ij}(n)$, as follows:

$$z_\alpha^{ij}(n) = \sigma_y(\alpha.\hat{x}_c^i(n) + \sqrt{1-\alpha^2}.\hat{w}_c^j(n)) \qquad (18)$$

In fact $z_\alpha^{ij}(n)$ is the model of *y(n)*. The noisy speech *y(n)* is expressed in relation (6). Moreover, $\sigma_y.\alpha\hat{x}_c^i(n)$ is the modeled speech, written making use of the *i*'th index of the speech codebook and α. In order to perform the speech enhancement, the optimum index of the speech codebook, i^*, and the optimum value of α, α^*, must be found. The details for finding i^* and α^* will be explained in the last paragraph of this section.

According to relations (15), (16) and (18), SNR of $z_\alpha^{ij}(n)$ is:

$$SNR_\alpha = 20\log_{10}\left(\frac{\alpha}{\sqrt{(1-\alpha^2)}}\right) \qquad (19)$$

Thus:

$$\alpha = \left(\frac{1}{1+10^{\frac{-SNR_\alpha}{10}}}\right)^{0.5} \qquad (20)$$

We assume that the SNR of the noisy speech varies from $SNR_{\min}$ to $SNR_{\max}$. So in our simulations, α varies from $\left(\frac{1}{1+10^{\frac{-SNR_{\min}}{10}}}\right)^{0.5}$ to $\left(\frac{1}{1+10^{\frac{-SNR_{\max}}{10}}}\right)^{0.5}$. If the number of α values in the simulations are chosen identical to N_α, the number of calculations of $z_\alpha^{ij}(n)$ will be $N_x \times N_w \times N_\alpha$. Besides, for detecting the silence and noise-free frames, both values of $\alpha = 0$ and $\alpha = 1$ are also taken into consideration in our simulations.

As mentioned in [1], combinations of codebooks entries that minimize the spectrally based distance function between the observed and modeled noisy speech are selected as the parameters of the speech and noise. In other words, the best combination of codebooks entries will be obtained through a spectral matching perspective [1]. For this purpose, we employ the perceptually weighted distance function, PWDF, in this paper. This function is introduced in appendix1. The advantages of this distance function are explained in [17]. The values of i, j and α that minimize PWDF will be adopted as shown in relation (21).

$$\{i^*, j^*, \alpha^*\} = \arg\min_{i,j,\alpha} PWDF(Y, Z_\alpha^{ij}) \qquad (21)$$

Y and Z_α^{ij} are the STFT amplitudes of $y(n)$ and $z_\alpha^{ij}(n)$, respectively. By using the selected index of the speech codebook, i^*, and the selected value of α, α^*, the enhanced speech (the best model of the speech) is obtained as follows:

$$x^*(n) = \sigma_y.\alpha^*.\hat{x}_c^{i^*}(n) \qquad (22)$$

that $\hat{x}_c^{i^*}(n)$ is obtained using relations (7), (13) and (15). In these relations i must be replaced by i^*. After determining all of the enhanced frames, the enhanced sentence is obtained by using both the overlap-add method (OVA) and phase synchronization technique [18]. The block diagram of the above enhancement method is shown in Fig.1. Compared to the conventional approaches, the advantage of our method is that it does not use any filtering in the speech enhancement. Besides, it does not estimate the noise from the silence or previous frames. In the proposed

method, the enhanced speech is synthesized by using the selected indexes of the speech codebook. Simulation results show that the proposed method can enhance the noisy speech with a low (negative) SNR.

III. Codebook Design

In this section, we propose a new method for the codebook design of the STFT parameters. Although the following relations are described for the speech codebook only, however theses formulas can be extended for the noise codebook similarly. Our training vectors, X^r, are the amplitudes of the STFT parameters. Firstly, we normalize all of the training vectors as follows:

$$\hat{X}^r = \frac{X^r}{\sqrt{\sum_{k=1}^{K} X^r(k)^2}}; \qquad r = 1,...,M \qquad (23)$$

where X^r and $\hat{X}^r$ are the training vector and normalized training vector, respectively. Moreover k and K are the frequency index and STFT length, respectively. The parameter M is the number of training vectors. Then we choose an arbitrary vector X^i from the training vectors and quantize other training vectors X^j ($j = 1,...,M$) to it if the condition of relation (24) is satisfied.

$$PWDF\left(\hat{X}^i, \hat{X}^j\right) < Th; \qquad 0 \le Th \le 1 \qquad (24)$$

In this relation $PWDF\left(\hat{X}^i, \hat{X}^j\right)$ represents a perceptually weighted distance measure between vectors $\hat{X}^i$ and $\hat{X}^j$ as explained in appendix 1. Besides, Th is the maximum allowable perceptually weighted distance between a vector and its quantized vector that is not heard by the human ear. The optimum value of Th is obtained by simulations. Therefore, a set of vectors is quantized to X^i with quantization error less than the threshold value, Th . In other words, X^i will be a vector of the codebook that we are designing. In order to find the other vectors of the codebook, we choose another arbitrary vector from the un-quantized training vectors and repeat the above quantization algorithm for un-quantized training vectors again. This procedure is continued until all of the training vectors are quantized. Making use of the above algorithm, the vectors of the codebook are selected from the training vectors. In other words, for extraction of the code-vectors we do not employ any averaging techniques while conventional methods such as LBG [19] utilize an averaging process to extract the code-vectors.

As shown in relations (7) and (8), in the proposed enhancement method of section 2 we need both the amplitude and phase of the STFT vectors. Therefore, while we are designing the codebook, both the amplitude and phase of the STFT must be selected. In contrast, in the LBG method the code-vectors are extracted by using an averaging process on the amplitudes of STFT parameters. Each of these code-vectors is not necessarily identical to one of the training vectors. Therefore, it is not possible to select the proper phases for them.

IV. Reduction of the Search Time

As mentioned in section 2, we must calculate $z_\alpha^{ij}(n)$ for all combinations of i, j, and α . Therefore, it is required searching through the large amounts of codebooks entries ($N_x \times N_w \times N_\alpha$ combinations of the codebooks entries). In this section as a modification, two methods are suggested to reduce the search time in the enhancement procedure of this paper.

A. Reduction of the codebooks size

If the size of the codebooks is reduced, the search time will be decreased. For this purpose, the designed codebook of section 3 is modified as follows: Firstly, considering the number of training vectors that are quantized to each code-vector, we sort the vectors of the previously designed codebook. In example, assume h_i and h_j $(i, j = 1....N_x)$ denote the number of training vectors that are quantized to code-vectors X_c^i and X_c^j, respectively. The sorting algorithm is performed as follows: if $i \le j$, then $h_i \ge h_j$. In fact $H_x = \left[h_1, h_2, \cdots, h_{N_x}\right]$ is the histogram of the code-vectors where for $i \le j$, the number of training vectors quantized to X_c^i are more than those quantized to X_c^j.

The vectors that have smaller probability will be heard in a sentence less than those have greater probability and vice versa. So the vectors having smaller probability can be quantized with a greater quantization error than those having more probability. Therefore, in order to modify the codebook, we quantize the training vectors again with a variable threshold as described in the following relations. In relation (25), corresponding to each code-vector X_c^m, the threshold $Th_c(m)$ is defined. The parameter Th denotes the threshold value that is used previously (section 3) in the design of the initial codebook. Moreover, γ is a parameter that its optimum value is obtained through the simulation.

$$Th_c(m) = \left(1 + \frac{m}{N_x}\gamma\right) Th \qquad m = 1,2,\cdots,N_x \qquad (25)$$

Then corresponding to each training vector, the threshold $Th_t(i)$ $(i = 1,2,\cdots,M)$ is defined as follows where $q(X^i)$ denotes the quantized vector of X^i .

$$Th_t(i) = Th_c(m) \quad if \quad q\left(X^i\right) = X_c^m \qquad (26)$$

In other words, all of the training vectors that are previously quantized to an identical code-vector will have an equal threshold.

After determination of $Th_t(i)$ for all values of i, we start to design the final codebook. The procedure for designing the final codebook is approximately identical to that of the initial codebook. The only difference between these procedures is that instead of

relation (24), the following relation is utilized for quantization.

$$q(X^j)=X^i \quad if \quad PWDF(\hat{X}^i,\hat{X}^j)\le Th_1(i) \qquad (27)$$

In fact, in the design of the final codebook the vectors with less probability have a greater threshold, leading to decrease the size of the codebook. Although the above relations are described for speech, however they can also be employed in order to reduce the size of the noise codebook.

B. Proposed search algorithm

In this section, the spectral center-of-gravity, COG, of the noisy speech $y(n)$ is investigated. Definition of COG for each frame of the noisy speech is as follows [20,21] where the variable p denotes the frame number.

$$COG_y(p)=\frac{\sum_{k=1}^{K} kY(k)^2}{\sum_{k=1}^{K} Y(k)^2} \qquad (28)$$

Similarly COG is obtained for both the speech $x(n)$ and noise $w(n)$, separately. For this purpose, in relation (28) the variable $Y(k)$ must be substituted by $X(k)$ and $W(k)$, respectively. Moreover, the subscript y must be substituted by x and w, correspondingly. According to appendix 2, we have:

$$\min\{COG_x(p),COG_w(p)\}\le COG_y(p)\le \max\{COG_x(p),COG_w(p)\} \qquad (29)$$

Making use of the above relation, in following paragraphs a new search algorithm is proposed that avoids searching in unprofitable combinations. For this purpose, both the speech and noise codebooks are sorted with respect to their center-of-gravity values as follows:

$$if\ i\le j \ \Rightarrow COG_{cx}(i)\ge COG_{cx}(j) \quad i,j=1,2,\cdots,N_x \qquad (30)$$

$$if\ i\le j \ \Rightarrow COG_{cw}(i)\ge COG_{cw}(j) \quad i,j=1,2,\cdots,N_w \qquad (31)$$

In the above relations the subscripts cx and cw denote the codebooks of the speech and noise, respectively. $COG_{cx}(i)$ and $COG_{cw}(i)$ are obtained according to relation (28). For this purpose the variable Y must be substituted by X_c^i and W_c^i, correspondingly. The subscript y must also be replaced by cx and cw, in the same way. It is important to note that in this step, we have never designed other codebooks. We only sorted the entries of the codebooks with respect to their center-of-gravities.

In the search algorithm, firstly the value of $COG_y(p)$ is calculated. Then it is tried to find the proper values of indexes $k_x(p)$ and $k_w(p)$ in the speech and noise codebooks so that the following relations are satisfied.

$$\begin{cases} COG_{cx}(i)\ge COG_y(p) & if \quad i\le k_x(p) \\ \& & \\ COG_{cx}(i)\le COG_y(p) & if \quad i\ge k_x(p) \end{cases} \qquad (32)$$

$$\begin{cases} COG_{cw}(j)\ge COG_y(p) & if \quad j\le k_w(p) \\ \& & \\ COG_{cw}(j)\le COG_y(p) & if \quad j\ge k_w(p) \end{cases} \qquad (33)$$

According to relations (29), (32), and (33), therefore it is sufficient to perform the search only in two groups of code-vectors. The first group is comprised of speech code-vectors with indexes $i\le k_x(p)$ and noise code-vectors with indexes $j\ge k_w(p)$. The second group is comprised of speech and noise code-vectors with indexes $i\ge k_x(p)$ and $j\le k_w(p)$, respectively. Making use of this approach, the search domain is restricted and consequently the search time is reduced.

V. Simulation Results

We use the Cooke database [22] consisting of 34 speakers uttering 500 sentences. We select four speakers including two male speakers (speakers 2 and 10) and two female speakers (speaker 7 and 16) from the database. We use 450 sentences of each speaker to train the codebooks. Besides, 4 different noises are separately utilized. They are babble noise, street noise from aurora database, factory noise, and volvo noise from noisex-92 database. The sampling rate of the sentences and noises of the databases are reduced to that of the noisy speech, 8 kHz. The frame length and frame shift are identical to 32ms and 10ms, respectively. Moreover, in order to extract the STFT parameters, the hamming window and 1024-point FFT are utilized. In the following simulations, 10 sentences from each of four speakers are used as the test sentences for evaluation of the speech codebooks. Of course, test sentences are different from training sentences. Besides, both the mentioned test sentences and unused signals of each noise are employed for generating the noisy speech with various SNR levels.

Firstly, the speech codebook is designed for various values of Th. For this purpose, the method of section III is employed. In figures 2 to 5, the average of PESQ [1] and the codebook size are depicted versus Th for speech codebook. Since PESQ=3.8 is an appropriate selection for the speech codebook [23], therefore Th is obtained as shown in table 1. Then the noise codebooks with several sizes are designed for each noise. These codebooks are designed by using of several values of Th, separately. Now we enhance the noisy speech by using of the selected speech codebooks and noise codebooks with several sizes, separately. In figures 6 to 9 the average of PESQ over all speakers is shown versus the noise codebook sizes for SNR equal to -10dB. As shown in figures 6 to 9, we choose noise codebook sizes according to table 2. The PESQ values of the enhanced speech are approximately saturated in the chosen noise codebook sizes.

In the following simulations, by using relation (20), the values of α are obtained for SNR levels from -12dB

to 12dB with the step 1 dB ($N_\alpha = 25$). In order to find the true indexes from codebooks, the relation (21) is utilized. Then making use of relation (22), the enhanced speech is synthesized. Before the presentation of enhancement results, we investigate the drawback of the proposed method and modify it. The drawback of the above method is that we must evaluate too many combinations in each frame, which is very time consuming. The number of combinations that must be calculated in each frame is equal to $N_x \times N_w \times N_\alpha$. This causes a time-consuming enhancement process for proposed method. Therefore, to reduce the search time, we will employ the proposed methods of section IV as follows.

TABLE 1. APPROPRIATE *TH* AND CODEBOOK SIZE FOR VARIOUS SPEAKERS

speakers	Spk1	Spk2	Spk3	Spk4
Th	0.145	0.147	0.15	0.130
Codebook size	18000	16700	16500	22800

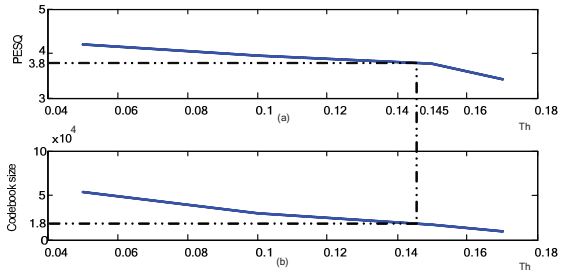

Figure 2. (a) PESQ and (b) codebook size versus *Th* for spk1

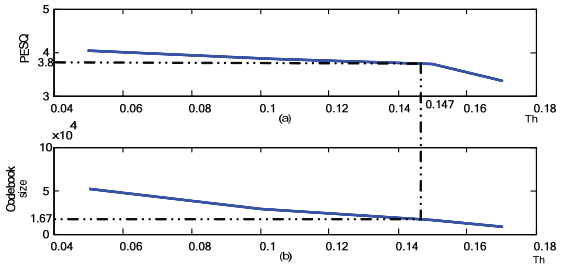

Figure3. (a) PESQ and (b) codebook size versus *Th* for spk2

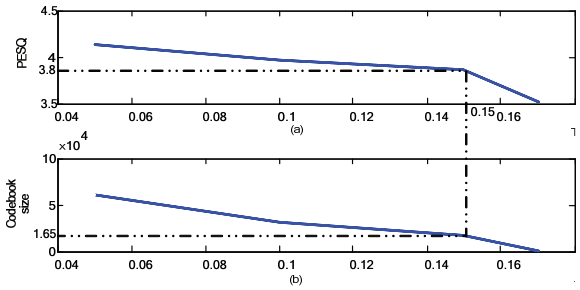

Figure 4. (a) PESQ and (b) codebook size versus *Th* for spk3

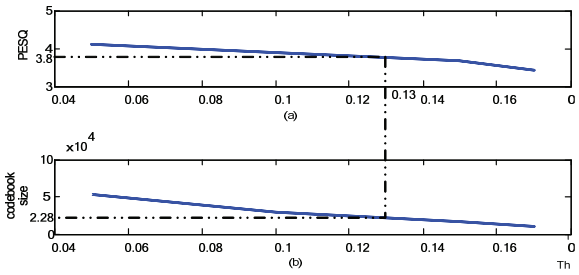

Figure5. (a) PESQ and (b) codebook size versus *Th* for spk4

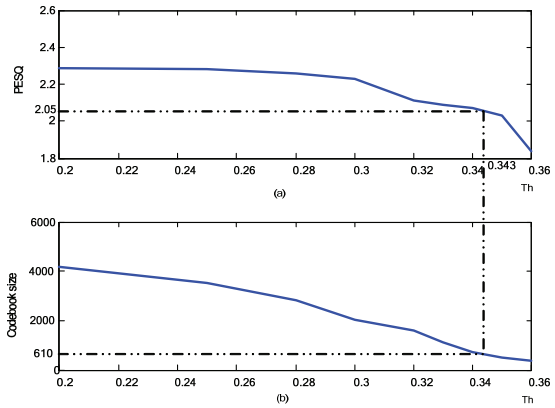

Figure 6. (a) PESQ and (b) codebook size versus *Th* for babble noise codebook

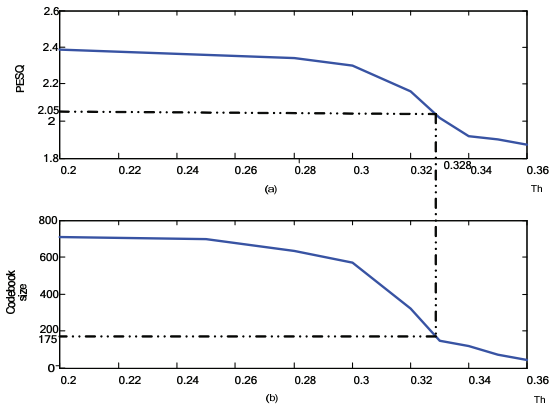

Figure 7. (a) PESQ and (b) codebook size versus *Th* for factory noise codebook

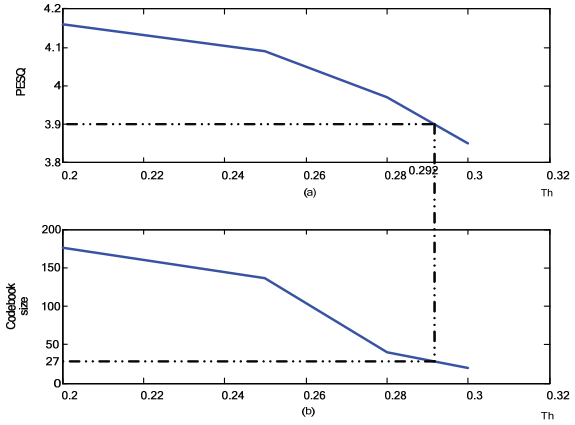

Figure 8. (a) PESQ and (b) codebook size versus *Th* for volvo noise codebook

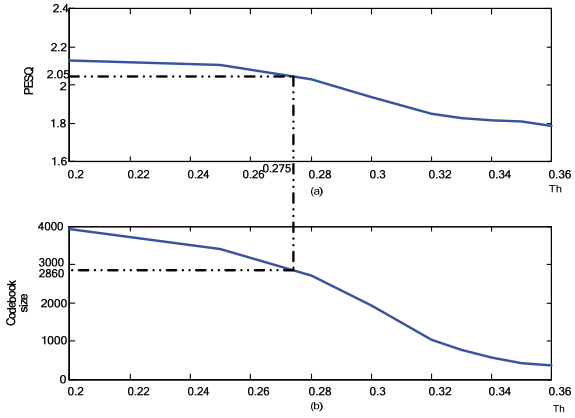

Figure 9. (a) PESQ and (b) codebook size versus *Th* for street noise codebook

In the design of the speech codebook based on the method described in section IV.A, it is necessary to find the optimum value of γ. For this purpose, the codebooks are designed for various values of *Th* and γ. Consequently, the average value of PESQ is obtained for each codebook. The simulation results are shown for various speakers in figures 10 to 13. Moreover, in figures 10 to 13 the speech codebook size is shown versus γ for various *Th* values. Considering the simulation results of figures 10 to 13, it is concluded that *Th* , γ and codebook size of table 3 lead to the appropriate PESQ of 3.8. Then the noise codebooks with several sizes are designed for each noise. These codebooks are designed by using of several values of *Th* and γ separately. Now we enhance the noisy speech by using of the selected speech codebooks and noise codebooks with several sizes, separately. In figures14 to 16 the average of PESQ over all speakers is shown versus the noise codebook sizes for SNR equal to -10dB. As shown in figures 14 to 16, we choose noise codebook sizes according to table 4. The PESQ values of the enhanced speech are approximately saturated in the chosen noise codebook sizes. As shown in table 2, the volvo noise codebook size is small enough. Therefore we do not reduce it.

TABLE.2. APPROPRIATE *TH* AND CODEBOOK SIZE FOR

Th	0.343	0.328	0.292	0.275
Codebook size	610	175	27	2860

As shown in the table 3, codebook sizes of speakers are not identical. So we reduce codebook sizes to 8192 as follow. The general histogram of k^{th} speaker is shown in figure 17. In this figure "*i*" denotes code-vectors indexes and *H* is the histogram. The nearest vector index *I* (*I*<8192) is determined for each *J* (*J*>8192). We define $X_w = \frac{H(I)X_c^I + H(J)X_c^J}{H(I)+H(J)}$. Then the nearest train-vector X_c^{IJ} to X_w is selected. Then X_c^I and X_c^J are replaced with the new code-vector X_c^{IJ} . So all vectors with *I*>8192 are omitted and the codebook size is reduced to 8192. After this codebook size reduction the PESQ of speakers is reduced a little as shown in table 5. Like this method the codebook sizes of noise codebooks are reduced to values of table 6. The average value of PESQ of final noise codebooks is reduced about 0.02.

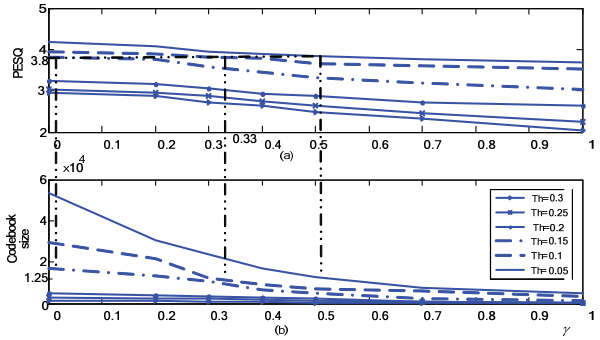

Figure 10. (a) PESQ and (b) codebook size of spk1 versus γ for various *Th* values

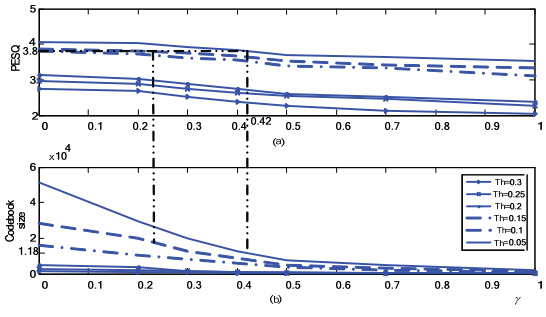

Figure 11. (a) PESQ and (b) codebook size of spk2 versus γ for various *Th* values

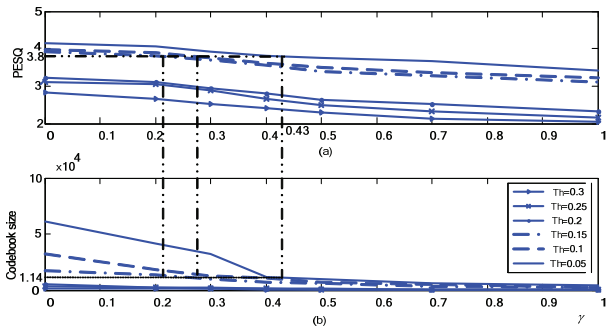

Figure 12. (a) PESQ and (b) codebook size of spk3 versus γ for various Th values

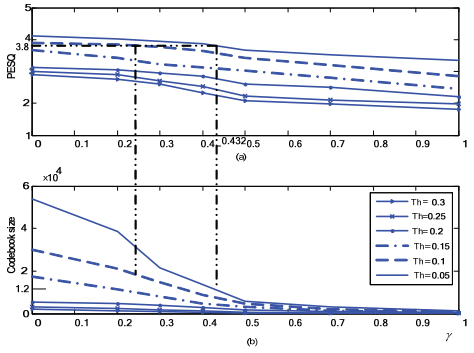

Figure 13. (a) PESQ and (b) codebook size of spk4 versus γ for various Th values

TABLE 3. APPROPRIATE *TH*, γ AND CODEBOOK SIZE FOR VARIOUS SPEAKERS

speakers	*Th*	γ	Codebook size
Spk1	0.1	0.33	12500
Spk2	0.05	0.42	11800
Spk3	0.05	0.43	11400
Spk4	0.05	0.43	12000

In section IV.B, based on relations between the spectral center-of-gravities of the speech, noise, and noisy speech, an algorithm to reduce the search time is proposed. This algorithm avoids searching in unprofitable combinations of the codebooks entries. Simulation results prove this method approximately halves the number of combinations that must be searched. This fact is more clarified in the following paragraphs.

TABLE 4. APPROPRIATE *TH*, γ , AND CODEBOOK SIZE FOR

Noise	babble	factory	volvo	street
Th	0.3	0.3	0.292	0.25
γ	0.33	0.38	0.00	0.25
Codebook size	562	165	27	1630

TABLE 5. PESQ OF SPEAKERS AFTER CODEBOOK SIZE REDUCTION.

Speakers	Spk1	Spk2	Spk3	Spk4
PESQ	3.72	3.75	3.76	3.74

TABLE 6. FINAL CODEBOOK SIZE OF NOISES

noise	babble	factory	volvo	street
codebook size	256	128	8	512

TABLE 7. AVERAGE IMPLEMENTATION TIME OF VARIOUS METHODS (IN HOUR)

#1	#2	#3	#4	WF	LPC-CB
18.4	2.3	9.1	1.1	0.004	0.17

We enhance the noisy speech by six different methods as follows: 1) the codebooks are designed based on the method described in section III and the search algorithm is performed on all combinations of the codebooks entries (method #1). 2) Speech codebook size is equal to 8192 and noise codebook size is equal to values of table 6. The search algorithm is performed on all combinations of the codebooks entries (method #2). 3) The codebooks are designed making use of the method described in section III and the search algorithm is done according to section IV.B (method #3). 4) Speech codebook size is equal to 8192 and noise codebook size is equal to values of table 6. The search algorithm is performed like section IV.B (method #4). In addition, the following methods are implemented for comparison with our methods. 5) Wiener filter that is implemented according to [3] and utilizes the noise estimation of [10] (WF method). 6) The method of [1] that uses the codebooks of LPC coefficients and wiener filtering to enhance the noisy speech (LPC_CB method).

All of the above methods are implemented by MATLAB2010 in a personal computer with Dual-Core 2.67GHz-CPU. The average required time for implementation of each method is explained in table.7. It is considered that the methods #2, #3 and #4 approximately reduce the search time by a factor of eight, two and seventeen, respectively.

The average values of the segmental SNR (SSNR)[1], SD, and PESQ of the enhanced speech are computed for each method. The average values of SSNR, SD, and PESQ are summarized in tables 8 to 10 for babble, factory, street and volvo noises. These results are shown for SNR values of -5dB and -10dB. In addition, as our subjective measures, we conducted a Mean Opinion Score (MOS) informal listening test to measure the perceived quality of the enhanced speech. Ten people were asked to give score between 1-5 to the enhanced speech (5 represents the clean speech score). The average of the listeners' scores is summarized in table.11. These results are shown for SNR values of -5 and -10dB. It is considered that our methods (methods #1 to #4) can enhance the noisy speech with low (negative) SNR. The methods #1 to #4 have

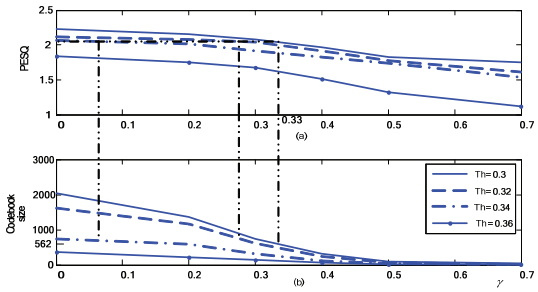

Figure 14. (a) PESQ and (b) codebook size of babble noise versus γ for various Th values

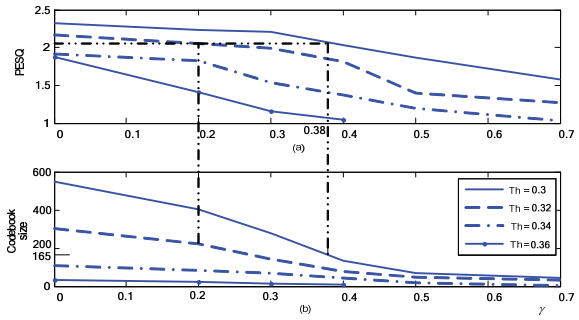

Figure 15. (a) PESQ and (b) codebook size of factory noise versus γ for various Th values

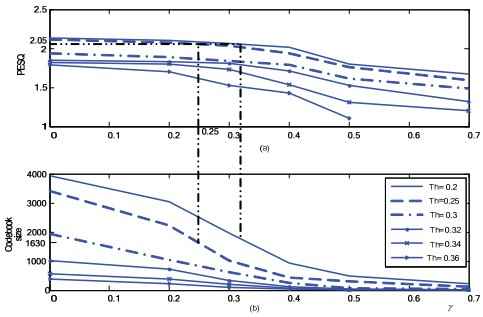

Figure 16. (a) PESQ and (b) codebook size of street noise versus γ for various Th values

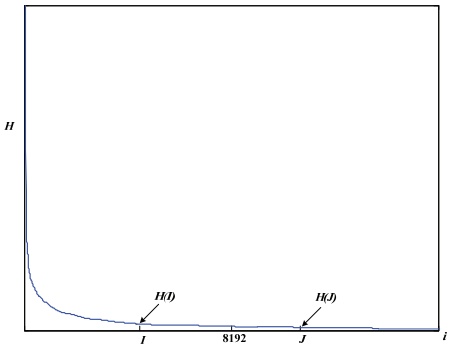

Figure 17. General histogram of spk_k

significantly better performance than methods WF and LPC_CB.

The noise that contaminates the speech is not heard in the enhanced speech. The artifact in the enhanced speech is because of the distortion of the synthesis process. At low SNR the error of the estimation of indexes are increased. So the distortion of the enhanced speech is increased and consequently its intelligibility is decreased. Despite the proposed method, the enhanced speech by using WF and LPC-CB methods lacks intelligibility at low SNR (below -5dB). In contrast, making use of the proposed method (method#4) 53% of words of the enhanced speech are heard correctly at SNR values of -10dB.

Considering tables 8 to 11 it is known that the method #1 has the best perceptual quality and the most implementation time among the proposed methods. On the other hands, the method #4 has less perceptual quality and smaller implementation time compared to other proposed methods. The above simulations and measurements are also repeated for SNR values of 0, 5 and 10 dB. The results are shown in tables 12 to 14. Although all of methods have acceptable performance in these simulations, however the methods # 1 to #4 have better performance than WF method and LPC_CB method.

The disadvantage of the proposed methods is that they need to a time-consuming search process compared to method LPC_CB method. WF method does not need any search process and it is implemented very quickly while our methods do not have this capability. In fact our methods can be employed in off-line applications. As the limitations of our methods, they are speaker-dependent. They also assume that we know the noise type contaminating the speech. The above disadvantage and limitations are the price that we pay for improving the performance in low (negative) SNR conditions.

TABLE.8. PESQ OF THE ENHANCED SPEECH FOR VARIOUS METHODS (SNR VALUES ARE -10 DB AND -5 DB.)

Noise	Input SNR(dB)	PESQ						
		Noisy	#1	#2	#3	#4	WF	LPC-CB
babble	-10	0.49	1.85	1.76	1.83	1.75	0.66	0.90
	-5	0.67	2.05	1.97	2.05	1.96	0.84	1.36
factory	-10	0.61	1.75	1.66	1.73	1.65	0.73	0.94
	-5	0.88	2.28	2.19	2.27	2.18	1.01	1.37
street	-10	0.49	1.74	1.66	1.72	1.63	0.63	0.83
	-5	0.67	2.16	2.07	2.15	2.06	0.98	1.34
volvo	-10	1.88	3.01	2.93	3.00	2.92	1.94	2.12
	-5	2.29	3.12	3.03	3.10	3.01	2.33	2.67

TABLE 9.SD OF THE ENHANCED SPEECH FOR VARIOUS METHODS (SNR VALUES ARE -10 DB AND -5 DB.)

Noise	Input SNR(dB)	SD						
		Noisy	#1	#2	#3	#4	WF	LPC-CB
babble	-10	16.1	7.3	7.8	7.5	7.9	10.1	9.2
	-5	14.5	6.5	7.2	6.9	7.4	8.9	8.1
factory	-10	14.1	7.8	8.4	8.0	8.6	9.5	9.4
	-5	12.8	6.3	7.0	6.6	7.1	8.7	8.5
street	-10	14.7	7.5	8.2	7.9	8.3	9.7	9.6
	-5	13.9	6.8	7.3	6.9	7.3	8.8	8.5
volvo	-10	8.9	5.5	6.0	5.6	6.2	4.5	4.3
	-5	7.8	5.3	5.8	5.4	6.0	4.4	4.2

Tables 8 to 14 show a little reduction in the perceptual quality of methods that use the COG characteristics of the noisy speech (method #3 and method #4). It is because of the approximation that is used in appendix.2.

VI. Conclusion

In this paper, making use of the STFT parameters, two codebooks are designed for speech and noise, independently. Each of these codebooks contains both the amplitude and phase of the STFT. For particular speakers contained in the speech codebook, the speech enhancement is performed as follows. Utilizing a perceptually weighted distance function and employing a search algorithm within the codebooks entries, the true indexes are found and consequently the enhanced speech is synthesized by using the speech codebook. Besides, as a modification of the proposed method, a new method for reduction of the codebook size is described. In this method, the training vectors having smaller probability are quantized with a greater quantization error than those having more probability. Finally the speech codebook sizes are reduced to 8192. Similarly the codebook sizes of noise are also reduced. Furthermore, by utilizing the relation between the spectral center-of-gravities of the speech, noise and noisy speech, another method is proposed for reducing the search area within the codebooks entries. Simulation results prove that both the above approaches reduce the search time by a factor of seventeen. What makes our enhancement method outstanding compared to the conventional approaches is that this method synthesizes the enhanced speech by using the STFT parameters of the speech codebook without any need to the filtering operations. Besides, it does not require estimating the noise from the silence or previous frames. Simulation results prove that the proposed method can enhance a noisy speech with low (negative) SNR.

TABLE 10.SSNR OF THE ENHANCED SPEECH FOR VARIOUS METHODS (SNR VALUES ARE -10 DB AND -5 DB.)

Noise	Input SNR (dB)	SSNR(dB)						
		Noisy	#1	#2	#3	#4	WF	LPC-CB
babble	-10	-13.6	5.6	5.0	5.4	4.7	-0.6	0.5
	-5	-8.7	8.0	7.3	7.7	7.2	0.3	1.1
factory	-10	-13.9	4.7	4.0	4.3	3.8	-0.3	0.5
	-5	-9.0	7.8	7.3	7.6	7.1	0.9	1.8
street	-10	-14.1	4.7	4.0	4.5	3.9	-0.4	-0.4
	-5	-9.3	7.2	6.6	7.1	6.5	1.1	0.8
volvo	-10	-13.4	11.2	10.5	10.9	9.8	-0.3	4.3
	-5	-8.6	13.0	12.3	12.8	12.2	1.8	4.5

TABLE 11. MOS OF THE ENHANCED SPEECH FOR VARIOUS METHODS (SNR VALUES ARE -10 DB AND -5 DB.)

Noise	Input SNR(dB)	MOS						
		Noisy	#1	#2	#3	#4	WF	LPC-CB
babble	-10	1.00	2.03	1.92	2.02	1.89	1.04	1.14
	-5	1.07	2.24	2.18	2.21	2.15	1.11	1.52
factory	-10	1.03	2.08	1.95	2.06	1.95	1.08	1.15
	-5	1.16	2.37	2.31	2.36	2.28	1.23	1.62
street	-10	1.00	1.93	1.84	1.90	1.81	1.03	1.12
	-5	1.06	2.32	2.25	2.28	2.21	1.24	1.58
volvo	-10	1.81	3.31	3.24	3.27	3.22	2.65	2.81
	-5	2.46	3.52	3.36	3.47	3.35	2.90	3.21

TABLE 12. PESQ OF THE ENHANCED SPEECH FOR VARIOUS METHODS (SNR VALUES ARE 0 DB, 5DB AND 10 DB.)

Noise	Input SNR(dB)	PESQ						
		Noisy	#1	#2	#3	#4	WF	LPC-CB
babble	0	0.93	2.61	2.54	2.60	2.53	1.34	1.52
	5	1.30	2.84	2.78	2.83	2.77	1.70	1.87
	10	1.76	3.13	3.07	3.13	3.06	2.13	2.22
factory	0	1.27	2.79	2.73	2.78	2.72	1.82	1.87
	5	1.73	2.88	2.81	2.86	2.80	1.91	2.02
	10	2.16	3.27	3.20	3.27	3.17	2.34	2.53
street	0	0.95	2.66	2.60	2.66	2.60	1.85	1.94
	5	1.37	2.82	2.76	2.82	2.76	2.07	2.30
	10	1.86	3.11	3.05	3.11	3.04	2.34	2.68
volvo	0	2.66	3.16	3.09	3.14	3.07	3.10	3.28
	5	3.02	3.33	3.26	3.31	3.23	3.21	3.41
	10	3.21	3.45	3.38	3.43	3.36	3.36	3.52

TABLE 13. SD OF THE ENHANCED SPEECH FOR VARIOUS METHODS (SNR VALUES ARE 0 DB, 5DB AND 10 DB.)

Noise	Input SNR(dB)	SD(dB)						
		Noisy	#1	#2	#3	#4	WF	LPC-CB
babble	0	12.8	6.4	6.8	6.4	6.9	6.5	6.4
	5	10.2	5.7	6.1	5.8	6.2	4.9	4.6
	10	9.1	5.2	5.5	5.3	5.6	4.1	3.9
factory	0	12.5	6.1	6.4	6.1	6.5	5.8	5.6
	5	12.1	5.7	6.1	5.7	6.2	4.6	4.1
	10	11.3	5.2	5.5	5.3	5.7	3.5	3.1
street	0	12.9	6.4	6.9	6.5	6.9	5.7	5.4
	5	12.1	6.0	6.4	6.1	6.5	4.6	4.5
	10	10.8	5.1	5.5	5.2	5.6	3.5	3.3
volvo	0	6.9	5.2	5.6	5.3	5.7	4.3	4.0
	5	6.3	5.1	5.3	5.2	5.4	3.1	2.8
	10	5.7	5.0	5.2	5.1	5.3	2.7	2.5

TABLE 14. SSNR OF THE ENHANCED SPEECH FOR VARIOUS METHODS (SNR VALUES ARE 0 DB, 5DB AND 10 DB.)

Noise	Input SNR(dB)	SSNR						
		Noisy	#1	#2	#3	#4	WF	LPC-CB
babble	0	-4.1	9.3	8.7	9.2	8.6	3.7	3.9
	5	1.1	11.3	10.7	11.2	10.6	6.5	6.6
	10	4.7	13.2	12.7	13.1	12.6	7.1	7.1
factory	0	-4.4	10.5	9.9	10.4	9.8	4.7	5.8
	5	1.3	12.8	12.2	12.6	12.1	6.5	6.7
	10	4.7	13.4	12.9	13.4	12.8	7.8	7.9
street	0	-4.6	8.8	8.3	8.7	8.2	4.2	4.5
	5	1.6	12.1	11.5	11.9	11.4	6.5	6.6
	10	4.4	13.3	12.7	13.1	12.6	7.7	8.2
volvo	0	-4.0	13.4	12.9	13.4	12.8	4.7	4.8
	5	1.0	13.6	13.2	13.5	13.1	6.5	6.7
	10	4.9	13.9	13.5	13.7	13.4	8.9	8.5

VII. ACKNOWLEDGMENT

The authors would like to thank Dr. Hossein Shamsi, the assistant professor of K.N Toosi University of Technology, Tehran, Iran, for his help at the edit of the paper.

APPENDIX.1: PERCEPTUALLY WEIGHTED DISTANCE FUNCTION

In [17], a perceptually weighted distance function is proposed for vector quantization (VQ) of the STFT amplitudes of the speech signal. In [17], the perceptually weighted distance function between vectors $X^i=[X^i(1),X^i(2),.....,X^i(K)]$ and $X^j=[X^j(1),X^j(2),.....,X^j(K)]$ is defined as shown in relation (A1). In this relation $X_p{}^i$ and $X_p{}^j$ are the preprocessed vectors of X^i and X^j, respectively. The preprocessing operations are described in [17]. Besides, V^i denotes a coefficient that is interpreted as the perceptual weight.

$$D_V(X^i,X^j)=\sum_{k=1}^{K}V^i(k)\,|\,X_p{}^i(k)-X_p{}^j(k)\,|^2 \quad \text{(A1)}$$

In order to limit $D_V(X^i,X^j)$ between 0 to 1, $X_p{}^i$, $X_p{}^j$ and V^i are normalized as follows where $\hat{X}_p{}^i$, $\hat{X}_p{}^j$ and $\hat{V}^i$ are the normalized vectors.

$$\hat{X}_p{}^i(k)=\frac{X_p^i(k)}{\sqrt{\sum_{k=1}^{K}X_p^{i\,2}(k)}},\qquad \hat{X}_p{}^j(k)=\frac{X_p^j(k)}{\sqrt{\sum_{k=1}^{K}X_p^{j\,2}(k)}} \quad \text{(A2)}$$

$$\hat{V}^i(k)=\frac{V^i(k)}{\sum_{k=1}^{K}V^i(k)} \quad \text{(A3)}$$

According to (A2) we have:

$$|\,\hat{X}_P^i\,|=\sqrt{\sum_{k=1}^{K}\hat{X}_P^{i\,2}(k)}=1,\qquad |\,\hat{X}_P^j\,|=\sqrt{\sum_{k=1}^{K}\hat{X}_P^{j\,2}(k)}=1 \quad \text{(A4)}$$

Therefore, by using the well-known triangle inequality theorem, the following relation is obtained:

$$|\,\hat{X}_P^i-\hat{X}_P^j\,|^2\le 4 \quad \text{(A5)}$$

On the other hand, since $0<\hat{V}^i(k)\le 1$, thus:

$$\hat{V}^i(k)\,|\,\hat{X}_P^i(k)-\hat{X}_P^j(k)\,|^2\le|\,\hat{X}_P^i(k)-\hat{X}_P^j(k)\,|^2 \quad \text{(A6)}$$

Finally the perceptually weighted distance function between X^i and X^j is defined as shown in relation (A7). Considering relations (A5) and (A6), it is easily understood that $PWDF(X^i,X^j)$ is between 0 to 1.

$$PWDF(X^i,X^j)=\frac{1}{4}\sum_{k=1}^{K}\hat{V}^i(k)\,|\,\hat{X}_p^i(k)-\hat{X}_p^j(k)\,|^2 \quad \text{(A7)}$$

This relation is used in this paper as the perceptually weighted distance function.

APPENDIX.2: SPECTRAL CENTER-OF-GRAVITY

Assume $y(n)=x(n)+w(n)$, so for each frame we have:

$$\overline{Y}(k)=\overline{X}(k)+\overline{W}(k)\quad k=1,2,\cdots,K \quad \text{(A8)}$$

that $\overline{Y}$, $\overline{X}$, and $\overline{W}$ are the STFT of y, x, and w, respectively. Moreover, the variables n, k, and K represent the time index, frequency index and length of the STFT (FFT length), correspondingly. For each frame the spectral center-of-gravity of signal $y(n)$, $COG_y(p)$, is defined as follows where p denotes the frame number.

$$COG_y(p)=\frac{\sum_{k=1}^{K}k\left|\overline{X}(k)+\overline{W}(k)\right|^2}{\sum_{k=1}^{K}\left|\overline{X}(k)+\overline{W}(k)\right|^2} \quad \text{(A9)}$$

Using the approximation of $\left|\overline{X}(k)+\overline{W}(k)\right|^2\cong\left|\overline{X}(k)\right|^2+\left|\overline{W}(k)\right|^2$, we have:

$$COG_y(p)\cong\frac{\sum_{k=1}^{K}k\left|\overline{X}(k)\right|^2+\sum_{k=1}^{N}k\left|\overline{W}(k)\right|^2}{\sum_{k=1}^{K}\left|\overline{X}(k)\right|^2+\sum_{k=1}^{N}\left|\overline{W}(k)\right|^2} \quad \text{(A10)}$$

Similarly, spectral center-of-gravities of signals $x(n)$ and $w(n)$ are obtained as shown below:

$$COG_x(p)=\frac{\sum_{k=1}^{K}k\left|\overline{X}(k)\right|^2}{\sum_{k=1}^{K}\left|\overline{X}(k)\right|^2} \quad \text{(A11)}$$

$$COG_w(p)=\frac{\sum_{k=1}^{K}k\left|\overline{W}(k)\right|^2}{\sum_{k=1}^{K}\left|\overline{W}(k)\right|^2} \quad \text{(A12)}$$

Thus, relation (A10) is written as follows:

$$COG_y(p)=\frac{E_x(p)\cdot COG_x(p)+E_w(p)\cdot COG_w(p)}{E_x(p)+E_w(p)} \quad \text{(A13)}$$

where $E_x(p)$ and $E_w(p)$ are:

$$E_x(p)=\sum_{k=1}^{K}\left|\overline{X}(k)\right|^2 \quad \text{(A14)}$$

$$E_w(p)=\sum_{k=1}^{K}\left|\overline{W}(k)\right|^2 \quad \text{(A15)}$$

In addition, we have:

$$\min\{COG_x(p),COG_w(p)\}\le COG_x(p)\le\max\{COG_x(p),COG_w(p)\} \quad \text{(A16)}$$

$$\min\{COG_x(p),COG_w(p)\}\le COG_w(p)\le\max\{COG_x(p),COG_w(p)\} \quad \text{(A17)}$$

Performing a few manipulations on relations (A16) and (A17), the following formula is obtained:

$$\min\{COG_x(p),COG_w(p)\}\le\frac{E_x(p)\cdot COG_x(p)+E_w(p)\cdot COG_w(p)}{E_x(p)+E_w(p)}\le\max\{COG_x(p),COG_w(p)\} \quad \text{(A18)}$$

Finally, considering relations (A13) and (A18), we have:

$$\min\{COG_x(p),COG_w(p)\}\le COG_y(p)\le\max\{COG_x(p),COG_w(p)\} \quad \text{(A19)}$$

In order to verify relation (A19), $COG_y(p)$, $COG_x(p)$, and $COG_w(p)$ are determined in each frame for noisy speech, speech and noise, respectively. The noisy speech is a sentence from the Cooke database, which is contaminated by the babble noise. In figures A1.a to A1.c, COG_y, COG_x, and COG_w are shown for a noisy speech with SNR levels equal to 5dB, 0dB, and

-5dB, respectively. It is observed that for most of the frames the relation (A19) is satisfied. However, there are a few numbers of frames that do not exactly satisfy relation (A19). It is because of the approximation used in relation (A10). This inaccuracy can be neglected with a good approximation.

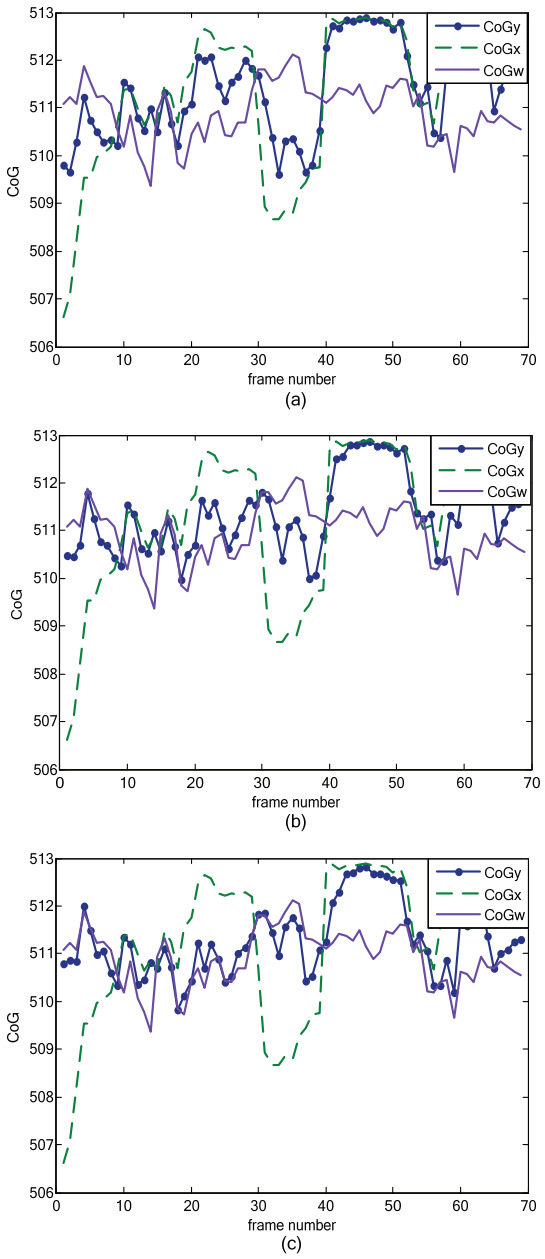

Figure. A1. COG_y, COG_x, and COG_w for a noisy speech with SNR levels equal to (a) 5 dB, (b) 0 dB, and (c)

REFERENCES

[1] Sriram Srinivasan, Jonas Samuelsson, and W. Bastiaan Kleijn, codebook Driven Short-Term Predictor Parameter Estimation for Speech Enhancement", IEEE Transactions on Speech and Audio Processing, vol.14, no.1, pp.163-176, January 2006.

[2] Saeed V. Vaseghi, "Advanced Digital Signal Processing and Noise Reduction", Second Edition, Wiley, 2000, ch. 11.

[3] Saeed. V. Vaseghi, "Advanced Signal Processing and Digital Noise Reduction", Wiley, 1998, ch. 6.

[4] Sriram Srinivasan, "Knowledge-Based Speech Enhancement", Ph.D. Thesis, School of Electrical Engineering KTH - Royal Institute of Technology, Department of Signals, Sensors and Systems, Sound and Image Processing Laboratory, 2005.

[5] Saeed Ayat, Mohamad T. Manzuri, Roohollah Dianat, and Jahanshah Kabudian, "An Improved Spectral Subtraction Speech Enhancement System by Using an Adaptive Spectral Estimator", Canadian Conference on Electrical and Computer Engineering, pp.261-264, May 2005.

[6] Yang Gui and H. K. Kwan, "Adaptive Subband Wiener Filtering for Speech Enhancement using Critical-Band Gammatone Filterbank", Midwest Symposium on Circuits and Systems, pp.732-735, August 2005.

[7] Y. Ephraim and H. L. van Trees, "A signal subspace approach for speech enhancement", IEEE Transactions on Speech and Audio Processing, vol.3, no.4, pp.251-266, July 1995.

[8] Chiung-Wen Li and Sheau-Fang Lei, "Signal subspace approach for speech enhancement in nonstationary noises", International symposium on communication and information technologies, pp.1580-1585, 2007.

[9] Sharon Gannot, David Burshtein and Ehud Weinstein, "Iterative and Sequential Kalman Filter-Based Speech Enhancement Algorithms", IEEE Transactions on Speech and Audio Processing, vol.6, no.4, pp.373-385, July 1998.

[10] Rainer Martin, "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics", IEEE Transaction on Speech and Audio Processing, vol.9, no.5, pp.504-512, July 2001.

[11] Y. Ephraim, D. Malah, and B. H. Juang, “On the application of hidden Markov models for enhancing noisy speech”, IEEE Transactions on Speech and Audio Processing, vol.37, pp.1846-1856, 1989.

[12] Hossein Sameti, Hamid Sheikhzadeh, Li Deng, and Robert L. Brennan, "HMM-Based Strategies for Enhancement of Speech Signals Embedded in Nonstationary Noise", IEEE Transactions on Speech and Audio Processing, vol.6, no.5, pp.445-455, September 1998.

[13] Athanasios Mouchtaris, Jan Van der Spiegel, Paul Mueller, and Panagiotis Tsakalides, “A Spectral Conversion Approach to Single-Channel Speech Enhancement, IEEE Transactions on audio, Speech and Language Processing, vol.15, no.4, pp.1180-1193, May 2007.

[14] Yi Hua, and Philipos C. Loizou, ” Environment-specific noise suppression for improved speech intelligibility by cochlear implant users”, J. Acoust. Soc. Am., vol. 127, no. 6, pp.3689-3695, June 2010.

[15] Alan V.Oppenheim, Ronald W. Schafer with John R.Buck, Discrete-Time Signal Processing, Second Edition, pp. 811_820, 1999.

[16] Roghayeh Doost, Abolghasem Sayadian, and Hossein Shamsi, "A New Method for Low SNR Estimation of Noisy Speech Signals Using Fourth-Order Moments", IEICE TRANS.INF.&SYST., vol. E93-D, no.6, pp.1599-1607, June 2010.

[17] Roghayeh Doost, Abolghasem Sayadian, and Hossein Shamsi, "A New Perceptually Weighted distance Measure for Vector Quantization of the STFT Amplitudes in the Speech Application", IEICE Electronic Express, vol.6, no.12, pp.824-830, June 25, 2009.

[18] Werner Verhelst, "Overlap-add methods for time-scaling of speech", vol.30, no.4, pp.207-221, April 2000.

[19] Y. Linde, A. Buzo, and R.M. Gray, "An Algorithm for Vector Quantizer Design", IEEE Transactions on Communications, vol. 28, no.1, pp.84-95, 1980.

[20] J. Janda, "Age Dependence of Children’s Speech Parameters", Acta Polytechnica, vol.49, no.2-3, pp.40-43, 2009.

[21] Youyi Lu and Martin Cooke, "Speech production modifications produced in the presence of low-pass and high-pass filtered noise", J. Acoust. Soc. Am., vol.126, no.3, pp.1495-1499, September 2009.

[22] M. Cooke, J.R. Hershey, and S.J. Rennie, “Monaural speech separation and recognition challenge,” Elsevier Computer

Speech and Language, vol. 24, no. 1, 2010, pp. 1–15.
[23] Thiago de M. Prego and Sergio L. Netto, "Efficient Search in the Adaptive Codebook for ITU-T G.729 Codec", IEEE Signal Processing Letters, vol.16, no.10, pp.881-884, October 2009.

Roghayeh Doost was born in Tehran, Iran, in Feb. 1980. She received her B.Sc. degree in electronics engineering from Iran University of Science and Technology, Tehran, Iran, in 2002, and M.Sc. and Ph.D. degrees in Communication Systems Engineering from Amirkabir University of Technology, Tehran, Iran in 2005 and 2011, respectively. Her interests include digital signal processing, speech processing and watermarking.

Abolghasem Sayadian received his B.Sc. degree from Amirkabir University of Technology in 1980, his M.Sc. degree from Isfahan University of Technology in 1987, and his Ph.D. degree from Tarbiat Modares University in 1993, all in electrical engineering. Dr. Sayadian has been an associate professor in the Electrical Faculty, Amirkabir University of Technology since 1991. He is the author or co-author of more than 120 international and national journal and conference publications on speech processing. His research interests include digital signal processing, speech processing and watermarking.