NOISE SUPPRESSION BASED ON AN ANALYSIS-SYNTHESIS APPROACH

R. F. Chen, C. F. Chan and H. C. So

Department of Electronic Engineering, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong
ruofechen2@student.cityu.edu.hk, itcfchan@cityu.edu.hk, hcso@ee.cityu.edu.hk

ABSTRACT

In noise suppression methods that are based on an analysis-synthesis approach, speech is enhanced by re-synthesis using acoustic cues extracted from harmonic noise model (HNM) analysis. In this paper, a complete analysis-synthesis framework using HNM is introduced and generalized for different noise conditions. Fine details in choosing and estimating HNM parameters are discussed. Techniques that contribute to robust estimation of HNM parameters are proposed. Evaluation results demonstrate the effectiveness of the proposed HNM-based noise suppression method in low signal-to-noise ratio environments.

1. INTRODUCTION

A noise suppression method based on an analysis-synthesis approach using harmonic noise model (HNM) is proposed in [1] to enhance the speech in the presence of car noise. This method is attractive as it can retrieve damaged speech structure and at the same time remove residual noises such as musical tones, provided that the modelling of HNM is accurate. In such cases, the choice and estimation of model parameters are crucial as they directly affect the perceptual quality of synthetic speech. In this paper, we focus on robust modelling of HNM. Details in choosing auxiliary estimation tools such as preliminary filter and voiced/unvoiced (V/UV) classifier are discussed. Techniques for pitch tracking and harmonic restoration are proposed to improve the robustness of estimating HNM parameters for speech enhancement in different noise environments.

This paper is organized as follows. Section 2 introduces a complete analysis-synthesis framework for applying HNM in speech enhancement. In Section 3, we look into details of each important component of the speech enhancement system and evaluate their interactions within the proposed analysis-synthesis framework. Section 4 shows and discusses the results of performance evaluation and Section 5 concludes this work.

2. ANALYSIS-SYNTHESIS FRAMEWORK

The proposed speech enhancement system comprises two stages, namely speech analysis stage and speech synthesis stage. At speech analysis stage, noisy speech is initially pre-cleaned by a classical speech enhancement algorithm. Acoustic features are extracted from the pre-cleaned speech through spectral analysis. Pitch frequency and harmonic magnitudes are estimated based on the spectrum matching between the pre-cleaned spectrum and the excitation spectrum. V/UV decision of each frame is made according to the total energy of pre-cleaned signals. Mixing function is calculated from the spectral envelope of pre-cleaned signals. Residual energy is obtained from both spectral envelope and mixing function. The refined pitch, harmonic magnitudes, residual energy, and mixing function are passed to the synthesis stage. At speech synthesis stage, voiced speech is synthesized in time domain to allow a smooth evolution of fundamental frequency from frame to frame. The amplitude function is linearly interpolated between frames with V/UV band information while a quadratic phase interpolation is resulted from linearly interpolated harmonic frequencies. Unvoiced speech is also synthesized in time domain. The weighted power spectrum is converted to autocorrelation data and then an all-pole linear predictive coding (LPC) model is fitted to the autocorrelation data to compute the synthesis filter's residual signal gain. Random Gaussian noise is generated and fitted into the synthesis filter to produce the unvoiced speech signal. The resulting synthesized speech is simply the sum of voiced and unvoiced speech. The block diagram of speech analysis and synthesis are shown in Figure 1.

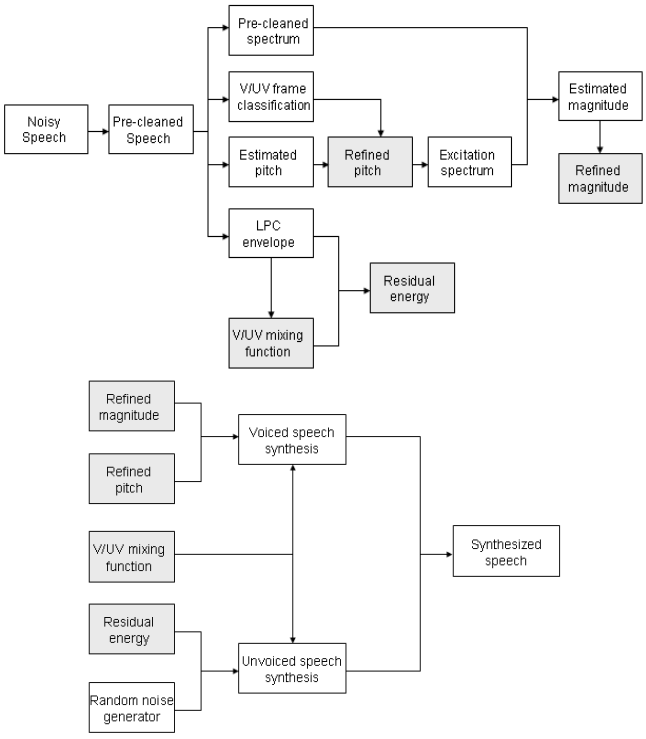


Figure 1 - Block diagram of speech analysis and synthesis

 ISSN 2076-1465

3. HNM MODELLING

3.1 Pre-cleaning

At the first stage, the degraded speech is initially enhanced by a classical speech enhancement algorithm. The major reason of doing so is that the distortion between clean speech spectrum and noise corrupted spectrum will be large in low signal-to-noise ratio (SNR) environment. For this reason, spectrum matching based pitch estimation method will have large errors if it is directly applied on original degraded speech. It is confirmed that pre-cleaning procedure becomes necessary when noise is aggressive (e.g. SNR<15dB). Hence the choice of the pre-cleaning algorithm is significant and it is selected based on the following three criteria: (i) the pre-cleaning algorithm should be able to restore some distorted harmonics; (ii) average noise level should be suppressed to a certain level; and (iii) suppression should not be too aggressive to distort those dominant formants. Speech enhancement algorithms used in [2] have been tested as a pre-cleaning tool to evaluate their interactions with the proposed HNM-based system. Results show that the minimum mean square error incorporating signal presence uncertainty (MMSE-SPU) algorithm [3] is the best choice based on the above criteria.

3.2 Pitch Estimation

The proposed analysis-synthesis framework is developed based on the prototype of multi-band excitation (MBE) coding. In MBE analysis, the optimum pitch period is obtained by minimizing an error function returned from the spectrum matching between an input spectrum $S(k)$ and an excitation spectrum $E(\tau,k)$ obtained from Fourier transform magnitude of a windowed impulse train with pitch period τ. An improved measure is proposed in [4] and adopted in this work to reduce the gross pitch errors owing to pitch doublings. The improved error measure is defined as

$$\epsilon(\tau) = \frac{1}{1-\tau B}\left[\frac{\sum_{m=1}^{M(\tau)}\sigma_m(\tau)[1-\varphi_m(\tau)]}{\sum_{m=1}^{M(\tau)}\sigma_m(\tau)} + \frac{\sum_{m=1}^{M(\tau)}[1-\varphi_m(\tau)]}{M(\tau)}\right] \quad (1)$$

where

$$\sigma_m(\tau) = \sum_{k=a_m(\tau)}^{b_m(\tau)} |S(k)|^2 \quad (2)$$

$$\varphi_m(\tau) = \frac{\left[\sum_{k=a_m(\tau)}^{b_m(\tau)} |S(k)||E(\tau,k)|\right]^2}{\alpha_m(\tau)\sum_{k=a_m(\tau)}^{b_m(\tau)} |E(\tau,k)|^2} \quad (3)$$

and B is a weighting factor for biasing the pitch dependent error, $M(\tau)$ is total number of bands in the speech spectrum, $a_m(\tau)$ and $b_m(\tau)$ are the lower and upper boundaries of the m-th harmonic band, respectively. This algorithm is very robust in clean environment and no extra pitch tracking procedure is required. However, in low SNR environment, the input spectrum would be severely distorted even after pre-cleaning. In order to reduce the mismatches, two post-processing techniques are proposed to improve the robustness of pitch estimation.

Initially, the algorithm searches the whole range of the pitch period that covers all possible fundamental frequencies of human voice. The assumption made here is that only one dominant speaker is talking at a time. Hence we assume that the evolution of pitch contour should be smooth, and it would not undergo abrupt fluctuation during a single voiced period. In this scenario, the first five consecutive robust pitch values estimated in a single voiced period, i.e., those pitch values which are derived with the normalized matching error less than a predefined threshold, are averaged to form a baseline pitch value. This value is updated using first order infinite impulse response (IIR) smoothing once a new robust pitch value is available and it is reset when a single voiced period is over. The pitch period searching is then refined to a narrower range based on the baseline value during this voiced period. There are two advantages of doing so: (i) it will reduce the gross pitch errors such as double pitch errors by neglecting the out-of-range values; and (ii) it will reduce the computational load significantly. To further reduce the gross pitch errors, we place an additional favor to those neighboring pitch periods of last estimated pitch period during searching process. Assume $\tau(l)$ is the pitch period in ms at l-th frame, and $\epsilon(\tau)$ is the improved error function, and then modified error function $\hat{\epsilon}(\tau)$ is defined as:

$$\hat{\epsilon}(\tau) = \kappa(i)\,\epsilon(\tau) \quad (4)$$

where

$$\kappa(i) = \begin{cases} \phi, & \text{for } \lfloor \frac{\alpha N}{f_s \tau(l-1)} - \beta \rfloor \le i \le \lfloor \frac{\alpha N}{f_s \tau(l-1)} + \beta \rfloor \\ 1, & \text{otherwise} \end{cases} \quad (5)$$

and i is the searching index, N is the fast Fourier transform (FFT) size, f_s is the sampling frequency, α is the upsampling factor, β is the offset with favor, ϕ is the weighting factor which is empirically set to 0.88, and $\lfloor . \rfloor$ stands for truncation to its lower closest integer value, respectively.

3.3 Harmonic Magnitude Estimation

Harmonic magnitudes are estimated in each frame for accurate restoration of harmonic structure in synthetic speech. Practically, for application of harmonic analysis in speech coding, band magnitudes are derived and encoded as the LPC spectral envelope of the original spectrum because of the compact representation of LPC parameters for low bit-rate transmission. However, in the application of speech enhancement, there is no strict limitation on the transmission rate, so the band magnitudes can be estimated directly from the original spectrum as

$$A_m(\tau_0) = \frac{\sum_{k=a_m(\tau_0)}^{b_m(\tau_0)} |S(k)||E(\tau_0,k)|}{\sum_{k=a_m(\tau_0)}^{b_m(\tau_0)} |E(\tau_0,k)|^2} \quad (6)$$

provided that the pitch $\tau_0 = \arg\min_{\tau}[\hat{\epsilon}(\tau)]$ has been accurately estimated. This approach offers accurate and robust estimation for noise-free spectra. However, in low SNR environment, it is typical that only those dominant harmonics remain after the pre-cleaning process. A post-processing technique is proposed to restore missing harmonics. Figure 2 illustrates the post-processing technique for magnitude restoration.

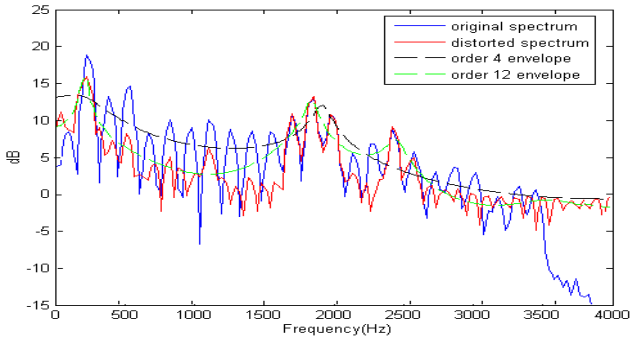


Figure 2 - Magnitude restoration using low-order LPC envelope

It is noticed that the region between dominant formants is over-suppressed after pre-cleaning. A lower order (e.g. order 4) LPC spectral envelope is applied to scale up the over-suppressed harmonics between formants (i.e. resample the harmonic magnitudes from the lower order envelope). Experiments have been carried out to evaluate this technique. It is testified that the perceptual quality is sensitive to the modification of magnitudes. Consequently, a conservative strategy of using this technique is adopted for magnitude post-processing, i.e., only those voiced frames with large drift in consecutive harmonics (e.g. difference >10dB) between formants are shaped using a lower order envelope. The employment of this post-processing technique generally gives better results in most of objective measures such as SNR gain and perceptual evaluation of speech quality (PESQ).

3.4 V/UV frame classification

In HNM analysis, a voice activity detector (VAD) is needed to classify voiced and unvoiced frames as only those frames labelled as voiced are applied with the aforementioned pitch estimation technique while for those unvoiced frames pitch values are set to zero. A total of three attempts are made for selecting a suitable VAD for HNM-based speech enhancement system. In clean speech analysis, the modified matching error function described in (4) is sufficient. However, some experiments have been conducted, showing that a good cut-off threshold that gives satisfactory V/UV classification in very low SNR environments still cannot be determined. A statistical VAD suggested in [1] is proven to be effective in most cases and no additional calculation is needed since it is already obtained in the pre-cleaning process. Alternatively, signal energy of pre-cleaned speech is found to be effective as a V/UV frame classification tool. This method generally offers a better classification than statistical VAD as it offers the flexibility to take into account which portion of the spectrum to compute the total energy. As a result, certain frequency regions can be exempted from VAD decision if it is deteriorated by strong colored noise.

3.5 V/UV harmonic band classification

V/UV harmonic band classification is used within an individual frame to guide harmonic magnitude interpolation in voiced speech synthesis. Basically, there are two approaches to label V/UV harmonic bands for different spectral regions. In the first approach, matching error returned in each harmonic band is used while in the second approach a V/UV mixing function adopted in [1] is employed. The latter approach is more robust in low SNR environment as illustrated in Figure 3.

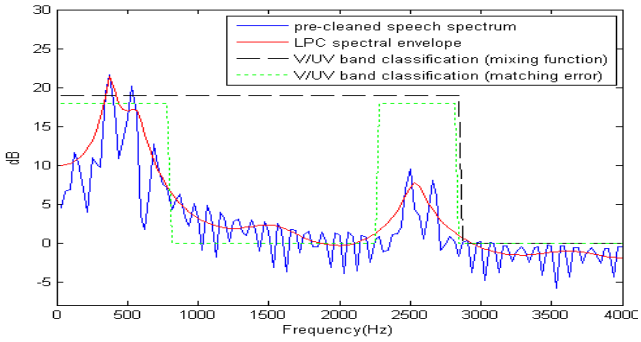


Figure 3 - Comparison of V/UV band classification

It shows a typical over-suppressed frame after pre-cleaning. It is observed that the dominant harmonics are retained while other minor harmonics are highly suppressed to insignificant level comparable to high-frequency noisy region. Consequently, the matching error approach may misclassify these over-suppressed harmonics as unvoiced, and hence they cannot be restored during synthesis stage. On the other hand, the mixing function approach is calculated based on less-sensitive LPC spectral envelope and the searching starts from the high-frequency region. Once the transition point is found, all of the harmonic bands in lower frequency regions are classified as voiced and the rest are classified as unvoiced.

4. PERFORMANCE EVALUATION

To correlate with the subjective evaluation conducted in [1], objective evaluations are carried out on the proposed method with improved HNM modelling in this work. The evaluation consists of a frequency weighted segmental SNR measure [5], a standardized PESQ measure [6], and a study of speech spectrograms. In all the experiments, speech signal is sampled at 8 kHz, FFT with length of 256 is used for analysis.

Overlap-save sectioning procedure is adopted and the percentage of overlapping is 75%. Three types of noisy files are used for objective measures: speech corrupted by car noise and street noise at SNR level ranges from 0dB to 15dB, with a step size of 5dB, are taken from NOIZEUS database [6], which contains standard IEEE sentences corrupted by real-world noise from AURORA database. Noisy speech of the third type is manually corrupted by white Gaussian noise also at same SNR levels, using ITU-T P.56 standard [8]. Evaluation results are averaged out using 10 utterances from the aforementioned NOIZEUS database. Half of the utterances are from male speakers and half are from female speakers.

Based on the results reported in [2] and the results of our previous experiment [1], the MMSE-SPU method [3] and the log-spectral minimum mean square error (LOGMMSE) method [9] generally give better subjective results. In this work, we compare the proposed method, which we label as HNM method, with the original MMSE-SPU method [3], the LOGMMSE [9] method and the corresponding noisy speech without enhancement (WE). Figures 4 and 5 show the results of PESQ measure and the frequency weighted segmental SNR measure, respectively. It is observed that the proposed noise suppression method obtains an average of around 0.2 point improvement in PESQ score (1.0 is worst and 4.5 is best) and an average of 1dB gain in frequency weighted segmental SNR measure over conventional methods in low SNR environments. For PESQ measure, the proposed method achieves substantial gain over conventional methods for all three types of noise at low SNR levels, particularly for colored noise (street noise and car noise). It suffers little degradation at relatively high SNR levels, comparing to conventional methods. For frequency weighted segmental SNR measure, the proposed method also achieves obvious improvement over conventional method for all three types of noise, at low SNR levels. However, obvious degradation is observed at relatively high SNR levels for colored noise. The major reason for performance gain in low SNR environments is that the proposed HNM-based method is able to eliminate residual noises, which can not be effectively suppressed by conventional methods. On the other hand, it is able to compensate the over-suppression caused by conventional methods at low SNR levels. Nevertheless, in relatively high SNR environments, the objective measures are more sensitive to the variability between the clean and re-synthesized speech signals. Consequently, the penalty on variability cancels out the potential gain obtained by better noise suppression ability. Practically a toggle may be placed so that HNM processing is enabled when estimated noise is aggressive. Otherwise, pre-cleaning is sufficient. Figure 6 demonstrates the effectiveness of the proposed method in retrieving damaged speech structure. It is observed from the spectrograms that substantial improvement has been made to restore the harmonic structure using the proposed method. At the same time, residual noises such as musical tones are greatly mitigated.

5. CONCLUSION

In this work, a complete analysis-synthesis framework for speech enhancement in different noise conditions is introduced. Improved HNM modelling, which includes estimation techniques for better tracking the pitch contours, post-processing technique for restoring missing harmonics are proposed. Choice of different HNM parameters is discussed. Simulation results, in terms of frequency weighted segmental SNR and PESQ score, have shown that the improved HNM-based speech enhancement system achieves considerable improvement over conventional methods in low SNR environments.

REFERENCES

[1] R. F. Chen, C. F. Chan, H. C. So, J. S. C. Lee, and C. Y. Leung, "Speech enhancement in car noise environment based on an analysis-synthesis approach using harmonic noise model," *in Proc. ICASSP-2009*, pp.4413-4416, Taipei, Taiwan, Apr. 2009.

[2] Y. Hu and P. C. Loizou, "Subjective comparison of speech enhancement algorithms," *in Proc. ICASSP-2006*, pp.I153-I156, Toulouse, France, May 2006.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, pp. 1109-1121, Dec. 1984.

[4] C.-F. Chan and E.W.M. Yu, "Improving pitch estimation for efficient multiband excitation coding of speech," *IEE Electronics Letters*, vol. 32, no. 10, pp. 870-872, May 1996.

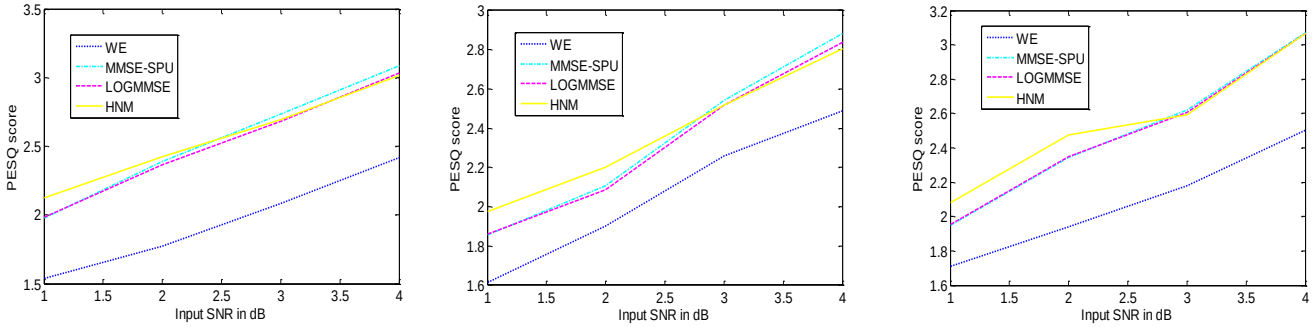
[5] J. Tribolet, P. Noll, B. McDermott and R. Crochiere, "A study of complexity and quality of speech waveform coders," *in Proc. ICASSP-1978*, pp.586-590, Tulsa, USA, Apr. 1978.

[6] ITU-T P.862, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *ITU-T Recommendation*, 2001.

[7] P. C. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, Boca Raton, USA, 2007.

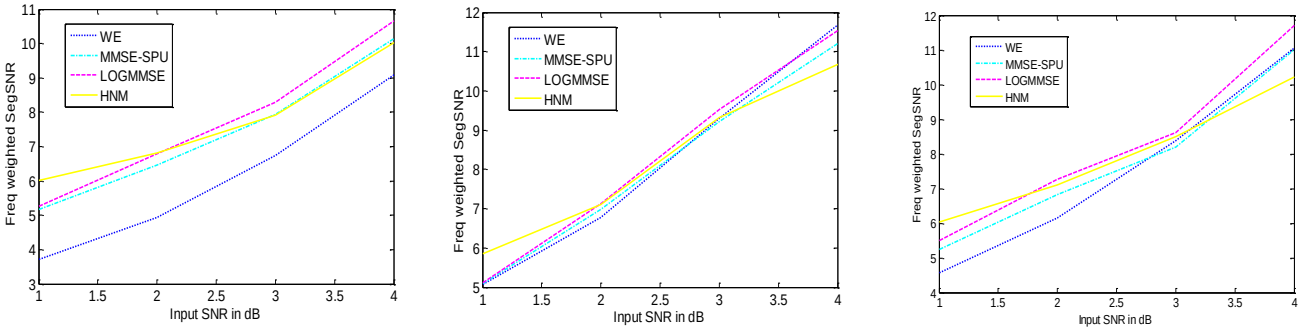
[8] ITU-T P. 56, "Objective measurement of active speech level," *ITU-T Recommendation,* 1993.

[9] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-23(2), pp. 443-445, 1985.



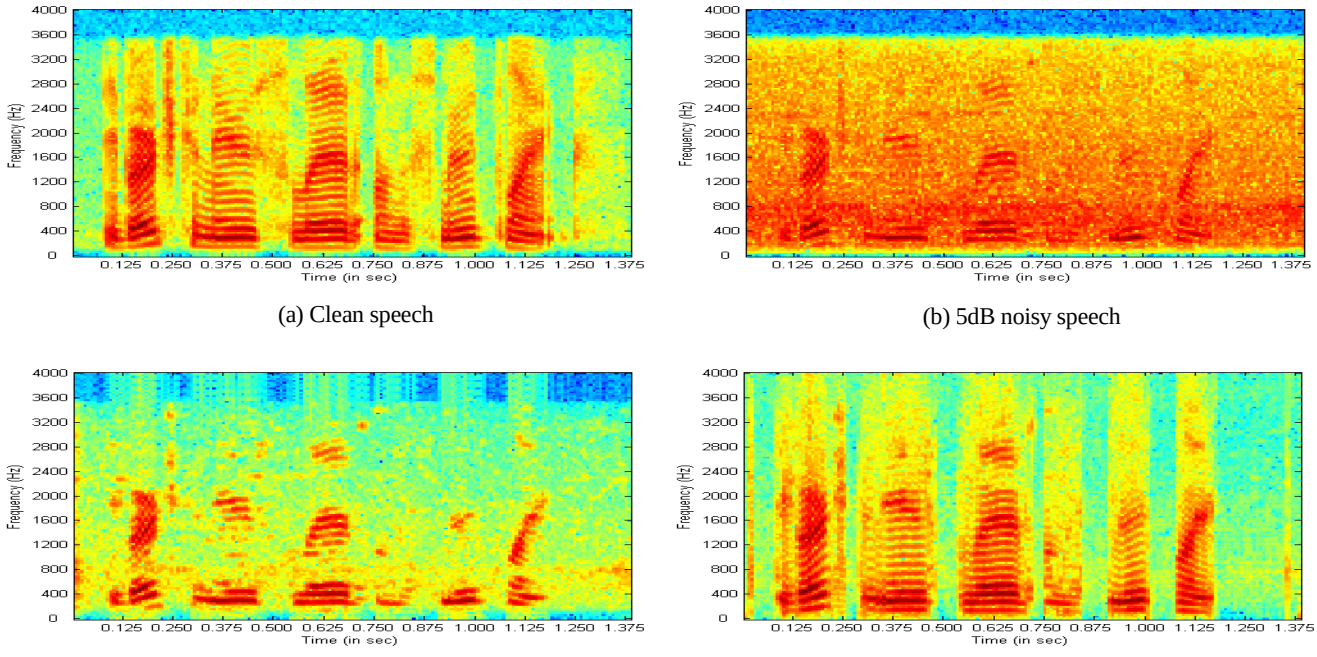
(a) white noise (b) street noise (c) car noise

Figure 4 - Comparison of PESQ scores in different noise and SNR settings



(a) white noise (b) street noise (c) car noise

Figure 5 - Comparison of frequency weighted segmental SNR in different noise and SNR settings



(a) Clean speech (b) 5dB noisy speech

(c) Pre-cleaned speech (d) HNM enhanced speech

Figure 6 - Comparison of spectrograms of a male speech segment at input SNR level of 5dB