NEW SPEECH HARMONIC STRUCTURE MEASURE AND IT APPLICATION TO POST SPEECH ENHANCEMENT

An-Tze Yu and Hsiao-chuan Wang

Department of Electrical Engineering
National Tsing Hua University, Hsinchu, Taiwan, ROC
yuat@ms25.hinet.net， hcwang@ee.nthu.edu.tw

ABSTRACT

This paper proposes a set of hierarchical harmonicities to analyze the harmonic structure of speech signal. The high redundant harmonic structure in voiced speech makes the perception of speech more robust in noisy environments. It is possible to recover a distorted harmonic structure. The systematic information of harmonic structure is represented by a set of harmonicities, including grid, temporal, spectral and segmental harmonicities. By using this harmonic measure, the speech quality after performing the enhancement can be evaluated. It gives an indicator of whether the further enhancement is necessary. A post speech enhancement based on the reconstruction of harmonic structure is proposed to enhance the quality of voiced speech. The experiment shows the effectiveness of the proposed method.

1. INTRODUCTION

Many speech enhancement algorithms have been proposed. However, few of them took the factor of speech production into account. The knowledge of speech production and perception needs to be applied into the enhancement system to further improve the speech quality. The voiced speech possesses high redundant harmonic structure. This high redundancy makes the perception of speech more robust in noisy environments and makes it possible to partly or entirely recover the distorted harmonic structure [1]. Thus the harmonic structure can be used for enhancing the speech quality.

When additive noise emerges, the harmonic structure of a noisy speech is partially distorted. The type and the level of corrupting noise determine the degree of distortion. Conventional harmonicity, which measures the integrity of harmonic structure, is often used to determine voicing degree or noise level [2]. To provide systematic and detailed information for harmonic structure analysis, a set of hierarchical harmonicities, including grid, temporal, spectral and segmental harmonicities, are proposed.

By analyzing the harmonic structure of enhanced speech, the speech quality after performing the enhancement can be evaluated. It gives an indicator of whether the further enhancement is necessary. A post speech enhancement based on the reconstruction of harmonic structure is proposed to enhance the quality of voiced speech. Temporal and spectral harmonicities provide detailed information for the reconstruction. The energy re-assignment of harmonics and spectral harmonicity shaping are the crucial elements for the reconstruction. The comb filter with adjustable gain and bandwidth adapted by the factors derived from temporal and spectral harmonicities is a tool for this purpose.

Through harmonicity evaluation and careful harmonic structure reconstruction, the proposed method achieves higher speech quality than conventional systems. The experiment shows the effectiveness of the proposed method.

2. NEW HARMONIC STRUCTURE MEASURES

To provide systematic and detailed information for harmonic structure analysis, a set of hierarchical harmonicities are proposed.

2.1. Grid harmonicity

The grid harmonicity measures the energy ratio between a harmonics and its surrounding noise. It involves three factors: (1) the local spectral dominance, (2) the temporal correlation, and (3) the harmonic spectral correlation. The local spectral dominance for *m*-th harmonics of a signal evaluated at frame *n* is defined as below,

$$h_D(n,m) = \frac{\sum_{k=(m-\eta)k_0}^{(m+\eta)k_0} S(n,k)\phi(k-mk_0)}{\sum_{k=(m-\eta)k_0}^{(m+\eta)k_0} S(n,k)\big(1-\phi(k-mk_0)\big)}, \qquad (1)$$

where $S(n,k)$ is the magnitude spectrum of an analyzed signal. k represents the frequency bin index,. k_0 is the frequency bin index of fundamental frequency. η is set to 0.5. $\phi(k)$ is a harmonics selector defined as

$$\phi(k) = e^{-\frac{k^2}{2\sigma^2}} , \tag{2}$$

where σ controls the width of harmonics selector.

The temporal variation of harmonics is measured by,

$$\begin{aligned} D_T(n,m) = & \frac{|S(n,mk_0) - S(n-1,mk_0)|}{S(n,mk_0) + S(n-1,mk_0)} \\ & + \frac{|S(n+1,mk_0) - S(n,mk_0)|}{S(n+1,mk_0) + S(n,mk_0)} \end{aligned} . \tag{3}$$

The temporal correlation is obtained from,

$$h_{tc}(n,m) = \frac{1}{1 + e^{a(D_T(n,m) - b)}} , \tag{4}$$

where a and b control the property of transformation and are experimentally set to 5 and 0.3, respectively. They are insensitive to testing data.

Harmonic correlation is measured by referencing its neighborhood local spectral dominances,

$$h_{hc}(n,m) = h_D(n,m-1) + h_D(n,m+1) . \tag{5}$$

Then the grid harmonicity is expressed as

$$h(n,m) = h_D(n,m) h_{Tc}(n,m) h_{hc}(n,m) , \tag{6}$$

The spectrogram and grid harmonicities for an utterance are displayed in Figure 1.

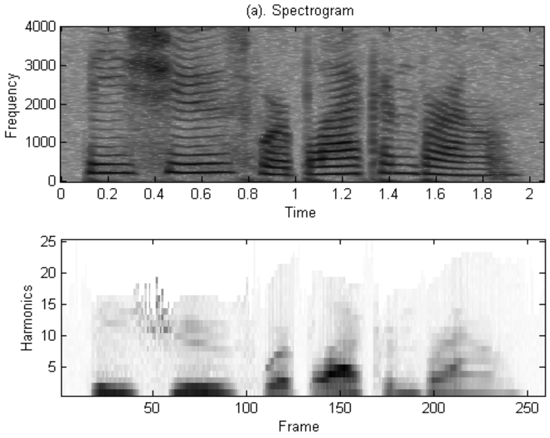

Figure 1. (a) Spectrogram (b) Grid harmonicities

2.2. Temporal harmonicity

Temporal harmonicity represents the variation of speech harmonicity with time. It is derived by summing the grid harmonicities over all harmonics,

$$H_T(n) = \sum_{m=1}^{M(n)} h(n,m) w_h(n,m) , \tag{7}$$

where $M(n)$ is the number of harmonics at frame n. $w_h(n,m)$ is a weighting factor and expressed as,

$$w_h(n,m) = \frac{e(n,m)}{E_T(n)} , \tag{10}$$

where $e(n,m)$ and $E_T(n)$ are grid and temporal energies, respectively. They are computed by

$$e(n,m) = \sum_{k=(m-\eta)k_0}^{(m+\eta)k_0} S(n,k)\phi(k - mk_0) , \tag{8}$$

and

$$E_T(n) = \sum_{m=1}^{M(n)} e(n,m) , \tag{9}$$

Figure 2 illustrates a speech waveform and its temporal harmonicity. High harmonicities parts correspond to voided speech regions.

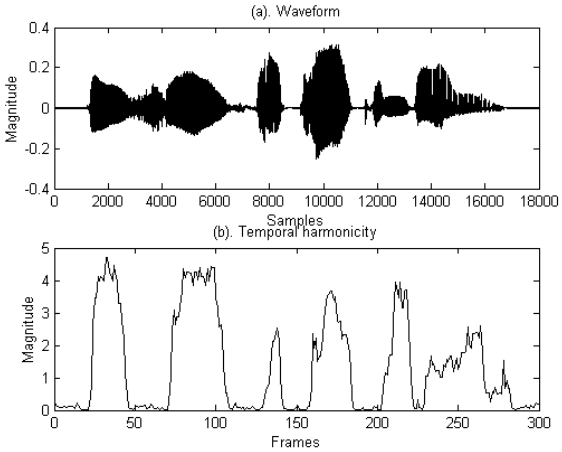

Figure 2. (a)Waveform (b)Temporal harmonicities

2.3. Spectral harmonicity

The spectral harmonicity evaluates the integrity of harmonics in a speech segment. It is computed by summing the grid harmonicities over the specified segment,

$$H_H(s,m) = \sum_{n=SegBegin(s)}^{SegEnd(s)} h(n,m) w_t(n,m) , \tag{11}$$

where SegBegin(s) and SegEnd(s) represent the first and last frame indexes belonging to segment s, respectively. $w_t(n,m)$ is a weighting factor defined as

$$w_t(n,m) = \frac{e(n,m)}{E_H(s,m)} , \tag{12}$$

where $E_H(s,m)$ is computed by.

$$E_H(s,m) = \sum_{n=SegBenin(s)}^{SegEnd(s)} e(n,m) , \tag{13}$$

Figure 3 shows the spectral harmonicities for the first three phonemes of an utterance pictured at Figure 1 and Figure 2.

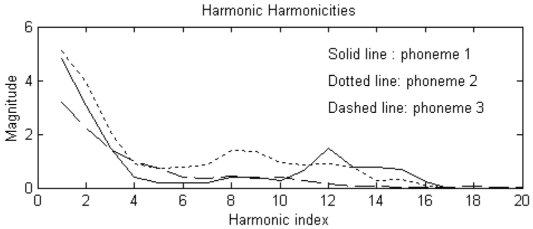

Figure 3. Spectral harmonicities

2.4. Segmental (phoneme) harmonicity

It is a complete evaluation for harmonic structure within a segment. The segmental or phoneme harmonicity is proposed and calculated as follows

$$\begin{aligned} H(s) &= \sum_{n=SegBegin(s)}^{SegEnd(s)} H_T(n) W_T(s,n) \\ &= \sum_{m=1}^{M(s)} H_H(s,m) W_H(s,m) \end{aligned}, \tag{14}$$

where $M(s)$ is the number of harmonics at segment *s*. $W_T(s,n)$ and $W_H(s,m)$ are the weighting factors for temporal and spectral harmonicities, respectively. They are computed by

$$W_T(s,n) = \frac{E_T(n)}{E_S(s)}, \tag{15}$$

and

$$W_H(s,m) = \frac{E_H(s,m)}{E_S(s)}, \tag{16}$$

where $E_S(s)$ is the segmental energy defined as ,

$$E_S(s) = \sum_{n-SegBegin(s)}^{SegEnd(s)} E_T(n) = \sum_{m=1}^{M(s)} E_H(s,m), \tag{17}$$

The segmental harmonicities for six phonemes in an utterance displayed in Figure 1 are shown in Figure 4.

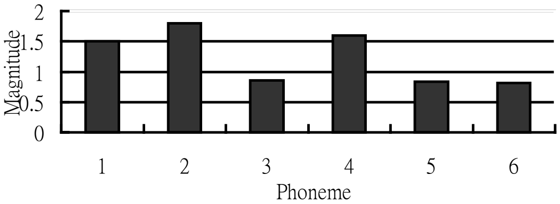

Figure 4. Segmental Harmonicity

3. POST SPEECH ENHANCEMENT

General objective performance measures, like SNR or spectral distortion, need a reference signal for quality evaluation. However, the reference signal is generally unavailable for most real testing environments. This results in the situation that conventional enhancement processing cannot guarantee the quality of enhanced speech. By analyzing the harmonic structure of the enhanced speech, the speech quality after enhancement can be evaluated. The result gives an indicator of whether the further enhancement is necessary. A post speech enhancement based on the reconstruction of harmonic structure is proposed to enhance the quality of voiced speech.

Robust fundamental frequency estimation is the basic requirement for the post enhancement [3]. Temporal and spectral harmonicities provide detailed information to guide the reconstruction of harmonic structure. The energy re-assignment of harmonics and spectral harmonicity shaping are the crucial elements for the reconstruction. The comb filter adapted by the factors derived from temporal and spectral harmonicities is a tool for this purpose. The comb filtering is implemented in frequency domain,

$$\Phi(n,k) = \sum_{m=1}^{M(n)} g(n,m)\phi(k - mk_0), \tag{18}$$

where $g(n,m)$ is the gain function defined as

$$g(n,m) = \frac{\hat{e}(n,m)}{e(n,m)}, \tag{19}$$

where $\hat{e}(n,m)$ is the adjusted grid energy after energy re-assignment and derived from

$$\hat{e}(n,m) = E_H(s,m)\frac{E_T(n)}{E_S(s)}, \tag{20}$$

$\phi(k,m)$ is the kernel function of comb filters defined as

$$\phi(k,m) = e^{-\frac{k^2}{2\sigma^2(m)}}, \tag{21}$$

where $\sigma(m)$ controls the width of comb filter and is set inverse proportional to a reference spectral harmonicity, which is measured from available speech database. Figure 5 plots a comb filter with adjustable gain and bandwidth.

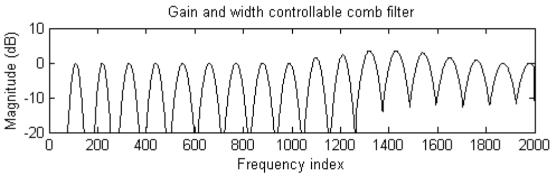

Figure 5. Spectrum of a gain and width controllable comb filter

4. PERFORMANCE EVALUATION

The evaluation consists of a subjective study of speech spectrograms and informal listening test. The performance results are averaged out using six different utterances corrupted with two types of noises, i.e., the white noise and the car noise. The speech signals, sampled at 8 KHz, are degraded by the two noise types with segmental SNRs -5 and 5 dB. The speech frame contains 256 samples and with 64 samples frame update step. A 2048 points FFT with zero padding transforms the speech data into frequency domain. A spectral subtraction algorithm is applied as the front-end enhancement system. The temporal harmonicity detects the speech pause and the harmonic tunneling estimates the noise spectrum during speech-activated regions [4].

The mean segmental harmonicities of noisy speech, spectral subtraction enhanced speech and post enhanced speech are shown at Table 1. Table 2 displays the segmental SNR improvements of enhancing algorithms. The experimental results show that the proposed post enhancement can significantly improve the performances. Figure 6 and Figure 7 respectively show the waveform and spectrogram of clean speech, noisy speech, and post enhanced speech. Figure 7 clearly shows that the distorted speech harmonic structures are rebuilt. The informal listening test reveals that the harmonics reconstruction has greatly reduced the unpleasant perception of noise.

Table 1. Mean segmental harmonicities of enhanced speeches

SNR	White			Car		
	Noisy	Enhanced	Post En	Noisy	Enhanced	Post En.
5dB	0.75	0.84	1.06	0.80	0.88	1.09
-5dB	0.27	0.37	0.84	0.41	0.55	0.99

Table 2. Segmental SNR improvement of enhancing algorithms

SNR	White		Car	
	Enhanced	Post En	Enhanced	Post En.
5dB	3.95	4.36	6.18	7.53
-5dB	5.72	7.09	7.22	8.69

5. CONCLUSSION

This paper proposes a harmonic structure measure, including grid, temporal, spectral and segmental harmonicities, to provide systematic and detail information for harmonic structure analysis. With these measures, a post speech enhancement system is implemented. The experiment shows that the proposed system can achieve higher speech quality than the conventional systems.

ACKNOWLEDGEMENT

This research was partially sponsored by the National Science Council, Taiwan, under contract number NSC-92-2213-E-007-036.

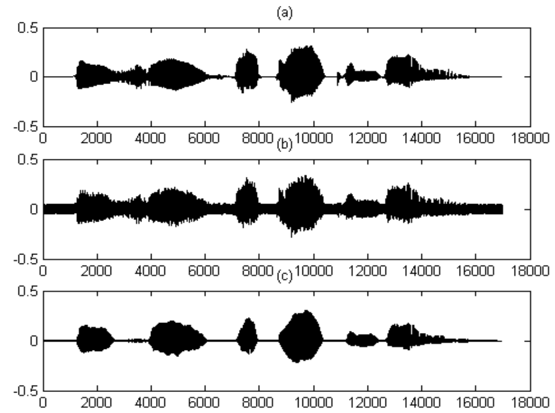

Figure 6. Waveforms of (a)Clean speech (b)Noisy speech (c)Enhanced speech (post enhanced).

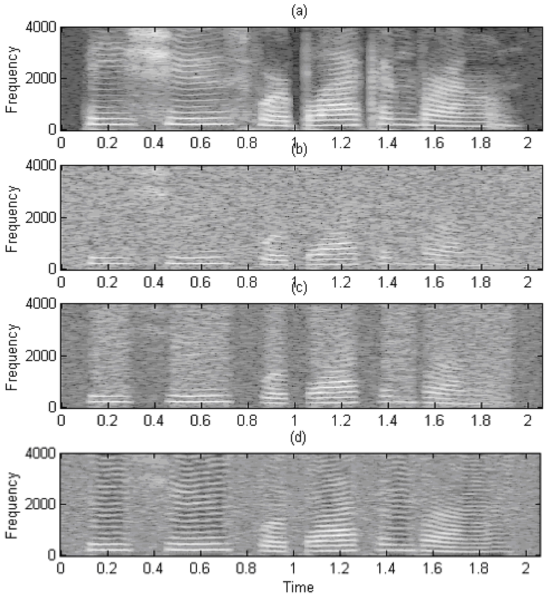

Figure 7. Spectrograms of (a)Clean speech (b)Noisy Speech (c)Enhance speech (d)Enhanced speech (post enhanced)

6. REFERENCES

[1] Jesper Jensen and John H. L. Hansen," Speech enhancement Using a Constrained Iterative Sinusoidal Model ", *IEEE Trans. ON SPEECH AND AUDIO PROCESSING*, Vol. 9, NO. 7, pp. 731-740, 2001

[2] Dhany Arifianto, Takao Kobayashi, "IFAS-based Voiced/Unvoiced classification of speech signal," *Proc. ICASSP*, Vol. I, pp. 812-815, 2003.

[3] M. J. Ross, H. L. Shaffer, A. Cohen, R. Freudberg, and H. J. Manley, "Average magnitude difference function pitch extractor," IEEE Trans. ASSP, vol. 22, pp. 353-362, 1974.

[4]. Ealey, D., Kelleher, H., Pearce, D., "Harmonic tunneling:tracking non-stationary noises during speech", in *EUROSPEECH*, Aalborg, pp. 437–410, Sep. 1999.