Noisy Speech Enhancement Based on Long Term Harmonic Model to Improve Speech Intelligibility for Hearing Impaired Listeners

Dongmei Wang, Philipos C. Loizou, John H. L. Hansen

Dept. Electrical Engineering, University of Texas at Dallas
800 West Campbell Rd, Richardson, Tx 75080, USA
{dongmei.wang, john.hansen}@utdallas.edu

10.21437/Interspeech.2014-468

Abstract

This study proposes a speech enhancement algorithm to improve speech intelligibility for hearing impaired listeners in adverse conditions. The proposed algorithm is based on a long term harmonic model, where the harmonics of target speech are more distinguished from noise spectrum interference. Our method consists of two stages: i) Prominent pitch estimation based on long term harmonic feature analysis and neural network classification. ii) Target speech spectrum estimation with pitch information based on long term noise spectrum extraction. The listening experiment with EAS vocoder speech shows that our algorithm is substantially beneficial for cochlear implant recipients to perceive speech in noisy environment in terms of word recognition rate.

Index Terms: speech enhancement, long term harmonic model, long term noise spectrum, speech intelligibility improvement,

1. Introduction

Understanding speech in noisy background is still difficult for hearing impaired listeners with hearing aids and cochlear implants. Therefore designing effective speech enhancement algorithm that could improve the intelligibility for the hearing impaired community is of great importance. Considering the limitations of electronic integration of hearing aids or cochlear implant devices, single channel speech enhancement algorithms are more desirable than multichannel methods.

Numerous single channel speech enhancement techniques have been developed over the past several decades [1]. Some of them assume specific statistical properties of noise. For example, minimum mean square error (MMSE) algorithm [2] is effective for stationary noise where the noise characteristics can be estimated from the silence/pause sections between speech parts. However, MMSE is not well suited for non-stationary noises such as babble noise seen in daily life. Furthermore, preserving the harmonics for the spectrum after MMSE processing is investigated, and substantial improvement of speech intelligibility is obtained over conventional MMSE processing [3]. Codebook methods that train a model of the harmonic structure were also proposed for more accurate harmonic estimation for speech enhancement [4]. Nevertheless, codebook techniques are more suitable for the speaker-dependent scenarios since it is difficult to train a general harmonic structure model. Spectral weighting function is derived by constrained optimization to suppress noise in the frequency domain [5]. In addition, a number of studies have been focusing on speech segregation, inspired by research on auditory scene analysis (ASA) [6, 7, 8].

In this study, we propose a speech enhancement algorithm based on a long term harmonic model. Since the long term speech signal has higher frequency resolution, the target speech and noise are more separable from each other. First we estimate the pitch for the target speech based on long term harmonic model [9] and neural network classifier. The estimated pitch is further used to estimate the target harmonics and noise component in the long term spectrum. The long term noise spectrum is converted into short term one for the short term spectrum gain function estimation. Note that the noise spectrum in unvoiced parts can be used as prior information to estimate the noise spectrum in the voiced parts. Finally, we smooth the gain function for the target speech with previous frame information.

The rest of this paper is organized as follows. Sec. 2 describes pitch estimation algorithm. Sec. 3 presents the target spectrum estimation. In Sec. 4, we evaluate the performance of the proposed method. Finally, Sec. 5 discusses the results and conclusion.

2. Pitch Estimation

Pitch estimation is performed as the first stage of the algorithm. It is based on long term harmonic feature analysis combined with neural network classification. Our algorithm overview is shown in Fig. 1 and described in detail in the following subsections.

2.1. Pitch candidates extraction

We propose to extract the pitch candidates from the long term spectrum which has a better harmonic resolution than a short term signal. When the analyzing frame is longer, the main lobe of spectrum peaks will be narrower. This indicates that a long term spectrum analysis is able to alleviate the spectrum interference between speech and noise. In this way pitch estimation can be performed more accurately.

After spectrum analysis, we perform spectrum compression and summation (*CS*). A function is defined to calculate *CS* as follows,

$$P_{\hat{r}}(\omega) = \sum_{k=1}^{K} \left| S_{\hat{r}}(k\omega) \right|^2 \quad (1)$$

where $S_{\hat{r}}(\omega)$ is the original noisy speech spectrum, and $P_{\hat{r}}(\omega)$ is obtained as the *CS* spectrum. The motivation of generating the *CS* spectrum is that compressing the spectrum along frequency axis by integer factors will cause multiple harmonics to coincide enforcement at the fundamental frequency position.

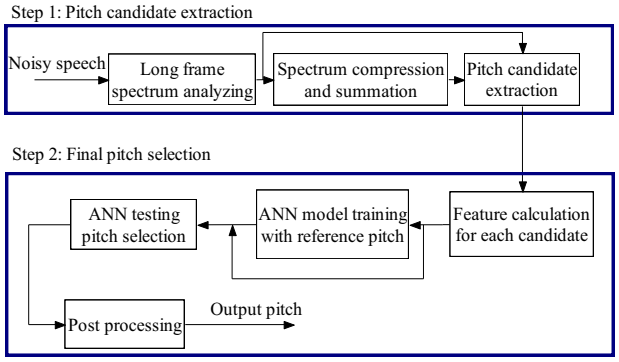


Fig. 1. Overall F0 algorithm estimation process

Both the *CS* spectrum and original spectrum will be used for pitch candidate extraction to reduce the true pitch missing rate. First, in the original spectrum, the peaks which exceed 10% of the maximum peak are selected. Second, in the *CS* spectrum, the peaks exceeding 50% of maximum value are selected as the additional pitch candidates. The frequencies of the selected peaks are considered as pitch candidates. Meanwhile, we set the pitch range from 50Hz to 400Hz.

2.2. Target pitch selection

We propose to select the target pitch from the pitch candidates list based on the long term harmonic features analysis and a neural network classification. Five harmonic features (illustrated in Table 1) are computed for each F0 candidate. It is noted from Table 1 that each feature is related to the harmonic structure. For clean speech, each of features for pitch examination can be used independently. However, for noisy speech, it is expected that a combination of these features will provide more accurate estimation. As each of these features may be affected by the noise, differently and hence, these effect can partially compensate for each other in feature fusion. The contribution weight of each feature for candidate selection should be estimated corporately. In our case, the ANN [10] is trained to model the relation between the input harmonic features and the ground truth pitch value. Once the training is finished, ANN model is ready to perform target pitch selection for unseen data. The backpropagation algorithm is used to learn the weights for the ANN model.

The input of the ANN model is the harmonic feature vector [***Hd Er O2e Rh Rc***] (Table 1). The output of the ANN indicates the input pitch candidate as true or false. The output in the training phase is defined based on the difference between each candidate and the corresponding ground truth. If the difference is within 20% of the ground truth value, the output is set as true, otherwise false.

Considering the ANN structure, three layers of sigmoid units are sufficient to express a rich variety of target functions [10]. According to this, we use three layers with sigmoid activation functions and one input layer with linear activation function. We use the cross validation method to find the optimal number of hidden units in each layer. We set 10 linear units in the input layer, 6 sigmoid units in the first hidden layer, 5 sigmoid units in the second hidden layer, and one output neuron.

After establishing the network structure, we feed the neural network with the prepared training data to obtain the weight values for each feature. Once the training step is completed we proceed with the testing step. During the testing step, in each frame, the candidate with the maximum output value is selected as the initial estimated target pitch.

Table 1 Long term harmonic feature

Hd	The average frequency deviation of the detected harmonics from the corresponding ideal harmonics.
Er	The ratio of the detected harmonic spectrum energy over the entire spectrum energy.
O2e	The energy ratio between odd harmonics and even harmonics. O2e can be used effectively to control the half-pitch error.
Rh	The ratio between the number of detected harmonics and total number of harmonic order.
Rc	The ratio of *CS* amplitude at the pitch candidate position and its maximum amplitude value.

2.3. Pitch tracking

Pitch tracking is performed to increase accuracy rate. The overview of pitch tracking algorithm is illustrated in Fig. 2. The objective of our algorithm is to first detect the major pitch frequency range (MPFR) from the initial estimated pitch, and then re-estimate the pitch within that detected frequency range. The MPFR of a particular utterance is obtained by calculating which frequency band has the maximum summation of the ANN output value. The re-estimated pitch is obtained by selecting the entry with the maximum ANN output value within the MPFR. Pitch tracking is performed on the re-estimated pitch based on continuity in time. The average ANN output value is computed for each pitch track, and the one with maximum average output will be chosen as the final estimated pitch.

3. Target Spectrum Estimation

In this section we estimate the target spectrum. The voiced parts and unvoiced parts are processed in different ways. In voiced parts the target spectrum is estimated by obtaining the gain function. In the unvoiced parts, noise spectrum is estimated with Gaussian filter and then subtracted from noisy spectrum to obtain the target speech.

3.1. Voice part processing

We define $|\hat{S}(\omega)| = G(\omega) \cdot |X(\omega)|$, where $|X(\omega)|$ is the observed noisy spectrum, $G(\omega)$ is the gain function, and $|\hat{S}(\omega)|$ is the estimated target speech spectrum. We assume $|X(\omega)| = |S(\omega)| + |N(\omega)|$, where $|S(\omega)|$ and $|N(\omega)|$ are the target and noise spectrum respectively. From above deduction, the gain function can be denoted as $\hat{G}(\omega) = 1 - |\hat{N}(\omega)| / |X(\omega)|$. Therefore, we need to estimate the noise spectrum in order to get the gain function.

Our noise estimation algorithm is also inspired by the fact that the spectrum interference between target speech and noise is milder in long term signal than the short one. Consequently, noise spectrum is able to be estimated more accurately from the long term signal. One might be wondering why we do not estimate target speech from the long term spectrum directly. The reason is that some fast variation of the speech component could be lost due to the long term average. Accordingly, the

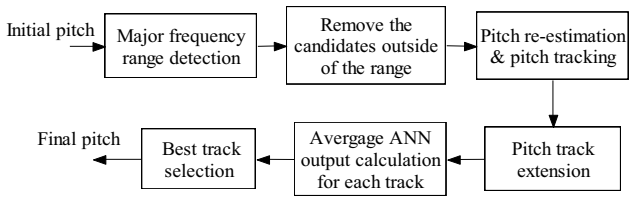


Fig. 2 Overview of pitch post-processing

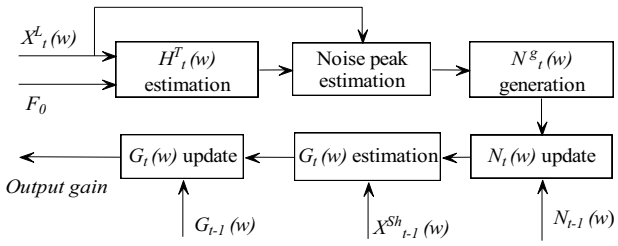


Fig. 3 Algorithm overview of gain function estimation

speech intelligibility would be affected. Nevertheless, this concern regarding long term noise estimation is less important as we do not seek to estimate the details for noise. The detailed algorithm is presented in Fig. 3 and described below.

i) Long term harmonic estimation. Firstly, spectrum peaks are extracted from the long term noisy spectrum $X_L(\omega)$. The target harmonics are estimated by selecting the spectrum peak which locates closest to the corresponding ideal harmonic frequency nF0.

ii) Long term noise spectrum peak estimation. Noise spectrum peaks are obtained by removing the estimated harmonics from long term spectrum peaks.

iii) Short term noise spectrum ($|N^g_t(\omega)|$) generation. The long term noise spectrum peaks from step ii) are used to form $|N^g_t(\omega)|$. In our case hamming window is used to analyze the time domain signal. Thus, $|N^g_t(\omega)|$ is generated by multiplying the estimated noise spectrum peaks with the normalized hamming window spectrum. In addition, the amplitude of $|N^g_t(\omega)|$ is adjusted by multiplying the length ratio between the short term and the long term.

iv) Noise spectrum updating. Considering the fact that noise activity changes are continuous between neighboring frames, we do the smoothing filtering on noise spectrum with first-order recursive filter:

$$|\hat{N}_t(\omega)| = \alpha|N^g_t(\omega)| + (1-\alpha)|\hat{N}_{t-1}(\omega)| \quad (2)$$

where α is the smoothing coefficient, and $|\hat{N}_{t-1}(\omega)|$ is the spectrum estimated in the previous frame. In the first frame of each voiced segment, $|\hat{N}_{t-1}(\omega)|$ is obtained by calculating the average noise in the previous unvoiced segment.

v) **Gain function estimation**. The gain function is initially estimated as: $\hat{G}_t(\omega) = 1 - |\hat{N}_t(\omega)| / |X_t(\omega)|$.

And subsequently updated with the average factor β:

$$\hat{G}_t(\omega) = \beta\hat{G}_t(\omega) + (1-\beta)\hat{G}_{t-1}(\omega) \quad (3)$$

3.2. Unvoiced part processing

It is known that in unvoiced parts of noisy speech, the low frequency spectrum is usually dominated by noise. In contrast, the higher frequency spectrum is dominated by speech. Here we consider performing stronger suppression in lower frequencies compared to higher frequencies. The noise spectrum is estimated by low pass filtering noisy spectrum with a 2-dimention Gaussian filter. We set the filter matrix size (*Hsize*) differently between low frequency and high frequency (3*3 for low frequency, and 5*5 for high frequency). Then we subtract the estimated noise spectrum from the original spectrum. After this, pre-emphasis is performed to make the high frequency spectral component more prominent.

4. Experiments and Results

In order to evaluate the entire algorithm, we demonstrate the results for both pitch estimation and speech intelligibility. IEEE sentence database [1] is used for both evaluations. Three types of noise are used for simulating the noisy condition, including 4-talker babble, 9-talker babble, and speech-shaped noise.

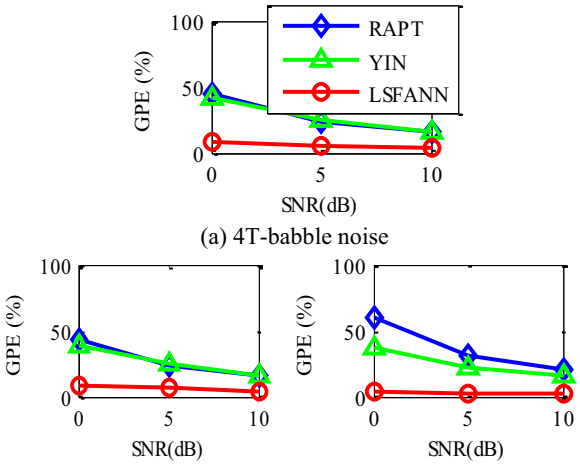
4.1. Pitch estimation results

Keele database [11] is used for the ANN model training. It provides ground truth pitch labels, which can be used as reference in the training session. The noisy speech in the training phase is obtained by adding the 9-talker babble noise into clean speech. IEEE sentences are used for the testing. The reference pitch of clean IEEE sentence is obtained with pitch estimation tool wavesurfer [12] plus manual correction. We present the global pitch error (GPE) [13] in Fig. 4. From Fig. 4 we see the proposed algorithm outperforms other two compared approaches, YIN [14] and RAPT [15] in all conditions. Low GPE results will ensure more accurate spectrum estimation in the next step.

4.2. Speech intelligibility results

Listening test is conducted to evaluate the performance of the proposed speech enhancement algorithm. The electrical and acoustical simulation (EAS) processing is used to simulate the bi-model cochlear implant system for the situation when the patients have residual hearing in low frequency. Seven normal hearing subjects are asked to dictate the EAS vocoder speech. All of the subjects are graduate school students at the University of Texas at Dallas, and receive financial compensation for participating in the experiment. The listening test is carried out in the double-walled sound attention booth (Acoustic System, Inc.). Before the formal listening test, there is a training session to get the subjects familiar with the EAS vocoder speech.

The performance is measured in terms of word recognition rate (WRR). The results are presented in Table 2. It can be seen that the WRR rate is improved by our algorithm in non-stationary noise condition (babble noise). The improvement is larger in low SNR level than high SNR level. For speech-shaped noise, neither the proposed method nor the MMSE algorithm brings any improvement. The results indicate our method is effective in babble noise, which is



(a) 4T-babble noise

(b) 9T-babble noise (c) Speech-shaped noise

Fig. 4 GPE results

Table 2 word recognition rate

SNR	Original	MMSE	Harmonic	Max WRR +
4T-babble noise				
0dB	24.6%	25.4%	39.2%	14.6%
5dB	41.9%	51.2%	60.9%	19%
10dB	62.8%	59.2%	69.8%	7%
9T-babble noise				
0dB	22.5%	25.8%	35.7%	13.2%
5dB	55.3%	53.4%	64.2%	8.9%
10dB	73.5%	68.1%	67.5%	-
speech-shaped noise				
0dB	61.2%	53.7%	58.0%	-
5dB	87.5%	80.0%	75.2%	-

contributed from suppression of the background harmonics makes target harmonics prominent.

Furthermore, the outperformance of our approach over MMSE is attributed to two aspects: i) We use cues from both target speech (pitch, harmonics) and noise (long term noise spectrum) in our algorithm. However MMSE only track the noise spectrum without using the cues of speech. ii) The more accurate harmonics estimation of our method is beneficial for patients who have residual hearing at low frequency [3].

5. Conclusion and Discussion

In this paper, we propose a speech enhancement algorithm to improve speech intelligibility for hearing impaired listeners. Our algorithm consists of two steps: pitch estimation and target spectrum estimation. The long term harmonics analysis was applied to achieve extracting more robust harmonic features for pitch estimation. In addition, artificial neural network is utilized as a practical model to estimate the weight for each harmonic feature. Long term noise spectrum estimation is performed with the estimated pitch information based on harmonic model. Subsequently it is matched into short term spectrum. The gain function for target spectrum is obtained with the smoothing filter technique. The listening test with normal hearing listeners on EAS vocoder speech shows that our algorithm can effectively improve the speech intelligibility in non-stationary noise conditions.

In our future work, we will focus on improving speech intelligibility in different types of noise.

6. Acknowledgement

Research supported by Grant No. R01 DC010494 from NIDCD/NIH1

7. References

[1] P. C. Loizou, Speech Enhancement: Theory and Practice. Boca Raton, FL: CRC, 2007.

[2] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean square error log-spectral amplitude estimator,” IEEE Trans. Acoust., Speech, Signal Process., vol. ASSP-33, no. 2, pp. 443-445, Apr. 1985.

[3] Y. Hu, P. C. Loizou, “On the importance of preserving the harmonics and neighboring partials prior to vocoder processing: Implications for cochlear implants,” J. Acoust. Soc. Am., vol. 127, no. 1, pp. 427-433, Jan. 2010.

[4] E. Zavarehei, S. Vaseghi, Q. Yan, “Noisy speech enhancement using harmonic-noise model and codebook-based post-processing,” IEEE Trans. Acoust., Speech, Signal Process., vol. 15, no. 4, pp. 1194-1203, May. 2007.

[5] W. Jin, X. Liu, M. S. Scordilis, L. Han, “Speech enhancement using harmonic emphasis and adaptive comb filtering,” IEEE Trans. Acoust., Speech, Signal Process., vol. 18, no. 2, pp. 356-368, Feb. 2010.

[6] G. Hu and D. L.Wang, “Monaural speech segregation based on pitch tracking and amplitude modulation,” IEEE Trans. Neural Network., vol. 15, no. 5, pp. 1135-1150, 2004.

[7] G. Hu and D. L. Wang, “A tandem algorithm for pitch estimation and voiced speech segregation,” IEEE Trans. Acoust., Speech, Signal Process., vol. 18, no. 8, pp. 2067-2079, Nov. 2010.

[8] A. S. Bregman, Auditory Scene Analysis. Cambridge, MA: MIT Press, 1990.

[9] Q. Huang, D. Wang, “Single-channel speech separation based on long-short frame associated harmonic model,” Digital Signal Processing, vol. 21, no. 4, July, 2011.

[10] T. M. Mitchell, Machine Learning, MIT Press and McGraw-Hill, 1997.

[11] F. Plante, G. Meyer, and W. A. Ainsworth, “A pitch extraction reference database,” in Proc. Eurospeech, 1995, pp. 837–840.

[12] http://www.speech.kth.se/wavesurfer/man.html

[13] W. Chu, A. Alwan, “SAFE: a statistical approach to F0 estimation under clean and noisy conditions,” IEEE Trans. Audio. Speech. Lang. Process., vol. 20, no. 3, pp. 933-944, Mar., 2012.

[14] A. Cheveigne, “YIN, a fundamental frequency estimator for speech and music,” J. Acoust. Soc. Am., vol. 111, no. 4, Apr., 2002.

[15] D. Talkin, “Robust algorithm for pitch tracking,” Speech Coding Synth, pp. 497-518, 1995.