Line Spectral Frequency-based Noise Suppression for Speech-Centric Interface of Smart Devices

Gil-Jin JANG[1], Jeong-Sik PARK[2], Ji-Hwan KIM[3], Yong-Ho SEO[4]
[1]*Ulsan National Institute of Science and Technology (UNIST), Ulsan, 689-798, South Korea*
[2]*Korea Advanced Institute of Science and Technology (KAIST), Daejeon, 305-701, South Korea*
[3]*Sogang University, Seoul, 121-742, South Korea*
[4]*Mokwon University, Daejeon, South Korea*
parkjs@kaist.ac.kr

***Abstract*—This paper proposes a noise suppression technique for speech-centric interface of various smart devices. The proposed method estimates noise spectral magnitudes from line spectral frequencies (LSFs), using the observation that adjacent LSFs correspond to peak frequencies of spectrum, whereas isolated LSFs are close to flattened valley frequencies retaining noise components. Over a course of segmented time frames, the logarithms of spectral magnitudes at respective LSFs are computed, and their distribution is then modeled by the Rayleigh probability density function. The standard deviation from the Rayleigh function approximates the noise spectral magnitude. The model is updated at every frame in an online manner so that it can deal with real-time inputs. Once the noise spectral magnitude is estimated, a time-domain Wiener filter is derived for the suppression of the estimated noise spectral magnitude, and this is then applied to the input noisy speech signals. Our proposed approach operates well on most smart devices owing to its low computational complexity and real-time implementation. Speech recognition experiments, conducted to evaluate the proposed technique, show that our method exhibits superior performance, with less distortion of original speech, when compared to conventional noise suppression techniques.**

***Index Terms*—noise measurement, noise reduction, speech enhancement, speech recognition, linear predictive coding.**

I. INTRODUCTION

Human speech provides a natural and intuitive interface for interaction with machines. As technical breakthroughs in the field of speech technology have been achieved, speech-centric interfaces are starting to be adopted in the overwhelming majority of smart devices, such as smart phones and tablet personal computer, and even in vehicles and aircraft [1][2].

Automatic speech recognition plays a major role in the speech-centric interface. Even though speech recognition helps to control devices with ease, it is obvious that the recognition performance is easily degraded in adverse conditions, particularly due to contamination by environmental noise. In practical situations, a great variety of background noises introduce intelligibility loss while modifying acoustic characteristics of the original speech [3]. For this reason, many researchers have made tremendous efforts in suppressing the noise in contaminated speech signals. Fig. 1 shows a simple illustration of speech-centric interaction with various smart devices, along with noise suppression. There is no doubt that noise-suppressed speech positively affects the performance of various speech processing tasks, leading to a more reliable user interface.

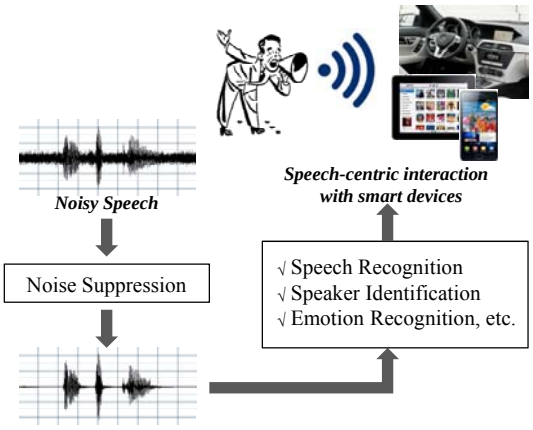

Figure 1. Noise suppression and its positive effect on the speech-centric interface of smart devices

In order to directly implement noise suppression features in these devices, the hardware limitations of the devices and the feasibility of real-time implementation should be carefully considered. This paper proposes a novel approach that requires lower computational complexity and memory capacity than conventional noise suppression techniques, while maintaining the suppression performance and real-time implementation. In particular, the proposed method directly utilizes line spectral frequencies (LSFs), which are equivalent to linear predictive coding (LPC) parameters and are thus capable of being effectively embedded in the LP-based voice coders that are adopted in most audio devices.

The remainder of this paper is organized as follows. Section II reviews several previous research outcomes related to noise suppression. In Section III, the properties of LSFs are illustrated and the detailed procedures of the proposed method are described. Section IV explains the setup and results of speech recognition experiments conducted to evaluate the proposed technique. Finally, Section V presents our conclusions.

Gil-Jin Jang was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (No. 2010-0025642). Ji-Hwan Kim was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2011-0005419). Corresponding author is Jeong-Sik Park.

Digital Object Identifier 10.4316/AECE.2011.04001

II. Conventional Approaches for Noise Suppression

Numerous techniques have been proposed to suppress background noise, mostly based on the spectral subtraction method first introduced by Boll in 1979 [4]. Spectral subtraction is very simple, yet powerful enough to be widely adopted in most speech processing applications [5]. The main concept of this technique is straightforward: by assuming that background noise is additive to the clean speech signal and changes slowly over time, the spectral magnitudes of the noise are approximated by the average spectral magnitudes estimated during non-speech periods, and then the estimated magnitudes are uniformly subtracted across the spectrum of noisy speech during speech periods. Hence, its performance is closely related to how accurately the non-speech periods are detected as well as how reliably the noise magnitudes are estimated. The detection of speech and non-speech periods is carried out by a voice activity detector (VAD). The VAD enables noise estimates to be updated whenever non-speech periods are detected, but the performance of the VAD varies significantly according to the types and conditions of noise. Recently, a quantile-based noise estimation method, which does not require the VAD, has been investigated [6]. This approach arranges a set of spectral magnitudes estimated during certain periods in ascending order, and finds the noise spectral magnitude based on the value at a given quantile. Usually, the 50% quantile is used, which means that the noise is generally regarded as the median of the input spectra.

The conventional approaches mentioned above have clear drawbacks when applied to smart devices. Those based on spectral subtraction require a sufficient number of speech streams in order to detect non-speech periods, and their real-time implementation is therefore fraught with difficulties. The principal downside of the quantile-based approach is the additional memory requirement and computational overhead to keep a sufficient amount of past input data in order. As a result, the use of these approaches is problematic in real-world devices.

This paper proposes a new paradigm for estimating the noise spectral magnitudes without the use of a VAD and computationally intensive algorithms.

III. Noise Suppression based on Line Spectral Frequencies

In wireless telecommunication applications, acoustic signals are converted to compressed digital forms and transmitted over the air. Most conventional voice coders adopt a framework of LPC derived from a human speech production model [7]. Such LP-based coders are required to yield maximized compression under permissible intelligibility loss to make efficient use of limited bandwidth. LSFs are an alternative representation of LPC coefficients, and successfully satisfy such requirements [8].

From the basic LSF derivation formulae, it is observed that the peak frequencies of LPC spectra are found near the adjacently located LSFs, whereas relatively flattened valley frequencies are located around the isolated LSFs [9][10]. In other words, the spectral magnitudes at LSFs are considered to be representative of the peaks and valleys of the corresponding LPC spectra. In consideration of this property, LSFs can support reliable noise estimation, since noise components generally exist around the valley frequencies of LPC spectra while acoustic characteristics of speech signals are exhibited in the peak frequencies. Especially, LSFs are obtained from every short duration speech frame, thus enabling real-time noise suppression. Based on these observations, this paper proposes a new approach to noise suppression utilizing LSFs.

A. Properties of LPC Spectra at LSFs

The proposed method makes use of the properties of LPC analysis. The input speech signal is decomposed into spectral envelope and excitation signal, such that

$$x[n] = \sum_{k=1}^{P} a_k \cdot x[n-k] + G \cdot e[n], \tag{1}$$

where n is a digitized sample index, $x[n]$ is the sampled input speech, a_k are the prediction filter coefficients of order P, $e[n]$ is the excitation signal, and G is a scalar gain so that the excitation signal has unit variance.

Equation (1) is equivalently described in the frequency domain as

$$X(z) = G \cdot E(z) / A(z), \tag{2}$$

where $X(z)$ and $E(z)$ are impulse responses of x and e, and a LP polynomial $A(z)$ satisfies $A(z) = 1 - \sum_{k=1}^{P} a_k z^{-k}$. $E(z)$ is spectrally flattened, whereas $A(z)$ represents the spectral envelope of a given input speech frame.

For the transmission purpose of a voice coder, $A(z)$ is expressed by the following two reciprocal polynomials [8]:

$$\begin{aligned} P(z) &= A(z) + z^{-(P+1)} A(z^{-1}), \\ Q(z) &= A(z) - z^{-(P+1)} A(z^{-1}). \end{aligned} \tag{3}$$

The roots of these two auxiliary polynomials are called line spectral frequencies, and are known to be most efficient in coding LPC coefficients due to their stability and insensitivity to quantization error [7][11].

As LSFs are the roots of $P(z)$ and $Q(z)$, both of which are monotonic between any pair of neighboring LSFs, $A(z)$ is close to their local minima [8]. Fig. 2 illustrates the behavior of $A(z)$ at given LSFs. The two dotted lines in the figure are the frequency responses of $|P(z)|$ and $|Q(z)|$, the black solid line is the magnitude of the LP filter response expressed by $|A(z)| = 0.5|P(z) + Q(z)|$, and the lightly colored line is the spectral envelope approximated by $|1/A(z)|$. A pre-emphasis filter, $1 - 0.97z^{-1}$, has been applied to boost high frequency energies. Downward triangles are drawn on $|A(z)|$ and upward triangles are drawn on $|1/A(z)|$ at the root frequencies of $P(z)$ and $Q(z)$. When a single LSF is adjacent to its neighbors, e.g., at around 0.5 kHz in Fig. 2, $|A(z)|$ decreases and hence becomes more resonant around the corresponding frequencies [9][10]. In contrast, if an LSF is isolated, being separated from its neighbors, $|P(z)|$ and $|Q(z)|$ change slowly so that $|A(z)|$ becomes relatively flattened. Such

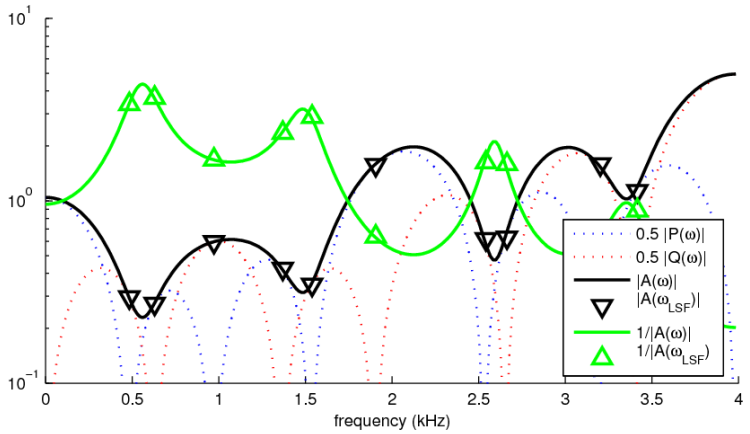

Figure 2. Properties of a LPC spectrum at line spectral frequencies.

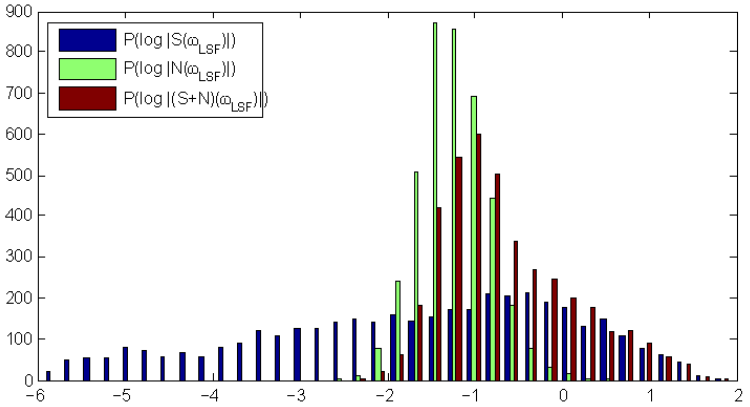

Figure 3. Distribution of log of LPC spectrum at LSFs multiplied by excitation gain. The log spectral magnitudes are chosen at LSFs only.

properties confirm that adjacently located LSFs correspond to the formant frequencies of speech signals in the LPC spectra and provide spectral magnitudes of speech. In contrast, the spectral characteristics of background noise are given around the isolated LSFs. We exploit these properties of LSFs for noise suppression.

B. Distribution of Spectral Magnitudes at LSFs

From the definition of a discrete Fourier transform, the impulse response of an LP polynomial at frequency ω is described by

$$A(\omega) = \sum_{k=0}^{P} a_k \exp(-j\omega k), \tag{4}$$

where a_0 = 1. Let us denote as ω_{it} the i th LSF of frame t. The smoothed spectral magnitude at ω_{it} is approximated by multiplying its LPC spectral magnitude $|A_t(\omega_{it})|$ by the scalar gain defined in (2), as follows:

$$|X_t(\omega_{it})| \approx G_t / |A_t(\omega_{it})|, \tag{5}$$

where G_t is the frame-dependent gain at frame t, which represents relative magnitude differences across time frames.

In order to use a single noise estimate regardless of frequencies, the LP spectral envelope is globally whitened by multiplying it with the long-term average of $A(z)$. It is then approximated in the logarithmic domain as

$$\begin{aligned} \log|Y_t(\omega_{it})| &= \log|X_t(\omega_{it}) \cdot \overline{A}_t(\omega_{it})| \\ &\approx \log G_t - \log|A_t(\omega_{it})| + \log|\overline{A}_t(\omega_{it})|, \end{aligned} \tag{6}$$

where the long-term average frequency response of LPC filters, $\overline{A}(z)$, is instantaneously updated by

$$\overline{A}_t(z) = (1-\alpha) \cdot \overline{A}_{t-1}(z) + \alpha \cdot A_t(z), \tag{7}$$

with an initial value of $\overline{A}_0(z) = 1$. According to experiments, the adaptation rate α performs best at a value of $\alpha = 0.02$.

The distribution of the log spectral magnitudes at LSFs is shown in Fig. 3. The sources used are male speech and factory noise from the signal processing information base (SPIB) database, which is available at http://spib.rice.edu/. The *x*-axis is quantized into histogram intervals from the log spectral magnitudes, and the *y*-axis is the number of frames whose *x*-value is in the corresponding interval.

The $S(z)$ are male speech and the noise spectra $N(z)$ are from factory noise signals. These two signals are added together to obtain the mixed frequency response $S(z) + N(z)$. Since the spectral energy of the factory noise is relatively stationary over time, there is a significant peak between -2 and -1 on the *x*-axis. The speech spectral magnitude is relatively scattered and varies considerably. The mixed distribution, expressed by lightly colored bars, has a peak around that of the noise distribution, and the portion of low energy components is significantly reduced. This is because the noise spectra which are consistent over time conceal the tiny spectral magnitudes of the speech signals.

C. *Noise Estimation and Suppression*

To find the noise estimate from the mixed distribution of spectral magnitudes at LSFs, we develop the following algorithm. We use $\log|Y_t(\omega_{it})|$ at the LSFs ω_{it} in (6) as the main input to the algorithm, and use the compact notation $y_{it} = \log|Y_t(\omega_{it})|$ in the following.

1. Compute the mean of the total data, $\theta = E[y_{it}]$.
2. Classify all of the input data samples as either noise, if they are smaller than the mean: $\{n_{it}\} = \{y_{it} \mid y_{it} < \theta\}$, or as speech otherwise: $\{s_{it}\} = \{y_{it} \mid y_{it} \geq \theta\}$.
3. Since the distribution of each half is single-sided, it can be modeled by the Rayleigh probability density function (pdf). The Rayleigh function is used to model the distribution of a nonnegative random variable. For noise, we can approximate the pdf by

$$p(n_{it}) = \frac{n_{it}}{\sigma_n^2} \exp\left[-\frac{(\theta - n_{it})^2}{2\sigma_n^2}\right], \tag{8}$$

where the maximum likelihood estimate of noise standard deviation is computed by the past inputs, yielding on-line implementation:

$$\sigma_n = \sqrt{\frac{1}{2N} \sum_{\text{previous } n_{it}} (\theta - n_{it})^2}. \tag{9}$$

Since all n_{it} are smaller than θ, the condition that $\theta - n_{it}$ is nonnegative is satisfied for the Rayleigh pdf. Similarly, we can approximate the speech pdf by

$$p(s_{it}) = \frac{s_{it}}{\sigma_s^2} \exp\left[-\frac{(s_{it} - \theta)^2}{2\sigma_s^2}\right], \tag{10}$$

where $\sigma_s = \sqrt{\frac{1}{2N} \sum_{\text{previous } s_{it}} (s_{it} - \theta)^2}$. Note that all s_{it} are larger than θ, so the condition that $s_{it} - \theta$ is nonnegative is satisfied as well.

4. Estimate noise spectral magnitudes at frame t by

$$|\tilde{N}_t(\omega)| = \exp[\theta - \sigma_n]. \tag{11}$$

We assume that the noise spectrum is the same over all frequencies, so the frequency parameter ω is left out.

5. A Wiener filter at frame t and frequency ω, suppressing the noise estimate from the spectral magnitude of noisy speech signal, is derived by

$$W_t(\omega) = \frac{|Y_t(\omega)|^2 - |\tilde{N}_t(\omega)|^2}{|Y_t(\omega)|^2} = 1 - \frac{|\tilde{N}_t(\omega)|^2}{|Y_t(\omega)|^2} \tag{12}$$

$$= 1 - \frac{1}{\exp[2(y_t(\omega) - \theta + \sigma_n)]}.$$

$W_t(\omega)$ is then converted to a minimum-phase time domain filter and applied to the noisy speech signal, $x[n]$ in (1), and overlap-added with trapezoidal windowing.

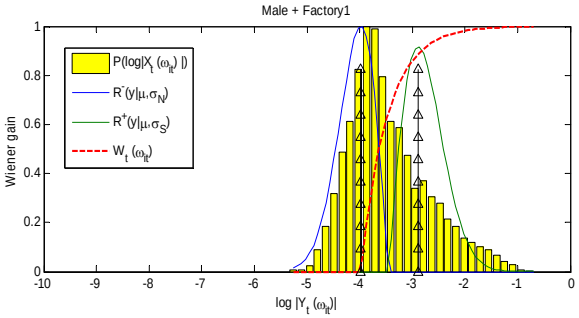

Figure 4. Noise spectral mean and Wiener filter estimation result by Rayleigh probability density functions for an additive mixture of male speech and factory noise.

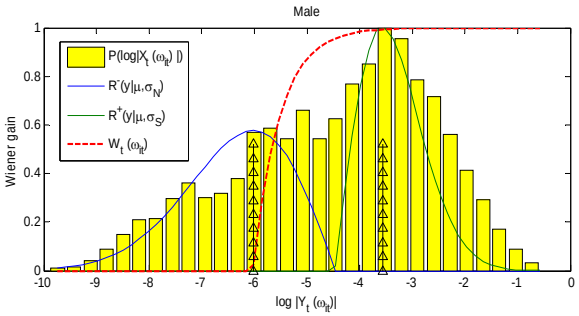

Figure 5. Noise spectral mean and Wiener filter estimation result by mixture of Rayleigh probability density functions for male speech only.

Fig. 4 illustrates the result of noise estimation and Wiener filter derivation for a mixture of male speech and factory noise. The lightly colored histogram bars approximate the probability distribution of the spectral magnitudes at LSFs in the log domain. The *x*-axis is $\log|Y_t(\omega)|$ in (6) at LSF i and frame t, and the *y*-axis shows both the normalized histogram and Wiener filter gains. The distribution of $\log|Y_t(\omega)|$ is displayed by histogram bars, and the estimated Rayleigh density functions are overdrawn on them by dashed curves. The leftmost curve is the Rayleigh pdf of noise spectral magnitudes, and the rightmost one is the pdf of speech spectral magnitudes. The Wiener filter $W_t(\omega)$ obtained by (12) is plotted by a thick, dashed line. The Wiener gain is zero when $\log|Y_t(\omega)|$ is smaller than the estimated noise. The distribution has a sharp peak around -4, which is well approximated by σ_n. A smaller peak is located around -3, approximated by σ_s. The spectral magnitudes of the speech signal vary much more than those of the noise, so its peak location is less distinct than the noise peak. The distribution in Fig. 5, illustrating male speech only, does not have a sharp peak, and the noise estimate is around -3 with much bigger variance. The noise mean estimate is shifted by about -2 when compared to Fig. 4, while the speech estimate in both figures is very close.

D. *Advantages of the Proposed Method in Smart Devices*

In order to directly implement the noise suppression features in consumer smart devices, two important requirements should be satisfied: the feasibility of real-time implementation and low computational complexity due to the hardware limitations of the devices.

In our approach, the distributions of noise spectral magnitudes and speech spectral magnitudes are respectively

modeled by Rayleigh pdfs for a given frame. The noise estimate is then obtained by the standard deviation parameter of the Rayleigh distribution of the lower half. Thus, our approach does not require computationally intensive tasks such as a VAD, which are necessary in most conventional noise suppression methods. In addition, the proposed method is operated in an online manner by adaptively updating the Rayleigh distribution parameters. For these reasons, the proposed approach allows real-time implementation and also enables the handling of real-time inputs.

The input in our method is the set of LPC coefficients, LSFs, and excitation gains. All of these are fundamental voice parameters estimated in LP-based voice coders. Hence, the proposed method can be efficiently integrated into the LP-based voice coders adopted by most audio devices, as additional parameter estimation is unnecessary, consequently yielding computational efficiency.

In contrast to the approaches in CDMA's enhanced variable rate codec (EVRC) and European Telecommunications Standards Institute (ETSI), both of which are commercial standards for noise suppression in wireless communication environments [12][13], the proposed method does not use fixed filter banks. Instead, variable filter banks are chosen according to the number of LSFs. In the case of an 8 kHz sampling frequency, the number of filter banks is 23 for EVRC and 16 for ETSI, but the number of LSFs is generally 10. These properties make our approach very efficient in terms of hardware usage. From simulation results on more than 20,000 test samples, the computation time of the proposed method is about ten times smaller than ETSI, due to the elimination of complicated modules such as VAD, the reduction from 16 filter bank energy computations to 10 discrete Fourier transforms at LSFs, and the help of optimized LPC and LSF computations.

IV. Experimental Results

A. Speech Recognition Experiments and Results

The efficiency of the proposed method is verified by a comparison with the conventional method, using automatic speech recognition experiments on the Speech Separation Challenge (SSC) database [14]. In SSC, speakers say sentences of exactly six words in the format "command–color–preposition–letter–number–adverb", such as "bin blue at F two now". The format is well matched to command-and-control situations in consumer electronics devices. The database has a training set of 17,000 utterances, spoken by 34 different speakers. All training files are recorded in a quiet environment without any background noise. The hidden Markov models (HMMs) are obtained by the hidden Markov model toolkit (HTK), as suggested by the coordinators of SSC [15]. The adopted features are 12 Mel-frequency cepstral coefficients (MFCCs) plus log energy, plus their velocities and accelerations, resulting in a 39-dimensional vector extracted at 10 ms intervals. A separate testing set of 600 utterances is also provided. There are no overlaps between the training and the test data.

The original recordings do not contain environmental noise. To evaluate the validity of the proposed method on various noise conditions, four different noise sources (airport, car, restaurant, and train) were chosen from the AURORA2 database [16] and added to clean test files. These four types are quite common in telecommunication scenarios. The simulated signal-to-noise ratios (SNRs) are 20, 16, 12, and 8 dBs. Deployment of a speech recognition system is not practical for SNRs lower than 8 dB due to the severe distortion of speech caused by the noise. For a fair evaluation, all HMMs are trained using the original clean speech.

For the performance comparison, we investigated the noise suppression front-end in the ETSI standard. The ETSI standard uses Mel-warped filter bank energies in voice activity detection and noise estimation. The source code is publicly available from the distributor.

The results of speech recognition experiments are summarized in Table I. Under 20 dB SNR condition, the proposed method was on a par with or slightly better than None (no processing), whereas the conventional ETSI was 1–2% worse than the others. In the other SNR conditions, the proposed method always exhibited a better recognition rate than ETSI and None by 1–5%. As the SNR became lower, so the improvement increased. In summary, the speech recognition results prove that the proposed method is quite stable, and much better than the conventional method with various noise types and various noise levels.

TABLE I. Comparison of speech recognition performance on the testing set with several types of noise from the AURORA2 database. Three noise suppression methods are applied: None (no processing), ETSI standard, and the LSF-based (proposed) method

Noise Type	Noise Suppression Methods	20dB	16dB	12dB	8dB
Airport	None	95.9%	93.3%	85.1%	73.1%
	ETSI	94.4%	92.7%	87.3%	76.3%
	LSF-based	96.2%	94.9%	89.6%	81.1%
Car	None	95.3%	90.3%	79.4%	64.3%
	ETSI	94.6%	91.2%	85.4%	71.8%
	LSF-based	95.2%	93.7%	88.3%	76.9%
Restaurant	None	94.1%	90.3%	83.3%	67.5%
	ETSI	93.1%	89.9%	84.2%	72.1%
	LSF-based	95.0%	92.3%	87.9%	77.4%
Train	None	95.6%	93.4%	86.9%	75.4%
	ETSI	94.4%	92.6%	88.2%	79.4%
	LSF-based	96.1%	94.6%	90.3%	83.1%

B. Spectral Analysis

Next, we analyzed the change in spectral figures after processing noise suppression. Fig. 6 and Fig. 7 represent the noise suppression results of male speech contaminated by factory noise and male speech only, respectively. The top panels show the spectrogram of the input signal, whose SNR is computed by the ratio of the male speech to the added noise. The middle panels display Wiener filter gains computed by the proposed method at each frequency and time. As the gain becomes larger, it is colored darker. The dark region follows the male speech spectrogram in the mixture well enough to keep most of the speech signal energies. The spectrogram of the final noise suppression result is shown in the bottom panels. Fig. 6 shows that the

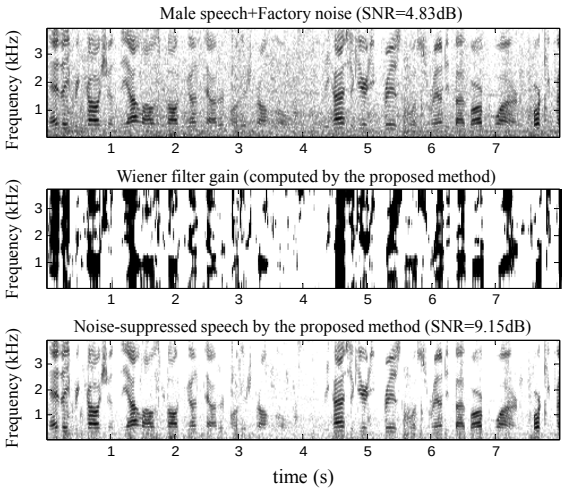

Figure 6. Noise suppression result of a mixture of male speech and factory noise. In all three images, the *x*-axis is time (s), and the *y*-axis is frequency (kHz).

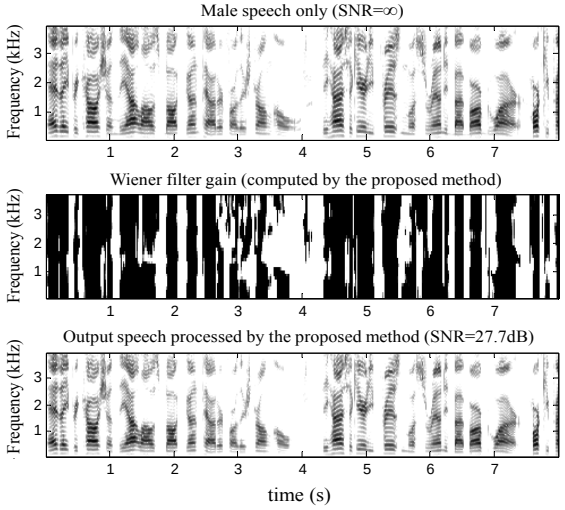

Figure 7. Noise suppression results of male speech only. The SNR of the original speech (male speech only, without additive noise) is ∞.

input SNR is 4.83 dB, but this increases to 9.15 dB after the suppression. In Fig. 7, most of the Wiener filter gains are close to unity, so the middle image is much darker than that of Fig. 6. The output SNR is 27.7 dB, implying that the input is almost clean.

V. Conclusions

This paper proposed a novel method to suppress background acoustic noise in smart devices. The proposed method estimated spectral information of noise in each frame based on line spectral frequencies, and used that information to suppress unwanted noise in the corresponding frame. The proposed method requires lower computational overhead and hardware resources than conventional noise suppression methods, and operates in real-time. To evaluate our method, we performed speech recognition experiments using noise-contaminated speech data. Experimental results indicated that the proposed approach considerably reduced the background noise under various SNR conditions and contributed to achieving superior recognition performance. This was proved by a comparison with the ETSI noise suppression standard, which is adopted in many distributed speech recognition applications.

For these reasons, it is strongly expected that the proposed method will operate efficiently in the speech-centric interfaces of a variety of smart devices, especially smart phones, smart pads, TVs, and other hands-free devices.

References

[1] M. Schuricht, Z. Davis, M. Hu, S. Prasad, P. Melliar-Smith, and L. Moser, "Managing multiple speech-enabled applications in a mobile handheld device," International Journal of Pervasive Computing and Communications, vol. 5, no. 3, pp. 332-359, Sep. 2009. [Online]. Available: http://dx.doi.org/10.1108/17427370910991884

[2] L. Deng, A. Acero, Y. Wang, K. Wang, H. Hon, et al., "A speech-centric perspective for human-computer interface," IEEE Workshop on Multimedia Signal Processing, pp. 263-267, Dec. 2002. [Online]. Available: http://dx.doi.org/10.1109/MMSP.2002.1203296

[3] K. Kim and M. Kim, "Robust speaker recognition against background noise in an enhanced multi-condition domain," IEEE Transactions on Consumer Electronics, vol. 56, no. 3, pp. 1684-1688, Aug. 2010. [Online]. Available: http://dx.doi.org/10.1109/TCE.2010.5606313

[4] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 27, no. 2, pp. 113-120, Apr. 1979. [Online]. Available: http://dx.doi.org/10.1109/TASSP.1979.1163209

[5] K. Wu and P. Chen, "Efficient speech enhancement using spectral subtraction for car hands-free applications," Proc. of International Conference on Consumer Electronics, pp. 220-221, Jun. 2001. [Online]. Available: http://dx.doi.org/10.1109/ICCE.2001.935283

[6] V. Stahl, A. Fischer, and R. Bippus, "Quantile based noise estimation for spectral subtraction and wiener filtering," Proc. of ICASSP, vol. 3, pp. 1875-1878, Jun. 2000. [Online]. Available: http://dx.doi.org/10.1109/ICASSP.2000.862122

[7] A. Kindoz and A. Kondoz, Digital speech; coding for low bit rate communication systems, John Wiley & Sons, Inc., New York, NY, USA, Jan. 1994.

[8] P. Kabal and R. Ramachandran, "The computation of line spectral frequencies using chebyshev polynomials," IEEE Transactions on Acoustics, Speech, Signal Processing, vol. 34, no. 6, pp. 1419-1426, Dec. 1986. [Online]. Available: http://dx.doi.org/10.1109/TASSP.1986.1164983

[9] M. Lee, H. Kim, S. Choi, and H. Lee, "On the use of LSF intermodel interlacing property for spectral quantization," Proc. of IEEE Workshop on Speech Coding, pp. 43-45, Jun. 1999. [Online]. Available: http://dx.doi.org/10.1109/SCFT.1999.781478

[10] M. Lee, H. Kim, and H. Lee, "A new distortion measure for spectral quantization based on the LSF intermodel interlacing property," Speech Communication, vol. 35, no. 3-4, pp. 191–202, Oct. 2001. [Online]. Available: http://dx.doi.org/10.1016/S0167-6393(00)00080-7

[11] T. Backstrom and C. Magi, "Properties of line spectrum pair polynomials—a review," Signal Processing, vol. 86, pp. 3286-3298, Nov. 2006. [Online]. Available: http://dx.doi.org/10.1016/j.sigpro.2006.01.010

[12] Telecommunications Industry Association (TIA), "Enhanced variable rate codec, speech service option 3 for wideband spread spectrum digital systems," Technical Report, TIA/EIA/IS-127-2, Dec. 1999.

[13] European Telecommunications Standards Institute, "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithm," Technical Report, ES 202 050 v1.1.5, Jan. 2007.

[14] M. Cooke, J. Hershey, and S. Rennie, "Monaural speech separation and recognition challenge," Computer Speech & Language, vol. 24, no. 1, pp. 1-15, 2010. [Online]. Available: http://dx.doi.org/10.1016/j.csl.2009.02.006

[15] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, et al., Hidden Markov model toolkit (HTK), ver. 3.4, Dec. 2006. [Online]. Available: http://htk.eng.cam.ac.uk

[16] D. Pearce and H. Hirsch, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy condition," Proc. of ICSLP, Oct. 2000. [Online]. Available: http://aurora.hsnr.de