EXPLOITING THE HARMONIC STRUCTURE FOR SPEECH ENHANCEMENT

Eunjoon Cho[*†], *Julius O. Smith III*[*], *Bernard Widrow*[†]

[*] Center for Computer Research in Music and Acoustics, Stanford University,
[†] Department of Electrical Engineering, Stanford University

ABSTRACT

We provide a single channel speech enhancement method leveraging the harmonic structure of voiced speech. A sinusoidal model, based on the pitch of the speaker, is used to filter noisy speech and remove any noise components that lie between the harmonics. To remove noise that lie on each harmonic frequency, we use a noise estimation procedure that exploits spectral sparsity of voiced speech. By measuring the power spectrum at frequencies that correspond to the zero crossings of the windowing function, we can estimate the noise levels even in frames that have voiced speech. We also provide a constrained linear least squares formulation to reduce "musical noise" which arises from difficulty in estimating speech and noise power spectral densities. We show that our method yields high perceptual performance over existing methods, and can easily adapt to conditions in which the noise characteristics are constantly changing.

Index Terms— Speech enhancement, Noise estimation, Harmonic filter

1. INTRODUCTION

Although there have been approaches to use multiple channels in enhancing speech [1], noise reduction on a single channel still remains a challenge in most situations.

An indicator one uses to pick out the speaker's voice from background noise is the speaker's pitch. Adaptive comb filters [2] or sinusoidal modeling [3] have been used to preserve the harmonic structure of the speaker's voice. We apply a similar method by applying a harmonic structured filter on the noisy signal to reduce any noise that might exist in between harmonics. An essential issue with harmonic filters is that any noise that lie specifically on the pitch and its harmonics should be reduced as well. Since humans are sensitive to the spectral peaks and these are what constitute the formant structure of speech, it is important that an accurate noise estimate is available at the peaks.

A large group of algorithms that perform noise estimation emphasizes temporal sparsity of speech. These range from methods using voice activity detection (VAD) to methods that estimate noise during speech activity using minimum statistics or recursive time averaging [4]. These methods assume the speech power is low at a recent speech frame in the past, and uses this frame to update the noise estimate.

We show that an accurate noise estimate can be made in each time frame by exploiting sparsity in the other domain, i.e., the spectral domain. The harmonic structure of voiced speech enables us to estimate the noise levels in-between harmonics, where the speech power is low. Previous works [5, 6] briefly touch on this idea for different applications. In this paper, we show that the exact frequencies in which no speech should theoretically exist can be calculated from the zero crossings of the windowing function in each harmonic band. With this method, an accurate yet adaptive noise estimate can be achieved in every time frame of the noisy speech, even if it is voiced.

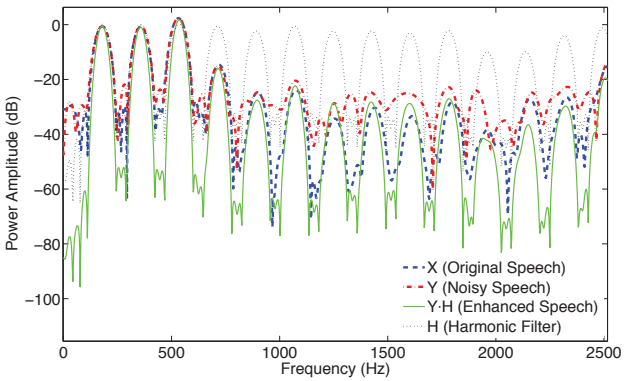

Fig. 1. Spectrum of underlying speech signal X, noisy speech signal Y, pitch harmonic filter H, and the filtered output signal $Y \cdot H$

From the noise estimate we adjust the gain of our harmonic filter in each harmonic frequency band with respect to the corresponding SNR. Apart from the difference that the filter is modeled to have a harmonic structure, our method largely falls into the general category of spectral subtraction methods. A well known shortcoming of these methods is that they introduce "musical noise" which appear from the spectral islands of the de-noised spectrum and is caused by the half-wave rectification process [7]. To minimize this, we formulate our problem as a linear constrained least squares problem, in which the difference between neighboring harmonic peaks are constrained. This prevents consecutive harmonic gains from fluctuating and helps make the de-noised signal more pleasing to the human ear.

2. MODELING THE HARMONIC FILTER

Given a noisy speech signal $y(n) = x(n) + e(n)$, we aim to recover the underlying speech signal, $x(n)$, given no information about the noise, $e(n)$.

It is well known that voiced speech can be decomposed as an impulse train generated at the vocal cords and is convoluted with a filter, $v(n)$, caused by the vocal tract. Given the hop size is R, the $m'th$ windowed frame of the signal $y(n)$ can be expressed as,

$$\begin{aligned} y_m(n) &= y(n+mR)w(n) \\ &= (x(n)+e(n))_m\, w(n) \\ &= \left(\sum_{k=-\infty}^{\infty} \delta(n-kT) * v(n) + e(n)\right)_m w(n) \end{aligned} \quad (1)$$

The Fourier Transform of this signal is,

$$Y_m(f) = \frac{1}{T}\sum_{k=-\infty}^{\infty} V_m\left(\frac{k}{T}\right) W\left(f-\frac{k}{T}\right) + \tilde{E}(f) \quad (2)$$

We apply a time varying harmonic filter $H_m(f)$ on the noisy speech spectrum $Y_m(f)$. This harmonic filter has the same shape as the underlying voiced speech. i.e.,

$$H_m(f) = \sum_{k=-\infty}^{\infty} A_k W\left(f-\frac{k}{T}\right) \quad (3)$$

The idea of the harmonic filter is to capture the voiced speech components in Eq. (2) and remove any frequency components that do not correspond to the pitch and its harmonics. Fig. 1 illustrates this idea of applying a harmonic filter on the noisy signal. Methods for estimating the pitch is discussed in Sec. 2.1.

With this procedure we are able to reduce noise at frequencies that cannot be explained with the voiced speech model. We, however, also have to reduce any noise that lie specifically on the harmonic frequencies. The parameter, A_k, governs the gain at the harmonic peaks and is determined by the noise level in the k'th harmonic band. We propose a method that exploits the sparsity of voiced speech to estimate the noise in Sec. 2.2.

Details of how to incorporate the pitch and noise estimate in modeling the harmonic filter is presented in Sec. 2.3. Additional linear constraints to minimize "musical noise" can be included in a least squares problem and is described in Sec. 2.4. Finally, although the pitch harmonic filter is modeled for voiced speech, we discuss in Sec. 2.5 how it can handle de-noising of unvoiced speech components.

2.1. Pitch estimation

Pitch detection is by itself a widely studied field, and there is a vast variety of work on effective and accurate pitch detection methods. Although an accurate estimate of the pitch is important, the focus of this paper is not on searching effective methods of pitch detection. We have thus implemented a simple algorithm where we search for the highest peak of the log power spectrum in a pre-determined frequency range using quadratic interpolation, and set that as our estimate of the speaker's pitch. To be more robust we can look at the second and third harmonics, and refine our estimate. An average of these values is used as the final pitch estimate.

Even under highly noisy conditions this simple method showed to be effective because most of the noise usually exist in high frequencies, and thus the peak of the fundamental frequency is usually detectable. We have also used other methods of pitch estimation using autocorrelation or cepstrums, and have observed similar quality estimates.

2.2. Noise estimation

The noise power at each harmonic frequency band can be estimated by exploiting the spectral sparsity of voiced speech signals. This enables us to maintain an accurate estimate of the noise even in frames where the voiced speech is dominant.

First, we assume C_k to be a limited set of frequencies at zero-crossings of the window centered at the $k'th$ harmonic frequency. Assuming the pitch is $1/T$, and using a Hann window of width B,

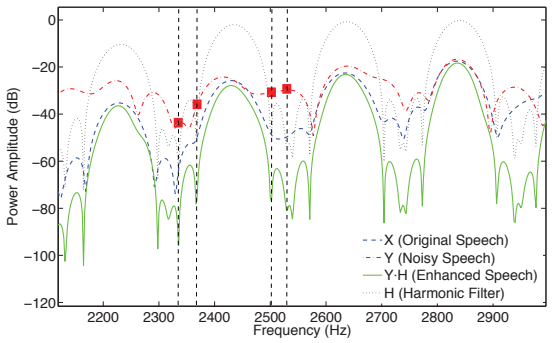

Fig. 2. Noise estimation of frequency band, k, centered at 2430Hz. The vertical dashed lines correspond to the zero crossings of the filter centered at frequency 2430Hz. The 4 red squares are the values used to approximate the noise level in this frequency band.

$$C_k = \left\{\frac{k}{T} \pm \frac{m}{B} \mid m = 2,3,4,....; \frac{m}{B} \leq \frac{1}{2T}\right\} \quad (4)$$

From this we can estimate the noise level at the k'th harmonic frequency band by averaging the power spectrum at neighbor frequencies corresponding to C_k.

$$\left|\hat{E}_m\left(\frac{k}{T}\right)\right|^2 = \frac{1}{|C_k|}\sum_{f'\in C_k}\left|Y_m\left(f'\right)\right|^2 \quad (5)$$

$$\approx \frac{1}{|C_k|}\sum_{f'\in C_k}\left|X_m\left(f'\right)\right|^2 + \left|E_m\left(f'\right)\right|^2 \quad (6)$$

$$\approx \frac{1}{|C_k|}\sum_{f'\in C_k}\left|E_m\left(f'\right)\right|^2 \quad (7)$$

Eq. (6) follows from the assumption that speech and noise is uncorrelated, and Eq. (7) is from Eq. (2) where we have shown that the spectrum of a clean speech frame is zero at the zero crossings of the windowing function. Fig. 2 illustrates this procedure.

To acquire a more stable estimate of the noise, we recursively smooth our estimate over previous time frames. i.e.,

$$\left|\hat{\hat{E}}_m\left(\frac{k}{T}\right)\right|^2 = \alpha\left|\hat{\hat{E}}_{m-1}\left(\frac{k}{T}\right)\right|^2 + (1-\alpha)\left|\hat{E}_m\left(\frac{k}{T}\right)\right|^2 \quad (8)$$

The tradeoff of this procedure is that we lose temporal resolution and are more susceptible to abrupt changes in noise characteristics. We use a α value of 0.8 in our experiments.

2.3. Model of the harmonic filter

We now show how the noise estimate in Sec. 2.2 can be used to set the gain, A_k, of the harmonic filter. Given that we want the filter applied signal to be equivalent to the underlying signal, we have:

$$\left|Y_m\left(\frac{k}{T}\right)H_m\left(\frac{k}{T}\right)\right|^2 = \left|X_m\left(\frac{k}{T}\right)\right|^2$$

$$\left|Y_m\left(\frac{k}{T}\right)\sum_{i=-\infty}^{\infty}A_i W\left(\frac{k}{T}-\frac{i}{T}\right)\right|^2 = \left|X_m\left(\frac{k}{T}\right)\right|^2$$

$$\left|Y_m\left(\frac{k}{T}\right)A_k W(0)\right|^2 \approx \left|X_m\left(\frac{k}{T}\right)\right|^2 \quad (9)$$

$$|A_k|^2 \approx \frac{\left|Y_m\left(\frac{k}{T}\right)\right|^2 - \left|E_m\left(\frac{k}{T}\right)\right|^2}{\left|Y_m\left(\frac{k}{T}\right)\right|^2} \quad (10)$$

We assume in Eq. (9) that the side lobe levels of the pitch filter will be significantly small and can be ignored at all indices where $i \neq k$. Eq. (10) is from the fact that A_k and $W(0)$ are real, and that we use a normalized filter such that $W(0) = 1$. Also, we assume that the signal and noise is uncorrelated, i.e., $|Y_m(f)|^2 = |X_m(f)|^2 + |E_m(f)|^2$.

From the noise estimate $\hat{\hat{E}}_m(\frac{k}{T})$, in the previous section, we can calculate the gain in each harmonic peak of our filter. We lower bound $|A_k|^2$ to a value greater than 0 such that the gain is positive. This half-wave rectification causes uneven spectral islands and results in "musical noise", which is a well known phenomenon in spectral subtraction based algorithms. This artifact can be prevented by restricting the values of $|A_k|^2$ such that it does not vary too much across harmonics. We show in the next section that all of this can be modeled as a simple least squares optimization problem.

2.4. Least squares formulation

From Eq. (10), we can formulate a least squares problem.

$$\mathbf{x} = \begin{bmatrix} |A_1|^2 \\ |A_2|^2 \\ \vdots \\ |A_K|^2 \end{bmatrix}, \ \mathbf{y} = \begin{bmatrix} 1 - \frac{|\hat{\hat{E}}(1/T)|^2}{|Y(1/T)|^2} \\ 1 - \frac{|\hat{\hat{E}}(2/T)|^2}{|Y(2/T)|^2} \\ \vdots \\ 1 - \frac{|\hat{\hat{E}}(K/T)|^2}{|Y(K/T)|^2} \end{bmatrix}$$

The first constraint is on $\mathbf{x}$, i.e., $\delta \leq x_k \leq 1$ for all k. The lower bound δ controls how much noise reduction is performed, where the smaller δ means a stronger noise reduction is applied. In our experiments we used a δ value of 1e-4.

The second constraint is to enforce neighboring values of $\mathbf{x}$ to be small such that the fluctuation of gains across harmonics is small, i.e., $|x_k - x_{k+1}| < \epsilon$ for $k = 1, ..., K-1$.

We can apply this constraint as follows.

$$\mathbf{Cx} = \begin{bmatrix} 1 & -1 & 0 & ... & 0 & 0 \\ 0 & 1 & -1 & ... & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & ... & 1 & -1 \\ -1 & 1 & 0 & ... & 0 & 0 \\ 0 & -1 & 1 & ... & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & ... & -1 & 1 \end{bmatrix} \mathbf{x} < \begin{bmatrix} \epsilon \\ \epsilon \\ \vdots \\ \vdots \\ \vdots \\ \epsilon \\ \epsilon \end{bmatrix}$$

Given all of this our least squares problem is simply,

$$\text{minimize } ||\mathbf{x} - \mathbf{y}||^2 \quad (11)$$

$$\text{subject to } \mathbf{Cx} < \epsilon \quad (12)$$

$$\delta \leq \mathbf{x} \leq 1 \quad (13)$$

which can be solved using any standard least squares solver.

2.5. Unvoiced speech

Although majority of speech is voiced, preserving unvoiced speech, such as fricatives and plosives, helps increase the intelligibility. Since our filter emphasizes the harmonic structure of voiced speech, it is natural to believe that it will over-estimate the noise powers for unvoiced speech and thus suppress the speech more than it should for unvoiced speech.

A way to handle this would be to use a voice activity detector that can classify frames into voiced, unvoiced and noise only frames. Different levels of suppression can be used once a VAD can classify the speech frame into one of these classes. One simple method that we have used is to check the power level at low frequencies and the difficulty in finding the pitch as an indicator of whether the speech frame is voiced. If we estimate that the frame is not voiced, we apply a higher lower bound than voiced frames, so that the level of suppression is not as strong.

3. EVALUATION

Our algorithm can be divided into two stages: noise estimation using harmonic sparsity (Sec. 2.2) and a harmonic filter based on the noise estimate (Sec. 2.3). To understand the effectiveness of each stage properly, we first conducted experiments on the noise estimation method itself, and then evaluated the overall speech enhancement method. All methods were evaluated using 8 different speeches in the CMU ARCTIC database [8], and each speech was combined with either white or babble noise from the NOISEX-92 database [9].

To independently assess the quality of our noise estimation method, we compared the similarity between the estimated and true underlying noise power spectra, as done in [7]. We use the median normalized squared error to assess how similar the estimation is. Table 1 shows the comparison of our method against other algorithms [4, 10, 11].

A shortcoming of this measure is that we cannot really compare it against the true noise PSD, and it does not show how it actually performs with speech enhancement algorithms. We have therefore used these noise estimation algorithms on a spectral subtraction based speech enhancement method [12] and compared its PESQ (Perceptual Evaluation of Speech Quality) measure. Various studies have been conducted to provide different objective measures on the quality of enhanced speech. The PESQ measure has shown to have high correlation with subjective MOS (Mean Opinion Score) measures conducted on listeners [13], and is the reason for selecting it as a performance measure.

Fig. 3 shows the performance of noise estimation methods on different noise type and power combinations. The harmonic based noise estimation method shows good performance in most cases except in low SNR babble noise environments in which the pitch of the speaker is more difficult to detect.

To evaluate the overall speech enhancement method, we compare the PESQ of our method against other literature. To achieve a balanced comparison, we have selected a spectral subtraction method [14], a Log-MMSE method [12] and a method that uses

Noise	Martin [4]	Cohen [10]	Rangachari [11]	Harmonic
White	0.727	0.834	2.062	0.603
Babble	0.875	0.772	0.838	0.809

Table 1. Average MedSE of harmonic noise estimation compared with other methods in literature.

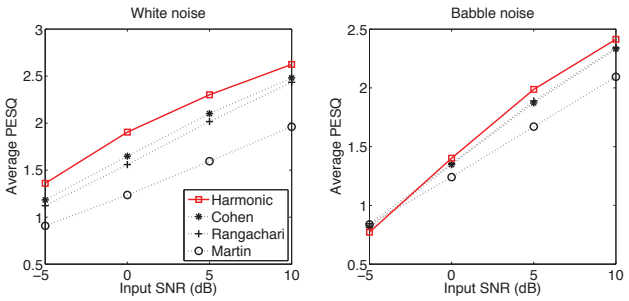

Fig. 3. Comparison of harmonic based noise estimation with other noise estimation methods. A standard spectral subtraction speech enhancement algorithm was used with the noise estimates.

subspace projection [15]. Fig. 4 shows how our algorithm yields notable improvement over competing methods.

4. CONCLUSION

We have proposed a speech enhancement algorithm that relies on the harmonic structure of the underlying speech. A filter is modeled exploiting the pitch of the speaker and we have used it to reduce any noise that lie between the harmonics of the speaker's voice. The noise at each harmonic frequency is estimated by sampling nearby frequencies that should ideally be zero because of the window spectrum. The noise estimate and some additional constraints are formulated as a least squares problem to find the optimal gain at each harmonic frequency.

Our method benefits in that a real-time noise estimate can be made in every time frame even if it is voiced speech. The harmonic filter also preserves the underlying structure of the voiced speech and thus helps reduce background noise without harming the speech itself. We have focused mostly on exploiting spectral sparsity in this paper, but one could look more into how this could be integrated with previous noise estimation methods that exploit temporal sparsity of speech.

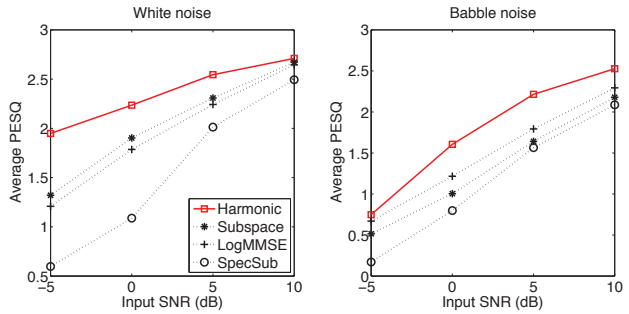

Fig. 4. Comparison of our harmonic speech enhancement method with other algorithms

5. REFERENCES

[1] J. Bitzer, K. U. Simmer, and K. Kammeyer, "Muti-microphone noise reduction techniques as front-end devices for speech recognition," *Speech Communication*, vol. 34, pp. 3–12, 2001.

[2] W. Jin, X. Liu, and M. S. Scordilis, "Speech enhancement using harmonic emphasis and comb filtering," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 2, pp. 356–368, 2010.

[3] J. Jensen and J. H. L. Hansen, "Speech enhancement using a constrained iterative sinusoidal model," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 7, pp. 731–740, 2001.

[4] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Audio Processing*, pp. 504–512, 2001.

[5] J. Beh and H. Ko, "A novel spectral subtraction scheme for robust speech recognition: spectral subtraction using spectral harmonics of speech," in *IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP '03)*, 2003, pp. 648–651.

[6] J.A. Mrales-Cordovilla, Ning Ma, V. Sanchez, J.L. Carmona, A.M. Peinado, and J. Barker, "A pitch based noise estimation technique for robust speech recognition with missing data," in *IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, 2011, pp. 4808 – 4811.

[7] P. Loizou, *Speech enhacement: theory and practice*, CRC Press, 2007.

[8] J. Kominek and A. W. Black, "Cmu arctic databases for speech synthesis," Tech. Rep., CMU, 2003.

[9] A. Varga, H. Steeneken, M. Tomlinson, and D. Jones, "The noisex–92 study on the effect of additive noise on automatic speech recognition," *Technical Report, DRA Speech Research Unit*, 1992.

[10] I. Cohen, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Processing Letters*, vol. 9, no. 1, pp. 12–15, 2002.

[11] S. Rangachari and P. Loizou, "A noise estimation algorithm for highly nonstationary environments," *Speech Communication*, vol. 28, no. 220-231, 2006.

[12] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, pp. 443–445, 1985.

[13] Y. Hu and P. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008.

[14] M. Berouti, M. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP '79)*, 1979, pp. 208–211.

[15] Y. Hu and P. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 334-341, 2003.