A METHOD OF SPEECH PERIODICITY ENHANCEMENT BASED ON TRANSFORM-DOMAIN SIGNAL DECOMPOSITION

Feng Huang[1], *Tan Lee*[1] *and W. Bastiaan Kleijn*[2]

[1]Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong, China
[2]School of Electrical Engineering, KTH - Royal Institute of Technology, Stockholm, Sweden
{fhuang, tanlee}@ee.cuhk.edu.hk, bastiaan@kth.se

ABSTRACT

Periodicity is an important property of speech signals. It plays a critical role in speech communication, especially when strong background noise is present. This paper presents a novel framework of periodicity enhancement for noisy speech. The enhancement operates on the linear prediction error (residual) signal. The residual signal goes through a constant-pitch time warping process and two sequential lapped frequency transforms, by which the periodic component is concentrated in the first modulation band. By emphasizing the respective transform coefficients, periodicity enhancement of noisy residual signal is achieved. The enhanced residual signal and estimated linear prediction filter parameters are used to synthesize the speech output. The effectiveness of the proposed method is confirmed consistently by various objective measures and subjective listening tests. It is observed that the enhanced speech can restore the harmonic structure of the original speech.

1. INTRODUCTION

Periodicity is an important property of speech signals. In the time domain, it is defined by the repetition of signal waveforms. In the frequency domain, periodicity is reflected by the appearance of strong spectral components at equally spaced harmonic frequencies. From the perspective of speech production, periodicity in acoustic signal is caused by periodic vibration of vocal cords when voiced speech is produced. It determines the pitch of speech, which is essential in speech communication. Important high-level linguistic information, for examples, intonation, lexical tones, stress and focus, is conveyed in the pitch contour of an utterance. Periodicity also means that there exists a great deal of redundancy in both the time and frequency domain. This contributes to the robustness of speech communication in noisy environments.

There have been numerous attempts to restore the periodicity of noisy speech signal, with the goal of improving perceptual quality. The approaches can be broadly categorized as spectral-domain harmonicity restoration techniques and time-domain waveform periodicity enhancement methods. Comb-filtering was a commonly used method, which attenuates signal components that are not harmonics [1]. In [2], a regeneration method was proposed to recover the harmonic structure of original speech. In [3], harmonicity enhancement was performed based on the harmonic+noise model of speech. In recent studies, harmonicity enhancement was typically applied as a post-processing step in general speech enhancement systems.

There have been relatively fewer 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 has been shown effective in improving pitch and tone perception. The commonly used techniques include increasing modulation depth and simplifying waveforms [4]. These methods generally cause severe nonlinear distortion and therefore lead to degradation of speech quality.

In this paper, we describe a new method of periodicity enhancement by exploiting a recently proposed speech representation model [5]. This speech model was developed to achieve a compact and complete representation of speech signals. The redundancy related to waveform periodicity is the basis of such representation. The speech model can be used for a wide range of applications including speech coding and prosodic modification. Our work on periodicity enhancement is based on an important property of this method, which is the effective periodic-aperiodic decomposition. The decomposition is applied on the residual signal of linear prediction (LP) analysis, which is considered to be the primary carrier of periodicity-related information in speech. The LP residual signal undergoes a series of transformations in a pitch-synchronous manner. Some of the transform coefficients represent the periodic component while the other coefficients represent the aperiodic component. Because noise signals generally do not have the same periodicity characteristic as speech, periodicity enhancement of noise-corrupted speech can be achieved by adjusting the relative contributions of the periodic and aperiodic components.

In Section 2, we first review the signal transformations as proposed in [5] and give illustrative examples of transformed and decomposed signals. Then the principle of periodicity enhancement is explained and some practical issues are discussed. In Section 3, the complete framework of speech periodicity enhancement is described. The problem of estimating LP parameters from noisy speech is addressed. Section 4 contains the experimental results in terms of both objective quality measures and results of subjective listening tests.

2. TRANSFORM-DOMAIN PERIODICITY ENHANCEMENT

Pitch or periodicity is caused by long-term dependencies in the speech signals, which are associated with the excitation source. It is carried primarily by the residual signal of LP analysis of speech. On the other hand, the LP filter coefficients characterize short-term dependencies that are caused by vocal tract resonances.

Periodic-aperiodic decomposition can be achieved on the

© EURASIP, 2010 ISSN 2076-1465

LP residual signal using the approach described in [5]. A brief review is given below.

2.1 Constant-pitch warping and lapped frequency transforms

Let $e(t)$ denote the LP residual signal from a voiced speech segment. We first time-warp the signal to have a constant pitch. We assume that the pitch track of the segment is available. Then a continuous pitch track is defined using a spline representation. The resulting pitch track is used to re-sample $e(t)$ to obtain a constant-pitch signal with period P_0. We write this signal as $e(\boldsymbol{t}(\tau))$, where the monotonic mapping $\boldsymbol{t}$ maps the constant-pitch time scale τ to the original time scale t.

If a signal segment contains both periodic and aperiodic components, they are concentrated in low- and high-frequency bands, respectively. Thus, to obtain an intuitive representation with energy concentration, we first separate the signal $e(\boldsymbol{t}(\tau))$ in frequency *channels*. This *pitch-synchronous* transform is implemented by a DCT-IV transform. The window size is $2P_0$ with 50% overlap. For a speech segment of K pitch-synchronous frames, the output of this transform includes $K \cdot P_0$ coefficients, denoted by $f(k,l)$, where k is frame (pitch-cycle) index and $l = 1,2,\cdots,P_0$, indexes the channels.

The transform that follows next is central to our algorithm as it separates out the periodic component from the signal. The periodic component of a channel is the component that does not change significantly from one pitch cycle to the next. Thus, we perform a frequency transform on each channel. This *modulation* transform is implemented by a DCT-II transform. For the lth channel, the coefficients $f(1,l), f(2,l),\cdots, f(K,l)$ are transformed to generate K output coefficients denoted by $g(q,l)$, where $q = 1,2,\cdots,K$ is the modulation band index. The modulation transform is performed over segments that are selected to maximize energy concentration. The practical implementation is done as an iterative process. Initially we start with a segment of one single frame, on which a measurement of energy concentration is computed [6, pp.A12]. The segment length is then increased by a step size of one frame, and the new energy concentration measurement is compared with the previous one. If it is increased, the process goes on with segment length further increased. If not, the length of the current segment is determined and a new segment starts.

Fig. 1 gives an example of applying the constant-pitch warping and transforms on a voiced speech segment. It shows the original and warped LP residual signals, and the output of the transforms. It is noticed that, in the transform output, most of the energy is concentrated in the low modulation bands, especially the first band.

2.2 Periodicity enhancement

In the transform domain, the coefficients of the first modulation band represent the periodic component of the signal, while the remaining coefficients describe the aperiodic component. 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., $f(i,l) = f(j,l)$ for any $i,j = 1,2,\cdots,K$ and $l = 1,2,\cdots,P_0$. In this case 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 (DC). This property suggests that periodic-aperiodic decomposition can be achieved by separating the low modulation band coefficients from the others.

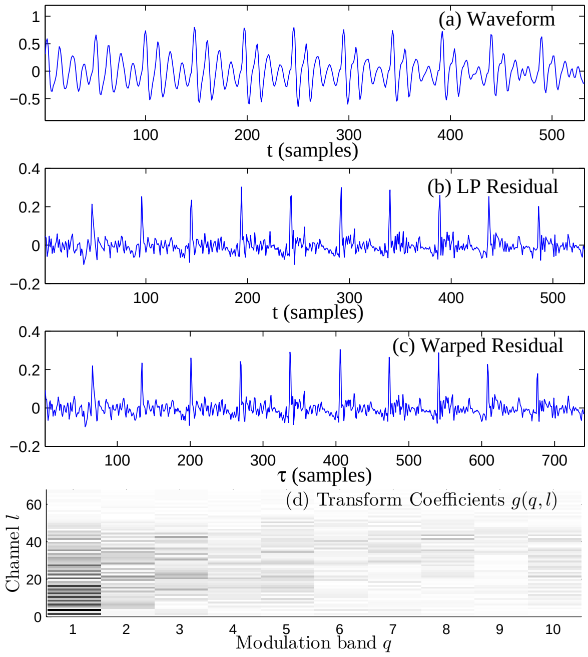

Figure 1: An example of constant-pitch warping and lapped frequency transforms of a voiced speech segment. $P_0 = 68$.

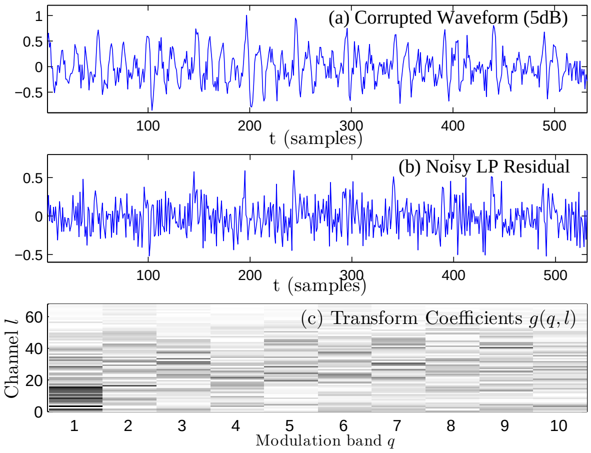

Figure 2: Effect of noise on transform coefficients.

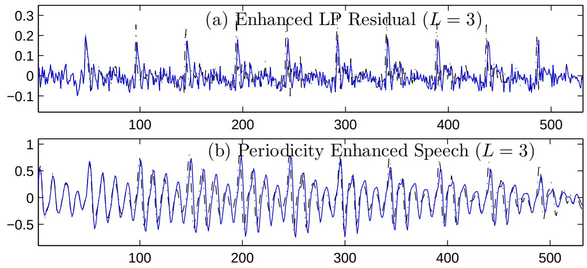

Figure 3: Periodicity-enhanced residual and speech waveforms. The blue solid lines are the results of enhancement while the black dashed lines are the clean counterparts.

In the presence of additive noise, the waveform periodicity of a speech signal is contaminated. Let us investigate how the transform-domain coefficients are affected by noise via an example as shown in Fig. 2. Fig. 2(a) shows the waveform of a noise-corrupted speech segment, which is obtained by adding white noise to the clean segment in Fig. 1(a). The SNR is 5dB. Fig. 2(b) gives the LP residual signal extracted from the noisy segment. Using the same pitch track as estimated from clean speech, we obtain the transform-domain coefficients as depicted in Fig. 2(c). Comparing Fig. 2(c) with Fig. 1(d), it is observed that the noise leads to an increase in energy in the high bands. Nevertheless, there is still a high level of energy concentration at the first modulation band, which represents the underlying periodic component.

Based on the above analysis, we propose to restore the periodicity of noise-corrupted speech by applying relatively heavier weights to the transform coefficients of the lower modulation bands and lighter weights to those of the higher bands. Let W_q denote the weighting factor for the qth modulation band. The modified transform coefficient $\hat{g}(q,l)$ is obtained as

$$\hat{g}(q,l) = W_q \cdot g(q,l). \tag{1}$$

Residual signal with enhanced periodicity is re-synthesized from $\hat{g}(q,l)$. In the experiments of this study, we use the following weighting scheme,

$$W_q = \begin{cases} \dfrac{L-q+1}{L} & q \leq L \\ 0 & q > L \end{cases} . \tag{2}$$

L is empirically set to 3. It can be seen that $0 \leq W_q \leq 1$ for all q. The periodic component is assigned the heaviest weight, i.e., $W_1 = 1$. For $1 < q \leq L$, the coefficients are attenuated. For $q > L$, the coefficients are discarded. By apply this weighting scheme to the example segment of Fig. 2, the enhanced residual and speech waveforms are given as in Fig. 3. It can be observed that the waveform periodicity is effectively restored.

2.3 Related issues

Pitch estimation The proposed method requires the pitch track of noise-corrupted input signal. An erroneous pitch track would cause problems in constant-pitch warping and affect the effectiveness of periodic-aperiodic decomposition. A typical pitch estimation algorithm [7] has a gross pitch error rate of about 5% at 0 dB SNR, i.e., 5% of the estimated pitch values[1] differ from the true values by 10 Hz or more.

In this study, a new algorithm of robust pitch estimation is used. The robustness is achieved by exploiting both pitch-related spectro-temporal information in speech and prior knowledge about pitch harmonics. Spectral peak pattern that is computed cumulatively over successive analysis frames is used as a robust feature for pitch estimation. The temporal cumulation effectively suppresses the effect of noise, which has irregularly located spectral peaks. For pitch estimation, the observed noisy pitch feature is assumed to be a sparse combination of clean feature exemplars. This combination is determined via an optimization procedure, which is very similar to the compressive sensing approaches [8]. Details of the algorithm is given in [9]. Preliminary experimental results show that the accuracy of pitch estimation at 0dB SNR is comparable to the noise-free case.

Segmentation and boundary smoothing In [5, 6], non-overlapping segments were used for speech coding applications. For periodicity enhancement, we observed that there were noticeable energy discontinuities at segment boundaries of the re-synthesized signal. This is due to the use of weighting factors at each individual segment. Subjective listening revealed that the discontinuities lead to severe perceptual distortion. To alleviate the problem, we impose a certain degree of overlapping between segments. At the synthesis stage, signals at the segment boundaries are smoothed by overlap-and-add with trapezoid windows.

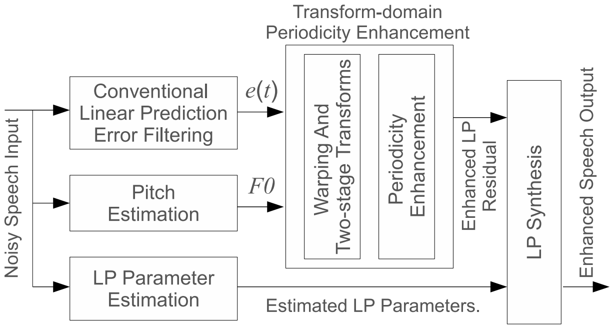

Figure 4: Complete framework of speech periodicity enhancement.

3. COMPLETE FRAMEWORK OF SPEECH PERIODICITY ENHANCEMENT

Fig. 4 gives the complete framework of the proposed method of speech periodicity enhancement. Noisy speech is first processed using conventional linear prediction error filtering (autocorrelation based). Noisy LP residual signal $e(t)$ is obtained for subsequent periodicity enhancement as described in Section 2. Meanwhile, a procedure for LP parameter estimation is carried out. The estimation aims to acquire an estimation of the clean LP parameters from the noisy speech input. Finally, the estimated LP parameters are used in conjunction with the enhanced residual signal to generate the speech output.

The problem of estimating LP parameters from noisy speech has been studied for many years. It aims at estimating the speech spectrum and the excitation gain. The representative approaches include noise compensation [10], codebook-driven estimation [11], and Kalman filtering [12, 13]. In this study, we adopt the codebook-driven approach [11] and the iterative Kalman filtering approach [12] for evaluation. The codebook method is data driven, where LP coefficients are estimated by searching over trained codebooks of clean speech and noise for a codeword pair that has the highest probability to produce the noisy observation. In the Kalman filter approach, LP filter coefficients are estimated iteratively. Each frame of speech is first enhanced by the Kalman filter initialized with the LP coefficients of noisy speech. A set of new coefficients are then estimated from the enhanced speech. The process goes on iteratively until convergence is reached [12].

4. EXPERIMENTS AND RESULTS

Performance of the proposed method is evaluated on two aspects: (1) effectiveness of periodicity enhancement of the LP residual signal, and (2) overall performance of speech periodicity enhancement with estimated LP parameters. The evaluation data consists of a total of 48 speech utterances from 3 different languages: American English, Mandarin and Cantonese. While English is used to represent western lan-

[1] In this paper, the terms "pitch" and "$F0$" (fundamental frequency) are used interchangeably.

Table 1: Performance of periodicity enhancement on LP residual signals under different input noise conditions.

	Input Speech SNR(dB)	Residual SegHarm			Residual SNR (dB)		
		Noisy	Enhanced		Noisy	Enhanced	
			R-$F0$	E-$F0$		R-$F0$	E-$F0$
White Noise	5	1.16	1.91	1.82	-8.10	-1.40	-1.72
	0	0.86	1.71	1.62	-11.91	-4.38	-5.04
	-5	0.61	1.47	1.38	-16.26	-8.16	-9.08
AR Noise	5	0.89	1.68	1.62	-4.71	0.93	0.30
	0	0.67	1.49	1.34	-8.45	-1.52	-2.04
	-5	0.56	1.32	1.19	-11.9	-4.39	-5.29
mean		0.79	1.60	1.51	-10.22	-3.15	-3.81

guages, Mandarin and Cantonese are among the most representative tonal languages, in which pitch is used to differentiate words. There are 16 utterances (equal number of male and female speakers) for each language. They are taken from TIMIT (English), 863 (Mandarin) and CUSENT (Cantonese), respectively. Mean duration of one utterance is about 4-5 seconds. Speech activity ratio[2] of the data set is 85% on average. Speech are down sampled to 8 kHz.

Reference pitch, denoted as R-$F0$, is obtained on clean speech using conventional time-domain autocorrelation method and the results are manually verified. Estimated pitch, denoted as E-$F0$, is obtained with the algorithm as described in Section 2.3. Twelfth-order LP analysis is applied to obtain the residual signals. The analysis frame is 20 ms long, with 50% overlap. For the proposed method, we applied the same weighting scheme on unvoiced segments as well as the voiced segments.

4.1 Periodicity enhancement of LP residual signal

In the first experiment, speech signals are degraded by two types of noise: white noise and first-order AR noise (simulating car noise [13]), at SNR of -5, 0 and 5 dB, respectively. Periodicity enhancement based on R-$F0$ and E-$F0$ is performed on the noisy LP residual signals.

We use the Mean Segmental Harmonicity (SegHarm) [14] and the global SNR of the residual signal as the performance indices. SegHarm measures the overall energy ratio between the harmonic peaks and their surrounding noise in the target signal. It is computed from all voiced segments in the utterances. Avarage SegHarm value of the clean residual signals is 1.72. Table 1 gives the SegHarm and global SNR of the residual signals before and after enhancement. Significant improvements can be observed on both types of noise at all input SNR levels. The average value of SegHarm increases from 0.79 to 1.60 and 1.51, when R-$F0$ and E-$F0$ are used respectively.

4.2 Objective quality assessment of enhanced speech

We also evaluate the quality of periodicity-enhanced speech. The methods being tested are "codebook-driven LP parameter estimation [11] + periodicity-enhanced LP residual" (**CB+PE**), "iterative Kalman filtering [12] + periodicity-enhanced LP residual" (**KF+PE**). They are compared with "clean LP parameters + periodicity-enhanced LP residual" (**CleanLP+PE**) and the comb-filter method (**CombF**) [1]. We are also interested to compare the two LP parameter estimation methods, i.e., "codebook-driven estimation" (**CB**), and "iterative Kalman filtering" (**KF**), without using enhanced residual signals.

Table 2: Performance of the evaluated speech enhancement methods.

	SNR (dB)	fwSNRseg (dB)	CEP	PESQ (MOS)
Input	0	2.43	6.19	1.56
CB	2.23	3.61	4.82	1.71
KF	1.37	2.74	5.27	1.64
CB+PE	3.38	4.40	4.15	2.41
KF+PE	1.72	3.67	4.86	2.00
CleanLP+PE	5.16	8.25	3.48	3.02
CombF	2.69	2.48	6.20	1.72

The speech utterances are corrupted by additive AR noise at 0 dB SNR. E-$F0$ is used for residual enhancement. For codebook-based LP parameter estimation, the speech codebooks are language-dependent. For each language, 24 utterances that are different from the test data are used to train a codebook with 2048 codewords. The size of noise codebook is 48. It is trained with a noise signal of 2-second length.

Global SNR, frequency-weighted segmental SNR (fwSegSNR), cepstrum distance (CEP) and the perceptual evaluation of speech quality (PESQ) are used as quality measures [15]. The results are shown in Table 2. It can be seen that both approaches of LP parameter estimation (**CB** and **KF**) can improve the speech quality to certain extent. **CB** is more effective than **KF**. With periodicity enhancement of residual signals, the speech quality is further improved. The PESQ value attained by **CB+PE** is 2.41, as compared to 1.71 by **CB** and 1.72 by **CompF**. The PESQ value of **CleanLP+PE**, i.e., 3.02, can be considered as the performance upper bound of the proposed approach in this noise condition.

Fig. 5 gives an example that shows the waveform and spectrograms of speech output enhanced by **CB** and **CB+PE**. It can be seen that **CB** is useful to recover the formant structure. With the use of periodicity-enhanced residual signal, the harmonic structure can be effectively restored. This is especially noticeable in the high-frequency region.

4.3 Subjective quality assessment of enhanced speech

Subjective listening tests were carried out on the enhanced speech as evaluated above. The tests are designed and conducted following the procedures in [16] and [17]. Half of the test utterances were used, i.e., 8 utterances for each language.

[2] Duration of speech (excluding silence) over duration of the whole utterance.

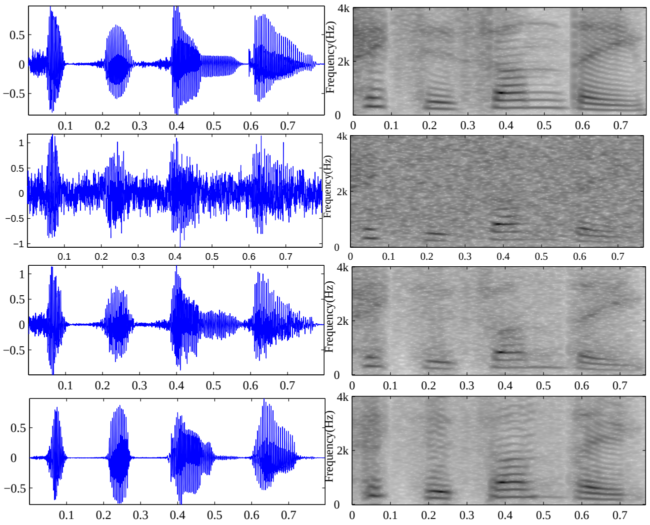

Figure 5: Waveforms and spectrograms of clean, noisy, **CB** enhanced and **CB+PE** enhanced speech (from top to bottom). Audio samples are available at http://www.ee.cuhk.edu.hk/~fhuang/pe.html .

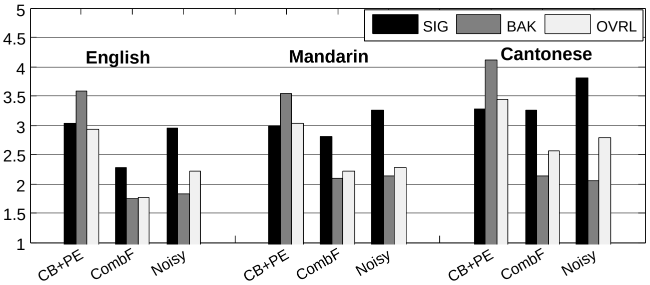

Figure 6: Results of subjective listening tests.

They were randomly selected from the full test set.

A total of 12 subjects participated into the listening tests. They include 6 native Cantonese speakers and 6 native Mandarin speakers, who were involved in the assessment of Cantonese and Mandarin utterances, respectively. In addition, all of the 12 subjects were asked to assess the English utterances. Each subject was required to rate a presented signal on [17]:

SIG: signal distortion, [5=very natural, 1=very unnatural];

BAK: background noise intrusiveness, [5=not noticeable,1=very conspicuous and very intrusive];

OVRL: overall effect, [5=excellent, 1=bad].

The test results are given as in Fig. 6. In terms of OVRL, the proposed method **CB+PE** significantly outperforms **CombF** for all of the three languages. The overall average of OVRL scores is 3.13, as compared with 2.18 for **CombF**. Both **CB+PE** and **CombF** introduce noticeable signal distortion and thus lead to lower SIG scores. **CB+PE** consistently attains a high BAK score, indicating that the effect of background noise has been effectively suppressed.

5. CONCLUSION

A novel framework of speech enhancement has been developed and evaluated. We have shown that enhancement of speech and/or suppression of noise can be effectively achieved by processing the LP parameters and the residual signal separately. The focus of this paper is on enhancing the pitch-related periodicity characteristic in the residual signal. With pitch track robustly estimated from noisy speech, the proposed method demonstrates significant improvement in both the signal-to-noise ratio and the perceptual quality of speech. Two previously proposed methods of LP parameter estimation have been adopted for evaluation. Quality of enhanced speech can be further improved with more accurate representation of speech spectrum and better restoration of the unvoiced segments.

6. ACKNOWLEDGEMENT

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.

REFERENCES

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

[2] C. Plapous, C. Marro, and P. Scalart, “Speech enhancement using harmonic regeneration,” in *Proc. ICASSP ’05*, March 18–23, 2005, vol. 1, pp. 157–160.

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

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

[5] W. B. Kleijn, “A frame interpretation of sinusoidal coding and waveform interpolation,” in *Proc. ICASSP ’00*, 2000, vol. 3, pp. 1475–1478.

[6] M. Nilsson, *Entropy and Speech*, Ph.D. thesis, Royal Institute of Technology (KTH), 2006.

[7] T. Shimamura and H. Kobayashi, “Weighted autocorrelation for pitch extraction of noisy speech,” *IEEE Trans. SAP*, vol. 9, no. 7, pp. 727–730, Oct. 2001.

[8] J. F. Gemmeke and B. Cranen, “Noise reduction through compressed sensing,” in *Proc. Interspeech ’08*, 2008, pp. 1785–1788.

[9] Feng Huang and Tan Lee, “Robust pitch estimation using harmonic peak cumualtion and compressive sensing technique,” *To be submitted.*

[10] C. E. Davila, “A subspace approach to estimation of autoregressive parameters from noisy measurements,” *IEEE Trans. SP*, vol. 46, no. 2, pp. 531–534, Feb. 1998.

[11] M. Kuropatwinski and W. B. Kleijn, “Estimation of the excitation variances of speech and noise AR-models for enhanced speech coding,” in *Proc. ICASSP 2001*, vol. 1, pp. 669–672.

[12] J. D. Gibson, B. Koo, and S. D. Gray, “Filtering of colored noise for speech enhancement and coding,” *IEEE Trans. SP*, vol. 39, no. 8, pp. 1732–1742, Aug. 1991.

[13] M. Kuropatwinski and W. B. Kleijn, “Estimation of the short-term predictor parameters of speech under noisy conditions,” *IEEE Trans. ASLP*, vol. 14, no. 5, pp. 1645–1655, Sept. 2006.

[14] A.-T. Yu and H.-C. Wang, “New speech harmonic structure measure and it application to post speech enhancement,” in *Proc. ICASSP ’04*, 17–21 May 2004, vol. 1, pp. I–729–32.

[15] Yi Hu and P. C. Loizou, “Evaluation of objective quality measures for speech enhancement,” *IEEE Trans. ASLP*, vol. 16, pp. 229–238, Jan. 2008.

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

[17] Yi Hu and Philipos C. Loizou, “Subjective comparison and evaluation of speech enhancement algorithms,” *Speech Commun.*, vol. 49, no. 7-8, pp. 588–601, 2007.