WEIGHTED CODEBOOK MAPPING FOR NOISY SPEECH ENHANCEMENT USING HARMONIC-NOISE MODEL

Esfandiar Zavarehei, Saeed Vaseghi, Qin Yan

Department of Electronic and Computer Engineering,
Brunel University, London, UK
{esfandiar.zavarehei, saeed.vaseghi, qin.yan}@brunel.ac.uk

10.21437/Interspeech.2006-70

Abstract

Most noisy speech enhancement methods result in partial suppression and distortion of speech spectrum. At instances when the local signal-to-noise ratio at a frequency band is very low speech partials are often obliterated. In this paper a method for enhancement and restoration of noisy speech based on a harmonic-noise model (HNM) is introduced. A HNM imposes a temporal-spectral structure that may reduce processing artifacts. The restoration process is enhanced through incorporation of a prior HNM of clean speech stored in a pre-trained codebook. The restored speech is a SNR-dependent combination of the de-noised observation and the speech obtained from weighted codebook mapping. The additional improvements of speech quality resulting from the proposed method in comparison to conventional and modern speech enhancement systems are evaluated. The results show that the proposed method improves the quality of noisy speech and restores much of the information lost to noise.

Index Terms: speech enhancement, codebook mapping, harmonic noise model

1. Introduction

Enhancement of speech quality impaired by background noise is one of the most challenging issues in the field of speech processing. The ongoing competitive research for development of more suitable mathematical models of speech and more robust methods for estimation (and perhaps tracking) of noise and speech parameters are evidence of the complexity of the problem and the importance of new solutions.

Common speech enhancement methods suppress the background noise through application of an adaptive gain applied to the short-time spectral amplitude (STSA) of the noisy speech signal [1][2]. Generally, the suppression gain of these estimators depends on the estimates of signal-to-noise ratio (SNR) spectrum. The inaccuracy of the estimates of the SNR, together with the non-optimal assumptions underlying the derivation of the estimators (e.g. speech and noise distributions) result in two major artifacts observed in the enhanced signal: i) the residual noise also known as musical noise, and ii) distortion and suppression of speech signal and in particular the harmonic structure [3][4]. These artifacts are particularly observed at instances when the SNR spectrum at a frequency band is very low.

One of the main issues addressed in this paper is a common form of processing distortion, namely severe suppression of parts spectrogram of speech, which is common to most speech enhancement methods in moderate to low SNR conditions. The major objectives of this work are: i) to apply HNM model of speech for reducing the amount of residual noise, and ii) to incorporate the prior information about the speech into the system in order to recover at least some of the distorted/suppressed speech.

Figure 1 shows the block diagram of the speech enhancement system. The system is divided to three sub-systems separated with dotted lines. The first sub-system is the simplified block diagram of a conventional noise reduction system. The *a priori* SNR estimation block [1] is not shown here and is assumed to be included in the noise reduction block. The second sub-system is the speech modelling module. This part is discussed in section 2. The last stage of the block diagram of Figure 1 is the reconstruction sub-system. At this stage the distorted/suppressed portions of speech are recovered using a weighted codebook mapping explained in section 3. In Section 4 the overall performance of the system is evaluated and practical issues are presented. Conclusions are drawn in section 5.

2. Harmonic-Noise Model

HNMs are widely employed in speech processing systems for source coding [5], speech enhancement [6], etc. In this paper a variant of HNM is used which produces high quality synthesized speech. Speech frames are analyzed and synthesized entirely in spectral amplitude domain. Three parameters are extracted from each harmonic sub-band: amplitude A_k, harmonicity V_k and harmonic central frequency f_k. The unprocessed phase spectrum is used for re-synthesis. The harmonic amplitudes represent the square-root energy of the corresponding sub-band and the harmonicity degree is a real-valued measure between 0 and 1 representing the voicing degree of each sub-band.

Given the set of HNM parameters for a speech frame the spectral amplitude is regenerated as a weighted summation of a Gaussian-shaped spectral amplitude function, $G(f)$ and a Rayleigh distributed [1] random spectral amplitude, $R(f)$:

$$X_{HNM}(f) = \sum_{k=1}^{N} A_k \left(V_k G(f - f_k) + (1 - V_k) R(f - f_k) \right) \qquad (1)$$

where $X_{HNM}(f)$ is the HNM-synthesized amplitude spectrum, N is the number of harmonics, A_k, V_k and f_k are the corresponding

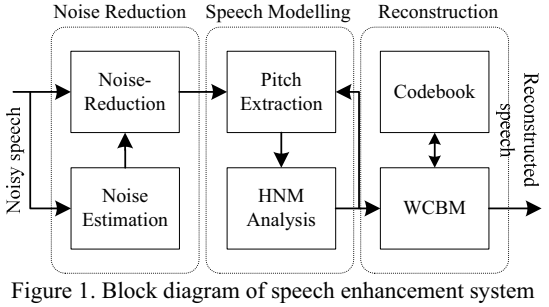

Figure 1. Block diagram of speech enhancement system

HNM parameters of the k^{th} harmonic sub-band.

2.1. Harmonic and Fundamental Frequency Tracking

The fundamental frequency (or pitch) track of the signal is extracted using a method that utilises the autocorrelation function for an initial estimate of pitch track and spectral matching for a subsequent refined estimate described in [7].

Due to possible inaccuracy of pitch estimation algorithm multiples of the pitch frequency might not exactly coincide with true harmonic frequencies. Although this displacement may be insignificant on its own, it affects the calculation of other HNM parameters. Hence the exact location of each harmonic frequency is extracted locally to maximize the harmonicity of that harmonic, i.e. to fit best to that harmonic:

$$f_k = \arg\max\left\{V_k\left(f_{search}\right)\right\} \quad kF_0 - \Delta F_H < f_{search} < kF_0 + \Delta F_H \quad (2)$$

where F_0 is the fundamental frequency and ΔF_H is the search range empirically set to 30Hz. The sensitivity of the algorithm to errors in the estimate of the fundamental frequency due to the background noise is discussed in section 4.

2.2. HNM Parameter Extraction

Given that $X(f)$ is the spectral amplitude of an arbitrary frame, HNM parameters are extracted as:

$$V_k = V_k\left(f_k\right) = 1 - \frac{\sqrt{\sum_{f=f_k-\Delta f}^{f_k+\Delta f}\left(X(f) - A_k G(f-f_k)\right)^2}}{A_k} \quad (3)$$

$$A_k = \sqrt{\sum_{f=f_k-\Delta f}^{f_k+\Delta f} X^2(f)} \quad (4)$$

where $2\Delta f$ is the harmonic bandwidth and:

$$G(f) = \alpha \exp\left(-\left(\frac{2.2f}{60}\right)^2\right) \quad (5)$$

$$\sum_{f=-\Delta f}^{\Delta f} G^2(f) = 1 \quad (6)$$

$$\sum_{f=-\Delta f}^{\Delta f} R^2(f) = 1 \quad (7)$$

Although the result of right hand side of equation (3) is not guaranteed to be between 0 and 1, in practice this is always the case. Equation (5) is the Gaussian-shaped function used to model the harmonics. In [7] a similar method is used to model the harmonics and make the voicing decisions, using the spectral amplitude of the hamming window. The parameter α in Equation (5) is calculated so that the energy of the $G(f)$ in the determined range of $[-\Delta f, \Delta f]$ is equal to 1. The value of α depends on the FFT size, sampling frequency and harmonic sub-band range Δf. In practice, one may use an adaptive value for Δf, based on the fundamental frequency, to synthesize the whole range between each two harmonics. However, as the value of $G(f)$ approaches zero for rather large values of f, we decided to use a fixed value of Δf=60Hz for analysis and synthesis purposes. The value of α for a FFT size of 1024 at a sampling rate of 8 KHz is calculated as α=0.4416. Figure 2 shows $G(f)$ for obtained using mentioned set of values.

The HNM model enforces a harmonic structure on speech signals. Synthesizing a noisy speech using its HNM parameters emphasizes the harmonic structure of speech and

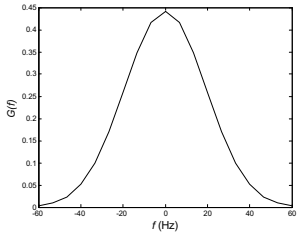

Figure 2. The Gaussian window used to model harmonics

smoothes out some of the background noise. The effect of such a process on quality of noisy speech signals is measured with Perceptual Evaluation of Speech Quality (PESQ) and illustrated for different SNRs in Figure 3. A set of 100 utterances are randomly selected from wall street journal (WSJ) database for evaluation purposes.

3. Weighted Codebook Mapping (WCBM)

Real-life noises often have a low-pass spectrum and most of their energy is concentrated in a rather limited frequency band. As a result of applying noise reduction these parts of speech signals are sometimes over-suppressed and distorted. In order to reconstruct the speech spectrum in these frequencies, a weighted codebook-mapping (WCBM) method is implemented to incorporate prior information on clean speech structure into the estimation process.

3.1. Codebook training

A codebook is trained on energy-normalized harmonic values of clean speech and used in the restoration of the harmonic amplitudes that have been severely distorted or drowned in noise. The harmonic amplitude values are normalized as

$$\mathbf{B} = \frac{\mathbf{A}}{\sqrt{\sum_k A_k^2}} \quad (8)$$

where $\mathbf{A}=[A_1, A_2, \ldots, A_N]$ and $\mathbf{B}=[B_1, B_2, \ldots, B_N]$ are the harmonic amplitude and normalized harmonic amplitude vectors respectively. The size of codebook is experimentally set to 1024 and the K-means algorithm is used for training the codebook. The rationale for normalizing energy of the harmonic amplitudes is that the codebook becomes energy-independent while preserving the shape of the spectrum. Euclidian distance is used for clustering:

$$l = \arg\min_m \left\{\sum_k \left(B_k - C_{k,m}\right)^2\right\} \quad \text{for training} \quad (9)$$

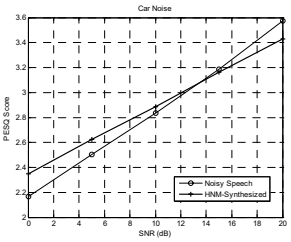

Figure 3. The effect of HNM analysis/synthesis on noisy speech (car noise)

where l is the nearest codeword to the data vector $\mathbf{B}$, and $\mathbf{C}_m$ is the m^{th} codeword of the codebook.

The speaker-independent normalized harmonic amplitude codebook is trained on speech utterances from WSJ database. A total of 420 utterances are randomly selected from different speakers. The total number of training (data) vectors is equal to 300107 where the time shift between successive vectors is 10ms, i.e. the total training speech is about 50 minutes.

Furthermore, another codebook is trained on the harmonicity degree of the harmonic sub-bands, V_ks, using the same database and using K-means algorithm.

3.2. WCBM Algorithm

The HNM parameters of the output of the noise reduction module are extracted. The weighted distance between HNM parameters of the noise-reduced signal and the codewords are estimated as:

$$D_m = \sum_k \left(W_k \left(B_{NR,k} - C_{k,m} \right) \right)^2 \tag{10}$$

where $B_{NR,k}$ is energy-normalized amplitude of the k^{th} harmonic obtained from "noise-reduced" HNM amplitude vector, $\mathbf{A}_{NR}$:

$$\mathbf{B}_{NR} = \frac{\mathbf{A}_{NR}}{\sqrt{\sum_k A_{NR,k}^2}} \tag{11}$$

and W_k is the weight of the k^{th} harmonic. The value of W_k is between 0 and 1 and is used as a measure of reliablility of the noise-reduced estimate. The *a priori* SNR of the signal may be used for obtaining these weights:

$$\begin{aligned} \xi_k &= \sum_{f=f_k-\Delta f}^{f_k+\Delta f} \xi(f) \\ W_{k,\xi} &= \left(\log(\xi_k) - \min\left[\log(\xi_k) \right] \right) \Big/ \max\left[\log(\xi_k) \right] \end{aligned} \tag{12}$$

where ξ_k and $\xi(f)$ are a priori SNR of the k^{th} sub-band and frequency f respectively. Several different algorithms are proposed in the literature for estimation of a priori SNR, from decision-directed method introduced by Ephraim [1], to the most recent algorithms such as non-causal a priori estimation [2]. Furthermore, to emphasize the effect of lower harmonics on estimation, the weights of Equation (12) are weighted with a function of frequency. A fixed frequency-dependent weight may be used for this purpose. In this work we use the following fixed weight function to modify the weights calculated in Equation (12):

$$W_k = 0.125\left(\cos\left(2\pi f_k / F_s \right) + 7 \right) . W_{k,\xi} \tag{13}$$

An estimate of the HNM amplitude vector is obtained from the L codewords with lowest distances as:

$$\mathbf{B}_{CB} = \sum_j q_j \mathbf{C}_j \tag{14}$$

where q_j is the weight of the codeword $\mathbf{C}_j$ and is proportional to the reciprocal of the distance of the codeword $\mathbf{C}_j$ from HNM normalized amplitudes. The resulting energy-normalized vector needs to be de-normalized, before combining with the noise reduced vector:

$$\mathbf{A}_{CB} = \alpha_{dn} \mathbf{B}_{CB} \tag{15}$$

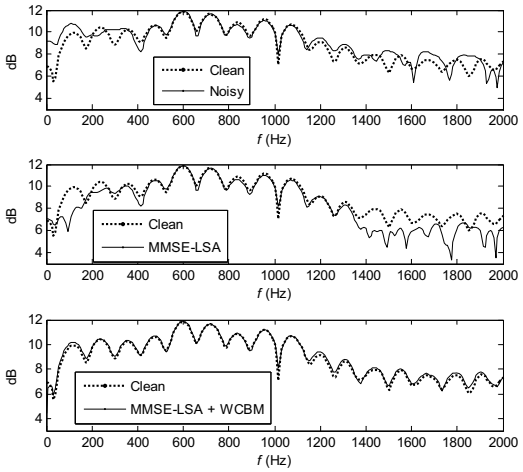

Figure 4. Spectral amplitude of clean speech superimposed on that of (top) noisy (middle) de-noised with MMSE-LSA and (buttom) de-noised with MMSE-LSA and restored using WCBM

$$\begin{aligned} \alpha_{dn} &= \arg\min_{\alpha} \left\{ \mathrm{E}\left[\sum_k \left(W_k \left(A_{NR,k} - \alpha B_{CB,k} \right) \right)^2 \right] \right\} \\ &= \frac{\sum W_k A_{NR,k} B_{CB,k}}{\sum W_k B_{CB,k}^2} \end{aligned} \tag{16}$$

Note that Equation (16) α_{dn} is calculated so that elements of A with corresponding higher weights are least changed. The resulting vector, $\mathbf{A}_{CB}$ is then combined with the noise-reduced vector, $\mathbf{A}_{NR}$ in a way that elements with higher weights, W_k, are less affected than those with lower weights:

$$\hat{A}_k = W_k A_{NR,k} + \left(1 - W_k \right) A_{CB,k} \tag{17}$$

The harmonicities of harmonics are mapped to the harmonicity codebook using a similar procedure, without using the energy-normalization.

4. Performance Evaluation

Figure 4 illustrates the amplitude spectrum of a sample speech frame from 0 to 2 KHz. The top figure shows the effect of noise spectrum (at SNR=5dB Train noise). The middle figure shows the effect of MMSE log spectral amplitude (LSA) estimator [1] noise reduction algorithm on the same noisy spectrum. It is evident that the harmonic structure of the spectrum is distorted and suppressed in frequency bands where noise is dominant. The bottom plot shows the effect of WCBM on this frame. The harmonic structure is restored in terms of amplitude and shape at most harmonics. Similar effects can be seen in Figure 5 where the spectrogram of the signal is illustrated in different states of processing.

The performance of the WCBM system when it is used as a post-processing module with a noise reduction module is evaluated. Three different methods are used for noise reduction and the performance of each of these modules on its own and when connected to the WCBM system is evaluated: i) MMSE-LSA ii) non-causal (NC) a priori SNR estimation with MMSE-LSA [2] and iii) DFT domain Kalman filter with correlated noise DFTK [8]. 100 utterances are randomly selected from WSJ database. Train and babble noises at

Table 1. PESQ score improvement of three different methods with and without WCBM

Noise Type	De-noise Method	Input SNR (dB) 0	5	10	15
Train Noise	MMSE-LSA	0.36	0.47	0.46	0.38
	MMSE-LSA+WCBM	0.48	0.51	0.47	0.38
	NC	0.23	0.32	0.32	0.27
	NC+WCBM	0.48	0.46	0.35	0.27
	DFTK	0.48	0.57	0.57	0.54
	DFTK+WCBM	0.91	0.84	0.75	0.61
Babble Noise	MMSE-LSA	0.11	0.10	0.10	0.09
	MMSE-LSA+WCBM	0.24	0.23	0.18	0.10
	NC	0.27	0.23	0.17	0.09
	NC+WCBM	0.69	0.48	0.32	0.12
	DFTK	0.27	0.22	0.18	0.10
	DFTK+WCBM	0.72	0.49	0.35	0.14

different SNRs are added to the speech signals. PESQ [8] scores and log-spectral distances (LSD) [2] of the noisy and de-noised signals are calculated and averaged. Tables 1 and 2 shows the improvement of PESQ and LSD of the de-noised and reconstructed speech compared to the noisy speech. It is evident that at low SNRs WCBM improves these measures significantly.

To determine the sensitivity of the proposed system to pitch errors, its performance is evaluated in the case when the pitch frequency is extracted from the clean speech. Note that the harmonic frequencies, including the fundamental, are then extracted using the noisy speech from Equation (2). MMSE-LSA is used as the core noise reduction method. A comparison between the results of Tables 1 to 3 shows that using the clean pitch track improves the quality of the resulting speech in terms of PESQ and LSD. Listening tests, however, reveal that the dissimilarities are only audible in instances where a substantial pitch error (e.g. double pitch) has occurred.

5. Conclusion

A weighted codebook mapping method for reconstruction of distorted de-noised speech is introduced. This method uses a simple heuristic technique to incorporate prior information for reviving severely distorted harmonic bands of speech spectrum lost to noise. The performance evaluation results show that

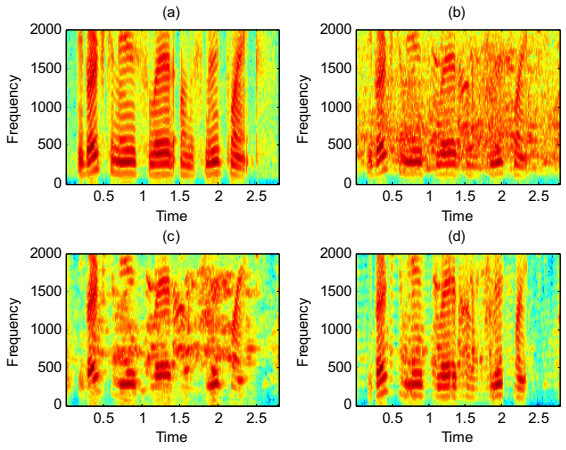

Figure 5. The spectrogram of: (a) clean speech (b) noisy speech (SNR=5dB restaurant noise) (c) MMSE-LSA and (d) MMSE-LSA + WCBM

Table 2. LSD improvement (reduction) of three different methods with and without WCBM

Noise Type	De-noise Method	Input SNR (dB) 0	5	10	15
Train Noise	MMSE-LSA	0.11	0.14	0.13	0.11
	MMSE-LSA+WCBM	0.15	0.15	0.14	0.11
	NC	0.13	0.11	0.11	0.08
	NC+WCBM	0.15	0.14	0.11	0.08
	DFTK	0.15	0.18	0.17	0.15
	DFTK+WCBM	0.27	0.25	0.21	0.19
Babble Noise	MMSE-LSA	0.03	0.03	0.03	0.03
	MMSE-LSA+WCBM	0.07	0.07	0.05	0.03
	NC	0.08	0.07	0.05	0.03
	NC+WCBM	0.21	0.15	0.10	0.04
	DFTK	0.08	0.06	0.05	0.03
	DFTK+WCBM	0.23	0.14	0.11	0.04

Table 3. PESQ and LSD improvement using MMSE-LSA and WCBM with clean pitch

Noise Type	Improvement when clean pitch is used	Input SNR (dB) 0	5	10	15
Train Noise	PESQ	0.61	0.57	0.49	0.38
	LSD	0.19	0.16	0.14	0.11
	Pitch Error %	17	11	6	4
Babble Noise	PESQ	0.33	0.28	0.20	0.15
	LSD	0.12	0.13	0.09	0.06
	Pitch Error %	21	13	7	3

WCBM improves the quality of processed signal. The method could be especially useful if used with vocoders based on HNM model of speech. This possibility is being investigated.

6. References

[1] Ephraim, Y., Malah, D., "Speech enhancement using a minimum mean square error log-spectral amplitude estimator," IEEE Trans. on Acoust., Speech, Signal Processing, vol. ASSP-33, pp. 443-445, Apr. 1985

[2] Cohen, I., "Relaxed Statistical Model for Speech Enhancement and a Priori SNR Estimation", Speech and Audio Processing, IEEE Transactions on vol. 13, Issue 5, Part 2, Sept. 2005 pp. 870 - 881

[3] Plapous, C., Marro, C., Scalart, P., "Speech Enhancement Using Harmonic Regeneration", IEEE Acoustics, Speech, and Signal Processing, 2005. Proc.. vol. 1, pp. 157 - 160

[4] Yu, A.T., Wang, H. C., "New speech harmonic structure measure and it application to post speech enhancement", IEEE Acoustics, Speech, and Signal Processing, 2004. Proc. vol. 1, pp. I - 729-32

[5] Griffin, D.W., Lim, J.S., "Multiband-excitation vocoder", IEEE Trans. Acoust., Speech, and Signal Processing, vol. ASSP-36:2, pp. 236-243, 1988.

[6] Raza, D.G., Cheung-Fat Chan, "Enhancing quality of CELP coded speech via wideband extension by using voicing GMM interpolation and HNM re-synthesis", IEEE Acoustics, Speech, and Signal Processing, 2002. Proc. vol. 1, pp. I-241-I-244

[7] Kondoz, A. M., *Digital Speech: Coding for Low Bit Rate Communication Systems*, John Wiley & Sons, Ltd., 1999

[8] E. Zavarehei, S. Vaseghi, Q. Yan, "Interframe Modelling of DFT Trajectories of Speech and Noise for Speech Enhancement Using Kalman Filters", in Speech Communication, Special Issue on Robustness, 2006, *in press*