SPEECH ENHANCEMENT IN CAR NOISE ENVIRONMENT BASED ON AN ANALYSIS-SYNTHESIS APPROACH USING HARMONIC NOISE MODEL

R. F. Chen [1,2], C. F. Chan [1], H. C. So [1], Jonathan S. C. Lee [2], C. Y. Leung [2]

[1] Department of Electronic Engineering, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong
[2] Avantwave Limited, 3/F Photonics Centre, No.2 Science Park East Avenue, HK Science Park, Shatin, N.T., Hong Kong

ABSTRACT

This paper presents a speech enhancement method based on an analysis-synthesis framework using harmonic noise model (HNM) in car noise environment. The major advantages of this method are effective suppression of car noise even in very low signal-to-noise ratio environments and mitigation of "musical tones" which are generally introduced by conventional methods. In this paper, we devise a complete analysis-synthesis based speech enhancement system, and give details in HNM modeling, parameter estimation, and car noise adaptation. Subjective evaluation results show that the proposed method exhibits better noise suppression ability over conventional approaches without obvious degradation of speech quality.

Index Terms— harmonic noise model, car noise, speech enhancement, speech synthesis.

1. INTRODUCTION

Conventional speech enhancement methods for typical communication systems consist of four classes of algorithms: spectral subtraction, subspace, statistical-model based and Wiener algorithms [1]. Experiments show that these methods only work well in relatively high signal-to-noise ratio (SNR) (e.g. ≥ 10dB) environment with relatively stationary background noise. However, the ambient noise in a typical real car driving environment is highly non-stationary and it usually exhibits very strong low frequency components [2]. In this case conventional methods undergo two common deficiencies. Firstly, the suppression rule is generally based on some form of segmental SNR, which is not so reliable in a highly non-Gaussian and non-stationary noise environment. As a result, the rigid suppression rule may exacerbate the processing artifacts, namely "musical tones". Secondly, those strong low frequency components typically in car noise can not be effectively eliminated using SNR-based suppression rule, even incorporating signal presence uncertainty [3].

To overcome this problem, we propose a new method based on an analysis-synthesis framework using harmonic noise model (HNM). In this approach, the speech corrupted by car noise is pre-cleaned by minimum mean-square error under signal presence uncertainty (MMSE-SPU) algorithm [3]. The output signal is then passed through the analysis-synthesis HNM framework to generate artificial signal. By taking into account the harmonic characteristic of speech, those strong low frequency components can be effectively discriminated and eliminated from voiced speech. Since the processed speech signal is artificially synthesized rather than modified by rigid suppression rule, processing artifacts "musical tones" can be completely avoided. Experimental results show obvious improvement by this new method in terms of subjective measures in car noise environment.

This paper is organized as follows. Section 2 will present the overall HNM framework of our method. Section 3 will provide speech analysis and synthesis details using the HNM framework. Section 4 will show experimental results using our method in comparison with conventional methods. Section 5 will be devoted to conclusion and future work.

2. HNM FOR SPEECH ENHANCEMENT

HNM assumes the speech signal to be composed of a deterministic/voiced part and of a stochastic/unvoiced part. The voiced part is assumed to contain only harmonically related sinusoids while the unvoiced part can be modeled using random signals [4]. HNM is extensively used in speech coding for its flexible and effective decomposition of speech and its compact parameter requirement. In speech coding, speech signal can be re-synthesized at the decoding stage with only the pitch, residual energy, and spectral envelope information. To apply HNM for speech enhancement, the accurately estimated pitch is a crucial parameter. On the other hand, original spectral information can be adopted to replace spectral envelope information for more accurate synthesis. Residual energy can be derived from linear prediction coefficient (LPC) of speech for unvoiced part synthesis.

At the first stage, noisy speech is passed through MMSE-SPU algorithm [3] for pre-cleaning. The output signal is directed for pitch detection. On the other hand, spectrum of pre-cleaned signal is calculated mainly for a spectral flatness measure in voiced/unvoiced (V/UV) mixing function [5]. The matching error of pitch detection

process is used together with a statistical voice activity detector (VAD) [6] obtained in pre-cleaning stage to classify the V/UV frame. The detected and post-processed pitch is used to form an excitation spectrum using Griffin's method [7]. This excitation spectrum is matched with pre-cleaned spectrum to estimate magnitude and phase of voiced speech. For unvoiced speech, LPC spectrum and V/UV mixing function are multiplied in frequency domain and transformed to time domain to form residual energy gain, which is multiplied with randomly generated Gaussian noise to synthesize unvoiced speech. Voiced and unvoiced speech is summed over in time domain to form enhanced speech. With the help of accurately estimated pitch and V/UV classification, synthesized speech can be very close to the original one and at the same time noise can be suppressed to a great extent.

3. SPEECH ANALYSIS AND SYNTHESIS

3.1. Speech Analysis

3.1.1 Pitch Detection and Post-Processing

In multiband excitation pitch analysis [7], the input speech spectrum $S(k)$ is matched to the synthetic harmonic spectrum by minimizing an error function $\alpha(\tau)$ with respect to the searching variable τ for the pitch period.

$$\alpha(\tau) = \frac{\sum_{m=1}^{M(\tau)} \sum_{k=a_m(\tau)}^{b_m(\tau)} [|S(k)| - A_m(\tau)|E(\tau,k)|]^2}{(1-\tau B)\sum_{m=1}^{M(\tau)} \sum_{k=a_m(\tau)}^{b_m(\tau)} |S(k)|^2} \quad (1)$$

where B is weighting factor for biasing the pitch dependent error, $A_m(\tau)$ is the harmonic magnitude, $S(k)$ is the input spectrum, and $E(\tau,k)$ is the synthetic harmonic spectrum. The total number of bands in the speech spectrum is $M(\tau)$ and the lower and upper boundaries of the *m*-th harmonic band are denoted as $a_m(\tau)$ and $b_m(\tau)$, respectively. However, this matching process may not be effective when directly applied in car noise environment. This is mainly because the low frequency components in car noise spectrum may also exhibit harmonic-like styles with large magnitudes, which will severely deteriorate the matching process. In this approach, MMSE-SPU pre-cleaned spectrum is used instead of original noisy spectrum for pitch detection, which will significantly improve the matching process. Nevertheless, gross pitch error still occasionally occurs since there are residual low frequency components and "musical tones" in the pre-cleaned spectrum. An improved pitch estimation error measure [8] $\epsilon(\tau) = \alpha(\tau) + \beta(\tau)$ is used to reduce the gross pitch error, where

$$\beta(\tau) = \frac{1}{M(\tau)(1-\tau B)} \sum_{m=1}^{M(\tau)} \left[\frac{\sum_{k=a_m(\tau)}^{b_m(\tau)} [|S(k)| - A_m(\tau)|E(\tau,k)|]^2}{\sum_{k=a_m(\tau)}^{b_m(\tau)} |S(k)|^2} \right] \quad (2)$$

To further improve the pitch estimation without frame delay, the searching index of pitch period is limited to a smaller range for each frame according to an indicating parameter to avoid fluctuations. The indicating parameter is changed only if a maximum drift of pitch period is reached. On the other hand, few number of magnitude bands can be exempted from matching process according to the specific car noise environment to minimize matching error.

3.1.2 V/UV Frame Classification

A statistical VAD [6] is incorporated with the matching error $\epsilon(\tau)$ in HNM pitch detection to form the V/UV frame classification. The statistical VAD is derived in the pre-cleaning stage using *a posteriori* SNR γ_k and *a priori* SNR ξ_k as follows:

$$\frac{1}{N} \sum_{k=1}^{N-1} \log \Lambda_k \underset{<H_1}{>H_0} \delta \quad (3)$$

where

$$\Lambda_k = \frac{1}{1+\xi_k} \exp\{\frac{\gamma_k \xi_k}{1+\xi_k}\} \quad (4)$$

where N is the size of the FFT, H_1 denotes the hypothesis of speech presence, H_0 denotes the hypothesis of speech absence, and δ is a fine-tuned threshold for different levels of car noise. Basically, the statistical VAD already gives satisfactory results. However, performance will be degraded when it is used to distinguish the situation of only soft speech presence and the situation of only strong engine noise presence. In this case, normalized matching error $\epsilon(\tau)$ is adopted to further classify them by exploiting harmonic structure.

3.1.3 V/UV Mixing Function

Within an individual frame, a V/UV mixing function is applied to improve the quality of synthesized speech. HNM creates very "clean" harmonics in the voiced region and hence introduces buzziness into the synthetic speech. By applying a V/UV mixing function [5] for both voiced and unvoiced part synthesis, the synthetic speech can achieve obvious improvement in terms of quality. It has been shown that the speech spectrum envelope inherently correlates to the degree of V/UV mixing [9]. Based on the characteristics of speech, high spectral flatness region can be classified as unvoiced and low spectral flatness region can be classified as voiced. A spectral flatness measure which covers the region from θ to π in speech spectrum $S(\omega)$ is defined as

$$f(\theta) = \frac{1}{\pi - \theta} \int_{\theta}^{\pi} [\log |S(\omega)| - m(\theta)]^2 d\omega \qquad (5)$$

where

$$m(\theta) = \frac{1}{\pi - \theta} \int_{\theta}^{\pi} \log |S(\omega)| d\omega \qquad (6)$$

By comparing $f(\theta)$ to a predefined threshold T_{uv}, a smooth V/UV mixing function is defined from the spectral flatness measure as:

$$v(\theta) = \begin{cases} 1 - \dfrac{f(\theta)}{2T_{uv}}, & f(\theta) < T_{uv} \\ \dfrac{T_{uv}}{2f(\theta)}, & f(\theta) > T_{uv} \end{cases} \qquad (7)$$

Note that the V/UV transition is at $v(\theta) = 0.5$, and the voiced and unvoiced regions are corresponding to $v(\theta) < 0.5$ and $v(\theta) > 0.5$, respectively.

3.2. Speech Synthesis

3.2.1 Voiced Speech Synthesis

A time domain approach is selected for voiced speech synthesis due to its advantage of allowing a continuous variation in fundamental frequency frame to frame. It can be synthesized as the sum of the outputs of a band of sinusoidal oscillators running at the harmonics of the fundamental frequency [7]. The voiced speech $V(t)$ is defined as:

$$V(t) = \sum_{m} A_m(t) \cos(\theta_m(t)) \qquad (8)$$

where $A_m(t)$ and $\theta_m(t)$ are the amplitude and phase of *m*-th harmonic band, respectively. The $A_m(t)$ can be obtained by setting $\partial\alpha(\tau)/\partial\tau = 0$ in (1), resulting

$$A_m(\tau) = \frac{\sum_{k=a_m(\tau)}^{b_m(\tau)} |S(k)||E(\tau,k)|}{\sum_{k=a_m(\tau)}^{b_m(\tau)} |E(\tau,k)|^2} \qquad (9)$$

The $\theta_m(t)$ can be derived in a similar manner using complex matching. 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.

3.2.2 Unvoiced Speech Synthesis

A frequency domain approach is selected for unvoiced speech synthesis for efficiency. The LPC envelope spectrum is weighted using the V/UV mixing function described above. The weighted power spectrum is converted to autocorrelation data and then an all-pole LPC model is fitted to the autocorrelation data to compute the synthesis filter's residual signal gain [7]. Random Gaussian noise is generated and fitted into the synthesis filter to produce the unvoiced speech signal.

4. EXPERIMENTAL RESULTS

To evaluate the performance of the proposed method over conventional methods, an informal listening test is designed to follow the testing procedure in [1]. NOIZEUS noisy speech corpus and algorithms for conventional methods are extracted from accompanying material in [6]. Those conventional methods with better performance in [1] are chosen to compare with our method. 10 noisy speech sentences (5 by male and 5 by female) are randomly extracted from NOIZEUS car noise class at a SNR of 5dB. 20 listeners are instructed to successively attend to and rate the enhanced speech on signal distortion (SIG)-[1=very unnatural, 5=very natural], background intrusiveness (BAK)-[1=very conspicuous, very intrusiveness, 5=not noticeable], and overall effect using the scale of mean opinion score (OVRL)-[1=bad, 5=excellent] as in [1],[12]. Fig.1 shows the results of listening test in terms of the three criteria. Algorithms in Fig.1 are named in accordance with those in [1], while HNM denotes our method.

It can be observed from Fig.1 that the SIG score of our method is slightly lower than those of some conventional methods. Since the speech processed by our method is artifically synthesized, signal distortion is unavoidable. The degradation is mainly due to the inaccurate estimation of acoustic cues (e.g. pitch, spectral envolope). On the other hand, the gain is obvious as outstanding BAK score indicates that our method surpass those convetional methods in terms of noise suppression ability. That is because with an analysis-synthesis structure, "music tone" problem is completely avoided. The OVRL score confirms that the proposed method attains an overall performance gain in terms of mean opinion score (MOS).

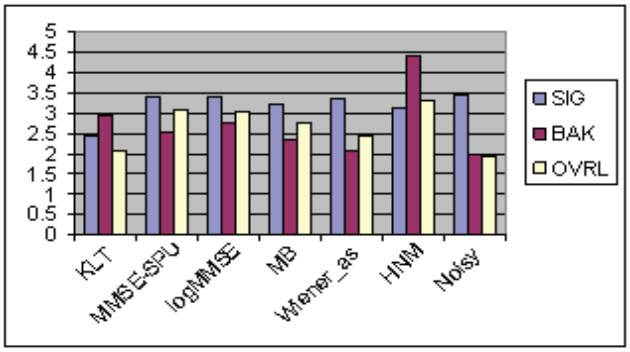
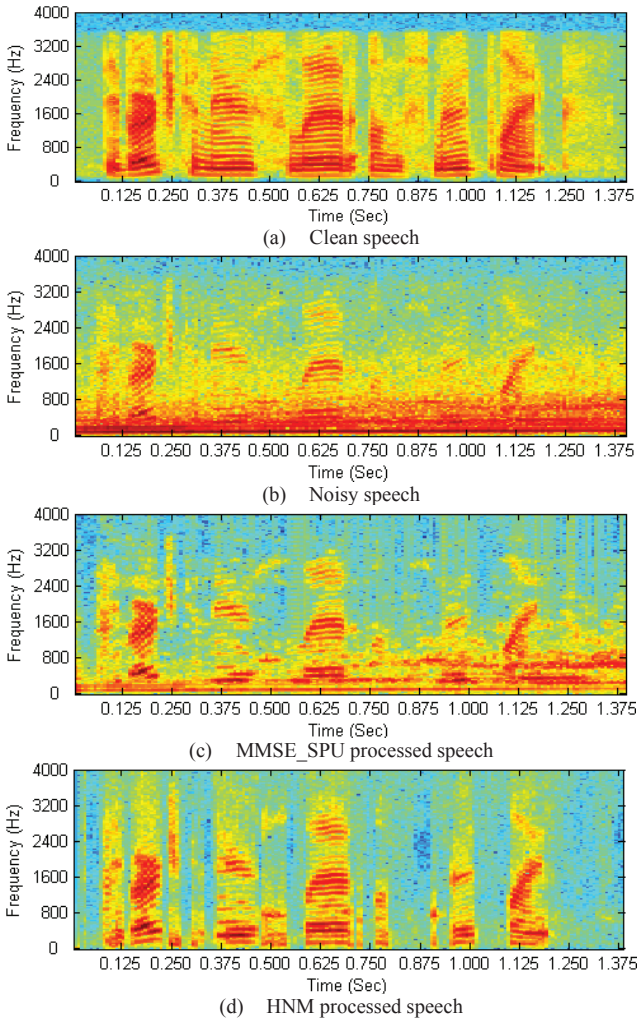


Fig.1. Listening test scores for different algorithms in 5dB car noise noisy speech from NOIZEUS.

To simulate an adverse car driving environment, real car noise is also recorded in an Alfa Romeo 155 series during acceleration from 60km/h to 100km/h on a highway with opened car windows, using an 8 kHz sampling. Clean speech sentences are extracted from IEEE corpus [10]. The recorded noise is mixed with clean speech at a SNR -10dB using ITU-T P.56 standard [11].

Fig.2(a) shows the clean speech spectrogram of a male speaking "The birch canoe slid on the smooth planks." While Fig.2(b) shows the real car noise corrupted speech spectrogram. It can be observed from Fig.2(b) that the car noise exhibits very strong components in low frequency region and it is highly non-stationary during an acceleration. Fig.2(c) shows the pre-cleaned spectrogram using MMSE_SPU algorithm. Note that "musical tones" (smeared spots) and car noise still reside on the spectrogram, which is similar as using other conventional methods. Fig.2(d) shows the spectrogram processed by our method. Most of "musical tones" and car noise are suppressed while clean harmonic structure is restored.



(a) Clean speech

(b) Noisy speech

(c) MMSE_SPU processed speech

(d) HNM processed speech

Fig.2. Spectrograms of clean, noisy and processed speech

5. CONCLUSION

In conventional speech enhancement methods, a suppression gain is generally computed based on some form of segmental SNR. These methods often suffer poor performance in low SNR and/or non-stationary noisy environment. In our approach, a complete analysis-synthesis based system exploiting HNM is used to replace traditional filter-based algorithms. We believe that this method is more promising as a post-processing tool to deal with very noisy and highly non-stationary environment. Pitch estimation and V/UV frame classification are crucial in this HNM-based method. Experiments show that with correctly estimated pitch and V/UV classification, very clean and high-quality speech can be synthesized even using amplitudes extracted from noisy spectrum. Future work may include more robust pitch estimation and V/UV frame classification in very noisy environment.

6. ACKNOWLEDGEMENT

The authors would like to acknowledge the grant from the Government of Hong Kong Special Administrative Region-Innovation and Technology Fund (ITF), University-Industry Collaboration Program (Project No. UIT/096, and CityU ref. 9440060).

7. REFERENCES

[1] Y. Hu and P.C. Loizou, "Subjective comparison of speech enhancement algorithms," *in Proc. ICASSP-2006*, vol.1, pp.153-156, Toulouse, France, May 2006.

[2] E. Hansler and G. Schmidt, *Topics in Acoustic Echo and Noise Control*, Springer-Verlag, Berlin Heidelberg, 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 Processing*, vol. ASSP-32, pp. 1109-1121, Dec. 1984.

[4] J. Laroche, Y. Stylianou and E. Moulines. "HNM: A simple, efficient harmonic+noise model for speech," *in Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics,* pp.169-172 New Paltz, NY, USA, Oct. 1993.

[5] E.W.M. Yu and C.-F. Chan, "Harmonic+noise coding using improved V/UV mixing and efficient spectral quantization," *in Proc. ICASSP-1999*, vol.1, pp. 477-480, Phoenix, USA, March 1999.

[6] P.C. Loizou, *Speech Enhancement: Theory and Practice,* Taylor&Francis Group, Boca Raton, 2007.

[7] D.W. Griffin and J. S. Lim, "Multi-band excitation vocoder," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-36, pp. 1223-1235, Aug. 1988.

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

[9] C.-F. Chan, "High-quality synthesis of LPC speech using multi-band excitation model," *in Proc. EUROSPEECH-93*, pp 535-538, Berlin, Germany, Sep. 1993.

[10] IEEE Subcommittee, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio and Electroacoustics*, pp. 225-246, 1969.

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

[12] ITU-T P. 835, "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm," *ITU-T Recommendation,* 2003.