TRANSFORM-DOMAIN WIENER FILTER FOR SPEECH PERIODICITY ENHANCEMENT

Feng Huang, Tan Lee

Department of Electronic Engineering
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong SAR

W. Bastiaan Kleijn

School of Electrical Engineering
KTH - Royal Institute of Technology
Stockholm, Sweden

ABSTRACT

In this paper, we present a transform-domain Wiener filtering approach for enhancing speech periodicity. The enhancement is performed on the linear prediction residual signal. Two sequential lapped frequency transforms are applied to the residual in a pitch-synchronous manner. The residual signal is effectively represented by two separate sets of transform coefficients that correspond to the periodic and aperiodic components, respectively. A Wiener filter operating on the transform coefficients is developed to restore periodicity and reduce noise. Different filter parameters are designed for the transform coefficients of the periodic and aperiodic components. A template-driven method is used to estimate the filter parameters for the periodic component. For the aperiodic components, the filter parameters are computed based on a local SNR for effective noise reduction. Experimental results confirm that the harmonic structure of the signal can be effectively restored with the proposed approach.

Index Terms— Speech periodicity enhancement, transform domain, Wiener filtering, sparse representation

1. INTRODUCTION

Periodicity is an important property of speech signals. Periodicity in speech signals is the result of periodic vibration of the vocal cords when voiced speech is produced. It determines the pitch of speech, which is essential to speech communication, especially for tonal languages. Important high-level linguistic information, e.g., intonation, lexical tones, stress and focus, is conveyed by the pitch contour of an utterance.

Restoring the periodicity of noise-corrupted speech is useful for improving perceptual quality of speech, in particular the perceptibility of pitch [1]. Comb-filtering is commonly used to suppress non-harmonic components in speech signals [2]. In [3] it was proposed to recover the harmonic structure of the original speech in the frequency domain. There have been relatively few studies on the enhancement of time-domain waveform periodicity. This is due to the difficulty of separating periodic and aperiodic components in a time-domain speech signal. In the area of hearing research, temporal periodicity enhancement was shown effective in improving pitch and tone perception [4]. However, severe nonlinear distortion was observed in the enhanced speech. In a preliminary study [5], we demonstrated the feasibility of waveform periodicity enhancement using a recently proposed speech representation model [6]. With the model, periodic and aperiodic components of a signal can be effectively separated in a transform domain. Periodicity enhancement is thus achievable.

In this study, we propose a Wiener filtering approach that operates in the transform domain for speech periodicity enhancement. The Wiener filter is employed to restore waveform periodicity and reduce noise. The enhancement is performed on the linear prediction (LP) residual signal. The LP residual is decomposed into periodic and aperiodic components using two-stage transforms. In the transform domain, the periodic component is concentrated and represented by a small portion of the coefficients. The Wiener filter for this subset of coefficients contributes to the restoration of the periodic impulse shape waveform. The filter parameters are estimated using a template-driven approach. A set of representative waveform cycles is learned from clean residuals. Their transform coefficients are used as templates, which are assumed to sparsely represent the coefficients of the periodic component. The constituent templates are estimated and used to determine the filter parameters. For the transform coefficients of the aperiodic components, the Wiener filter is mainly for noise reduction. The filter parameters are obtained based on local SNRs of the respective modulation bands. Our experimental results confirm that the harmonic structure of the signal can be effectively restored with the approach.

2. EFFECTIVE SIGNAL REPRESENTATION AND DECOMPOSITION

Let $e(n)$ denote the LP residual of an input speech signal. To allow effective separation of periodic and aperiodic components in the subsequent transforms, we first time-warp $e(n)$ to have a constant pitch period P_0 according to the speech pitch track [6]. Let $e(\nu)$ denote the warped residual signal and $e^{(k)}(\nu)$ denote the kth pitch-synchronous frame, i.e., $e^{(k)}(\nu) = e(kP_0+\nu)$, $\nu = 0, 1, \cdots, 2P_0-1$. The first transform is a modulated lapped transform, which aims to produce uncorrelated transform coefficients in different frequency channels. The DCT-IV transform is used. The transform coefficients $f(k,l)$ are computed by

$$f(k,l) = \sum_{\nu=0}^{2P_0-1} e^{(k)}(\nu) d(\nu) \sqrt{\frac{2}{P_0}} \cos\left(\frac{(2l+1)(2\nu-P_0+1)\pi}{4P_0}\right), \quad (1)$$

where $l = 0, 1, \cdots, P_0-1$ is the channel index and $d(\nu)$ is the square-root Hann window.

The second transform is a modulation transform that operates on $f(k,l)$. The purpose is to obtain a compact representation, as speech exhibits at times strong periodicity. This also means separating the periodic and aperiodic components of the signal. The DCT-II transform is used for effective energy concentration. Given a segment of Q pitch synchronous frames, the coefficients of the lth channel, i.e., $f(0,l), f(1,l), \cdots, f(Q-1,l)$, are transformed to generate Q output coefficients

$$g(q,l) = \sum_{k=0}^{Q-1} f(k,l)c(q)\sqrt{\frac{2}{Q}}\cos\left(\frac{(2k+1)q\pi}{2Q}\right), \qquad (2)$$

where $q = 0, 1, \cdots, Q-1$ is the *modulation band* index, $c(0) = \sqrt{1/2}$ and $c(q) = 1$ for $q \neq 0$. For the entire residual, Q of each individual segment for the modulation transform is determined based on an energy concentration measurement [5]. With the inverse transforms, the original signal can be reconstructed exactly from the transform coefficients.

Fig. 1 shows a segment of warped residual signal and the corresponding transform coefficients. It can be seen that most of the energy is concentrated in the low modulation bands, especially in the first band. The coefficients of the first modulation band actually represent the periodic component of the signal, while the remaining coefficients describe the aperiodic components. This can be easily understood by considering a strictly periodic signal. For such a signal, all pitch-synchronous frames are identical and hence the results of the pitch-synchronous transform are identical, i.e., $f(k_1,l) = f(k_2,l)$ for any $k_1, k_2 = 0, 1, \cdots, Q-1$. So the modulation transform for each channel is applied to a constant data sequence, and there is only one non-zero output coefficient at the first modulation band.

3. TRANSFORM-DOMAIN WIENER FILTERING

As the transform domain concentrates the signal energy, it is a particularly effective domain for the Wiener filtering operation. In this section, we discuss the principles of Wiener filtering in the transform domain for periodicity enhancement. Estimation of filter parameters will be presented in Section 4.

3.1. The MMSE optimal Wiener filter

LP residual signal is considered as the primary carrier of periodicity-related information in speech. Given a noisy residual signal, let us consider the task of recovering the underlying clean one. The noisy residual $r(\nu)$ is expressed as

$$r(\nu) = e(\nu) + u(\nu), \qquad (3)$$

where $e(\nu)$ is the clean residual and $u(\nu)$ is the noise. $e(\nu)$ and $u(\nu)$ are assumed to be uncorrelated. A common criteria for estimating $e(\nu)$ is to minimize the mean square error (MMSE) between the desired signal $e(\nu)$ and the estimated signal $\hat{e}(\nu)$. Wiener filter is optimal for this requirement. In the transform domain, the Wiener filter can be derived as

$$h(q,l) = \frac{g_e^2(q,l)}{g_e^2(q,l) + g_u^2(q,l)}, \qquad (4)$$

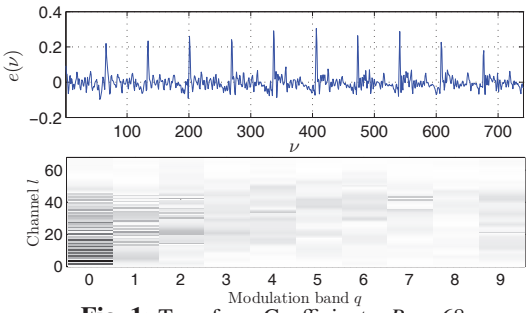

Fig. 1: Transform Coefficients. $P_0 = 68$.

where $g_e(q,l)$ and $g_u(q,l)$ are the transform coefficients of $e(\nu)$ and $u(\nu)$, respectively. The estimated or filtered coefficients for $\hat{e}(\nu)$ are then obtained by

$$\hat{g}_e(q,l) = h(q,l) \cdot g_r(q,l). \qquad (5)$$

In this paper, $h(q,l)$ is referred to as the filter parameter. Theoretically, $\hat{e}(\nu)$ reconstructed from $\hat{g}_e(q,l)$ is optimal in the sense that the signal-to-noise ratio (SNR) is maximized.

With the two-stage transforms, the signal is effectively represented by two sets of coefficients, i.e., coefficients in the first band for the periodic component and coefficients in the remaining bands for the aperiodic components. For the purpose of restoring waveform periodicity, we investigate the filter parameters for these two sets of coefficients separately.

3.2. Wiener filter for periodic component

For voiced speech, the residual is ideally an impulse train. Thus the coefficients of the first modulation band essentially represent the periodic impulse. For restoring waveform periodicity, we intend to obtain $\hat{g}_e(0,l)$ for each channel so that the desired waveform shape can be recovered. The filter parameter $h(0,l)$ plays an important role for *waveform shaping*. Denote $h(0,l)$ as $h_{\text{ws}}(l)$. For $l = 0, 1, \cdots, P_0 - 1$, we have

$$\hat{g}_e(0,l) = h_{\text{ws}}(l) \cdot g_r(0,l). \qquad (6)$$

3.3. Wiener filter for aperiodic components

For the coefficients of the high bands, it is observed that the noise energy becomes much higher than the energy of the speech residual. This is because the speech energy is concentrated in the first band and the energy left in the high bands is far weaker. On the other hand, for the noise, most energy is distributed in the high bands. Therefore, the Wiener filter for the high-band coefficients mainly contributes to *noise reduction*. For effective estimation of the filter parameters, we propose to determine the filter parameters based on a local SNR measure.

Denote SNR of the coefficient at channel l and band q as

$$\text{SNR}(q,l) = \frac{g_e^2(q,l)}{g_u^2(q,l)}. \qquad (7)$$

Eq. (4) can then be expressed as

$$h(q,l) = \left(1 + \frac{1}{\text{SNR}(q,l)}\right)^{-1}. \qquad (8)$$

Eq. (8) shows that $h(q,l)$ is determined by the SNR of the respective coefficient. Since the speech energy is weak, it is trivial to determine the SNR for each specific coefficient.

Considering that SNRs of different channels in the same high band are usually low and close, we employ a local SNR, known as the modulation band SNR, as defined below

$$\mathrm{SNR_b}(q) = \frac{\sum_l g_e^2(q,l)}{\sum_l g_u^2(q,l)}. \tag{9}$$

As the next section will show, $\mathrm{SNR_b}(q)$ can be effectively estimated. With $\mathrm{SNR_b}(q)$, the filter parameter for noise reduction in modulation band q is proposed as

$$h_{\mathrm{nr}}(q) = \left(1 + \frac{1}{\mathrm{SNR_b}(q)}\right)^{-1}. \tag{10}$$

Subsequently, the filtered coefficients are obtained by

$$\hat{g}_e(q,l) = h_{\mathrm{nr}}(q) \cdot g_r(q,l), \text{ for } q > 0. \tag{11}$$

4. PARAMETER ESTIMATION

To estimate $h_{\mathrm{ws}}(l)$ and $h_{\mathrm{nr}}(q)$, we assume that power spectrum of the noise is flat in the transform domain. This is reasonable because practical noise is normally aperiodic. In addiction, the speech pitch is used for signal warping. As a result, the transform coefficients of the noise usually do not have a particular portion with relatively high energy concentration. Based on this assumption, the filter parameters are estimated.

4.1. Filter parameters for aperiodic components

First, the noise variance $\hat{\sigma}_u^2$ is estimated with a codebook-driven method [7] that we use to estimate the LP coefficients for speech synthesis. The method is data-driven. For each codeword pair of LP parameters of clean speech and noise, optimal variances are computed for the speech and the noise. With corresponding optimal variances, the codeword pair that produces a spectrum closest to the noisy spectrum is selected as the result. From $\hat{\sigma}_u^2$, energy of the noise is obtained. For a block of signal $u(\nu_1), \cdots, u(\nu_2)$ used for the transforms, the total energy in modulation band q is estimated as

$$\xi_{\mathrm{b}} = \frac{(\nu_2 - \nu_1)\hat{\sigma}_u^2}{Q}, \tag{12}$$

where Q is the number of modulation bands as in Eq. (2). The modulation band SNR is then estimated as

$$\hat{\mathrm{SNR}}_{\mathrm{b}}(q) = \frac{\sum_l g_r^2(q,l) - \xi_{\mathrm{b}}}{\xi_{\mathrm{b}}}. \tag{13}$$

Substituting (13) into (10), we obtain

$$\hat{h}_{\mathrm{nr}}(q) = \max\left(\frac{\sum_l g_r^2(q,l) - \xi_{\mathrm{b}}}{\sum_l g_r^2(q,l)}, 0\right). \tag{14}$$

4.2. Filter parameters for periodic component

For the first-band coefficients, prior knowledge of clean residual is utilized to gain more accurate restoration of the impulse shape waveform. A template-based approach is proposed. A set of representative impulse waveforms is learned from clean (warped) residual signals. Each of them describes a cycle of impulse waveform of typical shape. Their variants with different phase shifts are also generated. Then for each waveform of one frame length, the transform coefficients are computed as a template. Denote the template in a vector form, i.e., $\bar{\mathbf{g}}_e = [\bar{g}_e(0,0)\ \bar{g}_e(0,1)\ \cdots\ \bar{g}_e(0,P_0-1)]^T$. From all templates, a prior information matrix $\mathbf{G}$ is composed as $\mathbf{G} = [\bar{\mathbf{g}}_e^{\{1\}}\ \bar{\mathbf{g}}_e^{\{2\}} \cdots \bar{\mathbf{g}}_e^{\{n\}} \cdots \bar{\mathbf{g}}_e^{\{N\}}]$, where $\bar{\mathbf{g}}_e^{\{n\}}$ denotes the nth template. $\mathbf{G} \in \mathcal{R}^{P_0 \times N}$ and $N \gg P_0$. $\mathbf{G}$ is used to represent the first-band coefficients of voiced residuals. Specifically, for a block of clean residual of Q pitch-synchronous frames, the transform coefficients $\mathbf{g}_e = [g_e(0,0)\ g_e(0,1)\ \cdots\ g_e(0,P_0-1)]^T$ are represented as a sparse linear combination of the templates,

$$\mathbf{g}_e = \mathbf{G}\mathbf{x}, \tag{15}$$

where $\mathbf{x} \in \mathcal{R}^{N \times 1}$ is a sparse weight vector. Eq. (15) is equivalent to representing the periodic impulse of the signal as an interpolation of a few similar ones from the templates. From the noisy coefficients $\mathbf{g}_r = [g_r(0,0)\ g_r(0,1)\ \cdots\ g_r(0,P_0-1)]^T$, the sparse weight $\mathbf{x}$ is then estimated by

$$\begin{aligned} \min_{\mathbf{x}} \quad & \|\mathbf{g}_r - \mathbf{G}\mathbf{x}\|_2^2 \\ \text{subject to} \quad & \|\mathbf{x}\|_1 \le \sqrt{Q}. \end{aligned} \tag{16}$$

With the estimated sparse weight $\hat{\mathbf{x}}$, we obtain

$$\hat{\mathbf{g}}_e = \mathbf{G}\hat{\mathbf{x}}. \tag{17}$$

The filter parameters are constructed from $\hat{\mathbf{g}}_e$ as

$$\hat{h}_{\mathrm{ws}}(l) = \frac{\hat{g}_e^2(0,l)}{\hat{g}_e^2(0,l) + \xi_{\mathrm{b}}/P_0}. \tag{18}$$

The constraint $\|\mathbf{x}\|_1 \le \sqrt{Q}$ in (16) is imposed to make sure that only the few templates with a similar impulse shape are selected [8]. The value $\sqrt{Q}$ is set based on the following considerations. For the sparse representation (15), if $\mathbf{g}_e$ is obtained from a single frame, we may require that all non-zero elements in $\mathbf{x}$ add up to 1 from an interpolation viewpoint. Consider the special case that the signal block consists of Q identical frames, then the corresponding $\mathbf{g}_e$ is simply a scaled version of the single frame coefficients with a scaling factor $\sqrt{Q}$. So in general, we impose $\|\mathbf{x}\|_1 \le \sqrt{Q}$.

5. EVALUATION

The proposed method is evaluated by objective assessment of the quality of enhanced residuals and enhanced speech signals. A previously proposed method [5] is used for comparison. For comparison of the quality of enhanced speech, the comb-filter method (**CombF**) [2] and the state-of-the-art codebook-driven (**CB**) method [7] are also used. For the approach in [5], $\hat{g}_e(q,l)$ is obtained by applying fixed weight (FxdWght) to the noisy coefficient, i.e., $\hat{g}_e(q,l) = w_q \cdot g_r(q,l)$, where $w_0 = 1$, $w_1 = \frac{2}{3}$, $w_2 = \frac{1}{3}$ and $w_q = 0$ for $q > 2$.

In the evaluation, 80 gender-balanced utterances from the CSLU-VOICES corpus [9] are used. They were down-sampled to 8 kHz. Half of the utterances are used for testing and the other half for parameter training. The template matrix $\mathbf{G}$ is trained in a speaker-dependent manner. For each speaker, the number of templates in $\mathbf{G}$ is 1000. Two types of noise are used. White noise was generated by software and car noise (VOLVO-340, 120km/h) was obtained from the NOISEX-92 database. Twelfth-order LP analysis, with a frame length of 24 ms and 50% overlap, is used to extract the residual signals.

Mean Segmental Harmonicity (SegHarm) [10] and global

SNR are used to assess the enhanced residual signals. SegHarm measures the overall energy ratio between the harmonic peaks and their surrounding noise. There are three kinds of pitch tracks used in the evaluation:

- F_0^{R}: obtained via manually labeling the clean waveforms. It is used as the reference for computing SegHarm. With F_0^{R}, SegHarm of the clean residuals is 1.91.
- F_0^{C}: estimated from clean speech using the method proposed in [11]. The gross pitch error rate is 2.5%.
- F_0^{N}: estimated from noisy speech using the method in [11]. Gross pitch error rates are 5.7% and 10.8% for the cases of 0 dB white noise and −10 dB car noise, respectively.

F_0^{C} and F_0^{N} are used for residual warping. For the warped residuals, the constant pitch period P_0 is 100 (warped-time samples). Objective measurements of the enhanced residuals are shown in Table 1. It can be seen that with both approaches of periodicity enhancement, SegHarm and SNR are significantly improved. The improvement implies that the harmonic structure of the residual signal is effectively restored and the noise is greatly suppressed. It is observed that the Wiener filtering method noticeably outperforms the fixed weight method. Particularly, the average SNR of the residuals by Wiener filter is significantly higher.

We further evaluate the quality of enhanced speech for the −10 dB car noise case. In the evaluation, F_0^{N} is used for residual warping. Output speech signals are synthesized with periodicity-enhanced (**PE**) residuals and LP coefficients estimated with the **CB** method. LP coefficients obtained from clean speech (**CleanLP**) are also used to illustrate the upper-bound performance. For comparison, speech signals enhanced by the **CombF** method and the **CB** method, respectively, are also obtained and evaluated. Global SNR, frequency-weighted segmental SNR (fwSegSNR) and the perceptual evaluation of speech quality (PESQ) are used to evaluate the enhanced speech [12]. Table 2 shows the results. It can be clearly seen that with periodicity enhancement, speech quality in terms of the objective measurements is significantly improved. With **PE**, performance of the **CB** method is noticeably improved. The Wiener filtering method surpasses the fixed weight method in all the results, especially in the SNR and PESQ results.

6. CONCLUSIONS

A Wiener filtering approach has been proposed for enhancing speech periodicity in a transform domain. The proposed Wiener filter operates on two separate sets of transform coefficients to restore waveform periodicity and reduce noise. Experimental results confirm that with the Wiener filter, the harmonic structure in the enhanced signals can be effectively restored and noise can be greatly suppressed. It is also confirmed that the Wiener filter method consistently outperforms the compared methods. With periodicity enhancement, performance of the codebook-driven method was noticeably improved. The improvement confirms the effectiveness of the transform-domain filtering approach.

Table 1: Objective measurements of periodicity-enhanced residuals

Noise	Cond.	Resid. SegHarm		Resid. SNR (dB)	
		F_0^{C}	F_0^{N}	F_0^{C}	F_0^{N}
White Noise (0 dB)	Noisy	1.19		−11.58	
	FxdWght	1.84	1.73	−5.67	−6.77
	Wiener	1.92	1.81	−2.02	−3.45
Car Noise (−10 dB)	Noisy	1.39		1.28	
	FxdWght	1.90	1.68	2.83	1.73
	Wiener	1.94	1.78	4.09	3.32

Table 2: Objective measurements of periodicity-enhanced speech

Method	Cond.	SNR (dB)	fwSNRseg (dB)	PESQ
Input		−10	3.74	1.93
CombF		−6.45	4.08	1.98
CB		−5.16	4.94	2.19
CB+PE	FxdWght	−3.20	6.43	2.26
	Wiener	0.21	7.08	2.61
CleanLP+PE	FxdWght	−1.36	9.16	2.65
	Wiener	2.11	10.45	3.15

ACKNOWLEDGMENT

This research is partially supported by the General Research Funds (Ref: CUHK 413507 & CUHK 414108) from the Hong Kong Research Grants Council, and a project grant from the Shun Hing Institute of Advanced Engineering, CUHK.

7. REFERENCES

[1] B. Cardozo and R. Ritsma, "On the perception of imperfect periodicity," *IEEE Trans. AE*, vol. 16(2), pp. 159 – 164, 1968.

[2] A. Nehorai and B. Porat, "Adaptive comb filtering for harmonic signal enhancement," *IEEE Trans. ASSP*, vol. 34, no. 5, pp. 1124–1138, 1986.

[3] E. Zavarehei, S. Vaseghi, and Q. Yan, "Noisy speech enhancement using harmonic-noise model and codebook-based post-processing," *IEEE Trans. ASLP*, vol. 15, no. 4, pp. 1194–1203, 2007.

[4] M. Yuan, T. Lee, and et al., "Effect of temporal periodicity enhancement on cantonese lexical tone perception," *JASA*, vol. 126, no. 1, pp. 327–337, 2009.

[5] F. Huang, T. Lee, and W. B. Kleijn, "A method of speech periodicity enhancement based on transform-domain signal decomposition," in *Proc. EUSIPCO '10*, Aug 2010, pp. 984–988.

[6] W. B. Kleijn, "A frame interpretation of sinusoidal coding and waveform interpolation," in *Proc. ICASSP '00*, Jun 2000, vol. 3, pp. 1475–1478.

[7] S. Srinivasan, J. Samuelsson, and W.B. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Trans. ASLP*, vol. 14(1), pp. 163–176, 2006.

[8] D. L. Donoho, "Compressed sensing," *IEEE Trans. IT*, vol. 52, pp. 1289–1306, 2006.

[9] A. Kain, "CSLU: VOICES," *Linguistic Data Consortium, Philadelphia*, 2006.

[10] A.-T. Yu and H.-C. Wang, "New speech harmonic structure measure and it application to post speech enhancement," in *Proc. ICASSP '04*, May 2004, vol. 1, pp. I–729–32.

[11] F. Huang and T. Lee, "Pitch estimation in noisy speech using accumulated peak spectrum and sparse estimation technique," *Submitted to IEEE Trans. ASLP*.

[12] P. C. Loizou, *Speech enhancement: theory and practice*, Signal processing and communications. CRC Press, 2007.