한국음향학회지 제31권 제3호 pp. 179~187 (2012)
The Journal of the Acoustical Society of Korea Vol.31, No.3 (2012)

Noise Spectrum Estimation Using Line Spectral Frequencies for Robust Speech Recognition

Gil-Jin Jang, Jeong-Sik Park*, and Sanghun Kim**

Ulsan National Institute of Science and Technology (UNIST)
*Mokwon University, Daejeon, South Korea
**Speech and Language Information Research Department Electronics and Telecommunications Research Institute

(Received May 20, 2011; revised February 14, 2012; accepted February 14, 2012)

ABSTRACT: This paper presents a novel method for estimating reliable noise spectral magnitude for acoustic background noise suppression where only a single microphone recording is available. The proposed method finds noise estimates from spectral magnitudes measured at line spectral frequencies (LSFs), under the observation that adjacent LSFs are near the peak frequencies and isolated LSFs are close to the relatively flattened valleys of LPC spectra. The parameters used in the proposed method are LPC coefficients, their corresponding LSFs, and the gain of LPC residual signals, so it suits well to LPC-based speech coders.
Key words: Line spectral frequencies (LSF), Noise suppression, Speech recognition
ASK subject classification: Speech Signal Processing (2.3)(2.5)

I. Introduction

The assumption of spectral subtraction [1] is that the noise is additive and changes slowly over time, so that noise spectrum should be approximated by an average spectrum in non-voice period. The error in estimating true noise spectrum directly accounts for either voice attenuation or less noise suppression, hence the performance is closely related to how reliable the noise spectrum estimates are. Most conventional methods rely on detecting whether the instantaneous input frame contains speech, called voice activity detector (VAD), which then enables updating noise estimates when background noise is present only. However, the performance of the VAD varies a lot according to various noise conditions.

This paper proposes a novel procedure for noise spectral magnitude estimation which also eliminates the use of VAD in a very efficient manner. From the basic LSF derivation formulae it is observed that the local maxima of LPC spectra are near the adjacently located LSFs [2,3], and relatively flattened valleys across frequency are around the isolated LSFs. In the proposed method the spectral magnitudes at LSFs are considered as representatives of the peaks and valleys of the corresponding LPC spectra, and participate in estimating noise spectral magnitude. Without any consideration of determining if the current analysis frame contains noise only, the distribution of the log spectral magnitudes at LSFs are modeled by mixture of dual Gaussian probability density functions. The Gaussian with smaller mean is then taken as noise distribution, so the mean is adopted as a noise spectral estimate. An online adaptation algorithm for the parameters of Gaussian distributions is also proposed so that it can handle real-time inputs. The noise Gaussian mean is updated at every time frame.

***Corresponding author:** Jeong-Sik Park (parkjs@kaist.ac.kr)
Mokwon University, 800 Doan-dong, Seo-gu, Daejeon 302-729, South Korea
(Tel: +82-42-350-7716)

A time-domain Wiener filter suppressing the estimated amount of noise spectral magnitude is computed for every time frame and frequency band, and applied to the input speech signal. The required parameters are LPC coefficients, LSFs, and excitation gains, which are all available in most LP vocoders. Therefore the proposed method can be easily integrated into LP vocoders with much less additional overhead than the other conventional noise suppression methods.

To assess the validity of the proposed method, automatic speech recognition experiments are carried out on *speech separation challenge* database. Results show significant improvement in speech recognition rates with relatively less speech distortion when compared to ETSI frontend and TIA's EVRC standard noise suppression.

II. Noise Spectrum Estimation

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

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

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 $e[n]$ has unit variance. Equation 1 is equivalently expressed in the frequency domain as

$$X(z) = \frac{G}{A(z)} E(z) = \frac{G}{1 - \sum_{k=1}^{P} a_k z^{-k}} E(z), \tag{2}$$

where $X(z)$ and $E(z)$ are z-transforms of x and e. $E(z)$ is spectrally flattened, so that $A(z)$ should contain most of the spectral envelope of the given input speech frame. For a transmission purpose, $A(z)$ is expressed by the two reciprocal polynomials [4]:

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

The roots of these two auxiliary polynomials are called line spectral frequencies (LSFs), and known to be most efficient in coding LPC coefficients due to its stability and little sensitivity to quantization error [5,6].

At LSFs either $P(z)$ or $Q(z)$ is zero, so $A(z)$ is close to its local minima since $P(z)$ and $Q(z)$ are monotonic between any pair of neighboring LSFs. Figure 1 illustrates the behavior of $A(z)$ at 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 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 on $1/|A(z)|$ at the root frequencies of $P(z)$ and $Q(z)$.

As adjacent LSFs become closer, for example around 500 Hz, $|A(z)|$ decreases and hence becomes more resonant around those frequencies [2,3]. However, when a single LSF is isolated, far from its neighbors,

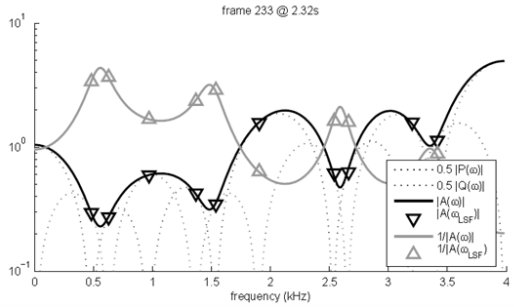

Fig. 1. Properties of LPC spectrum at line spectral frequencies.

$|P(z)|$ and $|Q(z)|$ change slowly so that $|A(z)|$ be relatively flattened. Therefore these LPC spectra at LSFs represent either spectral magnitude of speech at their formant frequencies between closely located LSFs, or background noise spectral magnitude at isolated LSFs. These properties are implicitly exploited by the proposed noise suppression procedure.

By the definition of discrete Fourier transform, the impulse response of A at frequency ω is expressed by

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

where $a_0 = 1$. The smoothed spectral magnitude at i^{th} LSF at frame t, denoted by ω_{it}, is approximated by multiplication of its LPC spectral magnitude and frame gain and expressed in log domain by

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

where G_t is gain at frame t defined in Equation 2 to represent relative magnitude difference across analysis frames. To model the global frequency characteristics of the input sounds, we approximate the long-term average of $X_t(\omega_{it})$ by the long-term average frequency response of LPC filters, denoted by $\overline{A}(z)$, is updated instantaneously by

$$\overline{A}_t(z) = (1-\alpha)\overline{A}_{t-1}(z) + \alpha A_t(z), \tag{6}$$

with an initial value $\overline{A}_0(z) = 1$. The adaptation rate $\alpha = 0.02$ gives a good performance in our experiments. The LP spectral envelope is then normalized by long-term average, and its log is approximated by the following equation:

$$\begin{aligned} \log|Y_t(\omega)| &= \log\left|\frac{X_t(\omega)}{\overline{A}_t(\omega)}\right| \\ &\approx \log G_t + \log|A_t(\omega)| - \log|\overline{A}_t(\omega)|. \end{aligned} \tag{7}$$

By using $\log|Y_t(\omega_{it})|$ instead of $X_t(\omega_{it})$, we can disregard the global shape of the noise, and a single noise estimate can be used regardless of frequencies.

The distribution of the log spectral magnitudes at LSFs is shown in Figure 2. The *x*-axis is quantized histogram intervals from the log spectral magnitude, and the *y*-axis is the number of frames whose *x*-value is in each interval. The speech spectra is from male speech, and the noise spectra is from car factory noise of SPIB database. The used sources are male speech and factory noise from Signal Processing Information Base (SPIB) database that is available at http://spib.rice.edu/. Since the spectral energy of the factory noise is near stationary over time, there is a significant peak between -2 and -1 on *x*-axis. Speech spectral magnitude is relatively scattered and varies a lot. The mixed distribution, expressed by lightly colored bars, has a peak around the one of the factory noise, and the portion of small energy components reduced a lot. This is because the noise spectra being consistent over time conceal tiny spectral magnitudes of speech signals. In the high energy regions (greater than -1.5 in Figure 2) where speech is present only, the spectral energy distribution of the speech signal dominates the mixed distribution.

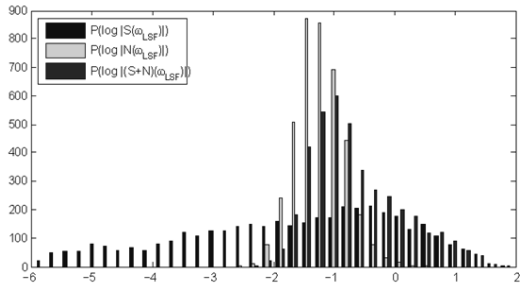

Fig. 2. Distribution of log of LPC spectrum at LSFs multiplied by excitation gain.

III. Proposed Noise Suppression Method

On the log spectral magnitudes in Equation 6 that are globally whitened, a mixture of dual Gaussian probability density functions is used to approximate mean spectral magnitude of noise. For each LSF, ω_{it}, a substitution $y_{it} = \log|Y_t(\omega_{it})|$ is taken for a compact notation. Denoting Φ_n as a set of parameters for noise Gaussian, and Φ_s for speech Gaussian, the posterior probability of y_{it} belonging to noise Gaussian is expressed by

$$P(\Phi_n|y_{it}) = \frac{P(\Phi_n)P(y_{it}|\Phi_n)}{P(\Phi_n)P(y_{it}|\Phi_n)+P(\Phi_s)P(y_{it}|\Phi_s)}, \quad (8)$$

where $P(\Phi_n)$ and $P(\Phi_s)$ are the prior probabilities of noise and speech presence, with a constraint that $P(\Phi_n)+P(\Phi_s)=1$. The likelihood of y_{it} given a set of parameters, $\Phi_n = \{\mu_{\mathrm{n}}, \sigma_{\mathrm{n}}^2\}$, is modeled by a univariate Gaussian density function:

$$P(y_{it}|\Phi_n) = \frac{1}{\sqrt{2\pi\sigma_n^2}}\exp\left[-\frac{(y_{it}-\mu_n)^2}{2\sigma_n^2}\right]. \quad (9)$$

The Gaussian parameters are updated online by the following adaptation rules:

$$\begin{aligned}\gamma_{n,it} &= \eta P(\Phi_{n,t-1}|y_{it})\\ \mu_{n,t} &= (1-\gamma_{n,it})\mu_{n,t-1}+\gamma_{n,it}y_{it}\\ \sigma_{n,t}^2 &= (1-\gamma_{n,it})\sigma_{n,t-1}^2\\ &\quad +\gamma_{n,it}(y-\mu_{n,t-1})^2\\ P(\Phi_{n,t}) &= (1-\gamma_{n,it})P(\Phi_{n,t-1})+\gamma_{n,it},\end{aligned} \quad (10)$$

where a positive constant $\eta \ll 1$ is step size, $\mu_{n,t-1}$ and $\sigma_{n,t-1}^2$ are Gaussian parameters of the previous frame. $\gamma_{n,it}$ is the computed adaptation rate for noise at frame t and i^{th} LSF. The same formula for speech can be derived by substituting Φ_n with $\Phi_s = \{\mu_{\mathrm{s}}, \sigma_{\mathrm{s}}^2\}$.

The adaptation rules for speech parameters are:

$$\begin{aligned}\gamma_{s,it} &= \eta P(\Phi_{s,t-1}|y_{it})\\ \mu_{s,t} &= (1-\gamma_{s,it})\mu_{s,t-1}+\gamma_{s,it}y_{it}\\ \sigma_{s,t}^2 &= (1-\gamma_{s,it})\sigma_{s,t-1}^2\\ &\quad +\gamma_{s,it}(y_{it}-\mu_{s,t-1})^2\\ P(\Phi_{s,t}) &= (1-\gamma_{s,it})P(\Phi_{s,t-1})+\gamma_{s,it}.\end{aligned} \quad (11)$$

From the mean of the noise Gaussian in Equation 10, $\mu_{n,t}$, noise spectral magnitude at frame t is approximated by

$$|\widetilde{N}_{t(\omega)}| = \exp[\mu_{n,t}]. \quad (12)$$

A Wiener filter suppressing the noise estimate from the spectral magnitude of the mixture signal is derived by

$$\begin{aligned}W_t(\omega) &= \frac{|Y_t(\omega)|^2-|\widetilde{N}_t(\omega)|^2}{|Y_t(\omega)|^2}\\ &= 1-\frac{|\widetilde{N}_t(\omega)|^2}{|Y_{t(\omega)}|^2}\\ &= 1-\frac{1}{\exp[2(y_t(\omega)-\mu_{n,t})]}.\end{aligned} \quad (13)$$

Then it is floored so that it should be always higher than a certain limit,

$$W_t^*(\omega) = \max\{\varepsilon, W_t(\omega)\}, \quad (14)$$

where a nonnegative constant ε is a minimum Wiener filter gain.

IV. Experimental Results

4.1 Database

The proposed method is compared to the conventional methods by automatic speech recognition performances on *speech separation challenge (SSC)* database [7]. The database is designed for assessing the effect of a noise suppression algorithm to a

simple speech recognition task. Talkers say short sentences of exactly 6 words, whose format is "*command color preposition letter number adverb*". For example, "bin blue at F 2 now". All of the original sound files are sampled at 25 kHz, and they are downsampled to 8 kHz since the EVRC and ETSI standards support 8 kHz only.

The database has a training set, which consists of 17,000 utterances (500 x 34 talkers). All training sound files are recorded in a quiet environment, i.e., without any background noise. The HMM models are obtained by HTK (hidden Markov model toolkit) [8] as suggested by the coordinators of SSC. The adopted feature is 12 MFCC plus log energy, plus their velocities and accelerations, resulting in a 39-dimensional vector at every 10 ms. A separate testing set of 600 utterances is also provided. There is no overlap between the training and the testing data. The original recordings do not contain environmental noise. Noisy data files are generated by adding speech-shaped noise (ssn), Gaussian random noise with their frequency responses modulated by the average of general speech signals. The simulated SNRs (signal-to-noise ratios) are clean (∞), 6, 0, -6, and -12 dBs, resulting in 3,000 (600 x 5) test sound files.

4.2 Implementation Details

The analysis settings of the proposed method are: sampling frequency 8 kHz, shift size 10 ms (80 samples), analysis frame length 20 ms (160 samples), and hamming windowing in LPC analysis of order 10, which results in 10 LSFs at every frame. A pre-emphasis filter defined by $1-0.97z^{-1}$ is used before the analysis to boost high-frequency energies. In re-synthesis, time-domain filters of order 48 are derived from the Wiener filters in Equation 14, applied to the input frames, and the resulting frames are overlap-added by trapezoidal windows of 24 samples overlaps between the neighboring frames. Among commercial standards, the noise suppression frontends in EVRC [9] and ETSI [11,12] standards are compared with the proposed method. They support 8 kHz sampling rate only, and mel-warped filter bank energies are used in voice activity detection and deriving noise estimates. The source codes of the two methods are publicly available by the distributors.

4.3 Illustrations of Noise Suppression Results

Figure 3 illustrates the noise spectrum estimation procedures for a mixture of male speech and factory noise. The *x*-axis is the log of spectral magnitude $Y(\omega_{it})$ at i^{th} LSF and frame t, and the *y*-axis is the normalized histogram and Wiener filter gains at the same time. The distribution of $\log|Y(\omega_{it})|$ is displayed by histogram bars, and the estimated Gaussian density functions are overdrawn on them by solid curves. The Gaussian mixture model generates a noise mean and a speech mean of log spectral magnitudes. The left one is noise Gaussian, $N(y|\mu_n,\sigma_n^2)$, and the right is speech, $N(y|\mu_s,\sigma_s^2)$, whose mean values are indicated by wedged vertical lines. The Wiener filter $W(\omega)$ obtained by Equation 13 is plotted by a thick, dashed line. Wiener filter gain (before flooring by Equation 14) is zero when $\log|Y(\omega_{it})|$ is smaller than the mean of noise Gaussian μ_n. The SNR of the input mixture is approximated by the distance between the two Gaussian means, $\mu_s-\mu_n=1.04=9.0\text{dB}$.

The distribution has a sharp peak around -1 which

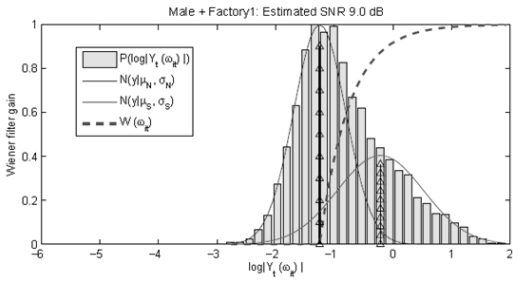

Fig. 3. Noise spectral mean and Wiener filter estimation result by mixture of Gaussian density functions, for an additive mixture of male speech and factory noise.

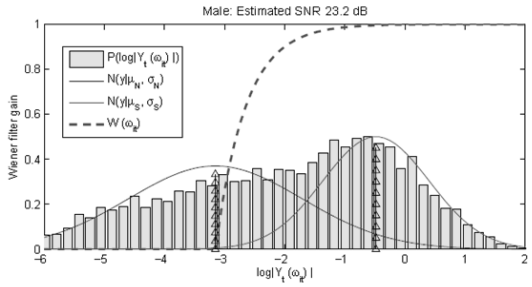

Fig. 4. Noise spectral mean and Wiener filter estimation result by mixture of Gaussian density functions for male speech only.

is well approximated by μ_n. A smaller peak is located around 0, approximated by μ_s. The spectral magnitude of speech signal varies much more than the noise, so its peak location is not as distinct as that of noise. The distribution in Figure 4, male speech only, does not have a sharp peak, and the noise Gaussian mean is around -3 with much bigger variance. The estimated SNR is $\mu_s - \mu_n = 2.67 = 23.2$ dB, which implies that the input signal is almost clean. The noise mean estimate is shifted to the left about 14 dB when compared to Figure 3, while speech Gaussian means of both Figures are very close.

4.4 Speech Recognition Performance Comparison

HMM models trained by clean training dataset is used for the experiments. The *ssn* mixtures of various SNRs are processed by EVRC standard, ETSI standard, and the proposed method. Speech recognition rates are compared in Table 1, where "bypass"' columns are the results without any processing. The proposed method significantly outperforms all the others in 6 dB and 0 dB, and slightly worse than ETSI in -6 dB and -12 dB conditions. One explanation is that the proposed method limits the minimum Wiener filter gain to -13 dB to obtain reasonable intelligibility loss. ETSI has been developed for high speech recognition performances in adverse environments [11], so it is expected to perform well in harsh noisy conditions. Although the proposed method does not have such features, the performance is not degraded in clean conditions.

Table 1. Comparison of speech recognition performances on the testing set with additive speech-shaped Gaussian noise (*ssn*).

Methods	clean	6 dB	0 dB	-6 dB	-12 dB
Bypass	97.6 %	60.7 %	27.1 %	12.8 %	11.2 %
EVRC	96.7 %	68.3 %	32.1 %	14.4 %	12.2 %
ETSI	96.9 %	76.4 %	43.4 %	21.6 %	14.1 %
Proposed	97.7 %	81.1 %	49.9 %	21.3 %	11.8 %

Since *ssn* is an artificial noise, a number of real noise cases are evaluated as well. From AURORA2 database [12], 8 different noise sources (airport, babble, car, exhibition, restaurant, street, subway, and train) are chosen, and added to clean test files. The measured speech recognition rates are in Table 2. Noise mixing levels are 12 dB, 6 dB, and 0 dB, and used as column indexes. Clean condition results are the same as Table 1, and negative SNRs are not considered since the speech recognition rates are too low to be meaningful. Row indexes are various noise suppression methods: bypass (no processing), EVRC standard, ETSI standard, and the proposed method. The last row summarizes speech recognition rates averaged over 8 noises. The top 2 highest values are boldfaced in each mixing SNR.

In terms of average recognition rates, the proposed method is always of higher recognition rates than the other methods with 6 dB and 0 dB SNR mixtures, by up to 7.7 %. In 12 dB SNR, the improvement over ETSI is about 3 %~4 %, and EVRC is the best but the difference to the proposed is only 0.1 %. The proposed method is always within the top 2 in all 3 SNR conditions. EVRC works well with relatively higher SNRs (12 dB and 6 dB), and ETSI is better suited to lower SNR cases (0 dB). However, the proposed method guarantees decent speech recognition performance in all noise levels. In terms of noise types, the proposed method significantly improves the perfor-

Table 2. Comparison of speech recognition performances on AURORA2 database.

Noise	Methods	12 dB	6 dB	0 dB
airport	bypass	85.1 %	65.2 %	37.6 %
	EVRC	89.4 %	69.0 %	35.8 %
	ETSI	86.9 %	68.4 %	39.3 %
	Proposed	90.3 %	74.7 %	48.1 %
babble	bypass	81.4 %	56.6 %	28.6 %
	EVRC	88.4 %	65.1 %	34.3 %
	ETSI	85.0 %	63.2 %	34.6 %
	Proposed	87.6 %	68.5 %	39.2 %
car	bypass	79.4 %	54.4 %	22.3 %
	EVRC	88.3 %	60.8 %	24.6 %
	ETSI	85.0 %	61.1 %	30.8 %
	Proposed	89.1 %	72.8 %	42.4 %
exhibition	bypass	73.7 %	41.9 %	18.4 %
	EVRC	83.4 %	64.8 %	31.0 %
	ETSI	82.1 %	59.6 %	30.2 %
	Proposed	83.7 %	60.5 %	30.7 %
restaurant	bypass	83.3 %	58.3 %	29.1 %
	EVRC	87.9 %	65.5 %	33.6 %
	ETSI	85.7 %	63.9 %	35.4 %
	Proposed	86.0 %	67.9 %	37.3 %
street	bypass	77.1 %	52.8 %	26.4 %
	EVRC	85.3 %	61.2 %	29.9 %
	ETSI	80.8 %	58.9 %	30.5 %
	Proposed	85.1 %	65.0 %	37.7 %
subway	bypass	76.8 %	43.9 %	19.1 %
	EVRC	85.6 %	65.9 %	32.3 %
	ETSI	81.3 %	60.1 %	30.6 %
	Proposed	85.1 %	63.0 %	31.1 %
train	bypass	86.9 %	67.1 %	38.5 %
	EVRC	89.7 %	67.9 %	35.6 %
	ETSI	87.9 %	71.7 %	41.5 %
	Proposed	90.7 %	77.4 %	51.6 %
Average	bypass	80.5 %	55.0 %	27.5 %
	EVRC	87.3 %	65.0 %	32.1 %
	ETSI	84.3 %	63.4 %	34.1 %
	Proposed	87.2 %	68.7 %	39.8 %

mance with airport, car, street, and train noises, about the same performance as ETSI's and EVRC's with babble and restaurant noises, and EVRC is slightly better with exhibition and subway noises, but the difference is not that large. In summary, the speech recognition results prove that the proposed method is quite stable, and much better than the conventional methods with various noise types and various noise levels.

V. Concluding Remarks

A novel method to reduce near-stationary acoustic noise added to a speech signal recorded by a single microphone is proposed. The noise spectral magnitude is estimated by the smaller mean of dual Gaussian mixture distributions for globally flattened LPC spectra at line spectral frequencies, and a Wiener filter suppressing the estimated noise is derived and applied to the input speech signal. The proposed method has several advantages.

Improved noise suppression performance: It suppresses additive noise with much less speech recognition performance degradations when compared to the conventional methods, proved by automatic speech recognition experiments on simulated mixtures of clean speech signals and diverse kinds of real noises in various SNRs. The characteristics of LPC spectral envelopes at line spectral frequencies are actively exploited to estimate noise spectral magnitude, and the need for voice activity detection has been eliminated.

Computational efficiency: ETSI and EVRC standards use fixed filter bank energies, and there are a lot of control parameters for reliable voice activity detection. The proposed algorithm does not use fixed filter banks, but variable filter banks are chosen according to line spectral frequencies. In 8 kHz sampling frequency, the number of filter banks are 23 for EVRC and 16 for ETSI, but the number of line spectral frequencies is only 10, which makes the proposed method much more computationally efficient.

Good match to voice coders: The algorithm is quite simple and requires only LPC coefficients, line spectral frequencies, and excitation gains, which are all available in LPC-based voice coders such as

QCELP. Being embedded with voice coders, it can be implemented much more efficiently.

Acknowledgment

This research was supported by the Ministry of Knowledge Economy, Korea (2008-S-019-02, Development of Portable Korean-English Automatic Speech Translation Technology), and by the 2010 Research Fund of the UNIST (Ulsan National Institute of Science and Technology).

Reference

1. S. F. 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, 1979.
2. Mi Suk Lee, Hong Kook Kim, Seung Ho Choi, and Hwang Soo Lee, "On the use of LSF intermodel interlacing property for spectral quantization," in *Proc. 1999 IEEE Workshop on Speech Coding*, June 1999, pp. 43-45.
3. Mi Suk Lee, Hong Kook Kim, and Hwang Soo 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, October 2001.
4. Peter Kabal and Ravi Prakash Ramachandran, "The computation of line spectral frequencies using chebyshev polynomials," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 34, no. 6, pp. 1419-1426, December 1986.
5. A. Kindoz and Ahmet M. Kondoz, Digital Speech; Coding for Low Bit Rate Communication Systems, *John Wiley & Sons, Inc.*, New York, USA, 1994.
6. Tom Bäckström and Carlo Magi, "Properties of line spectrum pair polynomials -- a review," *Signal Processing*, vol. 86, pp. 3286-3298, 2006.
7. Martin Cooke and Te-Won Lee, "Speech separation challenge," *INTERSPEECH*, 2006.
8. Steve Young, Gunnar Evermann, , Mark Gales, Thomas Hain, Dan Kershaw, Xun-ying (Andrew) Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev, and Phil Woodland, "Hidden Markov model toolkit (HTK) version 3.4," December 2006.
9. Telecommunications Industry Association (TIA), "Enhanced variable rate codec, speech service option 3 for wideband spread spectrum digital systems," Tech. Rep., TIA/EIA/IS-127, September 1996.
10. 3rd Generation Partnership Project, "Amr speech codec," Tech. Rep., 3GPP TS 26.071, V6.0.0, December 2004.
11. European Telecommunications Standards Institute (ETSI), "Speech processing, transmission and quality aspects (stq); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithm," Tech. Rep., ES 202 050 v1.1.1, October 2002.
12. David Pearce and Hans-Günter Hirsch, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy condition," in *Proc. ICSLP*, Beijing, China, October 2000.

Profile

▸ Gil-Jin Jang

Dr. Gil-Jin Jang is an assistant professor at Ulsan National Institute of Science and Technology (UNIST), South Korea. He received his B.S., M.S., and Ph.D. degrees in computer science from the Korea Advanced Institute of Science and Technology (KAIST), Daejon, South Korea in 1997, 1999, and 2004, respectively. From 2004 to 2006 he was a research staff at Samsung Advanced Institute of Technology and from 2006 to 2007 he worked as a research engineer at Softmax, Inc. in San Diego. From 2008 to 2009 he worked at Hamilton Glaucoma center at University of California, San Diego as a postdoctoral employee. His research interests include acoustic signal processing, pattern recognition, speech recognition and enhancement, and biomedical signal engineering.

▸ Jeong-Sik Park

Dr. Jeong-Sik Park received his B.E. degree in Computer Science from Ajou University, South Korea in 2001 and his M.E. and Ph.D. degree in Computer Science from KAIST (Korea Advanced Institute of Science and Technology) in 2003 and 2010, respectively. From 2010 to 2011, he was a Postdoctoral researcher in the Computer Science Department, KAIST. He is now an assistant professor in the Department of Intelligent Robot Engineering, Mokwon University. His research interests include speech emotion recognition, speech recognition, speech enhancement, and voice interface for human-computer interaction.

▸ Sanghun Kim

Dr. Sanghun Kim received the BS in electrical engineering from Yonsei University, Seoul, Korea, in 1990, and the MS degree in electrical and electronic engineering from KAIST, Daejeon, Korea, in 1992. He received his PhD from the Department of Electrical, Electronic, Information, and Communication Engineering at the University of Tokyo, Japan, in 2003. Since 1992, he has been with the Research Department of Spoken Language Processing Section of ETRI, Daejeon, Korea. Currently, he is a principal researcher in the Automatic Speech Translation Research Team. His interests include speech synthesis, speech recognition, and speech signal processing.